BREAST CANCER PREDICTION USING MACHINE LEARNING APPROCHES

BY

MST. RUKAIA FERDOUS

ID: 193-25-834

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

Supervised By
Dr. Md. Zahid Hasan
Associate Professor and Program Director of MIS
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH

17, January 2023

APPROVAL

This Theses titled “Breast Cancer Prediction using machine learning”, submitted by Mst. Rukaia Ferdous to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 17th January, 2023.

BOARD OF EXAMINERS

Dr. Rashed Haider Noori
Professor and Associate Head **Chairman**
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Md. Sadekur Rahman
Assistant Professor **Internal Examiner**
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Dr. Mohammad Shorif Uddin
Professor **External Examiner**
Department of CSE
Jahangirnagar University

DECLARATION

I hereby declare that, this project has been done by us under the supervision of **Md Zahid Hasan, Associate Professor & Program Director MIS, Department of CSE,** Daffodil International University**,** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree.

Supervised by:

Dr. Md. Zahid Hasan
Associate Professor & Program Director MIS
Department of CSE
Daffodil International University

Submitted by:

Rukaia Ferdous

Mst. Rukaia Ferdous
ID: 193-25-834
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I express my heartiest thanks and gratefulness to almighty God for His divine blessing makes me possible to complete the final year project successfully.

I really grateful and wish my profound my indebtedness to of **Dr. Md Zahid Hasan, Associate Professor & Program Director MIS, Department of CSE,** Daffodil International University, Dhaka. Deep Knowledge & keen interest of my supervisor in the field of web development influenced me to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

I would like to express my heartiest gratitude to Dr. Touhid Bhuiyan, Head, Department of CSE, Daffodil International University, Dhaka, for his kind help to finish my project and also to other faculty member and the staff of CSE department of Daffodil International University.

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

Finally, I must acknowledge with due respect the constant support and patience of my parents.

ABSTRACT

Breast Cancer is used as a kind of death every year. It's the most considered normal kind of disease in ladies and the main smite of death all around the world. Quite possibly of the most significant and fundamental assignment in AI and information mining is arrangement. A significant measure of exploration has been directed to characterize Breast Disease utilizing information mining and AI on various clinical datasets. There are a few snags along the way. After highlighting the dataset and creating a data frame, I employ a variety of machine learning classifiers. This study gives an overview of the opinion assessment challenges that am relevant to their approaches and strategies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF Tables

CHAPTER 1

INTRODUCTION

1.1 Introduction

The Breast Cancer is a sort of disease that begins in the human breast. It can begin in one or the two breasts of human. Disease creates when cells start to wildly multiply. The breast cancer fundamentally influences ladies; However, men can be impacted also. Breast tumors that am not malignant am strange developments that don't spread past the rest. Albeit not deadly, some harmless breast protuberances can raise a lady's gamble of creating breast cancer or cancer. A medical cam proficient ought to assess any breast protuberance or change to decide if it is harmless or dangerous (cancer) and whether it might influence ymy future disease risk. Breast cancer can foster in various amas inside the breast. The breast is an organ that is arranged over the upper ribs and chest muscles. Each side has two breasts, each with organs, channels, and greasy tissue. Ladies' breasts produce and convey milk to take cam of babies and newborn children. The size of the not entirely set in stone by how much greasy tissue present. Cancer cells spread to different pieces of the body when they enter the blood or lymph framework. The lymph (or lymphatic) framework is a piece of my body's insusceptible framework. It is an organization of lymph nodes (little bean-sized organs), pipes or vessels, and organs that cooperate to gather and move clear lymph liquid from body tissues to the circulatory system. The reasonable lymph liquid held inside the lymph vessels contains tissue side-effects and waste material, as Ill as invulnerable framework cells. Lymphatic vessels am liable for shipping lymph liquid away from the breast. On account of breast disease, cancer cells can enter lymph vessels and start to fill in lymph nodes. Breast cancer is the most widely recognized cancer in ladies, as Ill as the most Ill-known cancer around the world. More than 2.26 million new instances of breast cancer in ladies will be analyzed in 2020. Belgium had the most noteworthy level of breast disease in ladies in 2020. Barbados had the most noteworthy measure of female breast cancer mortality in 2020. Taking cam of breast cancer is unquestionably sensible, and early discovery is the underlying stage I can take, with AI filling in as the establishment. This' study will probably make an AI model that can foresee the probability of creating breast cancer.

1.2 Motivation

Modern technology is causing rapid change and is becoming more widely available in my country; it is being used in a wide range of industries. Smartphones enable us to access a wide range of services. Every year, the number of people diagnosed with cancer rises rapidly. Most people find out they have cancer at the very end because it doesn't have very common symptoms, people can’t away of it, am ignorant of it, and breast cancer testing isn't something common people do. Breast cancer is as yet the most recognizable and lethal disease among ladies in Bangladesh. It has turned into a secret Light, representing 69% of female passing’s. Breast disease is assessed to happen at a pace of 22.5 per 100,000 females of any age in Bangladesh; among Bangladeshi ladies matured 15-44 years, breast cancer has the most noteworthy commonness (19.3 per 100,000) contrasted with some other kind of cancer. Breast cancer treatment cost in my country is increasing day by day. However, the vast majority of my population is impoverished. As a result, detection is an important step to take in terms of both health and economic prospects. This is why I decided to create a Machine Learning (ML)-based model which can predict if there is a patient is likely to have breast cancer or not.

1.3 Rationale of the Study

Breast cancer research makes the way for better strategies for forestalling, recognizing, and treating breast cancer, as Ill as working on the personal satisfaction for cancer patients and survivors. Here am a portion of the primary am as of concentration in breast cancer research today, going from concentrating on makes and counteraction figuring out how to oversee and treat the illness. This examination makes a decent effect of human existence.

1.4 Expected Output

Following data collection, I pre-process the data and information. Then, at that point, I apply the algorithm I picked. For anticipating Breast Cancer, a couple of the relevant algorithms might give the best outcomes. The goal of this work is to illustrate the highest level of accuracy.

1.5 Management and Finance

Despite being timely and focused, this thesis is not financially supported or funded. Nonetheless, I might want to offer my ardent thanks to Daffodil International University for the entirety of their help.

1.6 Report Layout

This section offers insight into each of the six chapters that make up this report.

- Chapter 1, contains the introduction, motivation, and expected outcome of the study.
- Chapter 2, related study is discussed. It also gives Questions a field of study.
- Chapter 3, Data collection methodology and overall methodology.
- Chapter 4, describes my proposed system design.
- Chapter 5, contains the ethical aspects of the research and also about the social impact of the research.
- Chapter 6, provides the conclusion, future study, references.

CHAPTER 2

BACKGROUND

2.1 Introduction

Breast cancer is one of the most serious diseases that has been rapidly increasing in women over the last two decades. Many people use various algorithms to predict breast cancer.

2.2 Related Works

In The training process needs access to a representative dataset, which is essential. The robustness of the training performance will be improved by using more datasets to train the CNN model. It's also important to balance the amount of training data for the designated classes. In this study, they sought to synthesize the state of the art and advancements in human breast cancer checking out using thermography and CNNs. An overview of the publicly available breast thermal datasets is given after a discussion of the possibility of Breast Thermography in the very beginning identification of breast cancer. Talk about the similarities and differences betIen normal and malignant thermographic patterns. A simulated example illustrating feature learning is used to explain the classification of breast thermograms using a CNN model.

In This paper introduces they have planned an WBC (Wisconsin Breast Cancer) data was used to develop an interactive outfit voting scheme for breast cancer. My goal is to identify and explain how CNN and logistic algorithms can be used to recognize breast cancer with condensed variables. There am two varieties of tumors that can still be found here. There am two sorts of growths: benign and malignant, in which benign assumes non-cancerous and malicious means cancerous.

In this paper, the researchers presented a deep learning method that incorporates perfectly straight data to enhance breast cancer detection from the mammogram images. They discovered that when compamd to the baseline architecture, their proposed reduces FP predictions. At the candidate level, an AUC of 0.933 (p = 0.111) with a 95% confidence interval of [0.919, 0.947] was obtained, and my symmetry model achieved a CPM of 0.733 (p = 0.001) with a 95% confidence interval of [0.721, 0.823].

In this study compams Random Forest, kNN (k-Neamst-Neighbor), and Nave Bayes, three extensively used machine learning algorithms and methodologies for breast cancer prediction. The Wisconsin Diagnosis the Breast Cancer data set was used as a training set to compam the performance of different machine learning (ML) techniques on important metrics including accuracy and precision. Each algorithm was shown to be more than 94% accurate at identifying benign versus malignant tumors.

In his article portrays This dataset was exposed to information representation and AI strategies, for example, calculated relapse, k-nearst neighbors, support vector machine, credulous Bayes, decision tree, irregular backwoods, and revolution woods. R, and Python Ire decided to be utilized for AI and perception. The reason for the work was to lead a relative examination of breast tumor growth location and determination utilizing information representation and AI applications. For recognizing breast cancer, the indicative presentation of uses was similar. Information representation and AI procedures can altogether affect disease recognition and independent direction. A few AI and information digging methods for identifying breast cancer Ire make sense of in this paper. They had the most elevated exactness in order (98. 1%).In [6] Support Vector Machine (SVM), Random Forest (RF), and Bayesian Networks, three of the most popular machine learning (ML) algorithms for breast cancer detection and diagnosis, Ire presented in this research. The main characteristics methods of each of the three ML techniques Ire discussed. The effectiveness of the explored strategies was evaluated using the Original Wisconsin Breast Cancer Data collection. The results show that SVMs have the highest precision, specificity, and accuracy. The highest probability of classifying tumors correctly is with RFs.

In this study, I used fmy key algorithms on the Wisconsin Breast Cancer (original) datasets: SVM, NB, k-NN, and C4.5. To condition which algorithm got the best classification accuracy, I attempted to assess the efficiency and efficacy of various algorithms in terms of accuracy, precision, sensitivity, and specificity. SVM surpasses all other methods and has an accuracy of 97.13%. These all result of the experiment have the greatest degree of accuracy (97.28%) in categorizing dataset when comes to data mining method classification accuracies.

2.4 Scope of the Problem

I can conclude from the preceding discussion that some predictions am not very accurate, that some only use one or two measurement algorithms with few attributes, and that there am issues with the data set. The goal of this study was to foresee the occurrence of human breast cancer and determine in the best way to treat it.

2.5 Challenges

- Collecting data
- Data preprocessing
- Good quality of resource
- Implement current Colab Library
- The chapter structure and lowest am insufficient.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

I will discuss about entire study methodology in this section. Each analysis has a unique approach to solve it. First, I gathered the data from the Wisconsin breast cancer dataset from Kaggle. After creating the dataset and processing the data, I must eliminate certain all null value and all column which am not relevant to this research. Next, I choose the Machine learning (ML) algorithm that using. I've previously indicated that I'm employing three distinct machine learning methods, so in order to build the model and fit the algorithm, I must first create a data set. This data set is then used to train the model. On such foundation, feature selection is operating. After then, the information or data was partitioned into a training set and a testing set. It is now and again alluded to as the test informational collection and the train informational index. Then, subsequent to squeezing the information into different AI algorithm models and preparing the information with a training data set, I just get a critical piece of the information expected to test my model. The model is then evaluated to decide its accuracy. I can tell if someone is at risk of developing breast cancer with some of this level of precision. To provide an overview, I've have been using my standard process flow chart but to further understand how some of the algorithms operate, I'll go through them in general with equations and diagrams. Below is the process for the full research project, which gives a quick overview of the complete study project.

Subject of study and research components:

First of all, talk about the philosophical and theoretical aspects of how breast cancer is discovered. Machine learning models require a performance computer which is with a GPU and all other tool. Here is a list of the instruments needed for this research model.

Software and hardware:

- Intel Core i7 10th generation
- 1 TB Hard Disk Drive
- 4GB RAM

Development Tools:

- Python
- Pandas
- NumPy
- Matplotlib
- Seaborn
- Scikit-Learn

3.2 Preprocessing and Data Collection

I obtained the data set from Kaggle [15] and made a few minor changes in accordance with my specifications Each process must then use the dataset. Since each of the three algorithms I use operates differently, as I have already said, I must
To correctly fit the data set in my model, preprocess the data.

3.3 Statistical Analysis

1. There am 31 columns in total.
2.There am 570 rows in total.
3.70% of the data in my model was used to train it, while 30% was used for testing.
4.The dataset is saved as a csv file.

3.4 Proposed Methodology

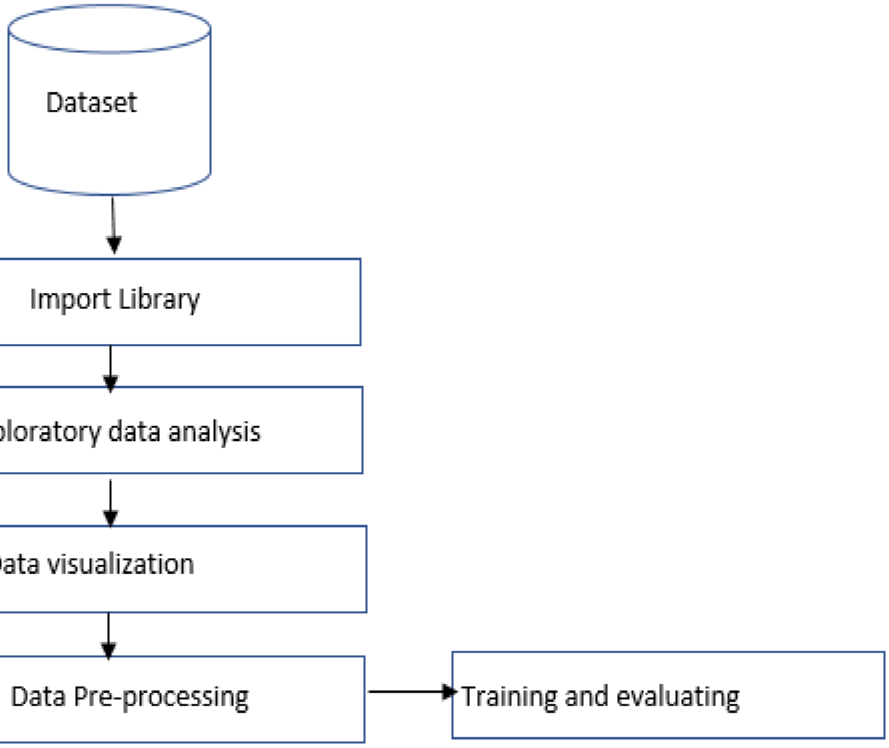


Fig 3.1: proposed methodology work flow

3.4.1 Dataset

Suggested, I have stated gathered data from many syces, yet various researchers use various types of data sets. That reason I prefer from utilizing one data set (information set) for many algorithms to identify or forecast whether each patient suffering from breast cancer disease or not, I must awfully select the data set. Because of this, I should aimfully select the dataset and properly pre - process it.

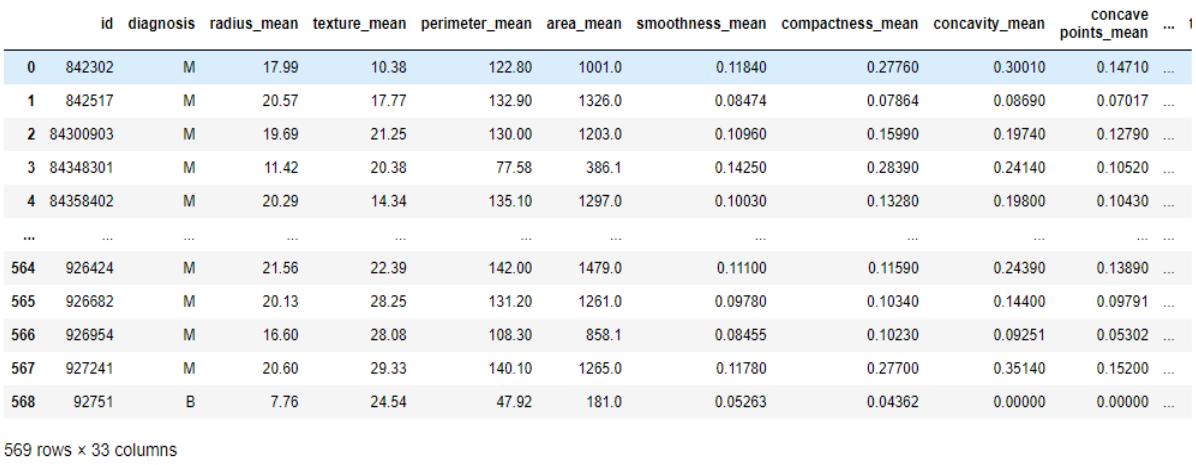
3.4.2 Import Library

In this project, here using many libraries, like as

- Data analysis library
- Data Visualization
- Machine learning packages
- Model training and evolution
- Some machine learning which am using algorithm

3.4.3 Exploratory Data analysis

Firstly, loading dataset and using many command lines for exploratory data analysis,31 columns and 570 rows. There am no null values in my data set I am checking for null values.



	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.30010	0.14710	...
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.08690	0.07017	...
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.19740	0.12790	...
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.24140	0.10520	...
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.19800	0.10430	...
...	...	...	...	...	...	...	...	...	...	...	...
564	926424	M	21.56	22.39	142.00	1479.0	0.11100	0.11590	0.24390	0.13890	...
565	926682	M	20.13	28.25	131.20	1261.0	0.09780	0.10340	0.14400	0.09791	...
566	926954	M	16.60	28.08	108.30	858.1	0.08455	0.10230	0.09251	0.05302	...
567	927241	M	20.60	29.33	140.10	1265.0	0.11780	0.27700	0.35140	0.15200	...
568	92751	B	7.76	24.54	47.92	181.0	0.05263	0.04362	0.00000	0.00000	...

569 rows × 33 columns

Fig 3.2: showing dataset

3.4.4 Data visualization

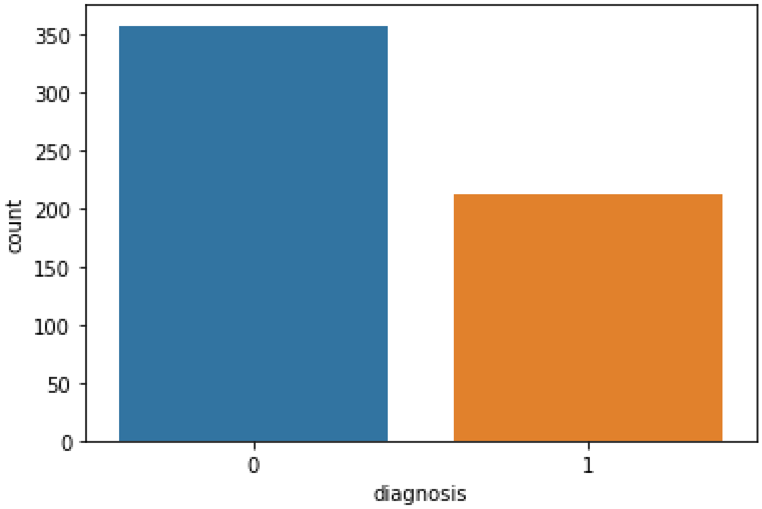


Fig 3.3: How many malignant or benign

357 people have not yet received a cancer diagnosis. 212 people have received a cancer diagnosis.

3.4.5 Data Pre-processing

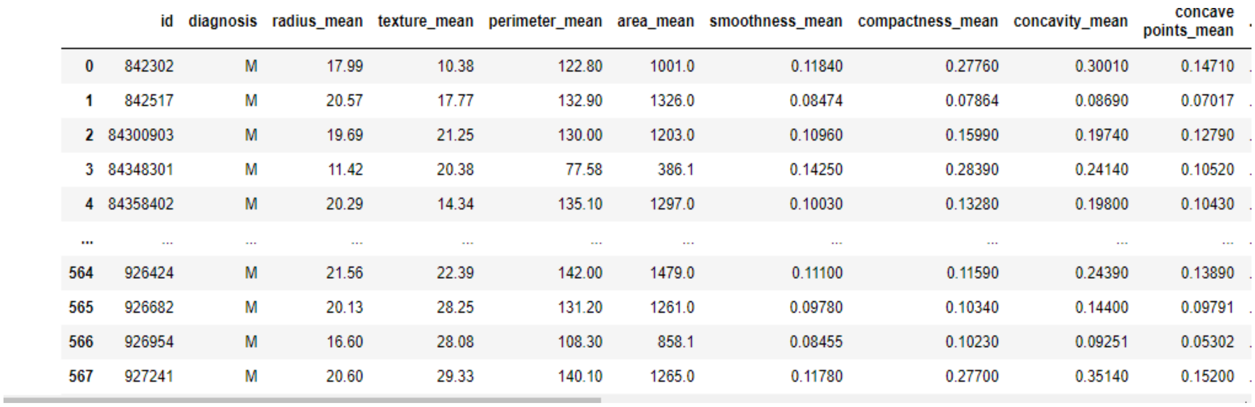
I have stated that I have gathered data or information from many source, still various many people am researching by using various types of data sets. Because I prefer from utilizing one data set many algorithms to identify or find out whether each patient has breast cancer or not, I must awfully select the data set. Because of this, I must comfily choose the dataset and correctly preprocess the data set.

Remove unnecessary columns:

A key step in data preparation is the elimination of pointless columns. I must determine whether or not there is a null result or value. Since the null value actually lowers accuracy, getting rid of it will improve my results since doing so will allow us to properly train data set and produce more accurate results. Because not all of the columns am essential for my model to fit, some of them must be removed.

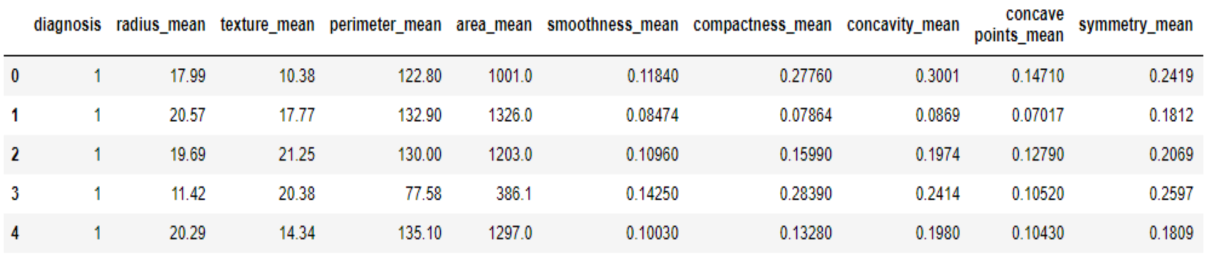
Remove null values and Replace values:

As I've already said, it's crucial to eliminate null values from data sets because if I don't do so, the algorithms I use to train the system will be improperly trained as a result of the null values. And occasionally the algorithm's model will produce less accurate results for these null values. Sometimes data can be entered incorrectly or missing entirely, thus such kinds of problems should be eliminated. When testing the algorithm on this dataset, I may observe the improved accuracy if I correctly eliminate all the null values before starting the training phase. In my data collection, the column name diagnostic has an attribute with two types of levels, one of which is "M":1, "B":0." Therefore, I must change the M values to 1 and the B values to 0.



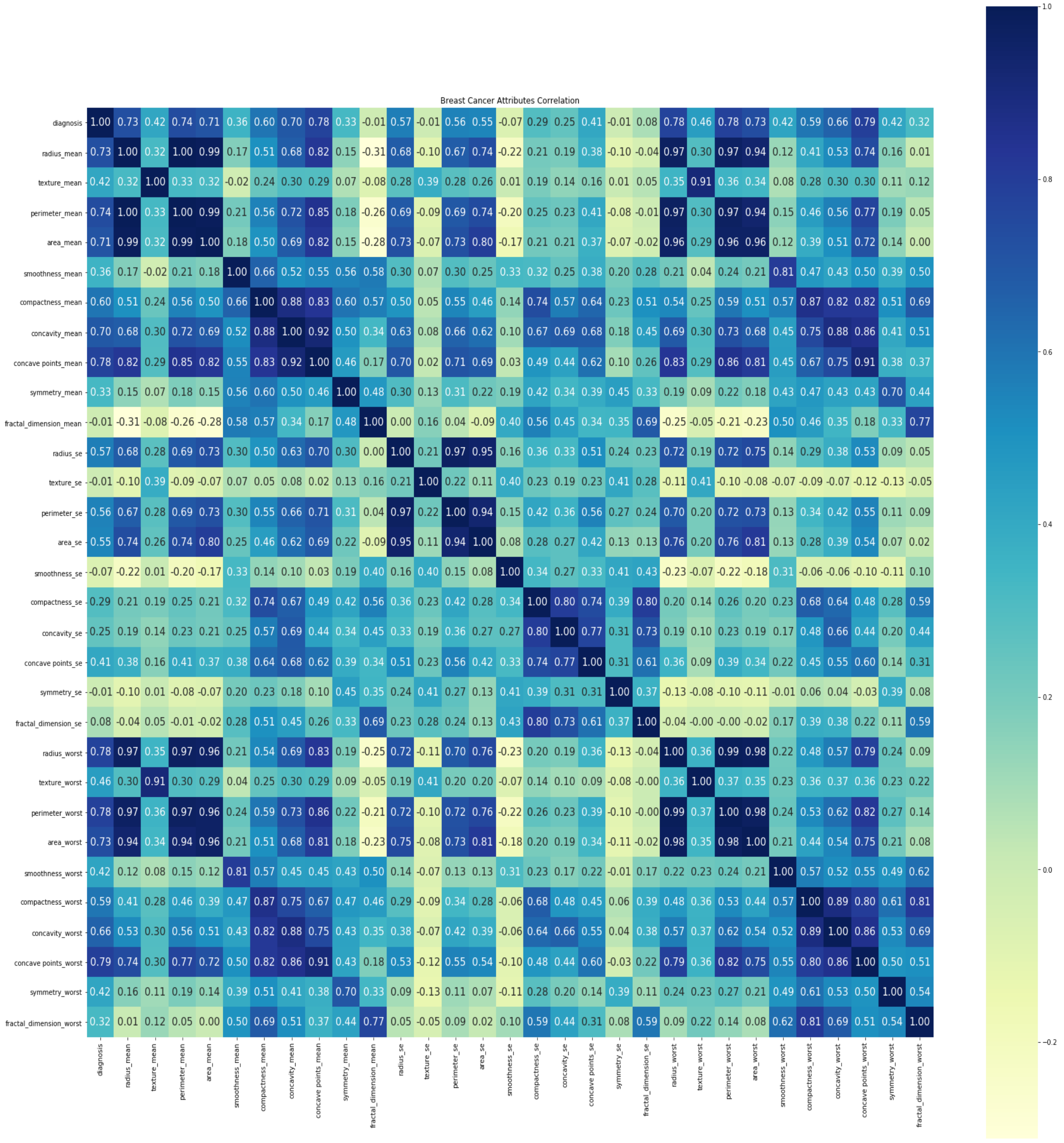
	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	.
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.30010	0.14710	.
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.08690	0.07017	.
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.19740	0.12790	.
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.24140	0.10520	.
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.19800	0.10430	.
...	...	...	...	...	...	...	...	...	...	...	.
564	926424	M	21.56	22.39	142.00	1479.0	0.11100	0.11590	0.24390	0.13890	.
565	926682	M	20.13	28.25	131.20	1261.0	0.09780	0.10340	0.14400	0.09791	.
566	926954	M	16.60	28.08	108.30	858.1	0.08455	0.10230	0.09251	0.05302	.
567	927241	M	20.60	29.33	140.10	1265.0	0.11780	0.27700	0.35140	0.15200	.

Fig 3.4: Before replacing values



	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean
0	1	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419
1	1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812
2	1	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069
3	1	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597
4	1	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809

Fig 3.5: After replacing values



	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fractal_dimension_mean	radius_se	texture_se	perimeter_se	area_se	smoothness_se	compactness_se	concavity_se	concave points_se	symmetry_se	fractal_dimension_se	radius_worst	texture_worst	perimeter_worst	area_worst	smoothness_worst	compactness_worst	concavity_worst	concave points_worst	symmetry_worst	fractal_dimension_worst
diagnosis	1.00	0.73	0.42	0.74	0.71	0.36	0.60	0.70	0.78	0.33	-0.01	0.57	-0.01	0.56	0.55	-0.07	0.29	0.25	0.41	-0.01	0.08	0.78	0.46	0.78	0.73	0.42	0.59	0.66	0.79	0.42	0.32
radius_mean	0.73	1.00	0.32	1.00	0.99	0.17	0.51	0.68	0.82	0.15	-0.31	0.68	-0.10	0.67	0.74	-0.22	0.21	0.19	0.38	-0.10	-0.04	0.97	0.30	0.97	0.94	0.12	0.41	0.53	0.74	0.16	0.01
texture_mean	0.42	0.32	1.00	0.33	0.32	-0.02	0.24	0.30	0.29	0.07	-0.08	0.28	0.39	0.28	0.26	0.01	0.19	0.14	0.16	0.01	0.05	0.35	0.91	0.36	0.34	0.08	0.28	0.30	0.30	0.11	0.12
perimeter_mean	0.74	1.00	0.33	1.00	0.99	0.21	0.56	0.72	0.85	0.18	-0.26	0.69	-0.09	0.69	0.74	-0.20	0.25	0.23	0.41	-0.08	-0.01	0.97	0.30	0.97	0.94	0.15	0.46	0.56	0.77	0.19	0.05
area_mean	0.71	0.99	0.32	0.99	1.00	0.18	0.50	0.69	0.82	0.15	-0.28	0.73	-0.07	0.73	0.80	-0.17	0.21	0.21	0.37	-0.07	-0.02	0.96	0.29	0.96	0.96	0.12	0.39	0.51	0.72	0.14	0.00
smoothness_mean	0.36	0.17	-0.02	0.21	0.18	1.00	0.66	0.52	0.55	0.56	0.58	0.30	0.07	0.30	0.25	0.33	0.32	0.25	0.38	0.20	0.28	0.21	0.04	0.24	0.21	0.81	0.47	0.43	0.50	0.39	0.50
compactness_mean	0.60	0.51	0.24	0.56	0.50	0.66	1.00	0.88	0.83	0.60	0.57	0.50	0.05	0.55	0.46	0.14	0.74	0.57	0.64	0.23	0.51	0.54	0.25	0.59	0.51	0.57	0.87	0.82	0.82	0.51	0.69
concavity_mean	0.70	0.68	0.30	0.72	0.69	0.52	0.88	1.00	0.92	0.50	0.34	0.63	0.08	0.66	0.62	0.10	0.67	0.69	0.68	0.18	0.45	0.69	0.30	0.73	0.68	0.45	0.75	0.88	0.86	0.41	0.51
concave points_mean	0.78	0.82	0.29	0.85	0.82	0.55	0.83	0.92	1.00	0.46	0.17	0.70	0.02	0.71	0.69	0.03	0.49	0.44	0.62	0.10	0.26	0.83	0.29	0.86	0.81	0.45	0.67	0.75	0.91	0.38	0.37
symmetry_mean	0.33	0.15	0.07	0.18	0.15	0.56	0.60	0.50	0.46	1.00	0.48	0.30	0.13	0.31	0.22	0.19	0.42	0.34	0.39	0.45	0.33	0.19	0.09	0.22	0.18	0.43	0.47	0.43	0.43	0.70	0.44
fractal_dimension_mean	-0.01	-0.31	-0.08	-0.26	-0.28	0.58	0.57	0.34	0.17	0.48	1.00	0.00	0.16	0.04	-0.09	0.40	0.56	0.45	0.34	0.35	0.69	-0.25	-0.05	-0.21	-0.23	0.50	0.46	0.35	0.18	0.33	0.77
radius_se	0.57	0.68	0.28	0.69	0.73	0.30	0.50	0.63	0.70	0.30	0.00	1.00	0.21	0.97	0.95	0.16	0.36	0.33	0.51	0.24	0.23	0.72	0.19	0.72	0.75	0.14	0.29	0.38	0.53	0.09	0.05
texture_se	-0.01	-0.10	0.39	-0.09	-0.07	0.07	0.05	0.08	0.02	0.13	0.16	0.21	1.00	0.22	0.11	0.40	0.23	0.19	0.23	0.41	0.28	-0.11	0.41	-0.10	-0.08	-0.07	-0.09	-0.07	-0.12	-0.13	-0.05
perimeter_se	0.56	0.67	0.28	0.69	0.73	0.30	0.55	0.66	0.71	0.31	0.04	0.97	0.22	1.00	0.94	0.15	0.42	0.36	0.56	0.27	0.24	0.70	0.20	0.72	0.73	0.13	0.34	0.42	0.55	0.11	0.09
area_se	0.55	0.74	0.26	0.74	0.80	0.25	0.46	0.62	0.69	0.22	-0.09	0.95	0.11	0.94	1.00	0.08	0.28	0.27	0.42	0.13	0.13	0.76	0.20	0.76	0.81	0.13	0.28	0.39	0.54	0.07	0.02
smoothness_se	-0.07	-0.22	0.01	-0.20	-0.17	0.33	0.14	0.10	0.03	0.19	0.40	0.16	0.40	0.15	0.08	1.00	0.34	0.27	0.33	0.41	0.43	-0.23	-0.07	-0.22	-0.18	0.31	-0.06	-0.06	-0.10	-0.11	0.10
compactness_se	0.29	0.21	0.19	0.25	0.21	0.32	0.74	0.67	0.49	0.42	0.56	0.36	0.23	0.42	0.28	0.34	1.00	0.80	0.74	0.39	0.80	0.20	0.14	0.26	0.20	0.23	0.68	0.64	0.48	0.28	0.59
concavity_se	0.25	0.19	0.14	0.23	0.21	0.25	0.57	0.69	0.44	0.34	0.45	0.33	0.19	0.36	0.27	0.27	0.80	1.00	0.77	0.31	0.73	0.19	0.10	0.23	0.19	0.17	0.48	0.66	0.44	0.20	0.44
concave points_se	0.41	0.38	0.16	0.41	0.37	0.38	0.64	0.68	0.62	0.39	0.34	0.51	0.23	0.56	0.42	0.33	0.74	0.77	1.00	0.31	0.61	0.36	0.09	0.39	0.34	0.22	0.45	0.55	0.60	0.14	0.31
symmetry_se	-0.01	-0.10	0.01	-0.08	-0.07	0.20	0.23	0.18	0.10	0.45	0.35	0.24	0.41	0.27	0.13	0.41	0.39	0.31	0.31	1.00	0.37	-0.13	-0.08	-0.10	-0.11	-0.01	0.06	0.04	-0.03	0.39	0.08
fractal_dimension_se	0.08	-0.04	0.05	-0.01	-0.02	0.28	0.51	0.45	0.26	0.33	0.69	0.23	0.28	0.24	0.13	0.43	0.80	0.73	0.61	0.37	1.00	-0.04	-0.00	-0.00	-0.02	0.17	0.39	0.38	0.22	0.11	0.59
radius_worst	0.78	0.97	0.35	0.97	0.96	0.21	0.54	0.69	0.83	0.19	-0.25	0.72	-0.11	0.70	0.76	-0.23	0.20	0.19	0.36	-0.13	-0.04	1.00	0.36	0.99	0.98	0.22	0.48	0.57	0.79	0.24	0.09
texture_worst	0.46	0.30	0.91	0.30	0.29	0.04	0.25	0.30	0.29	0.09	-0.05	0.19	0.41	0.20	0.20	-0.07	0.14	0.10	0.09	-0.08	-0.00	0.36	1.00	0.37	0.35	0.23	0.36	0.37	0.36	0.23	0.22
perimeter_worst	0.78	0.97	0.36	0.97	0.96	0.24	0.59	0.73	0.86	0.22	-0.21	0.72	-0.10	0.72	0.76	-0.22	0.26	0.23	0.39	-0.10	-0.00	0.99	0.37	1.00	0.98	0.24	0.53	0.62	0.82	0.27	0.14
area_worst	0.73	0.94	0.34	0.94	0.96	0.21	0.51	0.68	0.81	0.18	-0.23	0.75	-0.08	0.73	0.81	-0.18	0.20	0.19	0.34	-0.11	-0.02	0.98	0.35	0.98	1.00	0.21	0.44	0.54	0.75	0.21	0.08
smoothness_worst	0.42	0.12	0.08	0.15	0.12	0.81	0.57	0.45	0.45	0.43	0.50	0.14	-0.07	0.13	0.13	0.31	0.23	0.17	0.22	-0.01	0.17	0.22	0.23	0.24	0.21	1.00	0.57	0.52	0.55	0.49	0.62
compactness_worst	0.59	0.41	0.28	0.46	0.39	0.47	0.87	0.75	0.67	0.47	0.46	0.29	-0.09	0.34	0.28	-0.06	0.68	0.48	0.45	0.06	0.39	0.48	0.36	0.53	0.44	0.57	1.00	0.89	0.80	0.61	0.81
concavity_worst	0.66	0.53	0.30	0.56	0.51	0.43	0.82	0.88	0.75	0.43	0.35	0.38	-0.07	0.42	0.39	-0.06	0.64	0.66	0.55	0.04	0.38	0.57	0.37	0.62	0.54	0.52	0.89	1.00	0.86	0.53	0.69
concave points_worst	0.79	0.74	0.30	0.77	0.72	0.50	0.82	0.86	0.91	0.43	0.18	0.53	-0.12	0.55	0.54	-0.10	0.48	0.44	0.60	-0.03	0.22	0.79	0.36	0.82	0.75	0.55	0.80	0.86	1.00	0.50	0.51
symmetry_worst	0.42	0.16	0.11	0.19	0.14	0.39	0.51	0.41	0.38	0.70	0.33	0.09	-0.13	0.11	0.07	-0.11	0.28	0.20	0.14	0.39	0.11	0.24	0.23	0.27	0.21	0.49	0.61	0.53	0.50	1.00	0.54
fractal_dimension_worst	0.32	0.01	0.12	0.05	0.00	0.50	0.69	0.51	0.37	0.44	0.77	0.05	-0.05	0.09	0.02	0.10	0.59	0.44	0.31	0.08	0.59	0.09	0.22	0.14	0.08	0.62	0.81	0.69	0.51	0.54	1.00

Fig 3.6: Correlation heatmap

Here show the Breast cancer attributes correlation with Heatmap.

3.4.6 Training and evaluating mode

SVM (Support Vector Machine):

The "Support Vector Machine" (SVM) is a managed machine learning or AI algorithm that can be utilized for grouping and relapse undertakings. Notwithstanding, for the most part utilized in grouping issues. In the SVM algorithm, every information thing is addressed as a point in n-layered space, with the worth of each element being the worth of a particular co-ordinate. Scikit-learn is an machine learning (ML) algorithms am carried out utilizing a Python library. SVM is likewise remembered for the scikit-learn library and is utilized in a similar way.

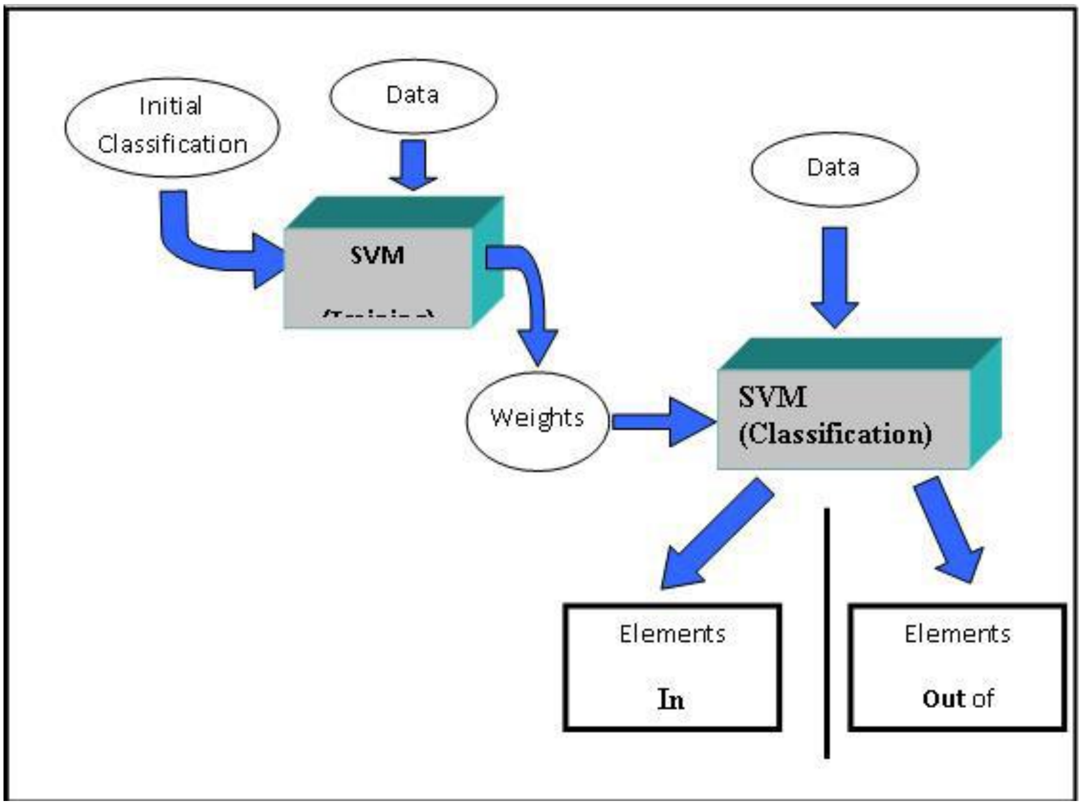


Fig 3.7: Support Vector Machine

3.4.7 Gaussian Naïve Bayes:

Nave Bayes is a probabilistic machine learning algorithm that depends on the Bayes hypothesis and is utilized for the majority grouping capabilities. Gaussian Nave Bayes is a guileless Bayes expansion. While different functions can be utilized to appraise data distribution, the Gaussian or typical dissemination is the least to carry out in light of the fact that we just have to work out the mean and standard deviation for the training data. The Gaussian probability density function can be used to make predictions by replacing the parameters with the new input value of the variable, and the Gaussian function will give an estimate for the probability of the new input value.

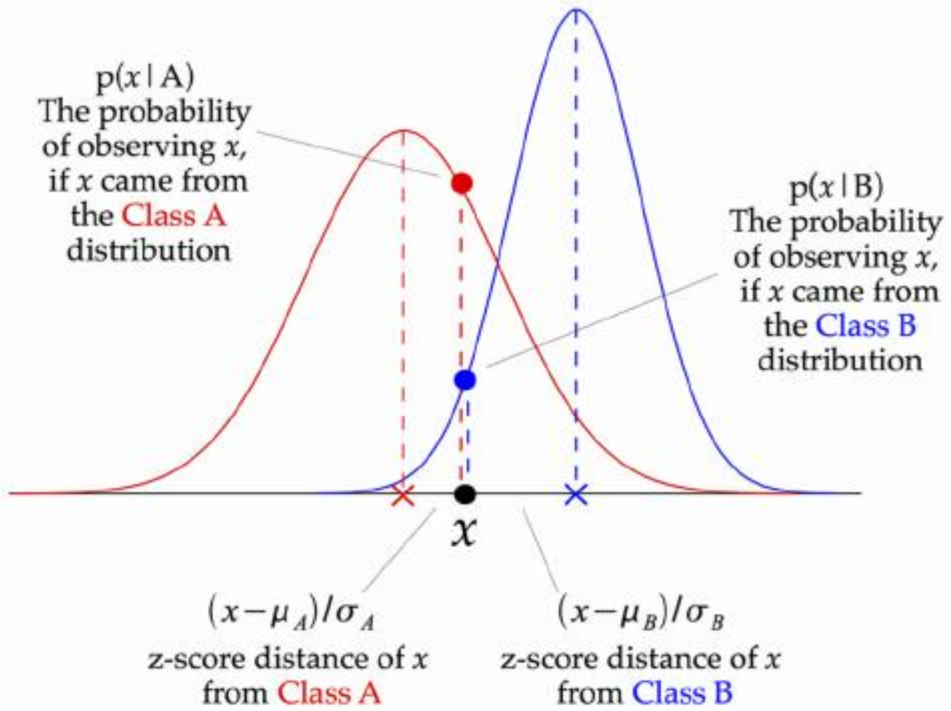
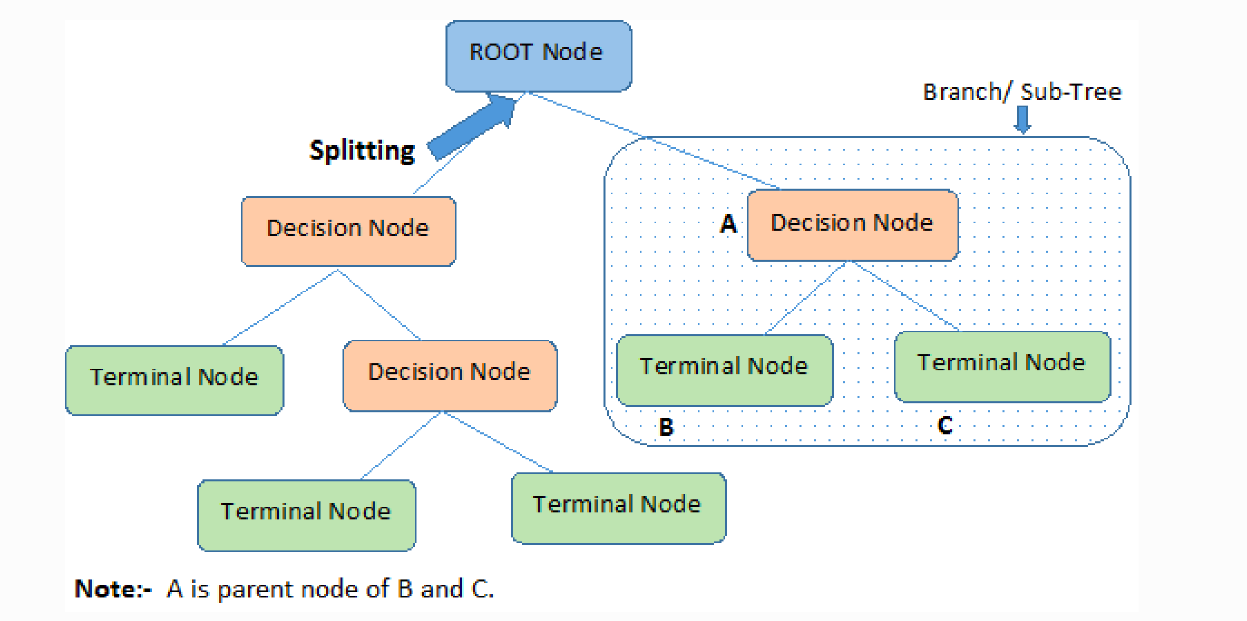

p(x | A)
The probability of observing x, if x came from the Class A distribution
p(x | B)
The probability of observing x, if x came from the Class B distribution
x
$(x-\mu_A)/\sigma_A$
z-score distance of x from Class A
$(x-\mu_B)/\sigma_B$
z-score distance of x from Class B

Fig 3.8: Gaussian Naïve Bayes

3.4.7 Decision Tree:

The Decision tree algorithm has a place with the group of regulated learning algorithms. Dissimilar to other managed learning algorithms, the decision tree algorithm can likewise be used to tackle order and relapse issues. The decision to make key parts fundamentally affects a tree's exactness. The decision models for arrangement and relapse trees am unmistakable. To choose whether to part a hub into at least two sub-nodes, decision trees utilize different algorithms. The formation of sub-nodes expands the homogeneity of the sub-nodes that outcome. As such, the immaculateness of the hub expansions comparable to the objective variable. The decision tree separates the nodes in view of every single accessible variable and afterward picks the split that creates the most homogeneous sub-nodes.



Note:- A is parent node of B and C.

Fig 3.9: Decision Tree

CHAPTER 4

EXPERIMENTAL ANALYSIS AND DISCUSSION

4.1 Introduction

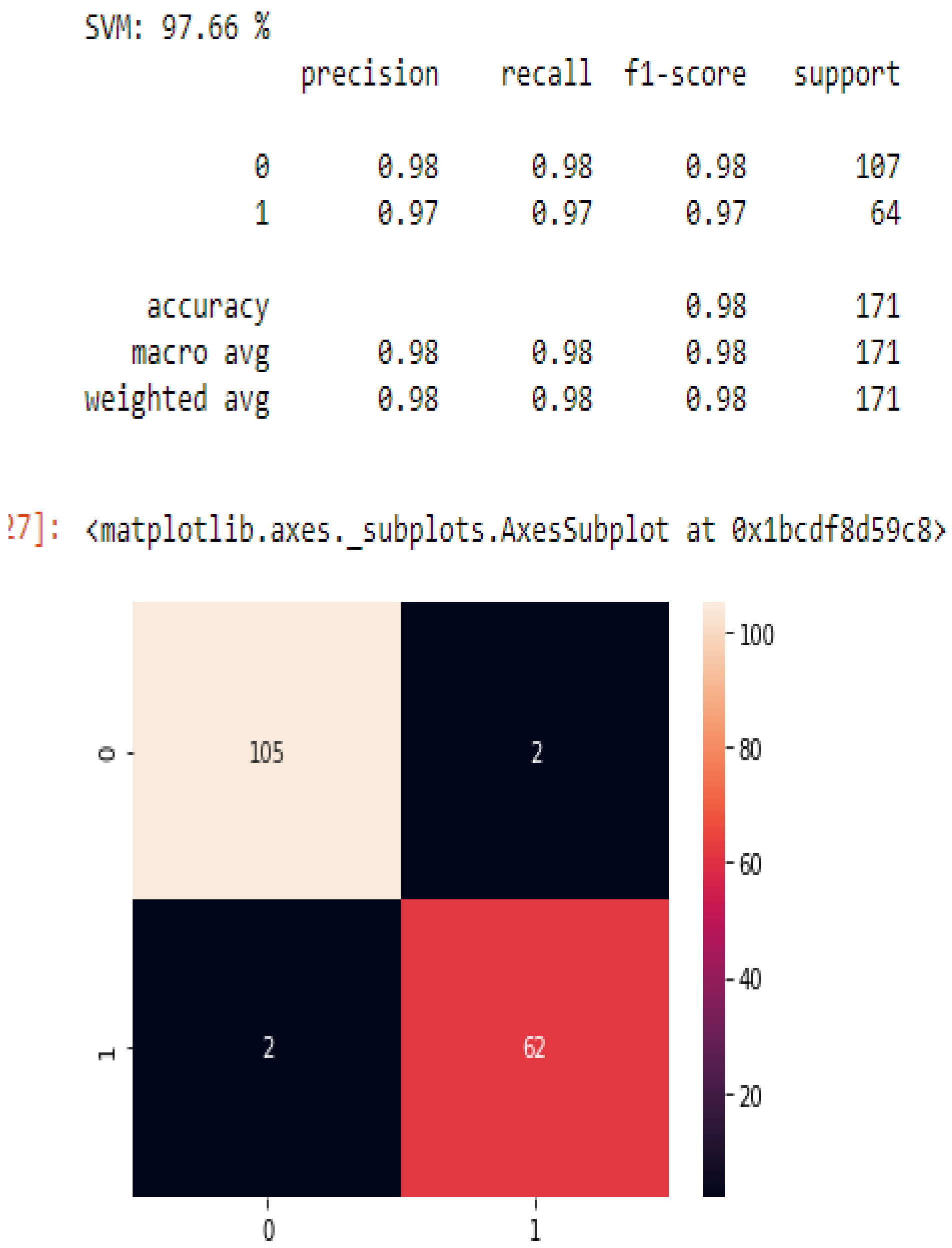
Here I ‘ill discuss outcome of this experiment in this part. Only one of the three algorithms I will utilize has been described. So, I can now assess how correctly those algorithms performed. And I'll evaluate the precision of each of the three algorithms.

Some process takes to complete this research am follows:
Step-1: Collecting dataset
Step-2: Import different libraries
Step-3: Data visualization
Step-4: Pre-processing
Step -5: Divide my dataset
Step-6: Make modals for all three algorithms.
Step-7: Train with all three machine learning algorithms.
Step-8: Determine the accuracy of each algorithm.

4.2 Experimental Result:

I am aware that no machine can produce a perfect result. In a similar vein, I may fine-tune my model's parameters during training to increase accuracy. However, the accuracy I obtain from various methods is rather good.

Below am a few images that provide a brief summary of my study activity. I am displaying precisions, recall, f1 scores, support, accuracy, and the heatmap in these images.



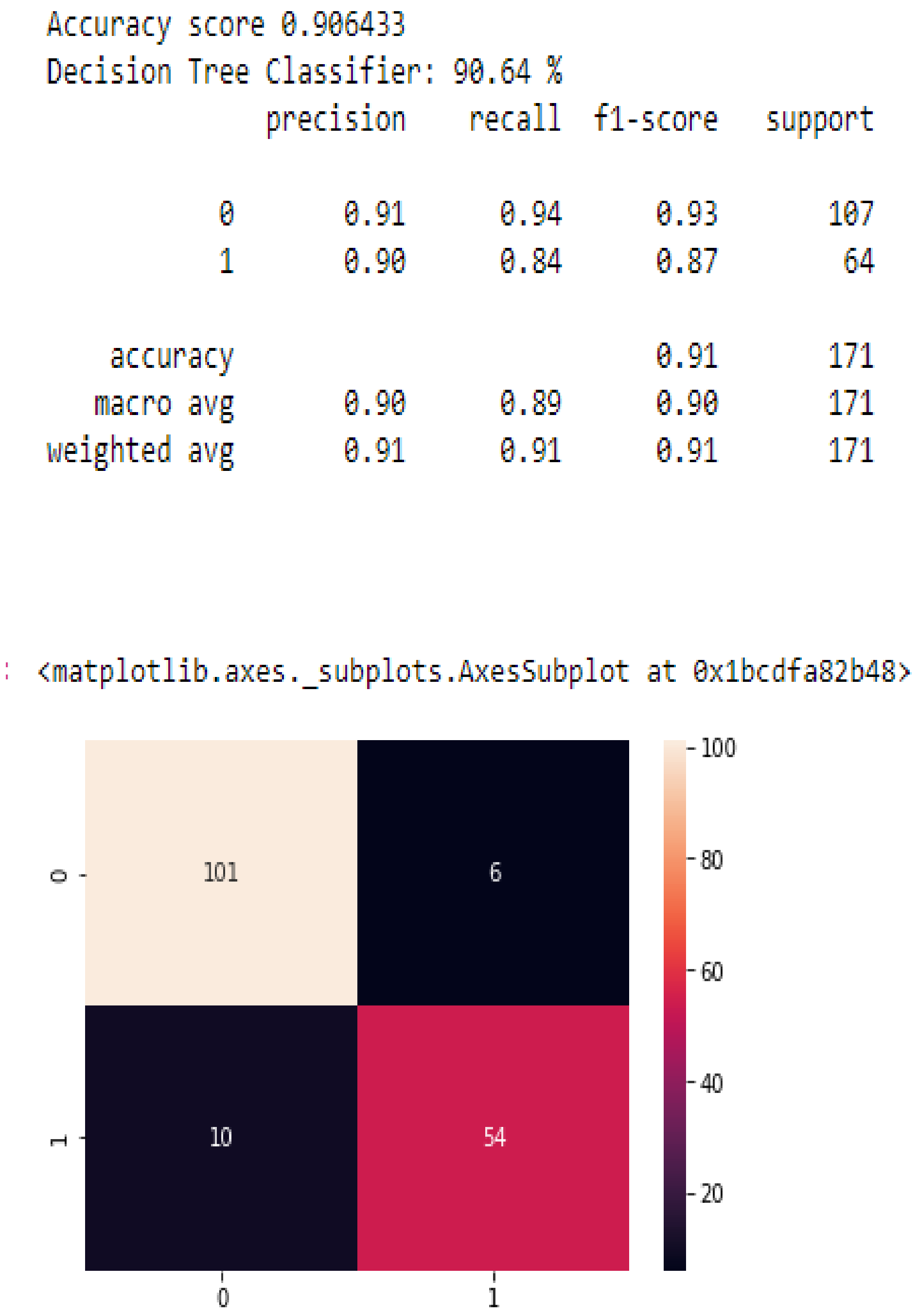
```
SVM: 97.66 %
              precision    recall  f1-score   support

           0       0.98      0.98      0.98       107
           1       0.97      0.97      0.97        64

    accuracy                           0.98       171
   macro avg       0.98      0.98      0.98       171
weighted avg       0.98      0.98      0.98       171
```

```
?7]: <matplotlib.axes._subplots.AxesSubplot at 0x1bcdf8d59c8>
```

Fig 4.2.1: Support Vector Machine result



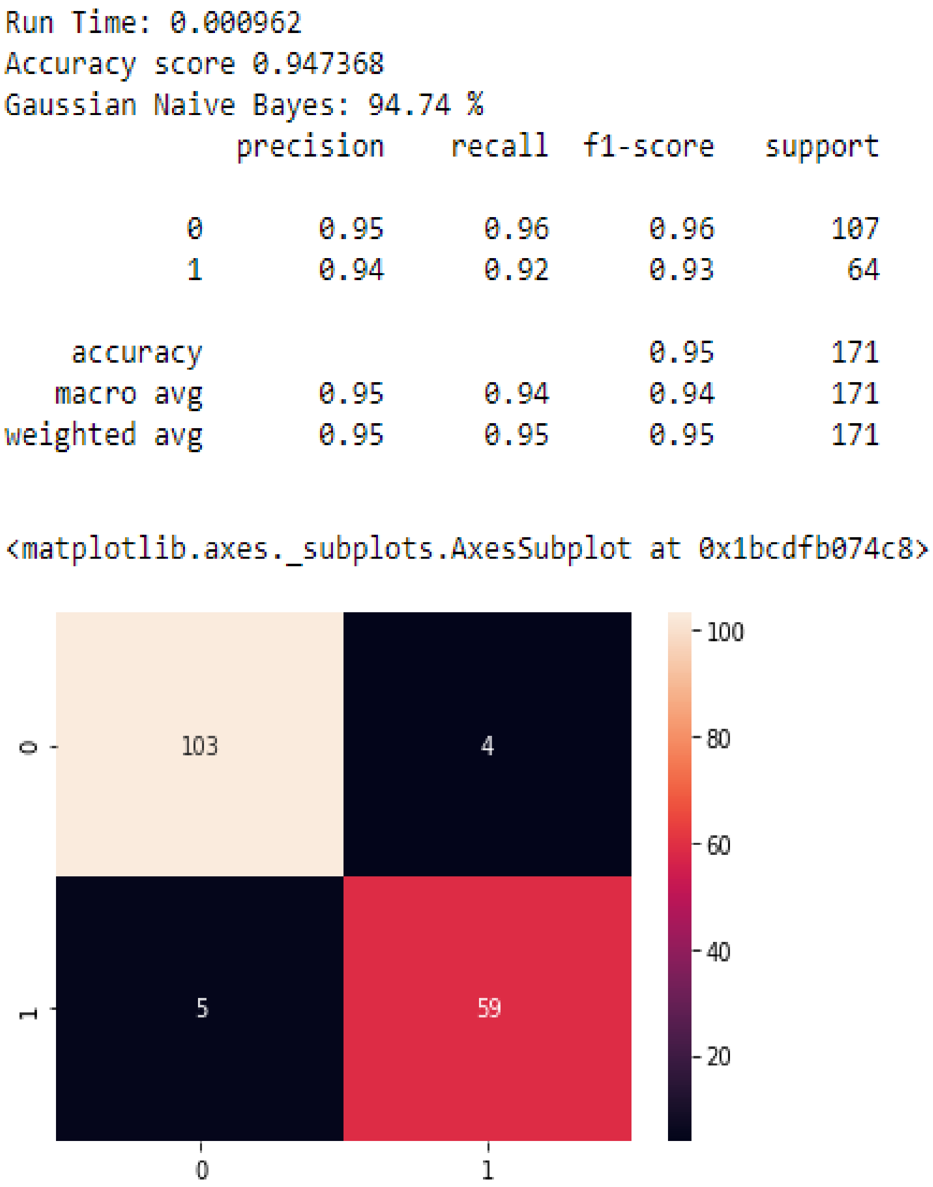
```
Accuracy score 0.906433
Decision Tree Classifier: 90.64 %
              precision    recall  f1-score   support

           0       0.91      0.94      0.93       107
           1       0.90      0.84      0.87        64

    accuracy                           0.91       171
   macro avg       0.90      0.89      0.90       171
weighted avg       0.91      0.91      0.91       171
```

```
: <matplotlib.axes._subplots.AxesSubplot at 0x1bcdfa82b48>
```

Fig 4.2.2: Decision Tree Classifier result



```
Run Time: 0.000962
Accuracy score 0.947368
Gaussian Naive Bayes: 94.74 %
              precision    recall  f1-score   support

           0       0.95      0.96      0.96       107
           1       0.94      0.92      0.93        64

    accuracy                           0.95       171
   macro avg       0.95      0.94      0.94       171
weighted avg       0.95      0.95      0.95       171

<matplotlib.axes._subplots.AxesSubplot at 0x1bcdfb074c8>
```

Fig 4.2.3: Gaussian Naïve Bayes result

Table 1: DIFFERENT ALGORITHM RESULTS

Algorithm Name	Precision		Recall		F1-score		Accuracy (%)
	0	1	0	1	0	1	
SVM	0.98	0.97	0.98	0.97	0.98	0.97	97.66
Gaussian Naïve Bayes	0.95	0.94	0.96	0.92	0.96	0.93	94.74
Decision Tree	0.91	0.90	0.94	0.84	0.93	0.87	90.64

The precision of three separate methods is seen. At least 90% accuracy is guaranteed for all algorithms. So, if I analyze all 3 algorithms, I may conclude that SVM is the most accurate at predicting or detecting breast cancer.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABLITY

5.1 Impact on Society

I accept that it will emphatically affect our society. Bosom malignant growth influences countless Bangladeshi ladies, a considerable lot of ladies. Early identification of the breast cancer utilizing Machine Learning could save an individual's life and assist with broadening their life expectancy over the long haul. By far most of Bangladeshis are indifferent about their wellbeing. It will, I accept, affect our general public.

5.2 Ethical Aspects

This study was done in a totally ethical way. The data I accumulated from the web was exclusively for the reasons for this review. All the more critically, the whole undertaking that I have embraced will help mankind. Subsequently, I don't believe the review to be unscrupulous.

5.3 Impact on Environment

There is no environmental impact from my work. There is no environmentally harmful process. The research was carried out using software.

5.4 Sustainability

This research was conducted with a long-term strategy in mind. People believe that several of these plans have been completed Ill. Data collection and data procedure can be difficult. However, if I am able to resolve these issues in the future, this project will undoubtedly improve and include advanced features that people require.

CHAPTER 6

CONCLUSION & FUTURE WORK

6.1 Conclusion

The scarcity about on prognosis frameworks makes it difficult for the doctors to develop a treatment plan which could increase patient in survival time. As a result, time ‘s required to develop a technique where the produces at the least number of errors for increasing accuracy. Breast cancer is extremely common among women people in Bangladesh. In my research, I came to a similar end. Early discovery of Breast Cancer is basic to a patient's capacity to carry on with a sound life. For this purpose, I developed a Machine Learning (ML)-based model.

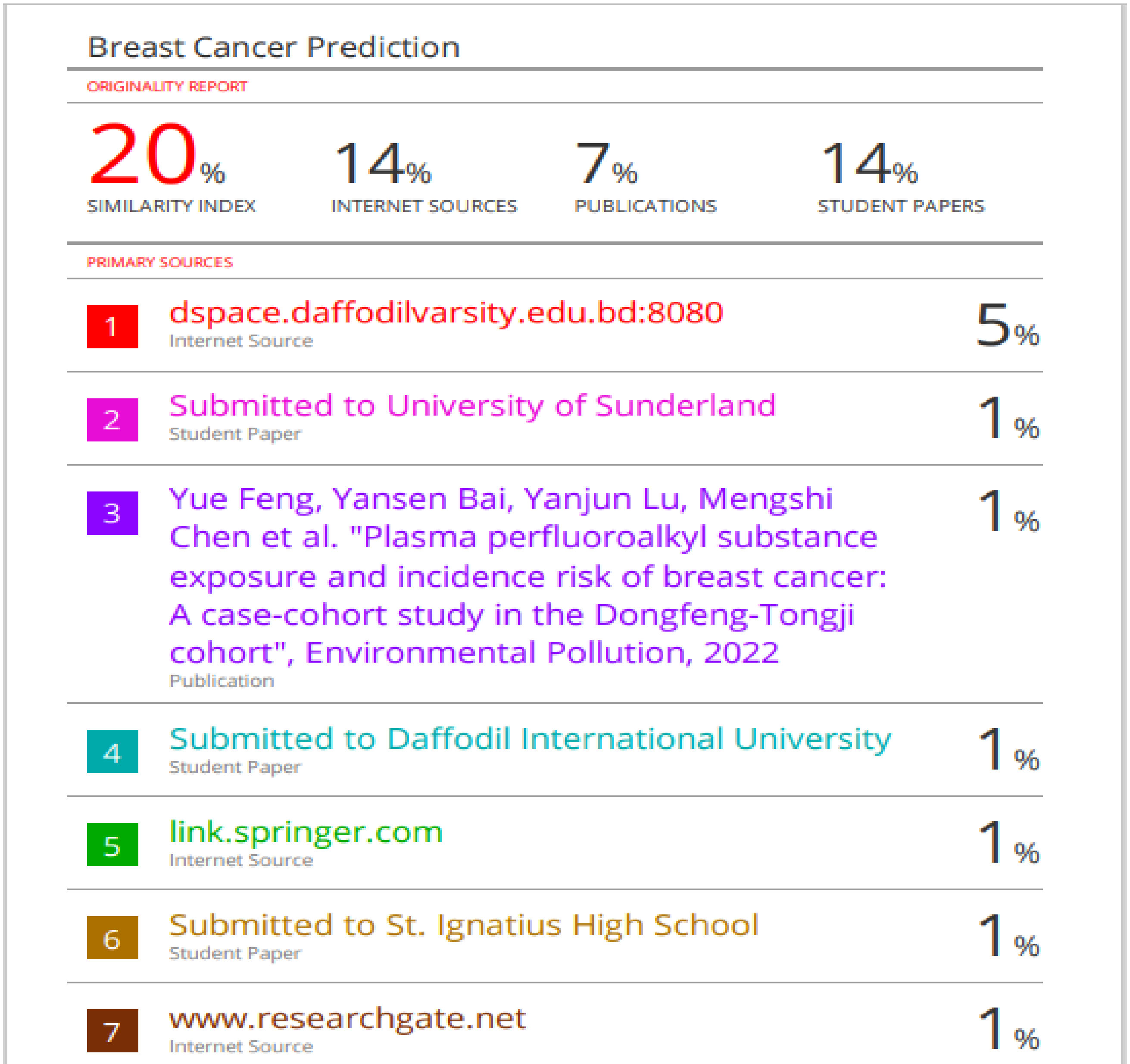
6.2 Future Work

The doors have been opened. More thorough and in-depth research is needed in this fields of breast cancer and Machine learning, particularly from the perspective of Bangladesh. There is a significant amount of room for improvement in this study. Collect a large amount of data and apply the deep learning method. The model can be used in manufacturing with more effective features and advanced techniques.

Reference

1. R. Roslidar, A. Rahman, R. Muharar, M. R. Syahputra, F. Arnia, M. Syukri, B. Pradhan, and K. Munadi, “A review on recent progress in thermal imaging and deep learning approaches for breast cancer detection,” *IEEE Access*, vol. 8, pp. 116176–116194, 2020.
2. S. H. Nallamala, P. Mishra, and S. V. Koneru , “Breast cancer detection using machine learning way,” *International Journal of Recent Technology and Engineering*, vol. 8, no. 2S3, pp. 1402–1405, 2019.
3. Y. Brhane Hagos, A. Gubern Mérida, and J. TeuIn, “Improving breast cancer detection using symmetry information with deep learning,” *Image Analysis for Moving Organ, Breast, and Thoracic Images*, pp. 90–97, 2018.
4. S. Sharma, A. Aggarwal, and T. Choudhury, “Breast cancer detection using machine learning algorithms,” *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, 2018.
5. M. F. Ak, “A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications,” *Healthcam*, vol. 8, no. 2, p. 111, 2020.
6. D. Bazazeh and R. Shubair, “Comparative study of machine learning algorithms for breast cancer detection and diagnosis,” *2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA)*, 2016.
7. H. Asri , H. Mousannif, H. A. Moatassime, and T. Noel, “Using machine learning algorithms for breast cancer risk prediction and diagnosis,” *Procedia Computer Science*, vol. 83, pp. 1064–1069, 2016.
8. A. H. Ralaidovy, C. Gopalappa, A. Ilbawi, C. Pretorius, and J. A. Lauer, “Cost-effective interventions for breast cancer, cervical cancer, and colorectal cancer: New results from WHO-DECISION,” *Cost Effectiveness and Resmyce Allocation*, vol. 16, no. 1, 2018.
9. “Breast cancer statistics: World cancer research fund international,” *WCRF International*, 14-Apr-2022. [Online]. Available: https://www.wcrf.org/cancer-trends/breast-cancer-statistics/. [Accessed: 19-Nov-2022].
10. “Breast cancer statistics: How common is breast cancer?” *American Cancer Society*. [Online]. Available: https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html#:~:text=It%20is%20about%2030%25%20(or,(DCIS)%20will%20be%20diagnosed. [Accessed: 19-Nov-2022].
11. M. A. R. Forazy, “Walshmedicalmedia, ” *Health Cam : Current Reviews*, 30-Nov--1. [Online]. Available: https://www.walshmedicalmedia.com/proceedings/incidence-of-breast-cancer-in-bangladesh-5168.html#:~:text=In%20Bangladesh%20the%20rate%20of,any%20other%20type%20of%20cancer. [Accessed: 19-Nov-2022].
12. S. Ray, “SVM: Support Vector Machine Algorithm in machine learning,” *Analytics Vidhya*, 26-Aug-2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/. [Accessed: 22-Nov-2022].

13. “Gaussian naive bayes: What we need to know?” *upGrad blog*, 16-Sep-2022. [Online]. Available: https://www.upgrad.com/blog/gaussian-naive-bayes/. [Accessed: 22-Nov-2022].
14. “Decision tree algorithm, explained,” *KDnuggets*. [Online]. Available: https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html. [Accessed: 22-Nov-2022].
15. “Your machine learning and Data Science Community,” *Kaggle*. [Online]. Available: https://www.kaggle.com/. [Accessed: 22-Nov-2022].
16. L. M. Co, E. C. Dee, M. A. Eala, S. D. Ang, and C. D. Ang, “Access to surgical treatment for breast cancer in the Philippines,” *Annals of Surgical Oncology*, vol. 29, no. 11, pp. 6729–6730, 2022.
17. M. C. Velarde, A. F. Chan, M. E. Sajo, I. Zakhamvich, J. Melamed, G. L. Uy, J. M. Teves, A. J. Corachea, A. P. Valparaiso, S. S. Macalindong, N. D. Cabaluna, R. B. Dofitas, L. C. Giudice, and R. R. Gerona, “Elevated levels of perfluoroalkyl substances in breast cancer patients within the Greater Manila Ama,” *Chemosphere* , vol. 286, p. 131545, 2022.
18. Y. X. Lim, Z. L. Lim, P. J. Ho, and J. Li, “Breast cancer in Asia: Incidence, mortality, early detection, mammography programs, and risk-based screening initiatives,” *Cancers*, vol. 14, no. 17, p. 4218, 2022.
19. M. Mulongo and C. J. ChibIsha, “Prevention of cervical cancer in low-resource African settings,” *Obstetrics and Gynecology Clinics of North America* , vol. 49, no. 4, pp. 771–781, 2022.
20. Q. Liu, J. Du, Y. Li, G. Peng, X. Wang, Y. Zhong, and R. Du, “Uncovering nasopharyngeal carcinoma from chronic rhinosinusitis and healthy subjects using routine medical tests via machine learning,” *PLOS ONE*, vol. 17, no. 9, 2022.



Breast Cancer Prediction

ORIGINALITY REPORT

20%	14%	7%	14%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	5%
2	Submitted to University of Sunderland Student Paper	1%
3	Yue Feng, Yansen Bai, Yanjun Lu, Mengshi Chen et al. "Plasma perfluoroalkyl substance exposure and incidence risk of breast cancer: A case-cohort study in the Dongfeng-Tongji cohort", Environmental Pollution, 2022 Publication	1%
4	Submitted to Daffodil International University Student Paper	1%
5	link.springer.com Internet Source	1%
6	Submitted to St. Ignatius High School Student Paper	1%
7	www.researchgate.net Internet Source	1%