Arab American University

Faculty of Graduate Studies

Prediction and Classification of Breast Cancer Depending on Risk Factors Using Data Mining Techniques: Case Study Palestine

By

Wala Amjad Faqeeh

Supervisor

Prof. Mohammed Awad

This thesis was submitted in partial fulfillment of the requirements for the Master's degree in Health Informatics

June 2022

© Arab American University – Palestine 2022.

All rights reserved

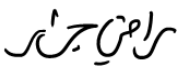
Thesis Approval

Prediction and Classification of Breast Cancer Depending on Risk Factors Using Data Mining Techniques: Case Study Palestine

By

Wala Amjad Faqeeh

This thesis was defended successfully on 00/07/2022 and approved by:

Committee Members	Signature
1. Prof. Dr. Mohammed Awad/ Supervisor:	
2. Internal Examiner: Dr. Yousef Mimi	
3. External Examiner: Dr. Radi Jarrar	

Declaration

This thesis was submitted in partial fulfillment of the requirement for Master's degree in Health Informatics.

I declare that the content of this thesis (or any part of the same) has not been submitted for a higher degree to any other university or institution.

Students Name: Wala Amjad Faqeeh

Sig: 

Date: / /

Dedication

I dedicate this study.

To my supervisor, Professor Dr. Mohammad Awad.

To my managers in the Palestinian Ministry of Health.

The Arab American University for learning and encouraging.

To my mother, father, sisters, brothers and friends.

The Palestinian Ministry of Health who gave me the chance to get a Master's degree.

To the Palestinian women who have suffered and continue to suffer from breast cancer.

Acknowledgments

This would be incomplete without the help and support of many people. Professor Mohammad Awad deserves our gratitude for his constant guidance and for sharing his knowledge and expertise in the field of Artificial Intelligence.

I would also like to thank the Department of Health Informatics, Dr. Yousef Al-Mimi and Dr. Shahnaz Al-Najjar, who commissioned us with a variety of research and projects.

As a result, I was able to achieve such a high level of achievement.

I would also like to thank the directors of the General Department of Information Technology at the Ministry of Health for their assistance in collecting the data.

In general, thanks to the Palestinian Ministry of Health.

Abstract

Breast cancer is the most frequent cancer in women worldwide. Because of increased life expectancy, the incidence of breast cancer is rising in developing countries. Breast cancer (BC) prediction and classification are important issues to detect cancer in the early stages. There are many risk factors for breast cancer. Due to global environmental problems, the way of life, and the increase in the world population, it is possible that in the coming years the number of people affected will increase. The majority of the women do not go to perform recurring screening mammograms.

This study used data mining techniques to predict and classify BC disease, and this is applied by Palestinian patients' datasets containing the diagnosis of cases (malignant, healthy and benign) and risk factors.

The dataset contains 1794 records; 635 malignant breast cancer cases, 570 healthy controls and 589 benign cases. Breast cancer cases were collected from mammography electronic Registry system in MOH Directorate.

Each record has 30 features (Variables) plus the class attributes. The class is formed as healthy, malignant, or benign cases.

Multiple data mining algorithms was used such as neural networks, boosted decision Tree, Random Forest and Support-Vector Machine. The performance of models was choosing by compared the data mining algorithms with each other.

The research evaluated and compared the performance of different machine learning (ML) algorithms in predicting breast cancer among Palestinian women and choose the best ML algorithm to develop a breast cancer prediction model.

The result shows that the best four models were developed for prediction the breast cancer. The first model for detection BC for three class (malignant, benign and healthy). The results of this model were 96.7% accuracy, 95.2% Precision, 95.1%Recall, 95.2% F1 Score and 97% AUC.

While the performance of second models for prediction the BC (Discrimination between malignant and healthy cases) was 98.5% Accuracy, 97.8% Precision, 99.3% Recall which meaning very low false negative, 98.6% F1 Score and 99.9% AUC.

The third model was developed to detection the BC (Malignant and benign) has performance with 95.0% Accuracy, 98.0% Recall and 99.5 % AUC. Finally, the model for predicting of BC (healthy and benign class) was developed with 89.4% accuracy, 90.9% Precision, 93.9% Recall which meaning very low false negative, 92.4% F1 Score and 96.1% AUC.

The models can be used for diagnosis and to assist doctors in determining whether or not a large number of women require early detection via mammography or other means. This allows women who are suspected of having breast cancer to be given priority for appointments at breast cancer clinics.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

BC	Breast Cancer
DM	Data Mining
MRI	Magnetic Resonance Imaging
AI	Artificial Inelegance
ANNs	Artificial Neural Networks
ML	Machine Learning
SVM	Support Vector Machine
AUC	Area Under Curve
ROC	Receiver Operating Characteristic Curve
TP	True Positive
TN	True Negative
FP	False-Positive
FN	False-Negative
TPR	True Positive Rate
TNR	True Negative Rate
NN	Neural network
DF	Random Forest

Chapter One: Introduction

1.1 Background

Breast cancer is top cancer in women in the world. The incidence of breast cancer is increasing in the developing world due to increased life expectancy. The mortality rates in developed countries have been dropping whereas rates in developing countries have been rising [1]. In 2020, over 2.26 million new incidences of breast cancer were diagnosed in women in the world shown in Figure 1.1 [2]. In Palestine, breast cancer is at the forefront of cancers that affect females, with 31.1% of all reported cancers, and an incidence rate of 40 new cases per year for every one hundred thousand females. While female cancer deaths reached 16.5% of all female cancer deaths in 2020[3]. Unfortunately, there is no way to completely prevent cancer. However, early detection increases the chance of breast cancer being cured and the availability of its treatment. But, if it is discovered too late, the opportunity for treatment is often missed. [4].

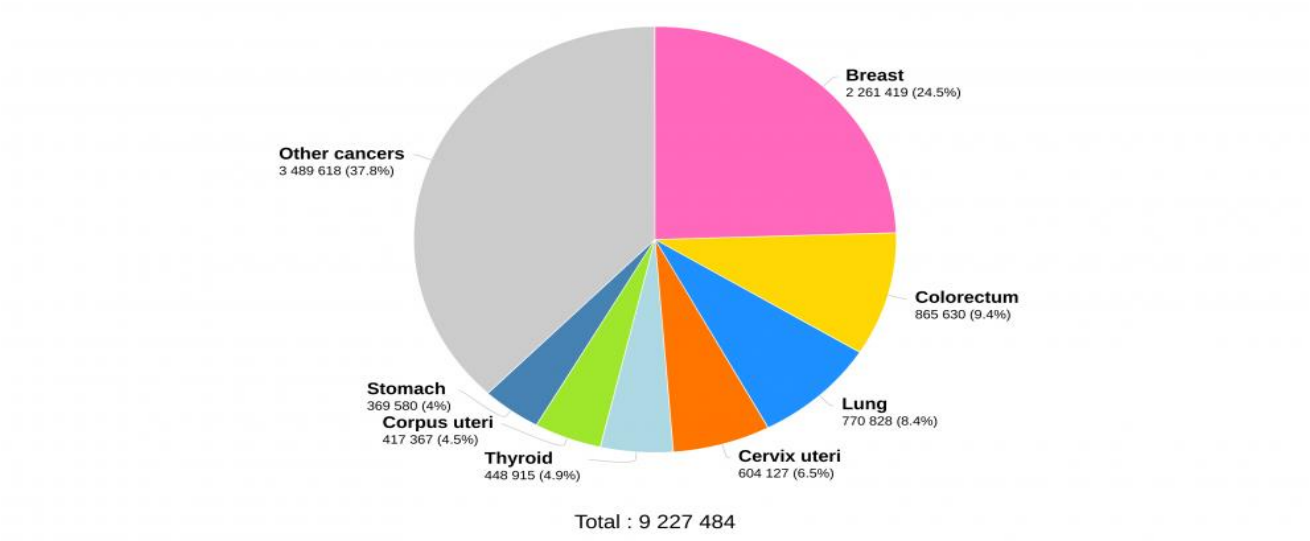


Figure 1.1: Estimated number of new cases in 2020 worldwide for females of all ages [2].

Early screening is essential for women who are more exposed to risk factors such as Family History of Breast Cancer, Existence of Benign Breast disease, Density, Age at First Birth, Age at Menopause, weight, Hight, waist circumference, Body Mass Index (BMI), Age at Menarche, Endogenous Estrogen Levels, Waist-Hip Ratio, Age, Smoking Frequency, Alcohol consumption, Occupational Hazard, Breast Feeding, hormone replacement therapy, number of deliveries, number of children breastfed, undergone IVM, Average duration of breastfeeding for children, previously used any family planning methods, Type of family planning methods, intrauterine device (IUD)-hormonal, duration of an intrauterine device (IUD) hormonal, physical activity and complaint of any chronic disease [5] [6].

The most common techniques used to detect breast cancer are mammography and magnetic resonance imaging (MRI). However, these techniques are very expensive, large, only available in several and large hospital facilities. As a result, finding a breast screening facility can be challenging, and there might be long wait times.

In addition, many of the barriers to screening techniques according to a study conducted in Saudi Arabia are the fear of detecting BC, fear of radiation exposure fear of BC detection, being busy all the time, and taking a long time to take a medical appointment a high cost of mammography [7].

There is currently an overwhelming amount of data. On a personal level, every choice we make and decision we make is recorded and stored. The important thing is to be able to find patterns in this chaotic world of information in data mining, amount data sets are utilized to find patterns and relationships that can be used to analyze data to solve discover the problems and challenges. It is important to do a thorough investigation to clarify the factors and facilitate the dissemination and awareness of this disease. To do this, data mining was used as a technique. Unlike the time-consuming and tedious, perhaps incomplete, manual data mining investigation allow for quick and direct analysis

after gathering and combining different sources of information. Data Mining (DM) is defined as the process of automatically discovering patterns in data, and trying to solve problems by analyzing existing databases. These discovered patterns must be significant and help explain something about the analyzed data. It can also be used as a predictive tool, to predict future or missing values of another variable [8].

In this thesis, a descriptive analysis of the breast cancer risk factors was performed, and different data mining techniques were used for the prediction of breast cancer and their performance compared to evaluate a more efficient and effective DM technique. The research used a dataset that contains independent variables to predict breast cancer by data mining tools.

1.2 Study Justification

Over the past few years, the incidence of breast cancer has increased in Palestine. It is the most common malignant tumor among Palestinian women. The prevalence of this disease reaches 29% of diagnoses in the West Bank and 20% in Gaza. In addition, breast cancer constitutes 32% of all cancers diagnosed in women last year and 16.5% of all cancers registered among Palestinians in 2020. Although breast cancer has one of the greatest survival rates of any malignancy, success is dependent on early detection. Unfortunately, more than 64.7% of breast cancer cases in Palestine are detected in late-stage in 2020, when cancer has metastasis. This means limiting treatment options and decreasing survival chances[3][9].

As a result of the delay in the early examination of women and the discovery of the disease, cancer deaths amounted to 12.9% of the deaths in Palestine in general, and the rate of cancer in the community reached an incidence rate of 19.1 per 100,000 population. Studies have shown that most

of the world's dead were in the late stages of infection. This is due to the spread of the disease completely [6].

In addition, studies have confirmed that early diagnosis of cancer saves lives, and more effective and low-cost treatments[10]. Figure 1.2 shows the importance of early diagnosis before the cells transform from healthy cells to cancerous cells[11]. So, the most important consideration is how to extract features from all risk factor data. Because there are many features and the data are huge, various machine learning techniques can be employed by pre-processing the data: data cleaning, clustering, filtering, and normalizing. Because risk factors play an important role in the diagnosis process, this leads to breast cancer prediction.

In medical applications, where the importance of the temporal component cannot be ignored, data mining techniques and machine learning have acquired great relevance. The applications of these techniques range from the intelligent visualization of large amounts of medical data to quality control in hospitals [5]. In detecting diseases such as breast cancer, the set of data mining techniques and methods turn out to be a very useful instrument in certain parts of the cancer detection and diagnosis process where it is necessary to classify information and find knowledge in large volumes of data.

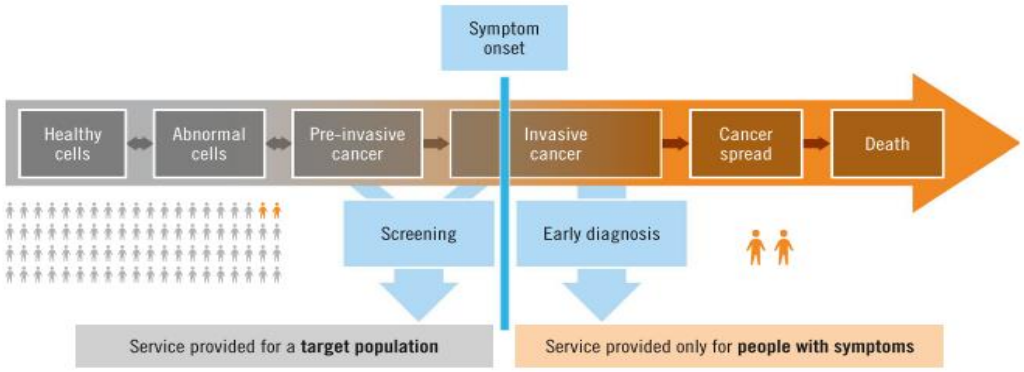


Figure 1.2: WHO Guide to Early Diagnosis shows how to screen from early diagnosis based on the onset of symptoms[11].

In Palestine, there is no breast cancer prediction framework. An efficient breast cancer prediction model based on data mining were proposed in this study. There are also various varieties of breast cancer, each with its stage of aggressiveness, progression, and genetic makeup [7]. Therefore, having a system that allows for early identification and prevention would be extremely beneficial in terms of improving breast cancer survival rates [12]. This decides to be taken as soon as possible for early detection and to reduce the mortality rate [13].

1.3 Problem Statement

Breast cancer is a serious disease that affects a large number of females all over the world. When a tumor is small, there are usually no symptoms, [14] which is why early detection in its early stages helps save women's lives [3]. Traditionally, statistical analysis has been a tool that, through the confirmation of hypotheses, has allowed the health field to generate knowledge and understand certain phenomena for the benefit of the development of technology and protocols for the care of diseases and recovery of health. Through the statistical analysis, primary variables and relationships are fundamentally considered, without considering the true interrelationships, which are usually hidden and can only be discovered using a more complex data treatment, which is only possible with data mining techniques [4]. Data mining techniques that include the use of supervised machine learning using risk factors and a patient's family history help in disease prediction and rapid decision-making for necessary examinations, thus helping to increase the early detection of disease. However, a machine learning study in breast cancer data analysis and prediction in the Ministry of Health is not available. Hence, there is an urgent need of developing a BC prediction model specifically for Palestinian women.

With this study, applying machine learning algorithms, we intend to identify and classify the greatest number of risk factors to avoid them or minimize their impact and to predict disease to increase early detection of disease.

1.4 Objective of the Study and Research Question

The general objective of this study is: Determine to what extent the application of machine learning techniques will improve the prediction process of the risk of breast cancer, and which of these risk factors are maybe the cause of breast cancer in Palestine.

1.4.1 The Objectives of This Work Can Be Summarized as Follows:

- Utilizing machine learning algorithms to predict breast cancer depending on the risk factors.
- Increasing the accuracy of breast cancer prediction based on a Palestinian risk factor dataset
- Design a framework for breast cancer diagnosis from a risk factor dataset.
- The breast cancer prediction model aids in the diagnosis of breast cancer in women who live in areas where mammography device is unavailable or far away.

1.4.2 The Main Questions of This Study are:

What is the accuracy of data mining algorithms used currently in the breast cancer prediction diagnosis?

How might data mining techniques be used to identify the risk factors that best characterize data for the goal of distinguishing between breast cancer malignant, benign and healthy?

how can missing feature values of breast cancer risk factors enhance the performance of breast cancer prediction?

How can the developed breast cancer data mining model be tested for classification accuracy?

1.5 Research Expected Outcomes

The expected outcome is the model with the highest possible performance for the prediction of the diagnosis of breast cancer and classification of Palestinian women into malignant, benign or healthy cases. It allows for predicting advance, in the specific case of breast cancer, and the factors that affect the women who suffer from this disease. The objective of the work is to classify the population of cancer patients using data mining techniques, hoping to find underlying relationships that cannot be identified by a classical statistical treatment. These outcomes will be aided in improved clinical decision-making in health care organizations when there is a high inter-individual variation in diagnosis. Also, it can help physicians find and validate the risk factors in Palestine, as well as enhance clinical decision-making in cases where the prognosis is heterogeneous.

1.6 Thesis Chapters Descriptions

This thesis contains six chapters as the follow

Chapter One: General Introduction: This chapter includes multiple sections such as study background, study justification, problem statement, study aim and objectives, ls and expected outcomes.

Chapter Two: **Literature Review:** This chapter includes a literature review of previous papers about the topic work or related to it.

Chapter Three: **Conceptual Framework:** contains a diagram for the relevant variables for the research and maps out how they might relate to each other.

Chapter Four: Methodology: This chapter provides a detailed account of our experiment and includes the study area and setting, study subjects, study design overview, sample size, data collection methods and measurements, data management, analysis plan, study limitations, and ethical consideration.

Chapter Five: Result: The results and discussion are presented. This chapter includes descriptive and visualization of the risk factors reported in datasets. this chapter is considered presenting and discusses the results of applying artificial intelligence (AI) tools to datasets.

Chapter Six: Conclusion and Discussion: The conclusion, recommendations, further work, and study limitations are presented.

Chapter Two: Literature Review

2.1 Introduction

This chapter reviews the literature related to the prediction and classification of breast cancer depending on risk factors using data mining predictions from previous research. The wide application of data mining in the cancer field helps to improve the accuracy of diagnosis, save the life of patients, and reduces treatment costs. In addition, this chapter focuses on the knowledge gap in the previous literature, and the conceptual framework of the study.

2.2 Female Breast Anatomy

The female breast has an extremely complex anatomy. A woman's breasts are mostly made up of fat and connective tissue. Lymph nodes, milk ducts, lobules, lobes, and arteries are some of the less visible elements of the female breast [15]. A network of tiny tubes connects the lobes, lobules, and bulbs (ducts). Milk is transported from the bulbs where it is generated to the dark region of skin in the center of the breast by ducts (areola). They unite to form larger ducts that end at the nipple, and milk is made available to the baby [16].

The blood in arteries and capillaries delivers oxygen, nutrients, and other life-sustaining nutrition to breast tissues. The lymphatic system, which consists of a network of veins, lymph ducts, and lymph nodes, aids in the battle against infection. The fluid that escapes from the capillaries into the lymph nodes under the armpit and behind the breastbone is drained through these veins [16][15]. Show in figure 2.1 the anatomy of the female breast.

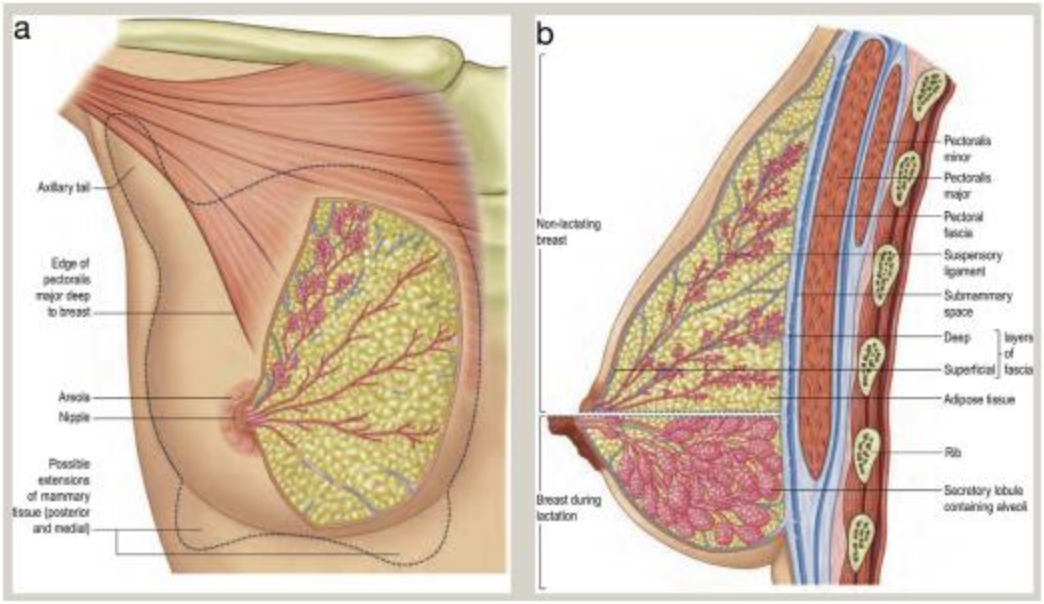


Figure 2.1: Anatomy of the female breast[15].

2.3 Breast Cancer in the World

Breast cancer grows when the breast tissue extends abnormally, causing the cells to grow more quickly than they should. In breast cancer, two different types of cancer develop: benign and malignant. Noncancerous tumors are benign, meaning they do not spread to other regions of the body, whereas cancerous tumors are malignant[17]. Benign tumors can be prevented from progressing to cancerous tumors with the right treatment, whereas malignant tumors cannot [18].

Tumors just proliferate and spread throughout the body. As a result, it's critical to spot the signs. Type of tumor at an early stage and use precision treatment to prevent it from progressing to cancerous tumors [4]. According to the World Health Organization (2020), cancer has been responsible for the deaths of millions of people worldwide. Figure 2.2

shows the number of female deaths for the year 2020 by type of cancer. According to the graph, breast cancer has recorded the highest number of female deaths (WHO).

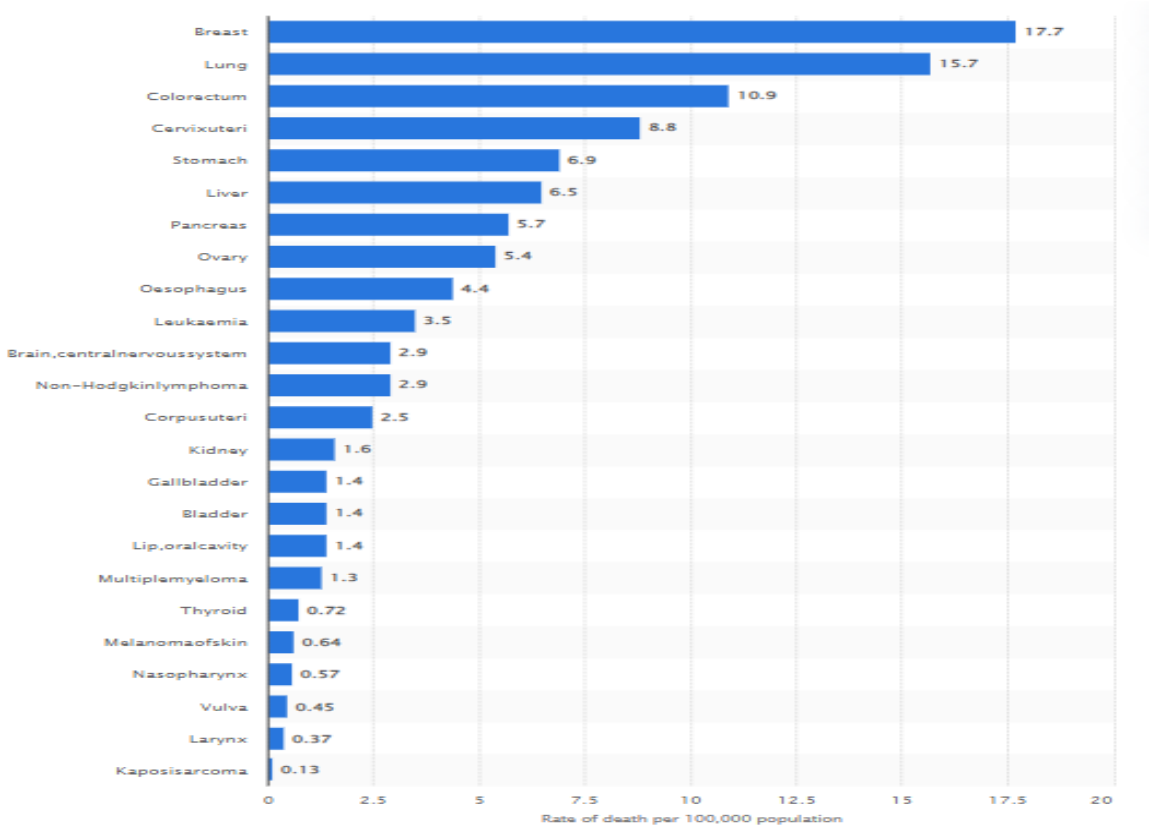


Figure 2.2: Rate of cancer deaths among females worldwide in 2020 by type of cancer per 100000 population sources: statista.com

Also, according to the American Cancer Society[19], roughly one in every 8 women may acquire breast cancer over their lifetime. If the disease is detected sooner and its severity is determined, the number of deaths may be decreased. Patients are treated and medicated in this way. As a result, it's critical to Identify a model that accurately forecasts the condition so that the risk of mortality in women can be lowered with adequate treatment.

2.4 Breast cancer in Palestine

2.4.1 Breast cancer incidents in Palestine

Breast cancer is the most common cancer with 526 cases reported in 2020; 16.5% of all cancer cases show Figure 2.3, with an incidence rate of 19.1 per 100,000 population shown in Figure 2.4[3].

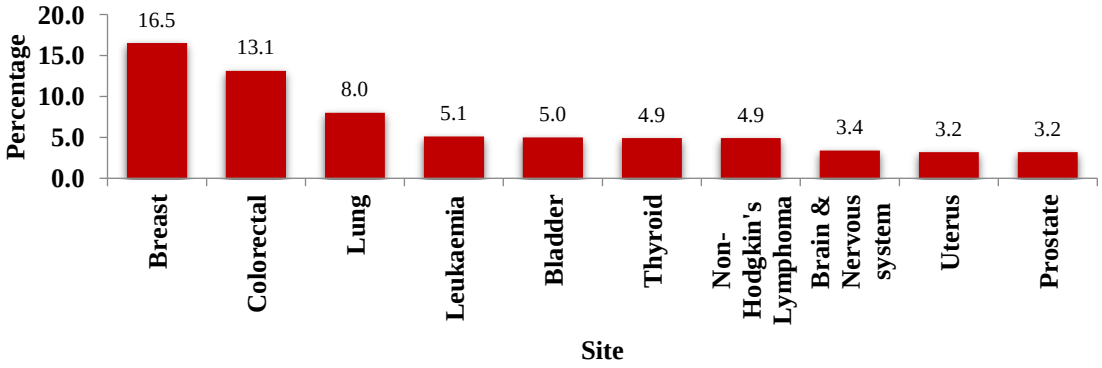


Figure 2.3: Distribution of Percentage of Top Ten Reported Cancers in all population, West Bank,Palestine[3]

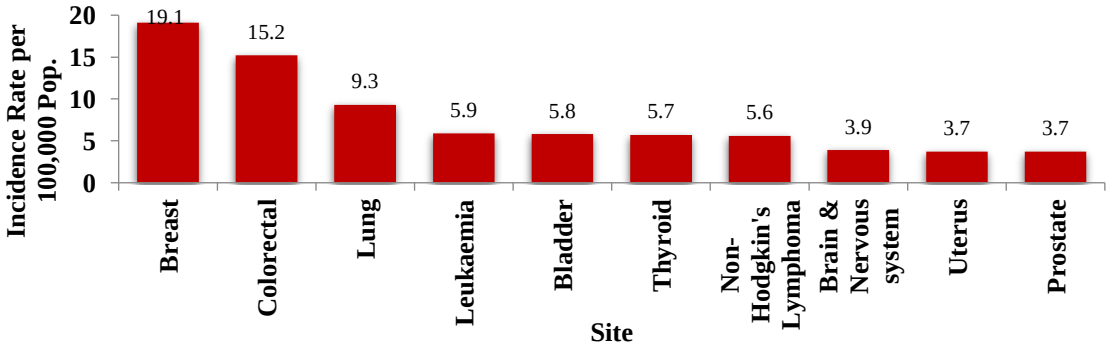


Figure 2.4: Incidence Rate of Top Ten Reported Cancers among Females, West Bank, Palestine 2020 [3]

Among females; breast cancer is the most common as 518 cases were reported in 2020, representing 32.0% of all cancer cases among females, with an incidence rate of 38.4 per 100,000 female population shown in figure 2.5 [3].

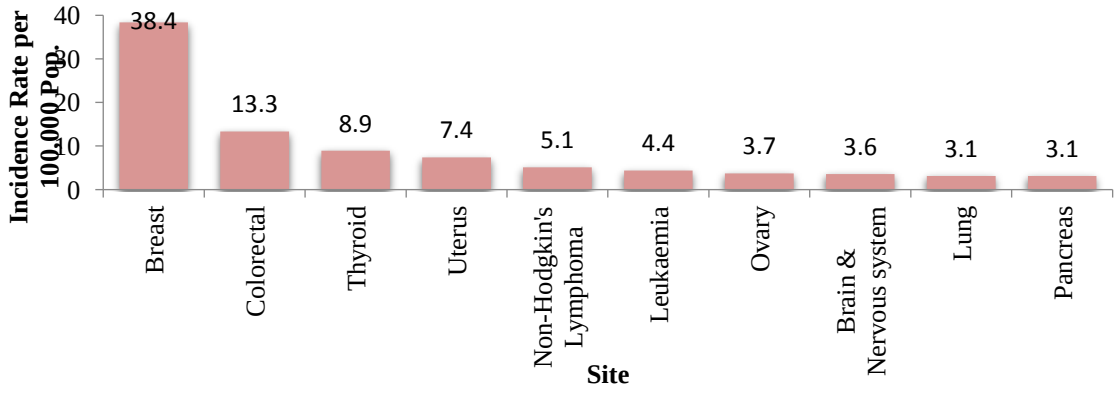


Figure 2.5: Incidence Rate of Top Ten Reported Cancers among Females per 100,000 Female populations, West Bank, Palestine 2020

2.4.2 Cancer Mortality in Palestine 2020

In 2020, the cancer deaths were 14.1% of the total reported deaths with 56.3% of cancer deaths among males, and 43.6% among females. 12.9% of all cancer deaths are caused by breast cancer shown figure 2.6 [3].

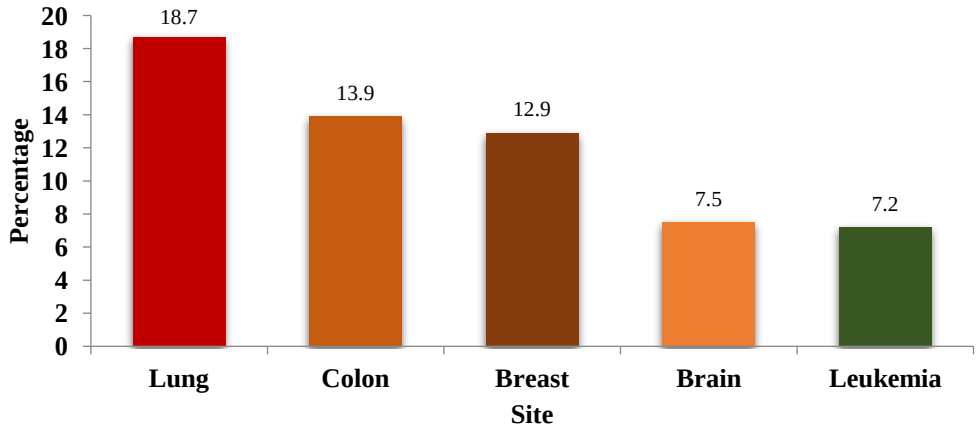


Figure 2.6: Cancer Mortality in Palestine 2020[3]

2.4.3 Breast Cancer Stages in Palestine

In 2020, the distance rate of breast cancer was 17.4% (Stage III and Stage IV) from total breast cancer this means metastatic breast cancer has occurred. While the region of breast

cancer was 47.3%, shown in Figure 2.7 increases in the distance and regional rate of breast cancer cases by type of stages from 2017-2020 [9].

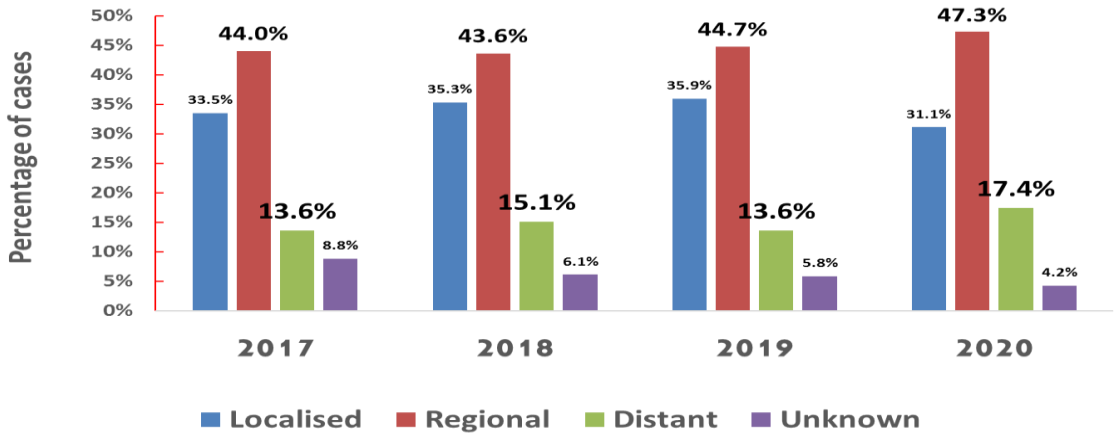


Figure 2.7: Percentage of breast cancer cases by stages from 2017-2020

2.4 Stages of Breast Cancer

There is a three-factor system for categorizing cancer into stages. Tumor size, lymph node involvement, and metastasis are all issues to consider.

The Stages May be Represented as Follows:

- **Early Breast Cancer:** Stage 0: Carcinoma in situ, or illness that has not infiltrated the basement membrane, Stage I: A narrow size distribution tumor with no invasion of lymph nodes, and Stage II: Regional lymph node involvement.
- **Locally Advanced Breast Cancer:** Stage III: Usually a large tumor with considerable nodal involvement in which the node or tumor is anchored to the chest wall; also includes inflammatory breast cancer, which progresses quickly.
- **Advanced or Metastatic:** Stage IV breast cancer Metastases in organs other than the main tumor [20].

2.5 Breast Cancer Risk Factors

Various factors can lead to developing breast cancer, in this section we reviewed studies related to BC risk factors, they are classified as either modifiable (those that can be controlled like habits, environmental hazards, etc.) or non-modifiable factors (those that cannot be controlled like, gender, family history etc.)

2.6 Breast Imaging Reporting and Data System (BI-RADS)

Various classifications are used to categorize breast lesions, and while they are all similar, the categorization proposed by the American College of Radiology (BI-RADS) is the most widely recognized [21]. Table 1 shows BI-RADS final assessment categories.

Table 1: BI-RADS final assessment categories

Category	Category1	Management	Likelihood of Cancer
0- Assessment	Incomplete Assessment	Additional imaging required	Not applicable yet
1	Negative	Routine annual screening	No cancer detected
2	Benign	Routine annual screening	0%
3	Probably Benign	Follow-up scan after 6 months or earlier, as advised by your doctor	0% to 2%
4	Probably Malign	Breast tissue biopsy recommended by the doctor	4A - 2% to 10% 4B - 10% to 50% 4C - 50% to 95%
5	Malignant	Biopsy to be done essentially	>95%
6	Biopsy- Proven Malignancy	Further treatment evaluation is done by the oncologist	Cancer already present

2.7 Data Mining and Machine Learning

Data mining (DM) is a technique that is used to find useful, hidden, and new patterns of knowledge from massive or large databases by a variety of techniques. There are various

data mining functions such as prediction, classification, clustering, concept descriptions, association rules and, sequence discovery to find useful patterns as shown in Figure 2.8.

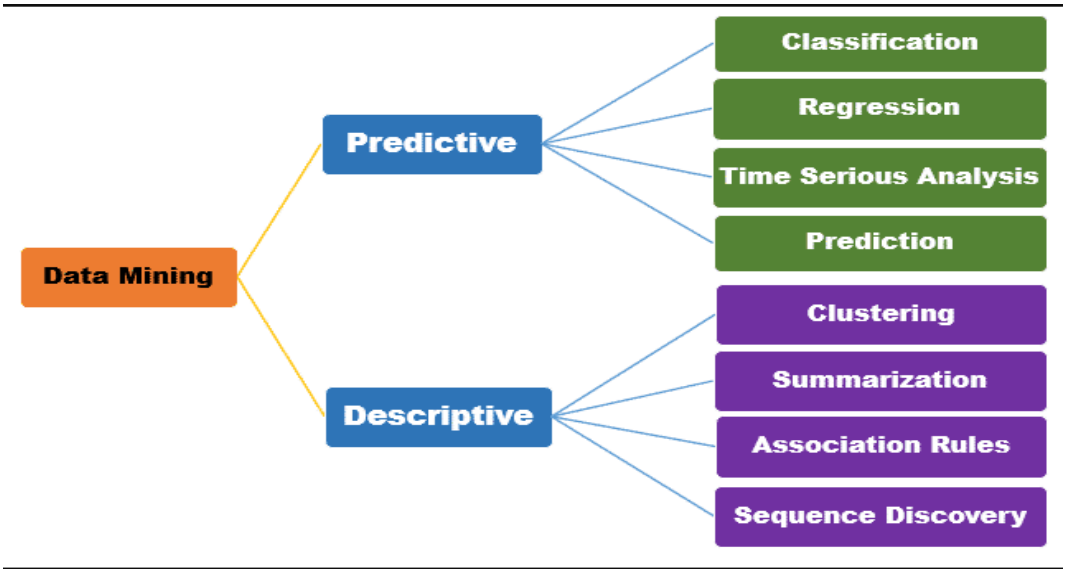


Figure 2.8: Types of Data Mining

Data Mining (DM) is an interdisciplinary field that includes database systems, machine learning, statistics, visualization, data science, and other disciplines shown in figure 2.9.

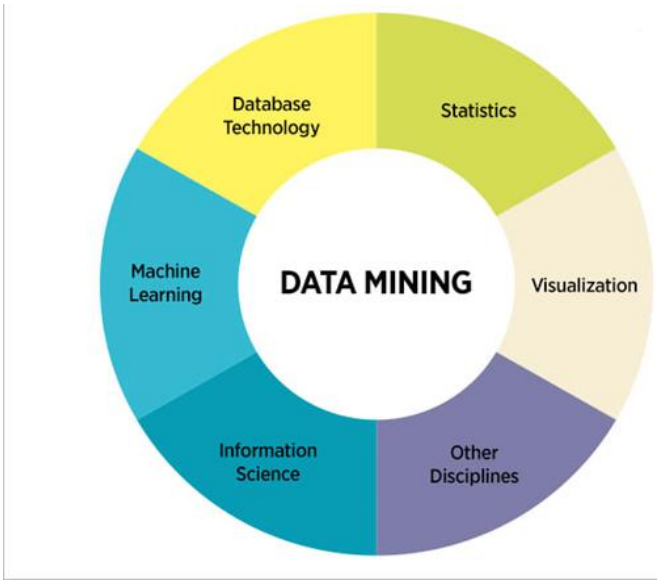


Figure 2.9: DM component.

In addition, data mining methods involve the use of complicated and sophisticated data analysis tools in the early detection of cancer. These tools include algorithms ,machine learning methods and statistical models [22].

2.7.1 Supervised Machine Learning and Classification

Classification is a supervised Machine Learning technique in which various objects or groups are assigned classes. There are two phases to classification: The first phase is model construction, which is described as an examination of a database's training records. The created model is then used for classification in the second stage. The percentage of correctly categorized test samples or records is used to calculate classification accuracy [23].

2.8 Related Works

Previous studies on BC prediction are covered in this chapter, as well as the DM and ML approaches utilized and which method produces the best results, dataset sources, number of features, and feature extraction and selection. In previous publications, researchers have used various data mining algorithms to predict breast cancer depending on risk factors.

Alfayez and colleagues (2020) investigated the machine learning approaches to construct predictive cancer risk models using scoping review approach. Their research compares the performance of several machine learning models. PubMed search was used with the following inclusion criteria: (1) asymptomatic adult population of both sexes (2) Any study that uses machine learning approaches to develop prediction models for future cancer risk in any context and any area of the world. As a result of their research, ten studies with significantly different results employed machine learning to predict future cancer risk in

asymptomatic people. However, all of these ten studies used the area under the curve (AUC) as a measure of model performance, but none of them used model calibration measures. In terms of performance, six studies achieved a 'poor' valuation, while four studies achieved a 'good' evaluation[24].

Can Hou et al. (2020) aimed to assess and compare the effectiveness of 4 machine learning algorithms in predicting breast cancer in Chinese women utilizing ten risk features for breast cancer. For model training and testing, they used a dataset of 7127 records from breast cancer patients and 7127 records from healthy controls. They calculated specificity, AUC, sensitivity, and accuracy as the evaluation of the model performance. The result showed the XGBoost (AUC 0.742) outperformed random forest (AUC 0.728) and deep neural network (AUC 0.728). The top-ranked variables in the three unique machine learning algorithms were the number of live births, main residence, menopause status, age, number of live births and age at first birth [25].

Kehinde Williams et al (2015) aimed to the prediction of BC risks among Nigerian women. They used a dataset collected from the Cancer-Registry of LASUTH. The dataset contains 69 instances of women and 17 attributes. The probability distribution is divided into three categories: improbable, likely, and benign. The dataset contains 1 dependent variable and 16 independent variables. The researcher two distinct classifications for data mining naive Bayes and decision trees methods. The performance of each technique was compared to choose the best classifier. 48 decision trees were chosen and showed a higher accuracy with lower error rates. The results of this study showed that 48 decision trees had higher accuracy of 94.2 percent and had accurate and wrong classifications of 65 and 4 respectively[26].

P. Ramachandra and his colleagues (2014) used the Support-Vector machine and K- means clustering algorithm to predict cancer patients and healthy. They used data from the Cancer registry Chennai. This data consists of 30 attributes and 746 records that were used to train the model and a portion that was utilized to test the system's relevance. The results showed the accuracy of the model was 99.8% with a mean absolute error of 0.0097 [27].

Shajahan et al. (2013) used decision trees to predict BC and predict the existence of cancer using data mining approaches. The data obtained included 699 patient records with ten variables and a malignant or benign output class. utilized input cell size, clump thickness, shape homogeneity, cell growth, and sample code number additional physical examination findings on the use of a supervised learning algorithm yielded the following results: showing the random tree method had the highest accuracy (100%) and error rate (0), whereas CART had the lowest accuracy of 92.99 percent, while nave bayes' had the highest accuracy of 97.42 percent [28].

Chang Ming and colleagues used eight generated datasets and two retrospective samples to compare the performance of eight different ML techniques. N = 1143 BC patients and their healthy female relatives, and N = 2481 Swiss breast cancer patients and cancer-free women requesting genetic examination and/or testing. The purpose of their study was to compare the accuracy of ML-based estimates against a pair of established methods “the Breast Cancer Risk Assessment Tool “(BCRAT) and “Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm “(BOADICEA) models shown in figure 5 the input of BCRAT and BOADICEA.

for the U.S. population-based sample, predictive accuracy (AU-ROC curve) for ML-Adaptive Boosting and ML-random forest was 88.28 percent and 88.89 percent,

respectively, vs 62.40 percent for BCRAT. For the Swiss clinic-based sample, predictive accuracy with ML-adaptive boosting was 90.17 percent and 89.32 percent with ML-Markov chain Monte Carlo generalized linear mixed model was 89.32 percent versus 59.31 percent with BOADICEA [29].

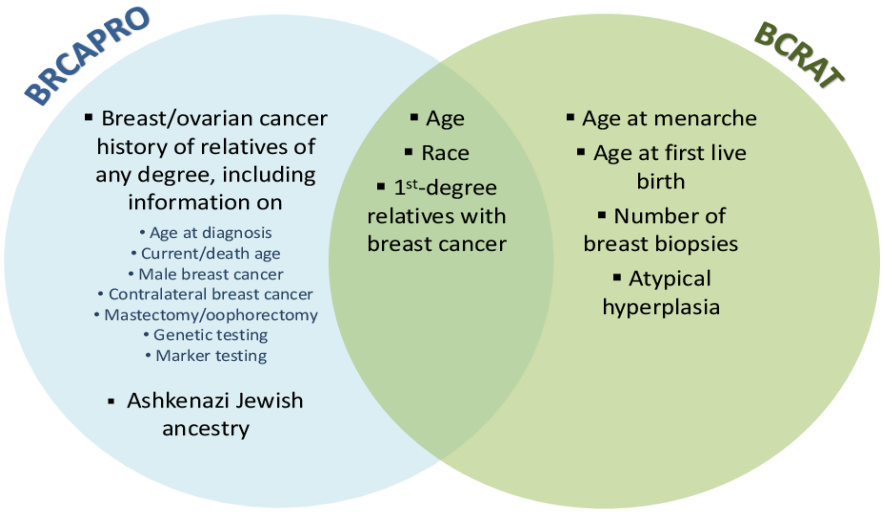


Figure 2.10 : Attributes to BRCAPRO and BCRA [30].

For the effective and accurate classification of benign and malignant mammography images, HuaLi and colleagues propose an improved neural network model. Shanxi Medical University's mammography images were used to create this dataset. The experimental results are analyzed and compared using pre-processed mammogram datasets. The results show that the neural network model has better classification performance than other network models when using the 10-fold cross validation method. The model's average accuracy is 94.55 percent, which improves the accuracy of mammogram image benign and malignant classification. At the same time, it demonstrates that the model is generalizable [31].

Anwar Alhazmi and his colleagues (2021) aimed to forecast a person's chance of acquiring oral cancer based on risk factors, clinic-pathological parameters, and medical history. They

developed an artificial inelegance of the neural network model. The researchers used a dataset collected from Prince Muhammad bin Nasser Hospital in Saudi Arabia between 2017 and 2020. The model used a dataset including 29 variables associated with patients. The dataset was divided into two parts: a training dataset with 54 cases (75 percent) and a testing dataset with 19 cases (25 percent). The results showed that the performance of the models was achieved using the ANN algorithm for cancer prediction. The ANN accuracy rate for oral cancer was (78.95%), while the sensitivity was 85.71% and specificity (60.00%) [32].

2.9 The Knowledge Gap in the Previous Literature

The Knowledge gap in the studies is that the BC dataset contains a limited number of patient records datasets to get results closer to accuracy and can be generalized, no dataset from cases Palestinian published and no previous studies about the prediction of BC in Palestine. Also, In the previous studies feature choices were made on a limited number of variables as BC risk factors, resulting in a decline in the number of unknown potential variables. Our study used 30 risk factors that used as input in prediction models, which will be updated national data set from Palestinian MOH with more new variables.

The results obtained when carrying out the analyzes, can be summarized in a ranking of environmental risk factors, as well as the relationship between them. There were no studies that used three classes for cancer prediction; instead, all studies classified cases as healthy or malignant, in this study the diagnosis was classified as benign, malignant, or healthy.

2.10 A Conceptual Framework

This chapter presents the factors that are associated with the conceptual framework components for prediction breast cancer. These factors can be placed into a number of categories. The first is patient Demographic characteristic which identifies woman Age, Family history, gender, locality type, BMI and Martial status. Breast characteristics describe density of breast and number of mass in breast. Life style and radiation such as alcohol consumption, occupation Career, smoking status, previous chemotherapy, previous Chest CT-scan, previous chest x-ray, previous Mammogram. Medical history such as complain of any chronic diseases, previous Cyst aspiration, previous Heart, previous Hypertension, previous Lumpectomy ... etc. Finally, Reproductive and hormonal factors. The conceptual framework for this study can be followed in Figure 2.11.

The risk factors can categorize to two types depending on whether controlled or not:

- BC Uncontrolled risk factors:
  - ✓ Demographic factors.
  - ✓ Medical History.
  - ✓ Breast characteristic.
- BC Controlled risk factors
  - ✓ Lifestyle and radiation exposure
  - ✓ Reproductive and hormonal factors

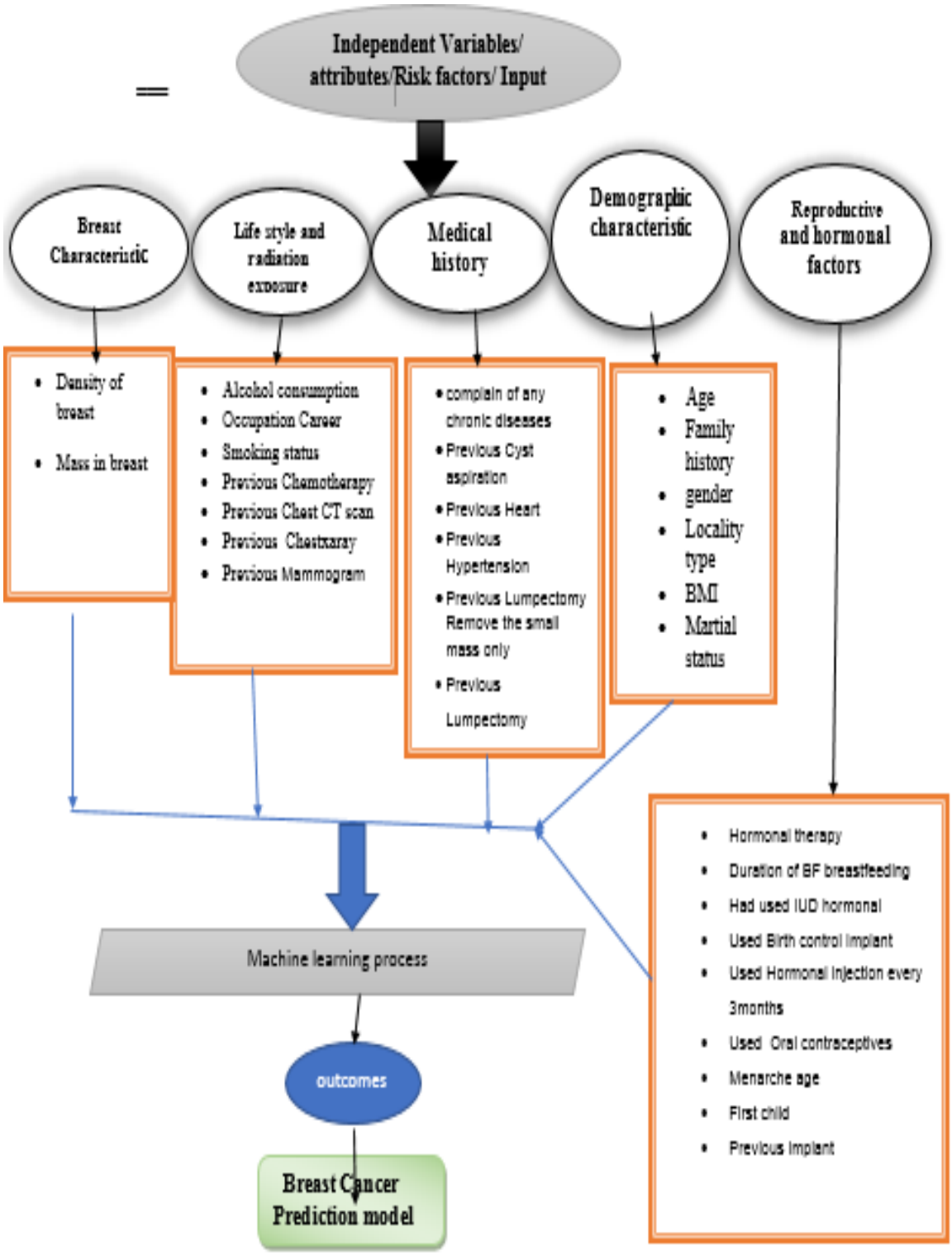


Figure 2.11: A conceptual framework of breast cancer

2.11 Summary

This chapter covered a literature review on BC prediction, how data mining is used widely in the medical profession to increase the quality of medical diagnosis while lowering costs, the knowledge gap in the prior research, and the study's conceptual framework.

Chapter Three: Methodology

3.1 Introduction

This chapter focuses on the study methodology, that was used to conduct this study. It includes research design, study settings, sample size, study population, study tools, datasets, data collection methods, data management, data analysis, Artificial intelligence framework, ethical considerations, and limitations of the study.

In this study, the study used the Quantitative methodology. The process of prediction and classification includes of extracting patterns in the form of rules or functions, from the data so that the user can analyze them. This process generally involves data collection, data preprocessing, data mining, evaluating and presenting results.

3.2 Study Design

A Cross-Sectional retrospective study will be used in this study, which used secondary data from the mammography e-Registry of PMOH. The date of incident was obtained from 2020 and 2021. This work employed experimental design, which is a common methodology used to predict breast cancer. Various data mining procedures will be engaged in this study, starting with proper data collecting from the Palestinian MOH database, as individual approaches are used in research of this magnitude to study the efficacy of the proposed solutions, evaluating experimental and developing models effectively and with high performance. The measures of the model performance with calculated, sensitivity, specificity, AUC, and accuracy. The methodology of this study was represented in Figure 3.1.

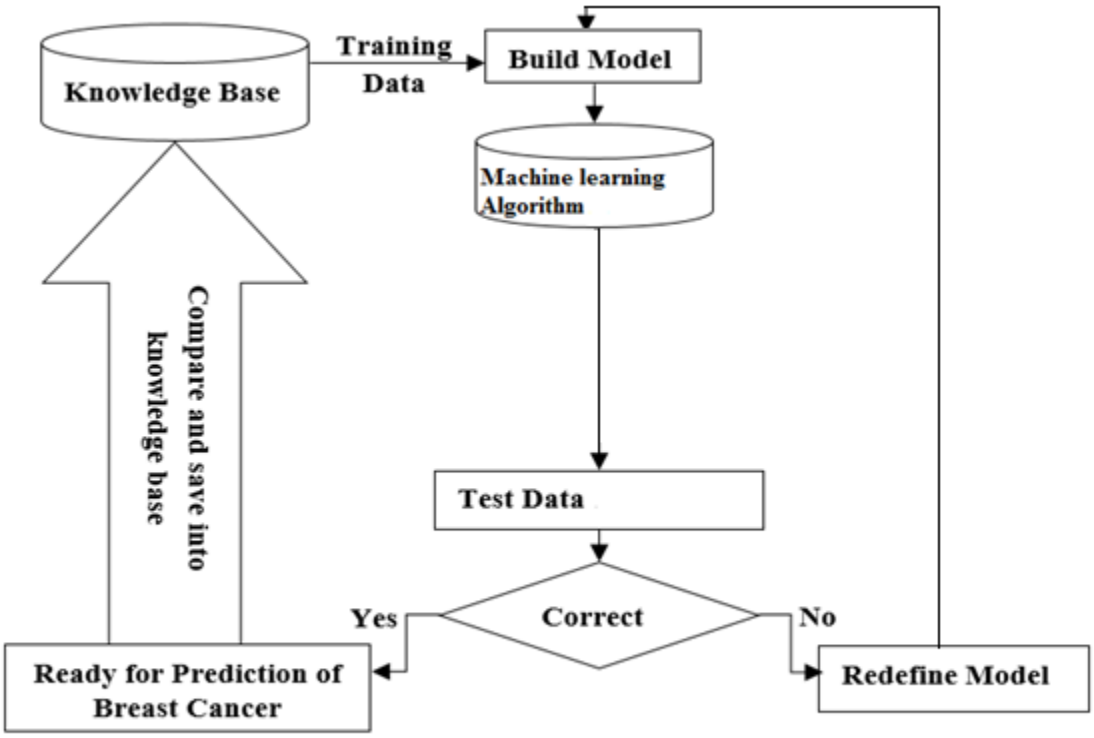


Figure 3.1: Study Methodology

3.3 Study Area/Setting

In 2013, PNIPH undertook research to assess the effectiveness of the screening program, and in collaboration with the Ministry of Health, produced the Mammogram e-Registry. The mammogram e-Registry has been designed to keep track of all women who get mammography screenings. After that, the patient's information is saved in a centralized electronic system.

The study collected secondary data from the electronic register of mammograms in several directorates of the Ministry of Health in the West Bank table 2. The study focused on cases registered for the incident date between 2020 and 2021.

Table 2: Distribution of hospitals in the areas and hospital types

NO	Moderate Name
1	Albireh
2	Al-Ram
3	Bethlehem Health Directorate
4	enter of Mammography \Tulkarm
5	Central Heath Directorate Salfit
6	General Health Directorate of Jericho
7	Halhul
8	Hebron Health Directorate
9	Jenin Main Center
10	Qalqilya Health Directorate
11	shohada Yetta Health Center
12	South Hebron Health Directorate
13	Tubas Central

3.4 Study Population and Sample Size

Study Population consisted of women who presented to the breast screening centers in the directorates of the Ministry of Health during the years 2020 and 2021. The study population amounted to 4199 women who presented to Screening Breast cancer clinics.

After ignoring incomplete records and incomplete assessments which were not diagnosed. The researcher obtained random sample contains 1794 records; 635 malignant breast cancer cases, 570 healthy controls and 589 benign cases. Breast cancer cases were collected from mammography electronic Registry system in MOH Directorate.

Each record has 30 features (Variables) plus the class attributes. The class is formed as healthy, malignant, or benign cases.

3.5 Study Tools and Data Collection

In 2013, PNIPH undertook research to assess the effectiveness of the screening program, and in collaboration with the Ministry of Health, produced the Mammogram e-Registry. The mammography e-Registry has been designed to keep track of all women who get

mammography screenings. After that, the patient's information is saved in a centralized electronic system. In the study, the dataset obtained from PMOH included all women who visited a breast cancer screening clinic in the Palestinian MOH directorates between 2020 and 2021.The objective of this stage is to obtain the internal data sources that serve as the basis for the data mining process. The dataset contains 1794 records with 29 risk factors of breast.

The class distribution is framed as benign, malignant and healthy. As a whole, there are 28 independent variables and one dependent variable, as shown in Table 3. While the non-modifiable factors are represented column 4 in table 3 below. The variables will be Family History of Breast Cancer, Existence of Benign Breast disease, Density, Age at First Birth, Age at Menopause, Body Mass Index (BMI), Age at Menace, Endogenous Estrogen Levels, Age, Smoking Frequency, Alcohol consumption, Occupational Hazard, Breast Feeding, hormone replacement therapy, undergone IVM, Average duration of fasting for children, previously used any family planning methods, Type of family planning methods, intrauterine device (IUD)-hormonal, duration of an intrauterine device(IUD) hormonal, physical activity, complain of any chronic disease.

Table 3: Variables/ attributes from dataset description

S/N	RISK FACTORS	Type variable	/ modifiable non-modifiable	Values
1	Density of breast	Qualitative nominal	non-modifiable	• almost entire fatty tissue • Scattered fibro glandular Density • Heterogeneous Dense • Extremely Dense
2	Mass in breast	Quantitative ratio	non-modifiable	0-10
3	Alcohol consumption	Qualitative nominal	modifiable	• No • yes

4	Occupation Hazard	Qualitative nominal	modifiable	• No • yes
5	Smoking status	Qualitative nominal	modifiable	• No • yes
6	Previous Chemotherapy	Qualitative nominal	Non -modifiable	• No • yes
7	Previous Chest CT scan	Qualitative nominal	Non -modifiable	• No • yes
8	Previous Chest x-ray	Qualitative nominal	Non -modifiable	• No • yes
9	Previous Mammogram	Qualitative nominal	modifiable	• No • yes
10	complain of any chronic diseases	Qualitative nominal	Non -modifiable	• No • yes
11	Previous Cyst aspiration	Qualitative nominal	Non -modifiable	• No • yes
12	Previous Heart	Qualitative nominal	Non -modifiable	• No • yes
13	Previous Hypertension	Qualitative nominal	Non -modifiable	• No • yes
14	Previous Lumpectomy Remove the small mass only	Qualitative nominal	Non -modifiable	• No • yes
15	Previous Lumpectomy	Qualitative nominal	Non -modifiable	• No • yes
16	Age group	Qualitative ordinal	Non - modifiable	• 1 = "24-34" • 2 = "34-44" • 3 = "44-54" • 4 = "54-55" • 5 = "64-64" • 6 = "74-84" • 7 = "84-94" • 8 = "more than 94"
17	Family history	Qualitative nominal	Non -modifiable	• No • yes
18	Locality type	Qualitative nominal	modifiable	• City • Village • camp • Other
19	Body Mass Index (BMI)	Qualitative nominal	modifiable	• 0 = "<=25" • 1 = ">25"
20	Marital status	Qualitative nominal	Non -modifiable	• single • married • Widowed, Divorced, Separated
21	Hormonal therapy	Qualitative nominal	modifiable	• NO • yes

22	Duration of BF breastfeeding	Qualitative nominal	modifiable	• Less or 6 months • More than 6 months
23	Had used IUD hormonal	Qualitative nominal	modifiable	• No • yes
24	Used Birth control Implant	Qualitative nominal	modifiable	• NO • yes
25	Used Hormonal injection every 3 months	Qualitative nominal	modifiable	• Yes • NO
26	Used Oral contraceptives	Qualitative nominal	modifiable	• Yes • NO
27	Menarche age	Qualitative nominal	Non - modifiable	• ≤13 years • >13 years
28	Age at first Children	Qualitative nominal	Non - modifiable	• <30 years • >=30 years
29	Previous Implants	Qualitative nominal	modifiable	• Yes • No
30	BI-RADS	Qualitative nominal		• BI-RADS 0 • BI-RADS 1 • BI-RADS 2 • BI-RADS 3 • BI-RADS 4 • BI-RADS 5 • BI-RADS 6
31	Breast Cancer diagnosis	Qualitative nominal	Class	• Healthy • Benign • malignant

3.6 Data Management and Analysis Framework: Data Mining Framework

After collecting the dataset from PMOH and extracting data from it, various data mining procedures were engaged in the ML algorithms described in this study, starting with proper data collecting from the Palestinian MOH database, as individual approaches are used in research of this magnitude to study the efficacy of the proposed solutions, evaluating experimental and developing models effectively and with high performance. The measures of the model performance were calculated, sensitivity, specificity, AUC, and accuracy. The proposed framework of this study is shown in Figure 3.2.

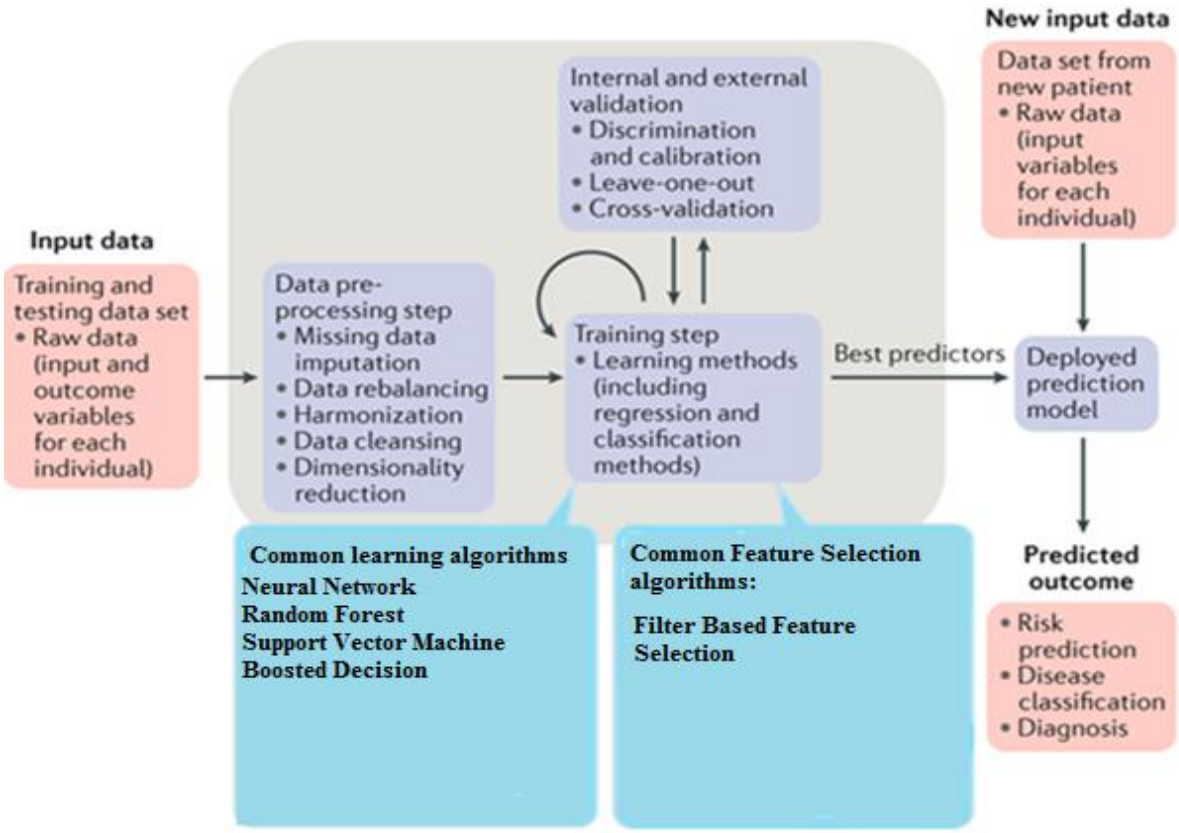


Figure 3.2: Study framework

3.6.1 Data Preprocessing

The objective of this stage is to obtain clean data, that is, data without null or anomalous values that allow obtaining quality standards. we will consider the relevance of certain attributes for the research. The collected dataset will need to be processed, altered, excluded, and other steps before it is entered into a machine learning technology, as the processing process will affect the nature and accuracy of the result.

3.6.1.1 Data Cleaning

In data mining, there must be a high quality of data before starting to build a model. Data quality problems can occur in any information system. Therefore, data cleaning can help solve these issues. To obtain high-quality and suitable data and Principal-Component

Analysis (PCA) were used in the data cleaning process. This improves the accuracy of the final data analysis [33]. In this study, the researcher used data cleaning to correct and remove data from a dataset for a variety of reasons, including irrelevant observations, incorrect, duplicate, outliers, incorrectly formatted corrupted, or incomplete data within the dataset. The data cleaning process is as follows:

- Remove duplicate or irrelevant data
- Structural bug fixes
- Filter out unwanted outliers
- Lost data processing

Following the data cleaning process, we estimate that the data set contained 4199 records, which has been reduced to 1794 complete records. the low of number record due to incomplete diagnosis of number of patients record. Also, the total number of variables was reduced from 70 to 30 variables due to \missing values for multiple variables at the same of record.

3.6.1.1.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is one of the most commonly used unsupervised machine learning algorithms across a variety of applications: exploratory data analysis, dimensionality reduction, information compression and data de-noising.

Principal Component Analysis (PCA) is a dimension reduction technique that reduces a big collection of variables to a small set that includes the majority of the information from the larger set [34]. PCA depends on a mathematical approach that applies an orthogonal transformation to reduce a large number of correlated variables into a smaller number of

uncorrelated linear variables called principal components [35]. After applying PCA, the dataset is reduced to 20 principal components from the previous 30 attributes which represent each observation.

3.6.1.2 Data Transformation

Data transformation is a method of converting raw data into a format that makes data mining and retrieving relevant data much easier. To transform the data into the appropriate form. In this step, the data is transformed as per the requirement of the model. Here are different data transformation techniques that can be used. The researcher used Standardization (Normalization) transform the data. *Normalization "is a data preparation technique that is frequently used in machine learning. The process of transforming the columns in a dataset to the same scale is referred to as normalization. Every dataset does not need to be normalized for machine learning. It is only required when the ranges of characteristics are different".* The goal of Standardization is to put different features on the same scale. This concept allows comparing scores between different types of variables. The following equation presents the data standardization:

$$Z = \frac{x - \min(x)}{\max(x) - \min(x)} \qquad 3.1$$

Where x is the data. In this study, the dataset normalized all variables except the class variable because of the range between 0 and 1.

3.6.1.3 Feature Selection

There are many factors or variables, which can be considered major or minor from a medical perspective. we also need to apply feature selection using statistical models, because the most practical data will be prioritized and the rest of the variables will be compared in terms of inclusion and weight, or significance value. Data might include a lot of features, some of which aren't always useful in getting results. The researcher in our study identified the most essential features that influence the goal variable (BC diagnosis). Because there are extraneous features that may degrade the performance of the models by training them on irrelevant attributes, getting the intended output is critical. There are numerous feature selection techniques available, but the filter feature selection tool from the Azure was employed in this investigation. This tool used a variety of methods, including chi-square, Pearson, Fisher and Spearman to choose features, as shown in the table below. Table 4 shows the order of attributes by the method.

Table 4: The attributes of the method

Id	Chi-Squared	Pearson Correlation	Spearman Correlation	Fisher Score
1	Mass	Mass	Mass	Mass
2	Density	density	density	density
3	Alcohol consumption	Alcohol consumption	Alcohol consumption	Alcohol consumption
4	Had used IUD hormonal	Had used IUD hormonal	Had used IUD hormonal	Had used IUD hormonal
5	age group	Previous Chest x-ray	Previous Chest x-ray	age group
6	Previous Chest x-ray	age at first child	age at first child	Previous Chest x-ray
7	age at first child	family history of the disease	family history of the disease	age at first child
8	family history of the disease	Lumpectomy Remove the small mass only	Lumpectomy Remove the small mass only	family history of the disease
9	Lumpectomy Remove the small mass only	Previous Heart	Previous Heart	Lumpectomy Remove the small mass only

10	Previous Chest CT scan	Previous Chest CT scan	Previous Chest CT scan	Previous Chest CT scan
11	Previous Heart	Previous Mammogram	Previous Mammogram	Previous Heart
12	BMI GROUP	BMI GROUP	BMI GROUP	BMI GROUP
13	Previous Lumpectomy	Previous Smoking status	age group	Previous Lumpectomy
14	Occupational hazard	Occupational hazard	Previous Smoking status	Occupational hazard
15	Previous Smoking status	age group	Occupational hazard	Previous Smoking status
16	Previous Mammogram	Used Birth control Implant	Used Birth control Implant	Previous Mammogram
17	Type of locality	Previous Lumpectomy	Previous Implants	Previous complain of any chronic diseases
18	Previous complain of any chronic diseases	Previous Implants	Previous Lumpectomy	Previous Hypertension
19	Previous Hypertension	Previous complain of any chronic diseases	Previous complain of any chronic diseases	MM_PRFRChem4apy
20	Previous Chemotherapy	Had used Hormonal injection every 3 months	Had used Hormonal injection every 3 months	Used Birth control Implant
21	Used Birth control Implant		Previous Cyst aspiration	Marital Status
22	Previous Implants		Previous Hypertension	Previous Implants
23	Marital Status		Marital Status	Had used Hormonal injection every 3 months
24	Had used Hormonal injection every 3 months		Type of locality	Previous Hormonal therapy
25			Duration of BF breast feeding	Previous Hormonal therapy
26	Previous Hormonal therapy		Previous Hormonal therapy	Duration of BF breast feeding
27	Duration of BF breast feeding		Previous Hormonal therapy	Previous Cyst aspiration
28	Previous Cyst aspiration		Had used Oral contraceptives	Type of locality
29	Menarche age		Menarche age	Menarche age
30	Had used Oral contraceptives		MM_PRFRChem4apy	Had used Oral contraceptives Had used Oral contraceptives

.

However, to identify critical features as risk factors for prediction breast cancer expert physician was consulted. The expert is Dr. Abla Fawaz Al-fari specializes in breast cancer She works in the Palestinian Ministry of Health in the Jenin Department.

The order of risk factors according to the doctor's opinion was arranged in the following table.

Table 5: Expert opinion on risk factors for breast cancer

Id	feature
1	History of personal breast cancer
2	Age
3	Genetic BRCA1/BRCA2
4	Marital status
5	A family member had cancer
6	Menarche early
7	First child (35-40)
8	Density
9	Occupation Hazard
10	Radiation X-ray CT
11	Hormones therapy
12	Oral contraption
13	implants
14	Obesity BMI
15	Alcohol
16	Chronic disease

We conclude that used feature selection in addition of PCA increase the performance of the models.

3.6.2 Data Partitioning

There are several ways to partition data for experimentation in machine learning. Training/test partitioning and cross-validation are two of the most common approaches. Data is typically partitioned into a training set and a test set in a specific ratio during training/test partitioning [36]. In the study, 70% of the data was used as the training set, while 30% of the data was used as the test set. This data partitioning was done either at

random or in a predetermined order. The first 70% of the data set's instances are assigned to the training set, while the rest are assigned to the test set. The instances were randomly assigned to training and test sets. Cross-validation is the process of testing a model against data that it hasn't been trained to estimate out-of-sample error. It is accomplished by first dividing data into folds. The data was divided into five folds for the study. After that we trained a model on the first- four folds and tested it on the fifth fold in the first iteration. The study was then trained on fold 3,4,5 and tested on fold 1 in the second iteration. After that, alternate which fold is the test fold until all of the folds are test folds (i.e. we train and test 5 times in total). Finally, we calculate the average performance of the models and compare them to each other shown Figure 3.3.

Cross-validation is regarded as an effective measure of an algorithm's learnability, or the degree to which the algorithm is capable of learning a high-quality model from supplied training data. This allows for the proper use of relevant learning algorithms to produce predictive models based on existing data

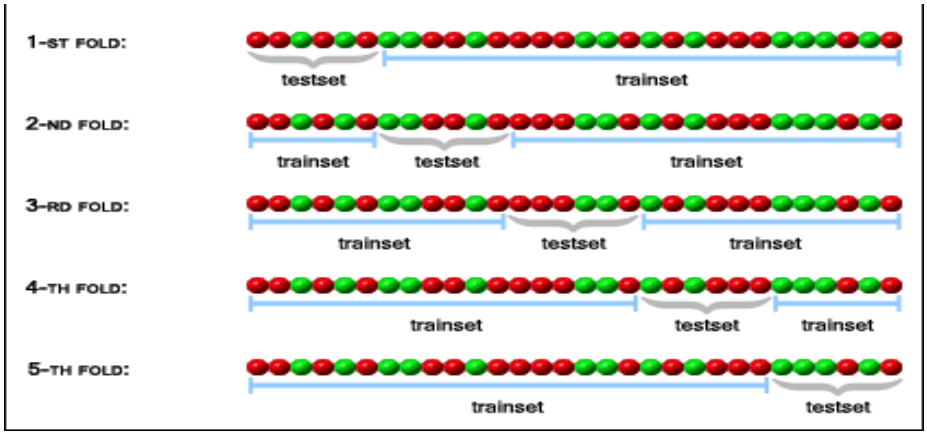


Figure 3.3: 5-fold cross-validation

3.6.3 Building Models Phase

The objective of the data mining stage is the search and discovery of unsuspected and hidden patterns by applying discovery tasks such as classification and prediction. Different data mining models were applied such as neural networks, Decision Tree (DT), Support-Vector Machine (SVM), and Decision forest. The final model is a classification of cancer into malignant, benign, and healthy cases.

3.6.3.1 Artificial Neural Network Multiclass Classifier

Artificial neural networks are types of neural networks (NNs) that are used in deep learning techniques. Their name and structure are inspired by the human brain, and they function similarly to organic neurons. It is comprised of a large number of simple, well-connected processing elements [37]. NNs contain a node layer containing an output layer, an input layer and, one or more hidden layers. Each node, or artificial neuron, is linked to the others and has its threshold and weight. If a node's output exceeds a certain threshold value, the node is activated, and data is sent to the next layer of the network. Aside from that, no data is sent on to the next layer of the network Figure 3.4 describes the Neural network [38].

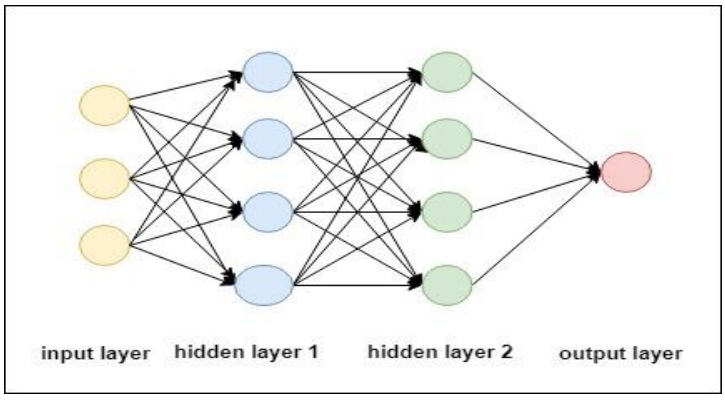


Figure 3.4: Neural network structure

So, every node has its linear regression model, which includes weights, input data, threshold (or a bias), and an out. The formula would look something like this:

NNs have two stages: the first stage is feed forward. In this stage, the output is predicted and it calculates the error then sent the error back to the backward prorogation stage. The second stage is backward propagation, through this stage; the error is propagated back through the network to modify the values of the weights and to make reduce the error value in the output layer. We can explain how NN works clearly through these main steps:

1. Input Data: In the first step we enter the training data for which we already know the correct output, after which it is processed through the hidden layers of the neural network.

2. Data processing by neurons (weights): At first, the input values are multiplied by the weights, according to this equation:

$$Y_{ij} = f\left(\sum_{i=1}^{n} Xi * w_{ij}\right) \quad 3.2$$

Where w_{ij} : is the connection weight between the *i*th node in the input layer and the *j*th node in the hidden layer, and x_i: is the *i*th input, and *f* is the activation function. It is worth noting that in each cell the layer is connected to all the cells in the next layer. The activation function *f* calculated as in the following equation:

$$Y_j = \frac{1}{1+e^{-y}} \quad 3.3$$

Where X_k is the input to the next layer' node.

3. Repeat the process with the other hidden layers: These operations are repeated with the other layers in the NNs until the last layer, which is the output layer.

4. Determine the output of the NN: The output values are ultimately determined by the type and function of the NNs, so we get the results and compare them with the real output.

5. Calculate the error: To break the process of training, there is a certain threshold θ is set depending on the error of the NN which represents the difference between the desired and actual output. The error is calculated using the following equation:

$$E = \frac{1}{2}\sum_{i}^{n}(y_d - y_j)^2 \quad 3.4$$

Where y_d is desired output, y_j is the actual output. So, by comparing the output with the results we already know.

6. Modify the weights to obtain more accurate results using the following equation:

$$w_{jk} = w_{jk} + \Delta w_{jk} \quad 3.5$$

$$\Delta w_{j+1} = \alpha . E . x_i$$

α is the learning rate, which is a constant value between 0.01 to 1. The back-propagation algorithm repeats until the error on the output node is minimize

3.6.3.2 Random Forest

Random Forests (DF) is a supervised machine learning algorithm used for regression, classification, and other tasks that works by training a large number of decision trees. For classification tasks, the random forest's output is the class chosen by the majority of trees. The mean or average prediction of the individual trees is returned for regression tasks. Random Forests address the problem of decision trees overfitting their training set. Random forests outperform decision trees in most cases. However, data characteristics can influence how well they perform shown in Figure 3.5 the DF diagram [39].

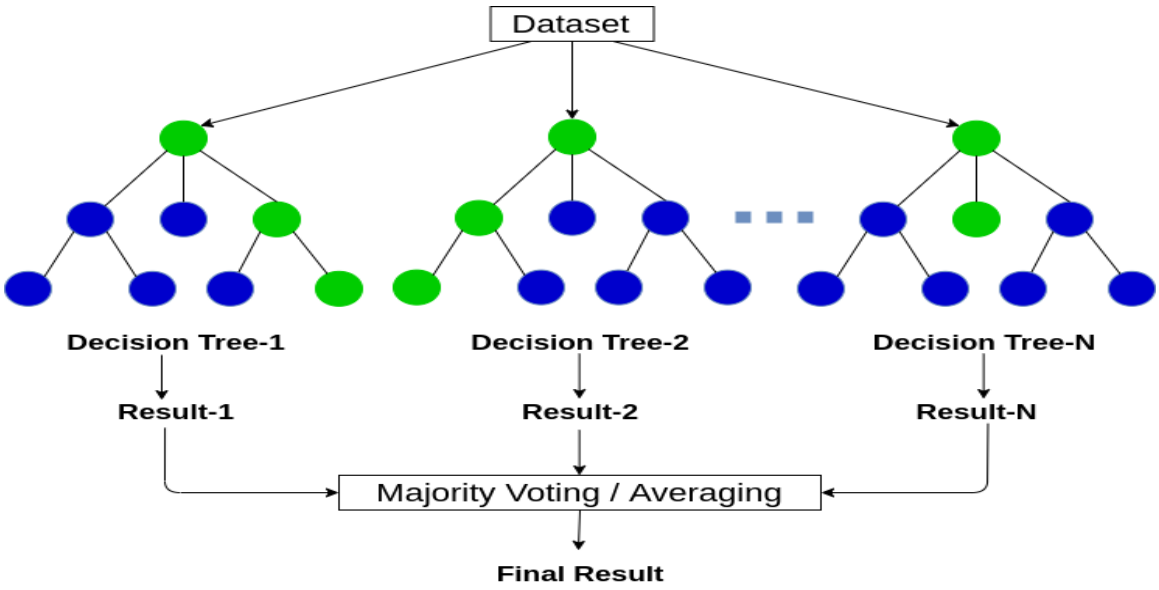


Figure 3.5: Random Forest

3.6.3.3 Multiclass Logistic Regression Classifier

In machine learning, logistic regression is a classification method. The dependent variable is represented by a logistic function. The dependent variable is dichotomous, which means that only two classes are conceivable. the observation can be categorized into one of two or one of many classes.

The purpose of binary logistic regression is to train a classifier that can determine the class of a new input observation using a binary choice. To compute p(yk = 1jx), the multinomial logistic classifier employs a generalization of the sigmoid . The sigmoid function transfers a vector of K arbitrary values z = [z1;z2;:::;zK] to a probability distribution, with each value in the range (0,1) and all values adding up to 1. It's an exponential function, like the sigmoid shown in figure 3.6 [40].

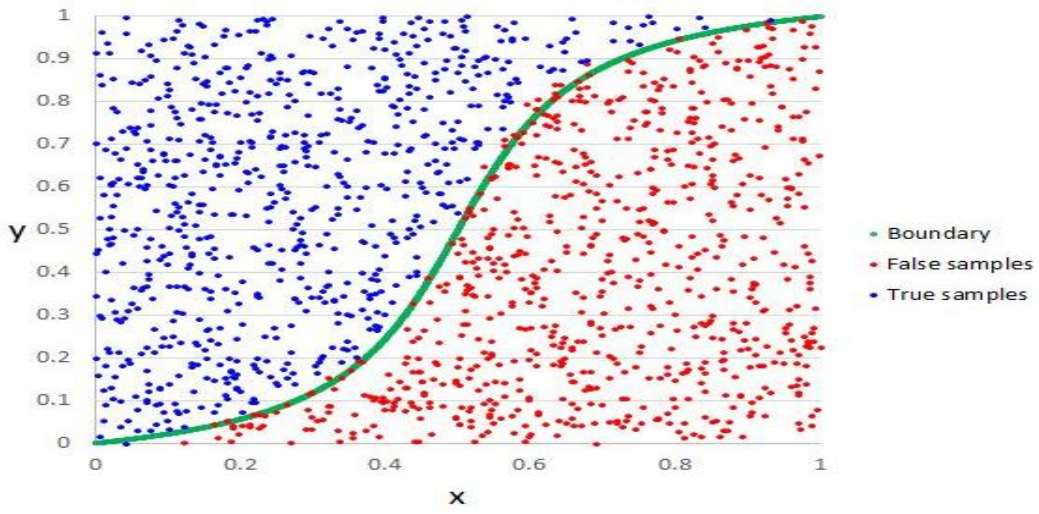


Figure 3.6: Logistic Sigmoid Function

3.6.3.4 Support Vector Machine Classifier

The "Support Vector Machine" (SVM) is a supervised machine learning technique that can solve classification and regression problems. however, mostly employed to solve classification issues. Each data item is plotted as a point in n-dimensional space (n meaning number of features in the dataset), also with the value of each feature becoming a value of a certain coordinate in the SVM algorithm. Then we accomplish classification by locating the hyper-plane that clearly distinguishes the two classes [41].

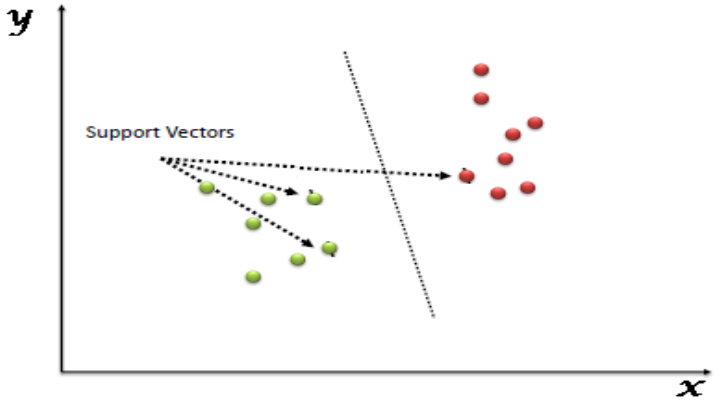


Figure 3.7: SVM hyper-plane

Supporting vectors are the data points closest to the hyper-plane and are the points that, if removed from the data set, will change the location of the hyper-plane that divides the data. So, these points can be considered the important elements in the data set. So, we want the points to be as far from the hyper-plane as possible keeping in mind that they stay on the right side of the dividing line. So, when we add new data to test it will be categorized based on which side it is on for the hyper-plane. The distance between the hyper-plane and the nearest point of any data set is known as the margin. The goal is to choose a hyper-plane with the largest margin between it and any point in the training data set, increasing the likelihood that any new data will be correctly classified. Figure (3.4-4) shows the margin, which is the distance between the hyperplane and its nearest points.

To understand hyper-plane equations and how to create them, given the training dataset:

$$(x_1,y_1), \ldots, (x_n,y_n),\ x_i \epsilon\ R^d \text{ and } y_i \epsilon\ (-1,1) \qquad 3.6$$

Where X_i is the symbol of the feature vector and Y_i is the class label (1 or 0) of training compound i. Hence, to express the optimum hyperplane equation:

$$wx^T + b = 0 \qquad 3.7$$

Where W is the weight vector, X is the feature vector and b is the bias.

The objective of the training dataset on SVM is to determine the suitable b and w to separate the data with optimum hyperplane and the max-margin distance [42]. SVM The Precision, works well on small and refined data sets. In contrast, SVM is not suitable for large datasets [43].

3.6.3.5 Boosted Decision Trees

Boosted decision trees are algorithms for improving a model's predictive value by iteratively improving its predictive value. The values mean weights, the coefficients, or biases. It is applied to each of the input features used to predict the output value. The output is adjusted in each iteration of the decision tree. The algorithms measure of the difference between the actual target and predicted values [44].

3.6.4 Models Evaluation

The performance of the model was tested using many criteria after machine learning algorithms implemented various strategies. The goal of evaluation is to have a better knowledge of the patterns by interpreting the data. The technique is accuracy, sensitivity, and specificity, which are concluded by the confusion matrix.

Actual class		Predicted Class	
		Positive	negative
	Positive (malignant or benign)	True Positive **436**	False Negative **9**
	Negative (Healthy)	False Positive **34**	True Negative **378**

Figure 3.8: Confusion matrix

The measures used to evaluate classification techniques are as follows:

3.6.4.1 Accuracy

The model's accuracy is defined as the proportion of correct predictions to total predictions. The accuracy formula is given in:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad 3.8$$

Table 6: Terms used to utilize Performance [45].

Outcome of the diagnostic test	Condition As determined by the Standard of Truth		
	Positive	Negative	Row Total
Positive	TP	FP	TP+FP (Total number of subjects with positive test)
Negative	FN	TN	FN + TN (Total number of subjects with negative test)
Column total	TP+FN (Total number of subjects with given condition)	FP+TN (Total number of subjects without given condition)	N = TP+TN+FP+FN (Total number of subjects in study)

The abbreviations of terms are:

- Positive (P): Observation is positive (positive cancer diagnosed or cancer).
- Negative (N): Observation is not positive or negative (for example: negative cancer diagnosed or healthy).
- True Positive (TP): Observation is positive, and is predicted to be positive.
- False Negative (FN): Observation is positive, but is predicted negative.
- True Negative (TN): Observation is negative, and is predicted to be negative.
- False Positive (FP): Observation is negative, but is predicted positive.

3.6.4.2 Sensitivity

Sensitivity known as the True Positive Rate, is the proportion of positive classes correctly classified as positive in comparison to all positive classes in the data. The formula in Equation 3.2 is used to calculate sensitivity:

$$\text{Sensitivity} = \frac{number\ of\ true\ positives}{number\ of\ true\ positives + number\ of\ false\ negatives} \quad 3.9$$

$$= \frac{number\ of\ true\ positives}{total\ number\ of\ sick\ individual\ in\ population}$$

$(probability\ of\ positive\ \ test\ given\ that\ the\ patient\ has\ disease\ )$

3.6.4.3 Specificity

Specificity known as the False Positive Rate is the ratio of the negative classes that are wrongly classified as positive concerning all the negative classes in the data. Specificity is calculated with the formula given in Equation below[45]:

$$\text{Sensitivity} = \frac{number\ of\ true\ negatives}{number\ of\ true\ negatives + number\ of\ false\ positives} \quad 3.10$$

$$= \frac{number\ of\ true\ negatives}{total\ number\ of\ well\ individuals\ in\ population}$$

$(probability\ of\ negative\ \ test\ given\ that\ the\ patient\ is\ well\ )$

3.6.4.4 Precision

Precision is the total number of correctly classified positive samples divided by the total number of actual positive samples.

$$\text{Precision} = \frac{TP}{(TP + FP)} \quad 3.11$$

3.6.4.5 Recall

The recall is the ratio of the total number of correctly classified positive samples divided by the predicted total number of positive samples.

$$\text{Recall} = \frac{TP}{TP + FN} \quad 3.12$$

3.6.4.6 F-score

The F1 score is the harmonic mean of precision and recall. It can be used as a single measure of performance of the test for the positive class. F1 score can be defined as the harmonic mean of recall and precision. It is used as a single indicator of test performance for the positive class. F1 score calculated based on the following Equation:[46]

$$\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \qquad 3.13$$

3.6.4.7 Area Under Curve (AUC)

AUC is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The AUC indicates how well the model distinguishes between positive and negative classes. The greater the AUC, the better model. So, plots the true positive rate (Sensitivity) against the false positive rate (100-Specificity) as its discrimination threshold is varied. All the points on the ROC curve can be regarded as a sensitivity/specificity pair corresponding to a particular decision threshold.

3.7 Ethical Consideration

Written approval was obtained from the Arab American University to conduct the study shown in Appendix 1. After that Approval was obtained from the Ministry of Health to get the risk factors dataset of breast cancer from the MOH database shown in Appendix 2.

3.8 Limitation Of The Study

The study has several limitations. First, because the dataset was derived using an observational study approach, selection bias cannot be ignored. Deriving the dataset from a big cohort study with enough breast cancer patients would be a preferable option.

However, given the high frequency of breast cancer, utilizing such a dataset would result in unbalanced classes, which would need the use of statistical approaches. Second, owing to the dataset's restrictions, only 20 breast cancer risk variables were selected to develop the model, and certain crucial risk factors such as a history of other breast illnesses were left out, which might have altered the models' performance. Furthermore, because our machine learning models were trained on cases and nonaccess in the West Bank, they may not be applicable elsewhere such as in Jerusalem and Gaza. Third There is no integration between the mammography electronic Registry system and his system or cancer system so can’t, so inability to acquire medical tests and link risk variables to the patient's medical examinations to predict disease. Fourth Many attributes were reused some values were lost during the data entering procedure in the health directorates. Fifth Full medical information about the results of biopsy or Fine-needle aspiration tests is not recorded in the tissue department and therefore maybe it is difficult to use this feature in our study, which leads us to be satisfied with predicting cancer by risk factors attributes.

In addition, the mammogram images were not used for patient privacy, as the patient's ID number is on the mammogram images and it is not possible to remove it from the mammography company. In addition to the scarcity of full-time doctors to evaluate and diagnose mammogram images.

3.9 Summary

This chapter discussed the study's settings, population and sample size, data management and ML Framework, feature selection, study methods, data collection techniques, data management and analysis framework, machine learning models’ ethical issues, and study limitations

Chapter Four: Results and Discussion

4.1 Introduction

The results of the various analytic techniques conducted on the BC dataset are presented in this chapter. This study used descriptive analysis and data mining techniques. In addition, the results of the ML algorithms were presented and discussed.

The results obtained with data mining techniques explore data, make sense of it, into knowledge to interpret a phenomenon and make decisions according to the needs.

4.2 Descriptive Statistics and Visualization

In this section, descriptive dataset analysis and visualization were implemented Using SPSS version-20. The study analysis all variables to summarize the frequency and percentage for each variable. Table 7 shows the frequency of benign, malignant, and control cases for each risk factor. The analysis of data was applied to balanced dataset. It contains 570 out of 1794 (31.8%) Healthy cases, 589(32%) benign and 635(35%) malignant (table 7 and figure 4.1).

Table 7: the frequency of benign, malignant, and control cases.

		Frequency	Percent	Cumulative Percent
Valid	healthy	570	31.8	31.8
	benign	589	32.8	64.6
	malignant	635	35.4	100.0
	Total	1794	100.0	

.

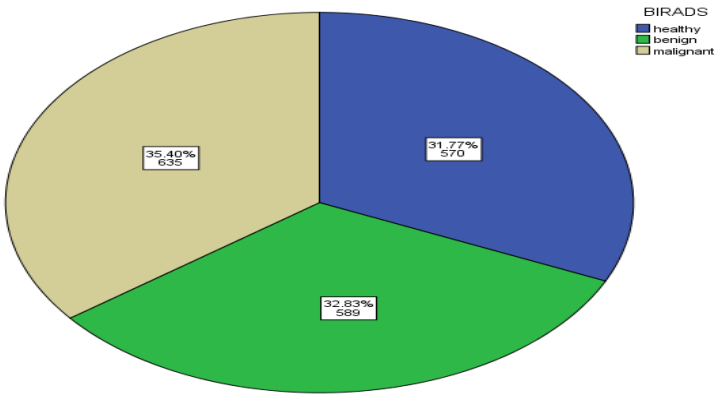


Figure 4.1: Classification of cases by diagnosis (BIRADS)

4.2.1 Sociodemographic Characteristics

The study found 63.27% cases from the villages, 31.22% the from cities, 5.24% from camps, 1 (and (0.28%) from another locality show Figure 4.2.

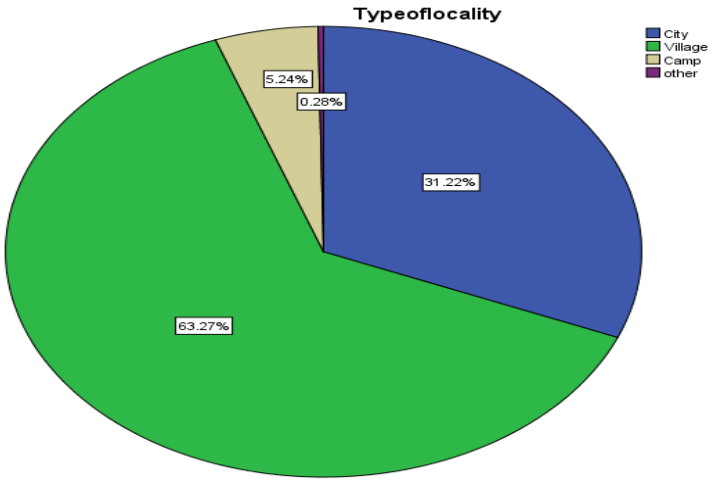


Figure 4.2: Distribution of cases across the type of locality.

Overall, 556 malignant cases were of married women, while 75 malignant cases were of widows, separated, and divorced women. As the study found 509 benign cases were of married women, while 80 benign cases were of widows, separated and divorced women shown in figure 4.3.

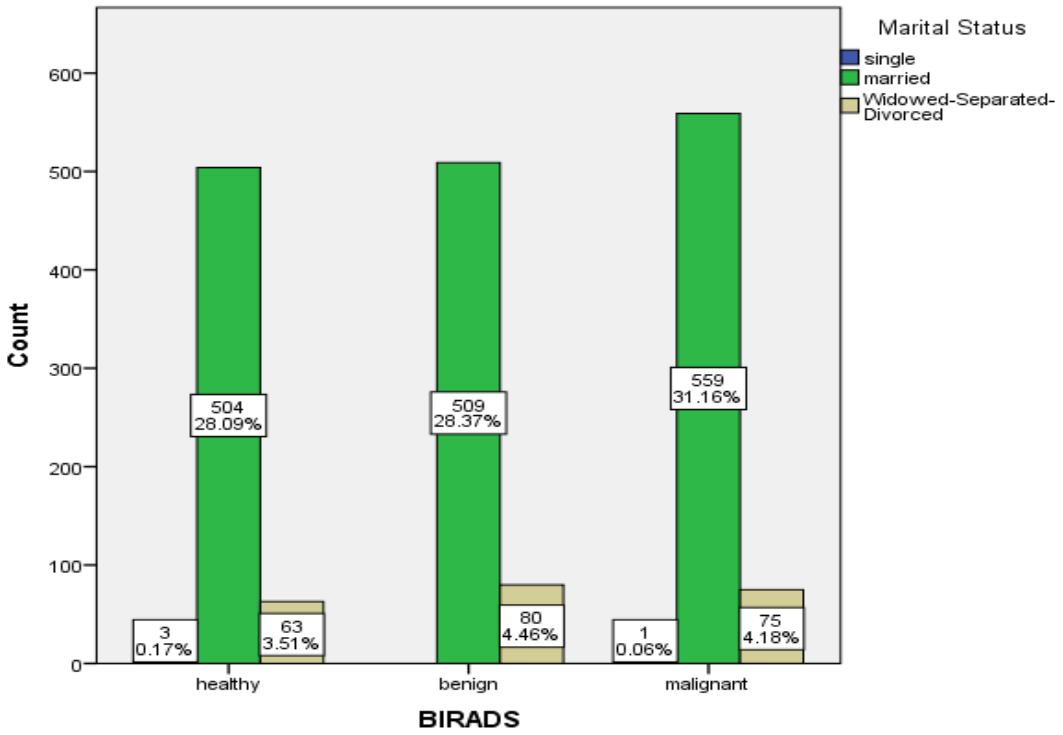


Figure 4.3: Distribution of cases across the type of marital status.

The study included women aged 24 to more than 94, these ages were divided into groups, each and group containing 10 ages. In this study, the highest age-group frequency for women who underwent cancer screening in 2021 and 2020 were 44-54, 54-64, and 44-43, respective shown in figure 4.4.

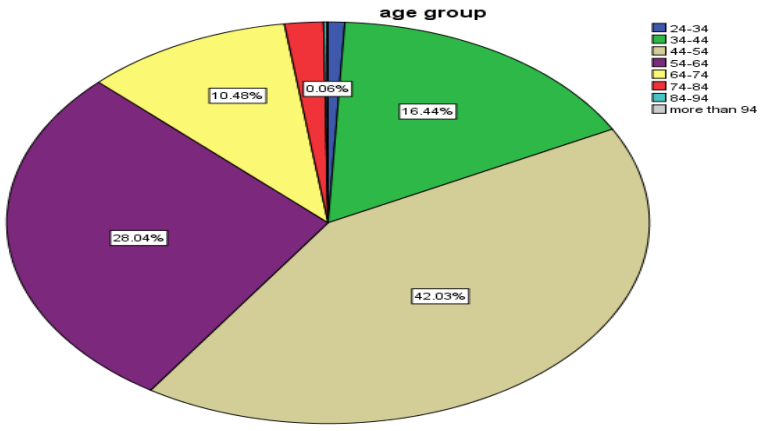


Figure 4.4: age group of the study

During 2020-2021, the median age at the time of breast cancer diagnosis of women in the study was 51.3 years. This means that half of the Palestinian women have malignant breast cancer are below of 51.3 years. Show figure 4.5 the age distribution depending on diagnosis cases.

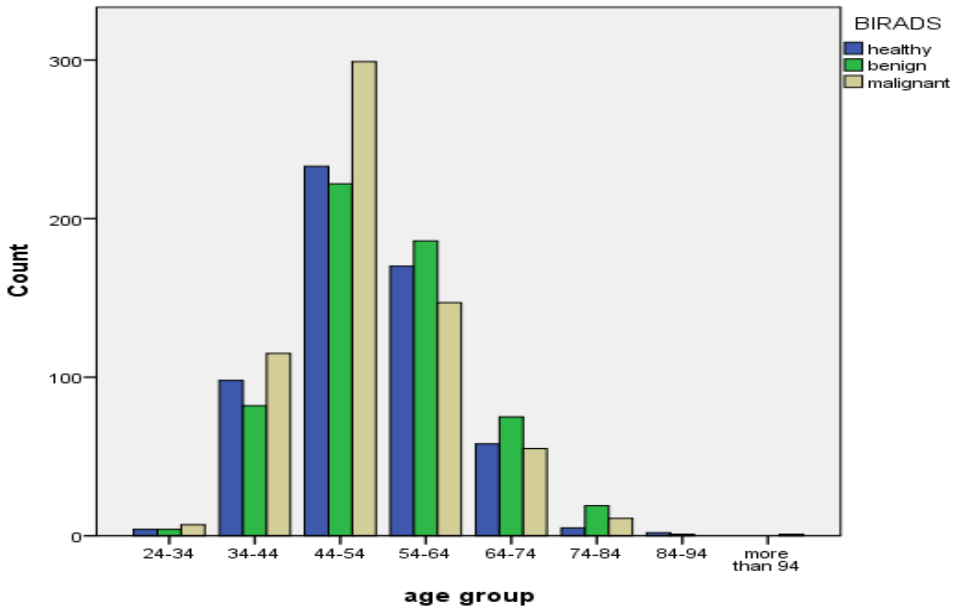


Figure 4.5: age of women in study dataset.

In the study, malignant breast cancer is concentrated in the ages ranging from 44-54, where there was 299 out 6 of 35 malignant cases, while there are 147 out 635 in the age group 54-64 years. Also, benign breast cancer is concentrated in the ages ranging from 44-54 where there were 186 malignant cases, while there are 147 in the age group 54-64 years. Therefore, the highest incidence of malignant cancer was in the age group 44-54 then 54-64 years show in figure 4.6.

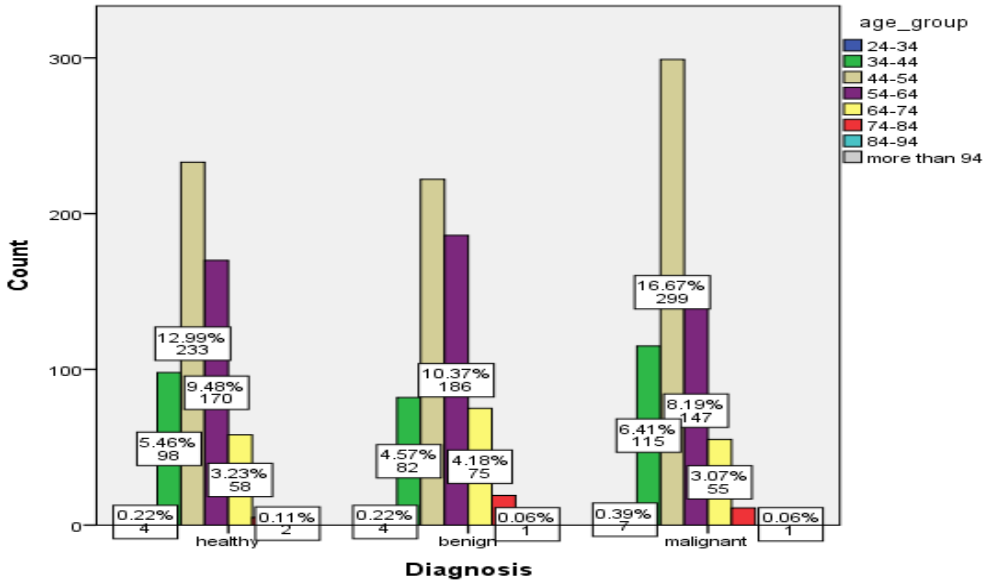


Figure 4.6: the women's age distribution on the diagnosis.

There are 493 patients with malignant breast cancer who were age is >= 30 at the time of their first birth, 419 with benign breast cancer, and 376 with a healthy diagnosis. There were 142 malignant cases, 170 benign cases, and 194 benign cases among women under the age of 30 figure 4.7. This is confirmed by previous studies that the younger a woman is when she begins childbearing, the lower risk of breast cancer [47]..

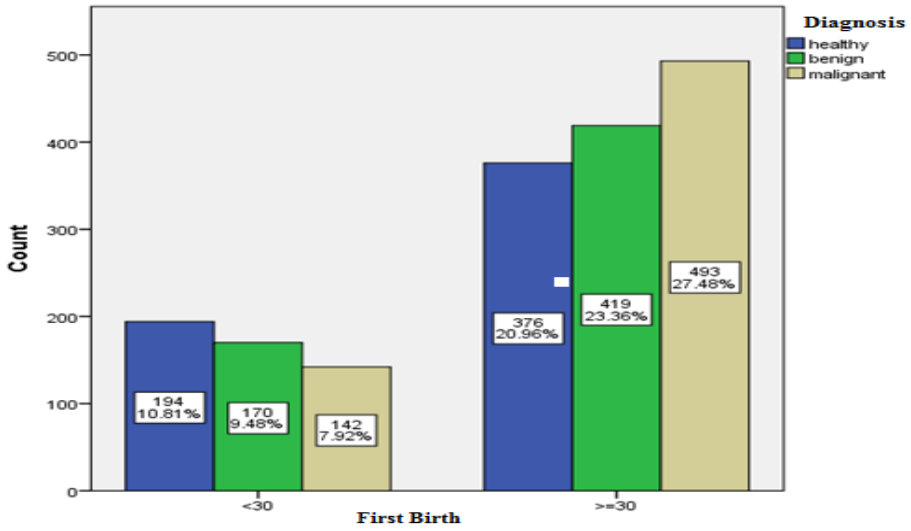


Figure 4.7: First birth of women

4.2.2 Density of Breast

The density is one Factor of BC, the study analyzed the density as shown in Figure 4.8. There are 95 cases of heterogeneous dense and 18 cases of extremely dense from malignant cases. While 34 cases of heterogeneous dense and 2 cases of extremely dense from benign cases. Also, 10 cases of heterogeneous dense and 0 cases of extremely dense from healthy cases shown in Figure 4.8. This confirmed study that high density increases the incidents of BC [48].

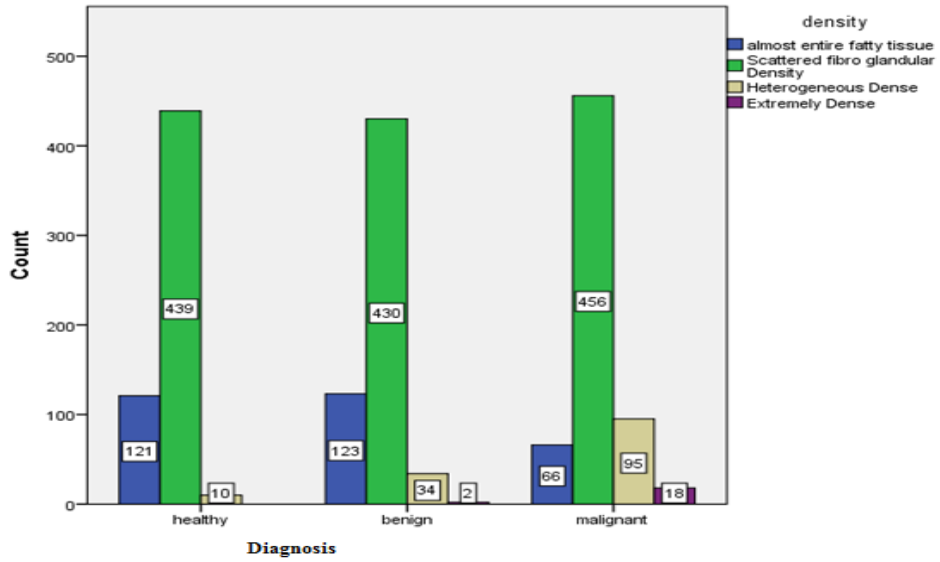


Figure 4.8: result of density breast cancer

4.2.3 Menarche Age

Previous studies confirmed that early menarche is known to increase women's risk of developing breast cancer. Early age at menarche is associated with a 4% per-year increase in the relative risk of breast cancer[49]. In our study, 60 percent of malignant breast cancer patients had a menstrual age of fewer than 13 years, while 39.69 percent had a menstrual age of more than 13 years shown in figure 4.9.

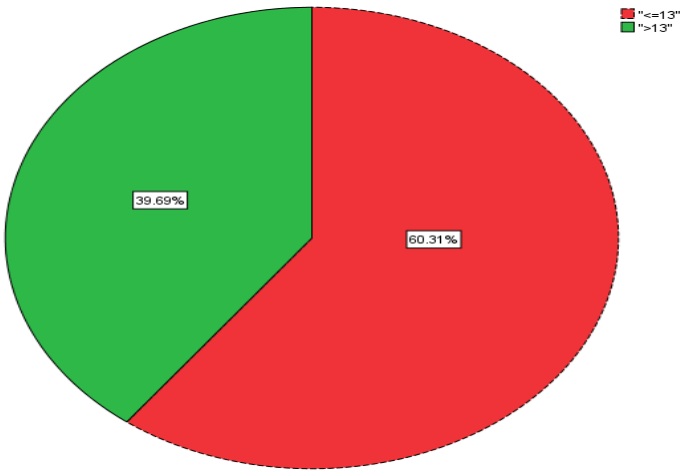


Figure 4.9: Menarche age of malignant cases

Especially among women of childbearing age, masses are extremely common. Over 25% of all women will develop breast disease at some point in their lives, and the massive majority of cases will display as a new breast mass in primary care [14]. In the study that 202 malignant cases contain one mass, 34 cases contain two masses, 10 cases contain three masses, 6 cases contain four masses, and five cases contain 5 masses. While 36 benign cases contain one mass and 6 cases contain two masses. However, there are no masses in healthy cases shown in Figure 4.10 number of masses in benign and malignant.

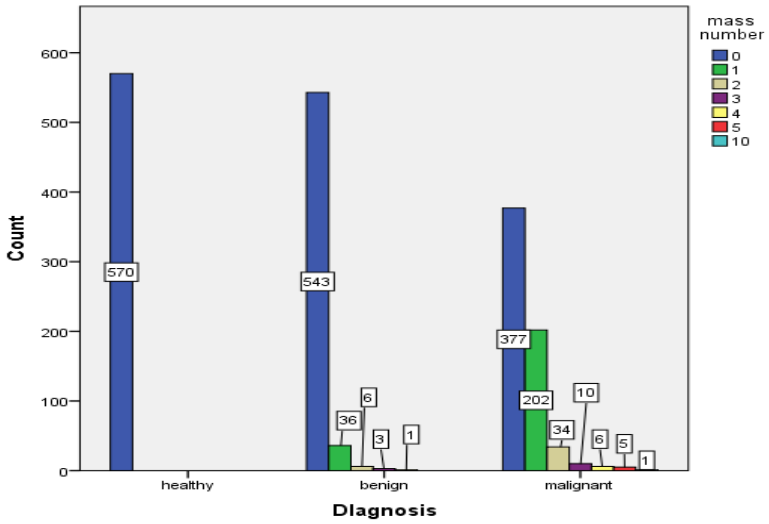


Figure 4.10: Mass of breast

4.2.4 Occupational Hazard

In the study, occupations were divided based on occupational factors and the risks arising from work Based on a previous study evaluating occupations according to exposure to radiation, pesticides, or others. Occupations are divided into occupations with risk factors and occupations with no risk factors. The study found that those who worked in professions with risk factors were more likely to develop malignant breast cancer [50]. In the study, 107 cases of malignant breast cancer and 9 cases of benign breast cancer in people who worked in professions with risk factors show Figure 4.11

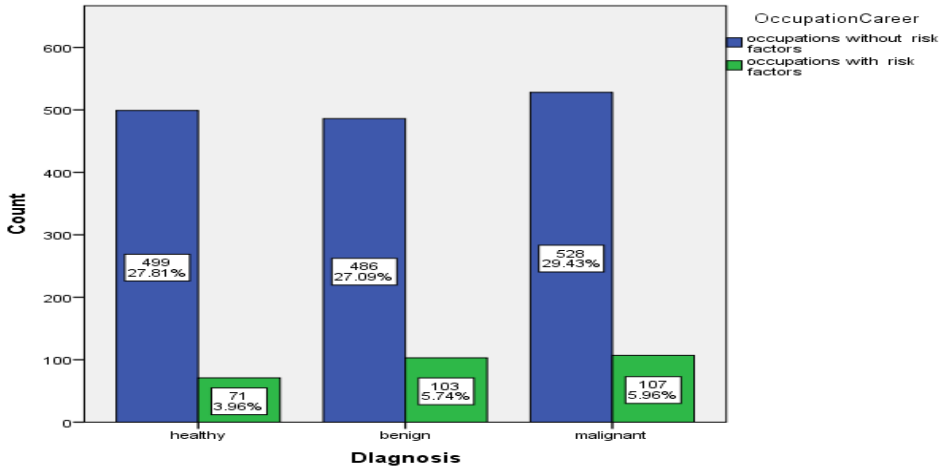


Figure 4.11: Occupational hazard

4.2.5 BMI Risk Factors

When looking at the link between breast cancer and weight, most current research employ BMI [14][10]. However, it is possible to have a normal BMI and a high amount of body fat. Women with excess body fat, even if their BMI is acceptable, have an increased risk of breast cancer, according to new research. A BMI of less than 18.5 is underweight. Between 18.5 and 24.9, you fall into the normal range for your BMI for your height. A

BMI above 25 is considered overweight, and over 30 is considered obese. In the study, BMI was divided into more than 25 to be considered overweight and less than 25 no obesity or overweight. the result was 547 BC malignant and 542 BC benign that BMI over 25 shown in figure 4.12

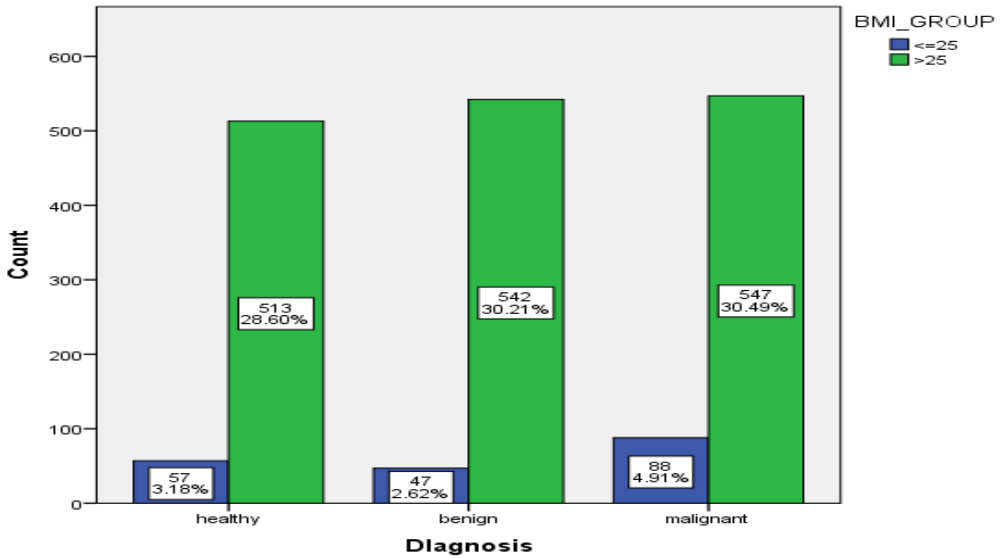


Figure 4.12: BMI risk factor

4.2.6 Previous Using Mammogram

The study showed that there are 240 (13.3%) malignant cases and the mammogram was used previously. While 201 (11.2%) of the data set were benign cases show figure 4.13.

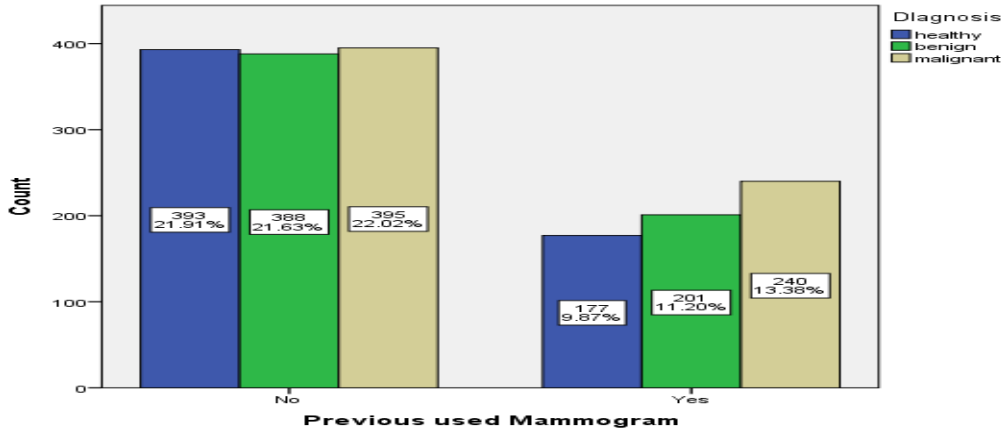


Figure 4.13: Previous mammograms

4.2.7 Family history

The study shown that 265 cases malignant breast cancer history. While 220 cases benign had breast cancer family history shown Figure 4.14.

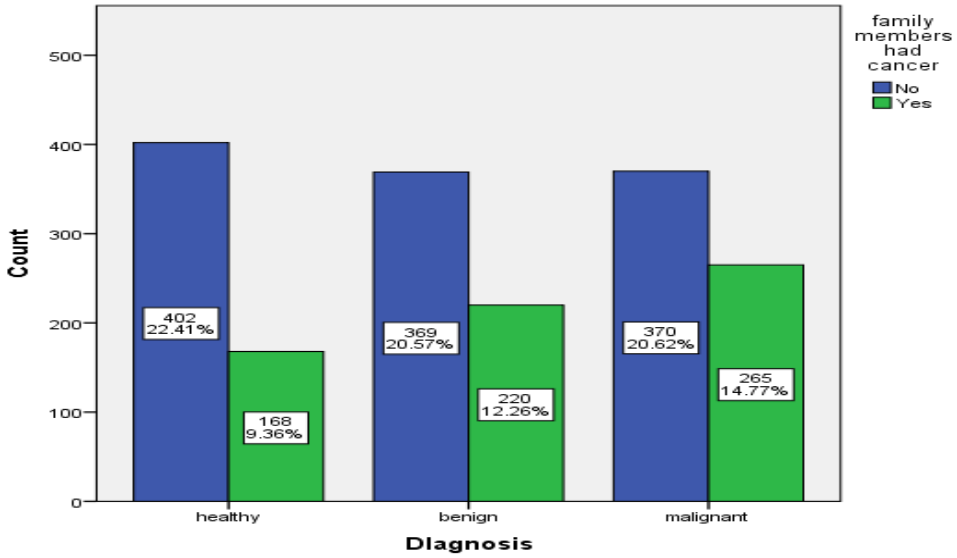


Figure 4.14: Family had history

4.2.8 Other Risk Factors

There are other risk factors such as oral contraceptive use. Figure 4.15 shows that there is approximately the same percentage for diagnosis and use of oral contraceptives.

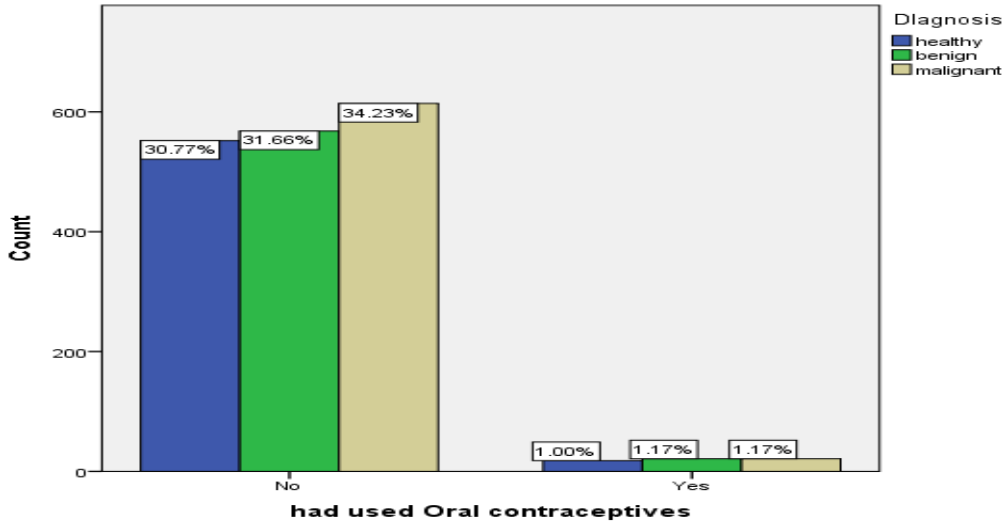


Figure 4.15: oral contraceptives risk factor

On another side, the data analysis for all variables (30 feature risk factors) table 6 show Frequencies distribution of 30 risk factors among malignant and benign cases and controls. Also, the correlation between the dependent variables (diagnosis) and the independent variables (risk factors). Table 8 shows the frequencies distribution of 30 risk factors among malignant and benign cases and controls.

Table 8: Frequencies distribution of 30 risk factors among malignant and benign cases and controls.

		Diagnosis					
		healthy		benign		malignant	
		Count	Column N %	Count	Count	Column N %	Count
Hormonal therapy	No	96.3%	559	549	96.3%	559	96.1%
	Yes	3.7%	30	21	3.7%	30	3.9%
Duration of breastfeeding	Less or 6 months	9.8%	71	56	9.8%	71	10.6%
	More than 6 months	90.2%	518	514	90.2%	518	89.4%
Density	almost entire fatty tissue	21.2%	123	121	21.2%	123	10.4%
	Scattered fibro glandular Density	77.0%	430	439	77.0%	430	71.8%
	Heterogeneous Dense	1.8%	34	10	1.8%	34	15.0%
	Extremely Dense		2			2	2.8%
Have family members had cancer	No	70.5%	369	402	70.5%	369	58.3%
	Yes	29.5%	220	168	29.5%	220	41.7%
Marital Status	Single	0.5%		3	0.5%		0.2%
	married	88.4%	509	504	88.4%	509	88.0%
	Widowed-Separated-Divorced	11.1%	80	63	11.1%	80	11.8%
Occupational hazard	Not Working	87.5%	486	499	87.5%	486	83.1%
	Working	12.5%	103	71	12.5%	103	16.9%
Type of locality	City	31.9%	174	182	31.9%	174	32.1%
	Village	63.5%	386	362	63.5%	386	60.9%
	Camp	4.2%	28	24	4.2%	28	6.6%
	other	0.4%	1	2	0.4%	1	0.3%
Smoking status	Non smoker	83.0%	486	473	83.0%	486	87.2%
	smoker	17.0%	103	97	17.0%	103	12.8%
	No	63.3%	336	361	63.3%	336	59.7%

Complain of any chronic diseases	Yes	36.7%	253	209	36.7%	253	40.3%
used IUD hormonal	No	69.1%	346	394	69.1%	346	49.3%
	Yes	30.9%	243	176	30.9%	243	50.7%
Had used Birth control Implant	No	100.0%	588	570	100.0%	588	99.5%
	Yes		1			1	0.5%
GOHRhadusedHormonalinjectionevery3months	No	98.4%	584	561	98.4%	584	99.1%
	Yes	1.6%	5	9	1.6%	5	0.9%
Had used Oral contraceptives	No	96.8%	568	552	96.8%	568	96.7%
	Yes	3.2%	21	18	3.2%	21	3.3%
mass	0	100.0%	543	570	100.0%	543	59.4%
	1		36			36	31.8%
	2		6			6	5.4%
	3		3			3	1.6%
	4		1			1	0.9%
	5						0.8%
	10						0.2%
Alcohol consumption	No	71.1%	494	405	71.1%	494	90.6%
	Yes	9.4%	95	60	9.4%	95	28.9%
Previous Chemotherapy	No	94.9%	545	541	94.9%	545	94.6%
	Yes	5.1%	44	29	5.1%	44	5.4%
Previous Chest CT scan	No	96.8%	536	552	96.8%	536	92.6%
	Yes	3.2%	53	18	3.2%	53	7.4%
Previous Chest x-ray	No	74.4%	396	424	74.4%	396	61.9%
	Yes	25.6%	193	146	25.6%	193	38.1%
Previous Cyst aspiration	No	99.3%	586	566	99.3%	586	99.7%
	Yes	0.7%	3	4	0.7%	3	0.3%
Previous Heart	No	98.6%	568	562	98.6%	568	94.3%
	Yes	1.4%	21	8	1.4%	21	5.7%
Previous Hormonal therapy	No	96.3%	559	549	96.3%	559	96.1%
	Yes	3.7%	30	21	3.7%	30	3.9%
Previous Hypertension	No	73.9%	405	421	73.9%	405	71.3%
	Yes	26.1%	184	149	26.1%	184	28.7%
Previous Implants	No	99.8%	589	569	99.8%	589	100.0%
	Yes	0.2%		1	0.2%		
Previous Lumpectomy	No	98.2%	560	560	98.2%	560	96.9%
	Yes	1.8%	29	10	1.8%	29	3.1%
	No	98.6%	559	562	98.6%	559	93.5%

Previous Lumpectomy Remove the small mass only	Yes	1.4%	30	8	1.4%	30	6.5%
Previous Mammogram	No	68.9%	388	393	68.9%	388	62.2%
	Yes	31.1%	201	177	31.1%	201	37.8%
Age group	24-34	0.7%	4	4	0.7%	4	1.1%
	34-44	17.2%	82	98	17.2%	82	18.1%
	44-54	40.9%	222	233	40.9%	222	47.1%
	54-55	29.8%	186	170	29.8%	186	23.1%
	64-64	10.2%	75	58	10.2%	75	8.7%
	74-84	0.9%	19	5	0.9%	19	1.7%
	84-94	0.4%	1	2	0.4%	1	
	more than 94						0.2%
BMI_GROUP	<=25	10.0%	47	57	10.0%	47	13.9%
	>25	90.0%	542	513	90.0%	542	86.1%
Menarche age	<=13	60.7%	349	346	60.7%	349	60.3%
	>13	39.3%	240	224	39.3%	240	39.7%
First child	<30	34.0%	170	194	34.0%	170	22.4%
	>=30	66.0%	419	376	66.0%	419	77.6%

There was a significant relationship (p-value <=0.05) between the features of multiple risk factors and diagnosis of breast cancer such as; density, family members had cancer, Occupational hazard, Type of locality, Smoking status, had used IUD hormonal, used Hormonal injection every 3months, first birth. However, there are no relation between the Complain of any chronic diseases, had used, Birth control Implant, had used Oral contraceptives, Previous Chemotherapy, previous Cyst aspiration, previous Hormonal therapy, Implants, BMI, Hypertension the p-values for these variables are more than 0.05 shown in table 9.

Table 9: P-value for BC risk factors

No	Variable	P Value
1	density	0.0000
2	family members had cancer	0.0000
3	Occupational hazard	0.031842

4	Type of locality	0.032
5	Smoking status	00.037378
6	Complain of any chronic diseases	0.194
7	Had used IUD hormonal	.000
8	Had used Birth control Implant	0.1
9	usedHormonalinjectionevery3months	0.32
10	Had used Oral contraceptives	0.883834
11	mass	0.0000
12	Previous Alcohol consumption	0.000
13	Previous Chemotherapy	0.835554
14	Previous Chest CT scan	0.01
15	Previous Chest Xray	0.000
16	Previous Cyst aspiration	0.341
17	Previous Heart	0.000
18	Previous Hormonal therapy	0.819
19	Previous Hypertension	0.328
20	Implants	0.291006
22	Previous Lumpectomy	0.121
23	Previous Lumpectomy Remove the small mass only	0.0000
24	Previous Mammogram	0.014
25	Age group	0.053070
26	BMI group	0.13
27	Menarche age	0.891
28	First child	0.0000

4.3 Experimental Results

The study used Microsoft Azure ML. Azure Machine Learning can be defined as a cloud service that helps you manage and accelerate machine learning projects. It may be used by machine learning specialists, data scientists, and engineers in their daily workflows:

Models are trained and deployed, and ML operations are managed [51]. This section discusses the three models developed in Azure ML tools.

4.3.1 Machine Learning Algorithms on Three Classes Model

The dependent variable (diagnosis of BC) contains three values; malignant, benign, and healthy. After applying multiple machine learning algorithms, the results for the three-class model were as follows:

4.3.1.1 Random Forest

After applying the 5-folds cross-validation of Random Forest model for three-classes (malignant, benign, and healthy) The Accuracy was 95.8 %. the performance metrics Precision, Recall and F1 score was 0.93.9%, 93.7%, 93.8% respectively shown in Table 10.

Table 10: Evaluation of Random Forest Three classes

Model	Accuracy	Precision	Recall	F1 Score	AUC
Random Forest	0.958	0.938956	0.93757	0.938	0.998

In the matrix below the three class models that used random forest it is shown that, the true positive scored 94.9%, while the false negative scored is 5.6%. However, false positive 11%. So, the Sensitivity was 94.6% and specificity was 94.5% shown in Figure 36, which includes the confusion matrix produced after applying the Random Forest algorithms on three classes, where o is healthy ,1 is benign, and 2 malignant.

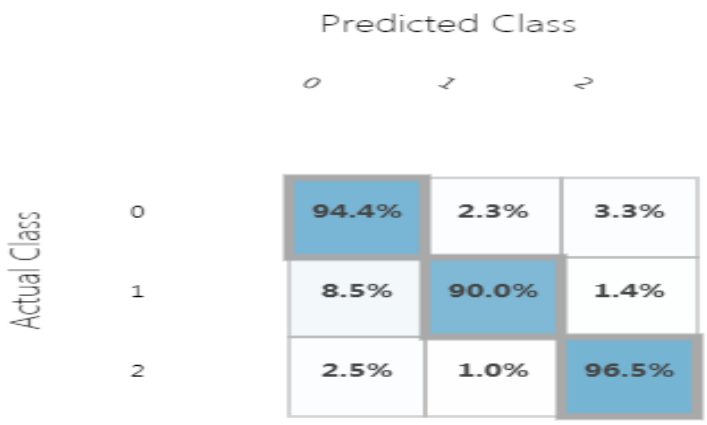


Figure 4.16: confusion matrix of Random Forest three classes

4.3.1.2 Neural Network Multiclass Classifier

Neural Network Multiclass Classifier was used on a three-class dataset for this thesis. Two tests were used in this model, with 50 and 100 hidden nodes, respectively. Table 4.10 shows that a Neural Network with 100 hidden nodes has a good testing accuracy, hence 100 hidden nodes were chosen for comparison with other ML models. Furthermore, the F1 score for 100 hidden nodes was 0.951, which is a reasonable result. The Precision, Recall, and F1 Score calculated in the Confusion Matrix Figure 4.17 were 0.951793, 0.951277, 0.951535 and, 0.923, respectively shown in table 11.

Table 11: Neural Network Multiclass-Classifier Evaluation Results

Algorithm	# Nueones	Accuracy	Precision	Recall	F1 Score
neural network	50	0.960609	0.942014	0.941848	0.94193
	100	0.967298	0.951793	0.951277	0.951535

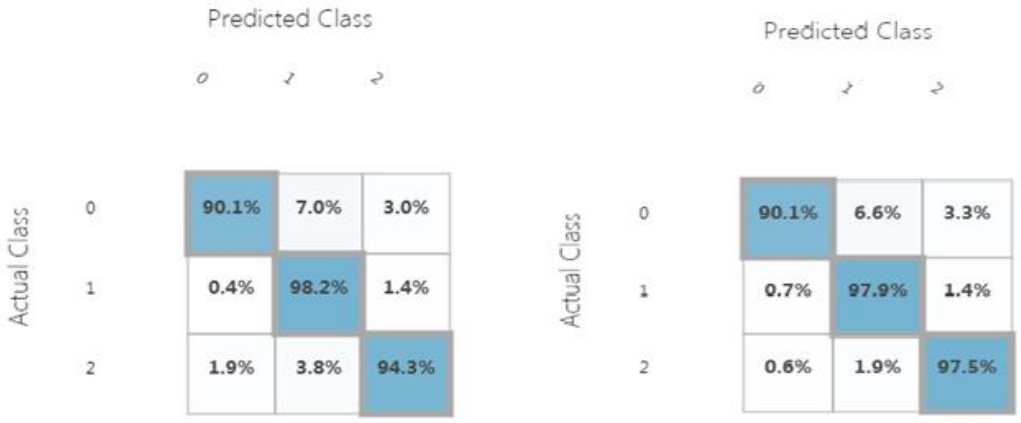


Figure 4.17: Confusion matrix for NN, where hidden node =50 ,100 respectively.

4.3.2 Machine Learning Algorithms on Two Classes (Malignant and Control Cases (Healthy))

In this section, machine learning is applied to two classes: one malignant and the other control (healthy cases). The best algorithms were as follows after applying Machin learning algorithms:

4.3.2.1 Random Forest for Healthy-Malignant

On a malignant and healthy dataset created specifically for this study, the DF algorithm was used. DF algorithms require changing a number of parameters when using the model. The number of decision trees is 8, while the maximum depth of the decision trees when applied to the model was 32. The Resampling Method was Bagging and Random Split Count was 128. The result of DF with Appling the setting on DF algorithms. Accuracy, Recall, F1 Score and AUC were 0.942**,** 0.942, 0.971, 0.917, 0.943, and 0.988 respectively. table 12 shows the result of DF.

Table 12: Result of DF tool.

algorithm	Accuracy	Precision	Recall	F1 Score	AUC
Random Forest	0.942	0.971	0.917	0.943	0.988

In the confusion matrix, there are 408 true positives and 387 true negatives shown in figure 4.18. Further, the true positive number was 408, while the false-positive number was 12. However, the true negative percentage scored 387 and the false-negative scored 37, as shown in figure 4.19. Thus, the sensitivity for DF is 1 and the specificity is 0.96.

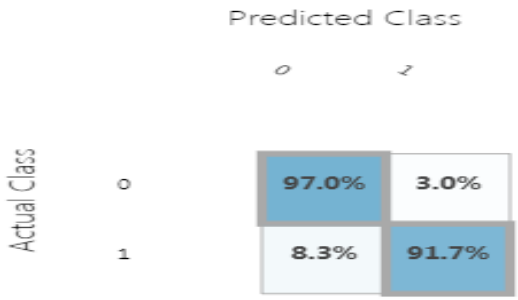


Figure 4.18: Confusion matrix for Random Forest (malignant and healthy)

		Predicted Class	
Actual class		True Positive **408**	False Negative **37**
		False Positive **12**	True Negative **387**

Figure 4.19: Confusion matrix for Random Forest (malignant and healthy)

4.3.2.2 Two Class Neural Network

Two class neural network was used on a two-class dataset for this thesis. 100 Hidden nodes were used in this model. Table 11 shows that a Neural Network with 100 hidden nodes with high accuracy can be achieved, hence 100 hidden nodes were chosen for comparison with other ML models. Furthermore, the F1 score for 100 hidden nodes was 0.986, which is a reasonable result. The accuracy, Precision, Recall, and F1 Scores calculated in the Confusion Matrix shown Figure 4.20 were 0.985, 0.978, 0.993, 0.993, and 0.986, respectively shown in Table 11. While the AUC was calculated sensitively and the

sensitivity was 0.999 shown in figure 4.21 AUC chart. Table 13: Result of NN two class (malignant and healthy)

Table 13: Result of NN two class (malignant and healthy)

algorithm	Accuracy	Precision	Recall	F1 Score	AUC
neural network	0.985	0.978	0.993	0.986	0.999

	Predicted Class	
Actual class	True Positive **442**	False Negative **3**
	False Positive **10**	True Negative **389**

Figure 4.20: Confusion matrix for neural network (malignant and healthy)

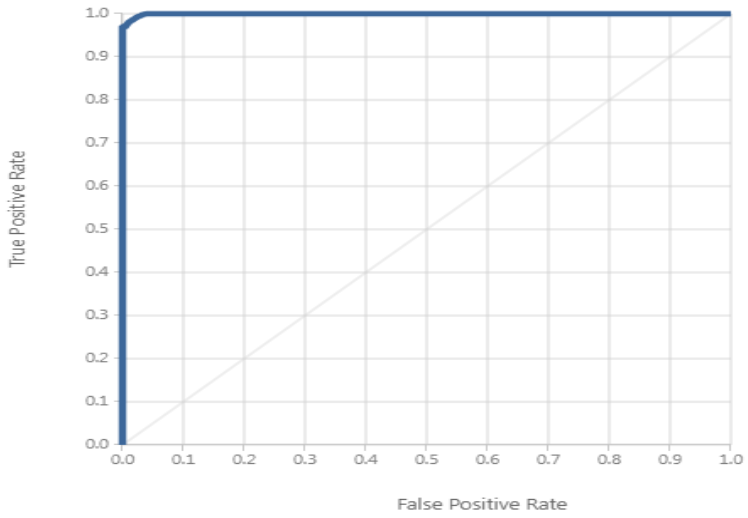


Figure 4.21: AUC of neural network (malignant and healthy)

4.3.2.3 Boosted Decision Tree

The BDT used the same dataset of (malignant and healthy cases). When applying the boosted decision to the model, used multiple parameters: the maximum number of leaves per tree and the minimum number of samples per leaf node and, when applied to the model,

were 32, and 20 respectively. The accuracy, Precision, Recall, and F1 Score calculated in the Confusion Matrix Figure 4.22 were 0.975, 0.980, 0.973, and 0.976, respectively shown in table 14. While the AUC was calculated from sensitivity and sensitivity was 0.997 shown in figure 4.23 AUC chart.

Table 14: Result of Boosted Decision two class (malignant and healthy)

algorithm	Accuracy	Precision	Recall	F1 Score	AUC
Boosted Decision	0.975	0.980	0.973	0.976	0.997

		Predicted Class
Actual class	True Positive **433**	False Negative **12**
	False Positive **9**	True Negative **390**

Figure 4.22: Confusion matrix for Boosted Decision (malignant and healthy)

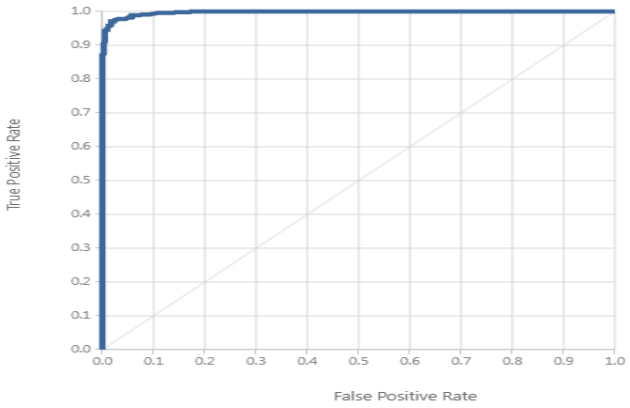


Figure 4.23: AUC chart (malignant and healthy)

4.3.2.4 Support Vector Machine Classifier

Support Vector Machine Classifier used the same dataset of (malignant and healthy cases). When applying the support vector machine classifier to the model. The Accuracy, Precision, Recall, and F1 Score calculated in the Confusion Matrix Figure 4.24 were 0.959,

0.968, 0.953, and 0.960, respectively shown in table 15. While the AUC calculated from sensitively and sensitivity was 0.997 shown in figure 4.25 AUC chart.

Table 15: Result of Support Vector Machine Classifier two class (malignant and healthy)

algorithm	Accuracy	Precision	Recall	F1 Score	AUC
Deep Support Vector Machine	0.959	0.968	0.953	0.960	0.987

	Predicted Class	
Actual class	True Positive **424**	False Negative **21**
	False Positive **14**	True Negative **385**

Figure 4.24: Support Vector Machine Classifier

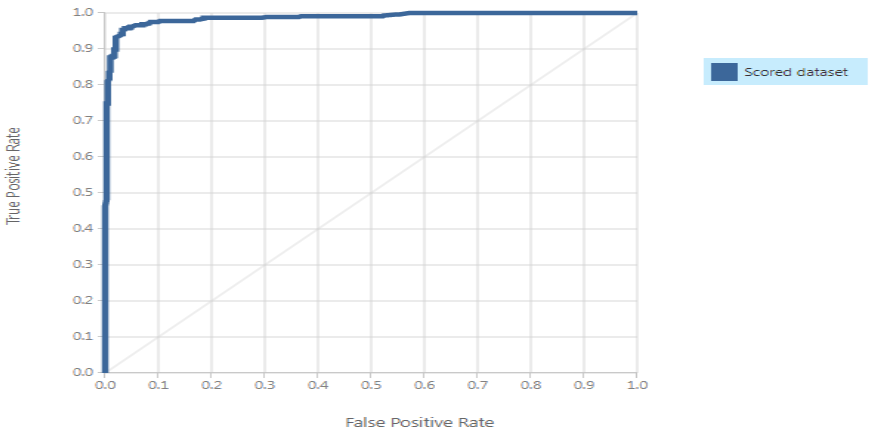


Figure 4.25: AUC chart (malignant and healthy)

4.3.3 Machine Learning Algorithms on Two Class (Malignant and Benign Class)

In this section, machine learning is applied to two classes: malignant and benign cases. The best algorithms were as follows after applying machine learning algorithms:

4.3.3.1 Neural Network

Two class neural network was used on a two-class dataset for this thesis. 100 Hidden nodes were used in this model. Table 14 shows that a Neural Network with 100 hidden nodes with high accuracy can be achieved, hence 100 hidden nodes were chosen for comparison with other ML models. Furthermore, the F1 score for 100 hidden nodes was 0.986, which is a reasonable result. The Accuracy, Precision, Recall, and F1 Score calculated in the Confusion Matrix Figure 4.26 were 0.095 0.928, 0.980, 0.950, and 0.953 respectively shown in table 16. While the AUC calculated using sensitivity and specificity was 0.995 shown in figure 4.27 AUC chart.

Table 16: the result of Neural Network for malignant and benign class

algorithm	Accuracy	Precision	Recall	F1 Score	AUC
Neural Network	0.950	0.928	0.980	0.953	0.995

Actual class	**Predicted Class**	
	True Positive **436**	False Negative **9**
	False Positive **34**	True Negative **378**

Figure 4.26: confusion matrix of Neural Network for malignant and benign class

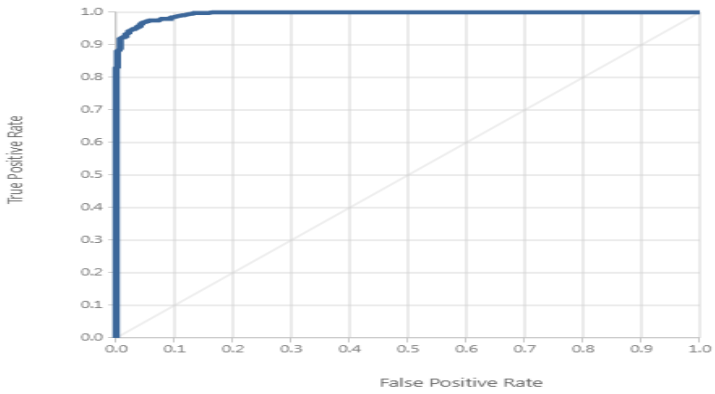


Figure 4.27: AUC OF Neural Network for malignant and benign class

4.3.3.2 Random Forest

On a malignant and benign dataset created specifically for this study, the DF algorithm was used. DF algorithms require changing several parameters when using the model. The number of decision trees is 8, while the maximum depth of the decision trees, when applied to the model, was 32. The Resampling Method was Bagging and Random Split Count was 128. The result of DF with Appling the setting on DF algorithms. Accuracy, Precision, Recall, F1 Score and AUC were 0.949, 0.965, 0.935, 0.950 and 0.988 respectively. show table 17 the result of DF show table 15 the result of the ML tool and show the figure 4.29 confusion matrix for this tool.

Table 17: the result of Random Forest for malignant and benign class

algorithm	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**AUC**
Random Forest	0.949	0.965	0.935	0.950	0.988

Actual class	**Predicted Class**	
	True Positive **424**	False Negative **21**
	False Positive **14**	True Negative **385**

Figure 4.28: confusion matrix of Random Forest for malignant and benign class

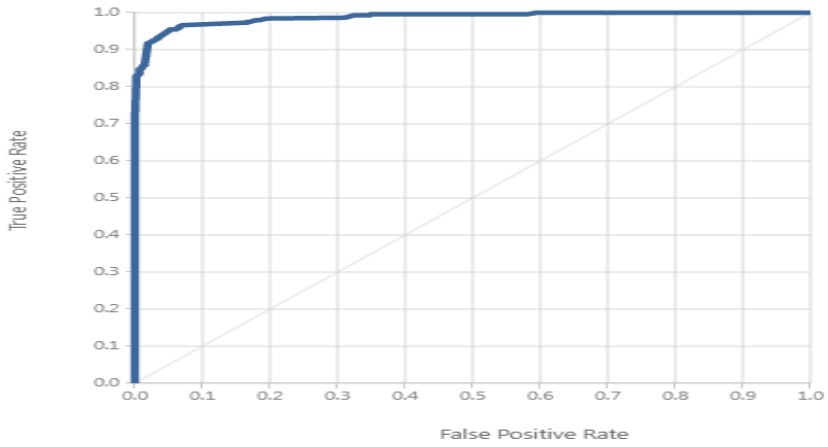


Figure 4.29: AUC for malignant and benign class

4.3.4 Machine Learning Algorithms on Two Class (Healthy and Benign Class)

In this section, machine learning is applied to two classes: healthy and benign cases. The best algorithms were as follows after applying machine learning algorithms.

4.3.4.1 Random Forest

On a healthy and benign dataset created specifically for this study, the DF algorithm was used. DF algorithms require changing several parameters when using the model. The number of decision trees is 8, while the maximum depth of the decision trees, when applied to the model, was 32. The Resampling Method was Bagging and Random Split Count was 128. The result of DF with Appling the setting on DF algorithms. Accuracy, Precision, Recall, F1 Score and AUC were 0.894, 0.909, 0.939, 0.924 and 0.961 respectively. Table 16 shows the result of DF show table 18 the result of the ML tool and show the figure 4.30 confusion matrix for this tool. Also, show Figure 4.31 the result of AUC. As mentioned previously, benign refers to a tumor or noncancerous development. It does not spread to other sections of the body in this way. It does not spread to neighboring tissues. As a result, having this model with such performance is good.

Table 18: the result of Random Forest for healthy and benign class

algorithm	Accuracy	Precision	Recall	F1 Score	AUC
Random Forest	0.894	0.909	0.939	0.924	0.961

Actual class	**Predicted Class**	
	True Positive **805**	False Negative **52**
	False Positive **81**	True Negative **318**

Figure 4.30: confusion matrix of Random Forest for healthy and benign class

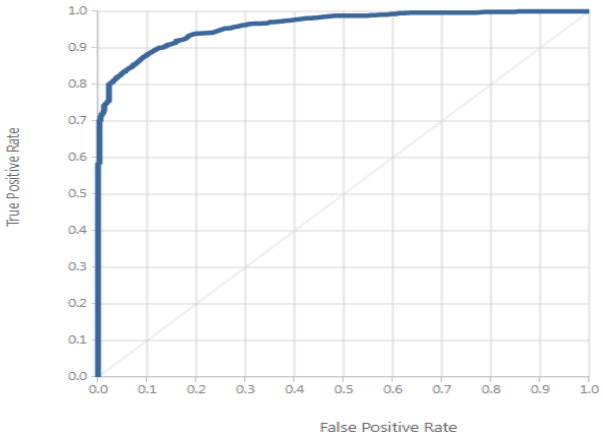


Figure 4.31: AUC of for healthy and benign class

4.3.4.2 Neural Network

Two class neural network was used on a two-class dataset for this thesis. 100 Hidden nodes were used in this model. Table 19 shows that a Neural Network with 100 hidden nodes with high accuracy can be achieved, hence 100 hidden nodes were chosen for comparison with other ML models. Furthermore, the F1 score for 100 hidden nodes was 0.986, which is a reasonable result. The Accuracy, Precision, Recall, and F1 Score calculated in the Confusion Matrix Figure 4.32 were 0.877, 0.86, 0.968, 0.950, and 0.915 respectively

shown in table 14. While the AUC calculated sensitively and specify was 0.961. As previously stated, benign refers to a tumor or noncancerous development. It does not spread to other sections of the body in this way. It does not spread to neighboring tissues. As a result, having this model with such performance is good.

Table 19: the result of Neural Network for healthy and benign class

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**AUC**
Neural Network	0.877	0.867	0.968	0.915	0.961

Actual class	**Predicted Class**	
	True Positive **830**	False Negative **27**
	False Positive **127**	True Negative **272**

Figure 4.32: confusion matrix of NN for healthy and benign class

4.4 Experimental Result Summary

After completing the deployment of all machine learning algorithms for breast cancer detection using the Palestinian dataset, the results may be compared from the table using the performance metrics.

4.4.1 Three Classes

The neural network and Random Forest models outperform other algorithms that used in three classes. From Table 20, it is clear that the neural network has the best performance

result in all metrics than other algorithms for this data set with 96.7% accuracy, 95.2% Precision, 95.1%Recall, 95.2% F1 Score, and 97% AUC.

Table 20: Comparison of Scores of Various Models

Algorithms		Accuracy	Precision	Recall	F1 Score	AUC
Random Forest		0.958	0.93896	0.93757	0.938	0.968
Neural network		0.9673	0.95179	0.95128	0.95154	0.97

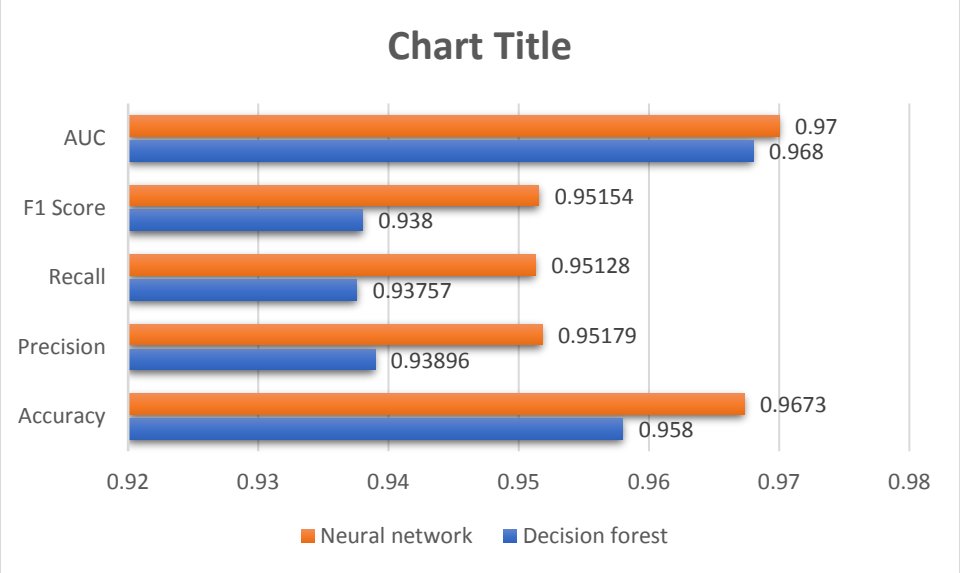


Figure 4.33: Comparison between machine learning for malignant, benign and healthy models.

4.4.2 Two Classes Malignant and Control Cases (Healthy)

The Boosted Decision, Support Vector Machine, neural network, and Random Forest models outperform other algorithms in two classes malignant and control cases (healthy). From Table 21, it is clear that the neural network has the best performance result in all metrics than other models for this data set with 98.5% accuracy, 97.8% Precision, and 99.3% Recall which means very low false negative, 98.6% F1 Score and 99.9% AUC. Although the neural network has the best performance model, also the other models

achieved high performance. The boosted decision has a perfect Precision score (98.0%) and perfect AUC (99.7%).

Table 21: Comparison of results of two classes malignant and control cases (healthy)

Algorithms	Accuracy	Precision	Recall	F1 Score	AUC
Random Forest	0.942	0.971	0.917	0.943	0.988
neural network	0.985	0.978	0.993	0.986	0.999
Boosted Decision	0.975	0.980	0.973	0.976	0.997
Deep Support Vector Machine	0.959	0.968	0.953	0.960	0.987

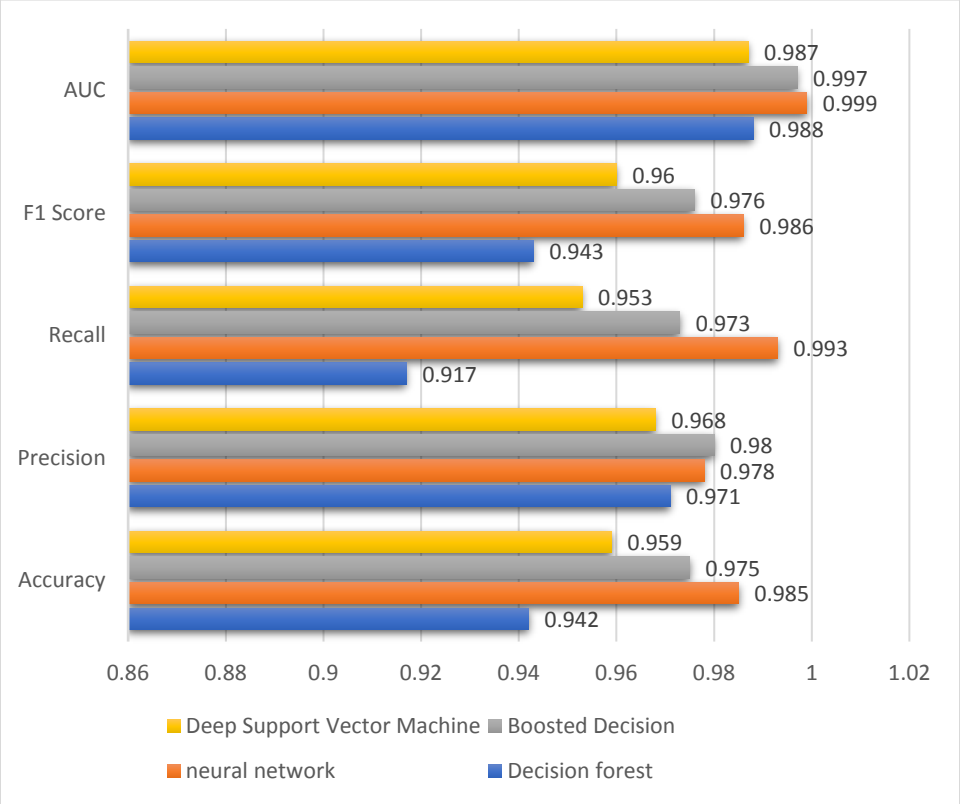


Figure 4.34: Comparison of results of two classes malignant and control cases (healthy)

4.4.3 Two Classes Malignant and Benign

The Random Forest and neural network models outperform other algorithms in two classes malignant and benign. From Table 22, it is clear that the neural network and decision predictions have convergent performance on all scales. Random Forest has a perfect

Precision score (96.5%) and a high of 94.9% Accuracy, which are more important in terms of cancer detection. While the Neural network has high-performance metrics of 95.0% Accuracy, 98.0% Recall, and 99.5 % AUC.

Table 22: Comparison of results of two classes malignant and benign

Algorithms	Accuracy	Precision	Recall	F1 Score	AUC
Neural Network	0.950	0.928	0.980	0.953	0.995
Random Forest	0.949	0.965	0.935	0.950	0.988

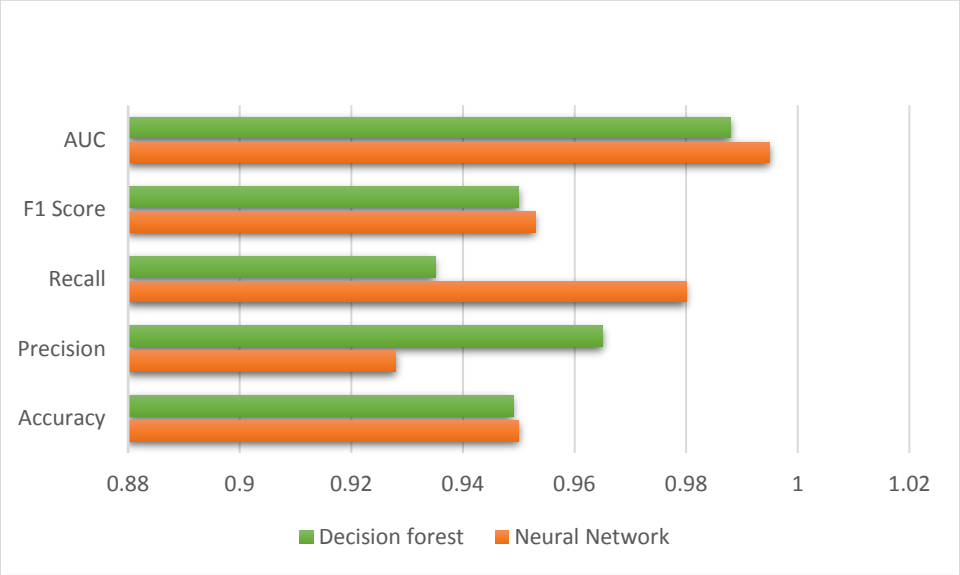


Figure 4.35: Comparison of results of two classes malignant and benign

4.4.4 Two Classes (Healthy and Benign)

The neural network and Random Forest models outperform other algorithms in two classes healthy and benign. From Table 23, it is clear that the Random Forest has the best performance result in all metrics of other models for this data set with 89.4% accuracy, 90.9% Precision, and 93.9% Recall which means very low false negative, 92.4% F1 Score and 96.1% AUC. Also, it is clear that the neural network has a perfect 96.1% AUC, and recall score (96.8%) which means a very low false negative shown Figure 4.36.

Table 23: Comparison of results of two classes healthy and benign class

Algorithms	Accuracy	Precision	Recall	F1 Score	AUC
Random Forest	0.894	0.909	0.939	0.924	0.961
neural network	0.877	0.867	0.968	0.915	0.961

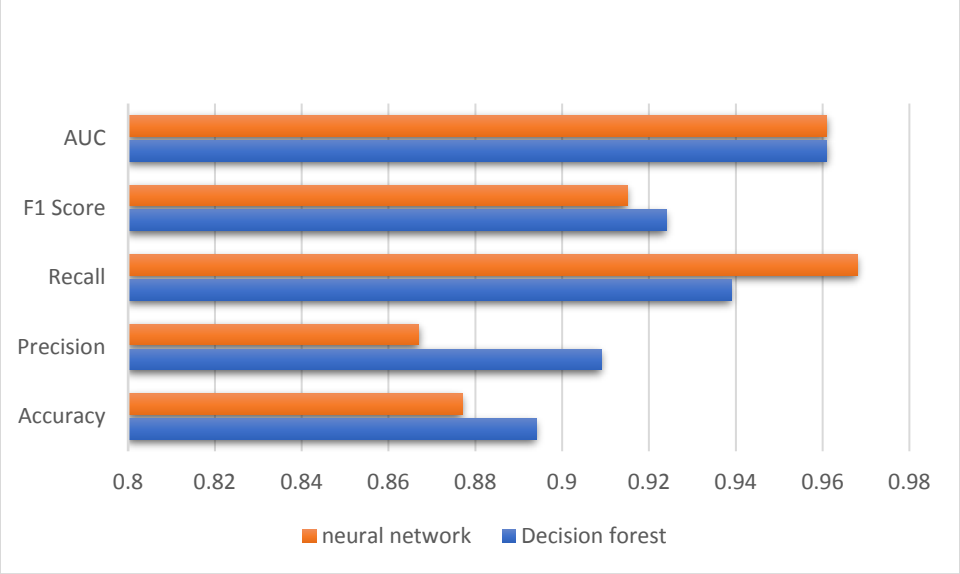


Figure 4.36: Comparison of results of two classes malignant and benign

4.5 Result Summary

It is clear from the results that data mining techniques can be used to predict breast cancer risk, and that the neural network, Random Forest, and Boosted Decision have higher accuracy than the other models in the three different cancer applications shown in the result in the previous section. The deep support vector machine, which is among the top three algorithms in two categories of malignant and control (healthy) conditions, also achieved high accuracy results. The prediction model for malignant and healthy breast cancer is the best performance model. While the performance of healthy-benign model is the lowest

performance of other breast cancer prediction models. This is because benign (changes) fibrous tissue are thought to be natural, so generally no treatment is needed unless they are causing discomfort [52]. Figure 4.37 shows the top performance results for three models based on the type of class used.

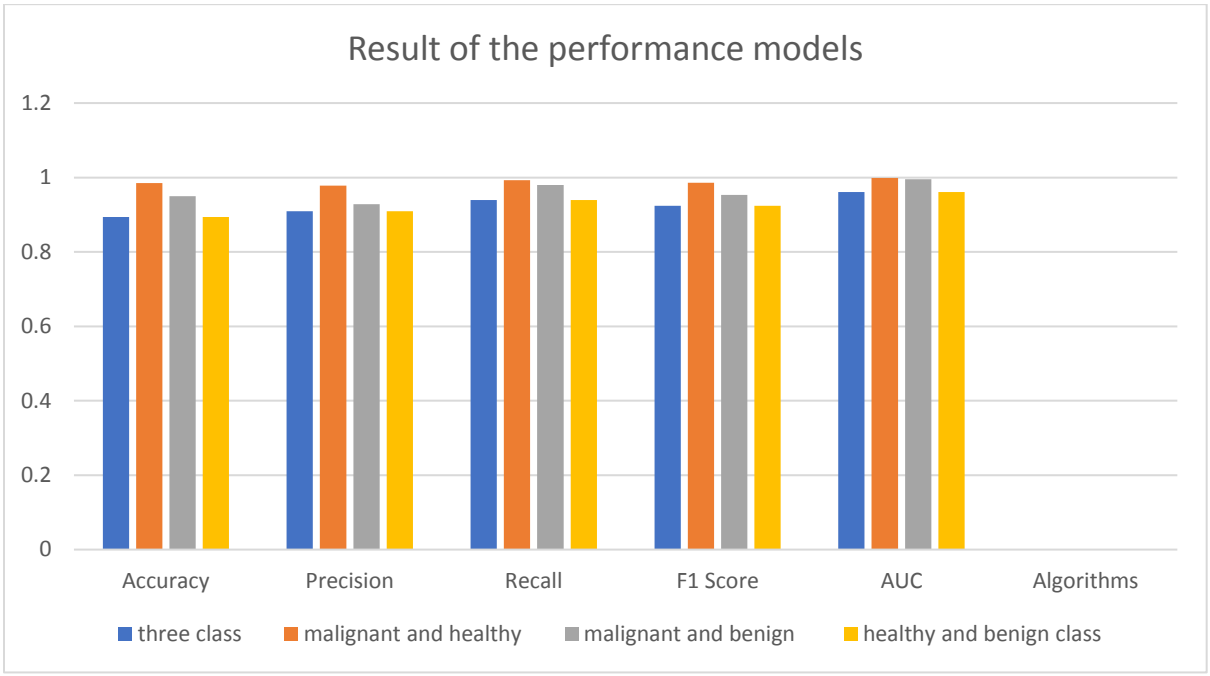


Figure 4.37: shows the results of the top three models based on the type of class used.

Chapter Five: Conclusion and Recommendations

5.1 Introduction

In this chapter, the study conclusion and recommendations are presented. Future work and planned studies for future work, as well as the study strengths are presented.

5.2 Conclusion and Recommendation

The fact that breast cancer, the most common cancer among women, is responsible for 69% of cancer-related deaths (WHO). However, the early detection of BC can increase the probability of survival for a large number of BC patients by early treatments on time. This thesis compared the effectiveness of various machine learning algorithms in predicting breast cancer depending on a variety of risk factors. The primary goal of this research is to find a model that can accurately detect BC in patients. The research is based on the Palestinian dataset. Then, four machine learning models were applied, and the performance metrics score was used to assess the performance of the models. The result shows that the best four models were applied for predicting breast cancer.

The First model (NN model) for detection of BC for three classes (malignant, benign, and healthy) with 96.7% accuracy, 95.2% Precision, 95.1%Recall, 95.2% F1 Score, and 97% AUC. While the performance of the second model (NN model) for prediction of the BC (Discrimination between malignant and healthy cases) was 98.5% accuracy, 97.8% Precision, 99.3% Recall which means very low false negative, 98.6% F1 Score, and 99.9% AUC. The third model (was developed to detect that BC (Malignant and benign) has performance with 95.0% Accuracy, 98.0% Recall, and 99.5 % AUC. Finally, the Random Forest model of prediction of BC (healthy and benign class) was developed with 89.4% accuracy, 90.9% Precision, 93.9% Recall which means very low false negative, 92.4% F1

Score, and 96.1% AUC. The fact that the models achieved high accuracy on data for Palestinian women leads us to use a system that has the ability to predict breast cancer with an accuracy close to a doctor's diagnosis. So, it can be said that the models can be used for diagnosis and to assist doctors in determining whether or not a large number of women require early detection via mammography or others. This allows women who are suspected of having breast cancer to be given priority for appointments at breast cancer clinics.

A large number of variables relating to patient risk factors were missing, so these features were removed, unfortunately, they may be important in BC prediction. This is necessary to attention to the quality of data through health directorates. also, recommend developing the system to force the introduction of important features by reviewing them. In addition, must publish BC prediction application for all Palestinian women to use and predict breast cancer. One of the great difficulties that arise in this kind of study is the poor quality of the data that many times, after the cleaning process, causes certain variables to be discarded due to the impossibility of obtaining their values , and that somehow influences the results of the data mining.

5.3 Future Works

With increased data availability and, most importantly, data growth, the model can be perfected in the future. Deep learning models perform well in proportion to the amount of data available, indicating that there is room for improvement with the availability of a large dataset. This research investigates the field of machine learning to incorporate it into the medical profession as a means of early breast cancer identification, which could eventually lead to a complete clinical system that saves lives. In the future will be developed the system data mining to predict cancer by images from a mammogram. It is necessary to

repeat the analysis using more data, preferably integrating them from a computerized medical record system. Finally publish web application prediction system for Palestinian women on internet.

5.4 Strengths of the Study

This is the first study in Palestine that uses machine learning techniques to predict breast cancer. The use of a large balanced dataset for model training and repeated k-fold cross-validation for model evaluation are both strengths of our research. Thirty features risk factors for breast cancer were used, and this number led us to discover new risk factors that cause this disease. Real data from the Palestinian dataset were used in BC prediction. Which enhances the ability to extract new features for Palestinian women with risk factors.

References

[1] F. Z. Francies, R. Hull, R. Khanyile, and Z. Dlamini, "Breast cancer in low-middle income countries: abnormality in splicing and lack of targeted treatment options.," *Am. J. Cancer Res.*, vol. 10, no. 5, pp. 1568–1591, 2020.

[2] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA. Cancer J. Clin.*, vol. 71, no. 3, pp. 209–249, 2021, doi: 10.3322/caac.21660.

[3] Ministry of Health, "2020 التقرير الصحي السنوي.Pdf." 2020.

[4] W. Y. Joko-Fru *et al.*, "Breast cancer survival in sub-Saharan Africa by age, stage at diagnosis and human development index: A population-based registry study," *Int. J. Cancer*, vol. 146, no. 5, pp. 1208–1218, 2020, doi: 10.1002/ijc.32406.

[5] A. J. Abdelwahab Yousef, "Male Breast Cancer: Epidemiology and Risk Factors," *Semin. Oncol.*, vol. 44, no. 4, pp. 267–272, 2017, doi: 10.1053/j.seminoncol.2017.11.002.

[6] Y. S. Sun *et al.*, "Risk factors and preventions of breast cancer," *Int. J. Biol. Sci.*, vol. 13, no. 11, pp. 1387–1397, 2017, doi: 10.7150/ijbs.21635.

[7] D. M. Abdel-Salam *et al.*, "Perceived barriers and awareness of mammography screening among Saudi women attending primary health centers," *Risk Manag. Healthc. Policy*, vol. 13, pp. 2553–2561, 2020, doi: 10.2147/RMHP.S277375.

[8] S. . Deshpande and V. . Thakare, "Data Mining System and Applications: A Review," *Int. J. Distrib. Parallel Syst.*, vol. 1, no. 1, pp. 32–44, 2010, doi: 10.5121/ijdps.2010.1103.

[9] *No Title*. .

[10] S. Winters, C. Martin, D. Murphy, and N. K. Shokar, *Breast Cancer Epidemiology, Prevention, and Screening*, vol. 151. Elsevier Inc., 2017.

[11] O. Ginsburg *et al.*, "Breast cancer early detection: A phased approach to implementation," *Cancer*, vol. 126, pp. 2379–2393, 2020, doi: 10.1002/cncr.32887.

[12] M. Huang, C. Chen, W. Lin, S. Ke, and C. Tsai, "SVM and SVM Ensembles in Breast Cancer Prediction," pp. 1–14, 2017, doi: 10.1371/journal.pone.0161501.

[13] R. S. Shukla and Y. Aggarwal, "Fourier Transform and Autoregressive HRV

Features in Prediction and Classification of Breast Cancer," *IETE J. Res.*, 2021, doi: 10.1080/03772063.2021.1878063.

[14] F. Labrèche, M. S. Goldberg, D. Hashim, and E. Weiderpass, "Breast cancer," *Occup. Cancers*, pp. 417–438, 2020, doi: 10.1007/978-3-030-30766-0_24.

[15] P. J. Bazira, H. Ellis, and V. Mahadevan, "Anatomy and physiology of the breast," *Surgery*, vol. 40, no. 2, pp. 79–83, 2021, doi: 10.1016/j.mpsur.2021.11.015.

[16] G. Bistoni and J. Farhadi, "Anatomy and Physiology of the Breast," *Plast. Reconstr. Surg. Approaches Tech.*, no. March 2015, pp. 477–485, 2015, doi: 10.1002/9781118655412.ch37.

[17] N. Harbeck *et al.*, *Breast cancer*, vol. 5, no. 1. 2019.

[18] A. L. W. Meisner, M. Houman Fekrazad, and M. E. Royce, "Breast Disease: Benign and Malignant," *Med. Clin. North Am.*, vol. 92, no. 5, pp. 1115–1141, 2008, doi: 10.1016/j.mcna.2008.04.003.

[19] American Cancer Society, "Breast Cancer What is breast cancer ?," *Am. Cancer Soc. Cancer Facts Fig. Atlanta, Ga Am. Cancer Soc.*, pp. 1–19, 2022.

[20] J. Adv, "March 2010 Revised: 19," vol. 1, no. 2, pp. 109–126, 2010.

[21] M. M. Eberl, C. H. Fox, S. B. Edge, C. A. Carter, and M. C. Mahoney, "BI-RADS classification for management of abnormal mammograms," *J. Am. Board Fam. Med.*, vol. 19, no. 2, pp. 161–164, 2006, doi: 10.3122/jabfm.19.2.161.

[22] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput. Sci.*, vol. 2, no. 3, pp. 1–21, 2021, doi: 10.1007/s42979-021-00592-x.

[23] R. Konieczny and R. Idczak, "Mössbauer study of Fe-Re alloys prepared by mechanical alloying," *Hyperfine Interact.*, vol. 237, no. 1, pp. 1–8, 2016, doi: 10.1007/s10751-016-1232-6.

[24] A. A. Alfayez, H. Kunz, and A. G. Lai, "Predicting the risk of cancer in adults using supervised machine learning: A scoping review," *BMJ Open*, vol. 11, no. 9, pp. 1–11, 2021, doi: 10.1136/bmjopen-2020-047755.

[25] C. Hou *et al.*, "Predicting breast cancer in chinese women using machine learning techniques: Algorithm development," *JMIR Med. Informatics*, vol. 8, no. 6, 2020, doi: 10.2196/17364.

[26] K. Williams, P. Adebayo Idowu, J. Ademola Balogun, and A. Ishola Oluwaranti, “Breast Cancer Risk Prediction Using Data Mining Classification Techniques,” *Trans. Networks Commun.*, vol. 3, no. 2, 2015, doi: 10.14738/tnc.32.662.

[27] P. Ramachandran, N. Girija, and T. Bhuvaneswari, “Early Detection and Prevention of Cancer using Data Mining Techniques,” *Int. J. Comput. Appl.*, vol. 97, no. 13, pp. 48–53, 2014, doi: 10.5120/17069-7492.

[28] S. S. Shajahaan, S. Shanthi, and V. Manochitra, “Application of Data Mining Techniques to Model Breast Cancer Data,” *Int. J. Emerg. Technol. Adv. Eng.*, vol. 3, no. 11, pp. 1–8, 2013.

[29] C. Ming, V. Viassolo, N. Probst-Hensch, P. O. Chappuis, I. D. Dinov, and M. C. Katapodi, “Machine learning techniques for personalized breast cancer risk prediction: Comparison with the BCRAT and BOADICEA models,” *Breast Cancer Res.*, vol. 21, no. 1, pp. 1–11, 2019, doi: 10.1186/s13058-019-1158-4.

[30] Z. Guan *et al.*, “Combining Breast Cancer Risk Prediction Models,” pp. 1–31, 2020.

[31] V. Punithavathi and D. Devakumari, “A Framework on Classification of Mammogram Images for Breast Cancer Detection using Image Processing with Data Mining Techniques,” vol. 8, no. 2, pp. 640–644, 2020.

[32] A. Alhazmi *et al.*, “Application of artificial intelligence and machine learning for prediction of oral cancer risk,” *J. Oral Pathol. Med.*, vol. 50, no. 5, pp. 444–450, 2021, doi: 10.1111/jop.13157.

[33] E. Rahm and H. H. Do, “Data Cleaning: Current Approaches and Issues,” *IEEE Data Eng. Bull*, vol. 23, no. 4, pp. 3–13, 2015.

[34] J. A. López del Val and J. P. Alonso Pérez de Agreda, “Principal components analysis,” *Aten. Primaria*, vol. 12, no. 6, pp. 333–338, 1993, doi: 10.5455/ijlr.20170415115235.

[35] S. Narasimhan and S. L. Shah, “Model identification and error covariance matrix estimation from noisy data using PCA,” *IFAC Proc. Vol.*, vol. 37, no. 1, pp. 511–516, 2004, doi: 10.1016/s1474-6670(17)38783-9.

[36] H. Liu and M. Cocea, “Semi-random partitioning of data into training and test sets in granular computing context,” *Granul. Comput.*, vol. 2, no. 4, pp. 357–386,

2017, doi: 10.1007/s41066-017-0049-2.

[37] C. Gershenson, “Artificial Neural Networks for Beginners,” no. September 2003, 2003.

[38] M. Galety, F. H. Al Mukthar, R. J. Maaroof, and F. Rofoo, “Deep Neural Network Concepts for Classification using Convolutional Neural Network: A Systematic Review and Evaluation,” *Tech. Rom. J. Appl. Sci. Technol.*, vol. 3, no. 8, pp. 58–70, 2021, doi: 10.47577/technium.v3i8.4554.

[39] M. Reza, S. Miri, and R. Javidan, “A Hybrid Data Mining Approach for Intrusion Detection on Imbalanced NSL-KDD Dataset,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 6, pp. 1–33, 2016, doi: 10.14569/ijacsa.2016.070603.

[40] H. Poirot, “Logistic Regression,” 2021.

[41] N. A. Bakshi, “Model Reference Adaptive Control of Quadrotor UAVs: A Neural Network Perspective,” *Adapt. Robust Control Syst.*, 2018, doi: 10.5772/intechopen.71487.

[42] Y. Tang, “Deep Learning using Linear Support Vector Machines,” no. Icml, 2013.

[43] X. Li and W. Yu, “Fast Support Vector Machine Classification for Large Data Sets,” *Int. J. Comput. Intell. Syst.*, vol. 7, no. 2, pp. 197–212, 2014, doi: 10.1080/18756891.2013.868148.

[44] Y. Coadou, “Boosted decision trees and applications,” *EPJ Web Conf.*, vol. 55, 2013, doi: 10.1051/epjconf/20135502004.

[45] W. Zhu, N. Zeng, and N. Wang, “Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS® implementations.,” *Northeast SAS Users Gr. 2010 Heal. Care Life Sci.*, pp. 1–9, 2010.

[46] H. Dalianis, “Evaluation Metrics and Evaluation,” *Clin. Text Min.*, no. 1967, pp. 45–53, 2018, doi: 10.1007/978-3-319-78503-5_6.

[47] V. Beral, D. Bull, R. Doll, R. Peto, and G. Reeves, “Breast cancer and breastfeeding: Collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50 302 women with breast cancer and 96 973 women without the disease,” *Lancet*, vol. 360, no. 9328, pp. 187–195, 2002, doi: 10.1016/S0140-6736(02)09454-0.

[48] J. S. Smilg, “Are you dense? The implications and imaging of the dense breast,”

South African J. Radiol., vol. 22, no. 2, pp. 1–4, 2018, doi: 10.4102/sajr.v22i2.1356.

[49] E. Amir, O. C. Freedman, B. Seruga, and D. G. Evans, "Assessing women at high risk of breast cancer: A review of risk assessment models," *J. Natl. Cancer Inst.*, vol. 102, no. 10, pp. 680–691, 2010, doi: 10.1093/jnci/djq088.

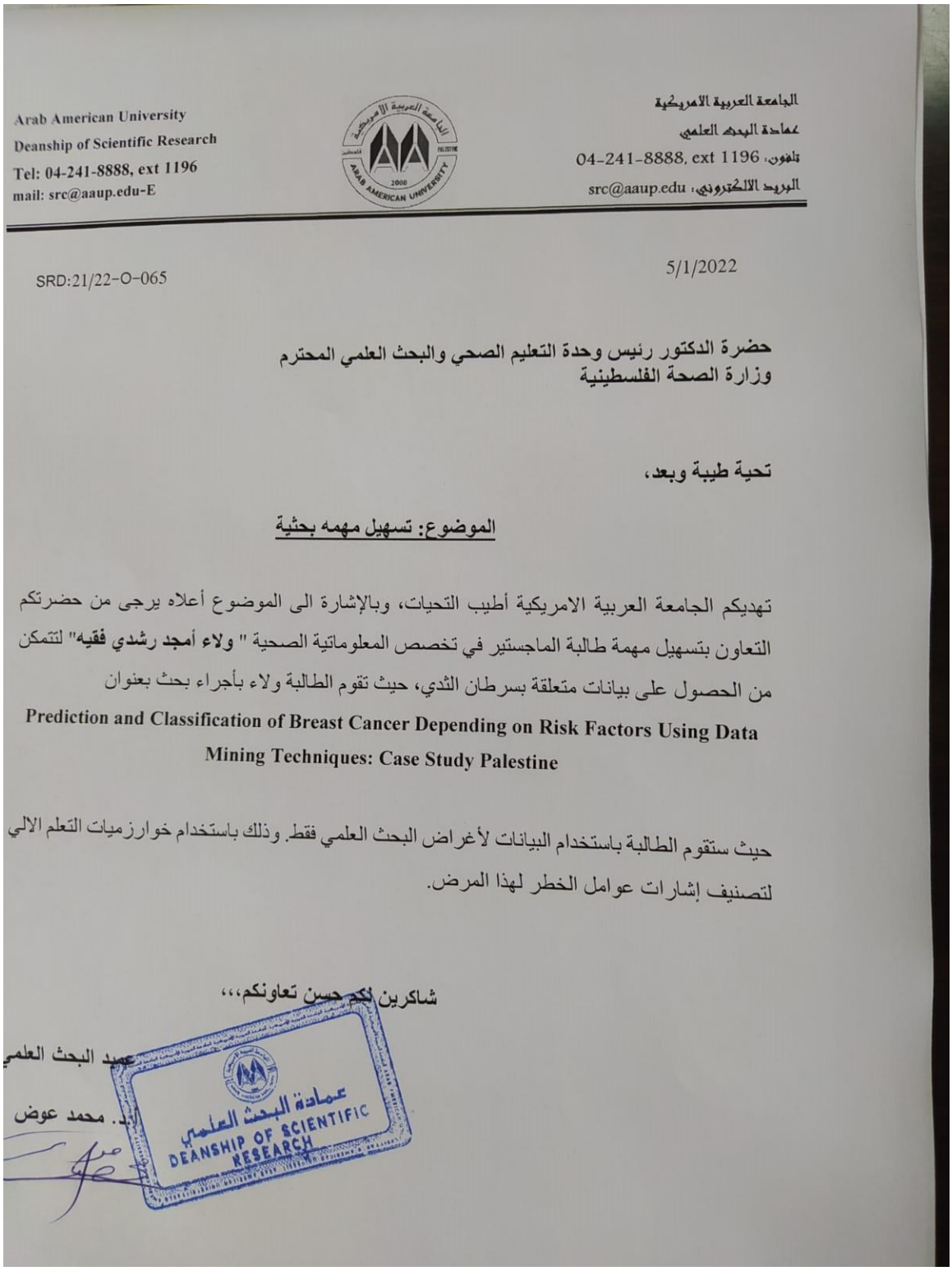
[50] C. Fenga, "Occupational exposure and risk of breast cancer (Review)," *Biomed. Reports*, vol. 4, no. 3, pp. 282–292, 2016, doi: 10.3892/br.2016.575.

[51] J. Barnes, *Azure Machine Learning Microsoft Azure Essentials*. 2015.

[52] American cancer society, "Non-cancerous Breast Conditions Fibrocystic Changes in the Breast," pp. 13–14.

Appendices

Appendix 1: Arab American University approval for Data collection.



Arab American University
Deanship of Scientific Research
Tel: 04-241-8888, ext 1196
mail: src@aaup.edu-E

الجامعة العربية الامريكية
عمادة البحث العلمي
تلفون: 04-241-8888, ext 1196
البريد الالكتروني: src@aaup.edu

SRD:21/22-O-065

5/1/2022

حضرة الدكتور رئيس وحدة التعليم الصحي والبحث العلمي المحترم
وزارة الصحة الفلسطينية

تحية طيبة وبعد،

الموضوع: تسهيل مهمه بحثية

تهديكم الجامعة العربية الامريكية أطيب التحيات، وبالإشارة الى الموضوع أعلاه يرجى من حضرتكم التعاون بتسهيل مهمة طالبة الماجستير في تخصص المعلوماتية الصحية " **ولاء أمجد رشدي فقيه**" لتتمكن من الحصول على بيانات متعلقة بسرطان الثدي، حيث تقوم الطالبة ولاء بأجراء بحث بعنوان

Prediction and Classification of Breast Cancer Depending on Risk Factors Using Data Mining Techniques: Case Study Palestine

حيث ستقوم الطالبة باستخدام البيانات لأغراض البحث العلمي فقط. وذلك باستخدام خوارزميات التعلم الالي لتصنيف إشارات عوامل الخطر لهذا المرض.

شاكرين لكم حسن تعاونكم،،،

عميد البحث العلمي

أ.د. محمد عوض

عمادة البحث العلمي
DEANSHIP OF SCIENTIFIC RESEARCH

Appendix 2: Ministry of health approval for Data collection

State of Palestine
Ministry of Health
Education in Health and Scientific Research Unit

دولة فلسطين
وزارة الصحة
وحدة التعليم الصحي والبحث العلمي

Ref.: ………………
Date:………………

الرقم: عتع/٢٣/٢٠٢٢
التاريخ: ١١/١١/٢٠٢٢

الأخ علـــي الحلـــو المحتـــرم
مديـر عـام الادارة العامة لتكنولوجيا المعلومات ،،،
تحية واحترام...

الموضـــوع: تسهيل مهمة بحث

يرجى التكـرم بتسـهيل مهمـة الطالبـة: **ولاء امجـد رشـدي فقيـه، ماجسـتير معلوماتيـة صحية– الجامعة العربية الامريكية**، لعمل بحث بعنوان:

"Prediction and Classification of Breast Cancer Depending on Risk Factors Using Data Mining Techniques: Case Study Palestine"

حيث ستقوم الطالبة بجمع معلومات من نظام الحاسوب المركزي DHIS2، دون التعرض للمعلومات التعريفية للمرضى، مع العلم أن مشرف الدراسة: د. محمد عوض.

وذلك في: **وحدة الحاسوب– قسم HIS**

على ان يتم الالتزام بالمحافظة على اخلاقيات البحث العلمي وسرية المعلومات.
على ان يتم الالتزام بجميع تعليمات واجراءات الوقاية والسلامة الصادرة عن وزارة الصحة بخصوص جائحة كورونا، وتحت طائلة المسؤولية. وابراز شهادة التطعيم قبل دخول مرافق وزارة الصحة.
على ان يتم تزويد الوزارة بنسخة PDF من نتائج البحث، التعهد بعدم النشر لحين الحصول على موافقة وزارة الصحة.

مع الاحترام...

د. عبد الله القواسمي
رئيس وحدة التعليم الصحي والبحث العلمي

نسخة: عميد البحث العلمي المحترم/ الجامعة العربية الامريكية

P.O .Box: 14
Telfax.:09-2333901
scientificresearch.dep@gmail.com
ص.ب. 14
تلفاكس: 09-2333901

الملخص

سرطان الثدي هو أكثر أنواع السرطانات شيوعا بين النساء في جميع أنحاء العالم. بسبب زيادة متوسط العمر المتوقع ، فإن حالات الإصابة بسرطان الثدي آخذة في الارتفاع في البلدان النامية. يُعد التنبؤ بسرطان الثدي وتصنيفه من القضايا المهمة للكشف عن السرطان في مراحله المبكرة. هناك العديد من عوامل الخطر للإصابة بسرطان الثدي. بسبب المشاكل البيئية العالمية، وطريقة الحياة، وزيادة عدد سكان العالم، من الممكن أن يزداد عدد الأشخاص المتضررين في السنوات القادمة. هناك مشكلة في غالبية النساء عدم التوجه لإجراء فحوصات تصوير الثدي بالأشعة السينية المتكررة الا بعد انتشار المرض وظهور الأعراض عليها.

في هذه الدراسة تم استخدام تقنية التنقيب عن البيانات للتنبؤ بمرض السرطان، وتصنيفه ، وذلك من خلال مجموعات بيانات للنساء الفلسطينيات التي تحتوي على تشخيص للحالات (الخبيثة ، الصحية ، الحميدة) وعوامل الخطر.

تحتوي مجموعة البيانات على 1794 سجلاً ؛ 635 حالة سرطان ثدي خبيث 570 حالة صحية و 589 حالة حميدة. تم جمع حالات سرطان الثدي من خلال نظام التسجيل الإلكتروني للتصوير الشعاعي للثدي في مديرية وزارة الصحة. يحتوي كل سجل على 30 سمة (متغيرات) بالإضافة إلى سمات الفئة. تتشكل الفئة كحالات صحية أو خبيثة أو حميدة.

يتم استخدام خوارزميات متعددة لاستخراج البيانات مثل الشبكات العصبية وشجرة القرار المعززة وغابة القرار وآلة المتجهات الداعمة. أداء النماذج مقارنة مع الخوارزميات لبعضها البعض وبالطرق

الإحصائية لتحديد أفضل نموذج لاستخراج البيانات في التنبؤ بسرطان الثدي لمجموعة البيانات الفلسطينية.

تظهر النتيجة أنه تم تطوير أفضل أربعة نماذج للتنبؤ بسرطان الثدي. النموذج الأول (نموذج NN) للكشف عن BC لثلاث فئات (خبيثة ، حميدة ، صحية) بالنتائج التالية :

96.7% accuracy, 95.2% Precision, 95.1%Recall, 95.2% F1 Score, and 97% AUC.

بينما كان أداء النموذج الثاني (نموذج NN) للتنبؤ بـ BC (التمييز بين الحالات الخبيثة والصحية) فكانت النتائج كالتالي:

98.5% accuracy, 97.8% Precision, 99.3% Recall which means very low false negative, 98.6% F1 Score, and 99.9% AUC.

النموذج الثالث تم تطويره لاكتشاف سرطان الثدي الخبيث والحميد فكانت النتائج كالتالي:

95.0% Accuracy, 98.0% Recall, and 99.5 % AUC.

أخيرًا ، تم تطوير نموذج مجموعة القرار للتنبؤ بين حالات المرض الصحية والحميدة فكانت النتائج كالتالي :

89.4% accuracy, 90.9% Precision, 93.9% Recall which means very low false negative, 92.4% F1 Score, and 96.1% AUC.

يمكن استخدام النماذج للتشخيص ولمساعدة الأطباء في تحديد ما إذا كان عدد كبير من النساء بحاجة إلى الكشف المبكر عن طريق التصوير الشعاعي للثدي أو غيره من الوسائل. يتيح ذلك إعطاء الأولوية للنساء المشتبه في إصابتهن بسرطان الثدي في المواعيد في عيادات سرطان الثدي.