Syrian Arab Republic

Ministry of Higher Education and Scientific Research

Syrian Virtual University

الجمهوريّة العربيّة السّوريّة

وزارة التّعليم العالي والبحث العلميّ

الجامعة الافتراضيّة السوريّة

Predicting Breast Cancer Prognosis Using Machine Learning

(A thesis submitted as a fulfilment of requirements for a Master's degree in Bioinformatics)

By

Ariana Younes

Supervised by

Prof. Dr. Majd Aljamali

2023 – 2022

Table of Contents:

Table of Abbreviations

Abbreviation	Meaning
1NN	1-Nearest Neighbor
AD	AdaBoost
AI	Artificial Intelligence
AIs	Aromatase Inhibitors
ALAN	Artificial light at night
ALND	Axillary Lymph Node Dissection
AUC	Area Under the Curve
BC	Breast Cancer
BCS	Breast-Conserving Surgery
BCSS	Breast Cancer-Specific Survival
BL1	Basal-Like subtype 1
BL2	Basal-Like subtype 2
BMBC	Bone Metastasis Breast Cancer
BMI	Body Mass Index
BRCA1	BReast CAncer gene 1
BRCA2	BReast CAncer gene 2
CDK	Cyclin-Dependent Kinase
CK	Cytokeratins
DCIS	Ductal Carcinoma in Situ
DDT	Dichlorodiphenyltrichloroethane
DNA	Deoxyribonucleic Acid
DT	Decision Trees
EGFR1	Epidermal Growth Factor Receptor Type 1
EMT	Epithelial–Mesenchymal Transition
ER	Estrogen Receptors
ESMO	European Society for Medical Oncology
FN	False Negative
FP	False Positive
HER2	Hormone Epidermal Growth Factor Receptor 2
HRT	Hormonal Replacement Therapy
IBC	Inflammatory Breast Cancer
IDC	Invasive Ductal Carcinoma
IHC	Immune-Histochemistry
ILC	Invasive Lobular Carcinoma
IM	Immunomodulatory
IMRT	Intensity-Modulated Radiotherapy
IORT	Intraoperative Radiation Therapy
LAR	Luminal Androgen Receptor
LCIS	Lobular Carcinoma In Situ
LED	Light-Emitting Diode
LumA	Luminal-A Subtype
LUMB	Luminal-B Subtype
MES	Mesenchymal
METABRIC	Molecular Taxonomy of Breast Cancer International Consortium
ML	Machine Learning
MLP	Multilayer Perceptron

MRI	Magnetic Resonance Imaging
mRNA	Messenger RNA
mTOR	Mechanistic Target of Rapamycin
N	Number
NB	Naive Bayes
NC	Not Classified
NSAIDs	Non-steroidal anti-inflammatory drugs
OS	Overall Survival
PAH	Polycyclic Aromatic Hydrocarbons
PCB	Polychlorinated Biphenyl
PoS tags	Part-of-Speech tagging
PR	Progesterone Receptors
Prolif	Proliferative Rate
PUFA	Polyunsaturated Fatty Acids
RBFN	RBF Network
RF	Random Forests
ROC	Receiver Operating Characteristic
SEER	Surveillance, Epidemiology, and End Results
SERDs	Selective Estrogen Receptor Degraders
SERMs	Selective Estrogen Receptor Modulators
SFN	Sulforaphane
Sig.	Significance
SLNB	Sentinel Lymph Node Biopsy
SNPs	Single Nucleotide Polymorphisms
SPSS	Statistical Package for the Social Sciences
SVM	Support Vector Machine
T-DM1	Trastuzumab combined with Emtasin
TN	True Negative
TNBC	Triple Negative Breast Cancer
TP	True Positive
TP53	Tumor Protein p53
TRF	Trees Random Forest
WHO	World Health Organization

Table of Figures

Table of Tables

Abstract

Background: Breast cancer is a significant global health issue, accounting for a large proportion of newly diagnosed cancers among women. Despite advancements in detection and treatment, breast cancer remains a leading cause of cancer-related deaths in women. Machine learning (ML), a branch of artificial intelligence, has emerged as a valuable tool for predicting breast cancer outcomes. ML models leverage diverse clinical and molecular features to improve the accuracy and reliability of prediction. Integration of genomic, proteomic, and imaging data has further enhanced the predictive capabilities of ML models. Previous studies have successfully developed ML models to predict various aspects of breast cancer, including tumor malignancy, survival probability, and risk of recurrence. These models provide valuable insights for clinical decision-making and have the potential to improve patient outcomes. Harnessing the power of ML in breast cancer prediction holds promise for advancing personalized medicine and optimizing treatment strategies.

Aim: The aim of the research is to find algorithms with high accuracy and sensitivity capable of predicting breast cancer prognosis and the cause of death in the study sample based on many variables in order to be able to intervene quickly in the patient's treatment protocol to reduce mortality as much as possible.

Materials and Methods: This study utilized the METABRIC database, containing targeted sequencing data of 1904 primary breast cancer samples, to predict breast cancer outcomes. Clinical and genetic attributes, such as age at diagnosis, type of surgery, chemotherapy, genetic expression levels, mutation data among others were analyzed using SPSS Statistics 25.0 and Python libraries. The dataset was split into training and test sets for model development and evaluation. Data preprocessing techniques were applied, and Python libraries facilitated data manipulation and analysis.

Results: The accuracy of the algorithms varied, with Logistic Regression achieving the lowest accuracy of 62.5%, while Decision Tree and Random Forest achieved perfect accuracies of 100%. SVM showed moderate performance with an accuracy of 75%. Confusion matrices provided additional insights into the classification performance of the algorithms. Decision Tree exhibited no errors in predictions, while Logistic Regression and SVM had misclassifications. The findings highlight the superior performance of Decision Tree and Random Forest algorithms in predicting breast cancer outcomes. These models demonstrated high accuracy and reliability, making them valuable tools for clinical decision-making. On the other hand, Logistic Regression showed lower accuracy, indicating the need for further improvement or exploration of alternative algorithms.

Conclusions: The study underscores the importance of selecting appropriate classification algorithms for predicting breast cancer patient outcomes. The Decision Tree and Random Forest algorithms offer promising results, while Logistic Regression may not be the most effective choice. These findings contribute to the field of breast cancer prognosis and provide insights for improving personalized treatment strategies. Future research can focus on exploring additional algorithms and incorporating more comprehensive datasets to further enhance predictive accuracy.

Theoretical Review

Chapter 1 – Breast Cancer

1.1. Introduction

Breast cancer has surpassed lung cancer as the most commonly diagnosed cancer worldwide. With 685,000 deaths and a total of 2.3 million new cases annually in both sexes combined, breast cancer accounts for 1 in every 4 cancer cases and 1 in every 6 cancer deaths in women, ranking first in incidence in the great majority of nations (159 of 185) and mortality in 110 countries (Fig. 1.1.) [1]. It was by far the most frequently diagnosed cancer in women in 2020, accounting for a quarter of all cancer cases in females, and its burden has been increasing in many regions of the world, particularly in transitional countries [2]. The primary reason breast cancer is incurable is that it is a metastatic disease that frequently spreads to distant organs such the bone, liver, lung, and brain. A positive prognosis and a high survival rate are advantages that come from early disease diagnosis [3].

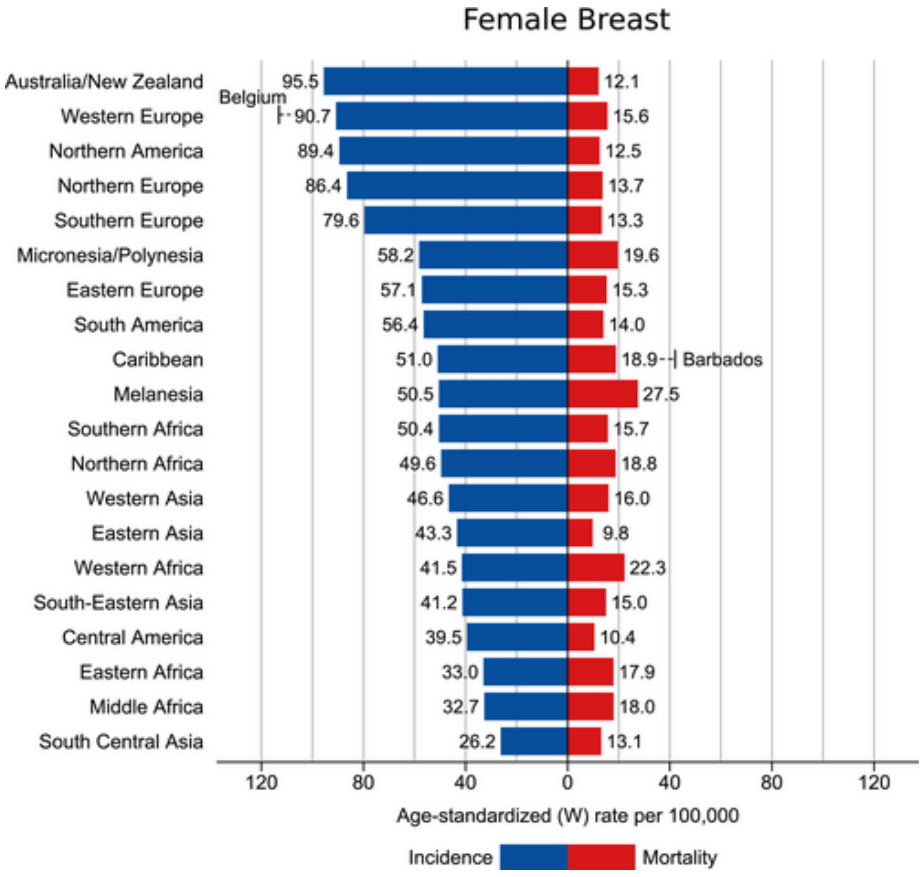

Figure 1.1. Region-Specific Incidence and Mortality Age-Standardized Rates for Female Breast Cancer in 2020 [1]

1.2. Types of Breast Cancer

Types of breast cancer include ductal carcinoma in situ, invasive ductal carcinoma, inflammatory breast cancer, and metastatic breast cancer [4].

1.2.1. Ductal Carcinoma in Situ (DCIS)

Ductal carcinoma in situ (DCIS) is a non-invasive cancer where abnormal cells have been found in the lining of the breast milk duct (Fig. 1.2.). The atypical cells have not spread outside of the ducts into the surrounding breast tissue. Ductal carcinoma in situ is very early cancer that is highly treatable, but if it's left untreated or undetected, it may spread into the surrounding breast tissue [5].

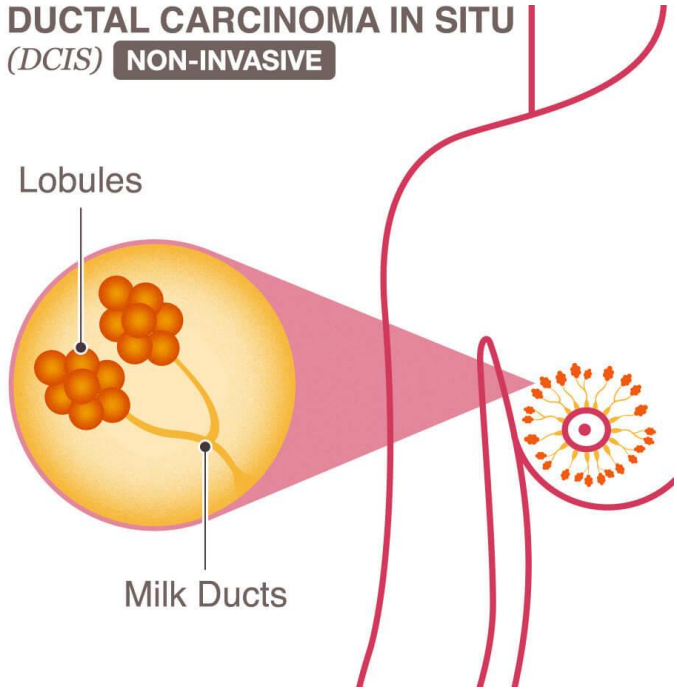

Figure 1.2. Non-invasive Ductal Cancinoma In Situ [5]

1.2.2. Invasive Ductal Carcinoma (IDC)

Invasive Ductal Carcinoma (IDC) is an invasive cancer where abnormal cancer cells that began forming in the milk ducts have spread beyond the ducts into other parts of the breast tissue. Invasive cancer cells can also spread to other parts of the body. It is also sometimes called infiltrative ductal carcinoma.

- IDC is the most common type of breast cancer, making up nearly 70- 80% of all breast cancer diagnoses.
- IDC is also the type of breast cancer that most commonly affects men [6].

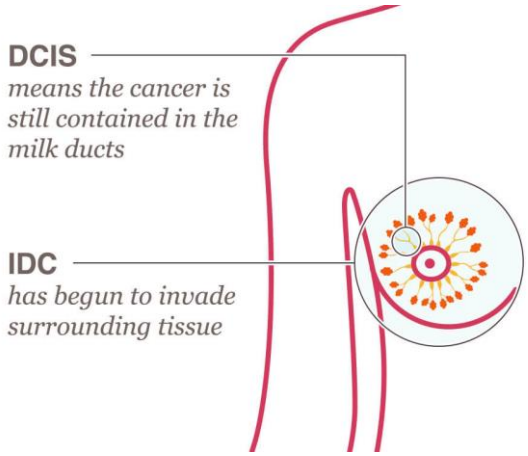

Figure 1.3. The Difference Between Invasive Ductal Carcinoma (IDC) And Ductal Carcinoma In Situ (DCIS) [6]

1.2.3. Lobular Carcinoma in situ (LCIS)

Lobular Carcinoma In Situ (LCIS) is a condition where abnormal cells are found in the lobules of the breast. The atypical cells have not spread outside of the lobules into the surrounding breast tissue [7].

1.2.4. Invasive Lobular Carcinoma (ILC)

Invasive breast cancer that begins in the lobules (milk glands) of the breast and spreads to surrounding normal tissue. It can also spread through the blood and lymph systems to other parts of the body. Invasive lobular breast cancer is the second most common type of breast cancer. Over 10% of invasive breast cancers are invasive lobular carcinomas. Though mammograms are helpful and important, they are less likely to detect invasive lobular breast cancer than other types of breast cancers. Invasive lobular cancer doesn't always appear clearly on a mammogram, instead an MRI might be needed [8].

1.2.5. Triple Negative Breast Cancer (TNBC)

A type of breast cancer in which the cells have tested negative for hormone epidermal growth factor receptor 2 (HER-2), estrogen receptors (ER), and progesterone receptors (PR). Since the tumor cells lack the necessary receptors, common treatments like hormone therapy and drugs that target estrogen, progesterone, and HER-2 are ineffective [9].

1.2.6. Inflammatory Breast Cancer (IBC)

Inflammatory breast cancer is aggressive and fast-growing breast cancer in which cancer cells infiltrate the skin and lymph vessels of the breast (Fig. 1.4.). It often produces no distinct tumor or lump that can be felt and isolated within the breast. But when the lymph vessels become blocked by the breast cancer cells, symptoms begin to appear [10].

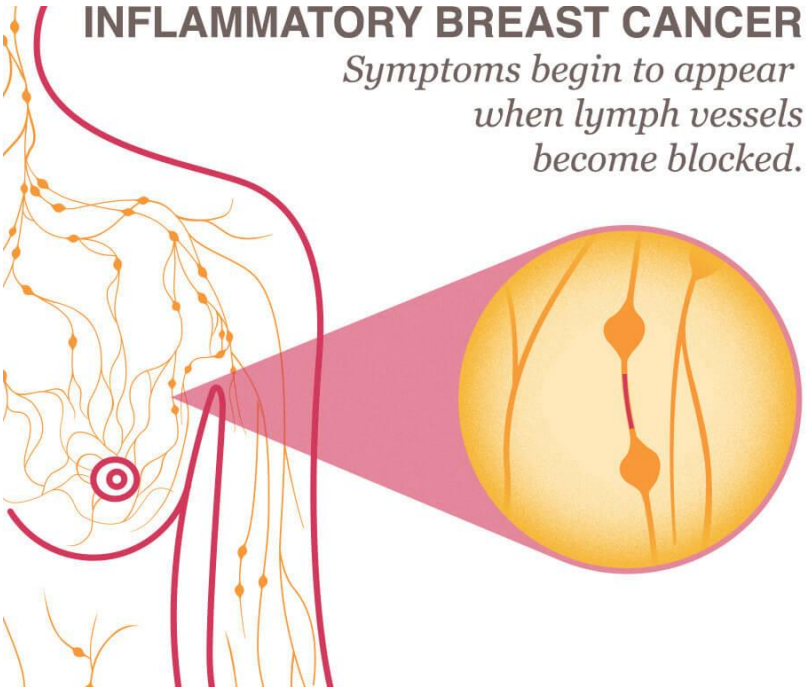

Figure 1.4. Inflammatory Breast Cancer [10]

1.2.7. Metastatic Breast Cancer

Metastatic breast cancer is also classified as Stage 4 breast cancer. The cancer has spread to other parts of the body. This usually includes the lungs, liver, bones or brain. The spread of cancer usually happens through one or more of the following steps:

- **Cancer cells invade nearby healthy cells.** When the healthy cell is taken over, it too can replicate more abnormal cells.
- **Cancer cells penetrate into the circulatory or lymph system.** Cancer cells travel through the walls of nearby lymph vessels or blood vessels.
- **Migration through circulation.** Cancer cells are carried by the lymph system and the bloodstream to other parts of the body.
- **Cancer cells lodge in capillaries.** Cancer cells stop moving as they are lodged in capillaries at a distant location and divide and migrate into the surrounding tissue.
- **New small tumors grow.** Cancer cells form small tumors at the new location (called micrometastases) [11].

1.2.8. Other Types

Although by far, the most common breast cancer type is ductal carcinoma in situ (DCIS), there are other types that are less commonly seen:

- **Medullary Carcinoma:** Medullary carcinoma accounts for 3-5% of all breast cancer types. The tumor usually shows up on a mammogram, but does not always feel like a lump. At times, it feels like a spongy change of breast tissue.
- **Tubular Carcinoma:** Making up about 2% of all breast cancer diagnosis, tubular carcinoma cells have a distinctive tubular structure when viewed under a microscope.

It is usually found through a mammogram and is a collection of cells that can feel like a spongy area of breast tissue rather than a lump. Typically this type of breast cancer is found in women aged 50 and above and usually responds well to hormone therapy.

- **Mucinous Carcinoma (Colloid):** Mucinous carcinoma represents approximately 1% to 2% of all breast cancers. The main differentiating features are mucus production and cells that are poorly defined. It also has a favorable prognosis in most cases.
- **Paget Disease of the breast or nipple:** This condition (also known as mammary Paget disease) is a rare type of cancer affecting the skin of the nipple and often the areola, which is the darker circle of skin around the nipple. Most people with Paget disease evident on the nipple also have one or more tumors inside the same breast; generally, either ductal carcinoma in situ or invasive breast cancer [12].

1.3. Molecular Subtypes of Breast Cancer

The molecular subtype of an invasive breast cancer is based on the genes the cancer cells express, which control how the cells behave. These subtypes are commonly grouped into four categories based on the immunohistochemical expression of hormone receptors: estrogen receptor positive (ER+), progesterone receptor positive (PR+), human epidermal growth factor receptor positive (HER2+), and triple-negative (TNBC), which is characterized by the lack of expression of any of the above receptors [13]. Estrogen receptor (ER) is an important diagnostic determinant, as approximately 70–75% of invasive breast carcinomas are characterized by significantly high ER expression [14, 15]. The progesterone receptor (PR) is expressed in more than 50% of ER-positive patients, and very rarely in those with ER-negative breast cancer. PR expression is regulated by ER [16]; therefore, physiological PR values inform about the functional ER pathway. However, both ER and PR are abundantly expressed in breast cancer cells, and both are considered diagnostic and prognostic biomarkers of breast cancer [17]. Higher expression of PR is positively associated with overall survival, time to recurrence, and time to treatment failure or progression, while lower levels are generally associated with a more aggressive course of disease, as well as poorer recurrence and prognosis [18].

Human epidermal growth factor receptor 2 (HER2) expression accounts for approximately 15–25% of breast cancers and its status is mainly relevant in the choice of appropriate treatment [19, 20]. HER2 overexpression is one of the earliest events during breast carcinogenesis [19]. HER2 increases the detection rate of metastatic or recurrent breast cancers by 50% and even 80%. Serum HER2 levels are considered a promising real-time marker for the presence or recurrence of tumors. HER2 amplification leads to increased overactivation of proto-oncogenic signaling pathways leading to uncontrolled cancer cell growth, which corresponds with worse clinical outcomes of HER2+ cases. HER2 overexpression also correlates with a significantly shorter disease-free period [21]. The Ki67 antigen is a cellular marker of proliferation and is an excellent marker for providing information on cell proliferation. The proliferative activities determined by Ki67 reflect the aggressiveness of the cancer along with response to treatment and time to recurrence [22]. Therefore, Ki-67 is crucial in terms of choosing the appropriate treatment therapy, and possible follow-ups for recurrence. It could also be considered as a possible prognostic factor. High expression of Ki67 also reflects lower survival rates [23, 24]. The need for molecular classification is to categorize patients who may benefit from targeted therapy, such as hormone therapy and anti HER2 therapy [25]. The characteristics of these four subtypes are presented bellow:

1.3.1. Luminal A Subtype

Luminal A tumors are characterized by the presence of ER and/or PR and the absence of HER2, and have a low expression of cell proliferation marker Ki-67 (less than 20%). Clinically they are low grade, slow growing, and have the best prognosis with less incidence of relapse and higher survival rate. These carcinomas present a high response rate to hormone therapy (tamoxifen or aromarase inhibitors), and a more limited benefit to chemotherapy [26].

1.3.2. Luminal B Subtype

Luminal B tumors are of higher grade and worse prognosis compared to Luminal A. They are ER positive and can be PR negative and have a high expression of Ki67 (greater than 20%). They are generally of intermediate/high histologic grade. These tumors may benefit from hormonal therapy along with chemotherapy. The elevated Ki67 makes them grow faster than luminal A and worse prognosis [27]. It constitutes 10–20% of luminal tumors. It has a moderately low expression of estrogen receptors, and increased expression of proliferation and cell cycle genes. It represents the group of luminal tumors with the worst prognosis. They benefit from hormone therapy and in a higher percentage from chemotherapy compared to the previous group [28].

1.3.3. HER2 Subtype

The HER2-positive group constitutes 10–15% of breast cancers and is characterized by high HER2 expression with absence of ER and PR. They grow faster than the luminal ones and the prognosis has improved after the introduction of HER2-targeted therapies. The HER2-positive subtype is more aggressive and fast-growing. Within this, two subgroups can be distinguished: luminal HER2 (E+, PR+, HER2+ and Ki-67:15–30%) and HER2-enriched (HER2+, E-, PR-, Ki-67>30%) [29]. They have a worse prognosis compared to luminal tumors, and require specific drugs directed against the HER2/neu protein, including trastuzumab, trastuzumab combined with emtasin (T-DM1), pertuzumab, and tyrosine kinase inhibitors such as lapatinib and neratinib, among others, in addition to surgery and treatment with precise chemotherapy [30]. They have a high response rate to chemotherapy schemes [31].

1.3.4. Triple-negative breast cancer (TNBC)

Triple-negative breast cancer is ER-negative, PR-negative, and HER2-negative. They constitute about 20% of all breast cancers. It is most common among women under 40 years of age, and in African-American women. The TNBC subtype is further classified into several additional subgroups including basal-like (BL1 and BL2), claudin-low, mesenchymal (MES), luminal androgen receptor (LAR), and immunomodulatory (IM), the first two being the most frequent with 50–70% and 20–30% of cases [32]. Moreover, each of these has unique clinical outcomes, phenotypes, and pharmacological sensitivities. TNBC presents an aggressive behavior and 80% of breast cancer tumors (tumor suppressor gene BRCA1 and BRCA2) belong to this group [33]. The risk of developing TNBC varies with genetics, race, age, overweight and obesity, breastfeeding patterns, and parity [32, 34]. TNBC is characterized by its aggressiveness, early relapse, and a greater tendency to present in advanced stages. It presents a high proliferation rate, alteration in DNA repair genes and increased genomic instability. Histologically, it is a poorly differentiated, highly proliferative, heterogeneous neoplasm, including subsets of variable prognosis. Immunohistochemically, they are

subdivided into basal and non-basal TNBC; the former characterized by expression of cytokeratins (CK)5/6 and human epidermal growth factor receptor type 1 (EGFR1), while the non-basal do not express CK5/6 cytokeratins.

1.4. Breast Cancer Risk Factors

The number of risk factors of breast cancer is significant and includes both modifiable factors and non-modifiable factors.

1.4.1. Non-Modifiable Factors

- **Female Sex**

Female sex constitutes one of the major factors associated with an increased risk of breast cancer primarily because of the enhanced hormonal stimulation. Unlike men who present insignificant estrogen levels, women have breast cells which are very vulnerable to hormones (estrogen and progesterone in particular) as well as any disruptions in their balance. Circulating estrogens and androgens are positively associated with an increased risk of breast cancer [35]. Less than 1% of all breast cancers occur in men. However, breast cancer in men is a rare disease that's at the time of diagnosis tends to be more advanced than in women. The average age of men at the diagnosis is about 67. The important factors increase a man's risk of breast cancer are: older age, BRCA2/BRCA1 mutations, increased estrogen levels, Klinefelter syndrome, family history of breast cancer, and radiation exposure [36].

- **Older Age**

Currently, about 80% of patients with breast cancer are individuals aged >50 while at the same time more than 40% are those more than 65 years old [37–39]. The risk of developing breast cancer increases as follows—the 1.5% risk at age 40, 3% at age 50, and more than 4% at age 70 [40]. Interestingly, a relationship between a particular molecular subtype of cancer and a patient's age was observed –aggressive resistant triple-negative breast cancer subtype is most commonly diagnosed in groups under 40 age, while in patients >70, it is luminal A subtype [37]. Generally, the occurrence of cancer in older age is not only limited to breast cancer; the accumulation of a vast number of cellular alternations and exposition to potential carcinogens results in an increase of carcinogenesis with time.

- **Family History**

A family history of breast cancer constitutes a major factor significantly associated with an increased risk of breast cancer. Approximately 13–19% of patients diagnosed with breast cancer report a first-degree relative affected by the same condition [41]. Besides, the risk of breast cancer significantly increases with an increasing number of first-degree relatives affected; the risk might be even higher when the affected relatives are under 50 years old [42–44]. The incidence rate of breast cancer is significantly higher in all of the patients with a family history despite the age. This association is driven by epigenetic changes as well as environmental factors acting as potential triggers [45]. A family history of ovarian cancer—especially those characterized by *BRCA1* and *BRCA2* mutations—might also induce a greater risk of breast cancer [46].

- **Genetic Mutations**

Several genetic mutations were reported to be highly associated with an increased risk of breast cancer. Two major genes characterized by a high penetrance are *BRCA1* (located on chromosome 17) and *BRCA2* (located on chromosome 13). They are primarily linked to the increased risk of breast carcinogenesis [47]. The mutations within the above-mentioned genes are mainly inherited in an autosomal dominant manner, however, sporadic mutations are also commonly reported.

- **Race/Ethnicity**

Disparities regarding race and ethnicity remain widely observed among individuals affected by breast cancer; the mechanisms associated with this phenomenon are not yet understood. Generally, the breast cancer incidence rate remains the highest among white non-Hispanic women [48, 49]. Contrarily, the mortality rate due to this malignancy is significantly higher among black women; this group is also characterized by the lowest survival rates [50].

- **Height**

Many studies have found that taller women have a higher risk of breast cancer than shorter women. The reasons for this aren't exactly clear, but it may have something to do with factors that affect early growth, such as nutrition early in life, as well as hormonal or genetic factors [51].

- **Starting menstrual periods early**

Women who have had more menstrual cycles because they started menstruating early (especially before age 12) have a slightly higher risk of breast cancer. The increase in risk may be due to a longer lifetime exposure to the hormones estrogen and progesterone [51].

- **Going through menopause later**

Women who have had more menstrual cycles because they went through menopause later (typically after age 55) have a slightly higher risk of breast cancer. The increase in risk may be because they have a longer lifetime exposure to the hormones estrogen and progesterone [51].

- **Reproductive History**

Numerous studies confirmed a strict relationship between exposure to endogenous hormones - estrogen and progesterone in particular - and excessive risk of breast cancer in females. Therefore, the occurrence of specific events such as pregnancy, breastfeeding, first menstruation, and menopause along with their duration and the concomitant hormonal imbalance, are crucial in terms of a potential induction of the carcinogenic events in the breast microenvironment. The first full-term pregnancy at an early age (especially in the early twenties) along with a subsequently increasing number of births are associated with a reduced risk of breast cancer [52, 53]. Besides, the pregnancy itself provides protective effects against potential cancer. However, protection was observed at approximately the 34th pregnancy week and was not confirmed for the pregnancies lasting for 33 weeks or less [54].

- **Density of Breast Tissue**

The density of breast tissue remains inconsistent throughout the lifetime; however, several categories including low-density, high-density, and fatty breasts have been established in clinical practice. Greater density of breasts is observed in females of younger age and lower BMI, who are pregnant or during the breastfeeding period, as well as during the intake of hormonal replacement therapy [55]. Generally, the greater breast tissue density correlates with the greater breast cancer risk; this trend is observed both in premenopausal and postmenopausal females [56]. It was proposed that screening of breast tissue density could be a promising, non-invasive, and quick method enabling rational surveillance of females at increased risk of cancer [57].

- **History of Breast Cancer and Benign Breast Diseases**

Personal history of breast cancer is associated with a greater risk of a renewed cancerous lesions within the breasts [58]. Besides, a history of any other non-cancerous alternations in breasts such as atypical hyperplasia, carcinoma in situ, or many other proliferative or non-proliferative lesions, also increases the risk significantly [59–61]. The histologic classification of benign lesions and a family history of breast cancer are two factors that are strongly associated with breast cancer risk [61].

- **Previous Radiation Therapy**

The risk of secondary malignancies after radiotherapy treatment remains an individual matter that depends on the patient's characteristics, even though it is a quite frequent phenomenon that arises much clinical concern. Cancer induced by radiation therapy is strictly associated with an individual's age; patients who receive radiation therapy before the age of 30, are at a greater risk of breast cancer [62]. The selection of proper radiotherapy technique is crucial in terms of secondary cancer risk [63]. Besides, the family history of breast cancer in patients who receive radiotherapy additionally enhances the risk of cancer occurrence [64].

1.4.2. Modifiable Factors

- **Drugs**

The intake of diethylstilbestrol during pregnancy is associated with an increased risk of breast cancer not only in mothers but also in the offspring [65]. This relationship is observed despite the expression of neither estrogen nor progesterone receptors and might be associated with every breast cancer histological type. The risk increases with age; women at age of ≥40 years are nearly 1.9 times more susceptible compared to women under 40. Moreover, breast cancer risk increases with greater diethylstilbestrol doses [66]. Numerous researches indicate that females who use hormonal replacement therapy (HRT) especially longer than 5 or 7 years are also at increased risk of breast cancer [67, 68]. Several studies indicated that the intake of chosen antidepressants, mainly paroxetine, tricyclic antidepressants, and selective serotonin reuptake inhibitors might be associated with a greater risk of breast cancer [69, 70].

- **Physical Activity**

Even though the mechanism remains yet undeciphered, regular physical activity is considered to be a protective factor of breast cancer incidence [71, 72]. Amongst females with a family history of breast cancer, physical activity was associated with a reduced risk of cancer but limited only to the postmenopausal period [73]. There are several hypotheses aiming to explain the protective role of physical activity in terms of breast cancer incidence; physical activity might prevent cancer by reducing the exposure to the endogenous sex hormones, altering immune system responses or insulin-like growth factor-1 levels [73, 74].

- **Body Mass Index**

According to epidemiological evidence, obesity is associated with a greater probability of breast cancer. This association is mostly intensified in obese post-menopausal females who tend to develop estrogen-receptor-positive breast cancer. Yet, independently to menopausal status, obese women achieve poorer clinical outcomes [75]. Females above 50 years old with greater Body Mass Index (BMI) are at a greater risk of cancer compared to those with low BMI [76]. Besides, the researchers observed that greater BMI is associated with more aggressive biological features of tumor including a higher percentage of lymph node metastasis and greater size. Obesity might be a reason for greater mortality rates and a higher probability of cancer relapse, especially in premenopausal women [77]. Increased body fat might enhance the inflammatory state and affects the levels of circulating hormones facilitating pro-carcinogenic events [78]. Thus, poorer clinical outcomes are primarily observed in females with BMI $\geq$ 25 kg/m2 [79]. Interestingly, postmenopausal women tend to present poorer clinical outcomes despite proper BMI values but namely due to excessive fat volume [80]. Greater breast cancer risk with regards to BMI also correlates with the concomitant family history of breast cancer [81].

- **Alcohol Intake**

Numerous evidences confirm that excessive alcohol consumption is a factor that might enhance the risk of malignancies within the gastrointestinal tract; however, it was proved that it is also linked to the risk of breast cancer. Namely, it is not alcohol type but rather the content of alcoholic beverages that mostly affect the risk of cancer. The explanation for this association is the increased levels of estrogens induced by the alcohol intake and thus hormonal imbalance affecting the risk of carcinogenesis within the female organs [82, 83]. Besides, alcohol intake often results in excessive fat gain with higher BMI levels, which additionally increases the risk. Other hypotheses include direct and indirect carcinogenic effects of alcohol metabolites and alcohol-related impaired nutrient intake [84]. Alcohol consumption was observed to increase the risk of estrogen-positive breast cancers in particular [85]. Consumed before the first pregnancy, it significantly contributes to the induction of morphological alterations of breast tissue, predisposing it to further carcinogenic events [86].

- **Smoking**

Carcinogens found in tobacco are transported to the breast tissue increasing the plausibility of mutations within oncogenes and suppressor genes (*p53* in particular). Thus, not only active but also passive smoking significantly contributes to the induction of pro-carcinogenic events [86]. Besides, longer smoking history, as well as smoking before the first full-term pregnancy,

are additional risk factors that are additionally pronounced in females with a family history of breast cancer [87–90].

- **Insufficient Vitamin Supplementation**

Vitamins exert anticancer properties, which might potentially benefit in the prevention of several malignancies including breast cancer, however, the mechanism is not yet fully understood. Attempts are continually made to analyze the effects of vitamin intake (vitamin C, vitamin E, B-group vitamins, folic acid, multivitamin) on the risk of breast cancer, nevertheless, the data remains inconsistent and not sufficient to compare the results and draw credible data [88]. In terms of breast cancer, most studies are currently focused on vitamin D supplementation confirming its potentially protective effects [91–93]. High serum 25-hydroxyvitamin D levels are associated with a lower incidence rate of breast cancer in premenopausal and postmenopausal women [92, 94]. Intensified expression of vitamin D receptors was shown to be associated with lower mortality rates due to breast cancer [95]. Even so, further evaluation is required since data remains inconsistent in this matter [88, 96].

- **Exposure to Artificial Light**

Artificial light at night (ALAN) has been recently linked to increased breast cancer risk. The probable causation might be a disrupted melatonin rhythm and subsequent epigenetic alterations [97]. According to the studies conducted so far, increased exposure to ALAN is associated with a significantly greater risk of breast cancer compared to individuals with lowered ALAN exposure [98]. Nonetheless, data regarding the excessive usage of LED electronic devices and increased risk of breast cancer is insufficient and requires further evaluation as some results are contradictory [98].

- **Intake of Processed Food/Diet**

According to the World Health Organization (WHO), highly processed meat was classified as a Group 1 carcinogen that might increase the risk of not only gastrointestinal malignancies but also breast cancer. Similar observations were made in terms of an excessive intake of saturated fats [99]. Ultra-processed food is rich in sodium, fat, and sugar which subsequently predisposes to obesity recognized as another factor of breast cancer risk [100]. It was observed that a 10% increase of ultra-processed food in the diet is associated with an 11% greater risk of breast cancer [100]. Contrarily, a diet high in vegetables, fruits, legumes, whole grains, and lean protein is associated with a lowered risk of breast cancer [101]. Generally, a diet that includes food containing high amounts of n-3 PUFA, vitamin D, fiber, folate, and phytoestrogen might be beneficial as a prevention of breast cancer [102]. Besides, lower intake of n-6 PUFA and saturated fat is recommended. Several in vitro and in vivo studies also suggest that specific compounds found in green tea might present anti-cancer effects which has also been studied regarding breast cancer [103]. Similar properties were observed in case of turmeric-derived curcuminoids as well as sulforaphane (SFN) [104, 105].

- **Exposure to Chemical**

Chronic exposure to chemicals can promote breast carcinogenesis by affecting the tumor microenvironment subsequently inducing epigenetic alterations along with the induction of pro-carcinogenic events [106]. Females chronically exposed to chemicals present significantly

greater plausibility of breast cancer which is further positively associated with the duration of the exposure [107]. The number of chemicals proposed to induce breast carcinogenesis is significant; so far, dichlorodiphenyltrichloroethane (DDT) and polychlorinated biphenyl (PCB) are mostly investigated in terms of breast cancer since early exposure to those chemicals disrupts the development of mammary glands [108, 109]. A potential relationship was also observed in the case of increased exposure to polycyclic aromatic hydrocarbons (PAH), synthetic fibers, organic solvents, oil mist, and insecticides [110].

- **Other Drugs**

Other drugs that might constitute potential risk factors for breast cancer include antibiotics, antidepressants, statins, antihypertensive medications (e.g., calcium channel blockers, angiotensin II-converting enzyme inhibitors), as well as NSAIDs (including aspirin, ibuprofen) [111–115].

1.5. Treatment Strategies

1.5.1. Surgery

There are two major types of surgical procedures enabling the removal of breast cancerous tissues and those include breast-conserving surgery (BCS) and mastectomy. BCS - also called partial/segmental mastectomy, lumpectomy, wide local excision, or quadrantectomy - enables the removal of the cancerous tissue with simultaneous preservation of intact breast tissue often combined with plastic surgery technics called oncoplasty. Mastectomy is a complete removal of the breast and is often associated with immediately breast reconstruction. The removal of affected lymph nodes involves sentinel lymph node biopsy (SLNB) and axillary lymph node dissection (ALND). Even though BCS seems to be highly more beneficial for patients, those who were treated with this technique often show a tendency for a further need for a complete mastectomy [116]. However, usage of BCS is mostly related to significantly better cosmetic outcomes, lowered psychological burden of a patient, as well as reduced number of postoperative complications [117]. Guidelines of the European Society for Medical Oncology (ESMO) for patients with early breast cancer make the choice of therapy dependent to tumor size, feasibility of surgery, clinical phenotype, and patient's willingness to preserve the breast [118].

1.5.2. Chemotherapy

Chemotherapy is a systemic treatment of BC and might be either neoadjuvant or adjuvant. Choosing the most appropriate one is individualized according to the characteristics of the breast tumor; chemotherapy might also be used in the secondary breast cancer. Neoadjuvant chemotherapy is used for locally advanced BC, inflammatory breast cancers, for downstaging large tumors to allow BCS or in small tumors with worse prognostics molecular subtypes (HER2 or TNBC) which can help to identify prognostics and predictive factors of response and can be provided intravenously or orally. Currently, treatment includes a simultaneous application of schemes 2–3 of the following drugs—carboplatin, cyclophosphamide, 5-fluorouracil/capecitabine, taxanes (paclitaxel, docetaxel), and anthracyclines (doxorubicin, epirubicin). The choice of the proper drug is of major importance since different molecular breast cancer subtypes respond differently to preoperative chemotherapy [119]. Preoperative chemotherapy is comparably effective to postoperative chemotherapy [120].

Even though chemotherapy is considered to be effective, its usage very often leads to several side effects including hair loss, nausea/vomiting, diarrhea, mouth sores, fatigue, increased susceptibility to infections, bone marrow supression, combined with leucopenia, anaemia, easier bruising or bleeding; other less frequent side effects include cardiomyopathy, neuropathy, hand-foot syndrome, impaired mental functions. In younger women, disruptions of the menstrual cycle and fertility issues might also appear. Special form of chemotherapy is electrochemotherapy which can be used in patients with breast cancer that has spread to the skin, however, it is still quite uncommon and not available in most clinics.

1.5.3. Radiation Therapy

Radiotherapy is local treatment of BC, typically provided after surgery and/or chemotherapy. It is performed to ensure that all of the cancerous cells remain destroyed, minimizing the possibility of breast cancer recurrence. Further, radiation therapy is favorable in the case of metastatic or unresectable breast cancer [121]. Choice of the type of radiation therapy depends on previous type of surgery or specific clinical situation; most common techniques include breast radiotherapy (always applied after BC), chest-wall radiotherapy (usually after mastectomy), and ‘breast boost’ (a boost of high-dose radiotherapy to the place of tumor bed as a complement of breast radiotherapy after BCS). Regarding breast radiotherapy specifically, several types are distinguished including: intraoperative radiation therapy (IORT), 3D-conformal radiotherapy (3D-CRT), intensity-modulated radiotherapy (IMRT) and brachytherapy - which refers to internal radiation in contrast to other above-mentioned techniques.

Irritation and darkening of the skin exposed to radiation, fatigue, and lymphedema are one of the most common side effects of radiation therapy applied in breast cancer patients. Nonetheless, radiation therapy is significantly associated with the improvement of the overall survival rates of patients and lowered risk of recurrence [122].

1.5.4. Endocrinal (Hormonal) Therapy

Endocrinal therapy might be used either as a neoadjuvant or adjuvant therapy in patients with Luminal–molecular subtype of BC; it is effective in cases of breast cancer recurrence or metastasis. Since the expression of ERs, a very frequent phenomenon in breast cancer patients, its blockage via hormonal therapy is commonly used as one of the potential treatment modalities. Endocrinal therapy aims to lower the estrogen levels or prevents breast cancer cells to be stimulated by estrogen. Drugs that block ERs include selective estrogen receptor modulators (SERMs) (tamoxifen, toremifene) and selective estrogen receptor degraders (SERDs) (fulvestrant) while treatments that aim to lower the estrogen levels include aromatase inhibitors (AIs) (letrozole, anastrazole, exemestane) [123, 124]. In the case of pre-menopausal women, ovarian suppression induced by oophorectomy, luteinizing hormone-releasing hormone analogs, or several chemotherapy drugs, are also effective in lowering estrogen levels [125]. However, approximately 50% of hormonoreceptor-positive breast cancer become progressively resistant to hormonal therapy during such treatment [126]. Endocrinal therapy combined with chemotherapy is associated with the reduction of mortality rates amongst breast cancer patients [127].

1.5.5. Biological Therapy

Biological therapy (targeted therapy) can be provided at every stage of breast therapy– before surgery as neoadjuvant therapy or after surgery as adjuvant therapy. Biological therapy is quite common in HER2-positive breast cancer patients; major drugs include trastuzumab, pertuzumab, trastuzumab deruxtecan, lapatinib, and neratinib [128–132].

In the case of Luminal, HER2-negative breast cancer, pre-menopausal women more often receive everolimus - mTOR inhibitor with exemestane while postmenopausal women often receive CDK 4–6 inhibitor palbociclib or ribociclib simultaneously, combined with hormonal therapy [133–135]. Two penultimate drugs along with abemaciclib and everolimus can also be used in HER2-negative and estrogen-positive breast cancer [136, 137]. Atezolizumab is approved in triple-negative breast cancer, while denosumab is approved in case of metastasis to the bones [138–140].

Chapter 2 - Machine Learning

2.1. Introduction to Machine Learning

In 1956, a group of computer scientists proposed that computers could be programmed to think and reason, they described this principle as “artificial intelligence” [141]. ML is a field that focuses on the learning aspect of AI by developing algorithms that best represent a set of data. In contrast to classical programming, in which an algorithm can be explicitly coded using known features, ML uses subsets of data to generate an algorithm that may use novel or different combinations of features and weights than can be derived from first principles (Fig. 2.1.) [142].

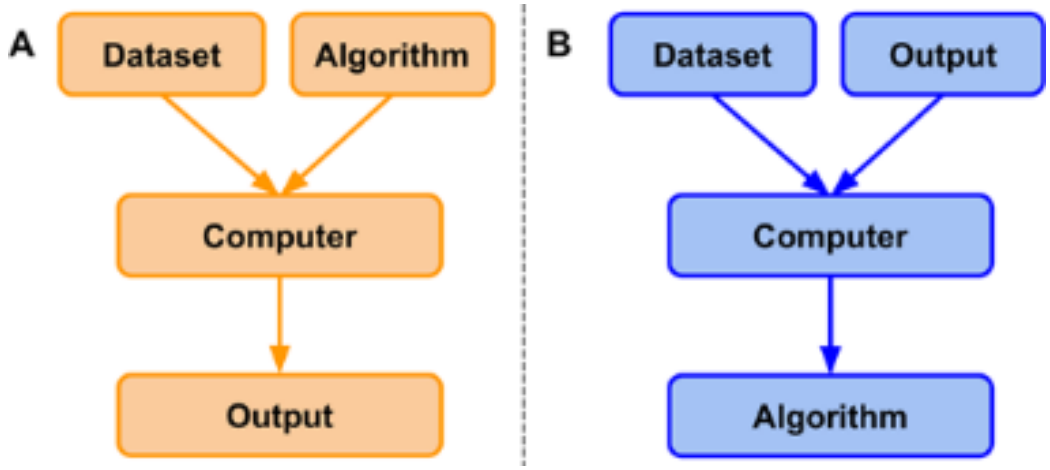

Figure 2.1 Classical programming versus machine learning paradigm. (A) Classical programming (B) Machine Learning [142]

Machine learning employs a variety of statistical, probabilistic and optimization techniques that allows computers to “learn” from past examples and to detect hard-to-discern patterns from large, noisy or complex data sets. This capability is particularly well-suited to medical applications, especially those that depend on complex proteomic and genomic measurements. As a result, machine learning is frequently used in cancer diagnosis and detection. More recently machine learning has been applied to cancer prognosis and prediction.

Machine learning, like statistics, is used to analyze and interpret data. Unlike statistics, though, machine learning methods can employ Boolean logic (AND, OR, NOT), absolute conditionality (IF, THEN, ELSE), conditional probabilities (the probability of X given Y) and unconventional optimization strategies to model data or classify patterns. These latter methods actually resemble the approaches humans typically use to learn and classify. Machine learning still draws heavily from statistics and probability, but it is fundamentally more powerful because it allows inferences or decisions to be made that could not otherwise be made using conventional statistical methodologies.

2.2. Learning methods used in Machine Learning

Machine learning is broadly classified as supervised, unsupervised, semi-supervised, and reinforcement learning. A supervised learning model has two major tasks to be performed, classification and regression. Classification is about predicting a nominal class label, whereas regression is about predicting the numeric value for the class label. Mathematically, building a regression model is all about identifying the relationship between the class label and the input predictors. Predictors are also called attributes. In statistical terms, the predictors are called independent variables, while the class label is called dependent variable. A regression model is a representation of this relationship between dependent and independent variables. Once this is learnt during the training phase, any new data is plugged into the relationship curve to find the prediction. This reduces the machine learning problem to solving a mathematical equation. The broad classification of machine learning is depicted in (Fig. 2.2.) [143].

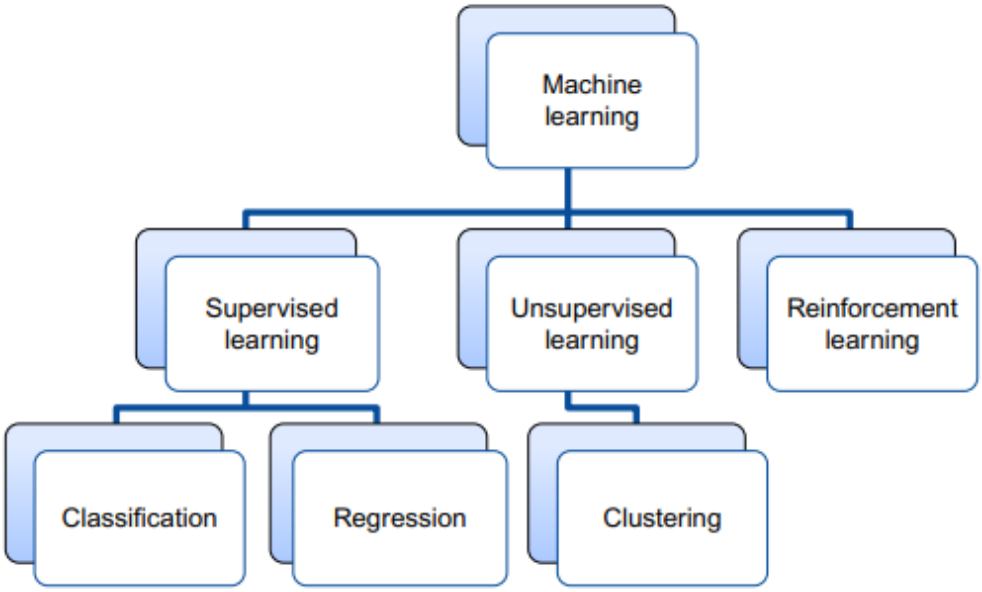

Figure 2.2 Types of learning [143]

2.2.1. Supervised Learning

Supervised learning is a learning model built to make prediction, given an unforeseen input instance. A supervised learning algorithm takes a known set of input dataset and its known responses to the data (output) to learn the regression/classification model. A learning algorithm then trains a model to generate a prediction for the response to new data or the test dataset. Supervised learning uses classification algorithms and regression techniques to develop predictive models. The algorithms include linear regression, logistic regression, and neural networks as well, apart from decision tree, Support Vector Machine (SVM), random forest, naive Bayes, and k-nearest neighbor.

Classification task predicts discrete responses. It is recommended if the data can be categorized, tagged, or separated into specific groups or classes. Classification models classify input data into categories. Popular or major applications of classification include bank credit scoring, medical imaging, and speech recognition. Also, handwriting recognition uses

classification to recognize letters and numbers, to check whether an email is genuine or spam, or even to detect whether a tumor is benign or cancerous.

Regression techniques predict continuous responses. A linear regression attempts to model the relationship between two variables by fitting linear equation to observed data. For example, say, a data is collected about how happy people are after getting so many hours of sleep. In this dataset, sleep and happy people are the variables. By regression analysis, one can relate them and start making predictions.

In Natural Language Processing, the input can contain an annotated text provided by humans. Annotated text is a metadata that is given along with the dataset to the machine. The annotations can be Part-of-Speech tagging (PoS tags), phrase, and dependency structures. For example, to determine whether the text "clean dishes" is a noun phrase or a verb phrase, the algorithm needs to be trained using annotated sentences like "Clean dishes are in the cupboard" or "Clean dishes before going to work." In the first case, the annotation says that it is a noun phrase and verb phrase in the second case [143].

2.2.2. Unsupervised Learning

In supervised learning, the goal is to learn mapping from the input to an output whose correct values are provided by a supervisor. But, in unsupervised learning, the goal is to find the regularities in the input such that certain patterns occur more often than others and to learn to see what generally happens and what does not. Examples on speech recognition, document clustering, and image compression go well with unsupervised learning. In document clustering, the aim is to group documents into various reports of politics, entertainment, sports, culture, heritage, art, and so on. Usually any document is represented as a "bag of words," that is, predefined lexicon of N words. Each document is an N-dimensional binary vector whose element "I" is 1. If the word "I" appears in the document, its suffixes "-s" and "-ing" are removed to avoid duplicates and stop words such as "of," "and," "the," "a" are also removed. Remaining terms in the document are then grouped, depending on the number of shared words [143].

2.2.3. Semi-supervised Learning

This learning technique is the combination of supervised and unsupervised learning and is used when less number of labeled data is identified for a particular application. It generates a function mapping from inputs of both labeled data and unlabeled data. The goal of semi-supervised learning is to classify some of the unlabeled data using the labeled information set. In a semi-supervised learning scenarios, the size of the unlabeled dataset should be substantially larger than the labeled data. Otherwise, the problem can be simply addressed by using supervised algorithms. Some real world examples like, protein sequence classification, web content classification and speech analysis where labeling audio files is a very intensive task and requires lot of human intervention [143].

2.2.4. Reinforcement Learning

Reinforcement learning involves interacting with the surrounding environment. Reinforcement learning addresses the issue of how an autonomous agent that senses and acts in its environment can learn to choose optimal actions to achieve its goals. An agent's behavior is rewarded based on the actions it takes in the environment. It considers the consequences of its actions and takes optimal steps further. A computer playing chess with the human, learning

to recognize spoken words, and learning to classify new astronomical structures are few examples of reinforcement learning [143].

2.3. Classification Algorithms

The classification process within a machine learning environment can be defined as the process of distributing data, which is called training data, into a slightly smaller training dataset and a separate validation dataset. The classification process is included in the supervised learning methods, since there are all the real values of all training data, and the classification process is the basis for the forecasting process through the models that are built during the classification process and related to the type of workbook used, and below we review the classifiers that have been used within this research that work more effectively with multiple classification issues, and the algorithms on which they are based to accomplish the various classification processes.

2.3.1. Logistic Regression

Logistic regression is a classification algorithm where the goal is to find a relationship between features and the probability of a particular outcome. Rather than using the straight line produced by linear regression to estimate class probability, logistic regression uses a sigmoidal curve to estimate class probability (Fig. 2.3.). This curve is determined by the sigmoid function, $y = \frac{1}{1+e^{-x}}$, which produces an S-shaped curve that converts discrete or continuous numeric features (x) into a single numerical value (y) between 0 and 1. The major advantage of this method is that probabilities are bounded between 0 and 1 (i.e., probabilities cannot be negative or greater than 1). It can be either binomial, where there are only two possible outcomes, or multinomial, where there can be three or more possible outcomes.

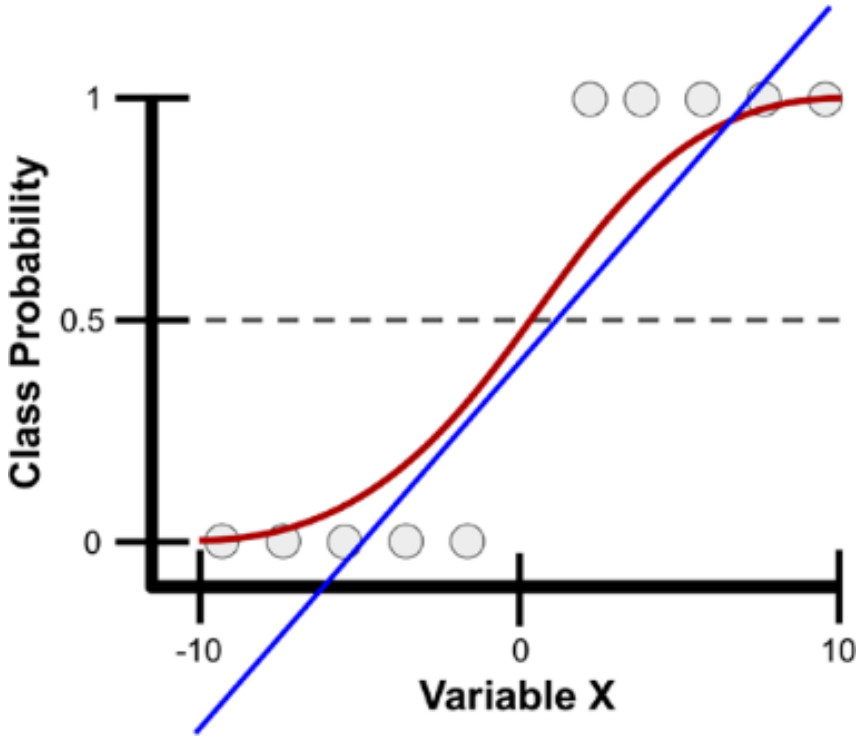

Figure 2.3 Example class probability prediction using linear and logistic regression

Presented are linear (*blue line*) and logistic (*red line*) regression models for predicting the probability of various samples (*gray circles*) as belonging to a particular class using a single variable, variable *X*, which ranges from -10 to 10. With logistic regression, variable *X* is transformed into class probabilities that are bounded between 0 and 1 using the sigmoid function. Simple linear regression attempts to estimate class probabilities, but is not bounded between 0 and 1; thus, it breaks a fundamental law of probability that does not allow for negative probabilities or those greater than 1.

2.3.2. Support Vector Machine (SVM)

SVM is one of the best known algorithms that would separate two classes using a hyperplane. SVM can not only perform linear classification but also can efficiently manage nonlinear data. An SVM model projects the samples of every label into a vector space. SVM then tries to separate the projected points such that they have a maximum distance between them. When a new sample is given to the model, it is projected into the vector space and the class/label to which it belongs is predicted based upon which side of the line it falls. In SVM, a decision surface is used to separate the classes and to maximize the margin between the classes. The following section explains different types of SVMs:

- **Linear SVM:** Linear SVMs use a linear decision boundary to separate the data points of different classes. When the data can be precisely linearly separated, linear SVMs are very suitable. This means that a single straight line (in 2D) (Fig. 2.4. A) or a hyperplane (in higher dimensions) can entirely divide the data points into their respective classes (Fig. 2.4. B). A hyperplane that maximizes the margin between the classes is the decision boundary.

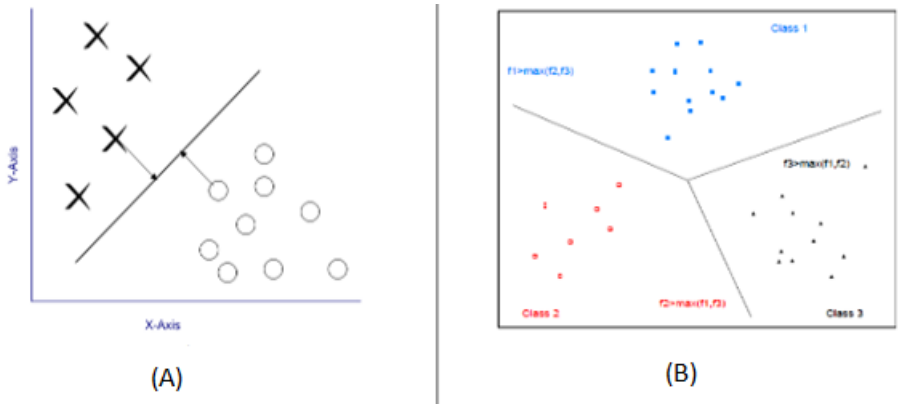

Figure 2.4. Linear Support Vector Machine. (A) 2D, (B) Higher dimensions

- **Non-Linear SVM:** Non-Linear SVM can be used to classify data when it cannot be separated into two classes by a straight line (in the case of 2D). By using kernel functions, nonlinear SVMs can handle nonlinearly separable data. The original input data is transformed by these kernel functions into a higher-dimensional feature space, where the data points can be linearly separated. A linear SVM is used to locate a nonlinear decision boundary in this modified space. Two cases can also be distinguished according to the categories and dimensions of the data, whether they are binary (2D) (Fig. 2.5. A) or higher dimensions (Fig. 2.5. B).

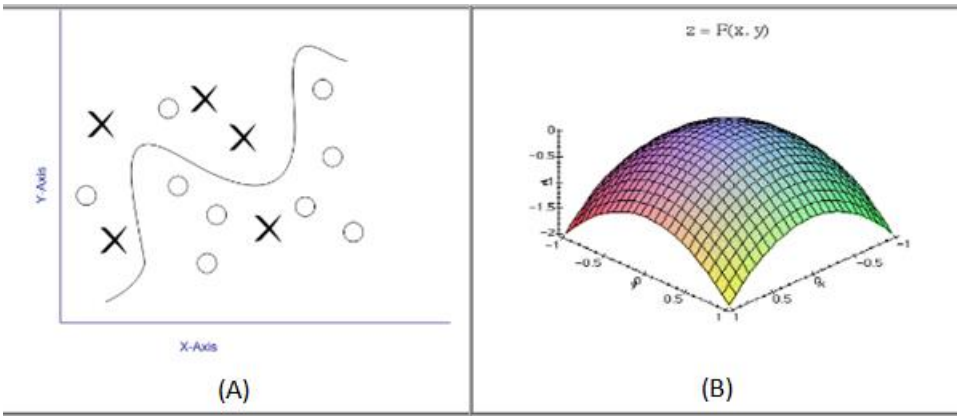

Figure 2.5. Non-Linear Support Vector Machine. (A) 2D, (B) Higher dimensions

2.3.3. Decision Trees (DT) and Random Forests (RF)

A decision tree is a supervised learning technique, primarily used for classification tasks, but can also be used for regression. A decision tree begins with a root node, the first decision point for splitting the dataset, and contains a single feature that best splits the data into their respective classes (Fig. 2.6.). Each split has an edge that connects either to a new decision node that contains another feature to further split the data into homogenous groups or to a terminal node that predicts the class. This process of separating data into two binary partitions is known as recursive partitioning. A random forest is an extension of this method, known as an ensemble method, that produces multiple decision trees. Rather than using every feature to create every decision tree in a random forest, a subsample of features is used to create each decision tree. Trees then predict a class outcome, and the majority vote among trees is used as the model's final class prediction. The models generated by this method of classification have high accuracy and speed in building the model, and can be applied to multi-category data [144].

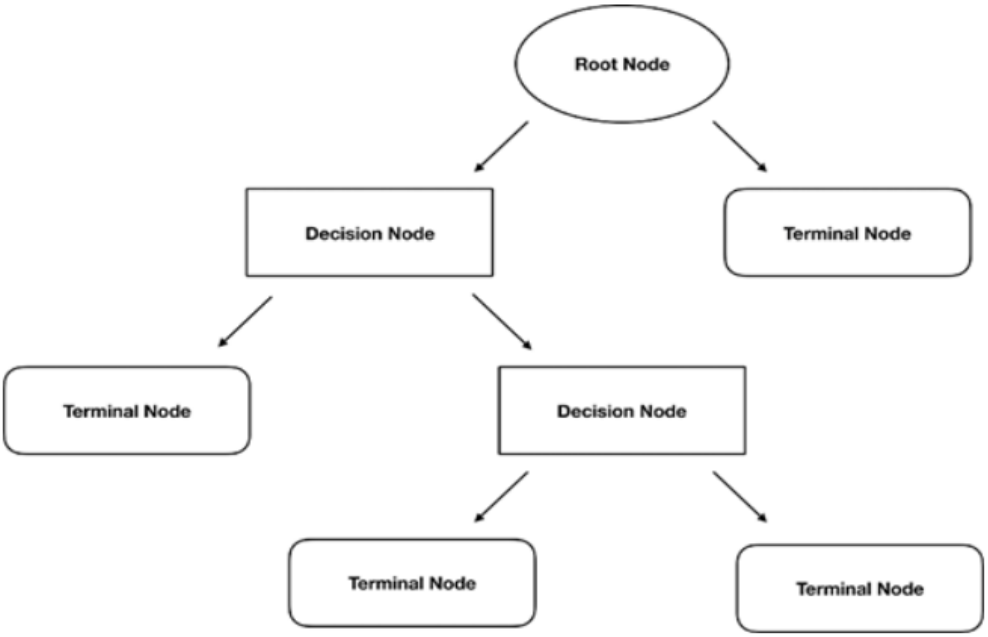

Figure 2.6. Structure of a decision tree [144]

2.3.4. Confusion Matrix

It is a table that records the number of instances in a dataset that fall in a particular category. The class label in a binary training set can take two possible values, which are called as positive class and a negative class. As seen in (Fig. 2.7.), the number of positive and negative instances that a classifier predicts correctly is called True Positives (TP) and True Negatives (TN), respectively. The misclassified instances are known as False Positives (FP) and False Negatives (FN). A learning model involuntarily decides which two classes in the dataset is the positive class. If the class labels of a dataset are strings, first the label is sorted alphabetically and at the first level, it is chosen as a negative class, while in the second level it is chosen as positive class. If the class labels are Boolean or integer in nature, then "1" or "true" labeled instances are assigned as positive class.

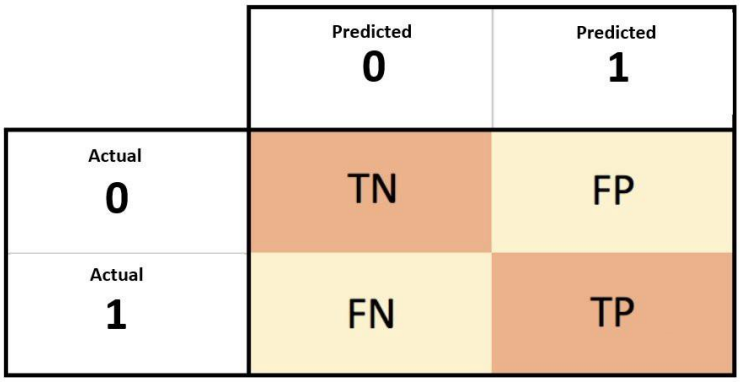

Figure 2.7. Typical 2X2 confusion matrix

2.3.5. Precision and Recall

Precision is the fraction of relevant instances among the retrieved instances.

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

Recall (opposite of precision) is not so much about answering questions correctly but more about answering all questions that have answer "true" with the answer "true." Therefore, if the models always answer "true," it is 100% recall.

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

It is important to note that measures precision and recall do not provide any information on the number of true negatives. This means it can be the case that a person has lousy precision and recall score (e.g., 50%) and still answered 99.99% of all questions correctly.

2.3.6. *F* Measure

In statistical analysis of binary classification, the *F* score (or *F* measure) is a metric of a test's accuracy. It takes into consideration the precision and the recall of the test to compute

its score. *F* measure is the harmonic average of precision and recall. *F* measure reaches its best value at 1 (perfect precision and recall) and worst at 0.

$$F\ Measure = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

2.3.7. Hyperparameter optimization

In order to tune the model parameters some techniques like grid search and cross validation are used to get to the model that gives us the best results. Tuned models are also matched with the data mining goals to see if we are able to get the desired results as well as performance. Model tuning is also termed as hyperparameter optimization in the Machine Learning world.

- **Grid Search**

Grid search is a hyperparameter optimization technique used to find the best combination of hyperparameter values for a machine learning algorithm. In most machine learning models, there are certain hyperparameters that are not learned during the training process but need to be set before training. These hyperparameters can significantly impact the performance of the model.

Grid search works by creating a grid of possible hyperparameter values for each hyperparameter of the model. It then exhaustively tries all possible combinations of these hyperparameter values and evaluates the model's performance using a performance metric (e.g., accuracy, F1 score, etc.) on a validation dataset. The combination of hyperparameter values that yields the best performance is selected as the optimal set of hyperparameters.

The grid search method is straightforward and systematic but can be computationally expensive, especially when dealing with a large number of hyperparameters or a wide range of possible values for each hyperparameter.

- **k-fold Cross-Validation**

K-fold cross-validation is a model evaluation technique used to assess the performance of a machine learning model and mitigate overfitting. It involves partitioning the original training dataset into "k" equally sized subsets (or folds). The model is then trained and evaluated "k" times, using a different fold as the validation set in each iteration and the remaining "k-1" folds as the training set.

The process can be summarized as follows:

a. Divide the training data into "k" subsets (folds). b. For each fold "i" (where "i" ranges from 1 to k), treat it as the validation set, and train the model on the remaining k-1 folds. c. Evaluate the model's performance on the validation set. d. Repeat steps b and c for all "k" folds, using a different fold as the validation set in each iteration. e. Calculate the average performance metric across all "k" iterations to obtain a more robust and reliable estimate of the model's performance.

K-fold cross-validation helps in obtaining a more generalizable evaluation of the model's performance because it uses different subsets of data for training and validation in each iteration. It also helps to make efficient use of the available data, especially when the dataset is limited in size. Common choices for the value of "k" are 5 or 10, but it can vary depending on the size of the dataset and computational resources available.

Chapter 3 – Reference Studies

3.1. First Study

Machine Learning and Deep Learning Approaches in Breast Cancer Survival Prediction Using Clinical Data

A study conducted by Kalafi et al. [145] at University of Malaya Medical Centre, Malaysia (2019). The primary objective of this study was to enhance breast cancer treatment protocols by accurately predicting patient survival prospects. To achieve this, the researchers aimed to compare the effectiveness of traditional machine learning models and more advanced deep learning algorithms in predicting breast cancer survival based on clinical data. The dataset employed in this study consisted of 4,902 patient records obtained from the University of Malaya Medical Centre Breast Cancer Registry. It included relevant clinical features such as marital status, menopausal status, presence of family history, race, methods of diagnosis, classification of breast cancer, laterality, cancer stage classification, grade of differentiation in tumor, oestrogen receptor (ER) status, progesterone receptor (PR) status, c-er-b2 status, primary treatment type, surgery status, type of surgery, method of axillary lymph node dissection, radiotherapy, chemotherapy, hormonal therapy, status (dead or alive), age, axillary lymph node, positive lymph nodes and tumor size. The researchers explored the application of various algorithms to predict breast cancer survivability using the clinical dataset. The algorithms evaluated in this study were: Multilayer Perceptron (MLP), Random Forest (RF), Decision Tree (DT) and Support Vector Machine (SVM). The results indicated that the multilayer perceptron (MLP), random forest (RF) and decision tree (DT) classifiers could predict survivorship, respectively, with 88.2 %, 83.3 % and 82.5 % accuracy in the tested samples. Support vector machine (SVM) came out to be lower with 80.5 %. In this study, tumor size turned out to be the most important feature for breast cancer survivability prediction. The study's results highlight the potential of machine learning and deep learning in predicting breast cancer survival, with the multilayer perceptron proving to be the most accurate. Further exploration of these methods could lead to enhanced treatment protocols and improved patient outcomes.

3.2. Second Study

Predicting factors for survival of breast cancer patients using machine learning techniques

The study was conducted by Ganggayah *et al.* [146], and was published in the BMC Medical Informatics and Decision Making journal in 2019. The researchers aimed to predict survival indicators for breast cancer patients using advanced machine learning methods instead of traditional statistical approaches. To carry out the study, the researchers collected a large dataset from the University Malaya Medical Centre in Kuala Lumpur, Malaysia. The dataset included information from 1993 to 2016 and consisted of 8,066 breast cancer cases. Among the data, there were 23 predictor variables (factors that could influence survival) and one dependent variable, which represented the survival status of the patients (whether they were alive or deceased). The researchers utilized several machine learning algorithms to build prediction models for detecting and visualizing significant prognostic indicators of breast cancer survival rate. The algorithms they used included decision tree, random forest, neural

networks, extreme boost, logistic regression, and support vector machine. To perform more advanced modeling, the dataset was clustered based on the receptor status of breast cancer patients, which was identified using immunohistochemistry. This allowed the researchers to refine the analysis and get more accurate results. Next, they ranked the important variables by using variable selection methods within the random forest algorithm. This step helped them identify the most critical factors influencing the survival rate of breast cancer patients. The results of the study showed that all the machine learning algorithms used in the analysis produced very similar outcomes in terms of both model accuracy and calibration measure. The decision tree had the lowest accuracy at 79.8%, while random forest had the highest accuracy at 82.7%. The important prognostic factors influencing the survival rate of breast cancer, as identified in the study, included cancer stage classification, tumor size, the number of total axillary lymph nodes removed, the number of positive lymph nodes, types of primary treatment, and methods of diagnosis. The study demonstrated that various machine learning algorithms, particularly in the Asian region, could be used as effective alternative predictive tools in breast cancer survival studies. The important factors influencing survival, validated by survival curves, are valuable and could be translated into decision support tools in the medical domain.

3.3. Third Study

Machine learning models in breast cancer survival prediction

The study was conducted by Montazeri *et al.*, and was published in the journal "Technology and Health Care" in 2016. The objective of this study was to develop an accurate and reliable system for early breast cancer diagnosis to improve survival rates. Breast cancer is one of the most common cancers with a high mortality rate among women, and early diagnosis is crucial for better outcomes. The researchers proposed a model that combined rules and various machine learning techniques to predict different types of breast cancer survival. Machine learning models are powerful tools that can analyze patterns and relationships within a large dataset of cases, enabling them to make predictions based on historical cases. For their analysis, the researchers used a dataset containing records of 900 patients, among whom 876 were females (97.3%) and 24 were males (2.7%). They employed several machine learning techniques such as Naive Bayes (NB), Trees Random Forest (TRF), 1-Nearest Neighbor (1NN), AdaBoost (AD), Support Vector Machine (SVM), RBF Network (RBFN), and Multilayer Perceptron (MLP) with a 10-cross fold technique. The performance of each machine learning technique was evaluated using various metrics, including accuracy, precision, sensitivity, specificity, and area under the Receiver Operating Characteristic (ROC) curve. The results of the study indicated that the Trees Random Forest (TRF) technique outperformed the other machine learning methods, including NB, 1NN, AD, SVM, RBFN, and MLP. TRF achieved an accuracy of 96%, sensitivity of 96%, and an area under the ROC curve of 93%. On the other hand, the 1NN machine learning technique exhibited poor performance with an accuracy of 91%, sensitivity of 91%, and an area under the ROC curve of 78%. The study highlighted that the Trees Random Forest (TRF) model, which is a rule-based classification model, provided the best results with the highest accuracy. Therefore, the researchers recommended TRF as a valuable tool for breast cancer survival prediction and medical decision-making.

3.4. Fourth Study

Machine learning predicts the prognosis of breast cancer patients with initial bone metastases

The study was carried out by Chaofan Li et al., and was published in Frontiers in Public Health in September 2022. The researchers obtained data for their analysis from the SEER (Surveillance, Epidemiology, and End Results) database covering the period from 2010 to 2019. They performed COX regression analysis to identify prognostic factors in breast cancer patients with bone metastases (BMBC). Using cross-validation, they constructed an XGBoost model for predicting survival in BMBC patients. Additionally, the study investigated the prognosis of patients treated with neoadjuvant chemotherapy plus surgical intervention compared to those treated with chemotherapy alone using propensity score matching and Kaplan-Meier survival analysis. The validation results of the XGBoost model showed high sensitivity, specificity, and correctness, making it the most accurate model for predicting the survival of BMBC patients. The area under the curve (AUC) for 1-year survival was 0.818, for 3-year survival was 0.798, and for 5-year survival was 0.791. Notably, patients with BMBC who started therapy ≥1 month after diagnosis had even better survival than those who began treatment immediately. The study also found that patients with higher income (≥USD$70,000) had better overall survival (OS) and breast cancer-specific survival (BCSS) compared to those with lower income (< USD$50,000). Furthermore, neoadjuvant chemotherapy plus surgical treatment significantly improved OS and BCSS in various molecular subtypes of BMBC patients, particularly in those with bone metastases only, bone and liver metastases, and bone and lung metastases. The researchers developed an artificial intelligence (AI) model that provides a quantitative method for predicting the survival of BMBC patients. The model's validation results indicated high reproducibility in a similar patient population. The study also identified potential prognostic factors for BMBC patients and suggested that primary surgery followed by neoadjuvant chemotherapy could improve survival in selected subgroups of patients. This study contributes valuable insights into the use of machine learning techniques for predicting breast cancer survival, specifically in patients with bone metastases.

Aim of the study

The aim of the research is to find algorithms with high accuracy and sensitivity capable of predicting breast cancer prognosis and the cause of death in the study sample based on many variables including: the patient's age at the time of diagnosis of the tumor, the state of menopause, the type of breast cancer based on histological examination of the tumor tissue, tumor size, lymph nodes, estrogen receptors (ER), progesterone receptors (PR), human epidermal growth factor receptors (HER2), the expression of a number of genes, some single nucleotide polymorphisms (SNPs) in a number of genes, in addition to clinical outcomes, including: information about patient survival and relapse, time and cause of death, time and state of relapse, in order to be able to intervene quickly in the patient's treatment protocol to reduce mortality as much as possible.

Practical Review

Chapter 4 - Methods and Materials

4.1. Study Sample

The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) database is a Canada-UK Project which contains targeted sequencing data of 1904 primary breast cancer samples. The dataset was downloaded from Kaggle and collected by Professor Carlos Caldas from Cambridge Research Institute and Professor Sam Aparicio from the British Columbia Cancer Centre in Canada and published on Nature Communications [147].

4.2. Clinical attributes in the dataset

4.2.1. Age at diagnosis

Age of the patient at diagnosis time.

4.2.2. Type of breast surgery

Breast cancer surgery type: 1- MASTECTOMY, which refers to a surgery to remove all breast tissue from a breast as a way to treat or prevent breast cancer. 2- BREAST CONSERVING, which refers to a surgery where only the part of the breast that has cancer is removed**.**

4.2.3. Cancer type detailed

Detailed Breast cancer types: 1- Breast Invasive Ductal Carcinoma 2- Breast Mixed Ductal and Lobular Carcinoma 3- Breast Invasive Lobular Carcinoma 4- Breast Invasive Mixed Mucinous Carcinoma 5- Metaplastic Breast Cancer.

4.2.4. Cellularity

Cancer cellularity post chemotherapy, which refers to the amount of tumor cells in the specimen and their arrangement into clusters.

4.2.5. Chemotherapy

Whether or not the patient had chemotherapy as a treatment (yes/no).

4.2.6. Pam50 + Claudin-low subtype

Pam 50: is a tumor profiling test that helps show whether some estrogen receptor-positive (ER-positive), HER2-negative breast cancers are likely to metastasize (when breast cancer spreads to other organs). The claudin-low breast cancer subtype is defined by gene expression characteristics, most prominently: Low expression of cell–cell adhesion genes, high expression of epithelial–mesenchymal transition (EMT) genes, and stem cell-like/less differentiated gene expression patterns.

4.2.7. ER status measured by IHC

To assess if estrogen receptors are expressed on cancer cells by using immune-histochemistry (a dye used in pathology that targets specific antigen, if it is there, it will give a color, it is not there, the tissue on the slide will be colored) (positive/negative).

4.2.8. ER status

Cancer cells are positive or negative for estrogen receptors.

4.2.9. Neoplasm histologic grade

Determined by pathology by looking the nature of the cells, do they look aggressive or not (It takes a value from 1 to 3).

4.2.10. HER2 status measured by SNP6

To assess if the cancer positive for HER2 or not by using advance molecular techniques (Type of next generation sequencing).

4.2.11. HER2 status

Whether the cancer is positive or negative for HER2.

4.2.12. Tumor other histologic subtype

Type of the cancer based on microscopic examination of the cancer tissue (It takes a value of 'Ductal/NST', 'Mixed', 'Lobular', 'Tubular/ cribriform', 'Mucinous', 'Medullary', 'Other', 'Metaplastic').

4.2.13. Hormone therapy

Whether or not the patient had hormonal as a treatment (yes/no).

4.2.14. Inferred menopausal state

Whether the patient is post-menopausal or pre-menopausal (post/pre).

4.2.15. Primary tumor laterality

Whether it is involving the right breast or the left breast.

4.2.16. Lymph nodes examined positive

Samples of the lymph node taken during the surgery and see if there were involved by the cancer.

4.2.17. Mutation count

Number of gene that has relevant mutations.

4.2.18. Nottingham prognostic index

It is used to determine prognosis following surgery for breast cancer. Its value is calculated using three pathological criteria: the size of the tumor; the number of involved lymph nodes; and the grade of the tumor.

4.2.19. PR status

Cancer cells are positive or negative for progesterone receptors.

4.2.20. Radio therapy

Whether or not the patient had radio as a treatment (yes/no).

4.2.21. Three Gene classifier subtype

Three Gene classifier subtype takes a value from 'ER-/HER2-', 'ER+/HER2- High Prolif', nan, 'ER+/HER2- Low Prolif','HER2+'.

4.2.22. Tumor size

Tumor size measured by imaging techniques.

4.2.23. Tumor stage

Stage of the cancer based on the involvement of surrounding structures, lymph nodes and distant spread.

4.2.24. Overall survival months

Duration from the time of the intervention to death.

4.2.25. Death from cancer

Whether the patient's death was due to cancer or not (yes/no).

4.2.26. Genetic attributes in the dataset

The genetics part of the dataset contains m-RNA levels z-score for 331 genes, and mutation for 175 genes. For mRNA expression data, the calculations of the relative expression of an individual gene and tumor to the gene's expression distribution in a reference population is done. That reference population is all samples in the study. The returned value indicates the number of standard deviations away from the mean of expression in the reference population (Z-score). This measure is useful to determine whether a gene is up- or down-regulated relative to the normal samples or all other tumor samples.

4.3. Practical study

The work was divided into two steps.

4.3.1. First Step

In the first step, we utilized SPSS Statistics 25.0 to characterize the study variables and analyze the significant relationships between them. This software allowed us to perform statistical analyses and explore the data in detail.

To explore the relationships between categorical variables, the Chi-Square Test was employed, while the t-test was utilized to analyze the relationship between dependent quantitative variables and independent categorical variables.

In all the conducted tests, statistical significance was observed at the predetermined significance threshold ($P<0.05$). This suggests that there are meaningful differences between the variables being studied. The obtained results indicate that the relationships between the variables are not likely to occur due to chance and are instead indicative of genuine associations.

4.3.1. Second Step

During the second step, we employed Google Colab, a software platform designed for writing scientific code in Python, to develop various models. In this phase, we worked with the data that had been processed in the first step. To handle missing values, we replaced them with the most frequent values in the dataset. Additionally, we performed data standardization, a technique used to ensure that variables with different scales do not disproportionately influence the results. By standardizing the data, all variables were transformed to a common scale with a mean of 0 and a standard deviation of 1.

To implement the models, we utilized several important Python libraries. The most significant libraries are reviewed in (Table 4.1.):

Table 4.1. The most important libraries used in the project.

Library	Description
NumPy	NumPy is one of the most widely used open-source Python libraries, focusing on scientific computation. It features built-in mathematical functions for quick computation and supports big matrices and multidimensional data. “Numerical Python” is defined by the term

	"NumPy." It can be used in linear algebra, as a multi-dimensional container for generic data, and as a random number generator, among other things. In Python, NumPy Array is preferred over lists because it takes up less memory and is faster and more convenient to use.
Pandas	Pandas are an important library for data scientists. It is an open-source machine learning library that provides flexible high-level data structures and a variety of analysis tools. It eases data analysis, data manipulation, and cleaning of data. Pandas support operations like Sorting, Re-indexing, Iteration, Concatenation, Conversion of data, Visualizations, Aggregations, etc.
SKlearn	It is a famous Python library to work with complex data. Scikit-learn is an open-source library that supports machine learning. It supports variously supervised and unsupervised algorithms like linear regression, classification, clustering, etc.
Seaborn	Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

The work followed a methodology in which the data was divided into a training set comprising 75% of the total data, and a test set comprising 25% of the total data.

Data splitting is an important step in statistical analysis and modeling. The training set is used to build statistical models and determine relationships between variables, while the test set is used to evaluate the performance of these models and their ability to predict unknown data.

By using a 75:25 ratio for the data split, a balance is achieved between having a sufficiently large training set for building strong models and a smaller test set that reflects the diversity and represents the overall data. This type of split helps objectively assesses model performance and provide an accurate estimation of their predictive ability on new data.

Chapter 5 – Results and Discussion

5.1. Results of first step

5.1.1. Descriptive Statistics

The average age of female patients at the time of cancer diagnosis is 61.09 years, with a standard deviation of 12.98, as shown in (Table 5.1.):

Table 5.1. Average age of breast cances patients

	N	Range	Minimum	Maximum	Mean	Std. Deviation
Age at diagnosis	1904	74.36	21.93	96.29	61.0871	12.97871

The distribution of age among breast cancer patients is shown in (Fig. 5.1.):

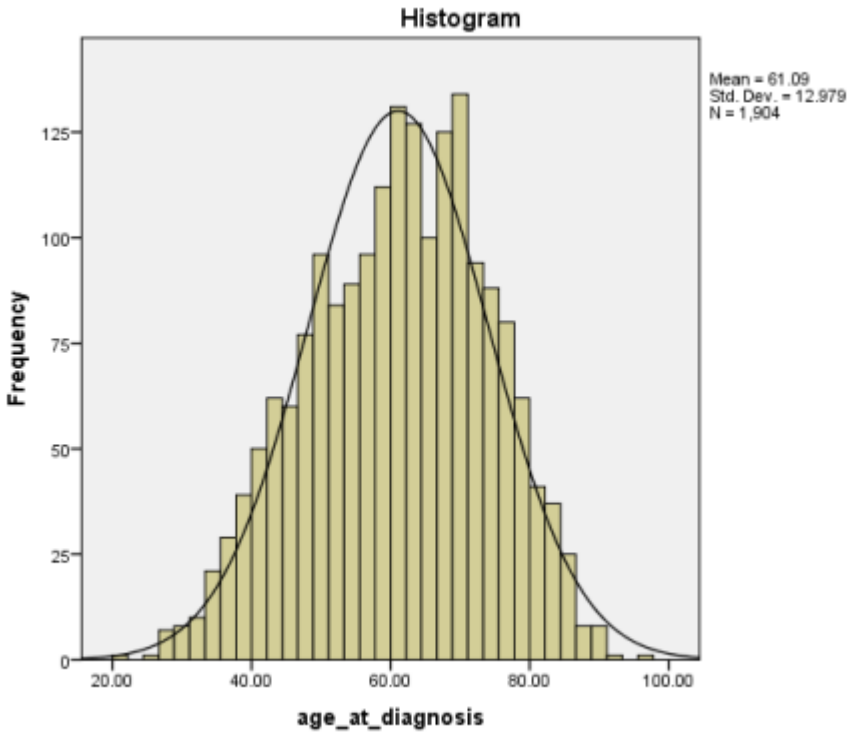

Figure 5.1. Breast cancer patients age distribution

Among the patients, the most prevalent cancer type is Breast Invasive Ductal Carcinoma, accounting for 78.9% of the cases. Following that, we have Breast Mixed Ductal and Lobular Carcinoma, representing 11.0% of the cases, and Breast Invasive Lobular Carcinoma, comprising 7.6% of the cases. Table 5.2. and Fig. 5.2. provide comprehensive insights into the distribution of patients across cancer types:

Table 5.2. Patients distribution according to the breast cancer type

Cancer Type Detailed	Frequency	Valid Percent
Breast Invasive Ductal Carcinoma	1503	78.9
Breast Mixed Ductal and Lobular Carcinoma	210	11.0
Breast Invasive Lobular Carcinoma	145	7.6
Breast Invasive Mixed Mucinous Carcinoma	22	1.2
Breast	20	1.1
Metaplastic breast Cancer	4	.2
Total	1904	100.0

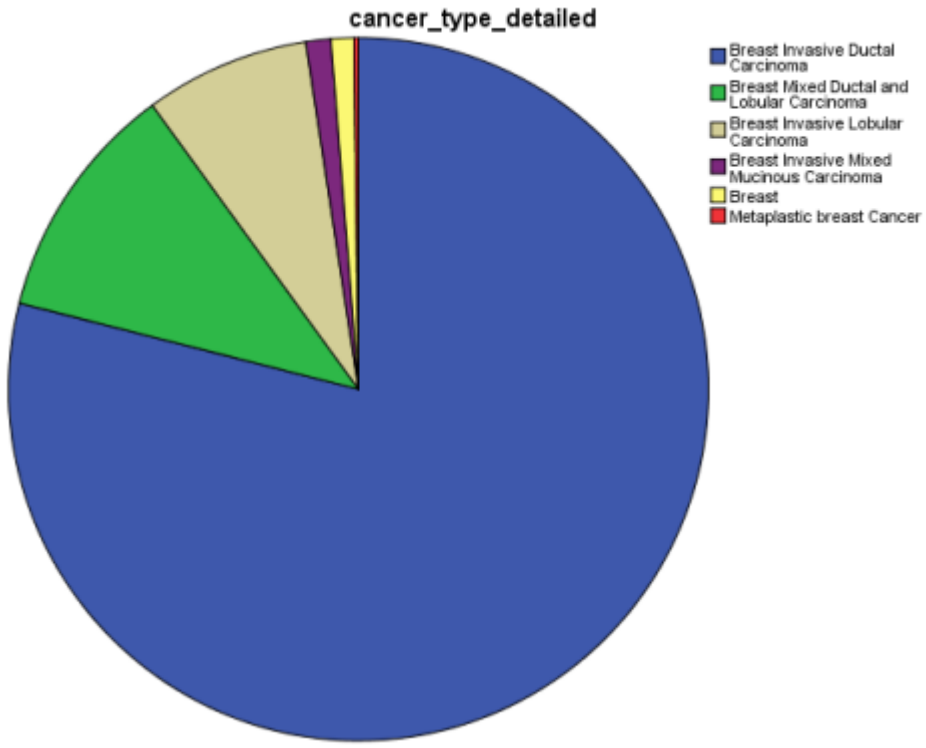

Figure 5.2. Patients distribution according to the breast cancer type

According to the data presented in (Table 5.3.), among the patients analyzed, the majority (59.9%) underwent mastectomy, a surgical procedure involving the removal of the breast tissue. On the other hand, a significant proportion of patients (40.1%) had breast-conserving surgery, a less invasive approach that aims to remove only the tumor while preserving the

breast. Fig. 5.3. presents a comprehensive analysis of the distribution of patients based on the type of surgery they underwent

Table 5.3. Distribution of patients based on the type of surgery they underwent

:

	Frequency	Valid Percent
MASTECTOMY	1127	59.9
BREAST CONSERVING	755	40.1
Total	1882	100.0

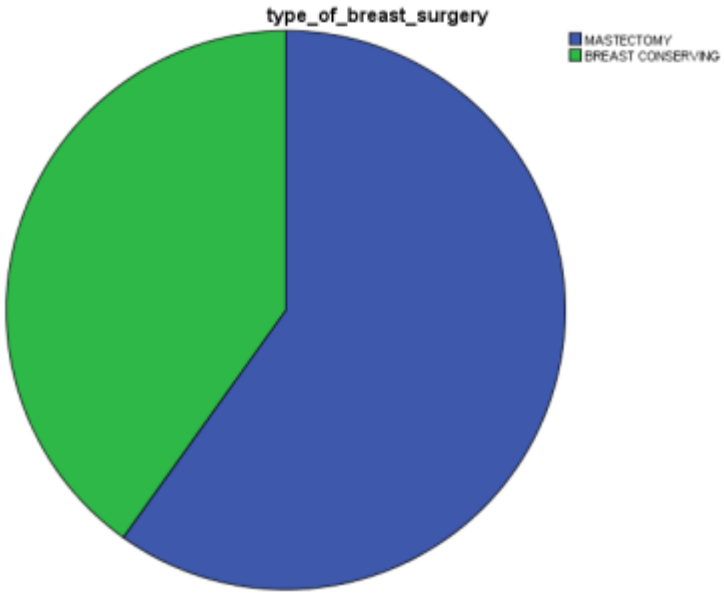

Figure 5.3. Distribution of patients based on the type of surgery they underwent

According to the data presented in (Table 5.4.), among the patients analyzed, the majority (79.2%) did not undergo chemotherapy, while a significant proportion (20.8%) received chemotherapy as part of their treatment plan.

Table 5.4. Percentage of patients who received chemotherapy

Chemotherapy	Frequency	Valid Percent
No	1508	79.2
Yes	396	20.8
Total	1904	100.0

Figure 5.4. provide a comprehensive analysis of the distribution of patients based on whether they underwent chemotherapy. These visual representations offer valuable insights into the frequency and proportion of patients who received chemotherapy as part of their treatment.

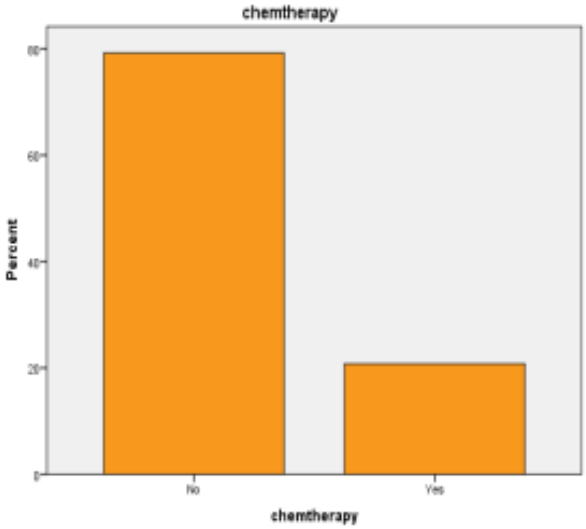

Figure 5.4. Distribution of patients based on whether they underwent chemotherapy

According to the data presented in (Table 5.5.), among the patients included in the analysis, a majority (61.7%) received hormone therapy as part of their treatment regimen. On the other hand, a proportion of (38.3%) did not undergo hormone therapy. Figure 5.5. provide a comprehensive analysis of the distribution of patients based on whether they underwent hormone therapy.

Table 5.5. Distribution of patients based on whether they underwent hormone therapy

Hormone therapy	Frequency	Valid Percent
No	730	38.3
Yes	1174	61.7
Total	1904	100.0

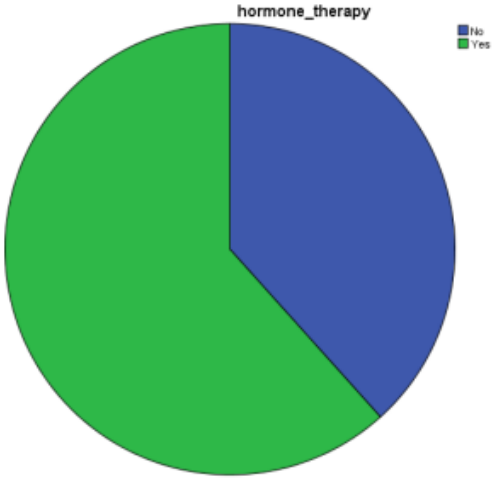

Figure 5.5. Distribution of patients based on whether they underwent hormone therapy

According to the data presented in (Table 5.6.), among the patients included in the analysis, a majority (59.7%) received radiotherapy as part of their treatment regimen, while a significant proportion (40.3%) did not undergo radiotherapy. The corresponding Figure 5.6. complements the analysis by visually presenting the distribution of patients in relation to their radiotherapy status.

Table 5.6. Distribution of patients in relation to their radiotherapy status

Radio therapy	Frequency	Valid Percent
0	767	40.3
1	1137	59.7
Total	1904	100.0

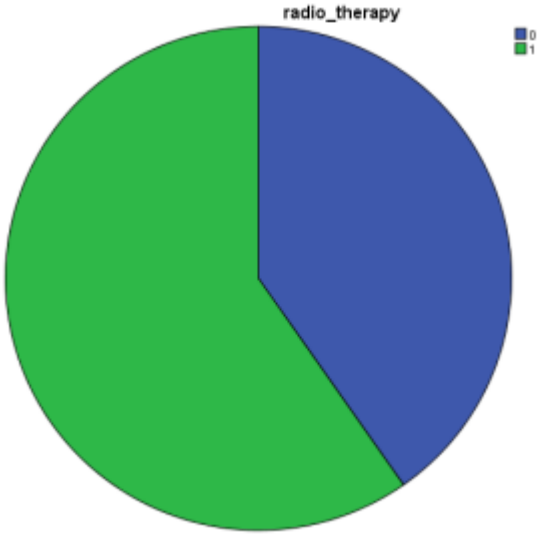

Figure 5.6. Distribution of patients in relation to their radiotherapy status

According to the statistical data provided in (Table 5.7.), the mean overall survival is calculated to be 125.1213235 months (10.4 years), indicating the average duration of survival among the patients. The standard deviation of 76.33414829 (6.4 years) signifies the spread or dispersion of survival times around the mean. This suggests that there is a degree of variability in the length of survival experienced by patients.

Table 5.7. Patients overall survival months

Overall survival months	
Mean	125.1213235
Std. Deviation	76.33414829
Minimum	.00000
Maximum	355.20000

According to the information provided in (Table 5.8.), among the patients included in the analysis, a significant proportion (57.9%) unfortunately passed away due to the disease. On the other hand, a considerable number of patients (42.1%) are categorized as "Living," indicating that they have survived at the time of data collection. Figure 5.7. offers a comprehensive analysis of the distribution of patients based on their outcomes in terms of mortality.

Table 5.8. Distribution of patients based on their outcomes in terms of mortality

	Frequency	Valid Percent
Living	801	42.1
Died of Disease	1102	57.9
Total	1903	100.0

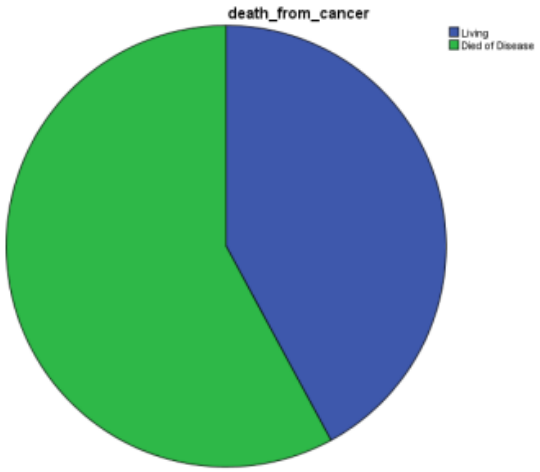

Figure 5.7. Distribution of patients based on their outcomes in terms of mortality

According to the information provided in (Table 5.9.), among the patients included in the analysis, approximately half of them (47.0%) tested negative for progesterone receptors. Conversely, the remaining patients (53.0%) tested positive for progesterone receptors. Figure 5.8. presents a comprehensive analysis of the distribution of patients based on the status of progesterone receptors.

Table 5.9. Distribution of patients based on the status of progesterone receptors

PR Status	Frequency	Valid Percent
Negative	895	47.0
Positive	1009	53.0
Total	1904	100.0

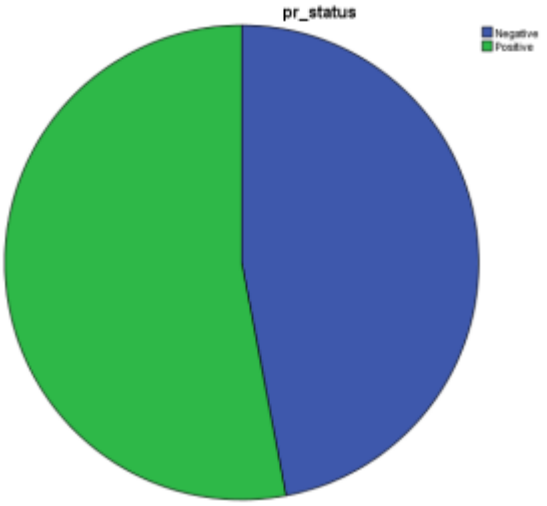

Figure 5.8. Distribution of patients based on the status of progesterone receptors

According to the information provided in Table (5.10.), among the patients included in the analysis, a significant proportion (50.8%) exhibited high cancer cellularity. Additionally, a considerable number of patients (38.4%) displayed moderate cellularity, while a smaller portion (10.8%) showed low cellularity. Figure 5.9. offers a comprehensive analysis of the distribution of patients based on the cancer cellularity.

Table 5.10. Distribution of patients based on the cancer cellularity

Cellularity	Frequency	Valid Percent
Low	200	10.8
Moderate	711	38.4
High	939	50.8
Total	1850	100.0

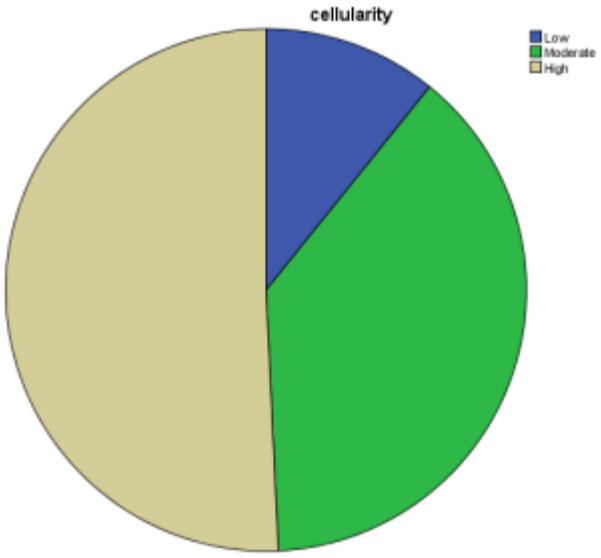

Figure 5.9. Distribution of patients based on the cancer cellularity

According to the information presented in (Table 5.11.), among the patients included in the analysis, the most prevalent subtype is LumA, accounting for 35.7% of the cases. Following that, we have LumB (24.2%), claudin-1 (10.5%), basal (10.5%), Her2 (11.6%) and normal (7.4%). Figure 5.10. provide a comprehensive analysis of the distribution of patients based on their breast cancer subtypes.

Table 5.11. Distribution of patients based on their breast cancer subtypes

Subtype	Frequency	Valid Percent
claudin-1	199	10.5
LumA	679	35.7
LumB	460	24.2
Her2	220	11.6
Normal	140	7.4
Basal	200	10.5
NC	6	.3
Total	1904	100.0

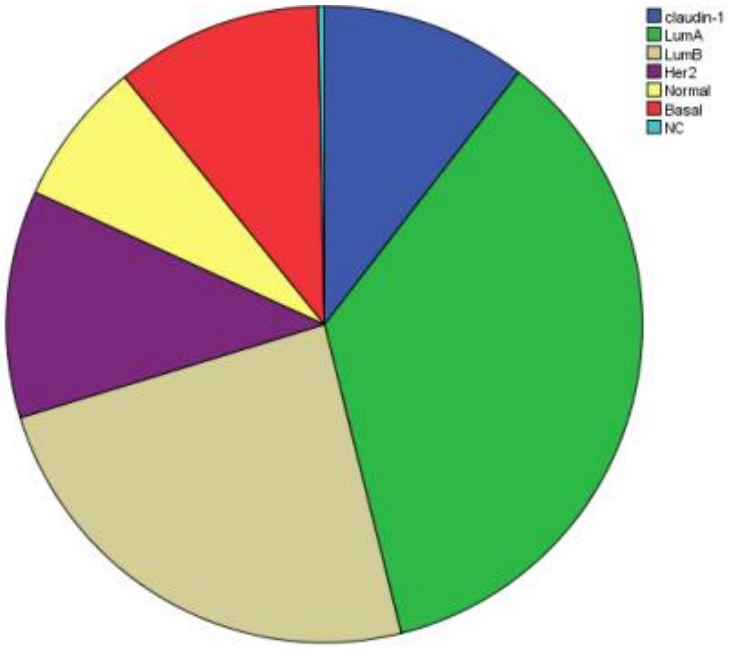

Figure 5.10. Distribution of patients based on their breast cancer subtypes

Understanding the distribution of tumor other histologic subtypes facilitates personalized treatment plans, as different subtypes may require specific interventions or targeted therapies. It also aids in evaluating disease progression, assessing the efficacy of treatment modalities, and predicting patient outcomes. According to the information provided in (Table 5.12.), among the patients included in the analysis, the most prevalent histologic subtype is Ductal/NST, accounting for 76.5% of the cases. The next most common subtype is Mixed, comprising 11.0% of the cases. Following that, we have Lobular (7.6%), Mucinous (1.3%), Medullary (1.5%), Tubular/cribriform (1.2%), and Other (.9%). These histologic subtypes represent different microscopic characteristics of the cancer tissue. Figure 5.11. present a comprehensive analysis of the distribution of patients based on tumor other histologic subtype.

Table 5.12. Distribution of patients based on tumor other histologic subtype

	Frequency	Percent	Valid Percent
Ductal/NST	1456	76.5	76.5
Mixed	210	11.0	11.0
Lobular	145	7.6	7.6
Mucinous	25	1.3	1.3
Medullary	28	1.5	1.5
Tubular/ cribriform	22	1.2	1.2
Other	18	.9	.9
Total	1904	100.0	100.0

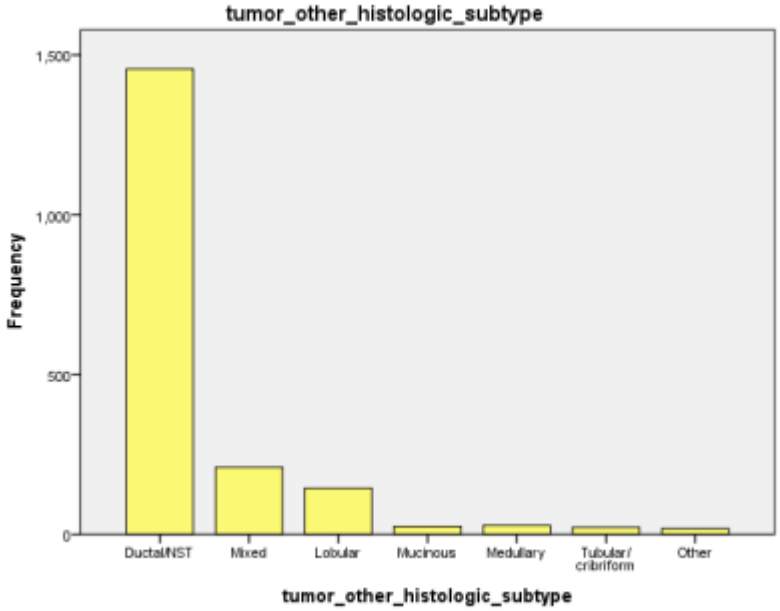

Figure 5.11. Distribution of patients based on tumor other histologic subtype

According to the information provided in (Table 5.13.), among the patients included in the analysis, a significant proportion (77.1%) tested positive for estrogen receptors, indicating the

presence of estrogen receptor expression. Conversely, a smaller percentage of patients (22.9%) tested negative for estrogen receptors. Figure 5.12. presents a comprehensive analysis of the distribution of patients based on the measurement of estrogen receptors using immune-histochemistry (IHC).

Table 5.13. Distribution of patients based on the measurement of estrogen receptors using immune-histochemistry (IHC)

ER status measured by IHC	Frequency	Valid Percent
Negative	429	22.9
Positive	1445	77.1
Total	1874	100.0

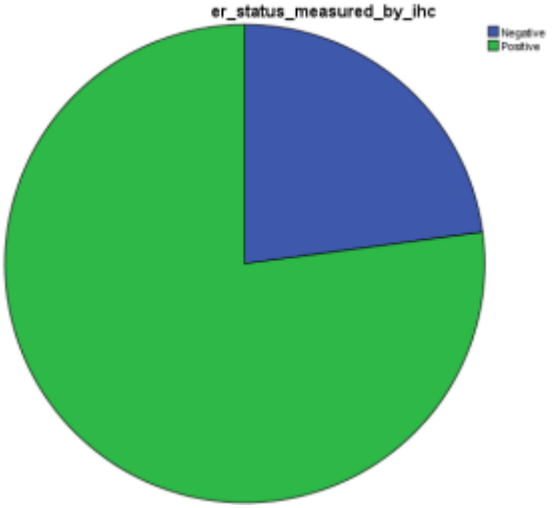

Figure 5.12. Distribution of patients based on the measurement of estrogen receptors using immune-histochemistry (IHC)

According to the information provided in (Table 5.14.), among the patients included in the analysis, approximately 76.6% tested positive for estrogen receptors, indicating the presence of estrogen receptor expression. Conversely, a smaller proportion of patients (23.4%) tested negative for estrogen receptors.

In comparison to (Table 5.13.), which measured estrogen receptor status using immune-histochemistry (IHC), we observe that the proportion of patients testing positive for estrogen receptors remains relatively similar. However, the percentage of patients testing negative for estrogen receptors is slightly higher in this analysis.

Table 5.14. Distribution of patients based on the measurement of estrogen receptors

ER status	Frequency	Valid Percent
Negative	445	23.4
Positive	1459	76.6

Neoplasm grade, determined by pathology, provides insights into the nature and aggressiveness of the cells observed in the tumor tissue. According to the information provided in (Table 5.15.), among the patients included in the analysis, 9.0% were classified as Grade 1, indicating a lower level of aggressiveness. A larger proportion of patients (40.4%) were assigned Grade 2, indicating intermediate aggressiveness. The highest proportion of patients (50.6%) were assigned Grade 3, representing the highest level of aggressiveness. Figure 5.13. offers a comprehensive analysis of the distribution of patients based on neoplasm histologic grade.

Table 5.15. Distribution of patients based on neoplasm histologic grade

Neoplasm histologic grade	Frequency	Valid Percent
1	165	9.0
2	740	40.4
3	927	50.6
Total	1832	100.0

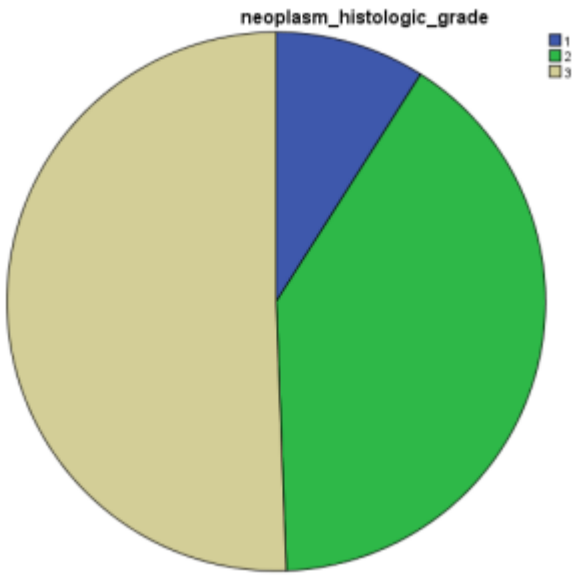

Figure 5.13. Distribution of patients based on neoplasm histologic grade

According to the information provided in Table (5.16.), among the patients included in the analysis, the majority (72.6%) were classified as HER2 status "NEUTRAL," indicating a normal or neutral HER2 gene amplification status. Additionally, a substantial number of patients (21.9%) showed HER2 status "GAIN," suggesting HER2 gene amplification. A smaller proportion of patients (5.3%) exhibited HER2 status "LOSS," indicating a loss of the HER2 gene. Furthermore, a negligible percentage of patients (0.2%) were categorized as "UNDEF" due to undefined or inconclusive HER2 status.

Table 5.16. Distribution of patients based on the assessment of HER2 status using advanced molecular techniques

HER2 status measured by snp6	Frequency	Valid Percent
LOSS	100	5.3
NEUTRAL	1383	72.6
GAIN	417	21.9
UNDEF	4	.2
Total	1904	100.0

According to the information provided in (Table 5.17.), among the patients included in the analysis, a majority (87.6%) tested negative for HER2. On the other hand, a smaller proportion of patients (12.4%) tested positive for HER2. Figure 5.14. present a comprehensive analysis of the distribution of patients based on their HER2 status.

Table 5.17. Distribution of patients based on their HER2 status

HER2 status	Frequency	Valid Percent
Negative	1668	87.6
Positive	236	12.4
Total	1904	100.0

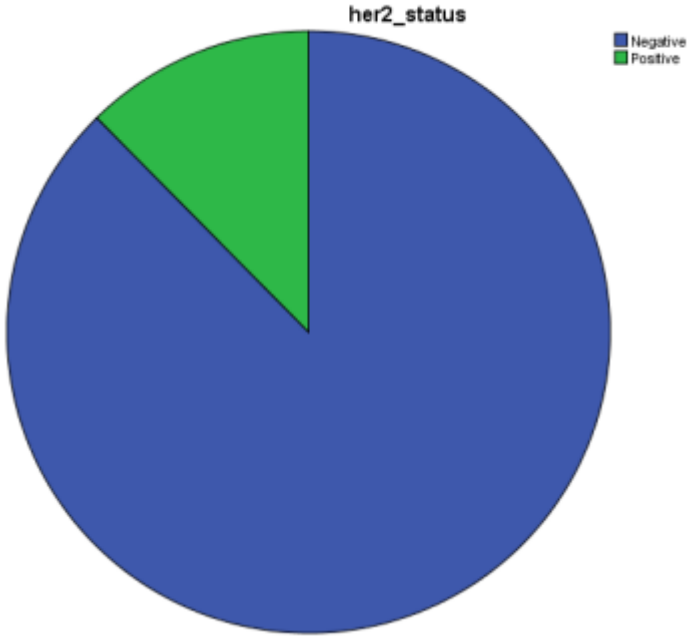

Figure 5.14. Distribution of patients based on their HER2 status

According to the information provided in (Table 5.18.), among the patients the majority (78.4%) were classified as post-menopausal. This indicates that these patients have gone through menopause, a natural biological process where menstrual periods cease, and reproductive hormone levels decline. On the other hand, a smaller proportion of patients (21.6%) were classified as pre-menopausal, indicating that they have not yet experienced menopause. Figure 5.15. presents a comprehensive analysis of the distribution of patients based on their inferred menopausal state.

Table 5.18. Distribution of patients based on their inferred menopausal state

Inferred menopausal state	Frequency	Valid Percent
Pre	411	21.6
Post	1493	78.4
Total	1904	100.0

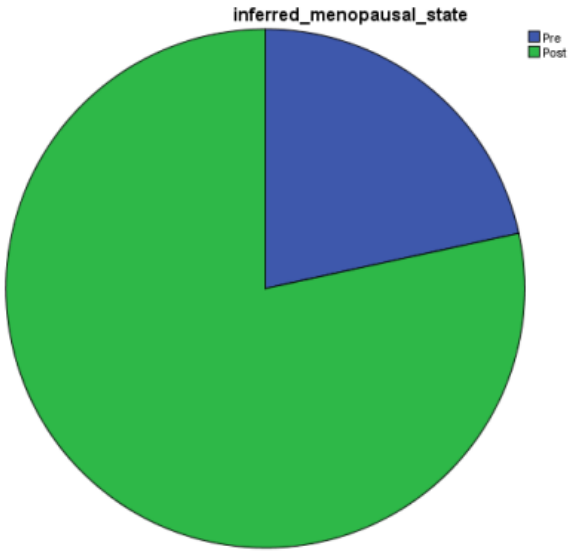

Figure 5.15. Distribution of patients based on their inferred menopausal state

Analyzing the distribution of primary tumor laterality can provide insights into potential asymmetries or patterns related to breast cancer occurrence. The information provided in (Table 5.19.) depicts that 48.0% of the patients had tumors involving the right breast, while a slightly higher proportion (52.0%) had tumors involving the left breast. Figure 5.16. illustrates the distribution of patients based on primary tumor laterality.

Table 5.19. Distribution of patients based on primary tumor laterality

Primary tumor laterality	Frequency	Valid Percent
Right	863	48.0
Left	935	52.0
Total	1798	100.0

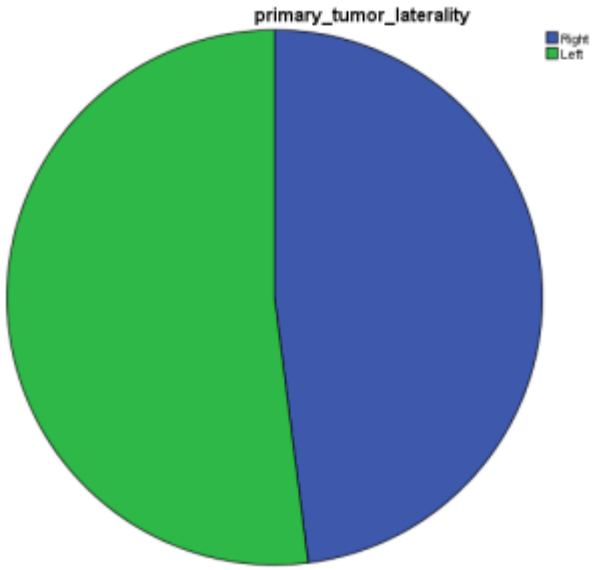

Figure 5.16. Distribution of patients based on primary tumor laterality

Tumor stage is a critical factor in cancer management, as it informs treatment decisions and contributes to optimizing patient care and outcomes as it reflects the extent of cancer involvement in surrounding structures, lymph nodes, and distant spread. According to the information provided in (Table 5.20.), among the patients included in the analysis, the majority (57.0%) were classified as Stage 2, indicating a significant extent of cancer involvement. Additionally, a substantial number of patients (33.9%) were categorized as Stage 1, representing localized cancer with limited spread. A smaller proportion of patients were classified as Stage 3 (8.2%), indicating regional spread, and an even smaller percentage were categorized as Stage 4 (.6%), signifying distant metastasis. Furthermore, there were a few cases (0.3%) categorized as Stage 0, representing carcinoma in situ or non-invasive cancer. Figure 5.17. presents a comprehensive analysis of the distribution of patients based on tumor stage.

Table 5.20. Distribution of patients based on tumor stage

Tumor stage	Frequency	Valid Percent
0	4	.3
1	475	33.9
2	800	57.0
3	115	8.2
4	9	.6

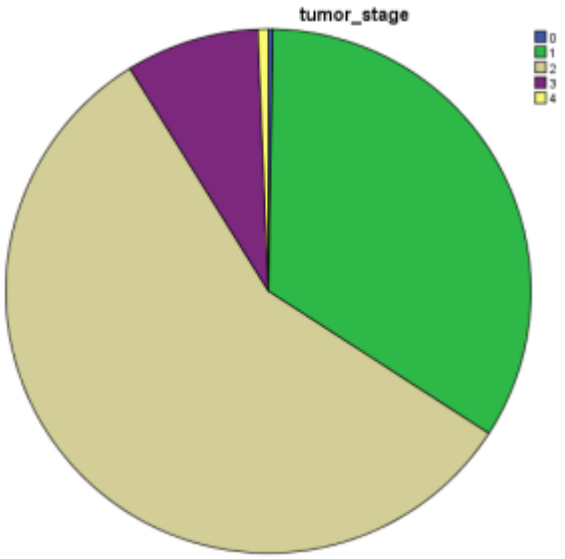

Figure 5.17. Distribution of patients based on tumor stage

The average expressions of all genes present in the sample were calculated in order to identify the genes most implicated in breast cancer (those with the highest and lowest expressions). Subsequently, these genes were ranked in descending order as presented in (Table 5.21.):

Table 5.21. Average expressions of the highest and lowest expressed genes present in the sample

Mutation	N	Minimum	Maximum	Mean
ahnak	**1904**	**-5.19820-**	**3.32900**	**-.0000027-**
cyp17a1	**1904**	**-2.85210-**	**6.53450**	**-.0000023-**
gldc	**1904**	**-.96670-**	**6.34670**	**-.0000018-**
nfkb1	**1904**	**-4.56350-**	**3.82130**	**-.0000018-**
kmt2d	**1904**	**-4.00810-**	**4.63600**	**-.0000016-**
psen2	**1904**	**-3.43050-**	**3.78140**	**.0000018**
jag1	**1904**	**-3.00580-**	**7.05010**	**.0000019**
pdgfrb	**1904**	**-4.71410-**	**2.69510**	**.0000020**
rad51c	**1904**	**-3.25670-**	**4.40540**	**.0000022**
nfkb2	**1904**	**-2.49350-**	**5.89230**	**.0000023**

It was observed that the genes with the lowest expressions were: ahnak, cyp17a1, gldc, nfkb1, and kmt2d. On the other hand, the genes with the highest expressions were psen2, jag1, pdgfrb, rad51c, and nfkb2.

5.1.2. Descriptive Statistics of Quantitative Variables

According to the information provided in (Table 5.22.), descriptive statistics were obtained for several variables related to breast cancer. These statistics offer valuable insights into different aspects of the disease within the study population.

First, we have the "Mutation Count," which represents the number of genes with relevant mutations in breast cancer patients. The range of mutation counts observed in the study is from 1 to 80, with a mean of 5.70. This indicates that, on average, patients in the study exhibit mutations in approximately 5 to 6 relevant genes.

Moving on to the "Nottingham Prognostic Index," this index is used to determine the prognosis following surgery for breast cancer. It is calculated based on three pathological criteria: tumor size, the number of involved lymph nodes, and the grade of the tumor. In this study, the index ranges from 5.36 to 6.36, with a mean of 4.0330187. These statistics highlight the variability in prognostic index scores within the study population.

Next, we have "Tumor Size," which refers to the size of the tumor measured using imaging techniques. The range of tumor sizes observed in the study is from 1.00 to 182.00, with a mean tumor size of 26.2387. These statistics demonstrate the variation in tumor sizes among the patients, ranging from very small (1 mm) to much larger (up to 182 mm) tumors.

Lastly, we examine "Lymph Nodes Examined Positive," which represents the number of lymph nodes examined during surgery to assess if they are involved by the cancer. The range of lymph nodes examined in the study is from 0 to 45, with a mean of 2.00. This data suggests that, on average, a small number of lymph nodes are examined, but there is considerable variation in the number of lymph nodes examined among the patients.

Table 5.22. Descriptive Statistics of Quantitative Variables (Mutation count, lymph nodes examined, Nottingham prognostic index and tumor size)

	N	Range	Minimum	Maximum	Mean
Mutation count	1859	79	1	80	5.70
Lymph nodes examined	1904	45	0	45	2.00
Nottingham prognostic index	1904	5.36000	1.00000	6.36000	4.0330187
Tumor size	1884	181.00	1.00	182.00	26.2387

5.1.3. Inferential Statistics

Initially, a comparison was made between the means of lymph nodes examined, mutation count, and tumor size among individuals who died from cancer and those who survived. From (Table 5.23.), we observe a statistically significant relationship between all the variables studied and death. These findings indicate that a higher number of lymph nodes examined,

higher mutation count, and larger tumor size are associated with a higher risk of death from cancer. High lymph node ratio was significantly associated with short overall survival and disease-free survival in breast cancer patients after neoadjuvant chemotherapy [148]. Notably, 5–10% of all breast cancer patients are genetically predisposed to cancers. Although the most common breast cancer susceptibility genes are *BRCA1* and *BRCA2*, which are also associated with the risk of developing ovarian and pancreatic cancer, advances in next-generation sequencing (NGS) analysis technology enabled the discovery of several non-*BRCA* genes responsible for breast and ovarian cancers [149]. Studies have shown that breast cancer mortality increases from 6.9% for tumours 1–10 mm in size to 60.4% for tumours 91–100 mm in size [150]. This highlights the importance of these factors in predicting prognosis and underscores the need for comprehensive assessment and management of these variables in cancer patients.

Table 5.23. Comparison between the means of lymph nodes examined, mutation count, and tumor size among individuals who died from cancer and those who survived

	Death from cancer	N	Mean	Std. Deviation	Sig. (2-tailed)
Lymph nodes examined	Living	801	1.21	2.721	
	Died of Disease	1102	2.58	4.755	.000
Mutation count	Living	771	5.32	3.343	
	Died of Disease	1087	5.96	4.477	.001
Tumor size	Living	794	23.3199	13.06156	.000
	Died of Disease	1089	28.3772	16.20347	

The relationship between HER2 status measured by snp6 and HER2 status with the breast cancer detailed type was examined as shown in (Table 5.24. and Table 5.25.). We observe a statistically significant association between the subtype of breast cancer and both variables. These findings suggest that the HER2 status measured by snp6 and HER2 status are closely related to the subtype of breast cancer. Understanding these associations can provide valuable insights into the molecular characteristics and potential treatment strategies for different subtypes of breast cancer.

Table 5.24. Relationship between HER2 status measured by snp6 and breast cancer detailed type

		HER2 status measured by snp6				Total	Asymptotic Significance (2-sided)
		LOSS	NEUTRAL	GAIN	UNDEF		
	Breast Invasive Ductal Carcinoma	82	1048	370	3	1503	.000

Cancer type detailed	Breast Mixed Ductal and Lobular Carcinoma	9	180	20	1	210	
	Breast Invasive Lobular Carcinoma	6	120	19	0	145	
	Breast Invasive Mixed Mucinous Carcinoma	0	18	4	0	22	
	Breast	3	13	4	0	20	
	Metaplastic breast Cancer	0	4	0	0	4	
Total		100	1383	417	4	1904	

Table 5.25. Relationship between HER2 status and breast cancer detailed type

		HER2 status		Total	
		Negative	Positive		Asymptotic Significance (2-sided)
Cancer type detailed	Breast Invasive Ductal Carcinoma	1287	216	1503	.000
	Breast Mixed Ductal and Lobular Carcinoma	203	7	210	
	Breast Invasive Lobular Carcinoma	136	9	145	
	Breast Invasive Mixed Mucinous Carcinoma	20	2	22	
	Breast	18	2	20	
	Metaplastic breast Cancer	4	0	4	
Total		1668	236	1904	

The relationship between breast cancer detailed type and HER2 status is depicted in (Figure 5.18.):

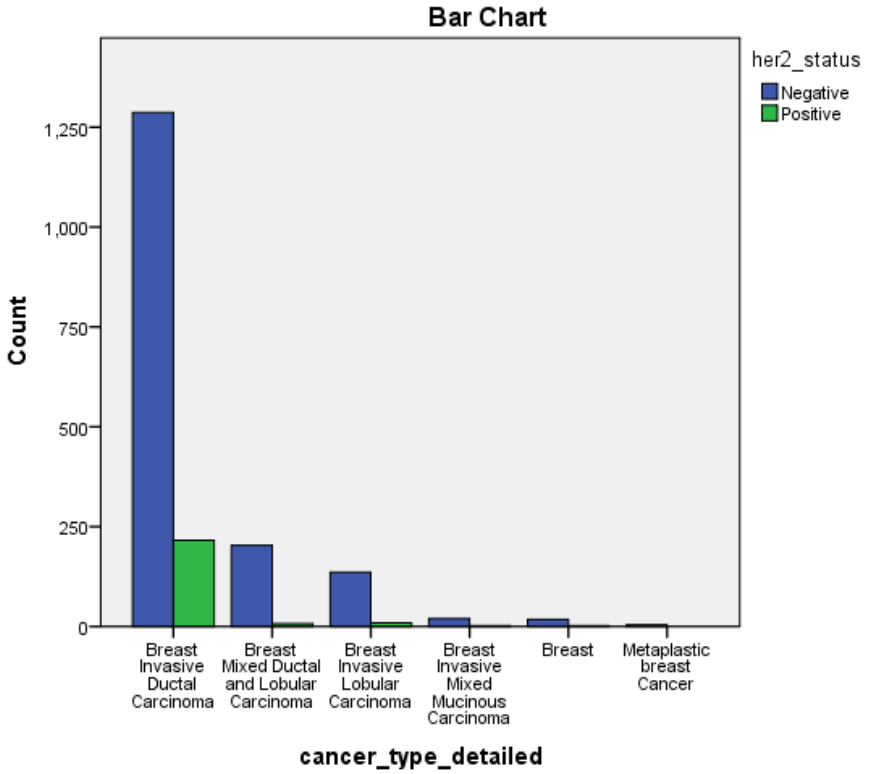

Figure 5.18. Relationship between HER2 status and breast cancer detailed type

The following (Table 5.26. and Table 5.27.) show the distribution of ER status based on two different measurement methods: ER status measured by IHC and ER status in the breast cancer detailed type. The observed counts indicate a significant association between the breast cancer type and both ER status variables. Studies have shown that ER status in mixed breast carcinomas, which consist of both ductal and lobular components, showed associations with lower grade and ER positivity compared to pure invasive ductal carcinoma (IDC) [151].

Table 5.26. Relationship between ER status measured by IHC and breast cancer detailed type

		ER status measured by IHC		Total	Asymptotic Significance (2-sided)
		Negative	Positive		
Cancer type detailed	Breast Invasive Ductal Carcinoma	390	1089	1479	
	Breast Mixed Ductal and Lobular Carcinoma	16	190	206	.000
	Breast Invasive Lobular Carcinoma	16	129	145	
	Breast Invasive Mixed Mucinous Carcinoma	1	19	20	

	Breast	5	15	20	
	Metaplastic breast Cancer	1	3	4	
Total		429	1445	1874	

Table 5.27. Relationship between ER status and breast cancer detailed type

		ER status		Total	Asymptotic Significance (2-sided)
		Negative	Positive		
Cancer type detailed	Breast Invasive Ductal Carcinoma	403	1100	1503	
	Breast Mixed Ductal and Lobular Carcinoma	14	196	210	.000
	Breast Invasive Lobular Carcinoma	20	125	145	
	Breast Invasive Mixed Mucinous Carcinoma	1	21	22	
	Breast	6	14	20	
	Metaplastic breast Cancer	1	3	4	
Total		445	1459	1904	

The relationship between breast cancer type detailed and ER status is depicted in (Figure 5.19.):

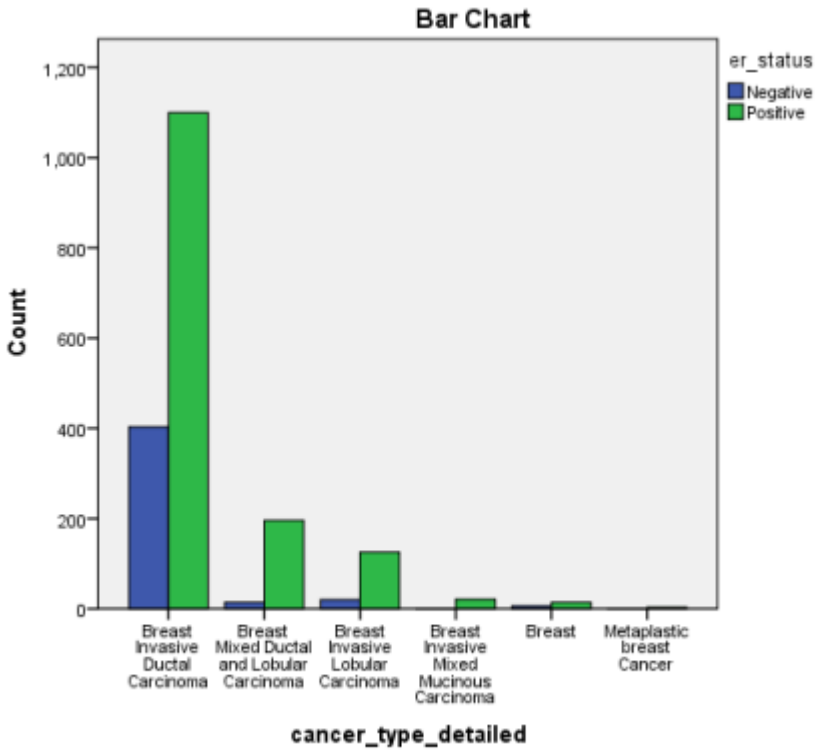

Figure 5.19. Relationship between HER2 status and breast cancer detailed type

The relationship between treatment modality (chemotherapy, radiotherapy, hormonal therapy) and occurrence of death was investigated in this study. The analysis examined the association between death from cancer and each treatment modality separately.

Table 5.28. reveals a statistically significant association between death from cancer and the use of chemotherapy (p = 0.048). It indicates that among the patients who received chemotherapy, a higher proportion (212 out of 396) died of the disease compared to those who did not receive chemotherapy (890 out of 1507). This suggests that chemotherapy may have an impact on patient survival, potentially indicating its effectiveness in treating certain types of cancer.

Table 5.28. Association between death from cancer and the use of chemotherapy

		Death from cancer		Total	
		Living	Died of Disease		Sig.
Chemotherapy	No	617	890	1507	
	Yes	184	212	396	.048
Total		801	1102	1903	

In contrast to the results for chemotherapy, Table 5.29. does not show a statistically significant association between death from cancer and hormonal therapy (p = 0.176). The number of deaths from disease is relatively similar between patients who received hormonal therapy (694 out of 1174) and those who did not (408 out of 729). This suggests that hormonal therapy may not have a significant impact on patient survival in the studied population.

Table 5.29. Association between death from cancer and hormonal therapy

		Death from cancer		Total	Sig.
		Living	Died of Disease		
Hormone therapy	No	321	408	729	
	Yes	480	694	1174	.176
Total		801	1102	1903	

Table 5.30. demonstrates a statistically significant association between death from cancer and the use of radiotherapy (p < 0.001). It shows that a higher proportion of patients who received radiotherapy (606 out of 1136) died of the disease compared to those who did not receive radiotherapy (496 out of 767). This finding suggests that radiotherapy may have implications for patient survival, potentially indicating its effectiveness in targeting and treating cancer cells.

Table 5.30. Association between death from cancer and the use of radiotherapy

		Death from cancer		Total	Sig.
		Living	Died of Disease		
Radio therapy	No	271	496	767	
	Yes	530	606	1136	.000
Total		801	1102	1903	

Overall, these findings highlight the importance of considering different treatment modalities in cancer management. While chemotherapy and radiotherapy show a significant association with death from cancer, the lack of a significant association with hormonal therapy suggests the need for further investigation and potential refinement of treatment approaches. The results emphasize the complexity of cancer treatment and the importance of individualized approaches based on patient-specific factors and tumor characteristics.

The present study investigated the relationship between occurrence of death and different histologic subtypes of tumors. Specifically, the association between death from cancer and tumor subtypes was examined. Table 5.31. displayed significant statistical associations between death from cancer and tumor subtypes. The crosstabulation analysis revealed that the histologic subtypes of Ductal/NST, Mixed, Lobular, Mucinous, Medullary, Tubular/Cribriform, and Other were all significantly associated with the occurrence of death (p = 0.004). The findings of this study suggest that the histologic subtypes of breast cancer play a role in determining patient outcomes. The significant associations observed between specific tumor subtypes and death from cancer indicate that the histologic characteristics of the tumor may influence disease progression and patient prognosis.

Table 5.31. Association between death from cancer and tumor subtypes

		Tumor other histologic subtypes							Total	Sig.
		Ductal/ NST	Mixed	Lobular	Mucinous	Medullary	Tubular/ cribriform	Other		
Death from cancer	Living	609	77	59	15	13	16	12	801	
	Died of Disease	846	133	86	10	15	6	6	1102	.004
Total		1455	210	145	25	28	22	18	1903	

This study aimed to investigate the relationship between death occurrence and the combined expression of Pam50 and Claudin-low breast cancer subtypes. The Pam50 test is utilized for profiling tumors to determine the likelihood of metastasis in certain estrogen receptor-positive (ER-positive), HER2-negative breast cancers. The Claudin-low subtype is characterized by specific gene expression patterns, including low expression of cell-cell adhesion genes, high expression of epithelial-mesenchymal transition (EMT) genes, and stem cell-like/less differentiated gene expression patterns. Table 5.32. demonstrates a statistically significant association between death occurrence and the combined expression of Pam50 and Claudin-low subtypes. The crosstabulation analysis revealed that the different subtypes, including Claudin-1, LumA, LumB, Her2, Normal, Basal, and NC, were all significantly associated with death from cancer (p = 0.000).

Table 5.32. Relationship between death occurrence and the combined expression of Pam50 and Claudin-low breast cancer subtypes

		Pam50 and Claudin-low							Total	Sig.
		Claudin-low	LumA	LumB	Her2	Normal	Basal	NC		
Death from cancer	Living	110	315	158	65	64	88	1	801	.000
	Died of Disease	89	363	302	155	76	112	5	1102	
Total		199	678	460	220	140	200	6	1903	

This study aimed to investigate the relationship between cancer cellularity, which represents the amount and arrangement of tumor cells in the specimen, and death occurrence in breast cancer patients. Cellularity is an important histopathological feature that can provide insights into tumor characteristics and behavior. Understanding its association with patient outcomes can contribute to risk stratification and treatment decision-making. Table 5.33. demonstrates that there was no statistically significant association between cancer cellularity and death occurrence (p = 0.351). The analysis included three categories of cellularity: low, moderate, and high. The frequencies of cellularity categories were assessed in relation to the outcome of living or death from disease. The findings of this study suggest that cancer cellularity alone may not be a strong predictor of death occurrence in breast cancer patients. The lack of a statistically significant association between cellularity and death suggests that other factors may have a greater influence on patient outcomes. The lack of a significant association between cancer cellularity and death occurrence emphasizes the complex nature of breast cancer and the need for a multi-dimensional approach to prognostication.

Table 5.33. Relationship between cancer cellularity and death occurrence in breast cancer patients

		Cellularity			Total	Sig.
		Low	Moderate	High		
Death from cancer	Living	92	288	393	773	
	Died of Disease	107	423	546	1076	.351
Total		199	711	939	1849	

The objective of this study was to examine the correlation between death occurrence and neoplasm histologic grade in breast cancer patients. Neoplasm histologic grade is determined through pathological examination and provides valuable information about the nature and aggressiveness of tumor cells. Table 5.34. reveals a statistically significant association between death occurrence and neoplasm histologic grade (p = 0.001). The analysis included three categories of histologic grade: grade 1, grade 2, and grade 3. The frequencies of each grade were evaluated in relation to the outcome of living or death from disease. Neoplasm histologic grade is a crucial factor in assessing tumor aggressiveness and predicting patient prognosis. It takes into account various histopathological features, including cellular differentiation, nuclear morphology, and mitotic activity. Higher histologic grades indicate more poorly differentiated and aggressive tumors, which are associated with an increased risk of disease progression and poorer outcomes.

Table 5.34. Relationship between death occurrence and neoplasm histologic grade

		Neoplasm histologic grade			Total	Sig.
		1	2	3		
Death from cancer	Living	90	326	363	779	
	Died of Disease	75	414	563	1052	.001
Total		165	740	926	1831	

We also explored the association between death and the laterality of the primary tumor in breast cancer patients. Breast cancer can occur in either the right or left breast, and the laterality of the tumor may have implications for disease progression and patient outcomes. By analyzing the data, we investigated whether there was a statistically significant relationship between death and the side of the primary tumor. The results of the study (Table 5.35.) revealed a significant association between death and primary tumor laterality. Specifically, higher mortality rates were observed among patients with tumors on the left side compared to those with tumors on the right side. This finding suggests that the location of the primary tumor may influence the prognosis and survival outcomes of breast cancer patients.

Table 5.35. Association between death and primary tumor laterality

		Primary tumor laterality		Total	Sig.
		Right	Left		
Death from cancer	Living	389	379	768	
	Died of Disease	473	556	1029	.028
Total		862	935	1797	

In this study we investigated the relationship between gene expression and the occurrence of death in a cohort of breast cancer patients. Gene expression levels of various genes (Table 5.36.), including BRCA1, BRCA2, TP53, RAD51C, MYC, JAG1, PSEN2, EGFR, NFKB1, NFKB2, PDGFRB, KMT2D, AHNAK, GLDC, and CYP17A1, were analyzed and compared between individuals who survived and those who died from the disease.

The results from the analysis revealed significant associations between gene expression levels and death. Specifically, elevated expression of BRCA1, MYC, NFKB1, NFKB2, PSEN2, and GLDC was associated with a higher risk of death. On the other hand, decreased expression of TP53 and RAD51C showed a potential association with increased mortality risk. However, the gene expressions of BRCA2, JAG1, EGFR, PDGFRB, KMT2D, AHNAK, and CYP17A1 did not show statistically significant relationships with death.

These findings suggest that altered gene expression patterns may play a role in determining the likelihood of death in individuals with the studied disease. The identified genes, such as BRCA1, MYC, NFKB1, NFKB2, PSEN2, TP53, and RAD51C, may have prognostic value and could serve as potential targets for further research and therapeutic interventions.

Table 5.36. Relationship between gene expression and the occurrence of death

	Death from cancer	N	Mean	Std. Deviation	Sig.
brca1	Living	801	-.0870728-	.98962620	.001
	Died of Disease	1102	.0639975	1.00376769	
brca2	Living	801	-.0000714-	.97952294	.966
	Died of Disease	1102	.0019047	1.01409246	

tp53	Living	801	.0682577	1.01488439	.011
	Died of Disease	1102	-.0502289-	.98721811	
rad51c	Living	801	.0422346	1.02976551	.123
	Died of Disease	1102	-.0293313-	.97704535	
myc	Living	801	.1807097	.96279760	.000
	Died of Disease	1102	-.1310606-	1.00734321	
jag1	Living	801	-.0139633-	1.02431332	.602
	Died of Disease	1102	.0102417	.98321299	
psen2	Living	801	-.0683052-	1.01190176	.011
	Died of Disease	1102	.0499275	.98962245	
egfr	Living	801	.0522915	.95902584	.053
	Died of Disease	1102	-.0377404-	1.02837382	
nfkb1	Living	801	.0868330	.98927834	.001
	Died of Disease	1102	-.0638754-	1.00400612	
nfkb2	Living	801	-.0721767-	.95121987	.007
	Died of Disease	1102	.0522275	1.03213504	
pdgfrb	Living	801	.0039082	1.05040963	.889
	Died of Disease	1102	-.0025799-	.96308435	
kmt2d	Living	801	.0426152	1.00797426	.115
	Died of Disease	1102	-.0306751-	.99434699	
ahnak	Living	801	.0050106	1.04245154	.845
	Died of Disease	1102	-.0040929-	.96926062	
gldc	Living	801	.0561856	1.09406828	.038
	Died of Disease	1102	-.0403789-	.92489353	
cyp17a1	Living	801	-.0218456-	.99139914	.436
	Died of Disease	1102	.0143554	1.00601786	

5.2. Results of second step

After completing the statistical analysis, Google Colab was utilized to build various models. The following libraries were utilized: numpy, pandas, matplotlib.pyplot, and seaborn. The final dataset consists of 1904 samples with 31 features and a single binary output variable (Survival status: Living or Died of disease).

To handle missing values, we replaced them with the most frequent values using the SimpleImputer library, and the following code was employed for this purpose:

```
from sklearn.impute import SimpleImputer
imputer= SimpleImputer(missing_values=np.nan, strategy="most_frequent")
imputer.fit(df.iloc[:,:9])
df.iloc[:,:9]=imputer.transform(df.iloc[:,:9])
```

Table 5.37. represents the variables that have missing values and the count of these missing values in each row:

Table 5.37. Input missing values

Variable	Number of missing values
Type of breast surgery	22
Cancer type detailed	4
Cellularity	54
ER status measured by IHC	30
Primary Tumor Laterality	106
Mutation count	45
Three Gene Classifier Subtype	204
Tumor size	20
Tumor stage	501
All remaining variables	0

A heatmap was generated to illustrate the relationships among the study variables using the following code:

```
plt.figure(figsize= (18,18))
sns.heatmap(df.iloc[:,:31].corr(),annot=True)
```

Figure 5.20. illustrates the heatmap, which provides visual insights into the strength and direction of the relationships between the variables.

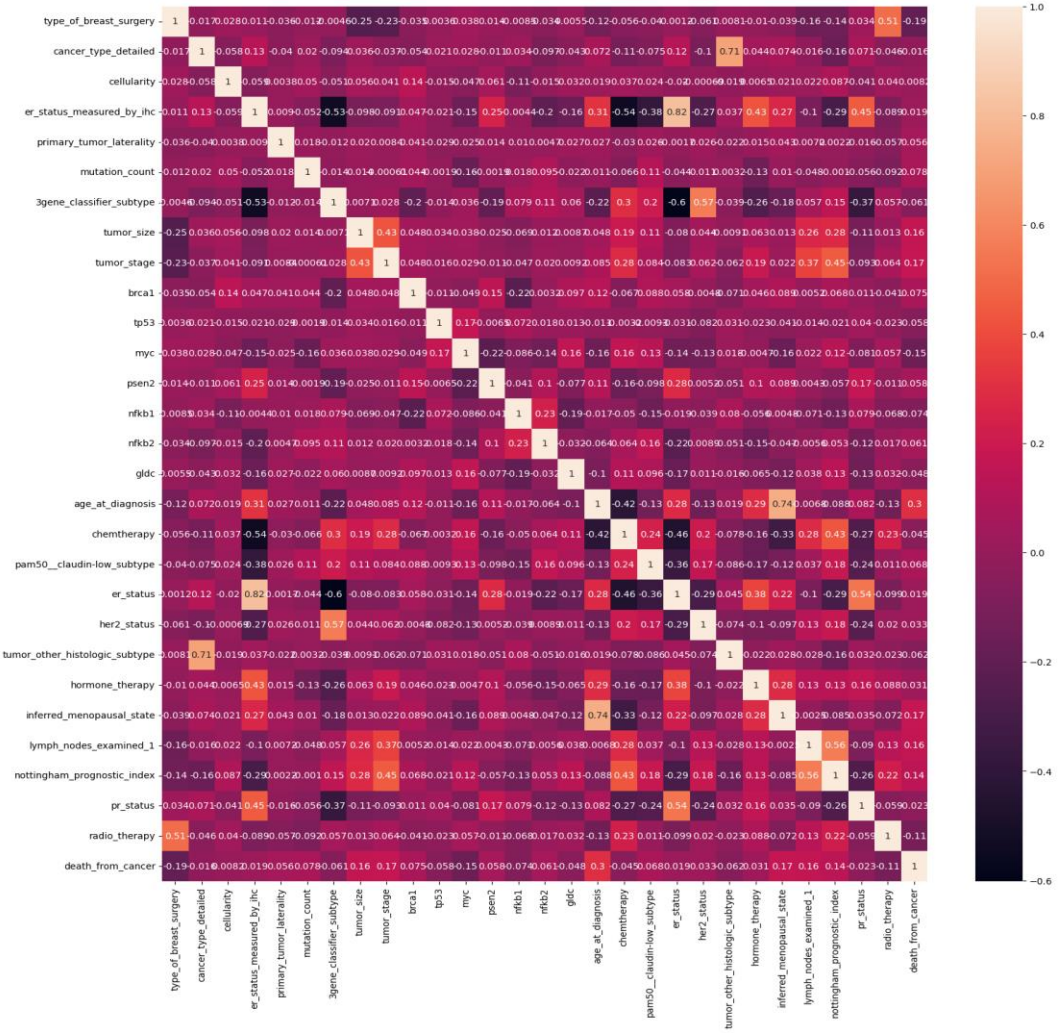

Figure 5.20. Heatmap of studied variables

Data was divided into a training set comprising 75% of the total data, and a test set comprising 25% of the total data using the following code:

```
from sklearn.model_selection import train_test_split

X_train,X_test, y_train,y_test= train_test_split(X,y,test_size=0.25, random_state= 0)
```

Data standarizition was performed, and the following code was employed for this purpose:

```
from sklearn.preprocessing import StandardScaler

sc= StandardScaler()


X_train= sc.fit_transform(X_train)

X_test= sc.fit_transform(X_test)
```

All machine learning algorithms demonstrated varying degrees of accuracy in predicting breast cancer patient outcomes (Table 5.38.). Not all algorithms performed equally well in

predicting breast cancer outcomes. Logistic Regression achieved an accuracy of 62.5%, which is relatively lower compared to other algorithms. Assuming a balanced dataset, a good accuracy score would be above 70%. On the other hand, the Decision Tree algorithm achieved a perfect accuracy of 100%, while the SVM algorithm achieved an accuracy of 75%.

Typically, overfitting occurs when we have an overly flexible model, which explains the high accuracy of the Decision Tree algorithm that did not predict any false values. Therefore, training was conducted using the Random Forest model, where each decision tree learns from a random sample of data points, ensuring that each tree is trained on a different sample. However, the perfect accuracy achieved by the Decision Tree raises concerns about potential overfitting, as it did not make any errors in predicting the test samples.

When determining the depth of the decision tree, the Random Forest algorithm achieved a remarkable accuracy of 100%. It is worth noting that there was no need to increase the depth of the forest because it already reached the highest level of accuracy. This indicates that the algorithm was able to capture the complex relationships within the dataset and make accurate predictions without the need for further depth.

Table 5.38. Accuracy of classification algorithms

		Accuracy Test=0.25
Logistic regression		62.5%
Support vector machine		75%
Decision tree		100%
Random forest	max_depth = 2	100%

It's worth noting that later, grid search was performed for the logistic regression, and k-fold cross-validation was conducted for the SVM algorithm. These techniques are commonly used to fine-tune hyperparameters and obtain a more robust evaluation of the model's performance. The following codes were used:

GridSearchCV:

```
  from sklearn.model_selection import GridSearchCV
  from time import *
  from sklearn.metrics import accuracy_score
  from sklearn.metrics import confusion_matrix
  from sklearn.metrics import classification_report
  clf = GridSearchCV(SVC(class_weight='balanced'), tuned_parameters,
cv=3,
                     scoring='accuracy')
```

K-fold validation:

```
  #Train the logistic regression model using the balanced weights

  lr_balanced = LogisticRegression( class_weight='balanced',
random_state=0, n_jobs=-1, max_iter=100)

  lr_balanced .fit(X_train, y_train)

  # fit the regression with X and Y data

  lr_balanced_model_cv = cross_validate(lr_balanced, X, y,
cv=StratifiedKFold(n_splits=7), return_train_score=True)
```

Overall, the grid search for logistic regression successfully found a set of hyperparameters that led to a perfect accuracy of 100%, indicating a strong performance on the development set. On the other hand, the k-fold cross-validation for the SVM algorithm resulted in an average accuracy of 88%, demonstrating good generalization capabilities. Both grid search and k-fold cross-validation are valuable techniques that help optimize hyperparameters and assess model performance in a robust manner.

To further delve into the performance of the four classification algorithms, it is necessary to refer to the respective confusion matrixes for the tested samples (n = 476). It was observed that the confusion matrix for the Decision Tree (DT) algorithm exhibited the best classification performance with no errors, as shown in (Table 5.39.). On the other hand, for the Logistic Regression algorithm with the lowest accuracy, we can observe from Table 5.40. that there were a few misclassifications, the algorithm correctly predicted 238 patients who would die and 59 patients who would survive. However, it made incorrect predictions for 120 patients, classifying them as survivors when they actually died, and for 59 patients, classifying them as deaths when they actually survived.

Analyzing the confusion matrices provides valuable insights into the classification performance of the algorithms and helps identify any specific areas of improvement or errors in predictions.

Table 5.39. Confusion matrix for the Decision Tree algorithm

Decision tree (DT)		**Predicted Outcome**	
		Died of Disease	Living
Actual outcome	Died of Disease	276	0
	Living	0	200

Table 5.40. Confusion matrix for the Logistic Regression algorithm

Logistic Regression		**Predicted Outcome**	
		Died of Disease	Living
Actual outcome	Died of Disease	238	120
	Living	59	59

As for the SVC (Support Vector Classifier), a Confusion Matrix was utilized to evaluate its performance, resulting in a prediction accuracy of 75% (Table 5.41). The algorithm

correctly predicted 299 patients who would die and 59 patients who would survive. However, it made incorrect predictions for 59 patients, classifying them as survivors when they actually died, and for 59 patients, classifying them as deaths when they actually survived. The accuracy of 75% indicates that the SVC algorithm has moderate performance in predicting patient outcomes. It is relatively successful in predicting patient deaths but struggles with predicting patient survival.

Table 5.41. Confusion matrix for the SVC algorithm

Support Vector Machine		**Predicted Outcome**	
		Died of Disease	Living
Actual outcome	Died of Disease	299	59
	Living	59	59

As for the k-fold cross-validation, a Confusion Matrix was utilized to evaluate its performance, resulting in a prediction accuracy of 88% (Table 5.42). The algorithm correctly predicted 370 patients who would die and 53 patients who would survive. However, it made incorrect predictions for 53 patients, classifying them as survivors when they actually died.

Table 5.42. Confusion matrix for the K-fold cross-validation

K-Fold Cross-Validation		**Predicted Outcome**	
		Died of Disease	Living
Actual outcome	Died of Disease	370	53
	Living	0	53

The results obtained from the classification algorithms in predicting breast cancer patient outcomes provide valuable insights into their performance. It is evident that not all algorithms performed equally well in this task. Logistic Regression achieved the lowest accuracy of 62.5%, indicating that it struggled in accurately predicting patient outcomes. The study conducted by Montazeri et al. [152] also reported lower accuracy for Logistic Regression compared to other techniques, indicating that Logistic Regression may not be the most effective algorithm for breast cancer survival prediction.

On the other hand, the Decision Tree algorithm achieved a perfect accuracy of 100%, suggesting that it performed exceptionally well in classifying patients. This aligns with the findings of different studies [145, 146, 152, 153], where Decision Tree models consistently demonstrated perfect accuracy in predicting breast cancer patient outcomes.

The Random Forest algorithm, which is an ensemble of decision trees, also achieved a perfect accuracy of 100%. This demonstrates the effectiveness of combining multiple decision trees to improve the overall predictive performance. The Random Forest algorithm showed no need for increasing the depth of the trees as it already reached the highest level of accuracy, indicating its ability to capture the complex relationships within the dataset.

The Support Vector Machine (SVM) algorithm achieved an accuracy of 75%. Similarly, the study conducted by Li et al. reported an accuracy of 75% for SVM in predicting the survival of breast cancer patients with bone metastases [153]. While this is relatively lower

compared to the Decision Tree and Random Forest algorithms, it still demonstrates moderate performance. The SVM algorithm was more successful in predicting patient deaths compared to patient survival.

Analyzing the confusion matrices provided a deeper understanding of the classification performance of the algorithms. The Decision Tree algorithm exhibited the best performance, with no errors in its predictions. However, the Logistic Regression algorithm had misclassifications, particularly in predicting patient survival. The SVC algorithm also had misclassifications, but it showed relatively better performance in predicting patient deaths.

In conclusion, the results highlight the varying performance of classification algorithms in predicting breast cancer patient outcomes. The Decision Tree algorithm, along with the Random Forest ensemble, demonstrated excellent accuracy without any false predictions. These algorithms can be considered reliable for breast cancer outcome prediction. On the other hand, the Logistic Regression algorithm had lower accuracy, indicating the need for further improvement or alternative algorithms.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F (2021) Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 71:209–249
2. Heer E, Harper A, Escandor N, Sung H, McCormack V, Fidler-Benaoudia MM (2020) Global burden and trends in premenopausal and postmenopausal breast cancer: a population-based study. Lancet Glob Heal 8:e1027–e1037
3. Sun YS, Zhao Z, Yang ZN, Xu F, Lu HJ, Zhu ZY, Shi W, Jiang J, Yao PP, Zhu HP (2017) Risk Factors and Preventions of Breast Cancer. Int J Biol Sci 13:1387–1397
4. Types of Breast Cancer - National Breast Cancer Foundation. https://www.nationalbreastcancer.org/types-of-breast-cancer/. Accessed 26 May 2023
5. Ductal Carcinoma In Situ (DCIS) - National Breast Cancer Foundation. https://www.nationalbreastcancer.org/dcis/. Accessed 26 May 2023
6. Invasive Ductal Carcinoma (IDC) - National Breast Cancer Foundation. https://www.nationalbreastcancer.org/invasive-ductal-carcinoma/. Accessed 26 May 2023
7. Lobular Carcinoma In Situ (LCIS) - National Breast Cancer Foundation. https://www.nationalbreastcancer.org/lobular-carcinoma-in-situ/. Accessed 26 May 2023
8. Invasive Lobular Cancer (ILC) - National Breast Cancer Foundation. https://www.nationalbreastcancer.org/invasive-lobular-cancer/. Accessed 26 May 2023
9. Triple Negative Breast Cancer - National Breast Cancer Foundation. https://www.nationalbreastcancer.org/triple-negative-breast-cancer/. Accessed 26 May 2023
10. Inflammatory Breast Cancer: Symptoms, Diagnosis, Treatment - National Breast Cancer Foundation. https://www.nationalbreastcancer.org/inflammatory-breast-cancer/. Accessed 26 May 2023
11. Metastatic Breast Cancer: What Is It, Symptoms, and More. https://www.nationalbreastcancer.org/metastatic-breast-cancer/. Accessed 26 May 2023
12. Other Types - National Breast Cancer Foundation. https://www.nationalbreastcancer.org/other-types-of-breast-cancer/. Accessed 26 May 2023
13. Shaath H, Elango R, Alajez NM (2021) Molecular classification of breast cancer utilizing long non-coding rna (Lncrna) transcriptomes identifies novel diagnostic lncrna panel for triple-negative breast cancer. Cancers (Basel) 13:5350
14. ZHANG MH, MAN HT, ZHAO XD, DONG N, MA SL (2014) Estrogen receptor-positive breast cancer molecular signatures and therapeutic potentials (Review). Biomed Reports 2:41–52
15. Miah S, Bagu E, Goel R, et al (2019) Estrogen receptor signaling regulates the expression of the breast tumor kinase in breast cancer cells. BMC Cancer 19:1–12

16. Hicks DG, Lester SC (2016) Hormone Receptors (ER/PR). Diagnostic Pathol Breast 430–439

17. Nicolini A, Ferrari P, Duffy MJ (2018) Prognostic and predictive biomarkers in breast cancer: Past, present and future. Semin Cancer Biol 52:56–73

18. Purdie CA, Quinlan P, Jordan LB, Ashfield A, Ogston S, Dewar JA, Thompson AM (2013) Progesterone receptor expression is an independent prognostic variable in early breast cancer: a population-based study. Br J Cancer 2014 1103 110:565–572

19. Iqbal N, Iqbal N (2014) Human Epidermal Growth Factor Receptor 2 (HER2) in Cancers: Overexpression and Therapeutic Implications. Mol Biol Int 2014:1–9

20. Vaz-Luis I, Winer EP, Lin NU (2013) Human epidermal growth factor receptor-2-positive breast cancer: Does estrogen receptor status define two distinct subtypes? Ann Oncol 24:283–291

21. Krishnamurti U, Hammers JL, Atem FD, Storto PD, Silverman JF (2009) Poor prognostic significance of unamplified chromosome 17 polysomy in invasive breast carcinoma. Mod Pathol 22:1044–1048

22. Haroon S, Hashmi AA, Khurshid A, Kanpurwala MA, Mujtuba S, Malik B, Faridi N (2013) Ki67 Index in Breast Cancer: Correlation with Other Prognostic Markers and Potential in Pakistani Patients. Asian Pacific J Cancer Prev 14:4353–4358

23. Hammond MEH, Hayes DF, Dowsett M, et al (2010) American Society of Clinical oncology/college of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. Arch Pathol Lab Med 134:907–922

24. Gupta P, Rai NN, Agarwal L, Namdev S (2018) Comparison of Molecular Subtypes of Carcinoma of the Breast in Two Different Age Groups: A Single Institution Experience. Cureus. https://doi.org/10.7759/CUREUS.2834

25. Sharma JD, Khanna S, Ramchandani S, Kakoti LM, Baruah A, Mamidala V (2021) Prevalence of Molecular Subtypes of Breast Carcinoma and Its Comparison between Two Different Age Groups: A Retrospective Study from a Tertiary Care Center of Northeast India. South Asian J Cancer 10:220–224

26. Higgins MJ, Stearns V (2009) Understanding Resistance to Tamoxifen in Hormone Receptor–Positive Breast Cancer. Clin Chem 55:1453–1455

27. Inic Z, Zegarac M, Inic M, Markovic I, Kozomara Z, Djurisic I, Inic I, Pupic G, Jancic S (2014) Difference between Luminal A and Luminal B subtypes according to Ki-67, tumor size, and progesterone receptor negativity providing prognostic information. Clin Med Insights Oncol 8:107–111

28. Lafcı O, Celepli P, Seher Öztekin P, Koşar PN (2023) DCE-MRI Radiomics Analysis in Differentiating Luminal A and Luminal B Breast Cancer Molecular Subtypes. Acad Radiol 30:22–29

29. Krishnamurti U, Silverman JF (2014) HER2 in breast cancer: A review and update. Adv Anat Pathol 21:100–107

30. Figueroa-Magalhães MC, Jelovac D, Connolly RM, Wolff AC (2014) Treatment of HER2-positive breast cancer. Breast 23:128–136

31. Wang J, Xu B (2019) Targeted therapeutic options and future perspectives for HER2-positive breast cancer. Signal Transduct Target Ther 2019 41 4:1–22

32. Kumar P, Aggarwal R (2015) An overview of triple-negative breast cancer. Arch Gynecol Obstet 2015 2932 293:247–269

33. Loibl S, Gianni L (2017) HER2-positive breast cancer. Lancet 389:2415–2429

34. Collignon J, Lousberg L, Schroeder H, Jerusalem G (2016) Triple-negative breast cancer: treatment challenges and solutions. Breast Cancer Targets Ther 8:93–107

35. TJ K, PN A, GK R, et al (2013) Sex hormones and risk of breast cancer in premenopausal women: a collaborative reanalysis of individual participant data from seven prospective studies. Lancet Oncol 14:1009–1019

36. Giordano SH (2018) Breast Cancer in Men. N Engl J Med 378:2311–2320

37. McGuire A, Brown JAL, Malone C, McLaughlin R, Kerin MJ (2015) Effects of age on the detection and management of breast cancer. Cancers (Basel) 7:908–929

38. Siegel R, Ma J, Zou Z, Jemal A (2014) Cancer statistics, 2014. CA Cancer J Clin 64:9–29

39. Benz CC (2008) Impact of aging on the biology of breast cancer. Crit Rev Oncol Hematol 66:65–74

40. (2004) Stat bite: Lifetime probability among females of dying of cancer. J Natl Cancer Inst 96:818

41. Beral V, Bull D, Doll R, et al (2001) Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. Lancet (London, England) 358:1389–1399

42. Baglia ML, Tang MTC, Malone KE, Porter P, Li CI (2018) Family History and Risk of Second Primary Breast Cancer after In Situ Breast Carcinoma. Cancer Epidemiol Biomarkers Prev 27:315–320

43. Brewer HR, Jones ME, Schoemaker MJ, Ashworth A, Swerdlow AJ (2017) Family history and risk of breast cancer: an analysis accounting for family structure. Breast Cancer Res Treat 165:193–200

44. Shiyanbola OO, Arao RF, Miglioretti DL, et al (2017) Emerging Trends in Family History of Breast Cancer and Associated Risk. Cancer Epidemiol Biomarkers Prev 26:1753–1760

45. Wu HC, Do C, Andrulis IL, et al (2018) Breast cancer family history and allele-specific DNA methylation in the legacy girls study. Epigenetics 13:240–250

46. Çelik A, Acar M, Erkul CM, Gunduz EG and M, Çelik A, Acar M, Erkul CM, Gunduz EG and M (2015) Relationship of Breast Cancer with Ovarian Cancer. A Concise Rev Mol Pathol Breast Cancer. https://doi.org/10.5772/59682

47. Shiovitz S, Korde LA (2015) Genetics of breast cancer: a topic in evolution. Ann Oncol Off J Eur Soc Med Oncol 26:1291–1299

48. Hill DA, Prossnitz ER, Royce M, Nibbe A (2019) Temporal trends in breast cancer survival by race and ethnicity: A population-based cohort study. PLoS One. https://doi.org/10.1371/JOURNAL.PONE.0224064

49. Yedjou CG, Sims JN, Miele L, Noubissi F, Lowe L, Fonseca DD, Alo RA, Payton M, Tchounwou PB (2019) Health and Racial Disparity in Breast Cancer. Adv Exp Med Biol 1152:31–49

50. Ghafoor: American Cancer Society breast cancer facts... - Google Scholar.

51. Breast Cancer Risk Factors You Can't Change. https://www.cancer.org/cancer/types/breast-cancer/risk-and-prevention/breast-cancer-risk-factors-you-cannot-change.html. Accessed 6 Jul 2023

52. Albrektsen G, Heuch I, Hansen S, Kvåle G (2005) Breast cancer risk by age at birth, time since birth and time intervals between births: exploring interaction effects. Br J Cancer 92:167–175

53. Bernstein L (2002) Epidemiology of endocrine-related risk factors for breast cancer. J Mammary Gland Biol Neoplasia 7:3–15

54. Husby A, Wohlfahrt J, Øyen N, Melbye M (2018) Pregnancy duration and breast cancer risk. Nat Commun. https://doi.org/10.1038/S41467-018-06748-3

55. Checka CM, Chun JE, Schnabel FR, Lee J, Toth H (2012) The relationship of mammographic density and age: implications for breast cancer screening. AJR Am J Roentgenol. https://doi.org/10.2214/AJR.10.6049

56. Kim EY, Chang Y, Ahn J, Yun JS, Park YL, Park CH, Shin H, Ryu S (2020) Mammographic breast density, its changes, and breast cancer risk in premenopausal and postmenopausal women. Cancer 126:4687–4696

57. Duffy SW, Morrish OWE, Allgood PC, et al (2018) Mammographic density and breast cancer risk in breast screening assessment cases and women with a family history of breast cancer. Eur J Cancer 88:48–56

58. Schacht D V., Yamaguchi K, Lai J, Kulkarni K, Sennett CA, Abe H (2014) Importance of a personal history of breast cancer as a risk factor for the development of subsequent breast cancer: results from screening breast MRI. AJR Am J Roentgenol 202:289–292

59. Wang J, Costantino JP, Tan-Chiu E, Wickerham DL, Paik S, Wolmark N (2004) Lower-category benign breast disease and the risk of invasive breast cancer. J Natl Cancer Inst 96:616–620

60. Dyrstad SW, Yan Y, Fowler AM, Colditz GA (2015) Breast cancer risk associated with benign breast disease: systematic review and meta-analysis. Breast Cancer Res Treat 149:569–575

61. Hartmann LC, Sellers TA, Frost MH, et al (2005) Benign breast disease and the risk of breast cancer. N Engl J Med 353:229–237

62. Ng J, Shuryak I (2014) Minimizing second cancer risk following radiotherapy: current perspectives. Cancer Manag Res 7:1–11

63. Zhang Q, Liu J, Ao N, Yu H, Peng Y, Ou L, Zhang S (2020) Secondary cancer risk after radiation therapy for breast cancer with different radiotherapy techniques. Sci Rep. https://doi.org/10.1038/S41598-020-58134-Z

64. Ng AK, Travis LB (2009) Radiation therapy and breast cancer risk. J Natl Compr Canc Netw 7:1121–1128

65. Hilakivi-Clarke L (2014) Maternal exposure to diethylstilbestrol during pregnancy and increased breast cancer risk in daughters. Breast Cancer Res 16:1–10

66. Palmer JR, Wise LA, Hatch EE, et al (2006) Prenatal diethylstilbestrol exposure and risk of breast cancer. Cancer Epidemiol Biomarkers Prev 15:1509–1514

67. Vinogradova Y, Coupland C, Hippisley-Cox J (2020) Use of hormone replacement therapy and risk of breast cancer: nested case-control studies using the QResearch and

CPRD databases. BMJ. https://doi.org/10.1136/BMJ.M3873

68. Narod SA (2011) Hormone replacement therapy and the risk of breast cancer. Nat Rev Clin Oncol 8:669–676

69. Steingart A, Cotterchio M, Kreiger N, Sloan M (2003) Antidepressant medication use and breast cancer risk: a case-control study. Int J Epidemiol 32:961–966

70. Wernli KJ, Hampton JM, Trentham-Dietz A, Newcomb PA (2009) Antidepressant medication use and breast cancer risk. Pharmacoepidemiol Drug Saf 18:284–290

71. Kyu HH, Bachman VF, Alexander LT, et al (2016) Physical activity and risk of breast cancer, colon cancer, diabetes, ischemic heart disease, and ischemic stroke events: systematic review and dose-response meta-analysis for the Global Burden of Disease Study 2013. BMJ. https://doi.org/10.1136/BMJ.I3857

72. Chen X, Wang Q, Zhang Y, Xie Q, Tan X (2019) Physical Activity and Risk of Breast Cancer: A Meta-Analysis of 38 Cohort Studies in 45 Study Reports. Value Health 22:104–128

73. Bernstein L, Ross RK (1993) Endogenous Hormones and Breast Cancer Risk. Epidemiol Rev 15:48–65

74. Hoffman-Goetz L (1998) Influence of physical activity and exercise on innate immunity. Nutr Rev. https://doi.org/10.1111/J.1753-4887.1998.TB01629.X

75. Kolb R, Zhang W (2020) Obesity and Breast Cancer: A Case of Inflamed Adipose Tissue. Cancers (Basel) 12:1–18

76. Wang X, Hui TL, Wang MQ, Liu H, Li RY, Song ZC (2019) Body Mass Index at Diagnosis as a Prognostic Factor for Early-Stage Invasive Breast Cancer after Surgical Resection. Oncol Res Treat 42:190–196

77. Sun L, Zhu Y, Qian Q, Tang L (2018) Body mass index and prognosis of breast cancer: An analysis by menstruation status when breast cancer diagnosis. Medicine (Baltimore). https://doi.org/10.1097/MD.0000000000011220

78. James FR, Wootton S, Jackson A, Wiseman M, Copson ER, Cutress RI (2015) Obesity in breast cancer--what is the risk factor? Eur J Cancer 51:705–720

79. Protani M, Coory M, Martin JH (2010) Effect of obesity on survival of women with breast cancer: systematic review and meta-analysis. Breast Cancer Res Treat 123:627–635

80. Iyengar NM, Arthur R, Manson JE, et al (2019) Association of Body Fat and Risk of Breast Cancer in Postmenopausal Women With Normal Body Mass Index: A Secondary Analysis of a Randomized Clinical Trial and Observational Study. JAMA Oncol 5:155–163

81. Hopper JL, Dite GS, MacInnis RJ, et al (2018) Age-specific breast cancer risk by body mass index and familial risk: prospective family study cohort (ProF-SC). Breast Cancer Res. https://doi.org/10.1186/S13058-018-1056-1

82. Rachdaoui N, Sarkar DK (2013) Effects of alcohol on the endocrine system. Endocrinol Metab Clin North Am 42:593–615

83. Erol A, Ho AMC, Winham SJ, Karpyak VM (2019) Sex hormones in alcohol consumption: a systematic review of evidence. Addict Biol 24:157–169

84. Alcohol consumption and the risk of breast cancer - PubMed. https://pubmed.ncbi.nlm.nih.gov/22218798/. Accessed 7 Jul 2023

85. Zeinomar N, Knight JA, Genkinger JM, et al (2019) Alcohol consumption, cigarette smoking, and familial breast cancer risk: findings from the Prospective Family Study Cohort (ProF-SC). Breast Cancer Res. https://doi.org/10.1186/S13058-019-1213-1

86. Liu Y, Nguyen N, Colditz GA (2015) Links between alcohol consumption and breast cancer: a look at the evidence. Womens Health (Lond Engl) 11:65–77

87. Couch FJ, Cerhan JR Cigarette smoking increases risk for breast cancer in high-risk breast cancer families - PubMed.

88. Misotti AM, Gnagnarella P (2013) Vitamin supplement consumption and breast cancer risk: a review. Ecancermedicalscience. https://doi.org/10.3332/ECANCER.2013.365

89. Jones ME, Schoemaker MJ, Wright LB, Ashworth A, Swerdlow AJ (2017) Smoking and risk of breast cancer in the Generations Study cohort. Breast Cancer Res. https://doi.org/10.1186/S13058-017-0908-4

90. Catsburg C, Miller AB, Rohan TE (2015) Active cigarette smoking and risk of breast cancer. Int J cancer 136:2204–2209

91. El-Sharkawy A, Malki A (2020) Vitamin D Signaling in Inflammation and Cancer: Molecular Mechanisms and Therapeutic Implications. Molecules. https://doi.org/10.3390/MOLECULES25143219

92. Atoum M, Alzoughool F (2017) Vitamin D and Breast Cancer: Latest Evidence and Future Steps. Breast Cancer (Auckl). https://doi.org/10.1177/1178223417749816

93. Cui Y, Rohan TE (2006) Vitamin D, calcium, and breast cancer risk: a review. Cancer Epidemiol Biomarkers Prev 15:1427–1437

94. Estébanez N, Gómez-Acebo I, Palazuelos C, Llorca J, Dierssen-Sotos T (2018) Vitamin D exposure and Risk of Breast Cancer: a meta-analysis. Sci Rep. https://doi.org/10.1038/S41598-018-27297-1

95. Huss L, Butt ST, Borgquist S, Elebro K, Sandsveden M, Rosendahl A, Manjer J (2019) Vitamin D receptor expression in invasive breast tumors and breast cancer survival. Breast Cancer Res. https://doi.org/10.1186/S13058-019-1169-1

96. Zhou L, Chen B, Sheng L, Turner A (2020) The effect of vitamin D supplementation on the risk of breast cancer: a trial sequential meta-analysis. Breast Cancer Res Treat. https://doi.org/10.1007/S10549-020-05669-4

97. Al-Naggar RA, Anil S (2016) Artificial Light at Night and Cancer: Global Study. Asian Pac J Cancer Prev 17:4661–4664

98. Johns LE, Jones ME, Schoemaker MJ, McFadden E, Ashworth A, Swerdlow AJ (2018) Domestic light at night and breast cancer risk: a prospective analysis of 105 000 UK women in the Generations Study. Br J Cancer 118:600

99. Dandamudi A, Tommie J, Nommsen-Rivers L, Couch S (2018) Dietary Patterns and Breast Cancer Risk: A Systematic Review. Anticancer Res 38:3209–3222

100. Fiolet T, Srour B, Sellem L, et al (2018) Consumption of ultra-processed foods and cancer risk: results from NutriNet-Santé prospective cohort. BMJ. https://doi.org/10.1136/BMJ.K322

101. Castelló A, Pollán M, Buijsse B, et al (2014) Spanish Mediterranean diet and other dietary patterns and breast cancer risk: case–control EpiGEICAM study. Br J Cancer 111:1454

102. Kotepui M (2016) Diet and risk of breast cancer. Contemp Oncol (Poznan, Poland)

20:13–19

103. Li MJ, Yin YC, Wang J, Jiang YF (2014) Green tea compounds in breast cancer prevention and treatment. World J Clin Oncol 5:520–528

104. Wright L, Frye J, Gorti B, Timmermann B, Funk J (2013) Bioactivity of turmeric-derived curcuminoids and related metabolites in breast cancer. Curr Pharm Des 19:6218–6225

105. Liu D, Chen Z (2013) The effect of curcumin on breast cancer cells. J Breast Cancer 16:133–137

106. Casey SC, Vaccari M, Al-Mulla F, et al (2015) The effect of environmental chemicals on the tumor microenvironment. Carcinogenesis 36 Suppl 1:S160–S183

107. Videnros C, Selander J, Wiebert P, Albin M, Plato N, Borgquist S, Manjer J, Gustavsson P (2020) Investigating the risk of breast cancer among women exposed to chemicals: a nested case–control study using improved exposure estimates. Int Arch Occup Environ Health 93:261

108. Eve L, Fervers B, Romancer M Le, Etienne-Selloum N (2020) Exposure to Endocrine Disrupting Chemicals and Risk of Breast Cancer. Int J Mol Sci 21:1–43

109. Rodgers KM, Udesky JO, Rudel RA, Brody JG (2018) Environmental chemicals and breast cancer: An updated review of epidemiological literature informed by biological mechanisms. Environ Res 160:152–182

110. Leso V, Ercolano ML, Cio DL, Iavicoli I (2019) Occupational Chemical Exposure and Breast Cancer Risk According to Hormone Receptor Status: A Systematic Review. Cancers 2019, Vol 11, Page 1882 11:1882

111. Zhang SM, Cook NR, Manson JE, Lee IM, Buring JE (2008) Low-dose aspirin and breast cancer risk: results by tumour characteristics from a randomised trial. Br J Cancer 98:989–991

112. Olsen JH, Sørensen HT, Friis S, McLaughlin JK, Steffensen FH, Nielsen GL, Andersen M, Fraumeni JF, Olsen J (1997) Cancer risk in users of calcium channel blockers. Hypertens (Dallas, Tex 1979) 29:1091–1094

113. J Brandes L Stimulation of malignant growth in rodents by antidepressant drugs at clinically relevant doses - PubMed.

114. Bjarnadottir O, Romero Q, Bendahl PO, et al (2013) Targeting HMG-CoA reductase with statins in a window-of-opportunity breast cancer trial. Breast Cancer Res Treat 138:499–508

115. Velicer CM, Lampe JW, Heckbert SR, Potter JD, Taplin SH (2003) Hypothesis: is antibiotic use associated with breast cancer? Cancer Causes Control 14:739–747

116. Morrow M, White J, Moughan J, Owen J, Pajack T, Sylvester J, Wilson JF, Winchester D (2001) Factors predicting the use of breast-conserving therapy in stage I and II breast carcinoma. J Clin Oncol 19:2254–2262

117. Rahman GA (2011) Breast Conserving Therapy: A surgical Technique where Little can Mean More. J Surg Tech Case Rep 3:1–4

118. Cardoso F, Kyriakides S, Ohno S, Penault-Llorca F, Poortmans P, Rubio IT, Zackrisson S, Senkus E (2019) Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up†. Ann Oncol Off J Eur Soc Med Oncol 30:1194–1220

119. Rouzier R, Perou CM, Symmans WF, et al (2005) Breast cancer molecular subtypes respond differently to preoperative chemotherapy. Clin Cancer Res 11:5678–5685

120. Fisher B, Bryant J, Wolmark N, et al (1998) Effect of preoperative chemotherapy on the outcome of women with operable breast cancer. J Clin Oncol 16:2672–2685

121. Jonathan Yang T, Ho AY (2013) Radiation therapy in the management of breast cancer. Surg Clin North Am 93:455–471

122. Joshi SC, Khan FA, Pant I, Shukla A (2007) Role of Radiotherapy in Early Breast Cancer: An Overview. Int J Health Sci (Qassim) 1:259

123. Tremont A, Lu J, Cole JT (2017) Endocrine Therapy for Early Breast Cancer: Updated Review. Ochsner J 17:405

124. F L, G L, SM B, U B, A B, V C (2011) Endocrine therapy of breast cancer. Curr Med Chem 18:469–491

125. Jones KL, Buzdar AU (2004) A review of adjuvant hormonal therapy in breast cancer. Endocr Relat Cancer 11:391–406

126. Drăgănescu M, Carmocan C (2017) Hormone Therapy in Breast Cancer. Chirurgia (Bucur) 112:413–417

127. Abe O, Abe R, Enomoto K, et al (2005) Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. Lancet (London, England) 365:1687–1717

128. Park JW, Liu MC, Yee D, et al (2016) Adaptive Randomization of Neratinib in Early Breast Cancer. N Engl J Med 375:11–22

129. Moreira C, Kaklamani V (2010) Lapatinib and breast cancer: current indications and outlook for the future. Expert Rev Anticancer Ther 10:1171–1182

130. Nguyen X, Hooper M, Borlagdan JP, Palumbo A (2021) A Review of Fam-Trastuzumab Deruxtecan-nxki in HER2-Positive Breast Cancer. Ann Pharmacother 55:1410–1418

131. Ishii K, Morii N, Yamashiro H (2019) Pertuzumab in the treatment of HER2-positive breast cancer: an evidence-based review of its safety, efficacy, and place in therapy. Core Evid 14:51–70

132. Maximiano S, Magalhães P, Guerreiro MP, Morgado M (2016) Trastuzumab in the Treatment of Breast Cancer. BioDrugs 30:75–86

133. Shah A, Bloomquist E, Tang S, et al (2018) FDA Approval: Ribociclib for the Treatment of Postmenopausal Women with Hormone Receptor-Positive, HER2-Negative Advanced or Metastatic Breast Cancer. Clin Cancer Res 24:2999–3004

134. Steger GG, Gnant M, Bartsch R (2016) Palbociclib for the treatment of postmenopausal breast cancer - an update. Expert Opin Pharmacother 17:255–263

135. Riccardi F, Colantuoni G, Diana A, et al (2018) Exemestane and Everolimus combination treatment of hormone receptor positive, HER2 negative metastatic breast cancer: A retrospective study of 9 cancer centers in the Campania Region (Southern Italy) focused on activity, efficacy and safety. Mol Clin Oncol. https://doi.org/10.3892/MCO.2018.1672

136. Royce ME, Osman D (2015) Everolimus in the Treatment of Metastatic Breast Cancer. Breast Cancer (Auckl) 9:73–79

137. Kwapisz D (2017) Cyclin-dependent kinase 4/6 inhibitors in breast cancer: palbociclib, ribociclib, and abemaciclib. Breast Cancer Res Treat 166:41–54

138. Tarantino P, Morganti S, Curigliano G (2020) Biologic therapy for advanced breast cancer: recent advances and future directions. https://doi.org/101080/1471259820201752176 20:1009–1024

139. Steger GG, Bartsch R (2011) Denosumab for the treatment of bone metastases in breast cancer: evidence and opinion. Ther Adv Med Oncol 3:233

140. Heimes AS, Schmidt M (2019) Atezolizumab for the treatment of triple-negative breast cancer. Expert Opin Investig Drugs 28:1–5

141. (PDF) The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years. https://www.researchgate.net/publication/220605256_The_Dartmouth_College_Artificial_Intelligence_Conference_The_Next_Fifty_Years. Accessed 25 Jun 2023

142. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Peter Campbell J (2020) Introduction to Machine Learning, Neural Networks, and Deep Learning. Transl Vis Sci Technol 9:14–14

143. Shobha G, Rangaswamy S (2018) Machine Learning. Handb Stat 38:197–228

144. Song YY, Lu Y (2015) Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry 27:130

145. Machine Learning and Deep Learning Approaches in Breast Cancer Survival Prediction Using Clinical Data - PubMed. https://pubmed.ncbi.nlm.nih.gov/32362304/. Accessed 30 Apr 2023

146. Ganggayah MD, Taib NA, Har YC, Lio P, Dhillon SK (2019) Predicting factors for survival of breast cancer patients using machine learning techniques. BMC Med Inform Decis Mak 19:1–17

147. Pereira B, Chin SF, Rueda OM, et al (2016) The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. Nat Commun 2016 71 7:1–16

148. Liu J, Li Y, Zhang W, et al (2022) The prognostic role of lymph node ratio in breast cancer patients received neoadjuvant chemotherapy: A dose-response meta-analysis. Front Surg 9:971030

149. Dhar SS, Lee MG (2021) Cancer-epigenetic function of the histone methyltransferase KMT2D and therapeutic opportunities for the treatment of KMT2D-deficient tumors. Oncotarget 12:1296

150. Sopik V, Narod SA (2018) The relationship between tumour size, nodal status and distant metastases: on the origins of breast cancer. Breast Cancer Res Treat 170:647

151. Rakha EA, Gill MS, El-Sayed ME, Khan MM, Hodi Z, Blamey RW, Evans AJ, Lee AHS, Ellis IO (2009) The biological and clinical characteristics of breast carcinoma with mixed ductal and lobular morphology. Breast Cancer Res Treat 114:243–250

152. Montazeri M, Montazeri M, Montazeri M, Beigzadeh A (2016) Machine learning models in breast cancer survival prediction. Technol Health Care 24:31–42

153. Li C, Liu M, Li J, et al (2022) Machine learning predicts the prognosis of breast cancer patients with initial bone metastases. Front Public Heal 10:3437