



بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ



**Sudan University of Science and Technology**

**College of Graduates Studies**

Thesis Title:

# **Developing the Classification of Breast Cancer System Using Random Forest Technique**

تطوير نظام تصنيف سرطان الثدي باستخدام تقنية الغابة العشوائية

Submitted in Partial Fulfillment of the Requirement of M.Sc.  
Degree in Biomedical Engineering

**BY:**

**Eithar Seifeldeen Mohammed Mustafa**

**SUPERVISED BY:**

**Dr. Eltahir Mohammed Hussein**

*October 2022*

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

## الآية

(وَيَسْأَلُونَكَ عَنِ الرُّوحِ قُلِ الرُّوحُ مِنْ أَمْرِ رَبِّي وَمَا أُوتِيتُمْ مِنَ  
الْعِلْمِ إِلَّا قَلِيلًا)

صدق الله العظيم

## DEDICATION

This work is dedicated to the man who gave me everything to get me where I am today, for he was the first chapter in my attainment of higher education (my beloved father: **Saifeldeen**), may God bless his soul.

To the one who sat me on the path of life, made me calm, and took care of me until I became the woman I am today, (my dear mother: **Sumia**), may God bless her soul.

To my sisters; **Nibras** and **Tsabyh**, those who had a great impact on my life and helped me to get over many obstacles and difficulties.

To my dear husband **Amro** who supported and did not hesitate to extend a helping hand to me, I dedicate this work to them.

In the end I dedicated this work for the strongest person I know,  
myself.

## **ACKNOWLEDGMENTS**

First and foremost, I would like to praise Allah Almighty my creator, my strong pillar, my source of inspiration, wisdom, knowledge and understanding.

Determination, health and granted me patience to successfully complete of this research.

I cannot express enough thanks to my supervisor **Dr. Eltahir Mohammed Hussein** for his continued guidance, support, unlimited help, and encouragement; I offer my sincere appreciation for the learning opportunities provided by you.

My completion of this research could not have been accomplished without the support of my parents, family, and friends. My deepest gratitude, and special appreciation and thanks to Breast cancer patients, **Eng. Waha Al Sharif**, and **Eng. Amro Ahmed** for their support and guidance. Your encouragement when the times got rough is much appreciated and duly noted.

# Table of Contents

<b>DEDICATION</b> .....	i
<b>ACKNOWLEDGMENTS</b> .....	ii
Table of Contents .....	iii
List of Figures .....	vi
List of Tables.....	viii
Abstract .....	ix
المستخلص .....	x
Chapter One.....	1
Introduction .....	1
1.1 General View: .....	1
1.2 Problem Statements:.....	1
1.3 The Objectives: .....	2
1.3.1 General Objective: .....	2
1.3.2 Specific Objectives are to: .....	2
1.4 Methodology: .....	2
1.5 Thesis Layout: .....	2
Chapter Two .....	3
Literature Reviews .....	3
Chapter Three.....	6
Theoretical Background .....	6
3.1 Anatomy of Female Breast: .....	6
3.2 Cancer: .....	7
3.3 Breast Cancer: .....	7
3.4 Breast Cancer Risk Factors: .....	7
3.5 Breast Cancer Staging:.....	8
3.6 Breast Cancer Symptoms: .....	9
3.7 Breast Cancer Treatment:.....	9
3.8 Breast Cancer Early Detection and Diagnosis: .....	10
3.9.1 Breast Cancer Biopsy Types: .....	10

3.9 Machine learning:.....	11
3.10 Types of Machine Learning Algorithms:.....	12
3.11.1 Supervised Learning: .....	12
3.11.2 Unsupervised Learning: .....	12
3.11 Machine learning Services: .....	12
3.12.1 Regression: .....	12
3.12.2 Classification:.....	13
3.12 Machine Learning Workflow:.....	13
3.13 Random Forest Algorithm: .....	14
3.14 Background on Python:.....	15
Chapter Four.....	16
Proposed System (Methodology).....	16
4.1 View of the Proposed System: .....	16
4.2 Dataset Description: .....	16
4.3 Data Preprocessing:.....	18
4.4 Feature Selection:.....	19
4.5 Dataset Splitting: .....	20
4.6 Classification:.....	20
4.7 Model Performance by Validation Measure: .....	20
4.8 Implementation Detail:.....	21
Chapter Five .....	22
Results and Discussion.....	22
5.1 Result of Breast Cancer Classification Using Random Forest: .....	22
5.2 Performance Evaluation of Random Forest Classification:.....	23
5.3 Result of Random Forest Feature Selection: .....	25
5.4 Result of Breast Cancer Classification Using Random Forest with Feature Selection: .....	26
5.6 Performance Evaluation of Random Forest Classification with Feature Selection: .....	29
5.7 Summary of Results and Discussion:.....	35
Chapter Six.....	37

Conclusion and Recommendation.....	37
6.1 Conclusion:.....	37
6.2 Recommendation:.....	37
References .....	38

## List of Figures

Figure 1. 1: Block diagram of breast cancer proposed system .....	2
Figure 3. 1: anatomy of female breast .....	6
Figure 3. 2: Various types of machine learning techniques.....	12
Figure 3. 3: Machine Learning Workflow .....	13
Figure 3. 4: diagram of random forest .....	14
Figure 4. 1: Flowchart of breast cancer proposed system.....	16
Figure 4. 2: Example of WBCD database details .....	17
Figure 4. 3: WBCD missing data .....	18
Figure 4. 4: WBCD duplicated data.....	19
Figure 4. 5: Example of WBCD dataset after preprocessed .....	19
Figure 5. 1: Results of accuracy of the breast cancer classification performed by random forest without feature selection.....	22
Figure 5. 2 : Sample of decision tree from the random forest .....	23
Figure 5. 3: Classification report for random forest classification without feature selection.....	23
Figure 5. 4: Confusion matrix for random forest classification without feature selection.....	24
Figure 5. 5: AUC of ROC for random forest classification without feature selection.....	25
Figure 5. 6: Results of the accuracy of breast cancer classification performed by random forest with eight feature selection.....	26
Figure 5. 7: Sample of decision tree from the random forest with eight attributes .....	27
Figure 5. 8: Results of the accuracy of breast cancer classification performed by random forest with six feature selection.....	27
Figure 5. 9: Sample of decision tree from the random forest with six attributes	28
Figure 5.10: Results of the accuracy of breast cancer classification performed by random forest with four feature selection .....	28
Figure 5. 11: Sample of decision tree from the random forest with four attributes .....	29
Figure 5. 12: Classification report for random forest classification with eight feature selection.....	29
Figure 5. 13: Confusion matrix for random forest classification with feature selection.....	30
Figure 5.14: AUC of ROC for random forest classification with eight feature selection.....	31



Figure 5. 15: Classification report for random forest classification with six feature selection.....	31
Figure 5. 16: Confusion matrix for random forest classification with feature selection.....	32
Figure 5.17: AUC of ROC for random forest classification with six feature selection.....	33
Figure 5. 18: Classification report for random forest classification with four feature selection.....	33
Figure 5. 19: Confusion matrix for random forest classification with feature selection.....	34
Figure 5.20: AUC of ROC for random forest classification with eight feature selection.....	35

## List of Tables

Table 2- 1: Some of prewise study.....	4
Table 3- 1: TNM staging .....	8
Table 3- 2: Anatomic Staging Summary .....	9
Table 4- 1: Feature of Wisconsin Breast Cancer Dataset with Description. ....	17
Table 5- 1: Result of random forest feature selection.....	26
Table 5- 3: Comparison between the proposal with literature reviews. ....	36

## Abstract

Breast cancer BC is the most cause of death in women around the world. Manual diagnosis is less effective due to physician uncertainty. For this reason, the purpose of this study was to use random forest algorithm for the classification of malignant and benign breast tumors based on cytological features.

In this study, a random forest RF algorithm was used to classify tumors of BC using the Wisconsin breast cancer WBCD dataset which consists of 699 instances, 11 real-world attributes, and two classification class from the University of California Irvine (UCI) machine learning repository, with selecting features and without it. In the first, RF is used to classify the WBCD without eliminating features. In second, the features are reduced from nine to: eight, six and four attributes, by using RF and then classified using RF either.

The RF model with six attributes obtained acceptable performance, where this model achieved an accuracy of 98.52%, 100% sensitivity, 98.04% specificity, and 0.99 AUC.

This research demonstrated that BC can be classified using RF classifier. Proposed RF model can be used to obtain fast automatic diagnostic system for any other diseases in the future.

## المستخلص

يعتبر مرض سرطان الثدي من أكثر الأسباب المؤدية للوفاة شيوعاً لدى النساء حول العالم. يعد التشخيص اليدوي أقل فعالية بسبب عدم اليقين عند الطبيب. لهذا السبب، تهدف هذه الدراسة إلى استخدام خوارزمية الغابات العشوائية لتصنيف الأورام الخبيثة والحميدة للثدي بناءً على السمات الخلوية.

في هذه الدراسة تم استخدام نموذج الغابات العشوائية لتصنيف أورام سرطان الثدي مع تحديد السمات المميزة مرة وبدون تحديد السمات المميزة مرة أخرى، حيث تم استخدام بيانات ويسكونسن الأصلية المكونة من 699 حالة و 11 سمة حقيقية وفنتين تصنيف والموجودة في جامعة كاليفورنيا إيفرين في مستودع التعليم الآلي لهذه الدراسة. في المرحلة الأولى استخدمت جميع السمات الخلوية الموجودة لتصنيف سرطان الثدي بالغابات العشوائية. وفي المرحلة الثانية تم تقليص تلك السمات من تسع إلى ثمانية ثم إلى ستة ثم إلى أربع سمات بواسطة الغابات العشوائية ومن ثم تم تصنيفها بواسطة الغابات العشوائية أيضاً.

حصل نموذج الغابات العشوائية باستخدام ست سمات خلوية على نتائج مقبولة مقارنة مع النماذج الأخرى، حيث كانت دقة هذا النموذج هي 98.52% والحساسية 100% والخصوصية 98.04% وكانت المساحة تحت منحنى خصائص التشغيل 0.99.

أظهر البحث أنه يمكن تصنيف سرطان الثدي باستخدام مصنف الغابات العشوائية. وكما يمكن اقتراح استخدام نموذج الغابات العشوائية للحصول على نظام تشخيص تلقائي سريع لأي امراض أخرى مستقبلاً.

# Chapter One

## Introduction

### 1.1 General View:

Cancer was defined as “a cell growing abnormal and uncontrollable”, according to the World Health Organization (WHO) in 2018, Breast Cancer (BC), the second most common Cancer in the world after lung Cancer and the first cause of death in women, and more than 0.6 million deaths due to BC worldwide. In Sudan, around 78% of BC cases were in the last stage of the disease. The factors which can cause BC in women were demographic, family history, reproductive factors, estrogen, and lifestyle [1-4].

Recently, there has been a great development in artificial intelligence, especially machine learning algorithms that were both interpretable and accurate which greatly improve the early detection and diagnosis of BC and increases the chance of treatment and survival [5].

Currently, a popular technique used for the detection of BC in the early stages was mammography and biopsy, the mammography methods include X-ray, ultrasound imaging, thermal imaging, and so on, and the biopsy method includes needle biopsy, surgery biopsy [6, 7]. The biopsy method has less harm, has painless, has a lower cost for patients, and has higher accuracy than the mammography method [6-8]. But according to cost and invasive fine needle aspiration biopsy (FNA) was the lowest, which “a thin and hollow needle is inserted into the breast mass to take a sample of cells (tissue or fluid) from breast then Collected samples of cells are then examined (analyzed) under a microscope”[5].

A machine learning technique can play an important role in reducing BC interpretation errors and diagnosis, thus, assisting physicians in decision-making by providing a second opinion[8].

### 1.2 Problem Statements:

Physicians take a lot of time to analyze cytological material in the breast by FNA, due to exhaustion or limited experience and knowledge, the occurrence of misdiagnosis leads to improper treatments and therefore scary results such as death.

### 1.3 The Objectives:

The objectives of this research are general objective and specific objectives.

#### 1.3.1 General Objective:

The main aim of this study was to use machine learning techniques for the classification of malignant and benign tumors for Wisconsin breast cancer.

#### 1.3.2 Specific Objectives are to:

1. classification of the BC tumor types, benign and malignant.
2. choose the best cytological features that estimates BC.
3. assess the accuracy, sensitivity, and specificity of the classifier.

### 1.4 Methodology:

The main aim of this study was to design a random forest model to classify BC. Firstly datasets [9] were been preprocessing, follow by feature selection, then split the reminding data into training and testing data, then data classified by random forest into benign or malignant, and finally, the accuracy and performance of classification were analyzed.

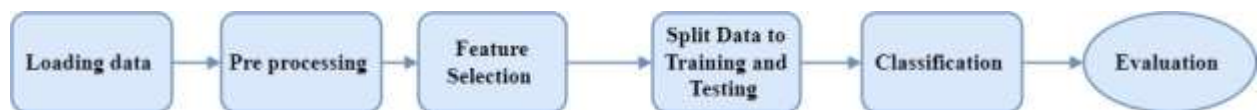


Figure 1. 1: Block diagram of breast cancer proposed system

### 1.5 Thesis Layout:

This research was organized into six various chapters:

Chapter one was an introduction, the related literature reviews were given in Chapter two, and followed by Chapter three presents the theoretical background. The proposed system was described in Chapter four. The penultimate Chapter five discusses the results and discussion, and finally, Chapter six is conclusions and recommendations which conclude the research and present future work.

## Chapter Two

### Literature Reviews

In the last decade, many machine learning algorithms have been applied for breast cancer (BC) classification, such as decision tree (DT), random forest (RF), support vector machine (SVM), neural network (ANN), linear regression (LR), K-nearest neighbor (KNN), and naive bayes (NB)[7, 10-17].

In this section some of the papers will be represented as the following:

R. Preetha and S. Vinila Jinny 2021 proposed early diagnosis BC system used a combination of principal component's analysis (PCA) and linear discriminate-analysis (LDA) for feature reduction of WDBC. An adaptive neuro-fuzzy inference system (ANFIS) as classification, the classifier compering with SVM and multilayer perceptron (MLP), but accuracy increased by using PCA-LDA-ANFIS to 98.6% [7].

N Rane et al. 2020 proposed compering six machine learning algorithms on WDBC and using the best to build the website that doctors can use to classify cancer. Nevertheless, the better algorithm was not specified [10].

Ed-Daoudy and Maalmi, 2020 applied association rules (AR) to eliminate irrelevant features in the WBCD. Four out of nine features were selected. Several classification algorithms were used. Performance metrics like accuracy, precision, and recall were calculated for each classifier. The SVM with threefold cross-validation produced the highest classification accuracy 98.00% [11].

Gouda I. Salma et al. in 2012 compared four classification algorithms DT, NB, MLP, KNN and sequential minimal optimization (SMO) on three different BC datasets, WBCD, WDBC, and WPBC. PCA was used for feature reduction transformation. found that MLP and J48 classifiers gave higher accuracy results as compared to other classifiers. SMO gave the best results in the WDBC dataset [12].

M. Kumari and V. Singh 2018 proposed a system for the prediction of WBCD. Applied Correlation-Based Measures (CBM) for feature selection, and k-folds cross-validation for validation performance. Used three different classification algorithms including KNN, LR, and SVM, and KNN achieved higher accuracy of 99.28%. Nevertheless, the optimal feature subset was not reported [13].

MH Alshayeji et al. 2022 proposed using the ANN model to diagnose and predict BC using datasets WBCD, and WDBC. The ANN models for WBCD achieved an accuracy of 99.85%, specificity of 99.72%, the sensitivity of 100the %, precision of 99.69%, and an F1 score of 99.84%. For BC detection using WDBC,

achieved an accuracy of 99.47%, specificity of 99.53%, sensitivity the 99.59%, precision of 98.71%, and F1 score of 99.13%. The AUC of the proposed model was 99.86% and 99.56% for the WBCD and the WDBC dataset, respectively [14].

OI Obaid et al. 2018 proposed three machine learning algorithms to classify WDBC. The SVM has the best accuracy 98.1% the and lowest false discovery rates [15].

G Gupta al el. 2021 implemented five classification algorithms, RF, NB, SVM, LR, and DT in three different data mining tools, Python, R, and Orange on WDBC dataset. RF proved to be the best classifier and outperformed all other classifiers used during the study and analysis as it gave 95% accuracy with different partition ratios (67:33 and 80:20) respectively [16].

Yixuan Li and Zixuan Chen in 2018 used BCCD and WBCD datasets to explore the relationship between BC and attributes, comparing the accuracy, F-measure metric, and ROC curve of five classification models, the RF achieve better performance and adaptation than other four methods [17].

Table 2- 1: Some of previse study.

Authors/ Years	Methods	Dataset	Classification Accuracy (%)
Yixuan Li & Zixuan Chen 2018[17]	DT	WBCD	96.1
	SVM		95.1
	RF		96.1
	LR		93.7
	ANN		95.6
MH Alshayeji et al. 2022[14]	ANN	WBCD	99.85
Ed-Daoudy & Maalmi 2020[11]	AR-SVM	WBCD	98.00
G Gupta al el. 2021[16]	RF	WDBC	95.0
	NB		93.0
	SVM		63.0
	LR		95.0
	DT		92.0
OI Obaid et al. 2018[15]	SVM	WDBC	98.1
	DT		93.7
	KNN		96.7



In previous studies, the authors focused on comparing the performance of different classification algorithms, without giving importance to finding the best subset of features, as they assumed that all features have equal importance for the classifier.

This research involves, classifying the whole dataset, identifying the most important features in this set, and evaluating the performance of the classifier effectiveness in classifying malignant and benign tumors.

# Chapter Three

## Theoretical Background

### 3.1 Anatomy of Female Breast:

The breasts in human females lie between the second and sixth rib anteriorly and extend from the lateral border of the sternum to the anterior mid-axillary line which all this area called a mammary region. The breast was centered by the nipple, which was located along the mid-clavicular line and the fourth rib. The areola was a circular region of pigment that surrounds the nipple. The volume and shape of the breast differ greatly among different women and during the lifetime depend on the amount of fat surrounding the glandular tissue[18, 19]

The major component of the breast was mammary glands, which were modified sweat glands located in the superficial fascia of the pectoral region, which produce and secrete milk during lactation for the nourishment of the infants, and have a complex tree structure.

The mammary gland contains 15 to 20 lobes arranged in a radial fashion. Each lobe is divided into 20 to 40 lobules, which consist of 10 to 100 hollow cavities known as alveoli, each lobe drains into lactiferous ducts which converge with the nipple, and each duct expands into a lactiferous sinus that reservoir of milk during lactation [20].

An overall view is given in Figure 3.1.

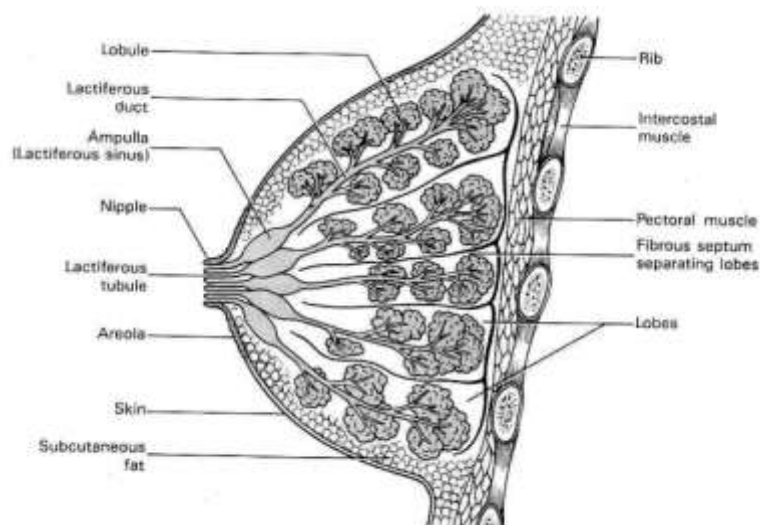


Figure 3. 1: anatomy of female breast [21]

### **3.2 Cancer:**

Cancer was defined as the disease caused by an uncontrolled division of abnormal cells in a part of the body. More than 100 types of cancers around the world, each has a unique set of symptoms and the second leading cause of death globally. One of the most common cancers was breast cancer (BC). According to UICC, in 2020, 25% of women diagnosed with cancer have BC [22-24].

### **3.3 Breast Cancer:**

Breast cancer (BC) was defined as a disease that arises in the epithelium cells of the lobules or ducts of the breast or in the stroma, which grows abnormal and out of control and affects other parts of the body. Ducts are tiny tubes that take milk away from lobules, which are glands that create milk. The stroma made up of everything other than epithelial tissue so, fibrous tissue, fatty tissue, blood vessels, which hold the breast and give a shape and size [23].

Most BC comes from epithelial cancers, which called carcinomas, about 80- 85% of BC comes from the ducts, and about 10-15%, comes from the lobules. According to WHO, BC is the second common leading cancer type, exceeded only by lung cancer. In Sudan 79.5% of patent diagnosis with cancer has invasive breast carcinoma [1, 23, 25].

### **3.4 Breast Cancer Risk Factors:**

Although the causes of BC were unknown, many risk factors (RF) have been identified, including genetics, environment, and lifestyle. RF can be divided into two main factors, the modifiable which can do something about, and the non-modifiable which was out of control. When looking at RF the two main factors are being a woman and getting older, which nothing can do about. According to research, people were diagnosed after the age of 50. Men still have a risk of getting BC but much lower than women. Also, hereditary including genetic factors and positive family history of BC can increase the risk of developing BC due to carry mutations or mistakes in genes. Endogenous hormones like, hormones made by the ovaries in the premenstrual, and the adrenal glands above the kidneys also produce hormones. These hormones can be converted to estrogen, which increases the risk of developing BC. Generally, little things that can be done about non-modifiable RF[4].

The modifiable RF includes lifestyle (alcohol consumption and smoking), obesity and overweight women after menopause), radiation (exposure to high doses of radiation), reproductive factors (late age of menopause and pregnancy characteristics), hormones (hormonal contraceptive method and postmenopausal hormone therapy), air pollution, night works, socioeconomic status, and diabetes.

Finally, when someone has all the mentioned factors, doesn't mean will sufferers and vice versa. In the case of cancer, if the diagnosis occurs quickly, the patient can be saved as there have been advances in cancer treatment[10].

### 3.5 Breast Cancer Staging:

The American Joint Committee on Cancer's Staging System or AJCC, which is the commission that oversees staging has a very simple scheme for almost all cancers, determines whether cancer has spread to other parts of the body and cancer stage. This scheme was called the Tumor-Node-Metastasis system or TNM for short, In TNM classification, (T) describes the size of the tumor; (N) describes whether cancer has spread to the lymph nodes, and (M) describes whether cancer has spread to a different part of the body or distant metastases outside of the breast and the lymph nodes. Putting the T, N, and M together gives the staging classification. Table (3-1) provides a complete view of this system. BC has five stages from zero to four described in Table (3-2) [23].

Table 3- 1: TNM staging [26]

<b>Classification</b>	<b>Definition</b>
<b>Primary tumor (T)</b>	
TX	Primary tumor cannot be assessed
T0	No evidence of primary tumor
Tis	Carcinoma in situ
T1	Tumor $\leq 2$ cm in greatest dimension
T2	Tumor $> 2$ cm but $\leq 5$ cm in greatest dimension
T3	Tumor $> 5$ cm in greatest dimension
T4	Tumor of any size with direct extension to chest wall or skin, only as described below
<b>Regional lymph nodes (N)</b>	
NX	Regional lymph nodes cannot be assessed (e.g., previously removed)
N0	No regional lymph node metastasis
N1	Metastasis in movable ipsilateral axillary lymph node(s)
N2	Metastases in ipsilateral axillary lymph nodes fixed or matted, or in clinically apparent* ipsilateral internal mammary nodes in the absence of clinically evident axillary lymph-node metastasis
N3	Metastasis in ipsilateral infraclavicular lymph node(s), or in clinically apparent* ipsilateral internal mammary lymph node(s) and in the presence of clinically evident axillary lymph-node metastasis; or metastasis in ipsilateral supraclavicular lymph node(s) with or without axillary or internal mammary lymph-node involvement

Distant metastasis (M)	
MX	Distant metastasis cannot be assessed
M0	No distant metastasis
M1	Distant metastasis

Table 3- 2: Anatomic Staging Summary [27]

Stage	TNM
Stage 0	Tis, N0, M0
Stage IA	T1, N0, M0
Stage IB	T0, N1mi, M0 T1, N1mi, M0
Stage IIA	T0, N1, M0 T1, N1, M0
Stage IIB	T2, N0, M0 T2, N1, M0 T3, N0, M0
Stage IIIA	T0, N2, M0 T1, N2, M0 T2, N2, M0 T3, N1, M0 T3, N2, M0
Stage IIIB	T4, N0, M0 T4, N1, M0 T4, N2, M0
Stage IIIC	Any T, N3, M0
Stage IV	Any T, Any N, M1

The main places where BC spreads to the bone. That's the most common. Then lungs, liver, brain occasionally, and then other places.

### 3.6 Breast Cancer Symptoms:

Symptoms of BC differ in persons. However, some common symptoms include dimpling or puckering in the breast, which indicator of cancers pulling or tugging in on that skin. Lump, whether can be seen or felt. Most BC comes from the ducts and goes to the nipple, nipple changes could be indicators of BC as flattened or inverted, scaling of the nipple, or nipple discharge. Finally, change in the breast skin [28].

### 3.7 Breast Cancer Treatment:

The treatment for BC based on the type, the tumor size, and stage of the cancer. But commonly involves: surgery, radiation therapy, chemotherapy, and hormonal therapy[29].

Localized tumors removed surgically by partial mastectomy, where the affected part is removed, and larger tumors which have spread to nearby tissue are removed by total mastectomy, where the entire breast is removed. In addition, nearby structures like lymph nodes may also be removed if the cancer has metastasized to them.

When the tumor has already spread in the body, the role of radiation therapy comes for killing cells using high energy waves. The other way is Chemotherapy,

which is the use of drugs, however, this treatment also has its side effect such as hair loss, early menopause, and fatigue.

Hormone therapy is used when tumor cell has hormone receptors like estrogen and HER2, and may include medications which block the formation or effects of estrogen.

Unfortunately, currently no cure for cancer completely. So, if earlier the diagnosis, then effective treatment.

### **3.8 Breast Cancer Early Detection and Diagnosis:**

All researchers agree that early diagnosis of BC remains an important factor in determining the stage of treatable [7, 30, 31].

BC is often detected by: breast self-examination (BSE)/clinical breast exam, x-rays imaging, ultrasound imaging (US), magnetic resonance imaging (MRI), and biopsy.

The BSE has been the simplest method for BC detection in women, this exam starts with an inspection, followed by palpation, which is feeling the breast tissue and the lymph nodes.

Beyond the breast exam, the x-ray can be used in screening for BC, the most common is mammography, which revolutionized dealing with BC. Because as opposed to BSE where a lump needs to be a reasonable size for the physician to detect by physical exam, on mammography physicians can often find cancers before ever feeling them.

When women have dense breast tissue, mammography can't be useful. so most commonly, the US was used to distinguish masses from being either cystic or solid, rely on the application of high-frequency sound waves directly into the breast tissue, and detect the reflected sound waves (echoes) [32].The only people who need an MRI for screening are those at very high risk.

#### **3.9.1 Breast Cancer Biopsy Types:**

Performing a biopsy process entails taking a sample of the patient's tissue and sending it to a pathologist for examination. Breast needle biopsy refers to the sampling of non-palpable or indistinct breast lesions by using techniques that enable the spatial localization of the lesion within the breast. Breast biopsies have three major types: The first is what's called a fine needle aspiration or the FNA. The second was a core needle biopsy. And the third is an excisional biopsy. All these types of biopsies enable the pathologist to collect cancer information, whether, invasive or in situ, coming from duct or lobules, and have receptors for hormones or other receptors.

### Fine Needle Aspiration:

The procedure known as fine needle aspiration cytology (FNAC) involves taking a sample of cells from the breast lump specifically for microscopic analysis using a thin gauge (25-22G) needle by a pathologist. Diagnostics by FNAC was still the first choice to evaluate breast lesions because simple, cheap, quick, and relatively accurate [33, 34].

This procedure was done by the following steps: “The skin was cleaned with an alcohol swab. A 25-22-gauge needle was used on a plastic disposable 10- or 20-mL syringe attached to a plastic or metal holder. The nodule was immobilized between the fingers, and the needle tip was rapidly directed through the skin into the nodule. Once the needle enters the mass, the needle was continuously aspirated while the needle was rapidly moved in and out to obtain the sample. Suction was then relieved, and the needle was withdrawn and detached from the syringe. Air was aspirated, and the material was expelled on glass slides. The material is gently but rapidly smeared on the slides and immediately dipped in alcohol fixative. Some of the slides can be left to air dry for Diff-Quick staining and rapid evaluation. The fixed material is stained by Papanicolaou or hematoxylin-eosin stains, which are both excellent for nuclear detail” [34].

### Core Biopsy Needle:

Core biopsy process involves taking a core of tissue from the breast lump using a needle that have the same thickness 14-gauge needles as a pencil lead required local anaesthetic [33, 35].

### Excision Biopsy:

Excision biopsy was also called a surgical biopsy. A whole organ or a whole lump is removed (excised) under local or general anaesthetic. Some surgeons prefer excisional biopsies of most breast lumps to ensure the greatest diagnostic accuracy [33, 36].

## **3.9 Machine learning:**

Machine learning (ML) defined it as “a subfield of artificial intelligence (AI) that given computers the ability to learn without being explicitly programmed”. ML composed of algorithms that educate computers to carry out functions required. ML characterized by automating the process of automation, where intakes data and output to create a program itself. Besides, data accessibility and computation power that have encouraged widespread usage of ML. Based on these key factors, ML models can extract patterns from huge amounts of data and developing its own algorithms as and when the data changes [37, 38].

In general, the aim of ML was to understand the structure of data and fit it into models that people can understand and use. Effectiveness and efficiency of solution ML depend on the nature and characteristics of data and the performance of the learning algorithms. In the area of ML algorithms, classification analysis,

regression, data clustering, feature engineering and dimensionality reduction, association rule learning, or reinforcement learning techniques exist to effectively build data-driven systems [38].

### 3.10 Types of Machine Learning Algorithms:

ML algorithms can be divided into three categories as supervised learning, unsupervised learning, and reinforcement learning.

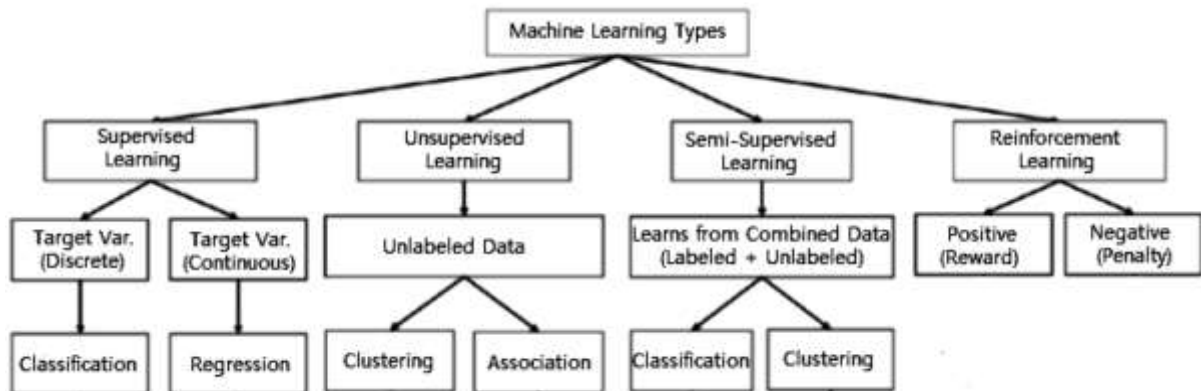


Figure 3. 2: Various types of machine learning techniques[38]

#### 3.11.1 Supervised Learning:

In supervised learning, the computer is provided with labeled inputs and outputs. The purpose of this method was for the ability of the algorithm to learn by comparing actual output and taught outputs to find errors and modify the model accordingly. Supervised learning, therefore, uses patterns to predict label values on additional unlabeled data. The most common supervised tasks are classification which separates the data, and regression which fits the data[38].

#### 3.11.2 Unsupervised Learning:

Unsupervised learning analyzes unlabeled datasets by finding commonalities among their input data. As unlabeled data are more abundant than labeled data. So, widely used for extracting generative features, identifying meaningful trends and structures, groupings in results, and exploratory purposes. The most common unsupervised learning tasks are clustering, density estimation, feature learning, dimensionality reduction, finding association rules, anomaly detection, etc. [38].

### 3.11 Machine learning Services:

All ML algorithms need an input data to achieve different goals. Supervised ML algorithms can further be divided into two subcategories, classification and regression. The types of services that can be solved by applying ML:

#### 3.12.1 Regression:

Regression tasks mainly deal with estimation of numerical values (continuous variables). Some of the following ML methods could be used for solving regressions problems:



- Kernel regression
- Gaussian process regression
- Regression trees
- Linear regression
- Support vector regression

### 3.12.2 Classification:

Classification can be defined as mapping data into predefined labels or categories, The outputs obtained are not in continuous form.

Classification services was simply related with predicting a category of a data (discrete variables). Some of the common use cases could be found in the area of healthcare such as whether a person is suffering from a particular disease or not. The ML methods such as following could be applied to solve classification tasks:

- Kernel discriminant analysis (Higher accuracy)
- K-Nearest Neighbors (Higher accuracy)
- Artificial neural networks (ANN) (Higher accuracy)
- Support vector machine (SVM) (Higher accuracy)
- Random forests (Higher accuracy)
- Decision trees
- Boosted trees
- Logistic regression
- Naive Bayes
- Deep learning

### 3.12 Machine Learning Workflow:

A Figure 3.3 depicted the ML workflow in the following diagram:

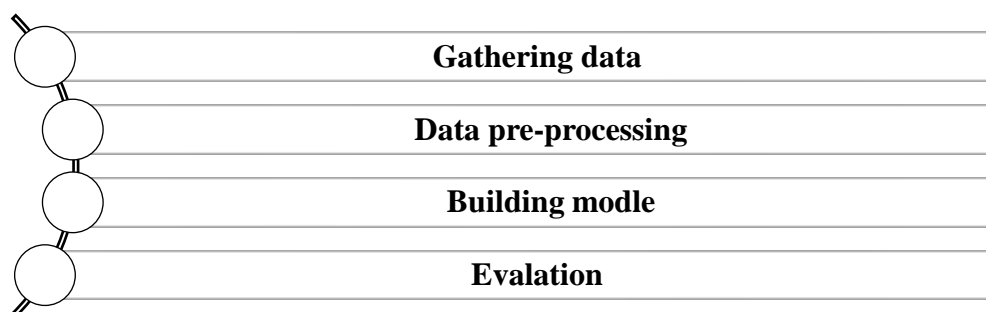


Figure 3. 3: Machine Learning Workflow

The sequence of workflow starts with gathering data ingestion that is followed by data preprocessing involves cleaning, verifying, and formatting data into a usable dataset. The building datasets phase involves the use of ML algorithm that generates a trained model, which starts breaking processed data into training, validating, and testing. Finally, the evaluation of the desired accuracy of the model is then deployed so that the application can use the model.

### 3.13 Random Forest Algorithm:

Random forest (RF) was created by Tin Kam Ho in 1995. A RF was a supervised machine learning technique that constructed from decision tree which means it used ensemble learning to solve the problem by combining many classifiers. Figure 3.4 shows a diagram of a RF [23, 39].

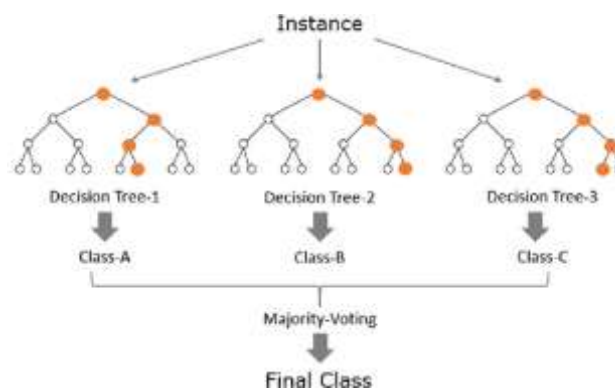


Figure 3. 4: diagram of random forest[40]

The random forest algorithm can split into two stages: Random Forest creation, Random Forest prediction.

The training process of random forests can be described by algorithm as follow:

1. Randomly select  $k$  features from total  $m$  features.

Where  $k \ll m$

2. Among the  $k$  features, calculate the node  $d$  using the best split point.
3. Split the node into daughter nodes using the best split.
4. Repeat 1 to 3 steps until  $l$  number of nodes has been reached, or until form the tree with a root node and having the target as the leaf node.
5. Build forest by repeating steps 1 to 4 for  $n$  number times to create  $n$  number of trees. Or to create  $n$  randomly created trees.

the testing process of the random forests:

1. Takes the test features and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome (target).
2. Calculate the votes for each predicted target.

3. Consider the high voted predicted target as the final prediction from the random forest algorithm.

To perform the prediction using the trained random forest algorithm. Need to pass the test features through the rules of each randomly created tree. Each random forest will predict different target (outcome) for the same test feature. Then by considering each predicted target, votes will be calculated. This concept of voting is known as majority voting.

### **3.14 Background on Python:**

Python is a general-purpose, interpreted high-level programming language. GuidoVan Rossum developed Python. a general-purpose language, which means it's designed to be used in a range of applications, including data science, software and web development, automation, and generally getting stuff done. Popular IDEs for Python are Spyder, IPython Notebook. Python is a programming language, using an open source library such as pandas, SciPy, NumPy, Sckikitlearn, Matplotlib [41].

## Chapter Four

### Proposed System (Methodology)

#### 4.1 View of the Proposed System:

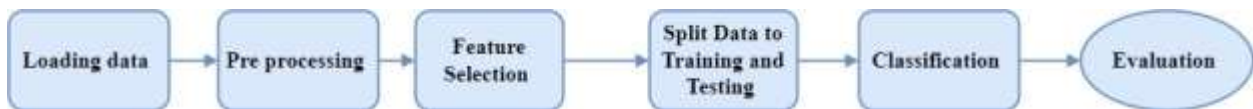


Figure 4. 1: Flowchart of breast cancer proposed system

The flowchart of the proposed system was shown above in Figure 4.1. The breast cancer (BC) dataset used has been taken from the University of California Irvine machine learning repository. Firstly, datasets were preprocessed by removing missing and duplication data, then follow by feature selection to reduce the attribute, then split the reminding data into training and testing data, then data classified by random forest (RF) into benign or malignant, and finally, the classification model performance has been evaluated. Additionally, the nine cytological attributes have been evaluated as being useful in separating benign from malignant, and a comparison of the proposed approach with the methods proposed by other researchers.

#### 4.2 Dataset Description:

The dataset used in this research; Breast Cancer Wisconsin Original Dataset abbreviated WBCD was the benchmark dataset. WBCD was created by Dr. William H. Wolberg at the University of Wisconsin Madison Hospitals Madison which took two years to collect using fine needle aspirate (FNA) from human breast tissue, the data was present in the UCI machine learning repository, which is publicly available. To create the dataset, Dr. Wolberg took a fluid sample from the breast mass of volunteers by a type of biopsy called FNA then the sample was placed under the microscope and converted into a digital image to extract the characteristics of the cell nuclei using image processing techniques.

The database consists of 699 instances, 11 real-world attributes, and two classification class, each attribute has a range of 1 to10, where the value 1 indicates that the feature was most non-cancerous or benign which represent 458 (34.5%) of data, while the value 10 indicates most cancerous or malignant which represent 241 (65.5%) of the dataset. Figure 4.2 below shows an example of WBCD database details. The complete details of all the eleven attributes were shown below in Table 4.1.

```
In [3]: Data
Out[3]:
```

id	clump_thickness	size_uniformity	shape_uniformity	marginal_adhesion	epithelial_size	bare_nucleoli	bland_chromatin	normal_nucleoli	mitoses	class
1000026	5	1	1	1	2	1	3	1	1	2
1002945	5	4	4	5	7	10	3	2	1	2
1015426	3	1	1	1	2	2	3	1	1	2
1018277	6	8	8	1	3	4	3	7	1	2
1017023	4	1	1	3	2	1	3	1	1	2
776716	3	1	1	1	3	2	1	1	1	2
841769	3	1	1	1	2	1	1	1	1	2
888820	5	10	10	3	7	3	8	10	2	4
897471	4	8	8	4	3	4	10	8	1	4
897471	4	8	8	5	4	5	10	4	1	4

ows x 11 columns

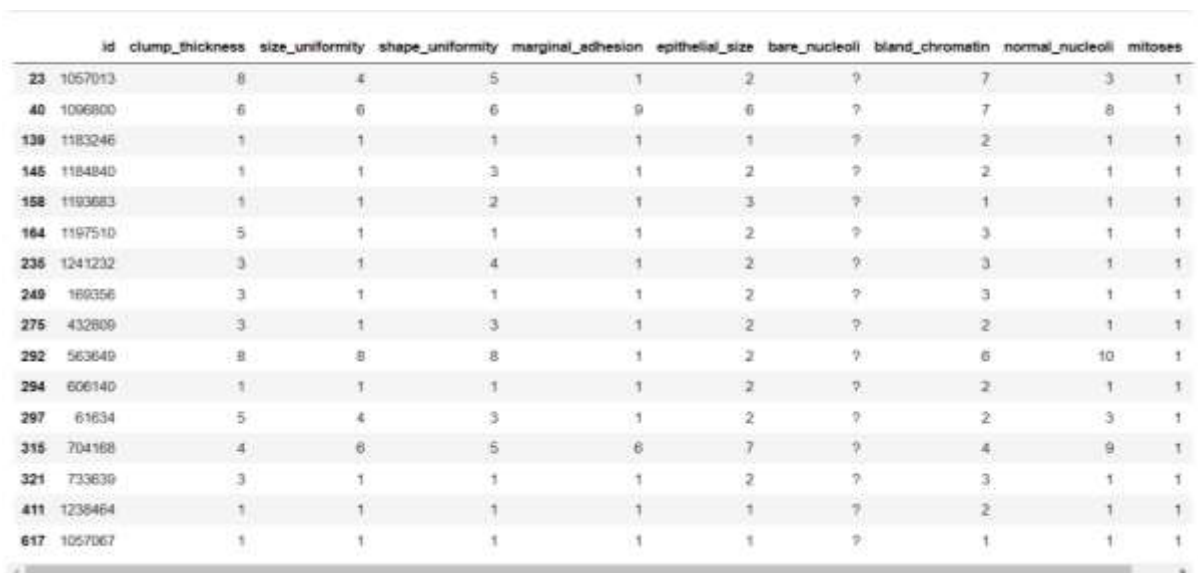
Figure 4. 2: Example of WBCD database details

Table 4- 1: Feature of Wisconsin Breast Cancer Dataset with Description.

Attribute	Domain	Description
Case Id	#ID	Identification-number
Clump thickness	1-10	Benign cells grouped in monolayer, while malignant in multilayer
Uniformity of cell size	1-10	Attribute plays a vital role in identifying cell is cancerous or not based on size
Uniformity of cell shape	1-10	Attribute plays a vital role in identifying cell is cancerous or not based on shape
Marginal adhesion	1-10	Normal cells have strong adhesion, while malignant cells lose adhesion
Single epithelial cell size	1-10	Large size epithelial cell indicate malignancy
Bare nuclei	1-10	Bare nuclei without cytoplasm coating which are found in benign tumors but not in malignant.
Bland chromatin	1-10	Represent uniform texture in benign cells and harsh texture in malignant
Normal nucleoli	1-10	Nucleaous appear small in normal cell while become more prominent in cancerous cell
Mitoses	1-10	It was the process in cell division by which nucleus divides.
class	Benign 2, Malignant 4	Benign: 458 (65.5%) Malignant: 241 (34.5%)

### 4.3 Data Preprocessing:

The WBCD dataset was obtained from the UCI machine learning repository as a (.CSV) file, then the (CSV) file has been reading in Python using the (Pandas) library using the keyword “read\_csv”. WBCD dataset consisted of 699 instances and 11 attributes. All 11 attributes provide precise information about the BC. Moreover, the dataset was checked for any unidentified values, discrepancies, or inaccurate data which can have a consequential effect on the interpretations that can be derived from the data. It has been previously reported that there were 16 instances with missing values, which all related to the "Bare Nuclei" attribute which was denoted by “?” in the WBCD dataset 15 instances were from the benign class, and one instance was from the malignant class. Since the number of missing values was small and for the interest of maintaining data consistency, these 16 instances were removed by replacing the missing value with the “NaN” constant from the (NumPy) library and then removed from the dataset by “dropna”. So, the data decreased to 683 instances. Figure 4.3 shows the 16 missing values.



	id	clump_thickness	size_uniformity	shape_uniformity	marginal_adhesion	epithelial_size	bare_nucleoli	bland_chromatin	normal_nucleoli	mitoses
23	1057013	8	4	5	1	2	?	7	3	1
40	1006800	6	6	6	9	6	?	7	8	1
139	1183246	1	1	1	1	1	?	2	1	1
145	1184840	1	1	3	1	2	?	2	1	1
158	1183683	1	1	2	1	3	?	1	1	1
164	1197510	5	1	1	1	2	?	3	1	1
236	1241232	3	1	4	1	2	?	3	1	1
249	189356	3	1	1	1	2	?	3	1	1
275	432809	3	1	3	1	2	?	2	1	1
292	563649	8	8	8	1	2	?	6	10	1
294	606140	1	1	1	1	2	?	2	1	1
297	61634	5	4	3	1	2	?	2	3	1
316	704168	4	6	5	6	7	?	4	9	1
321	733639	3	1	1	1	2	?	3	1	1
411	1238464	1	1	1	1	1	?	2	1	1
617	1057067	1	1	1	1	1	?	1	1	1

Figure 4. 3: WBCD missing data

It has been noticed that there were 16 instances duplicated by the “duplicated” keyword in Python. So, one instance was kept for each duplicate instance using the “drop\_duplicates” keyword. Figure 4.4 shows this duplicated data. Now the number of data was 675 instances.

id	clump_thickness	size_uniformity	shape_uniformity	marginal_adhesion	epithelial_size	bare_nucleoli	bland_chromatin	normal_nucleoli	mitoses	Class
1218860	1	1	1	1	1	1	3	1	1	2
1100524	6	10	10	2	8	10	7	3	3	4
1118118	9	10	10	1	10	8	3	3	1	4
1198541	3	1	1	1	2	1	3	1	1	2
320675	3	3	5	2	3	10	7	1	1	4
704097	1	1	1	1	1	1	2	1	1	2
1321942	5	1	1	1	2	1	3	1	1	2
466006	1	1	1	1	2	1	1	1	1	2

Figure 4. 4: WBCD duplicated data

The "class" attribute value has been replaced with 0, 1 instead of 2, and 4 for benign, and malignant respectively, turning it into a binary class dataset. So, 439 (65%) instances were benign and 236 (35%) instances were malignant, which ends the data preprocessing step. Figure 4.5 show an example of data after preprocessing.

id	clump_thickness	size_uniformity	shape_uniformity	marginal_adhesion	epithelial_size	bare_nucleoli	bland_chromatin	normal_nucleoli	mitoses	Class
100002E	5	1	1	1	2	1	3	1	1	0
1002945	5	4	4	5	7	10	3	2	1	0
1015425	3	1	1	1	2	2	3	1	1	0
1018277	6	8	8	1	5	4	3	7	1	0
1017023	4	1	1	3	2	1	3	1	1	0
...	...	...	...	...	...	...	...	...	...	...
778715	3	1	1	1	3	2	1	1	1	0
841789	2	1	1	1	2	1	1	1	1	0
888620	5	10	10	3	7	3	8	10	2	1
887471	4	8	8	4	3	4	10	8	1	1
887471	4	8	8	5	4	5	10	4	1	1

rows x 11 columns

BC.shape  
(675, 11)

Figure 4. 5: Example of WBCD dataset after preprocessed

#### 4.4 Feature Selection:

Feature selection (FS) was an important process in supervised machine learning and was always used before classification data. Because not all of the features were equally important and not all helped in the model accuracy. FS plays a significant role in enhancing data comprehensibility, data visualization as well as reducing the time to train a classification model, and improving the prediction results. In this research, the goal of performing the FS process was to identify the features of greatest importance in classification and thus improve classification accuracy, was applied the RF as a feature identification technique with deferent values using the "feature\_selection" library, especially "SelectFromModel".

The "case Id" attribute did not affect the "class" variable. So, to reduce the dimensionality of the dataset and to avoid incorporating insignificant features, the "case Id" column has been discarded.

## 4.5 Dataset Splitting:

Since any classification system needs to be training and testing data, splitting the data into training and testing data using stratified k-fold the “model\_selection” library, especially “train\_test\_split” by inter input, target, and testing or training or both percentages, which was performed nine times with different percentages of the training data utilized. The best was when WBCD dataset has been split into 10% for tests and 90% for training.

Each tree of the RF can calculate the importance of a feature according to its ability to increase the pureness of the leaves. The higher the increment in leaf purity, the higher the importance of the feature. This is done for each tree, then averaged among all the trees and, finally, normalized to 1. So, the sum of the importance scores calculated by a RF is 1.

## 4.6 Classification:

Random Forest (RF) was a popular machine learning algorithm that part of to the supervised learning technique. It can be used for both classification and regression problems. It was based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. RF classifier contains a number of decision trees on various subsets of the given dataset and takes the prediction from each tree, and based on the majority votes of predictions, it predicts the final output.

## 4.7 Model Performance by Validation Measure:

the performance of the proposed BC classification model was evaluated using the accuracy, sensitivity, specificity, receiver operating characteristic curve (ROC), and area under the ROC curve (AUC).

**Accuracy** defined as how close a measured value is to the actual (true) value, can be obtained by the following equation.

$$\text{Accuracy} = \frac{TN+TP}{TP+TN+FN+FP}$$

For TP, TN, FP, FN represents True Positive, True Negative, False Positive, and False Negative.

Where:

TP: Correctly classified as having BC and actually have.

TN: Correctly classified as not having BC and actually don't have.

FP: Classified as having BC but actually don't have BC

FN: Classified as not having BC but actually have BC.



**Sensitivity** defines the probability that the algorithm was classified malignant among those with the malignant cases, can be obtained by the following equation:

$$\mathbf{Sensitivity} = \frac{TP}{TP + FN}$$

**Specificity** defines the probability that the algorithm was classified benign among those with the benign cases, can be obtained by the following equation:

$$\mathbf{Specificity} = \frac{TN}{TN + FP}$$

**ROC curve** was a graph showing the performance of a classification model at all classification thresholds.

**AUC curve** measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the performance of the model at distinguishing between the malignant and benign classes.

#### **4.8 Implementation Detail:**

Python programming language was used for implementing the proposed RF classifier approaches. Furthermore, different packages and libraries such as Scikit-Learn library for statistical and machine learning analysis, also used Matplotlib library for plotting and visualization, and NumPy package used for numerical computing.

## Chapter Five

### Results and Discussion

The proposed work was performed in two parts, first, WBCD applied a machine learning algorithm especially random forest (RF) without reducing the features, using Python which uses the publicly available open-source machine learning software. Second, RF has been used to reduce the dimension of the feature space from nine to eight, six and four attributes based on extracted RF, then using RF for classification of the model. Finally, measure the performance of the classification model by applying accuracy, precision, and area under the curve (AUC) for both parts.

#### 5.1 Result of Breast Cancer Classification Using Random Forest:

A RF algorithm was used for the classification of 439 benign and 236 malignant cases with nine attributes to decisions making for diagnosing BC based on FNA instances. According to the result shown on the Figure 5.1, the optimal distinction between benign and malignant when 20% of training data, resulting in an accuracy of 97.77% with 15 forest trees. Conversely, the worst distinction was recorded at 94.11% with 10% training data with 15 trees. Figure 5.2 shows sample of decision tree from the RF.

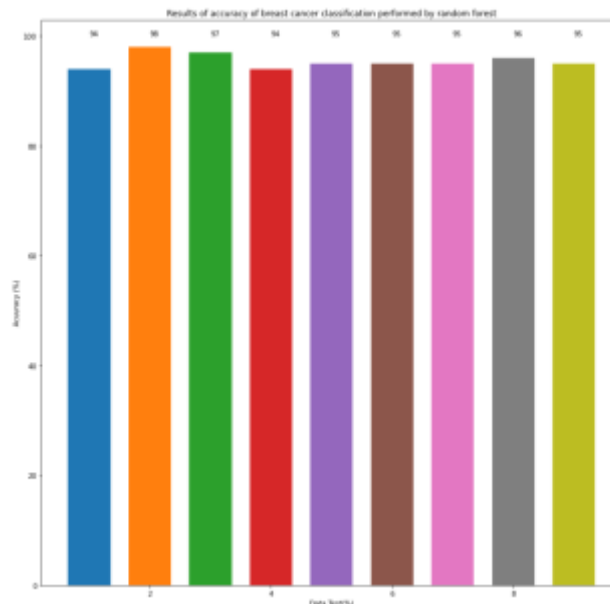


Figure 5. 1: Results of accuracy of the breast cancer classification performed by random forest without feature selection.

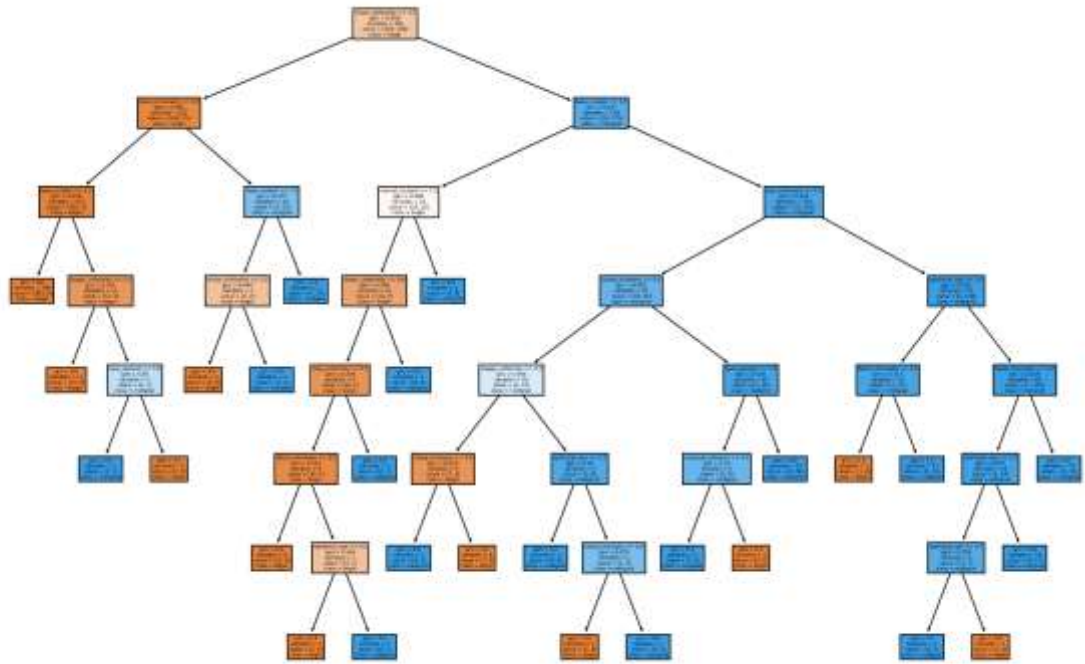


Figure 5. 2 : Sample of decision tree from the random forest

### 5.2 Performance Evaluation of Random Forest Classification:

The classification report shows a text report showing the main classification metrics (precision, recall, F1-score, support size (number of elements in each class), and accuracy) for the test data to measure the quality of predictions from the RF classification algorithm.

Random Forest classification report without feature selection				
	precision	recall	f1-score	support
benign	0.96	1.00	0.98	80
malignant	1.00	0.95	0.97	55
accuracy			0.98	135
macro avg	0.98	0.97	0.98	135
weighted avg	0.98	0.98	0.98	135

Figure 5. 3: Classification report for random forest classification without feature selection

Figure 5.3 shows that the accuracy of 80 instances of benign and 55 malignant which was 20% of test data was 98%.

The confusion matrix was summary of prediction results on a classification used to evaluate the quality of the output of a classifier on the WBCD, where the diagonal elements represent the amount of data for which the predicted class was equal to the true class, while off-diagonal elements that mislabeled by the classifier. The higher the diagonal values of the confusion matrix the better, indicating many correct predictions.

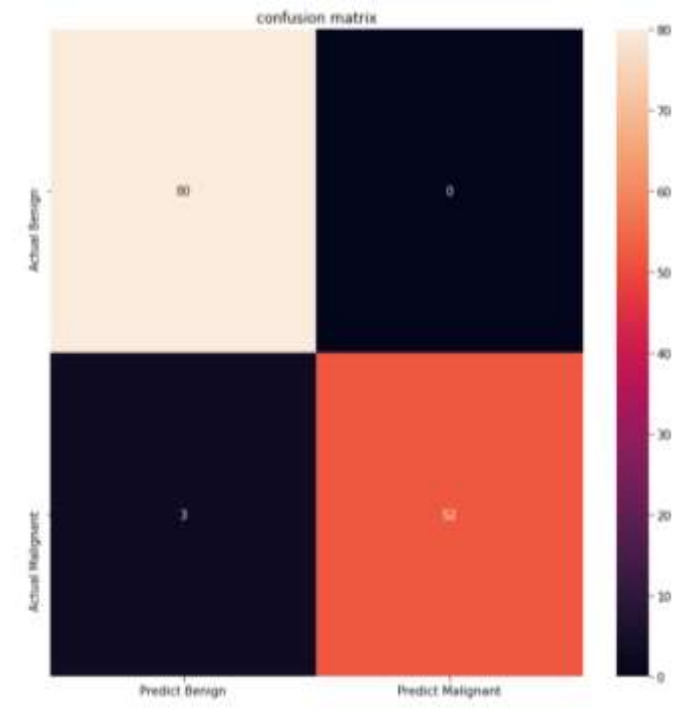


Figure 5. 4: Confusion matrix for random forest classification without feature selection

Figure 5.4 shows the confusion matrix without feature selection, in which 94.55% of data was classified correctly as malignant, and 100% of data was classified correctly as benign.

From confusion matrix, TP=52, TN=80, FP=0, and FN=3.

$$\text{Sensitivity} = \frac{TP}{TP+FN} \times 100 = \frac{52}{52+3} \times 100 = \mathbf{94.55\%} \quad (5.1)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \times 100 = \frac{80}{80+0} \times 100 = \mathbf{100\%} \quad (5.2)$$

$$\text{Accuracy} = \frac{TN+TP}{TP+TN+FN+FP} \times 100 = \frac{80+52}{80+52+3+0} \times 100 = \mathbf{97.77\%} \quad (5.3)$$

The AUC measure of the ability of a classifier to distinguish between classes and used as a summary of the ROC curve. The higher the AUC, the better the

performance of the model at distinguishing between the malignant and benign classes.

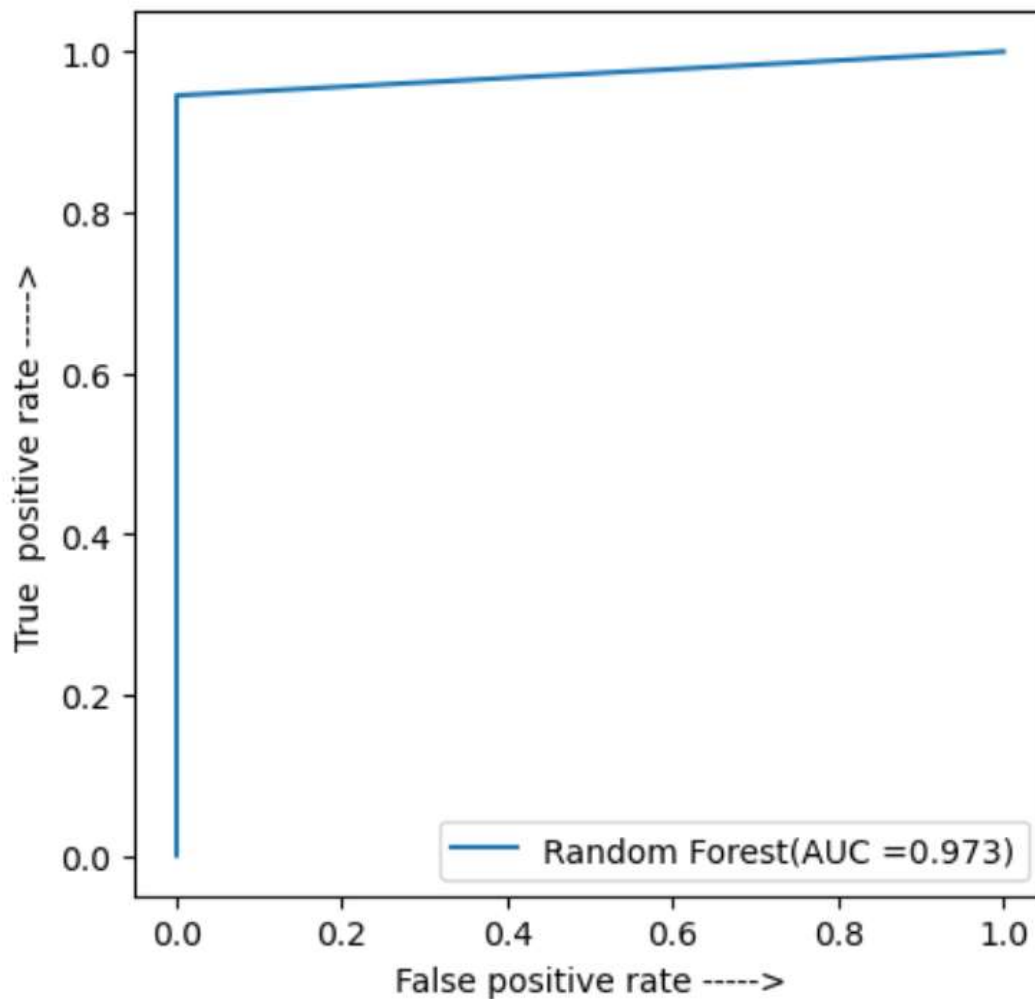


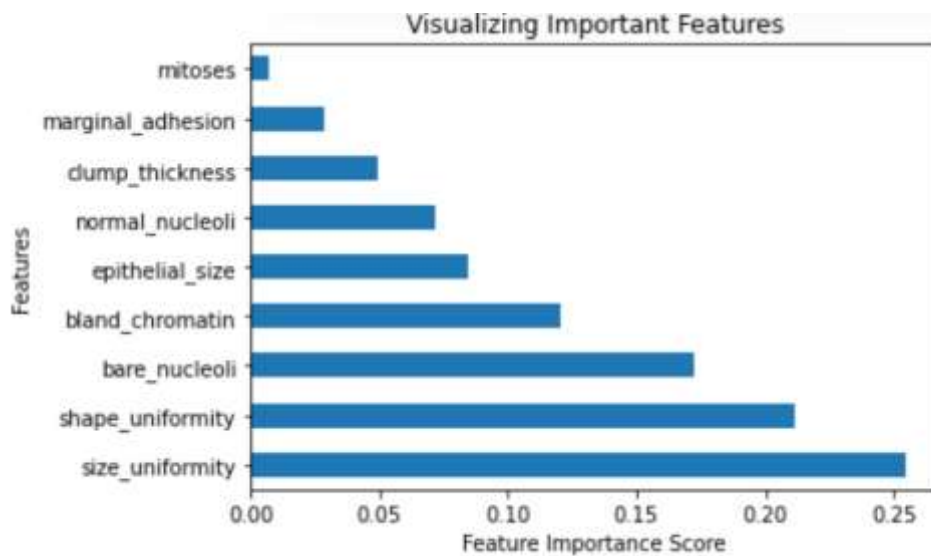
Figure 5. 5: AUC of ROC for random forest classification without feature selection

Figure 5.5 show the AUC of RF classifier without feature selection was 0.973, which indicator of ability of RF classifier to distinguish between benign and malignant tumor.

### 5.3 Result of Random Forest Feature Selection:

The result of feature selection determines whether the feature was important or not as shown in Table 5-1. According to the result shown in Table5-1, the important features were clump thickness, size uniformity, shape uniformity, epithelial size, bare nucleoli, and bland chromatin. However, the RF was used to label the important features.

Table 5- 1: Result of random forest feature selection



### 5.4 Result of Breast Cancer Classification Using Random Forest with Feature Selection:

The second part, firstly, used RF-based feature selection to select the best attributes, and then, RF algorithm was used for the classification of 213 benign and 236 malignant cases with eight, six and four attributes to decisions making for diagnosing BC based on FNA instances.

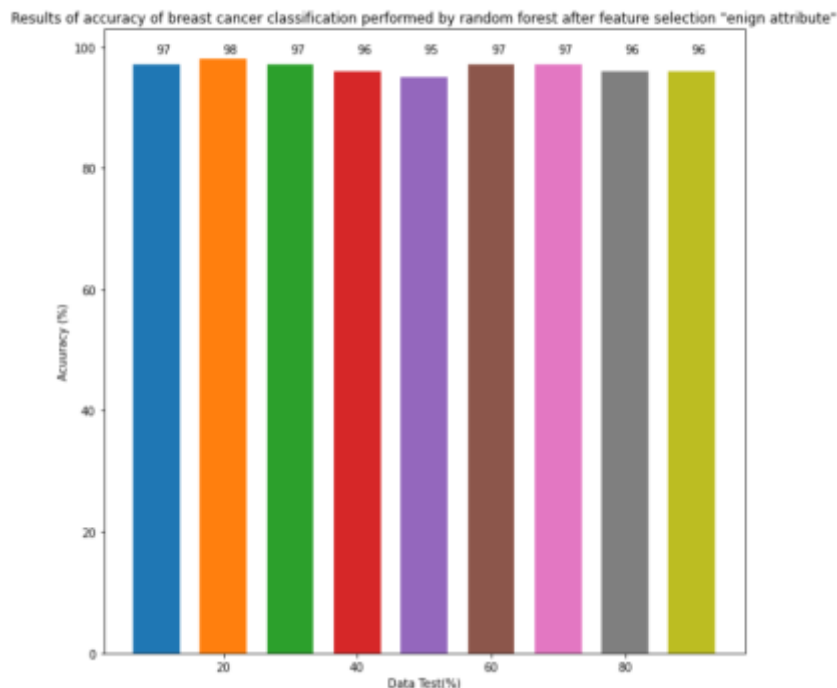


Figure 5. 6: Results of the accuracy of breast cancer classification performed by random forest with eight feature selection.

According to the result shown in Figure 5.6, the optimal distinction between benign and malignant when 20% of training data, resulting in an accuracy of 97.77% with 100 forest trees. Conversely, the worst distinction was recorded at 94.97% with 50% training data with 100 trees and eight attributes. Figure 5.7 shows a sample of the decision tree from the RF.

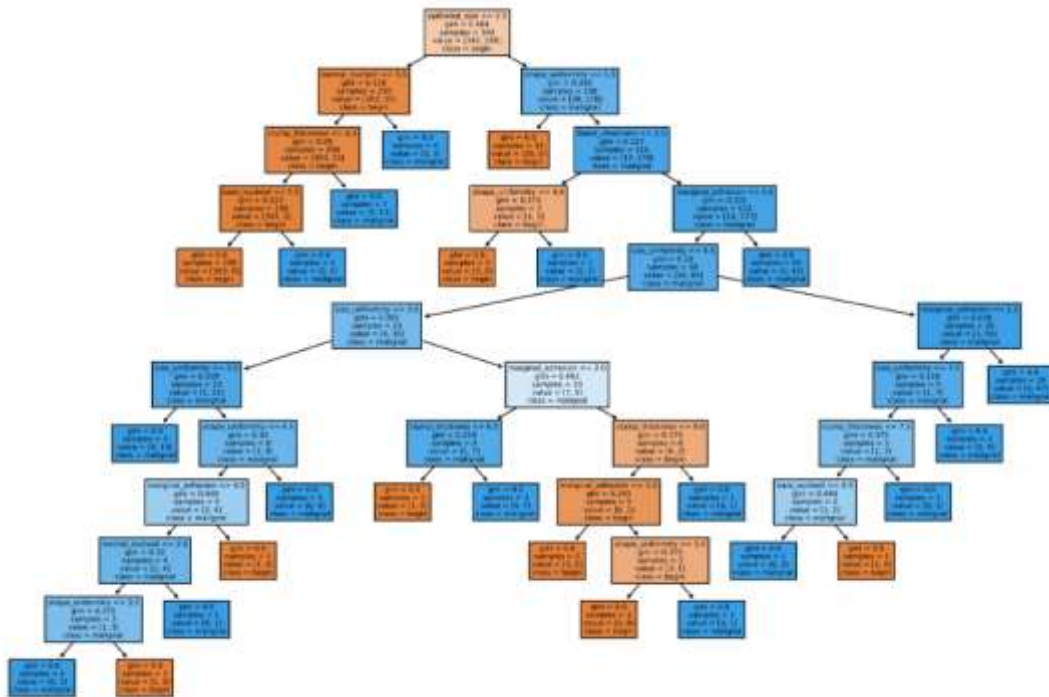


Figure 5. 7: Sample of decision tree from the random forest with eight attributes

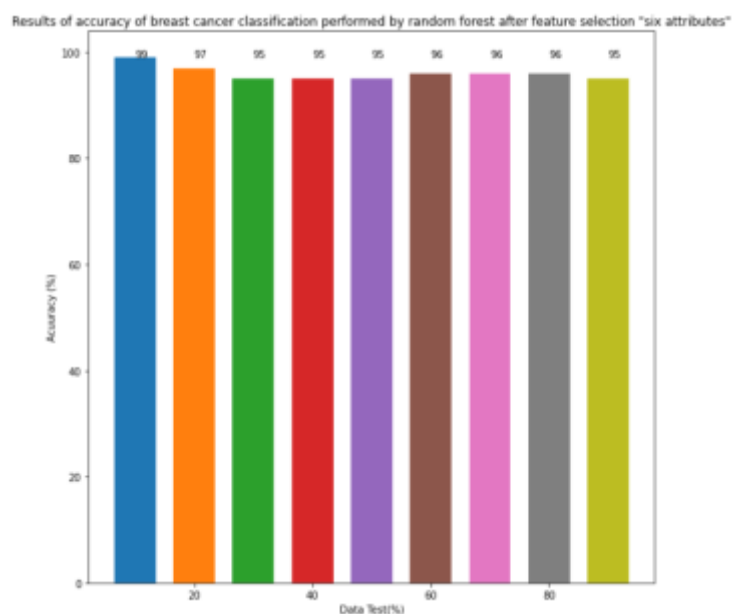


Figure 5. 8: Results of the accuracy of breast cancer classification performed by random forest with six feature selection.

According to the result shown in Figure 5.8, the optimal distinction between benign and malignant when 10% of training data, resulting in an accuracy of 98.52% with 100 forest trees. Conversely, the worst distinction was recorded at 94.67% with 50% training data with 100 trees and six attributes. Figure 5.9 shows a sample of the decision tree from the RF.

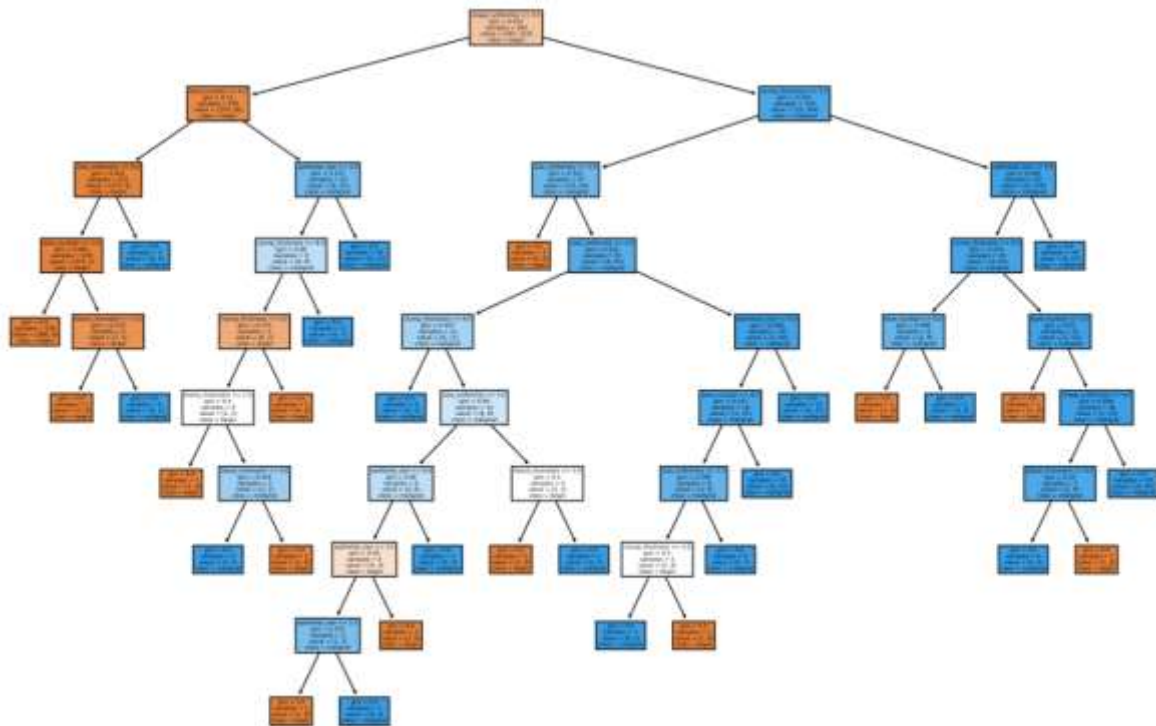


Figure 5. 9: Sample of decision tree from the random forest with six attributes

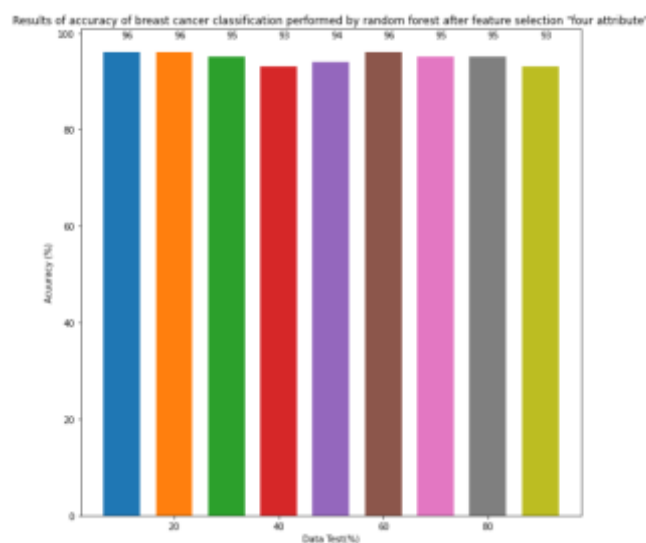


Figure 5.10: Results of the accuracy of breast cancer classification performed by random forest with four feature selection



According to the result shown in Figure 5.10, the optimal distinction between benign and malignant when 10% of training data, resulting in an accuracy of 95.58% with 100 forest trees. Conversely, the worst distinction was recorded at 93.33% with 40% training data with 100 trees and four attributes. Figure 5.11 shows a sample of the decision tree from the RF.

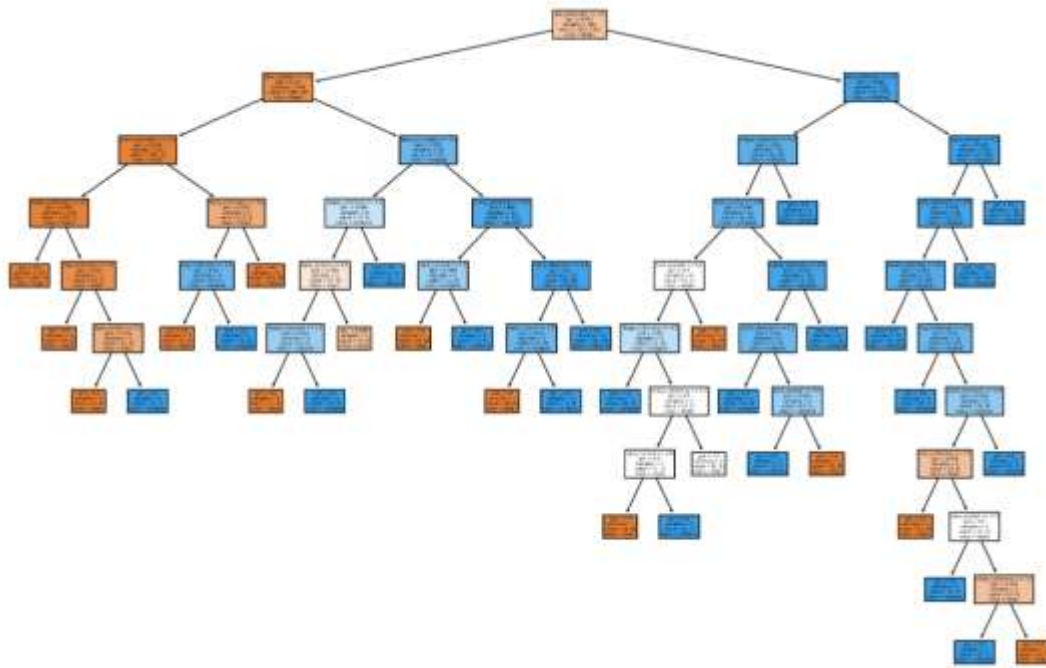


Figure 5. 11: Sample of decision tree from the random forest with four attributes

## 5.6 Performance Evaluation of Random Forest Classification with Feature Selection:

Random Forest classification report with featur selection				
	precision	recall	f1-score	support
benign	0.99	0.98	0.98	93
malignant	0.95	0.98	0.96	42
accuracy			0.98	135
macro avg	0.97	0.98	0.97	135
weighted avg	0.98	0.98	0.98	135

Figure 5. 12: Classification report for random forest classification with eight feature selection

Figure 5.12 shows that the accuracy of 93 instances of benign and 42 malignant which was 20% of test data was 98% when reduce the feature selection to eight attributes.

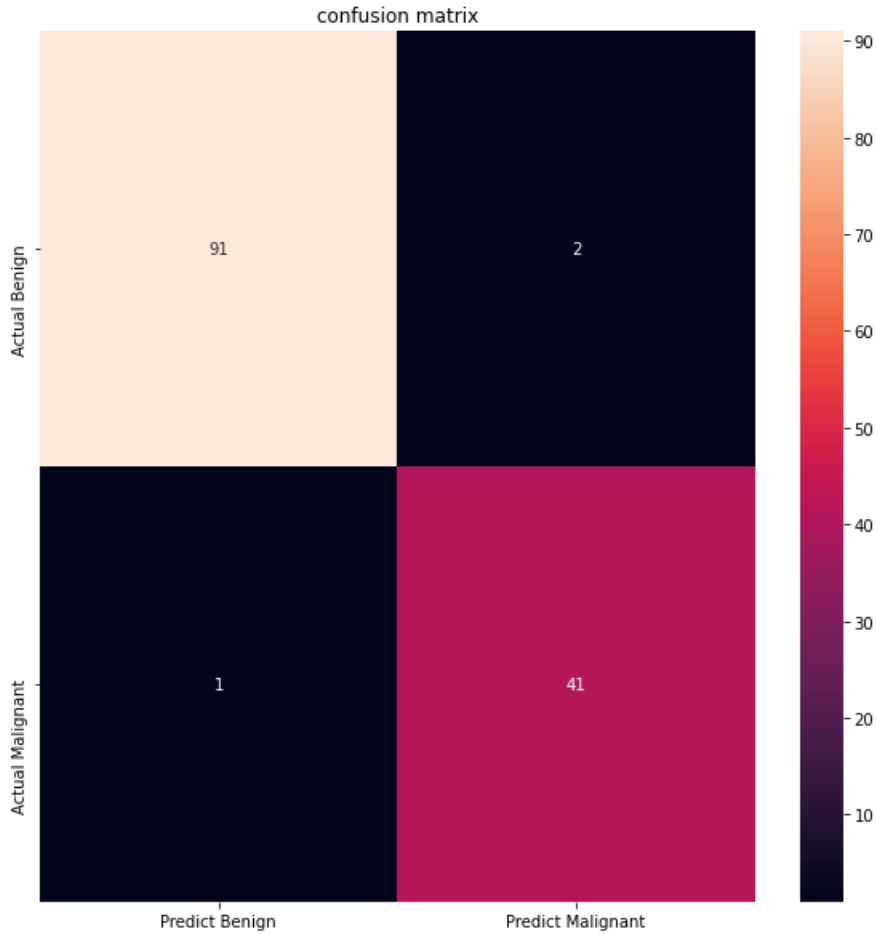


Figure 5. 13: Confusion matrix for random forest classification with eight feature selection

Figure 5.13 shows the confusion matrix with eight feature selection, in which 97.62 of data was classified correctly as malignant, and 97.85% of data was classified correctly as benign.

From confusion matrix, TP=41, TN=91, FP=2, and FN=1.

$$\text{Sensitivity} = \frac{TP}{TP+FN} \times 100 = \frac{41}{41+1} \times 100 = \mathbf{97.62\%} \quad (5.4)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \times 100 = \frac{91}{91+2} \times 100 = \mathbf{97.85\%} \quad (5.5)$$

$$\text{Accuracy} = \frac{TN+TP}{TP+TN+FN+FP} \times 100 = \frac{91+41}{91+41+1+2} \times 100 = \mathbf{97.77\%} \quad (5.6)$$

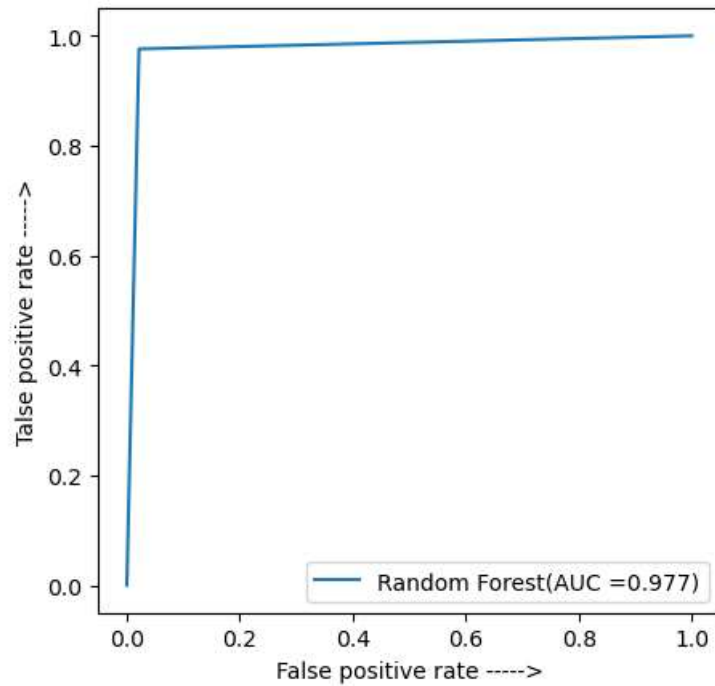


Figure 5.14: AUC of ROC for random forest classification with eight feature selection

Figure 5.14 show the AUC of RF classifier with eight feature selection was 0.977, which indicator of ability of RF classifier to distinguish between benign and malignant tumor.

Random Forest classification report with featur selection				
	precision	recall	f1-score	support
benign	1.00	0.98	0.99	51
malignant	0.94	1.00	0.97	17
accuracy			0.99	68
macro avg	0.97	0.99	0.98	68
weighted avg	0.99	0.99	0.99	68

Figure 5. 15: Classification report for random forest classification with six feature selection

Figure 5.15 shows that the accuracy of 51 instances of benign and 17 malignant which was 10% of test data was 99% when reduce the feature selection to six attributes.

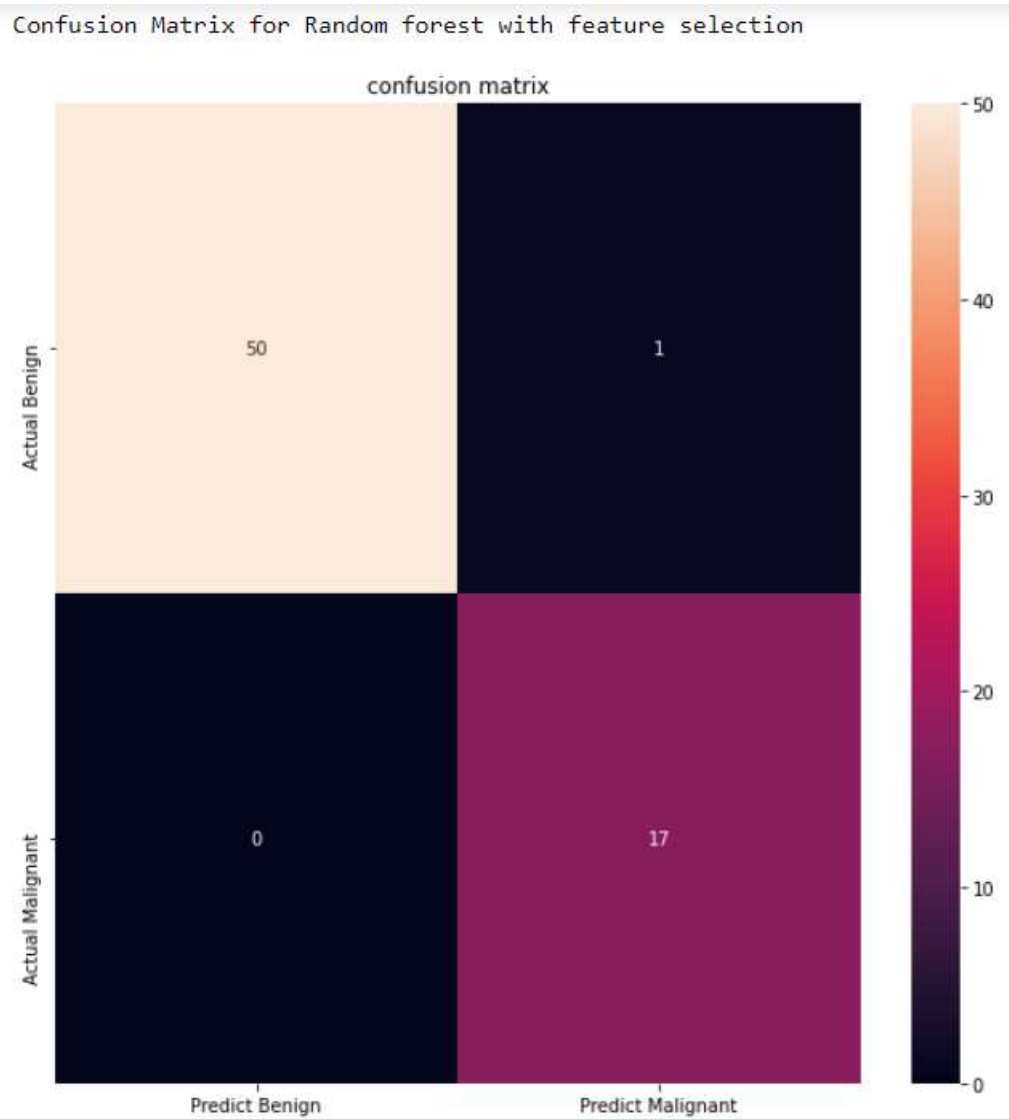


Figure 5. 16: Confusion matrix for random forest classification with six feature selection

Figure 5.16 shows the confusion matrix with six feature selection, in which 100% of data was classified correctly as malignant, and 98.04% of data was classified correctly as benign.

From confusion matrix, TP=17, TN=50, FP=1, and FN=0.

$$\text{Sensitivity} = \frac{TP}{TP+FN} \times 100 = \frac{17}{17+0} \times 100 = \mathbf{100\%} \quad (5.7)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \times 100 = \frac{50}{50+1} \times 100 = \mathbf{98.04\%} \quad (5.8)$$

$$\text{Accuracy} = \frac{TN+TP}{TP+TN+FN+FP} \times 100 = \frac{17+50}{17+50+1+0} \times 100 = \mathbf{98.52\%} \quad (5.9)$$

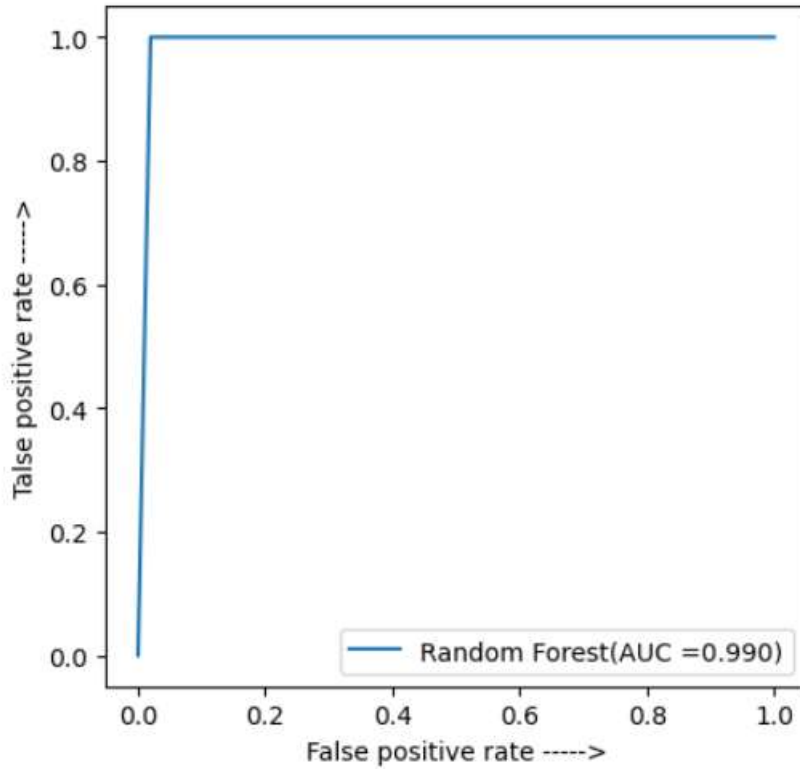


Figure 5.17: AUC of ROC for random forest classification with six feature selection

Figure 5.17 show the AUC of RF classifier with six feature selection was 0.990, which indicator of ability of RF classifier to distinguish between benign and malignant tumor.

Random Forest classification report with featur selection				
	precision	recall	f1-score	support
benign	1.00	0.94	0.97	51
malignant	0.85	1.00	0.92	17
accuracy			0.96	68
macro avg	0.93	0.97	0.94	68
weighted avg	0.96	0.96	0.96	68

Figure 5. 18: Classification report for random forest classification with four feature selection

Figure 5.18 shows that the accuracy of 51 instances of benign and 17 malignant which was 10% of test data was 96% when reduce the feature selection to four attributes.

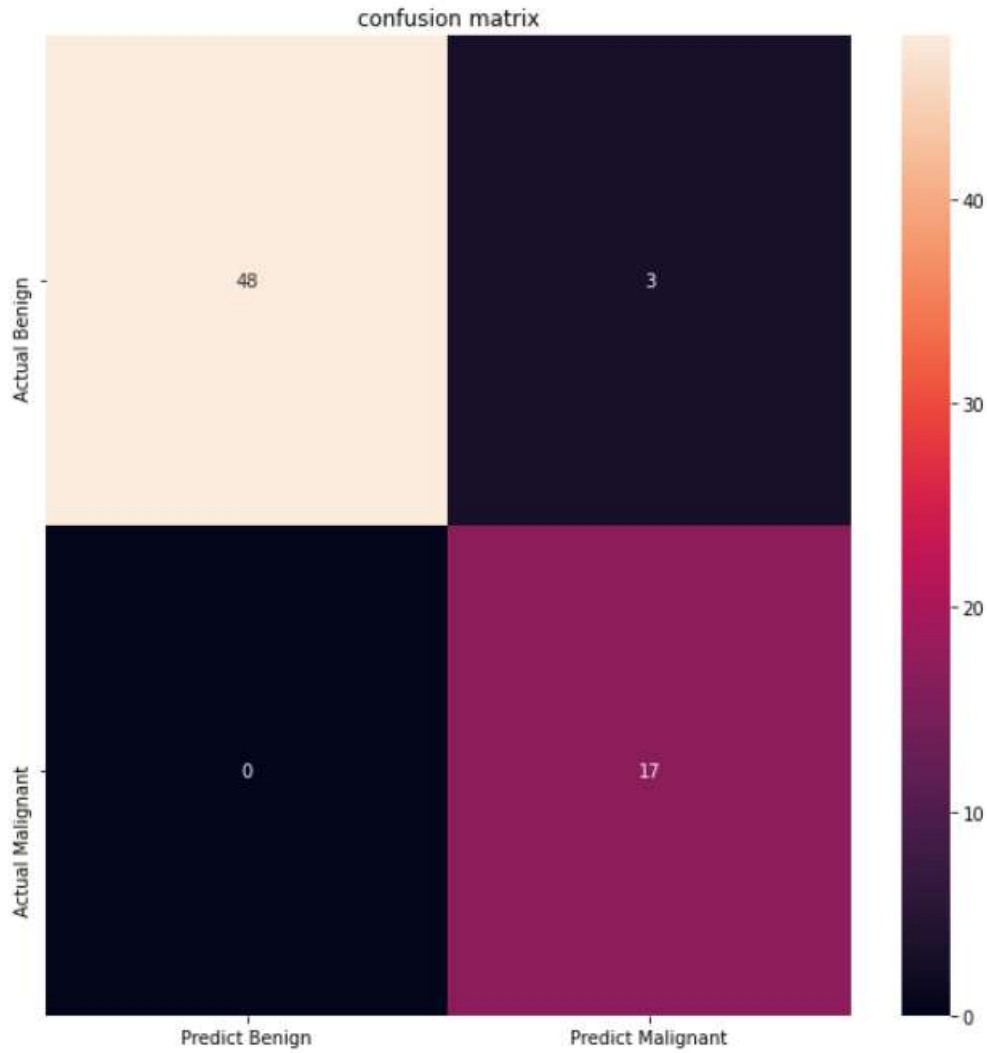


Figure 5. 19: Confusion matrix for random forest classification with four feature selection

Figure 5.19 shows the confusion matrix with four feature selection, in which 100% of data was classified correctly as malignant, and 94.12% of data was classified correctly as benign.

From confusion matrix, TP=17, TN=48, FP=3, and FN=0.

$$\text{Sensitivity} = \frac{TP}{TP+FN} \times 100 = \frac{17}{17+0} \times 100 = \mathbf{100\%} \quad (5.10)$$

$$\text{Specificity} = \frac{TN}{TP+FP} \times 100 = \frac{48}{48+3} \times 100 = \mathbf{94.12\%} \quad (5.11)$$

$$\text{Accuracy} = \frac{TN+TP}{TP+TN+FN+FP} \times 100 = \frac{17+48}{17+48+3+0} \times 100 = \mathbf{95.59\%} \quad (5.12)$$

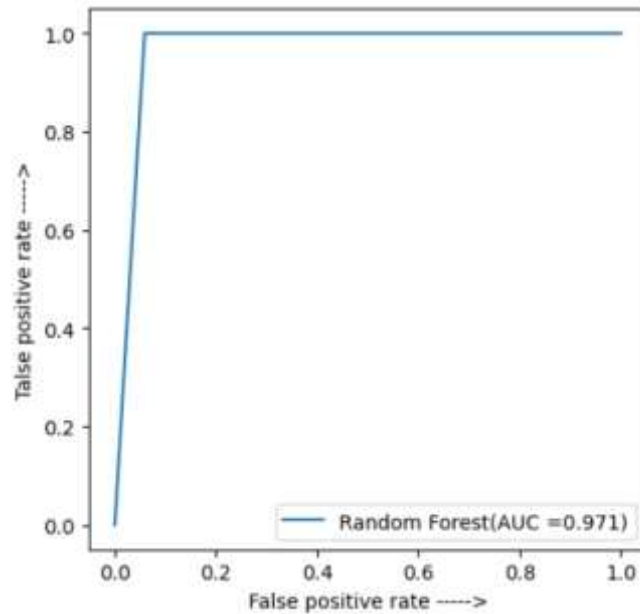


Figure 5.20: AUC of ROC for random forest classification with four feature selection

Figure 5.20 show the AUC of RF classifier with four feature selection was 0.971, which indicator of ability of RF classifier to distinguish between benign and malignant tumor.

### 5.7 Summary of Results and Discussion:

Table 5.2 summarizes the models. Although the proposed model gives the same result for accuracy when using feature selection with eight attributes or without. While the six attributes achieved the best accuracy, sensitivity, AUC, and acceptable specificity. the four attributes get the least measure with others except in specificity. Therefore, the final results that will be considered are the results of training the proposed model by using six attributes of feature selection. Table 5-3 shows the comparison between the proposed model and literature reviews.

Table 5-2: Summary of the proposed model.

model	Accuracy%	Sensitivity %	Specificity %	Precision	AUC
Nine attributes	97.77	94.55	100	98	0.973
Eight attributes	97.77	97.62	97.85	97	0.977
<b>Six attributes</b>	<b>98.52</b>	<b>100</b>	<b>98.04</b>	<b>97</b>	<b>0.990</b>
Four attributes	95.59	100	94.12	93	0.971

Table 5- 3: Comparison between the proposal with literature reviews.

Authors/ years	dataset	algorithms	Classification Accuracy (%)
<b>The Proposal model</b>	<b>WBCD</b>	<b>RF</b>	<b>98.52</b>
Yixuan Li & Zixuan Chen 2018[17]	WBCD	DT	96.1
		SVM	95.1
		RF	96.1
		LR	93.7
		ANN	95.6
R. Preetha & S. Vinila Jinny 2021[7]	WDBC	SVM	87.6
		MLP	86.3
		PCA-LDA-ANFIS	98.6
Ed-Daoudy & Maalmi 2020[11]	WBCD	AR-SVM	98.00
G Gupta al el. 2021[16]	WDBC	RF	95.0
		NB	93.0
		SVM	63.0
		LR	95.0
		DT	92.0
Gouda I. Salma et al. 2012[12]	WBCD	NB	95.9943
		MLP	95.279
		J48	95.1359
		SMO	96.9957
OI Obaid et al. 2018[15]	WDBC	SVM	98.1
		DT	93.7
		KNN	96.7

Depending on the Wisconsin breast cancer dataset, the proposed model has obtained an accuracy of 98.52%, compared with Yixuan Li et al., Ed-Daoudy et al, and Gupta al el. models, the proposed was outperformed in accuracy. And when looking at a model of Ed-Daoudy et al., obtained underperformed in accuracy, sensitivity, and specificity compared to the proposed model, except in F1-measure 98.4%, and precision 98.90%. while R. Preetha et al., MH Alshayeji et al., and M. Kumari models exceeded in accuracy by over 98.5%.

Table5-3 above shows this comparison.



## Chapter Six

### Conclusion and Recommendation

#### 6.1 Conclusion:

Breast cancer is a widespread disease that affects millions of people worldwide spatially women every year. Due to uncertainty of the physicians in decision-making during times. A machine learning technique with physician knowledge can reduce errors and improve diagnosis. This study used a random forest algorithm to classify Wisconsin breast cancer. Furthermore, a random forest was also used for the feature selection phase. This study was divided into two parts, firstly without feature selection, where the model gets 98% accuracy, 94% sensitivity, 100% specificity, 98% for both precision and f1-score, and final 0.973 AUC, second with feature selection, also divided into three models. First, the model produces 98% accuracy, 97.62% sensitivity, 97.85% specificity, 97% for both precision and f1-score, and a final 0.977 AUC when decreased the attribute from nine to eight. while the second model obtain 98.52% accuracy, 100% sensitivity, 98.04% specificity, 97% and 98 for precision and f1-score respectively, and finally 0.99 AUC when attributes was six. And the final model results were 96% accuracy, 100% sensitivity, 94% specificity, 93% precision, 94% f1-score, and a final 0.971 AUC when decreased the attribute from nine to four. According to the experiment result, RF scored the best accuracy at 98.52% using six attributes. Therefore, the results of this study are acceptable to distinguish the character of carcinoma.

#### 6.2 Recommendation:

The recommendation would include:

1. Apply the random forest model to anther diseases.
2. Apply a hybrid machine learning algorithm for this data.
3. Link the random forest model with a website and create a graphical user interface.

## References

- [1] W. H. Organization. "Cancer " [https://www.who.int/health-topics/cancer#tab=tab\\_1](https://www.who.int/health-topics/cancer#tab=tab_1) (accessed 17,sep,2022).
- [2] W. H. Organization. "Breast cancer." <https://www.who.int/news-room/fact-sheets/detail/breast-cancer> (accessed 17,sep,2022).
- [3] S. I. A. Elhassan, "The five-year survival rate of breast cancer at Radiation and Isotopes Centre Khartoum, Sudan," *Heliyon*, vol. 6, no. 8, p. e04615, 2020.
- [4] Y.-S. Sun *et al.*, "Risk factors and preventions of breast cancer," *International journal of biological sciences*, vol. 13, no. 11, p. 1387, 2017.
- [5] V. J. Kadam, S. M. Jadhav, and K. Vijayakumar, "Breast cancer diagnosis using feature ensemble learning based on stacked sparse autoencoders and softmax regression," *Journal of medical systems*, vol. 43, no. 8, pp. 1-11, 2019.
- [6] T. Octaviani and d. Z. Rustam, "Random forest for breast cancer prediction," in *AIP Conference Proceedings*, 2019, vol. 2168, no. 1: AIP Publishing LLC, p. 020050.
- [7] R. Preetha and S. V. Jinny, "Early diagnose breast cancer with PCA-LDA based FER and neuro-fuzzy classification system," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 7, pp. 7195-7204, 2021.
- [8] N. Naveed, H. T. Madhloom, and M. S. Husain, "Breast cancer diagnosis using wrapper-based feature selection and artificial neural network," *Applied Computer Science*, vol. 17, no. 3, 2021.
- [9] U. M. L. Repository. "Breast Cancer Wisconsin (Original) Data Set " <https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28original%29> (accessed 17,sep,2022).
- [10] N. Rane, J. Sunny, R. Kanade, and S. Devi, "Breast cancer classification and prediction using machine learning," *International Journal of Engineering Research and Technology*, vol. 9, no. 2, pp. 576-580, 2020.
- [11] A. Ed-daoudy and K. Maalmi, "Breast cancer classification with reduced feature set using association rules and support vector machine," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 9, no. 1, pp. 1-10, 2020.
- [12] G. I. Salama, M. Abdelhalim, and M. A.-e. Zeid, "Breast cancer diagnosis on three different datasets using multi-classifiers," *Breast Cancer (WDBC)*, vol. 32, no. 569, p. 2, 2012.
- [13] M. Kumari and V. Singh, "Breast cancer prediction system," *Procedia computer science*, vol. 132, pp. 371-376, 2018.

- [14] M. H. Alshayegi, H. Ellethy, and R. Gupta, "Computer-aided detection of breast cancer on the Wisconsin dataset: An artificial neural networks approach," *Biomedical Signal Processing and Control*, vol. 71, p. 103141, 2022.
- [15] O. I. Obaid, M. A. Mohammed, M. Ghani, A. Mostafa, and F. Taha, "Evaluating the performance of machine learning techniques in the classification of Wisconsin Breast Cancer," *International Journal of Engineering & Technology*, vol. 7, no. 4.36, pp. 160-166, 2018.
- [16] G. Gupta, M. Sharma, S. Choudhary, and K. Pandey, "Performance Analysis of Machine Learning Classification Algorithms for Breast Cancer Diagnosis," in *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*, 2021: IEEE, pp. 1-6.
- [17] Y. Li and Z. Chen, "Performance evaluation of machine learning methods for breast cancer prediction," *Appl Comput Math*, vol. 7, no. 4, pp. 212-216, 2018.
- [18] A. B. Rivard, L. Galarza-Paez, and D. C. Peterson, "Anatomy, Thorax, Breast," in *StatPearls [Internet]*: StatPearls Publishing, 2021.
- [19] R. Stevens, "Gray's Anatomy for Students," ed: The Royal College of Surgeons of England, 2006.
- [20] Y. S. Khan and H. Sajjad, "Anatomy, Thorax, Mammary Gland," 2019.
- [21] S. Verralls, *Anatomy and Physiology Applied to Obstetrics*. Churchill Livingstone, 2013.
- [22] T. U. f. I. C. C. s. (UICC). "Breast Cancer " <https://www.uicc.org/what-we-do/thematic-areas-work/breast-cancer> (accessed sep,19,2022)
- [23] C. Aroef, Y. Rivan, and Z. Rustam, "Comparing random forest and support vector machines for breast cancer classification," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 18, no. 2, pp. 815-821, 2020.
- [24] N. C. Institute. "What Is Cancer?" <https://www.cancer.gov/about-cancer/understanding/what-is-cancer> (accessed Sep,19,2022).
- [25] M. M. A. Elbasheer *et al.*, "Spatial distribution of breast cancer in Sudan 2010-2016," *PLoS One*, vol. 14, no. 9, p. e0211085, 2019.
- [26] S. E. Singletary and J. L. Connolly, "Breast cancer staging: working with the sixth edition of the AJCC Cancer Staging Manual," *CA: a cancer journal for clinicians*, vol. 56, no. 1, pp. 37-47, 2006.
- [27] S. Kalli, A. Semine, S. Cohen, S. P. Naber, S. S. Makim, and M. Bahl, "American joint committee on cancer's staging system for breast cancer: what the radiologist needs to know," *Radiographics*, vol. 38, no. 7, pp. 1921-1933, 2018.
- [28] C. f. D. C. a. P. (CDC). "What Are the Symptoms of Breast Cancer?" [https://www.cdc.gov/cancer/breast/basic\\_info/symptoms.htm](https://www.cdc.gov/cancer/breast/basic_info/symptoms.htm) (accessed 19,sep,2022).

- [29] C. f. D. C. a. P. (CDC). "How Is Breast Cancer Treated?" [https://www.cdc.gov/cancer/breast/basic\\_info/treatment.htm](https://www.cdc.gov/cancer/breast/basic_info/treatment.htm) (accessed 19,sep 2022).
- [30] M. Abdolahi, M. Salehi, I. Shokatian, and R. Reiazi, "Artificial intelligence in automatic classification of invasive ductal carcinoma breast cancer in digital pathology images," *Medical Journal of the Islamic Republic of Iran*, vol. 34, p. 140, 2020.
- [31] Erin V Newton. "Breast Cancer Screening." <https://emedicine.medscape.com/article/1945498-overview> (accessed 19,sep 2022).
- [32] A. C. Society. "Breast Cancer Early Detection and Diagnosis." (accessed 19,sep 2022).
- [33] Edward. Uthman. "All about biopsies." [https://training.seer.cancer.gov/treatment/surgery/all\\_about\\_biopsy\\_accessible.pdf](https://training.seer.cancer.gov/treatment/surgery/all_about_biopsy_accessible.pdf) (accessed).
- [34] P. E. Zapanta. "Fine-Needle Aspiration of the Salivary Glands." <https://emedicine.medscape.com/article/882291-overview#a4> (accessed 19,Sep 2022).
- [35] H. Singhal. "Breast Stereotactic Core Biopsy/Fine Needle Aspiration." <https://emedicine.medscape.com/article/1845123-overview#a5> (accessed 17,sep,2022).
- [36] M. Keating. "Diagnosing Breast Cancer " <https://www.mariekeating.ie/cancer-information/breast-cancer/diagnosing-breast-cancer/> (accessed 17,sep,2022).
- [37] P. Sodhi, N. Awasthi, and V. Sharma, "Introduction to machine learning and its basic application in python," in *Proceedings of 10th International Conference on Digital Strategies for Organizational Success*, 2019.
- [38] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN Computer Science*, vol. 2, no. 3, pp. 1-21, 2021.
- [39] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5-32, 2001.
- [40] J. Golze, S. Zourlidou, and M. Sester, "Traffic regulator detection using GPS trajectories," *KN-Journal of Cartography and Geographic Information*, vol. 70, no. 3, pp. 95-105, 2020.
- [41] G. Van Rossum, "Python Programming Language," in *USENIX annual technical conference*, 2007, vol. 41, no. 1: Santa Clara, CA, pp. 1-36.

