REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE MOHAMED BOUDIAF - M'SILA

FACULTE DE Technologie **FILIÈRE: Electronique**

DEPARTEMENT Electronique **OPTION: instrumentation**

N° :

Mémoire présenté pour l'obtention

Du diplôme de Master Académique

Par :

Filali Sabir

Salem Abdelhamid

THÈME :

Breast cancer classification using machine learning methods

Soutenu devant le jury composé de :

Dr.	Université M^{ed} Boudiaf –M'sila	Président
Dr.	Université M^{ed} Boudiaf –M'sila	Rapporteur
Dr.	Université M^{ed} Boudiaf –M'sila	Co-Rapporteur
Dr.	Université M^{ed} Boudiaf –M'sila	Examinateur

Année universitaire: 2022 /2023

ACKNOWLEDGEMENTS

we would like to thank the Almighty Allah for giving us the strength and support to complete this research work.

Also, We would like to express our sincere gratitude to our supervisor Dr. Mohamed Djerioui for his valuable and constructive suggestions throughout the process. In addition to Tawfiq beghriche for his willingness to give his time so generously has been very much appreciated.

Besides our advisor, we would like to thank both our families and especially our parents for their constant support and for believing in us throughout our entire careers.

Last but not least, we would like to thank our colleagues whom we've spent amazing time with and everyone we learnt something from throughout our lives.

DEDICATION

I am dedicating this thesis to my family who have meant and continue to mean so much to me, a big special feeling of gratitude to my loving beloved parents whose support and encouragement has been with been since the childhood.

To my brothers and sisters each by his and her name, who have been the source of inspiration and guidance throughout my career, thank you all so much.

With heartfelt gratitude and appreciation,

FILALI SABIR.

To my family, mentors, and contributors, this thesis is dedicated. Your support, guidance, and inspiration have been invaluable on this journey of knowledge. With heartfelt gratitude, we offer this work as a contribution to the pursuit of advancement and betterment.

Salem Abdelhamid.

TABLE OF CONTENTS

ABBREVIATION AND NOTATION

- **AI:** Artificial intelligence
- **ML:** Machine Learning
- **DL:** Deep Learning
- **SVM:** Support Vector Machine
- **DT:** Decision tree
- **LT:** logistic regression
- **RF:** Random forest
- **TP:** True positive
- **TN:** True negative
- **FP:** False positive
- **FN:** False negative
- **WHO:** World Health Organisation
- **WBCD:** Wisconsin Breast Cancer Dataset

LIST OF TABLES

LIST OF FIGURES

General Introduction

General Introduction

Breast cancer is a highly lethal form of cancer, ranking as the second most deadly malignancy among women worldwide, surpassed only by lung cancer. It predominantly affects the duct region of a woman's breast. Shockingly, in 2020 alone, there were 2.3 million women diagnosed with breast cancer, leading to 627,000 deaths attributed to the disease [1].

To detect breast cancer, medical professionals employ various diagnostic techniques such ultrasonography, mammography, magnetic resonance imaging (MRI), and biopsy. Based on the results of these tests, further investigations or treatments may be recommended. Early detection plays a pivotal role in breast cancer management, as it can significantly improve a patient's chances of survival. Therefore, extensive research is underway to accurately identify and classify patents into either: malignant or benign [2].

Machine Learning (ML) is a scientific field that focuses on computational algorithms and statistical models designed to surpass human intelligence. These computer-based techniques leverage prior knowledge to enhance performance and make accurate predictions. By learning from previous data, ML systems can forecast future outcomes [3].

Machine learning is being utilized across various industries to classify behavioral patterns and predict future events. In the realm of breast cancer diagnosis, extensive research has been conducted using a range of ML techniques, including Linear Regression, Decision Tree Regression, Random Forest classification, Light Gradient Boosted Machine, k-neare, and Logistic Regression. For the purpose of comparison and identifying the most efficient algorithm for breast cancer diagnosis, Logistic Regression, Decision Tree, Xgboost, Random Forest classification, and Support Vector Machine have been selected . All of these methods fall under the category of supervised machine learning models.

The research conducted in our study used the Breast Cancer Wisconsin (Diagnostic) Data Set, which was developed by Dr. William H. Wolberg, a physician at the University of Wisconsin Hospital in Madison, Wisconsin, USA. To account for missing and null values, the dataset underwent preprocessing. The "train-test-split" technique was utilized to divide the processed dataset into training and testing subsets. These subsets were then employed

to train and evaluate the selected machine learning algorithms for breast cancer diagnosis. The performance of the algorithms was evaluated using metrics such as accuracy, precision, recall, F1-score and specificity. Finally, compare the obtained results with state-of-the-art methods in the field which used the same data.

.

Chapter 1
Breast Cancer (BC)

Introduction

Breast cancer remains the predominant form of cancer detected among women, accounting for over 10% of new cancer cases annually. It stands as the second leading cause of cancer-related death among women worldwide [4].

Breast cancer typically progresses silently, often detected during routine screenings. However, some individuals may come across a breast lump, changes in breast shape or size, or nipple discharge by chance. Mastalgia, or breast pain, is not uncommon either. To confirm the presence of breast cancer, a comprehensive diagnostic process involving physical examination, imaging techniques (such as mammography), and tissue biopsy is necessary. Early detection significantly enhances the chances of survival, as breast tumors tend to spread through the lymphatic and hematological pathways, resulting in distant metastasis and a poorer prognosis. Consequently, the significance of breast cancer screening programs cannot be overstated [5].

1.1 Cancer

Cancer is a disease characterized by the uncontrolled growth and spread of abnormal cells in the body. It can originate in almost any part of the body, which is composed of countless cells that normally grow, multiply, and replace old or damaged cells. However, when the process of cell division becomes disrupted, abnormal cells can proliferate and form tumors [6].

There are two main categories of cancer:

- **Hematologic (blood)**: are cancers of the blood cells, including leukemia, lymphoma, and multiple myeloma.
- **Solid tumor cancers**: are cancers of any of the other body organs or tissues. The most common solid tumors are breast, prostate, lung, and colorectal cancers [7].

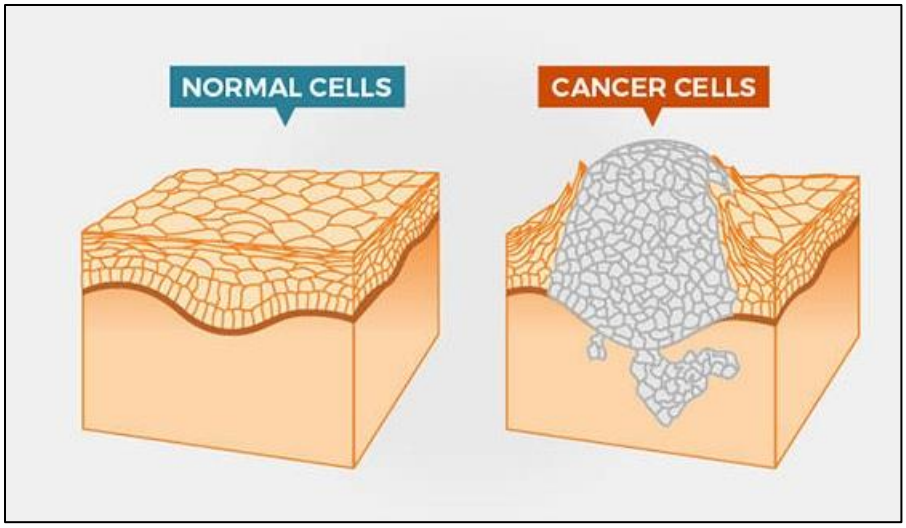


Figure 1. Difference between normal and cancer cells.

1.1.1 Tumor

A tumor is an abnormal mass of tissue that develops when cells divide and grow in an uncontrolled manner or fail to die when they should [6].

- **Benign:** Lumps that are not cancer are called benign (benign tumors do not spread to other parts of the body).
- **Malignan:** Lumps that are cancer are called malignant (malignant do spread to other parts of the body) [8].

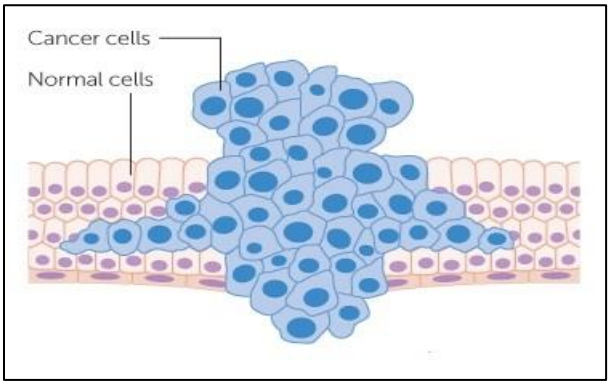


Figure 2. Cancer cells tumor.

1.2 Breast cancer

Structurally, the breast is positioned in front of the chest wall, housing milk-producing glands that rest upon the pectoralis major muscle. Supporting ligaments connect the breast to the chest wall, while 15 to 20 circularly arranged lobes constitute its form. The size and shape of the breast are primarily influenced by the adipose tissue enveloping these lobes. Each lobe consists of lobules, housing glands responsible for milk production when stimulated by hormones [4].

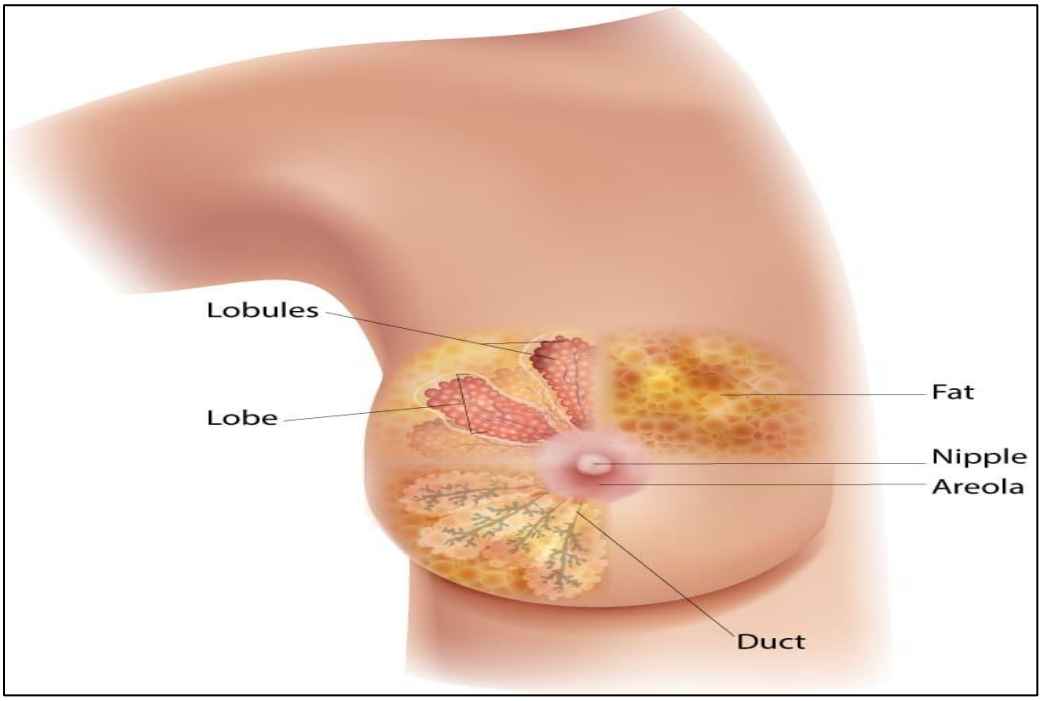


Figure 3. Diagram of the breast.

1.3 Epidemiology

Breast cancer is a significant global health issue, being the most commonly diagnosed cancer and the leading cause of cancer-related deaths in women. It accounts for 23% of total cancer cases and 14% of all cancer-related mortalities worldwide. The lifetime risk of developing breast cancer for women is 1 in 8, with a higher risk among those aged over 65. Interestingly, women aged over 70 have the highest risk, with a 1 in 15 chance of developing breast cancer. The number of elderly patients with breast cancer is expected to increase rapidly in the future, as the population aged over 65 is projected to reach over 20% by 2030. Improved disease screening and early detection methods have led to more cases

being diagnosed at an earlier age. As a result, there is an increasing number of patients, particularly elderly individuals, requiring long-term treatment and management for breast cancer. This review aims to explore the impact of modern investigations and treatment options for breast cancer, categorized by age group, and how these trends may shape future research and treatment approaches [9].

1.4 Breast Cancer Symptoms

The primary and most common symptom of breast cancer is the presence of a new lump or mass in the breast. It's important to note that the majority of breast lumps are not cancerous. However, if a lump is painless, feels hard, and has irregular edges, there is a higher likelihood that it could be cancerous. It's worth mentioning that breast cancers can also manifest as soft, round, tender, or even painful masses.

In addition to lumps, there are other potential symptoms of breast cancer that individuals should be aware of. These can include:

- Swelling of all or part of a breast (even if no lump is felt).
- Skin dimpling (sometimes looking like an orange peel).
- Breast or nipple pain.
- Nipple retraction (turning inward).
- Nipple or breast skin that is red, dry, flaking, or thickened.
- Nipple discharge (other than breast milk).
- Swollen lymph nodes under the arm or near the collar bone (Sometimes this can be a sign of breast cancer spread even before the original tumor in the breast is large enough to be felt) [6].

1.5 Breast cancer causes

While the exact causes of breast cancer remain elusive, the fully understanding of it is incomplete, which makes it challenging to determine why some women develop breast cancer while others do not.

Nevertheless, certain risk factors have been identified that can influence the likelihood of developing breast cancer. While some of these factors are beyond our control, there are others that we can modify to some extent [10].

1.5.1 Natural causes

- Age**:** Being 55 or older increases your risk for breast cancer.
- Sex**:** Women are much more likely to develop breast cancer than men.
- Family history and genetics**:** If you have parents, siblings, children or other close relatives who have been diagnosed with breast cancer, you are more likely to develop the disease at some point in your life. About 5% to 10% of breast cancers are due to single abnormal genes that are passed down from parents to children, and that can be discovered by genetic testing [11].

1.5.2 Hormones and hormone medicine

- Exposure to oestrogen: a female hormone can stimulate the growth of breast cancer cells. Factors such as early menstruation, late menopause, and not having children or having them later can increase the risk due to prolonged exposure to estrogen.
- Hormone replacement therapy (HRT): is associated with an increased risk of breast cancer, especially when used for longer than 1 year.
- Contraceptive pill: The use of contraceptive pills slightly increases the risk of breast cancer, but the risk decreases after discontinuation and returns to normal after 10 years [12].

1.5.3 Lifestyle factors

Certain lifestyle choices may increase the risk such as:

- Smoking: The use of tobacco is associated with an increased risk of various cancers, including breast cancer.
- Alcohol use: Research suggests that alcohol consumption can raise the risk of certain types of breast cancer.
- Obesity: Being obese is known to increase the risk of developing breast cancer and the likelihood of breast cancer recurrence [11].

1.5.4 Radiation Exposure

Previous radiation therapy to the chest area, particularly during childhood or adolescence, can elevate the risk of developing breast cancer later in life [6].

1.6 Breast cancer diagnosis

Breast cancer diagnosis involves a combination of methods and tests that healthcare providers use to evaluate and determine the presence of breast cancer. Here are some of the commonly used methods for breast cancer diagnosis:

1.6.1 Physical Examination

A healthcare provider performs a thorough examination of the breasts and surrounding areas to check for any abnormalities, such as lumps, changes in size or shape, skin changes, or nipple discharge.

1.6.2 Clinical Breast Imaging Techniques

1.6.2.1 MRI (Magnetic Resonance Imaging)

MRI is a medical imaging technique that produces images of different cross-sections by utilizing a strong magnetic field in combination with radiofrequency (RF) signals. To enhance the resolution of MRI images, a contrast agent can be administered. Breast MRI is recommended for individuals with a high risk of breast cancer, but it is not generally recommended for the general population due to various reasons.

The limitations of breast MRI include a high false-positive rate, elevated cost, time-consuming nature, limited availability of units, the requirement for experienced radiologists, and a lack of significant clinical utility. The American Cancer Society (ACS) has issued guidelines regarding the use of MRI as an additional tool alongside mammography. They recommend annual MRI tests for specific population groups, such as those carrying BRCA mutations or individuals at high risk of breast cancer.

In comparison to mammography and ultrasound, MRI exhibits higher sensitivity in detecting small tumors in individuals with a high risk of breast cancer. However, it is less specific, which means it may generate more false-positive results. Despite these

considerations, breast MRI serves as a valuable adjunct to other imaging methods in the targeted population groups mentioned above [13].

Figure 4. Medical-imaging technique MRI.

1.6.2.2 Ultrasound

Breast ultrasonography is a widely available and cost-effective screening tool used to detect tumors by utilizing acoustic waves that are bounced off breast tissue. It involves the application of an ultrasound transducer to measure the reflected acoustic waves and identify the structure of the human breast. While breast ultrasonography is effective in identifying cysts and solid masses and increases cancer detection rates for individuals at high risk of breast cancer, it is not as efficient as mammography.

In the case of subjects with high breast cancer risk, pregnant women, and those unable to undergo mammography, breast ultrasonography has been recommended as a supplementary screening method alongside mammography. When used as a supplement to mammography, breast ultrasonography enhances imaging sensitivity but at the cost of reduced specificity and increased biopsy rates. However, it should be noted that breast ultrasonography may fail to detect certain tumors due to the similarity in acoustic properties between healthy and cancerous tissues. Additionally, the accuracy of breast ultrasonography heavily relies on the expertise of experienced radiologists, which can significantly impact its sensitivity and specificity [14].

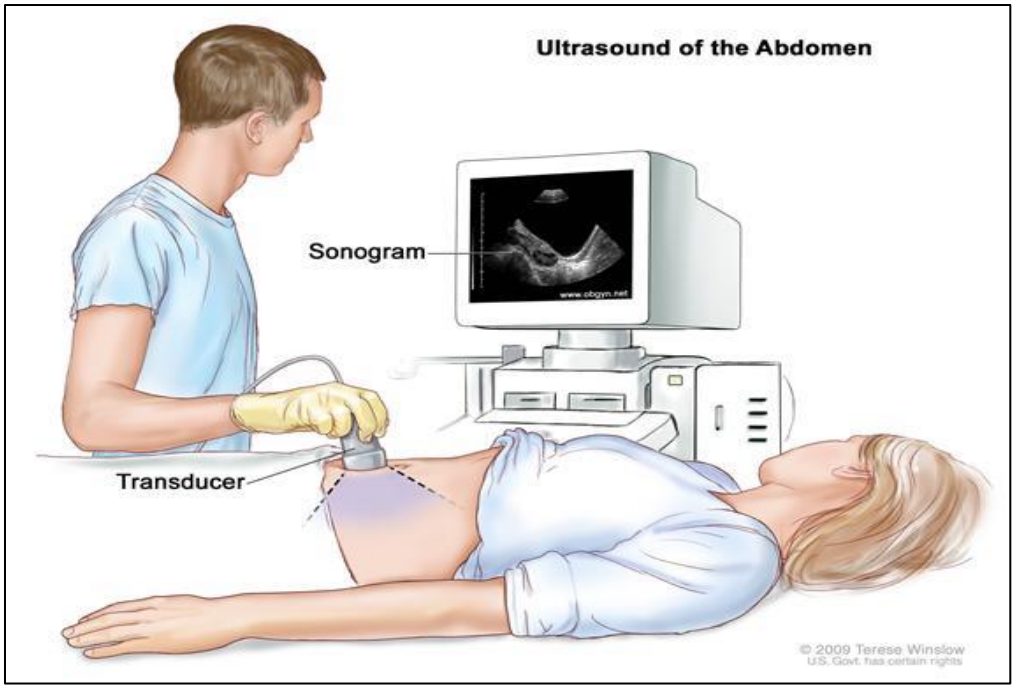


Figure 5. Medical imaging technique Ultrasound.

1.6.2.3 Mammography

The patient's breast is placed on a flat support plate and compressed with a parallel plate called a paddle. An x-ray machine produces a small burst of x-rays that pass through the breast to a detector located on the opposite side. The detector can be either a photographic film plate, which captures the x-ray image on film, or a solid-state detector, which transmits electronic signals to a computer to form a digital image. The images produced are called mammograms [15].

The American Cancer Society (ACS) recommends annual mammograms for females starting at age 40, with particular benefits for those between 40 and 74 years old. However, mammography has relatively high false-positive and false-negative rates, especially for patients with dense breasts, such as those under 40 years old .The sensitivity of mammography is influenced by factors such as age, ethnicity, personal history, radiologist's experience, and technique quality. It may have reduced sensitivity in women with dense breasts and those who are premenopausal. Mammography is associated with several drawbacks, including the use of ionizing radiation, limited suitability for subjects

with dense breasts, relatively high false-positive and false-negative rates, and discomfort during the examination. In fact, studies suggest that mammography only reduces breast cancer death rates by 0.0004%, raising questions about its overall usefulness .

Recently, contrast-enhanced (CE) digital mammography has emerged as an adjunct breast screening tool to traditional mammography. It relies on detecting tumor angiogenesis and involves intravenous iodinated contrast injections, resulting in slightly higher radiation exposure compared to conventional mammography. CE mammography improves sensitivity and performance compared to mammography and ultrasound, and it has shown enhanced detection accuracy when compared to mammography alone [16].

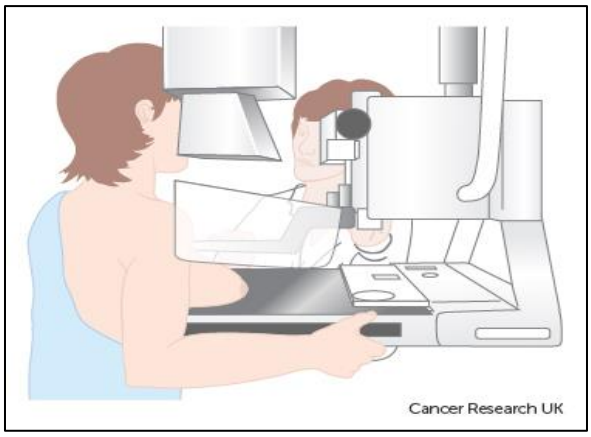


Figure 6. Medical imaging technique mammography.

1.6.2.4 Biopsy

A breast biopsy is a test that removes tissue or sometimes fluid from the suspicious area. The removed cells are examined under a microscope and further tested to check for the presence of breast cancer [17]. A biopsy is the only diagnostic procedure that can definitely determine if the suspicious area is cancerous. There are several types of breast biopsies, including needle biopsy, vacuum-assisted biopsy, fine needle aspiration, punch biopsy, and wire-guided excision biopsy.

- Needle biopsy: This is the most common type and involves using a hollow needle to remove small samples of breast tissue. It can be guided by imaging techniques such as ultrasound or mammogram.

- Vacuum-assisted biopsy: This procedure uses a special needle connected to a vacuum device to remove breast tissue. It allows for the removal of a larger sample and is guided by imaging techniques.
- Fine needle aspiration: It involves using a thin needle to extract fluid and cells from the breast tissue. It is often guided by ultrasound imaging and may not require anesthesia.
- Punch biopsy: A small cutting device is used to remove a sample of breast tissue from the layers just beneath the skin.
- Wire-guided excision biopsy: This type of biopsy is performed when an abnormal area is detected on imaging but cannot be felt during a physical examination. A thin wire is inserted into the breast to guide the surgeon during surgery to remove the specific area [18].

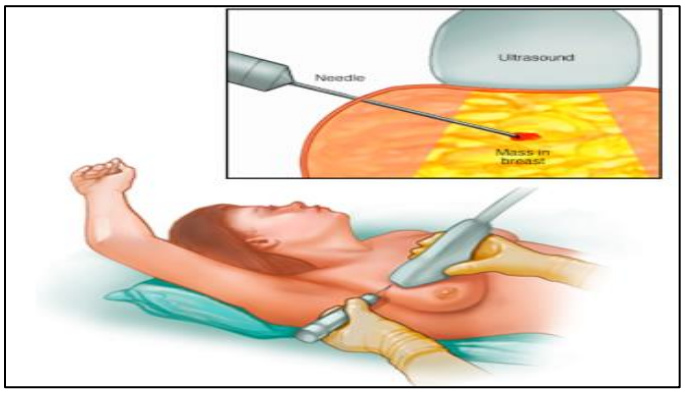


Figure 7. Breast biopsy test.

1.6.3 Molecular and Genetic Testing

These tests evaluate specific genes and molecules in the breast tissue to provide additional information about the tumor, its aggressiveness, and potential treatment options. Examples include testing for hormone receptors (estrogen and progesterone receptors) and HER2/neu gene amplification [6].

1.7 Classification

1.7.1 Histopathologic Types

Breast cancer is primarily classified based on its histological appearance. Most breast cancers originate from the epithelium lining the ducts or lobules, categorized as ductal or

lobular carcinoma. Carcinoma in situ refers to the growth of low-grade cancerous or precancerous cells within a specific tissue compartment without invading surrounding tissues. Invasive carcinoma spreads beyond the initial tissue compartment [19].

1.7.2 Grade

Grading compares the appearance of breast cancer cells to normal breast tissue. Normal cells in the breast are differentiated, meaning they have specific shapes and forms reflecting their organ function. Cancerous cells lose this differentiation. Grading is based on the degree of disorganization, uncontrolled cell division, and changes in cell nuclei.

- G1 (Low Grade): The cells appear to be growing slowly and resemble normal cells to some extent.
- G2 (Medium Grade): The cells look more abnormal than low-grade cells.
- G3 (High Grade): The cells appear highly abnormal and are more likely to grow quickly.

1.7.3 Stage

Breast cancer staging utilizes the TNM system, considering the :

- T (Tumor): It indicates the size of the tumor.
- N (Node): It signifies whether the cancer has spread to the nearby lymph nodes.
- M (Metastasis): It determines if the cancer has spread to distant parts of the body.

Larger tumors, nodal spread, and metastasis correspond to higher stage numbers and a worse prognosis. The main stages include:

- Stage 0: Ductal carcinoma in situ (DCIS), sometimes referred to as stage 0, where abnormal cells are present but have not spread beyond the milk ducts.
- Stage 1: The tumor measures less than 2cm and has not spread to the lymph nodes or other parts of the body.
- Stage 2: The tumor measures 2 to 5cm, may or may not have spread to the lymph nodes in the armpit, but has not spread elsewhere.
- Stage 3: The tumor measures 2 to 5cm, may be attached to surrounding tissues, and has affected the lymph nodes in the armpit, but has not spread beyond.

- Stage 4: The tumor can be any size, and cancer has spread to other parts of the body (metastasis).

Each stage can further be divided into subcategories (A, B, C) for more precise classification[6, 12].

1.7.4 Receptor Status

Breast cancer cells have receptors on their surface and in their cytoplasm and nucleus. Estrogen receptor (ER), progesterone receptor (PR), and HER2 are three important receptors. ER+ cancer cells depend on estrogen for growth and can be treated with drugs blocking estrogen effects. HER2+ breast cancers are generally more aggressive but can respond to drugs like trastuzumab. Breast cancers lacking these three receptors are called triple-negative, although they may express receptors for other hormones [13,14].

1.7.5 DNA Assays

DNA testing, such as DNA microarrays, compares normal cells to breast cancer cells. Specific changes in the DNA can be used to classify the cancer and guide treatment selection based on the DNA type.

1.8 Types of breast cancer

Breast cancer is a complex disease that can be classified into different types based on various factors, including the characteristics of the cancer cells and the presence or absence of specific receptors .the types of breast cancer can be classified into invasive and non-invasive types:

1.8.1 Benign Breast cancer

- Fibroadenoma: A common benign tumor composed of glandular and fibrous tissue. It usually feels like a firm, smooth, and rubbery lump in the breast [22].
- Fibrocystic Changes: Noncancerous changes in the breast characterized by the presence of cysts, fibrous tissue, and glandular changes. It can cause breast pain, tenderness, and lumpiness [23].
- Adenosis: A condition where the lobules of the breast become enlarged due to an increased number of glandular cells [24].

- Phyllodes Tumor: A rare tumor that develops in the connective tissue of the breast. It can be benign, borderline, or malignant [19].

1.8.2 Malignant Breast Cancer

In the malignant type of cancer , it can be classified into invasive and non-invasive types.

1.8.2.1 Non-invasive Breast Cancer

- Ductal Carcinoma In Situ (DCIS): DCIS is a non-invasive breast cancer where abnormal cells are found in the lining of the milk ducts. It is considered the earliest stage of breast cancer and has not spread to nearby tissues [26].
- Lobular Carcinoma In Situ (LCIS): LCIS is not a true cancer but an indication of an increased risk of developing invasive breast cancer. It starts in the milk-producing glands (lobules) of the breast [27].

1.8.2.2 Invasive Breast Cancer

- Invasive Ductal Carcinoma (IDC): IDC is the most common type of breast cancer. It starts in the milk ducts and invades the surrounding breast tissue. It may also spread to other parts of the body if left untreated [28].
- Invasive Lobular Carcinoma (ILC): ILC begins in the milk-producing lobules and then invades the nearby breast tissue. It accounts for about 10-15% of invasive breast cancers [29].
- Triple-Negative Breast Cancer (TNBC): TNBC is a subtype of breast cancer that lacks estrogen receptors (ER), progesterone receptors (PR), and human epidermal growth factor receptor 2 (HER2) protein. It tends to be more aggressive and may require different treatment approaches [30].
- Hormone Receptor-Positive (HR+) Breast Cancer: This type of breast cancer is characterized by the presence of estrogen receptors (ER) and/or progesterone receptors (PR) on the cancer cells. It can be treated with hormone therapy that targets these receptors [31].
- HER2-Positive Breast Cancer: HER2-positive breast cancer has an overexpression of the human epidermal growth factor receptor 2 (HER2) protein on the surface of

cancer cells. It tends to be more aggressive but can be treated with targeted therapies that block HER2 [32].

1.9 Statistics

These statistcs are from the latest world health organization cancer report of the year 2020 .

1.9.1 Breast Cancer Statistics Worldwide

- In 2020, there were an estimated 2.3 million new cases of breast cancer worldwide.
- Breast cancer accounted for approximately 11.7% of all new cancer cases.
- It was responsible for around 6.9% of all cancer-related deaths globally [33].
    - Breast cancer is the most commonly diagnosed cancer and the leading cause of cancer-related deaths in women worldwide.
    - It accounts for approximately 23% of total cancer cases and 14% of all cancer-related mortalities [34].

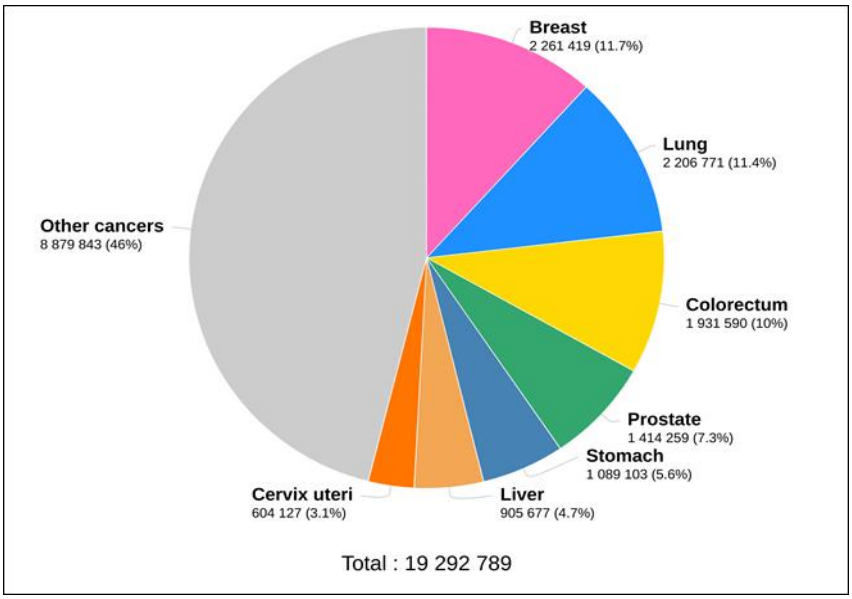


Figure 8. Estimated Number of New Cases in 2020 Worldwide, all Ages, both Sexes.

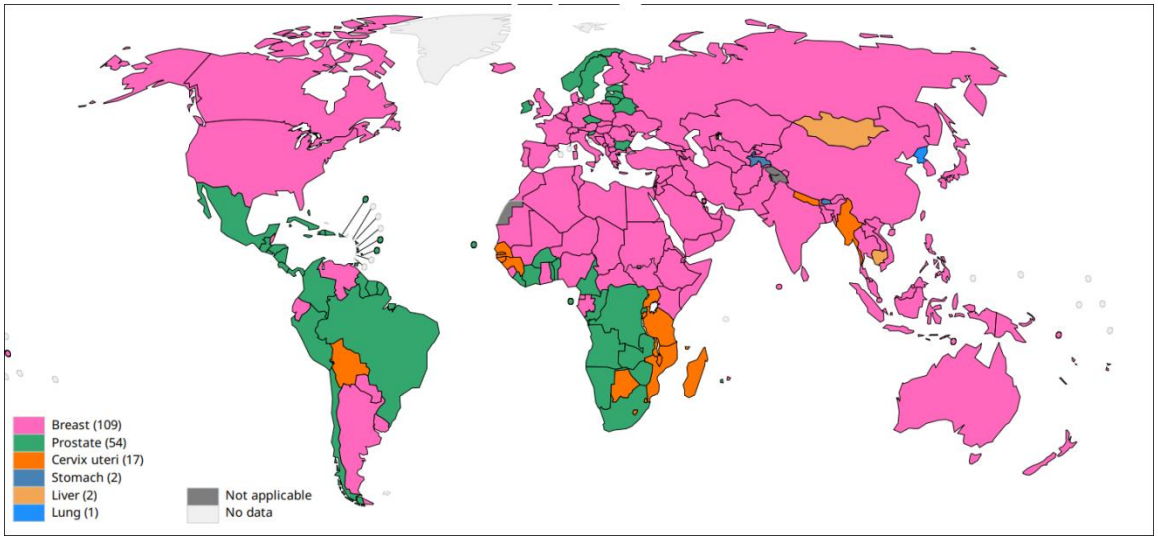


Figure 9. Top cancer per country estimated age-standardized incidence rates (World) in 2020, both sexes, all ages [35].

1.9.2 Breast Cancer Statistics in Algeria

Breast cancer is also a significant health concern in Algeria, as in many other countries. According to the Globocan 2020 report:

- breast cancer is the most common cancer among women in Algeria.
- In Algeria, breast cancer is estimated to account for about 30% of all cancer cases in women.
- The incidence rate of breast cancer in Algeria varies based on different reports and studies, ranging from 25 to 40 cases per 100,000 women.

The mortality rate due to breast cancer in Algeria is also a concern, and efforts are being made to improve early detection and access to appropriate treatment [34].

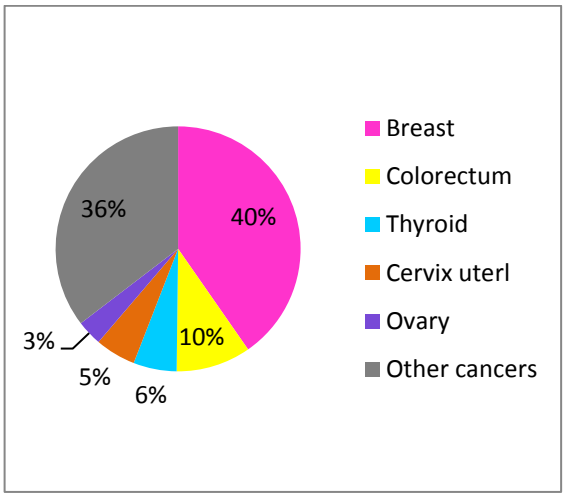


Figure 10. Number of New Cases in females, all Ages, in 2020.

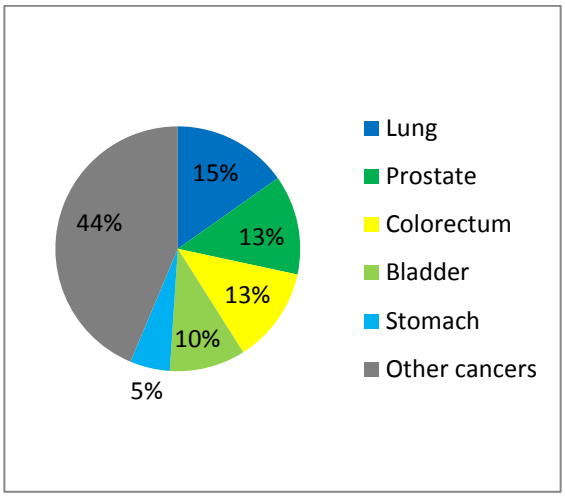


Figure 11. Number of New Cases in males, all Ages, in 2020.

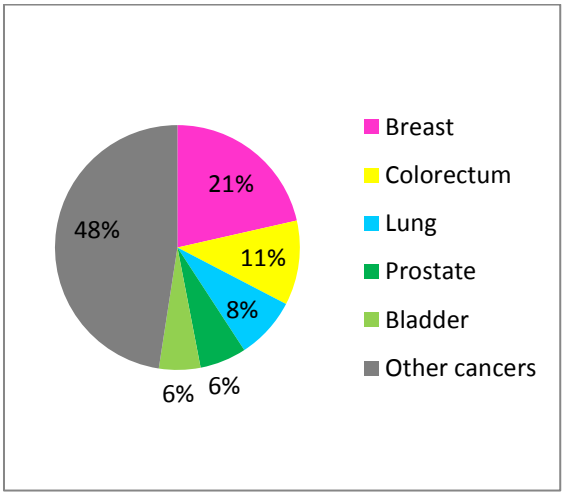


Figure 12. Number of New Cases in 2020, all Ages, both Sexes.

	Males	Females	Both sexes
Population	22,153,808	21,697,235	43,851,043
Number of new cancer cases	27,328	31,09	58,418
Age-standardized incidence rate (World)	129.7	141.7	135.3
Risk of developing cancer before the age of 75 years (%)	13.7	14.1	13.9
Number of cancer deaths	17,902	14,9	32,802

Age-standardized mortality rate (World)	84.2	68.5	76.1
Risk of dying from cancer before the age of 75 years (%)	8.6	7.3	7.9
5-year prevalent cases	64,379	86,011	150,39
Top 5 most frequent cancers excluding non-melanoma skin cancer (ranked by cases)	Lung Prostate Colorectum Bladder Stomach	Breast Colorectum Thyroid Cervix uteri Ovary	Breast Colorectum Lung Prostate Bladder

Table 1. Statistics summary.

Total population	Number of new cases	Number of deaths	Number of prevalent cases (5year)
43,851,043	58,418	32,802	150,390

Table 2. Numbers at a glance in Algeria.

Conclusion

This chapter on breast cancer provides an overview of the global and regional impact of this disease. It highlights that breast cancer is the most commonly diagnosed cancer and the leading cause of cancer-related deaths in women worldwide, as well as general statistics and facts about this disease : its signs, symptoms, causes and methods on how is diagnosed .

Based on screening and tests provided by the patients ,we can create a dataset which will help us a lot in the prevention and the classification of the disease using artificial intelligence, and machine learning techniques which we are going to be talking about in details in the next chapter.

Chapter 2

Machine Learning Techniques

Introduction

In today's information age, the concepts of artificial intelligence, machine learning and deep learning are no longer as lofty as they were more than ten years ago. They have gradually penetrated into every corner of life and become understood and used by everyone of us , devices such as GPS, voice assistant, Google maps available on the Internet is a typical representative of the application of artificial intelligence in life.

In the healthcare industry, there is a growing accumulation of large data sets, often stored as unstructured data in electronic health records (EHRs) [36]. To extract valuable insights from this data, machine learning with its techniques and algorithms are employed to reorganize the information into structured sets. This transformation enables healthcare professionals to efficiently derive actionable insights from the data, facilitating better decision-making and improving patient care.

This chapter provides a concise overview of the fundamental principles of using AI in healthcare, with a specific focus on machine learning techniques used in our breast cancer classification models where In this section, we will discuss various types of machine learning (ML) techniques and explore their functionalities as well as their importance.

2.1 Artificial Intelligence

Artificial intelligence (AI) is a field that combines computer science with the study of human intelligence. It focuses on developing computer systems that can exhibit intelligent behavior and achieve goals in a manner similar to humans [37]. Intelligence encompasses various abilities such as thinking, imagination, memory, understanding, pattern recognition, decision-making, adaptation, and learning from experience [38]. AI aims to make computers behave in a more human-like fashion, but with greater efficiency and speed compared to humans [39].

2.2 Stages of AI development

The development history of artificial intelligence can be summarized into three stages: incubation, formation, and development.

2.2.1 Incubation (pre-1956)

The pre-1956 period, referred to as "Incubation," highlights the contributions of thinkers like Aristotle, Francis Bacon, Gottfried Leibniz, George Boole, and Alan Turing, who laid the foundations for AI with their work on formal logic, reasoning computation, and symbolic language [40].

2.2.2 Formation (1956)

The formation of AI is marked by the Dartmouth workshop in 1956, where the term "artificial intelligence" was officially adopted. Notable achievements during this period include machine learning with the development of a checkers-playing program, theorem proving using computers, early neural network-based pattern recognition, and the creation of the General Problem Solver (GPS) for solving diverse problems [41].

2.2.3 Development (after 1960s)

This is the year where the field expanded globally. Expert systems and natural language processing gained prominence during the 1980s, with notable advancements in molecular structure analysis and machine translation. The establishment of international conferences and journals, such as the International Joint Conference on Artificial Intelligence (IJCAI), provided platforms for researchers to exchange ideas and further drive AI progress.

However, the path of progress in AI has not been without challenges. Despite significant achievements, researchers faced difficulties and encountered limitations. For instance, early machine translation systems failed to capture the complexities of language, resulting in flawed translations. Nevertheless, these challenges sparked further research and exploration, leading to ongoing advancements in the field of artificial intelligence [42].

2.3 Basic examples of AI

AI has been dominate in various fields from engineering to medicine to even gaming Here are some basic examples of AI applications that you may encounter in everyday life:

- Virtual personal assistants.
- Recommendation systems.
- Spam filters employ.
- Image recognition.
- Chatbots.
- Autonomous vehicles.
- Fraud detection.
- Language translation.
- Gaming.

2.4 Main branches of AI

The main branches of AI, or artificial intelligence, are:

- **Natural Language Processing (NLP)**

NLP is a multidisciplinary field that focuses on the research and application of techniques for computers to comprehend and manipulate natural language text or speech in order to accomplish practical objectives. NLP researchers strive to gain insights into how humans comprehend and utilize language, enabling the development of appropriate tools and methodologies for computer systems to understand and manipulate natural languages to perform desired tasks [43].

- **Computer Vision**

Computer vision is concerned with enabling computers to understand and interpret visual information, such as images and videos. It involves tasks such as image recognition, object detection, image segmentation, and video analysis. Computer vision enables machines to perceive and analyze visual data like humans [44] .

- **Robotics**

Robotics combines AI with physical systems to create intelligent machines that can interact with the physical world. It involves designing and programming robots to perform tasks autonomously or with human guidance. Robotics aims to create machines that can perceive the environment, make decisions, and manipulate objects effectively[45] .

2.5 AI in healthcare

AI applications in healthcare encompass a wide range of areas, providing support and enhancement to healthcare personnel rather than replacing their work. Some of the major applications of AI in healthcare include:

- Administrative Workflows: AI can automate administrative tasks such as scheduling appointments, managing electronic health records (EHRs), and streamlining billing and coding processes. This helps reduce the burden on healthcare staff and improves overall operational efficiency.
- Clinical Documentation and Patient Outreach: AI can assist in clinical documentation by automatically extracting relevant information from patient records and generating summaries or reports. It can also facilitate patient outreach by analyzing data to identify individuals at risk and providing personalized recommendations or interventions.
- Image Analysis: AI algorithms can analyze medical images such as X-rays, MRIs, and CT scans, assisting in the detection of abnormalities and aiding in diagnosis. This can help radiologists and other healthcare professionals in interpreting complex images more accurately and efficiently.

- Robotic Surgery: AI-powered robotic systems enable precise and minimally invasive surgical procedures. Surgeons can use robotic assistance to enhance their skills, perform delicate surgeries with greater precision, and reduce the risk of complications.
- Virtual Assistants: AI-based virtual assistants can provide patient support, answer questions, and offer basic medical advice. They can also assist healthcare professionals by retrieving relevant information quickly and accurately, allowing them to make informed decisions.
- Clinical Decision Support: AI algorithms can analyze large amounts of patient data, medical literature, and clinical guidelines to provide evidence-based recommendations for diagnosis, treatment plans, and medication selection. This supports healthcare providers in making well-informed decisions.
- Connected Machines and Dosage Error Reduction: AI can facilitate the connectivity of medical devices and machines, enabling real-time monitoring and data analysis. This helps in early detection of abnormalities and reduces the risk of medication errors by ensuring accurate dosages.
- Drug Development: AI algorithms can analyze vast amounts of biomedical data and accelerate the drug discovery and development process. AI can aid in identifying potential drug targets, predicting drug efficacy, and optimizing clinical trial designs, leading to faster and more efficient drug development.
- Ambient Assisted Living (AAL): AI technologies can be integrated into smart home environments to support independent living for the elderly and individuals with chronic conditions. These systems can monitor vital signs, detect falls, remind

individuals to take medication, and provide emergency assistance [46] .

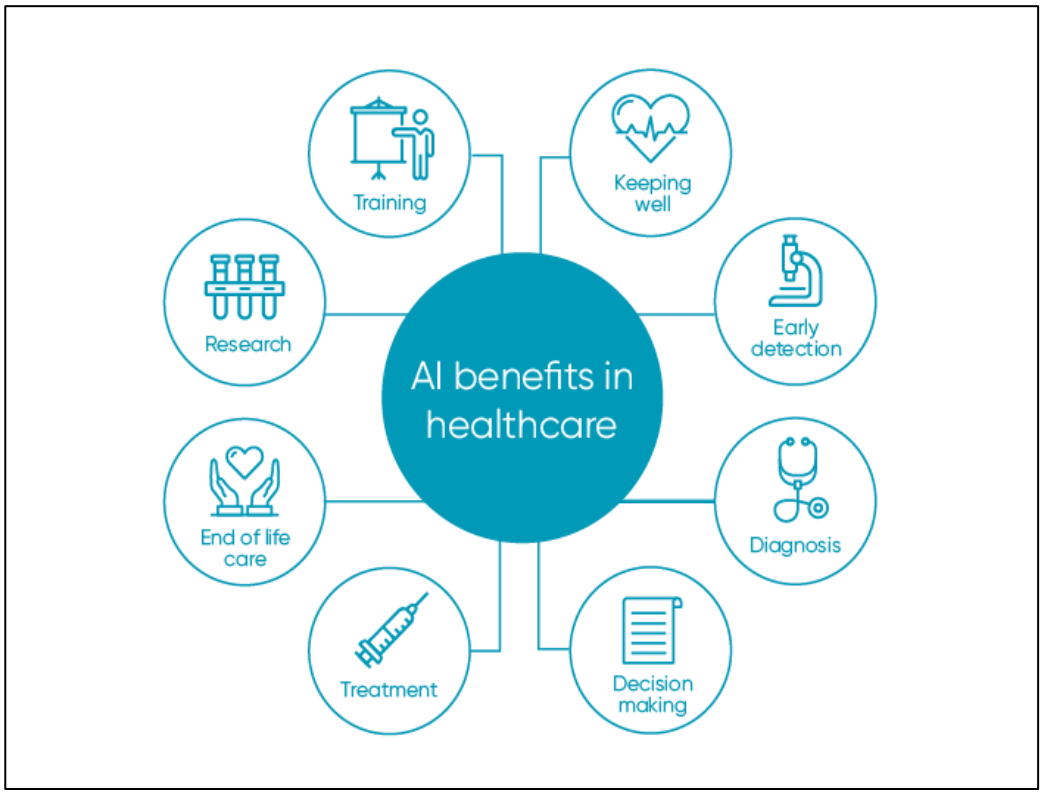


Figure 13. Benefits of AI in healthcare.

2.6 AI, ML, and DL

AI is the broadest concept of the three. It involves endowing machines with human-like intelligence. ML is a subfield of AI, and DL is a subfield of ML that emphasizes the use of deep neural networks which are neural network models with multiple hidden layers [47].

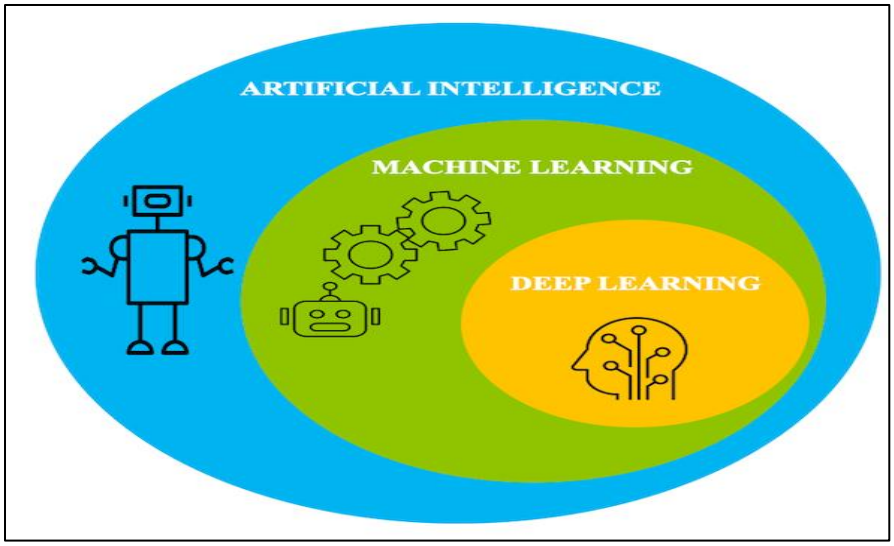


Figure 14. Relation between AI, ML, and DL.

2.7 Machine Learning

Machine learning (ML) is a discipline focused on developing techniques that enable machines to acquire knowledge and enhance their performance on specific tasks through data analysis. Instead of relying on explicit programming, ML algorithms construct models based on sample data, known as training data, to make predictions or decisions. By utilizing training data, machine learning algorithms can identify patterns and relationships within the data, allowing them to generalize and make accurate predictions on new, unseen data. The process involves extracting meaningful features from the input data and learning the underlying patterns to generate predictions or make informed decisions [48].

2.8 Deep learning

Deep learning is a form of machine learning that enables computers to learn from experience and understand the world in terms of a hierarchy of concepts. Because the computer gathers knowledge from experience, there is no need for a human computer operator formally to specify all of the knowledge needed by the computer. The hierarchy of concepts allows the computer to learn complicated concepts by building them out of simpler ones; a graph of these hierarchies would be many layers deep [49].

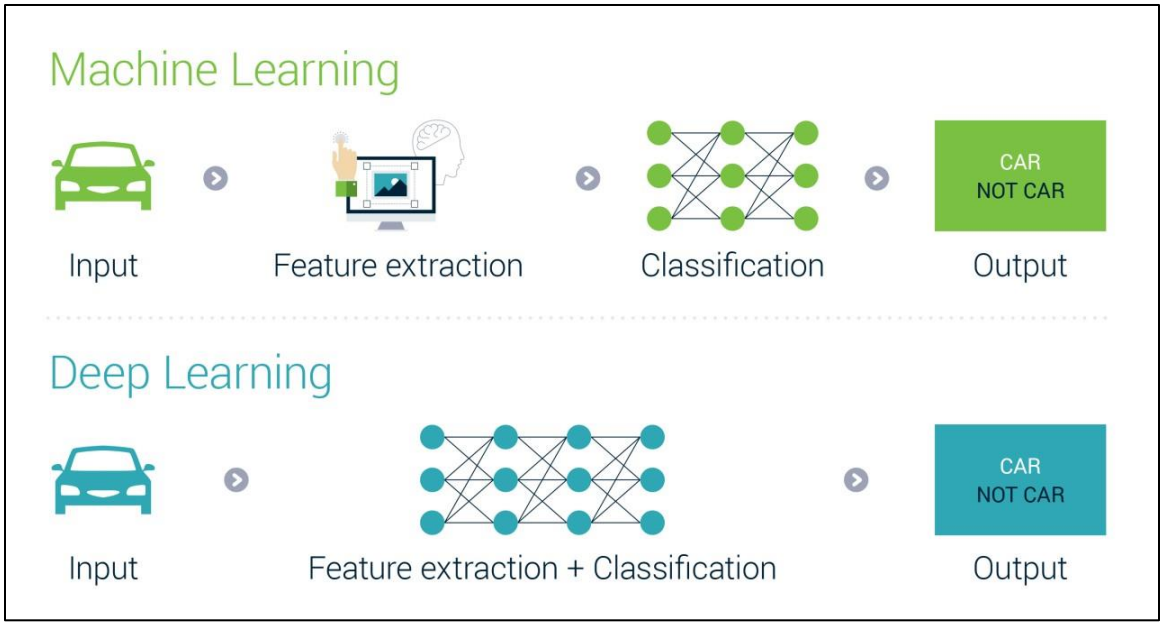


Figure 15. Diffrence between Machine learning Deep learning.

2.9 Proposed system workflow

The proposed system workflow, encompasses the step-by-step process involved in developing a machine learning model. Here is a general outline of the machine learning workflow:

1. **Data Collection**: Gather the relevant data required to train and evaluate your machine learning model. Ensure the data is representative, comprehensive, and accurately labeled.

2. **Data Preprocessing**: Clean and preprocess the collected data to handle missing values, outliers, and inconsistencies. This step may involve data cleaning, normalization, feature scaling, and handling categorical variables.

3. **Data Splitting**: Divide the preprocessed data into training, validation, and testing sets. The training set is used to train the model, the validation set helps in tuning hyperparameters, and the testing set evaluates the final model performance.

4. **Model Selection**: Choose an appropriate machine learning algorithm or model architecture based on the problem type (classification, regression, clustering, etc.) and the nature of the data.

5. **Model Training**: Train the selected model using the training data. The model learns patterns and relationships in the data to make predictions or classifications.

6. **Model Evaluation**: Assess the performance of the trained model using the validation set. Use suitable evaluation metrics (accuracy, precision, recall, F1-score, etc.) to measure the model's effectiveness.

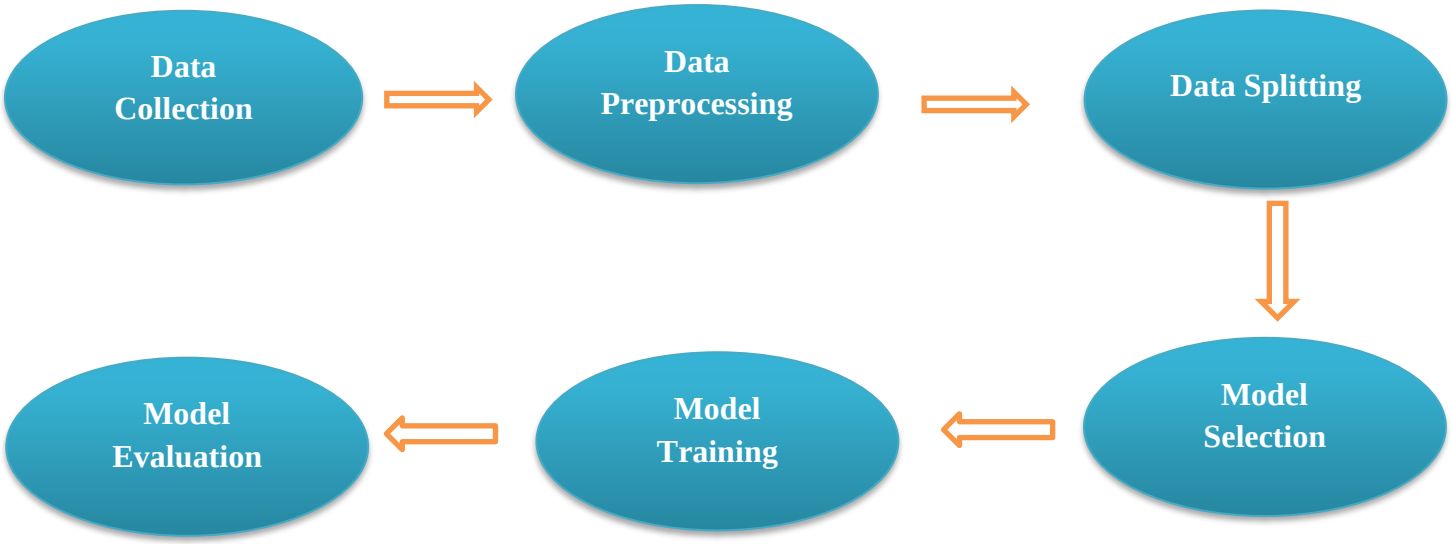


Figure 16. The proposed system's flowchart.

2.10 Types of Machine Learning Algorithms

There are various types of machine learning algorithms, each designed to tackle different learning tasks and data characteristics**.**

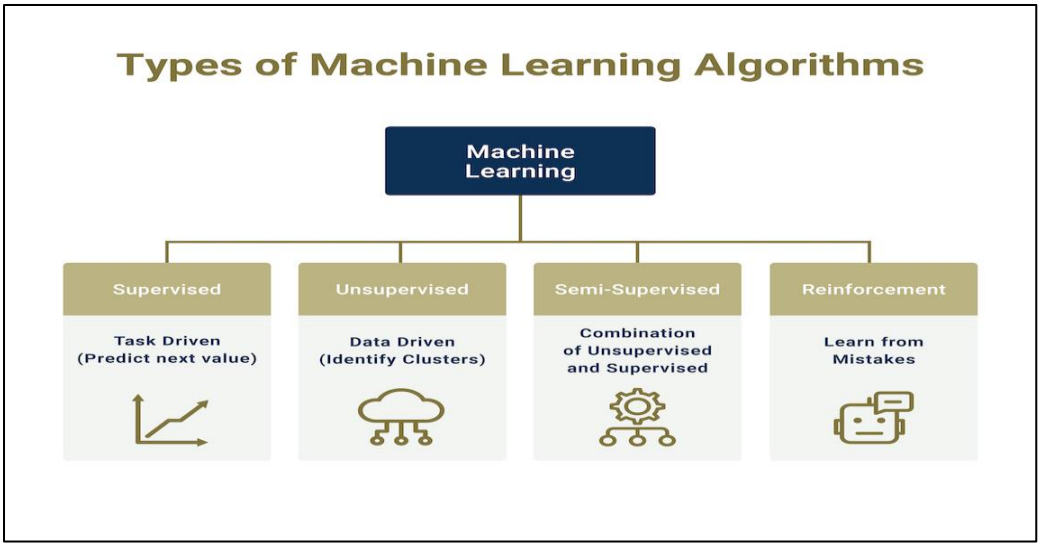


Figure 17. Types of machine learning.

2.10.1 Supervised Learning

Supervised learning involves comparing computed output with expected output to adjust and minimize errors. It aims to produce accurate predictions by learning from labeled data There are two types of supervised learning [50].

2.10.1.1 Regression

The target output can be a single real number or a vector of real numbers. For example, it can represent the predicted price of a stock in 6 months' time or the temperature at noon tomorrow [51].

2.10.1.2 Classification

the target output can be represented as a class label. In the simplest case, this involves making a binary choice between positive and negative [51].

2.10.2 Unsupervised Learning

Unsupervised learning involves learning patterns and structures from input data without labeled examples. It uses clustering algorithms to group similar data points together. An example is Google News, which clusters news stories based on their content to provide collective news stories .Unsupervised learning poses greater challenges as the computer needs to learn and perform tasks without explicit instructions on how to do so.

Consequently, defining the precise goal of this learning process becomes more challenging [52].

2.10.3 Semi-supervised

Semi-supervised learning is situated between unsupervised learning(which lacks any labeled training data) and supervised learning(which relies on fully labeled training data). It bridges the gap by utilizing a combination of labeled and unlabeled data, leading to improved learning accuracy compared to unsupervised learning alone [53].

2.10.4 Reinforcement Learning

Reinforcement learning focuses on an agent's interaction with an environment to maximize long-term rewards. In Reinforcement learning (RL) the output is an action or a sequence of actions and the only supervisory signal is an occasional scalar reward. The basic reinforcement method consists of [54]:

- a set of environment.
- a set of actions.
- rules of transitioning between states.
- rules that determine the scalar immediate reward of a transition.
- rules that describes what the agent observes.

It differs from supervised learning as it does not rely on labeled input/output pairs.

2.10.5 Recommender Systems

Recommender systems use learning techniques to personalize online experiences for users based on their preferences. They can provide ratings and recommendations for products or related items based on user behavior. There are two main approaches: content-based recommendation and collaborative recommendation, which help users discover relevant information and make intelligent recommendations. Many e-commerce sites utilize recommender systems [55].

2.11 ML techniques

There are various machine learning techniques that are commonly used to solve different types of problems, we are going to discuss only those used in our project.

2.11.1 Random Forest

Random forest was introduced by Leo Breiman in 2001 [56]. This algorithm uses a group of classification tress, each of which is built using a bootstrap sample of the data. A graphic overview of random forest classification is given in Figure 18. The figure shows multiple classification trees being used to achieve a classification of an entity. By means of majority voting the final classification of random forest is determined.

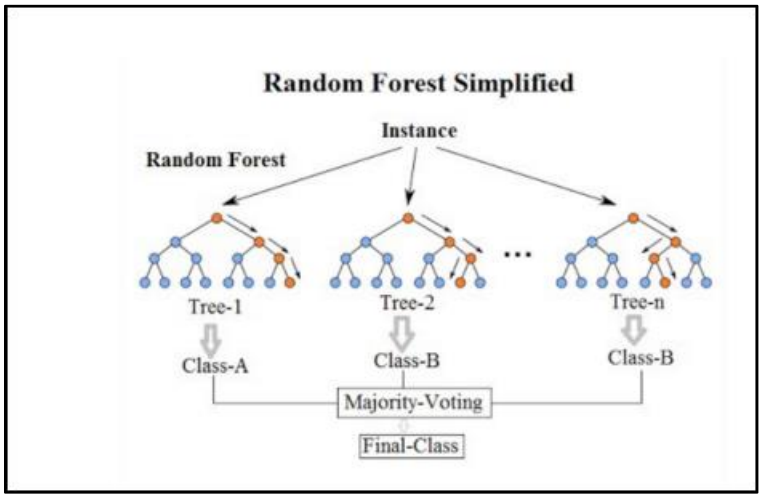


Figure 18. Random Forest structure.

In (Khalilia et al., 2011) [57]. random forest is used to predict the risk of several diseases from the medical diagnosis history of individuals. The authors show that random forest outperforms several other classification algorithms. The authors of (Xu et al., 2017) [58] apply random forest to determine the risk of cardiovascular problems in individual patients and in (Nguyen et al., 2013) [59] this classification method is used for breast cancer diagnosis and prognostic.

2.11.2 Support Vector Machine

Fundamentally, support vector machines (SVMs) search for the optimal separating hyperplane, where the margin between two different objects is maximal. To find this maximal margin, support vectors are used (Yoo et al., 2012) [60]. This concept can be seen in Figure 19, where the dashed lines represent the support vectors and the solid line represents the optimal hyperplane.

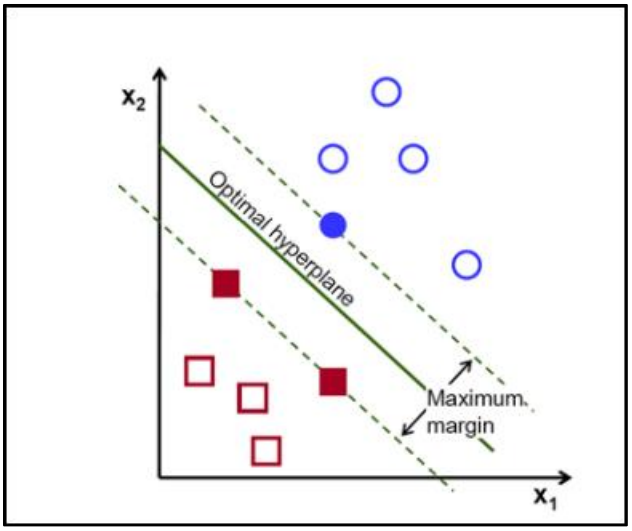


Figure 19. Support Vector Machine oprimal hyperplane.

In (Alajmani and Elazhary, 2019) [61], an SVM classifier is used to predict the likelihood of hospital readmission. This classification technique was shown to be the best performing on this data set. Similarly, the authors of (Zheng et al., 2015) [62], also show good results using an SVM classifier for risk prediction of hospital readmission. In (Rejani and Selvi, 2009) [63], SVM is used for early detection of breast cancer.

2.11.3 Logistic Regression

Logistic regression is a statistical regression model that utilizes a logistic function to represent and model a binary dependent variable. One of the key advantages of this technique is that it offers users explicit probabilities rather than solely providing class label information. By applying the logistic function, logistic regression enables the estimation of the likelihood of an event occurring, allowing for a more nuanced understanding of the data. This characteristic makes logistic regression a valuable tool in various fields, such as classification problems where the outcome is binary [64].

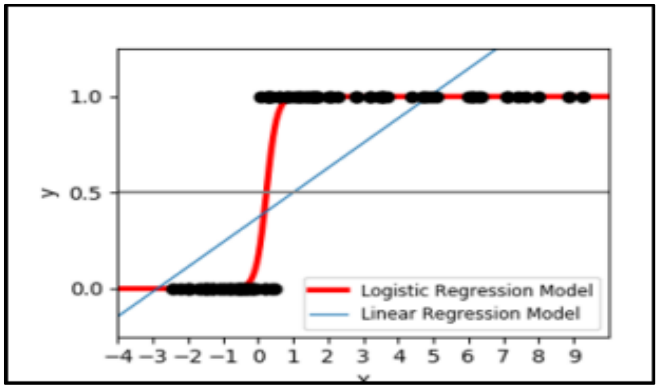


Figure 20. Logistic Regression statistical regression model.

2.11.4 Decision Tree Classifier

Decision tree classifiers build a tree-like structure in which each step involves identifying an attribute that best separates the data into pure partitions based on class values. The goal is to find the attribute that maximizes the homogeneity of each resulting subgroup in terms of class values [65]. By recursively splitting the data based on the selected attributes, decision tree classifiers create a hierarchical model that enables effective classification and prediction tasks. This approach allows for intuitive interpretation and can handle both categorical and numerical data, making decision trees widely used in machine learning and data analysis. A graphic overview of this classifier is given in Figure 21. In each decision node a distinction of the data is made based on its variables. In each leaf node the classification is given [66].

In (Alajmani and Elazhary, 2019) [61], a decision tree classifier is used to predict the likelihood of hospital readmission. Similarly, in (Sushmita et al., 2016) [67], decision tree is used to predict all-cause hospital readmission. In (K. Chen et al., 2014) [68],this classification technique is used for cancer classification using gene expression data.

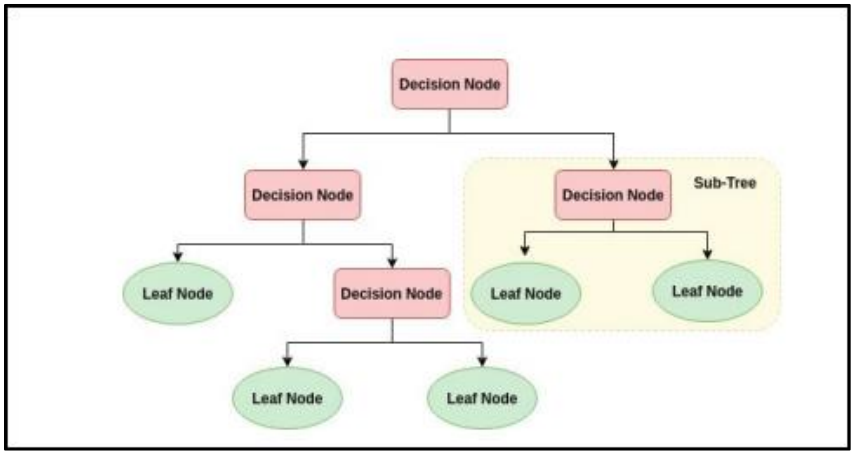


Figure 21. Decision Tree classifier's tree-like structure.

2.11.5 XGBoost (eXtreme Gradient Boosting)

XGBoost is a powerful and widely used Gradient Tree Boosting-based software that has been applied successfully in various research fields. It has gained popularity due to its superior performance, efficient processing of large-scale machine learning tasks, and its ease of use through its Python interface. it has achieved considerable success on both regression and classification problems [69].

The XGBoost algorithm is based on the concept of gradient boosting, where weak learners (usually decision trees) are iteratively added to a model to improve its predictive power. The algorithm optimizes a differentiable loss function by minimizing its value through an additive model, where each weak learner is trained to correct the mistakes of the previous learners. This process continues until a predefined stopping criterion is met [70].

Figure 22. XGBoost Gradient Boosting schematic representation.

Conclusion

This chapter talks about AI, Machine learning and deep learning and has provided an overview of various machine learning techniques and their applications in artificial intelligence. Their cases of use and importance in health care field. We mentioned the five techniques used in our research, explaining how they function and how can we put them in use. These techniques have proven to be powerful tools in solving a wide range of real-world problems.

In the next chapter, we are going to create a classification model based on five machine learning techniques classifiers and later moving to the evaluation metrics to test and comparing the result of this model with the state of the art.

Chapter 3
Application Development

Introduction

This chapter presents the methodology employed to achieve the objectives and address the research questions of the thesis. The experimentation process involved the evaluation and comparison of five machine learning algorithms using the provided dataset. The performance of these algorithms was assessed based on some well-known evaluation metrics such as accuracy, precision, specificity, recall, and F1 score.

3.1 PROCESS FLOW DIAGRAM

A six-phase methodology has been developed to enhance Breast Cancer classification . The methodology consists of the following phases:

1. **Data Collection**: obtained from Kaggle, This data set was created by Dr. William H. Wolberg.
2. **Preprocessing**: Relevant features are carefuly selected while removing irrelevant or incorrect information.
3. **Dataset splitting** : dividing the dataset into two for our classification models : training and testing.
4. **Training the Classification Model** : The preprocessed 10%,20% then 30% of dataset is used to train various machine learning algorithms.
5. **Testing the Classification Model**: The preprocessed 10%,20% then 30% of dataset is used to test various machine learning algorithms.
6. **Model Evaluation (performances)** : The trained model is assessed based on the five essential metrics ensuring the model's efficacy in detecting breast cancer.

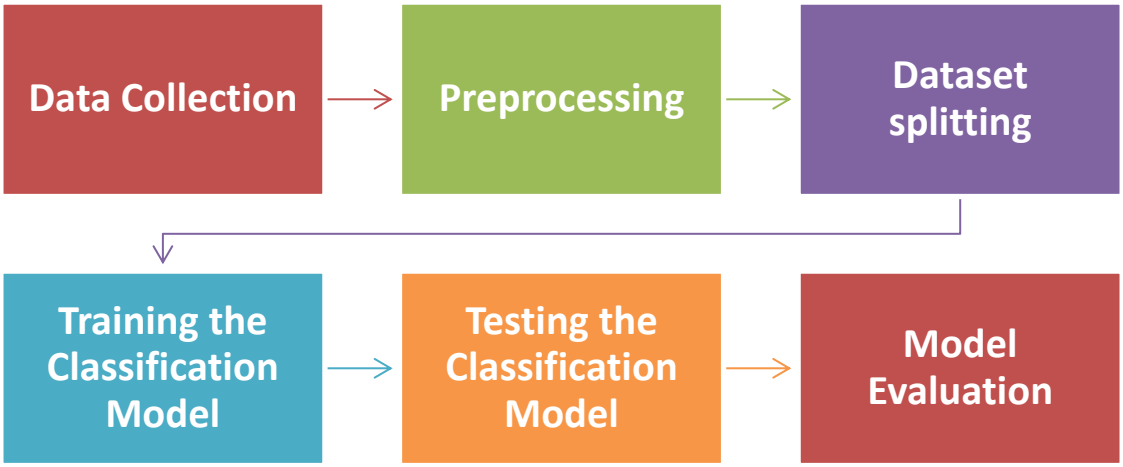
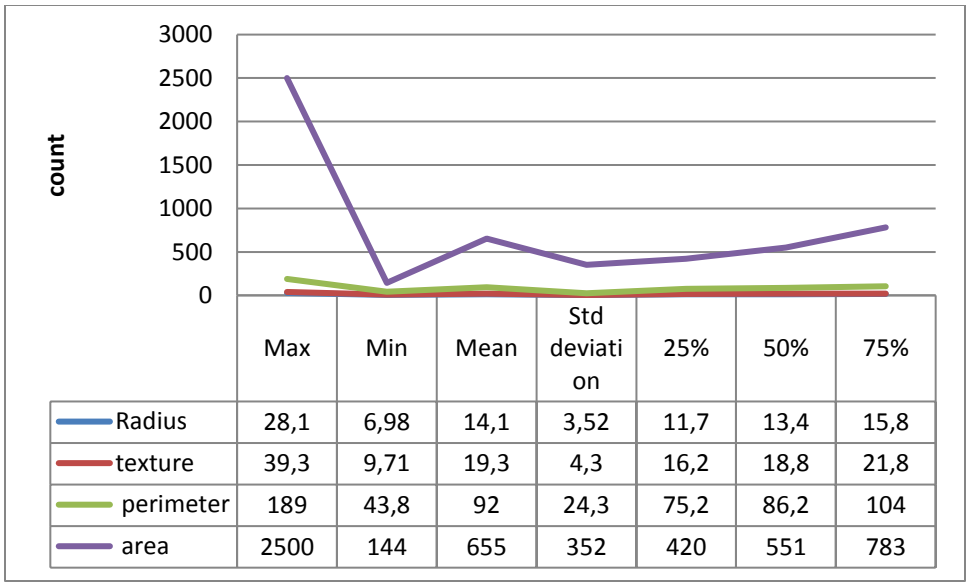


Figure 23. The employed methodology.

3.2 Dataset description

This description of the Wisconsin Diagnostic Breast Cancer (WDBC) is obtained from Kaggle, created by Dr. William H. Wolberg, physician at the University of Wisconsin Hospital at Madison, Wisconsin, USA.

This dataset consists of a total of 569 samples. There are a total of 32 features that characterize our samples, the first of which is the ID of the sample, the second is its class, and the remaining 30 are features that contain various information about the cells. The class label of our samples can be malignant (M) or benign (B). These are medical terms that refer to the benign and malignant tumor cells we talked about earlier.



	Max	Min	Mean	Std deviation	25%	50%	75%
Radius	28,1	6,98	14,1	3,52	11,7	13,4	15,8
texture	39,3	9,71	19,3	4,3	16,2	18,8	21,8
perimeter	189	43,8	92	24,3	75,2	86,2	104
area	2500	144	655	352	420	551	783

Figure 24. Top four features of the dataset.

3.2.1 Data Visualization

graphs are used to represent and interpret data. In our project, we employed Matplotlib to analyze the distribution of features within the data. Figure 25 provides insights into the representation of patients with benign tumors (represented by B) and patients with malignant tumors (represented by M).

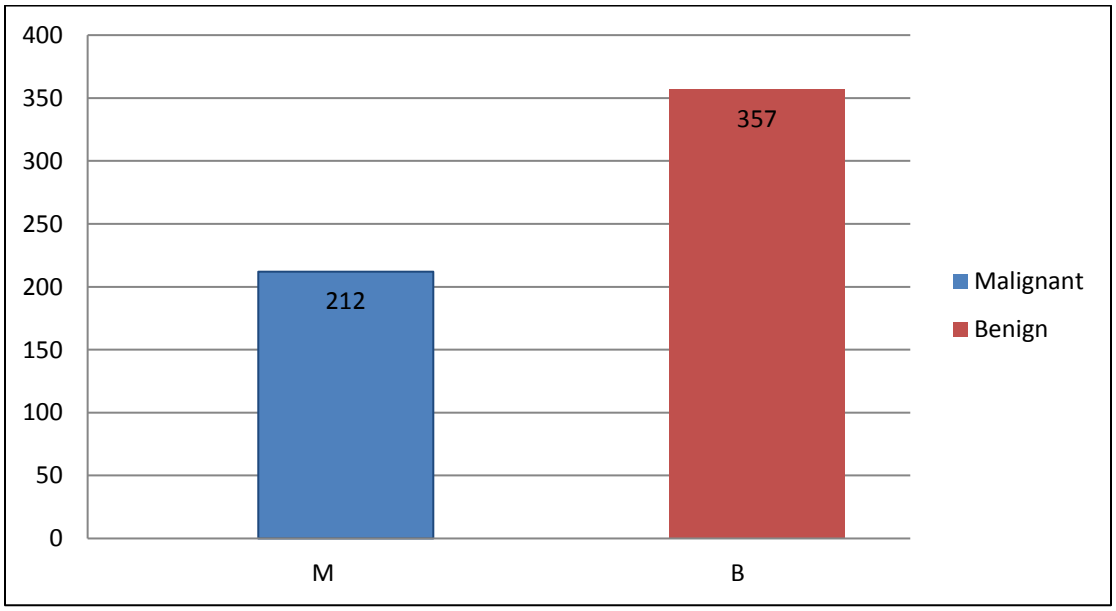


Figure 25. The WDBC dataset's target distribution.

3.2.2 Features' Correlation

We discovered the correlation between several features and created a chart using the Seaborn library's heatmap function. From Figure 26, we can observe that Mean radius, Mean area and Mean perimeter are one of the most highly correlated points to our worst perimeter feature. These types of correlations can also be seen with the target variable in the figure bellow.

Figure 26. Features Correlation.

3.3 Pre-processing

In breast cancer detection, relevant features are carefully selected while removing irrelevant or incorrect information. Numeric columns are standardized to a common scale, and noise is reduced by selecting only the relevant data. These processes enhance the performance of our machine learning algorithms. Overall, these steps contribute to accurate and effective breast cancer classification by preparing the dataset through feature analysis, data cleaning, missing value imputation, data normalization, and feature selection.

3.4 Evaluation Matrices

3.4.1 Confusion Matrix

The Confusion Matrix is a widely used and intuitive metric for assessing the accuracy and correctness of a model. It is particularly suitable for classification problems involving binary classes, making it highly relevant to this thesis. The matrix layout provides a visual representation of the algorithm's performance. In Table 3, each row corresponds to instances belonging to an actual class, while each column represents instances assigned to a predicted class, or vice versa . This layout allows for a comprehensive evaluation of the model's predictive capabilities and aids in understanding its performance in classifying different types of instances.

	Predictive Negative	**Predictive Positive**
Actual Negative	True Negative **(TN)**	False Positive **(FP)**
Actual Positive	False Negative **(FN)**	True Positive **(TP)**

Table 3. A Confusion Matrix in binary classification tasks.

Terms associated with Confusion matrix:

TPs: are cases where the actual class and predicted class are both true (1).

TNs: are cases where the actual class and predicted class are both false (0).

FPs: are cases where the actual class is false (0), but the predicted class is true (1).

FNs: are cases where the actual class is true (1), but the predicted class is false (0).

3.4.2 Classification report

The performance evaluation metric of a classification-based machine-learning model includes precision, recall, F1 score, and support. This metric provides insights into the overall performance of the trained model, offering a better understanding of its effectiveness.

3.4.2.1 Accuracy

It measures the overall correctness of a model's predictions by calculating the ratio of correctly classified instances to the total number of instances.

$$\text{Accuracy} = \frac{(TP_i+TN_i)}{(TP_i+FP_i+FN_i+TN_i)}$$

3.4.2.2 Precision

It measures the proportion of true positive predictions out of all predicted positive instances, providing insight into the accuracy of positive predictions.

$$\text{Precision} = \frac{TP_i}{(TP_i+FP_i)}$$

3.4.2.3 Recall

It also known as sensitivity, assesses the model's ability to correctly identify positive instances by calculating the ratio of correctly classified positive instances to the total number of actual positive cases.

$$\text{Recall} = \frac{TP_i}{(TP_i+FN_i)}$$

3.4.2.4 F1 score

It combines precision and recall into a single metric, providing a balanced evaluation of a model's accuracy in binary classification tasks.

$$F_1 - \text{score} = 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$

3.4.2.5 Specificity

It evaluates a model's ability to correctly identify true negative instances by calculating the proportion of actual negative instances that are classified as negative by the model.

$$\text{Specificity} = \frac{TN_i}{(TN_i + FP_i)}$$

3.4.2.6 Support

Support is the number of actual occurrences of the class in the dataset. It does not vary between models; it just diagnoses the performance evaluation process.

3.5 Experiment Environment

3.5.1 Software and libraries

We used Kaggle as an online platform due to its simplicity in using data and machine learning projects that facilitates collaboration, dataset sharing, model development, and participation in data science competitions. In a addition to Python which is a popular programming language used extensively in data analysis and machine learning. It offers a variety of libraries and frameworks that simplify tasks such as data manipulation, visualization, and modeling. As well as other libraries used in our data analysis project, including :

Seaborn: for statistical data visualization.

NumPy: for numerical data manipulation.

Matplotlib: for 2D data visualization.

Pandas: for data manipulation in tabular structures.

Scikit-learn (Sklearn): is a comprehensive library for machine learning, providing a wide range of algorithms and preprocessing tools.

Overall, Kaggle and Python, along with their associated libraries, provide a robust ecosystem for data scientists to explore, analyze, and model datasets, making it a popular choice in the field.

3.5.2 Training and testing the Model

Using the train-test-split method of the SciKit-Learn library, we are assessing the performance of a model on different testing scenarios. By varying in testing data portions (10%, 20%, and 30%) and the random states (0, 29, and 42). The dataset is fed through the Decision Tree, Random Forest classification, Support Vector Machine, XGboost, and Logistic Regression machine learning algorithms. These algorithms are tested individually on the dataset. These models were taken from various Sklearn modules. The fit() function is used to train each model on the training dataset, where it takes the training data as argument. Five evaluation metrics are used to determine the best performing model which are : accuracy, sensitivity, specificity, precision, F1-score, and specificity .

3.6 Models Performances

3.6.1 Logistic Regression

The model demonstrates good performance across various testing scenarios. The accuracy scores range from 0.9561 to 0.9912, indicating the overall correctness of the model's predictions. Sensitivity, or true positive rate, ranges from 0.9459 to 1.0, indicating the model's ability to correctly identify positive cases. Specificity, or true negative rate, ranges from 0.9166 to 1.0, indicating the model's ability to correctly identify negative cases. Precision, or the positive predictive value, ranges from 0.9428 to 1.0, indicating the proportion of correctly predicted positive cases out of the total predicted positives. The F1-scores, which consider both precision and sensitivity, range from 0.9629 to 0.9933, with higher values indicating better overall performance.

In summary, the model generally performed well across different data portions and random states.

Testing Portion (%)	Random states	Accuracy (%)	Sensitivity (%)	Specifity (%)	Precision (%)	F1-score (%)
10	0	96,49	100	91,66	94,28	97,05
	29	96,49	94,59	100	100	97,22
	42	96,49	97,5	94,11	97,5	97,5

20	0	95,61	95,58	95,65	97,01	96,29
	29	**99,12**	**98,68**	**100**	**100**	**99,33**
	42	98,24	97,26	100	100	98,61
30	0	96,49	98,11	93,84	96,29	97,19
	29	98,83	98,11	100	100	99,04
	42	97,66	97,27	98,36	99,07	98,16

Table 4. The performance of Logistic Regression classifier.

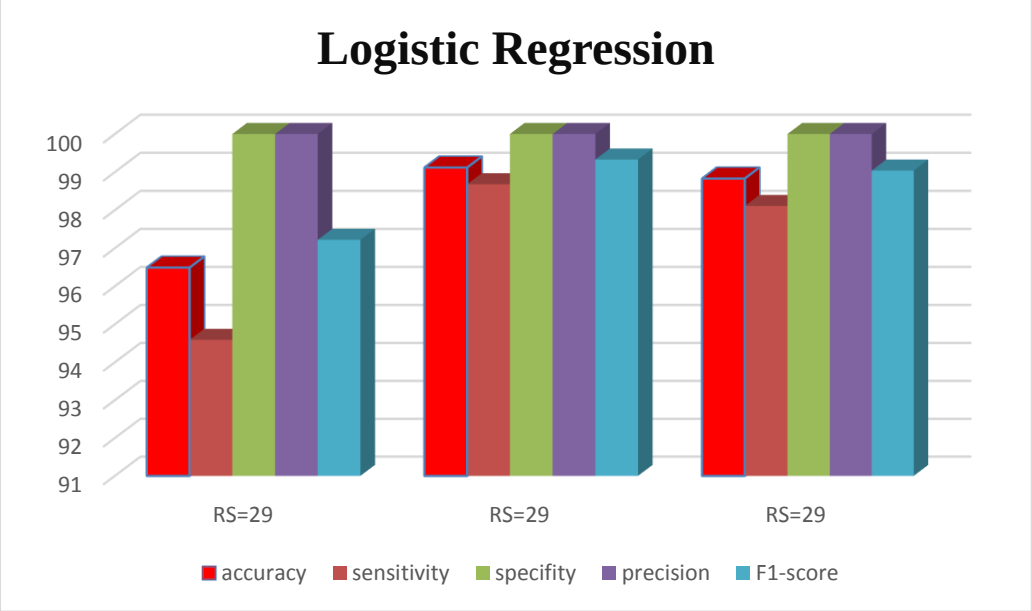


Figure 27. Logistic Regression best three performing models.

3.6.2 Decision tree

The model exhibits reasonably good performance across the different testing scenarios. The accuracy values range from 0.8771 to 0.9473, indicating the overall correctness of the model's predictions. Sensitivity (true positive rate) ranges from 0.9014 to 1.0, indicating the model's ability to correctly identify positive cases. Specificity (true negative rate) ranges from 0.8095 to 0.9523, indicating the model's ability to correctly identify negative cases. Precision, or the positive predictive value, ranges from 0.9 to 0.9714, indicating the proportion of correctly predicted positive cases out of the total predicted positives. The F1-

scores, which consider both precision and sensitivity, range from 0.9014 to 0.9577, with higher values indicating better overall performance.

In summary, the model's performance varies across different data portions and random states. The accuracy, sensitivity, specificity, precision, and F1-scores are generally good, indicating reasonable predictive capability. However, there are some variations in performance, suggesting that the model's performance may be affected by the data portion and the random state used for testing.

Testing Portion (%)	Random states	Accuracy (%)	Sensitivity (%)	Specifity (%)	Precision (%)	F1-score (%)
10	0	94,73	100	88	91,42	95,52
	29	**94,73**	**94,44**	**95,23**	**97,14**	**95,77**
	42	92,98	100	80,95	90	94,73
20	0	93,85	95,45	91,66	94,02	94,73
	29	93,85	95,94	90	94,66	95,3
	42	87,71	90,14	83,72	90,14	90,14
30	0	91,81	96,07	85,5	90,74	93,33
	29	93,56	96,03	90	93,26	94,63
	42	93,56	95,32	90,62	94,44	94,88

Table 5. The performance of Decision tree classifier.

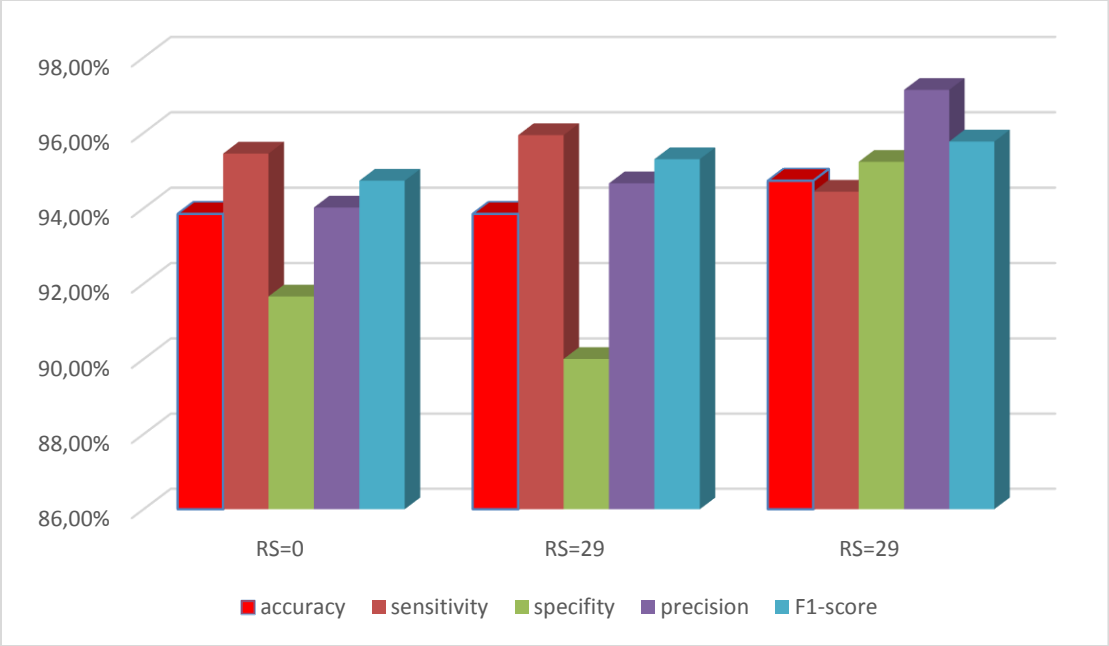


Figure 28. Decision tree confusion matrix.

3.6.3 Random Forest

The model's performance was evaluated on different data portions and random states to assess its robustness and generalizability. Across the various testing scenarios, the model consistently demonstrated good performance. The accuracy values ranged from 0.9298 to 0.9707, indicating a high level of correctness in its predictions. The sensitivity values ranged from 0.8974 to 1.0, suggesting that the model was able to accurately identify positive cases. Similarly, the specificity values ranged from 0.8384 to 1.0, indicating the model's ability to correctly identify negative cases. The precision values ranged from 0.8857 to 1.0, representing the proportion of correctly predicted positive cases out of all predicted positives. The F1-scores, which consider both precision and sensitivity, ranged from 0.9393 to 0.9907, with higher values indicating better overall performance. These findings collectively highlight the model's effectiveness in making accurate predictions across different testing scenarios.

Testing Portion (%)	Random states	Accuracy (%)	Sensitivity (%)	Specifity (%)	Precision (%)	F1-score (%)

10	0	92,98	100	84,61	88,57	93,93
	29	92,98	89,74	100	100	94,59
	42	96,49	100	89,74	95	97,43
20	0	95,61	95,58	95,65	97,01	96,29
	29	94,73	94,8	94,59	97,33	96,05
	42	96,49	95,89	97,56	98,59	97,22
30	0	96,49	98,11	93,84	96,29	97,19
	29	95,9	95,32	96,87	98,07	96,68
	42	**97,07**	**96,39**	**98,33**	**99,07**	**97,71**

Table 6. The performance of Random Forest classifier.

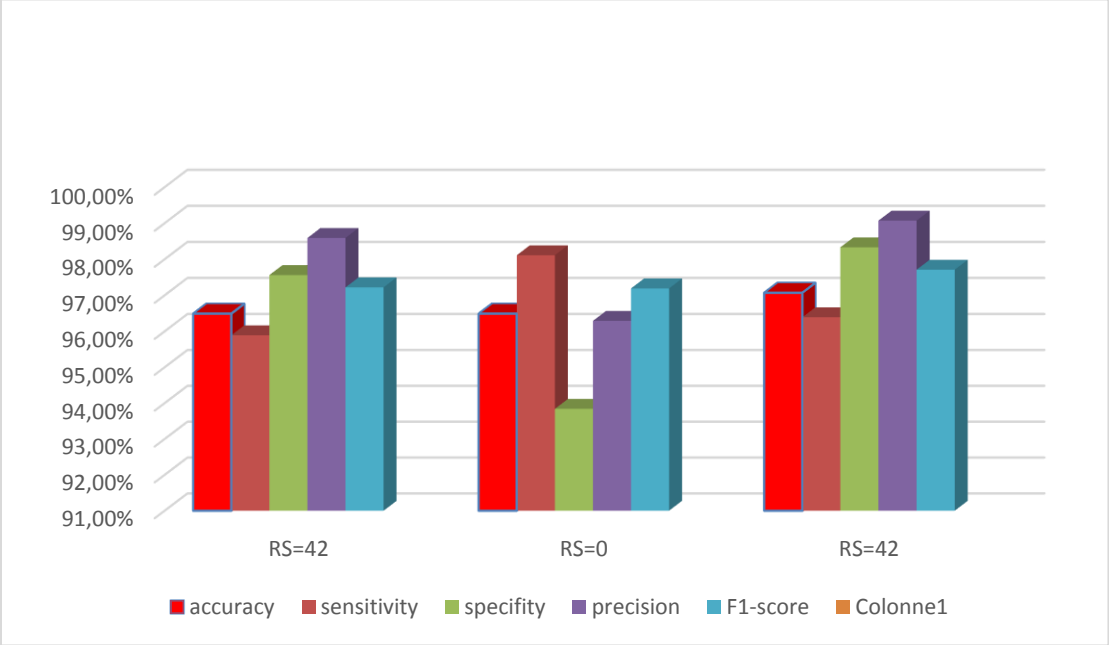


Figure 29. Random Forest best three performing models.

3.6.4 Support Vector Machine (SVM)

Across different testing scenarios, the model consistently demonstrates strong performance. The accuracy values range from 0.9473 to 0.9824, indicating a high level of correctness in its predictions. Sensitivity values range from 0.9459 to 1.0, indicating the model's ability to accurately identify positive cases. Specificity values range from 0.8888 to 1.0, indicating the model's ability to correctly identify negative cases. Precision values

range from 0.9351 to 1.0, representing the proportion of correctly predicted positive cases out of the total predicted positives. The F1-scores, ranging from 0.9573 to 0.9868, take into account both precision and sensitivity, with higher values indicating better overall performance. These findings collectively highlight the model's strong predictive capabilities across various testing scenarios.

Testing Portion (%)	Random states	Accuracy (%)	Sensitivity (%)	Specifity (%)	Precision (%)	F1-score (%)
10	0	98,24	100	95,65	97,14	98,55
	29	96,49	94,59	100	100	97,22
	42	94,73	97,43	88,88	95	96,2
20	0	97,36	97,05	97,82	98,5	97,77
	29	**98,24**	**97,4**	**100**	**100**	**98,68**
	42	97,36	95,94	100	100	97,93
30	0	94,73	98,05	89,7	93,51	95,73
	29	98,24	97,19	100	100	98,57
	42	97,66	97,27	98,36	99,07	98,16

Table 7. The performance of SVM.

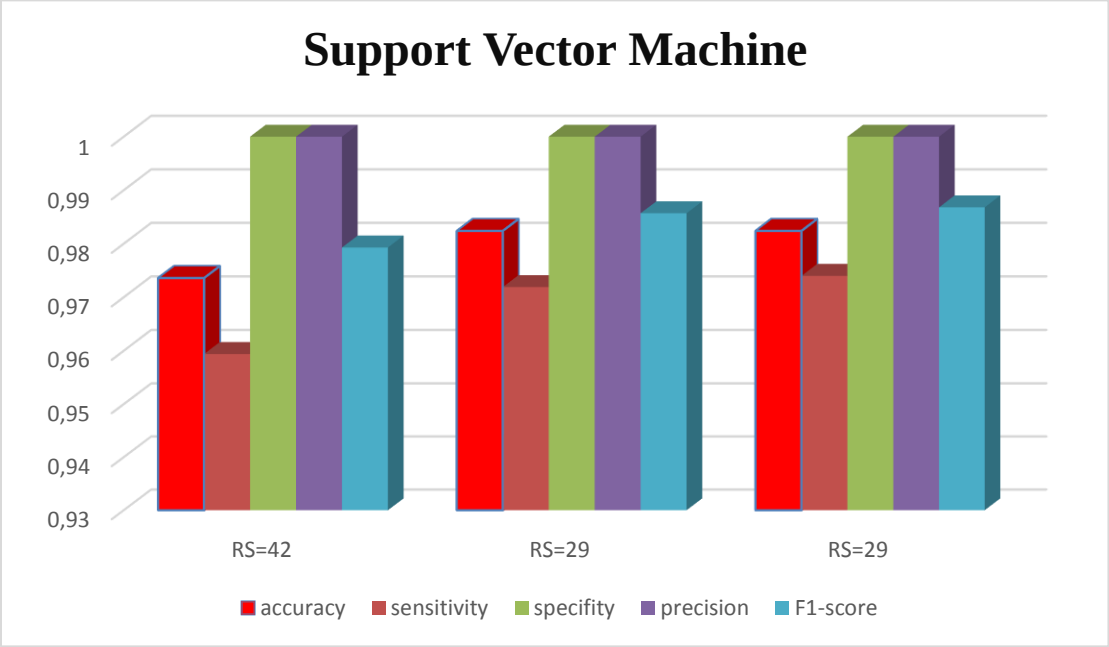


Figure 30.SVM best three performing models.

3.6.5 XGboost

The model consistently demonstrates good performance across different testing scenarios. The accuracy values range from 0.9298 to 0.9824, indicating the overall correctness of its predictions. Sensitivity values range from 0.8974 to 1.0, indicating the model's ability to correctly identify positive cases. Specificity values range from 0.88 to 1.0, indicating the model's ability to correctly identify negative cases. Precision values range from 0.9142 to 1.0, representing the proportion of correctly predicted positive cases out of the total predicted positives. The F1-scores, ranging from 0.9459 to 0.9873, take into account both precision and sensitivity, with higher values indicating better overall performance. These findings highlight the model's consistent and strong predictive capabilities across various testing scenarios.

Testing Portion (%)	Random states	Accuracy (%)	Sensitivity (%)	Specify (%)	Precision (%)	F1-score (%)
10	0	94,73	100	88	91,42	95,52
	29	92,98	89,74	100	100	94,59
	42	98,24	100	94,44	97,5	98,73
20	0	96,49	94,36	100	100	97,1
	29	98,24	98,66	97,43	98,66	98,66
	42	96,49	94,66	100	100	97,26
30	0	96,49	97,22	95,23	97,22	97,22
	29	98,24	98,09	98,48	99,03	98,56
	42	**98,24**	**97,29**	**100**	**100**	**98,63**

Table 8. The performance of XGboost.

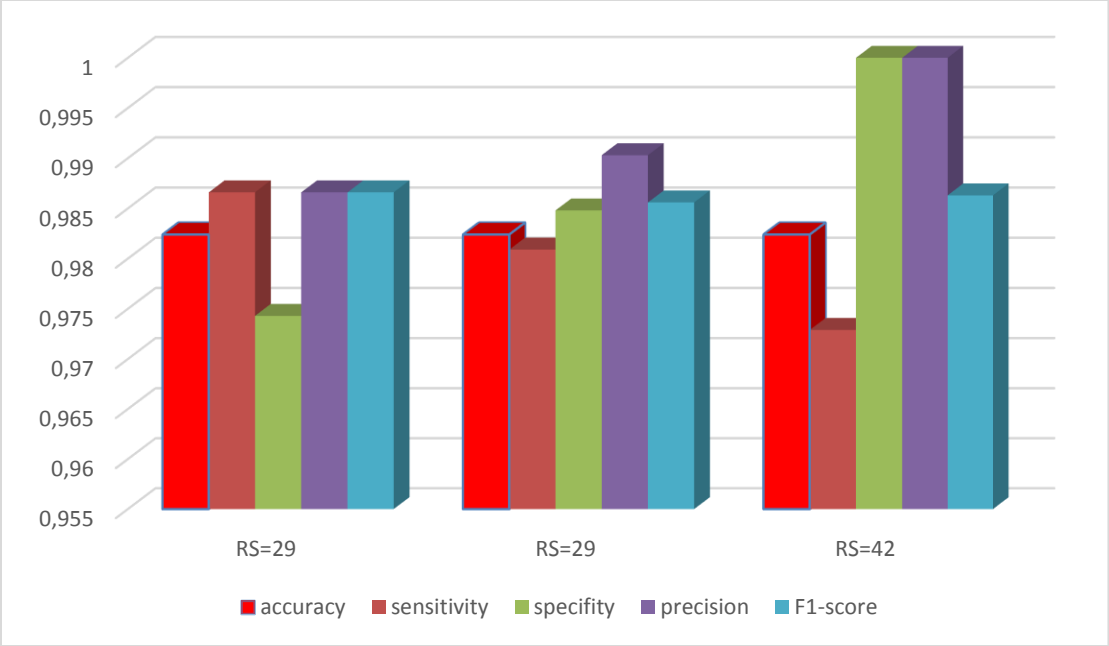


Figure 31. Xgboost best three performing models.

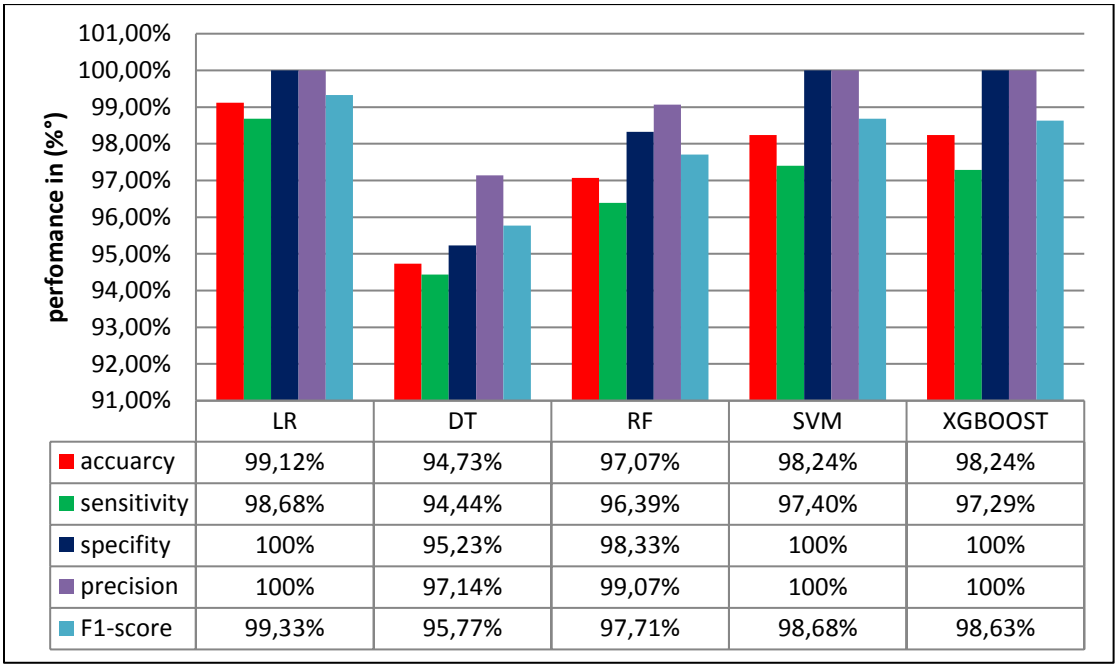
3.7 Performance comparison

After completing the implementation of all five algorithms for detecting breast cancer from the dataset, after choosing the best performing models of each model , the results can be compared from the figure 31 using the performance metrics discussed previously.

All five models demonstrate good performance across different data portions and random states, exhibiting high accuracy, sensitivity, specificity, precision, and F1-scores. Logistic Regression performs well across all metrics, consistently achieving high accuracy, sensitivity, specificity, precision, and F1-scores. It consistently outperforms other models in terms of specificity, indicating its ability to correctly identify negative cases. Additionally, Logistic Regression achieves high sensitivity, precision, and F1-scores, indicating its ability to accurately identify positive cases. Therefore, Logistic Regression can be considered the best-performing model among the five tested.

It's important to note that the choice of the best-performing model may depend on the specific requirements and objectives of the task at hand. Other factors such as

computational efficiency, interpretability, and generalizability should also be considered in selecting the most suitable model.



	LR	DT	RF	SVM	XGBOOST
accuarcy	99,12%	94,73%	97,07%	98,24%	98,24%
sensitivity	98,68%	94,44%	96,39%	97,40%	97,29%
specifity	100%	95,23%	98,33%	100%	100%
precision	100%	97,14%	99,07%	100%	100%
F1-score	99,33%	95,77%	97,71%	98,68%	98,63%

Figure 32. Evaluation matrices of the five top performing models.

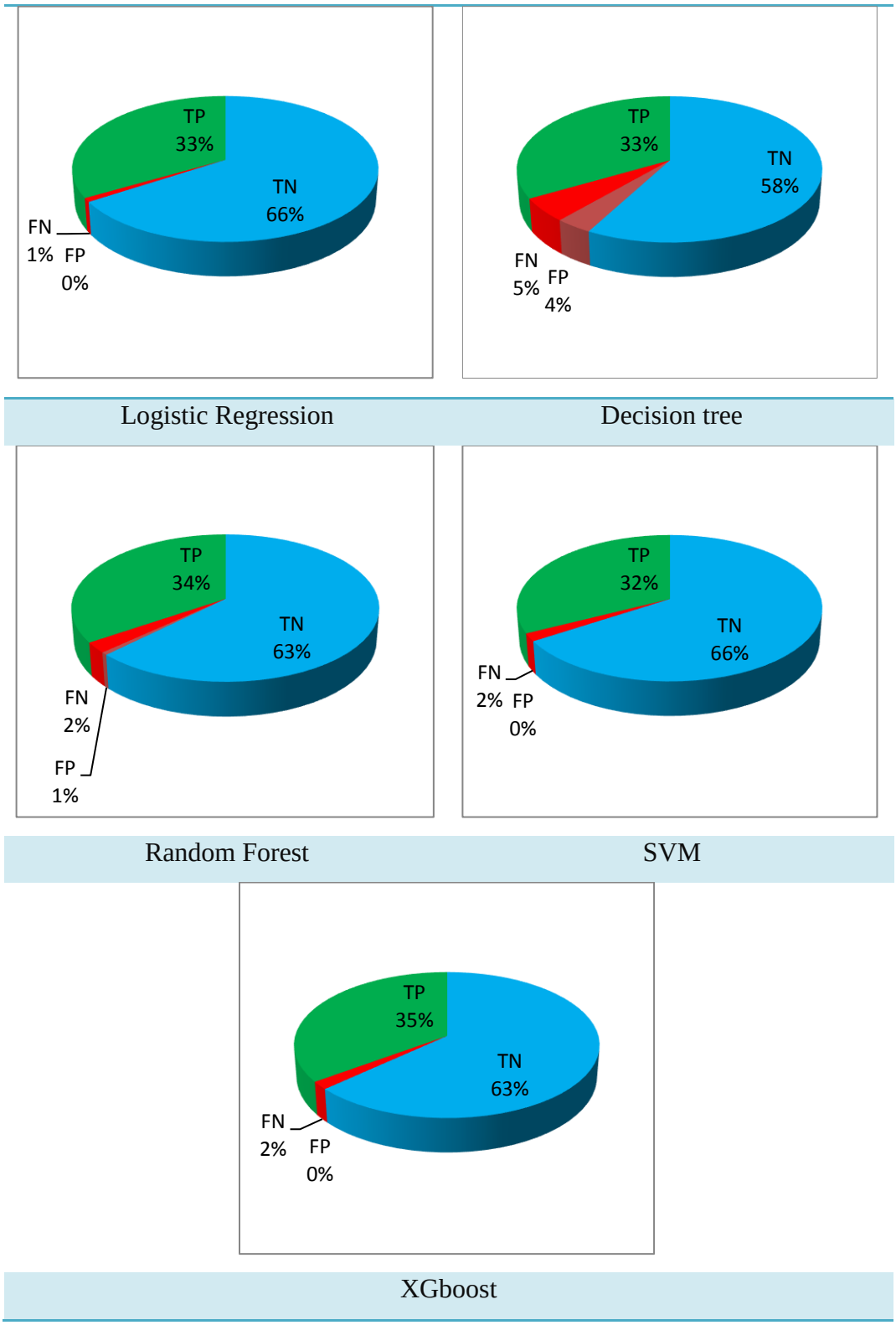


Figure 33. Confusion matrices of all five performing models.

3.8 Comparison with the state of art

To provide an overview of the performance of our breast cancer prediction system, we conducted a comparison with existing works from 2019 to 2023 that utilized similar performance measures. In Table 9 , we present a comprehensive comparison based on the five evaluation metrics . The results clearly demonstrate that the obtained results surpasses all the state-of-the-art techniques, indicating its superior performance.

Author	Datasets	Techniques (%)	Accuracy (%)	Sensitivity (%)	Specifity (%)	Precision (%)	F1-score (%)
1 Ajay Kumar , 2019 [71]	WDBC	DT RF SVM	92.97 95.95 97.89	93 96 97.9	/	93 96 97.9	93 95.9 97.9
2 Yopie Noor Hantoro, 2020 [72]	WDBC	RF SVM.	95.61 95.26	94.94 95.52	/	95.08 94.85	/
3 Harsha Gangadhara Vinay Kumar Donga, 2021 [73]	WDBC	DT RF SVM LR	96.1 96.6 97.8 96.8	97 96 99 92	/	98 97 98 98	90 96 98 95
4 Vattsal Singhal1 , 2022 [74]	WDBC	DT RF SVM LR.	95.1 96.5 97.2 95.8	92 94 94 91	/	94 96 98 98	93 95 96 94
5 Our work , 2023	WDBC	LR DT RF SVM Xgboost	**99.12** **94.73** **97.07** **98.24** **98.24**	**98.63** **94.44** **96.39** **97.40** **97.24**	**100** **95.23** **98.33** **100** **100**	**100** **97.14** **99.07** **100** **100**	**99.33** **95.77** **97.71** **98.68** **98.63**

Table 9. A comparison with the stat of the art.

Conclusion

In this chapter, we presented the models, libraries, and configurations used for our implementation, as well as the datasets (training,testing) we utilized and their respective evaluations. Then choosing the best performing model of the five and finally compare the models results with state of the art.

An accuracy score of over 99% is an optimistic result for cancer classification; however, the approaches we proposed need further improvement for generalization at a larger scale.

General conclusion

General conclusion

Breast cancer, which accounts for 69% of cancer-related deaths highlights the enormity of the issue among woman. Early detection of breast cancer plays a crucial role in improving survival rates by enabling individuals to receive on time clinical interventions And making use of machine learning and its algorithims.

This study focused on the classification of breast cancer using five machine learning techniques: Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), XGBoost, and Decision Tree (DT).

The first chapter provided an overview of breast cancer, while the second chapter discussed the fundamentals of AI and machine learning techniques. The third chapter presented the results obtained from applying the five models to the breast cancer classification task, with variations in the random state.

The results revealed that all five machine learning techniques were effective in classifying breast cancer. Among the models evaluated, the LR model demonstrated the highest accuracy and precision, making it the most reliable model for breast cancer classification. The LR model achieved an accuracy of 99.12%, sensitivity of 98.68%, specificity of 100%, precision of 100%, and an F1-score of 99.33%.

The findings of this study contribute to the field of medical diagnostics by showcasing the potential of machine learning techniques in accurately identifying breast cancer. These models can aid in early detection and treatment, leading to improved patient outcomes. Future improvements could include incorporating additional features for more precise prediction and recommendation of treatments based on the severity of the patient's condition.

As we look ahead, the future holds promise for refining the models through greater access to data, particularly with the exponential growth of information. Machine learning models exhibit enhanced performance in correlation with the volume of data, suggesting ample room for improvement with the availability of extensive datasets.

As written in the holy book of Islam, the Quran:

“Whoever saves one life,

it is written as if he has saved all humanity.”

References

[1] H. Sung et al., ‘Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries’, CA: A Cancer Journal for Clinicians, vol. 71, no. 3, pp. 209–249, 2021, doi: 10.3322/caac.21660.

[2] ‘How Is Breast Cancer Diagnosed?’, Centers for Disease Control and Prevention, Mar. 09, 2022. https://www.cdc.gov/cancer/breast/basic_info/diagnosis.htm.

[3] M. Javaid, A. Haleem, R. Pratap Singh, R. Suman, and S. Rab, ‘Significance of machine learning in healthcare: Features, pillars and applications’, International Journal of Intelligent Networks, vol. 3, pp. 58–73, Jan. 2022, doi: 10.1016/j.ijin.2022.05.002.

[4] ‘Alkabban FM, Ferguson T. Breast Cancer. [Updated 2022 Sep 26]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK482286/).’

[5] F. M. Alkabban and T. Ferguson, ‘Breast Cancer’, in StatPearls, Treasure Island (FL): StatPearls Publishing, 2023. Accessed: Jun. 07, 2023. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK482286/

[6] ‘What Is Cancer? - NCI’. https://www.cancer.gov/about-cancer/understanding/what-is-cancer.

[7] ‘What is cancer? | Cancer Research UK’. https://www.cancerresearchuk.org/about-cancer/what-is-cancer.

[8] A. Patel, ‘Benign vs Malignant Tumors’, JAMA Oncol, vol. 6, no. 9, p. 1488, Sep. 2020, doi: 10.1001/jamaoncol.2020.2592.

[9] A. McGuire, J. A. L. Brown, C. Malone, R. McLaughlin, and M. J. Kerin, ‘Effects of Age on the Detection and Management of Breast Cancer’, Cancers (Basel), vol. 7, no. 2, pp. 908–929, May 2015, doi: 10.3390/cancers7020815.

[10] nhswebsite, ‘Breast cancer in women’, nhs.uk, Oct. 20, 2017. https://www.nhs.uk/conditions/breast-cancer/.

[11] ‘Breast Cancer Overview: Causes, Symptoms, Signs, Stages & Types’. https://my.clevelandclinic.org/health/diseases/3986-breast-cancer.

[12] nhswebsite, ‘Breast cancer in women - Causes’, nhs.uk, Oct. 24, 2017. https://www.nhs.uk/conditions/breast-cancer/causes/.

[13] E. J. Schneble et al., ‘Future Directions for the Early Detection of Recurrent Breast Cancer’, J Cancer, vol. 5, no. 4, pp. 291–300, Mar. 2014, doi: 10.7150/jca.8017.

[14] R. J. Hooley, L. M. Scoutt, and L. E. Philpotts, ‘Breast ultrasonography: state of the art’, Radiology, vol. 268, no. 3, pp. 642–659, Sep. 2013, doi: 10.1148/radiol.13121606.

[15] ‘Mammography’, National Institute of Biomedical Imaging and Bioengineering. https://www.nibib.nih.gov/science-education/science-topics/mammography.

[16] ‘Sensors | Free Full-Text | Early Diagnosis of Breast Cancer’. https://www.mdpi.com/1424-8220/17/7/1572.

[17] ‘Breast Biopsy: Procedure Types, What to Expect & Results Guide’. https://www.nationalbreastcancer.org/breast-cancer-biopsy/.

[18] ‘Fine Needle Aspiration (FNA) of the Breast’. https://www.cancer.org/cancer/types/breast-cancer/screening-tests-and-early-detection/breast-biopsy/fine-needle-aspiration-biopsy-of-the-breast.html).

[19] ‘Cancer Diagnosis - Hematology and Oncology’, MSD Manual Professional Edition. https://www.msdmanuals.com/professional/hematology-and-oncology/overview-of-cancer/cancer-diagnosis?query=Histopathologic%20Type%20breast.

[20] C. Sotiriou and L. Pusztai, ‘Gene-expression signatures in breast cancer’, N Engl J Med, vol. 360, no. 8, pp. 790–800, Feb. 2009, doi: 10.1056/NEJMra0801289.

[21] J.-Q. Chen and J. Russo, ‘ERα-Negative and Triple Negative Breast Cancer: Molecular Features and Potential Therapeutic Approaches’, Biochim Biophys Acta, vol. 1796, no. 2, pp. 162–175, Dec. 2009, doi: 10.1016/j.bbcan.2009.06.003.

[22] ‘Breast Fibroadenoma - an overview | ScienceDirect Topics’. https://www.sciencedirect.com/topics/pharmacology-toxicology-and-pharmaceutical-science/breast-fibroadenoma.

[23] L. C. Hartmann et al., ‘Benign breast disease and the risk of breast cancer’, N Engl J Med, vol. 353, no. 3, pp. 229–237, Jul. 2005, doi: 10.1056/NEJMoa044383.

[24] ‘Benign Breast Disease in Women - Endotext - NCBI Bookshelf’. https://www.ncbi.nlm.nih.gov/books/NBK278994/.

[25] G. M. K. Tse, Y. Niu, and H.-J. Shi, ‘Phyllodes tumor of the breast: an update’, Breast Cancer, vol. 17, no. 1, pp. 29–34, 2010, doi: 10.1007/s12282-009-0114-z.

[26] B. A. Virnig, T. M. Tuttle, T. Shamliyan, and R. L. Kane, ‘Ductal carcinoma in situ of the breast: a systematic review of incidence, treatment, and outcomes’, J Natl Cancer Inst, vol. 102, no. 3, pp. 170–178, Feb. 2010, doi: 10.1093/jnci/djp482.

[27] D. L. Page, W. D. Dupont, L. W. Rogers, and M. S. Rados, ‘Atypical hyperplastic lesions of the female breast. A long-term follow-up study’, Cancer, vol. 55, no. 11, pp. 2698–2708, Jun. 1985, doi: 10.1002/1097-0142(19850601)55:11<2698::aid-cncr2820551127>3.0.co;2-a.

[28] E. A. Rakha, J. S. Reis-Filho, and I. O. Ellis, ‘Basal-like breast cancer: a critical review’, J Clin Oncol, vol. 26, no. 15, pp. 2568–2581, May 2008, doi: 10.1200/JCO.2007.13.1748.

[29] C. Li, B. Anderson, P. PP, S. Holt, J. Daling, and R. Moe, ‘Changing incidence rate of invasive lobular carcinoma among older women’, Cancer, vol. 88, pp. 2561–9, Jun. 2000, doi: 10.1002/1097-0142(20000601)88:113.3.CO;2-O.

[30] Z. Elsawaf and H.-P. Sinn, ‘Triple-Negative Breast Cancer: Clinical and Histological Correlations’, Breast Care (Basel), vol. 6, no. 4, pp. 273–278, Aug. 2011, doi: 10.1159/000331643.

[31] ‘Triple-negative breast cancer: clinical features and patterns of recurrence - PubMed’. https://pubmed.ncbi.nlm.nih.gov/17671126/ (accessed Jun. 07, 2023).

[32] D. J. Slamon, G. M. Clark, S. G. Wong, W. J. Levin, A. Ullrich, and W. L. McGuire, ‘Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene’, Science, vol. 235, no. 4785, pp. 177–182, Jan. 1987, doi: 10.1126/science.3798106.

[33] ‘Cancer today’. http://gco.iarc.fr/today/home .

[34] ‘Breast cancer’. https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

[35] ‘Global cancer data by country | World Cancer Research Fund International’, WCRF International. https://www.wcrf.org/cancer-trends/global-cancer-data-by-country/.

[36] ‘Electronic Health Records (EHRS) | NIOSH | CDC’, May 06, 2022. https://www.cdc.gov/niosh/topics/ehr/default.html .

[37] J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon, ‘A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955’, AI Magazine, vol. 27, no. 4, Art. no. 4, Dec. 2006, doi: 10.1609/aimag.v27i4.1904.

[38] ‘Human intelligence and brain networks - PubMed’. https://pubmed.ncbi.nlm.nih.gov/21319494/.

[39] D. Marr, ‘Artificial intelligence—A personal view’, Artificial Intelligence, vol. 9, no. 1, pp. 37–48, Aug. 1977, doi: 10.1016/0004-3702(77)90013-3.

[40] M. Haenlein and A. Kaplan, ‘A brief history of artificial intelligence: On the past, present, and future of artificial intelligence’, California management review, vol. 61, no. 4, pp. 5–14, 2019.

[41] B. Delipetrev, C. Tsinaraki, and U. Kostic, ‘Historical evolution of artificial intelligence’, 2020.

[42] T. Bench-Capon et al., ‘A history of AI and Law in 50 papers: 25 years of the international conference on AI and Law’, Artificial Intelligence and Law, vol. 20, pp. 215–319, Sep. 2012, doi: 10.1007/s10506-012-9131-x.

[43] K. Chowdhary and K. R. Chowdhary, ‘Natural language processing’, Fundamentals of artificial intelligence, pp. 603–649, 2020.

[44] ‘What is Computer Vision? | IBM’. https://www.ibm.com/topics/computer-vision.

[45] M. Soori, B. Arezoo, and R. Dastres, ‘Artificial intelligence, machine learning and deep learning in advanced robotics, a review’, Cognitive Robotics, vol. 3, pp. 54–70, Jan. 2023, doi: 10.1016/j.cogr.2023.04.001.

[46] A. Bohr and K. Memarzadeh, ‘The rise of artificial intelligence in healthcare applications’, Artificial Intelligence in Healthcare, pp. 25–60, 2020, doi: 10.1016/B978-0-12-818438-7.00002-2.

[47] P. B et al., Deep Learning for Intelligent Demand Response and Smart Grids: A Comprehensive Survey. 2021.

[48] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics). 2009.

[49] I. Goodfellow, Y. Bengio, and A. Courville, Deep learning. MIT press, 2016.

[50] E. G. Learned-Miller, ‘Introduction to Supervised Learning’.

[51] T. G. Dietterich, ‘Machine-Learning Research’, AI Magazine, vol. 18, no. 4, Art. no. 4, Dec. 1997, doi: 10.1609/aimag.v18i4.1324.

[52] T. O. Ayodele, ‘Types of Machine Learning Algorithms’, in New Advances in Machine Learning, IntechOpen, 2010. doi: 10.5772/9385.

[53] S. Laine and T. Aila, ‘Temporal Ensembling for Semi-Supervised Learning’. arXiv, Mar. 15, 2017. doi: 10.48550/arXiv.1610.02242.

[54] L. P. Kaelbling, M. L. Littman, and A. W. Moore, ‘Reinforcement Learning: A Survey’. arXiv, Apr. 30, 1996. doi: 10.48550/arXiv.cs/9605103.

[55] ‘Recommendation systems: Principles, methods and evaluation - ScienceDirect’. https://www.sciencedirect.com/science/article/pii/S1110866515000341.

[56] L. Breiman, ‘Random Forests’, Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[57] M. Khalilia, S. Chakraborty, and M. Popescu, ‘Predicting disease risks from highly imbalanced data using random forest’, BMC Medical Informatics and Decision Making, vol. 11, no. 1, p. 51, Jul. 2011, doi: 10.1186/1472-6947-11-51.

[58] S. Xu, Z. Zhang, D. Wang, J. Hu, X. Duan, and T. Zhu, ‘Cardiovascular risk prediction method based on CFS subset evaluation and random forest classification framework’, in 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA), Mar. 2017, pp. 228–232. doi: 10.1109/ICBDA.2017.8078813.

[59] C. Nguyen, Y. Wang, and H. N. Nguyen, ‘Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic’, JBiSE, vol. 06, no. 05, pp. 551–560, 2013, doi: 10.4236/jbise.2013.65070.

[60] J. Park, J. Kwon, J. Oh, S. Lee, J.-Y. Kim, and H.-J. Yoo, ‘A 92-mW Real-Time Traffic Sign Recognition System With Robust Illumination Adaptation and Support Vector Machine’, IEEE Journal of Solid-State Circuits, vol. 47, no. 11, pp. 2711–2723, Nov. 2012, doi: 10.1109/JSSC.2012.2211691.

[61] S. Alajmani and H. Elazhary, ‘Hospital Readmission Prediction using Machine Learning Techniques’, International Journal of Advanced Computer Science and Applications (IJACSA), vol. 10, no. 4, Art. no. 4, 30 2019, doi: 10.14569/IJACSA.2019.0100425.

[62] B. Zheng, J. Zhang, S. W. Yoon, S. S. Lam, M. Khasawneh, and S. Poranki, ‘Predictive modeling of hospital readmissions using metaheuristics and data mining’, Expert Systems with Applications, vol. 42, no. 20, pp. 7110–7120, Nov. 2015, doi: 10.1016/j.eswa.2015.04.066.

[63] A. Gunawan, P. Suardana, A. Sulaiman, A. Negara, A. Mahendra Putra, and A. Negara, ‘Classification of invasive lobular carcinoma (ILC) and invasive ductal carcinoma (IDC) using the support vector machine (SVM) method’, Applied Mathematical Sciences, vol. 16, pp. 261–271, Jan. 2022, doi: 10.12988/ams.2022.916783.

[64] S. K. Shevade and S. S. Keerthi, ‘A simple and efficient algorithm for gene selection using sparse logistic regression’, Bioinformatics, vol. 19, no. 17, pp. 2246–2253, Nov. 2003, doi: 10.1093/bioinformatics/btg308.

[65] M. K. Bardsiri and M. Eftekhari, ‘Comparing ensemble learning methods based on decision tree classifiers for protein fold recognition’, Int J Data Min Bioinform, vol. 9, no. 1, pp. 89–105, 2014, doi: 10.1504/ijdmb.2014.057776.

[66] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning. in Springer Series in Statistics. New York, NY: Springer, 2009. doi: 10.1007/978-0-387-84858-7.

[67] S. Sushmita et al., ‘Predicting 30-day risk and cost of" all-cause" hospital readmissions’, in Workshops at the thirtieth AAAI conference on artificial intelligence, 2016.

[68] K.-H. Chen et al., ‘Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm’, BMC Bioinformatics, vol. 15, no. 1, p. 49, Feb. 2014, doi: 10.1186/1471-2105-15-49.

[69] ‘Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost - ScienceDirect’. https://www.sciencedirect.com/science/article/abs/pii/S0167865520302129.

[70] T. Chen and C. Guestrin, ‘XGBoost: A Scalable Tree Boosting System’, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in KDD ’16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[71] A. Kumar, R. Sushil, and A. Tiwari, ‘Comparative Study of Classification Techniques for Breast Cancer Diagnosis’, International Journal of Computer Sciences and Engineering, vol. 7, pp. 234–240, Jan. 2019, doi: 10.26438/ijcse/v7i1.234240.

[72] Y. N. Hantoro, ‘Comparative Study of Breast Cancer Diagnosis using Data Mining Classification’, International journal of engineering research and technology, Jun. 2020, Accessed: Jun. 07, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Comparative-Study-of-Breast-Cancer-Diagnosis-using-Hantoro/fcaf6f4b0b306c2e22af2ab4253ed850a66f0831

[73] H. G. V. K. Donga, Comparing Machine Learning Models : For Diagnosis of Breast cancer. 2022.. [Online]. Available: https://urn.kb.se/resolve?urn=urn:nbn:se:bth-23391

[74] V. Singhal, Y. Chaudhary, S. Verma, U. Agarwal, and Mr. P. Sharma, ‘Breast Cancer Prediction using KNN, SVM, Logistic Regression and Decision Tree’, IJRASET, vol. 10, no. 5, pp. 1877–1881, May 2022, doi: 10.22214/ijraset.2022.42688.

Abstract

Breast cancer is a significant health concern, and early detection is crucial for effective treatment. Machine learning classification techniques have shown great efficiency in improving breast cancer diagnosis. In this research, we used five different algorothims : Random Forest (RF), Logistic Regression (LR), XGBoost, Support Vector Machine (SVM), and Decision Tree (DT) for breast cancer classification. The dataset used was the Wisconsin Diagnostic Breast Cancer dataset (WDBC).it is observed that the logistic regression outperform all other classifiers and achieves impressive scores across multiple performance metrics such as specifity of 100% , precision of 100% , sensitivity of 98.63% and Accuracy of 99.12% .As we conducted a thorough comparison with previous approaches, and our results demonstrated the superiority of our proposed model in breast cancer classification.

Keywords: Breast Cancer classification , Random Forest (RF), Logistic Regression (LR), XGBoost, Support Vector Machine (SVM), and Decision Tree (DT).

ملخص

سرطان الثدي هو قضية صحية هامة، والكشف المبكر أمر جوهري للعلاج الفعال. أظهرت تقنيات تصنيف التعلم الآلي كفاءة كبيرة في تحسين تشخيص سرطان الثدي . في هذا البحث، استخدمنا خمسة خوارزميات مختلفة: الغابات العشوائية (RF)، الانحدار اللوجستي(LR) ، تعزيز التدرج الشديد (XGBoost) ، آلة الدعم النوعي(SVM) ، وشجرة القرارات (DT) لتصنيف سرطان الثدي. تم استخدام مجموعة البيانات ويسكونسن (WDBC) لتشخيص سرطان الثدي . لوحظ أن الانحدار اللوجستي تفوق على جميع الطرق الأخرى كما حقق نتائج مثيرة للإعجاب في العديد من مقاييس التقييم مثل معدل التحديد بنسبة 100%، ودقة التنبؤ بنسبة 100%، وحساسية التحليل بنسبة 98.63%، ودقة التصنيف بنسبة 99.12%. أجرينا مقارنة شاملة مع النهج السابقة، أظهرت نتائجنا تفوق نموذجنا المقترح في تصنيف سرطان الثدي.

الكلمات المفتاحية: تصنيف سرطان الثدي، الغابات العشوائية (RF)، الانحدار اللوجستي (LR)، تعزيز التدرج الشديد (XGBoost) ، آلة الدعم النوعي (SVM)، وشجرة القرارات (DT).