

# **Revolutionizing Breast Cancer Management: Harnessing the power of Machine Learning in Diagnosis and Treatment**

A DISSERTATION  
SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE AWARD OF THE DEGREE  
OF

Master of Science

In

**Biotechnology**

Submitted by:

**Anukriti Yadav**

**2K21/MSCBIO/06**

Under the supervision of:

Prof. Yasha Hasija

Professor



**DEPARTMENT OF BIOTECHNOLOGY**  
**DELHI TECHNOLOGICAL UNIVERSITY**  
(Formerly Delhi College of Engineering)  
Bawana Road, Delhi - 110042

**DEPARTMENT OF BIOTECHNOLOGY**  
**DELHI TECHNOLOGICAL UNIVERSITY**  
(Formerly Delhi College of Engineering)Bawana Road, Delhi - 110042

## **CANDIDATE’S DECLARATION**

I Anukriti Yadav Roll Number: 2K21/MSCBIO/06, student of M.Sc. Biotechnology, hereby declare that the work which is presented in the Major Project entitled - “**Revolutionizing Breast Cancer Management: Harnessing the Power of Machine Learning in Diagnoses and Treatment**” in the fulfillment of the requirement for the award of the degree of Master of Science in Biotechnology and submitted to the Department of Biotechnology, Delhi Technological University, Delhi, is an authentic record of my own carried out during the period from January- May 2023, under the supervision of Prof. Yasha Hasija.

The matter presented in this report has not been submitted by me for the award for any other degree of this or any other Institute/University. The work has been accepted in SCI/SCI expanded /SSCI/Scopus Indexed Journal OR peer reviewed Scopus Index Conference with the following details:

**Title of the Paper:** Behavioral analysis using machine learning algorithms in healthcare sector

**Author Names:** Anukriti Yadav, Deepak Kumar and Yasha Hasija

**Name of Conference:** International Conference on Advancement in Computation and Computer Technologies (IEEE InCACCT-2023)

**Conference Date and Venue:** 6<sup>th</sup> May at Chandigarh University, Gharuan, Mohali (Punjab) - India

**Registration:** Done

**Status of Paper:** Accepted

**Date of Paper Communication:** 1<sup>st</sup> March 2023

**Date of Paper Acceptance:** 5March 2023

Date:

Anukriti Yadav

DELHI TECHNOLOGICAL UNIVERSITY  
(Formerly Delhi College of Engineering)  
Bhawana Road, Delhi-110042

**Certificate**

I hereby certify that the Project Dissertation titled “**Revolutionizing Breast Cancer Management: Harnessing the Power of Machine Learning in Diagnoses and Treatment**” which is submitted by **Anukriti Yadav (2K21/MSCBIO/06)** Department of Biotechnology, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Science is recorded for the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any degree or any diploma to this university or elsewhere.

Place: Delhi

Date:

**Prof. Yasha Hasija**

**(Supervisor)**

**Professor**

Department of Biotechnology

Delhi Technological University

**Prof. Pravir Kumar**

**Head of Department**

**Dean (International Affairs)**

Department of Biotechnology

Delhi Technological University

## Acknowledgement

I would like to express my gratitude towards my supervisor, Prof. Yasha Hasija, for giving me the opportunity to do research and providing invaluable guidance throughout this research. Her dynamism, vision, sincerity and motivation have deeply inspired me. She has motivated to carry out the research and to present my work works as clearly as possible. It was a great privilege and honor to work and study under her guidance. I am extremely grateful for what he has offered me. Her insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

I am extremely grateful to my parents for their love, prayers, caring and sacrifices for educating and preparing me for my future. I would also like to thank to Jaishree Mam for guiding me and helping me to complete mt thesis.

A special thanks to my lab mate Deepak Kumar for moral support, tolerance and help from the beginning to the end.

I would also like the institution Delhi Technological University, Delhi for giving me the opportunities throughout the tenure of study.

Finally, my thanks go to all the people who have supported me to complete the research work directly or indirectly.

Anukriti Yadav

## **Abstract**

Breast cancer is a worldwide health issue that demands precise and efficient management. By recognizing patterns and links in massive datasets and establishing personalized treatment plans, machine learning has shown promise for advancing breast cancer care. Early diagnosis of tumor can dramatically improve patients' prognoses and chances of survival by promoting timely therapeutic therapy. More accurate categorization of benign tumours may protect patients from needless therapies. As a result, substantial study is being conducted into the precise analysis of Breast Cancer and the organization of those diagnosed into malignant or benign types. However, there are several issues with machine learning, such as accessibility and quality, ethical concerns, and accountability. The purpose of this thesis is to present a complete assessment of the present state of advances in machine learning in breast cancer management as well as identify potential and obstacles connected with their incorporation into clinical practice. This thesis will add to the current attempts to revolutionize breast cancer treatment through the potential of machine learning by studying the most recent research and breakthroughs in this sector.

# Contents

Candidate's Declaration	i
Certificate	ii
Acknowledgement	iii
Abstract	iv
Contents	v
List of Figures	vii
List of Tables	viii
List of Abbreviations	ix
CHAPTER 1 INTRODUCTION	1
1.2 Purpose	3
1.3 Research Question	3
1.4 Motivation and Problem statement	3
CHAPTER 2 Background	4
2.2 Breast Cancer	5
2.3 Symptoms	6
2.4 Stages of breast cancer	7
2.4.1 Stage 0	8
2.4.2 Stage I	8
2.4.3 Stage II	9
2.4.4 Stage III	10
2.4.5 Stage IV	11
2.5 Traditional methods of breast cancer treatment	12
2.6 Risk Factors	13
2.6.1 Non modifiable risk factors	14
2.6.2 Modifiable risk factors	15
2.7 Commonly used diagnostic methods	16
CHAPTER 3 Machine Learning	
3.2 Types of Machine Learning	18
3.2.1 Supervised	19
3.2.2 Unsupervised	20
3.2.3 Reinforcement	20

3.3 ML in Healthcare	24
3.3.1 Benefits of Machine Learning	21
	23
CHAPTER 4 PROPOSED METHODOLOGY	25
4.1.1 Data collection	26
4.1.2 Data pre processing	27
4.1.3 Feature Selection	27
4.1.4 Model Selection	28
4.1.5 Training and Testing	28
CHAPTER 5 TECHNICAL APPROACH	29
5.1 Naïve Bayes	29
5.2 Random Forest	30
5.3 Decision Tree	31
5.4 ANN	32
CHAPTER 6 ALGORITHMS AND RESULTS	33
6.1 Naïve Bayes result	33
6.2 Random Forest result	34
6.3 ANN result	35
6.4 Decision Tree result	42
CHAPTER 7 CONCLUSION	44
7.1 Limitations	45

## **LIST OF FIGURES**

Fig. 1. Cell morphology during Cancer

Fig 2. Different stages of Breast Cancer

Fig 3. Classification of machine learning

Fig 4. Flowchart of experimental process

Fig 5. Overview of dataset

Fig 6. Split ray data into test and train dataset

Fig 7. Diagram depicting how naïve bayes classifier works

Fig 8. Decision Tree

Fig 9. Artificial neural network

Fig 10. Naïve Bayes result

Fig 11. Random Forest result

Fig 12. ANN result

Fig 13. Decision Tree result



## **LIST OF TABLES**

TABLE 1 - Morphology of cells during Breast Cancer

TABLE 2 - Description of dataset

TABLE 3 – Comparison of different algorithms

## **LIST OF ABBREVIATIONS**

RF – Random Forest

DC – Decision Tree

BC – Breast Cancer

NB – Naïve Bayes

ANN – Artificial Neural Network

BCR – Breast Cancer Research

SVM – Support Vector Machine

LR – Logistic Regression

CART – Classification and Regression Trees

WHO – World Health Organization

# Chapter 1

## **1.1 INTRODUCTION**

Breast cancer is the most frequent cancer in females worldwide, and it is an important issue for public health. In 2020, it was responsible for 30% of all cancer cases in women and 15% of fatalities caused by cancer in women. All benign breast lesions fail to develop to cancer, and all malignant lesions do not advance to cancer. BC has traditionally had an elevated death and fatality rate as the most communal cancer in women. According to the most recent cancer statistics, BC will account for 25% of every new diagnosis of cancer and 15% of all cancer deaths among women globally. Scientists were conscious of the dangers of BC from the beginning, consequently much early research in the therapy of BC has already been implemented. Researchers' efforts and early detection approaches have resulted in a consistent and lowering trend in mortality rates over the previous decades. According to Cancer Research UK statistics, the overall survival rate after five years for BC is nearly 100% if found early, but can be even lower as 15% if detected late. Recently, ML approaches have played a significant part in determining the cause and treatment of BC by using methods for classification to identify persons with BC, discriminate benign from malignant tumors, and predict prognosis. Accurate classification can also help medics prescribe the best effective treatment regimen.

A combination of preoperative treatments, such as medical screening, mammography, fine-needle cytology aspiration, and core needle biopsy, can, nevertheless, increase diagnosis accuracy. Despite considerable downsides, multiple diagnostic processes are more accurate, dependable, and acceptable than a single medical approach. Physical examination, imaging procedures (including mammography, ultrasound, and MRI imaging), and biopsy are all classic modalities of breast cancer care.

These methods have disadvantages, such as false positives, false negatives, and subjectivity. Furthermore, breast cancer treatment entails a variety of techniques, such as surgery, chemotherapy, radiation, and targeted therapy, and can be complicated and necessitate a personalised strategy. Numerous statistical examination and ML strategies for breast cancer analysis and classification have been developed over the last few decades, which may be separated into three major stages:

preprocessing, extraction of characteristics, and categorization.

Preprocessing mammography films improves visibility of the perimeter and intensity distribution, which aids in interpretation and analysis, and numerous approaches have been documented to aid in this process.

Machine learning is fetching more popular, and it is on its way to becoming a service. Unfortunately, machine learning remains a difficult field that frequently necessitates expert knowledge. Machine learning (ML) is now recognised as a viable strategy for enhancing breast cancer management by permitting for more objective, precise, and personalised analysis and treatment. ML algorithms can analyse enormous datasets and uncover patterns that traditional approaches may miss. ML has been used in breast cancer care in a variety of ways, including risk prediction, evaluation, therapy planning, and prognosis.

## **1.1 Purpose**

In order to diagnose breast cancer, this thesis compares and evaluates the effectiveness of different machine learning algorithms using a variety of data sources. In order to increase the precision and reliability of diagnosis, the thesis attempts to determine the most efficient algorithms for correctly identifying breast cancer cases, choosing pertinent features, and integrating numerous data sources. In order to create a computerised diagnostic system or a decision-support tool that can improve the accuracy, effectiveness, and affordability of breast cancer diagnosis, the thesis also aims to evaluate the additional benefits of machine learning methods compared to conventional diagnostic methods.

## **1.2 Research Question**

1. How may machine learning techniques be used to identify breast cancer?
2. How can the suggested machine learning model for breast cancer detection and diagnosis be compared?
3. What machine learning model would be the most effective for detecting breast cancer?

## **1.3 Motivation and Problem statement**

As a result of numerous important factors, machine learning has become a potent tool for diagnosing breast cancer. First of all, by analysing huge and varied datasets, such as medical imaging, patient records, and genetic data, algorithms for machine learning have the capability to greatly enhance accuracy. These algorithms can produce more accurate predictions and lower the probability of misdiagnosis by identifying patterns and signals that might be suggestive of breast cancer. Second, early detection is essential for successful treatment and better patient results. Machine learning methods can use complex data to find tiny patterns and signals linked to early-stage breast cancer, allowing for prompt treatment and improving the likelihood of success. Last but not least, given the complexity of breast cancer diagnosis and the way machine learning algorithms can handle complex data, they are ideally equipped for this task. These mathematical frameworks can provide more thorough and accurate assessments, assisting healthcare providers in determining what to do by combining numerous data sources and taking into account a wide range of aspects.

## **CHAPTER 2**

### **2.1 Background**

Breast cancer is a chief public health issue around the world, and there is growing awareness of the possible application of machine learning to boost its care [1]. One study discovered that machine learning methods outperformed human radiologists in identifying breast cancer in mammography pictures. In another study, machine learning was found to be useful for forecasting breast cancer risk based on a person's medical history and other risk factors.

Traditional breast cancer care procedures, like mammography and biopsy, have accuracy and efficiency limits. By discovering patterns and links in massive datasets, machine learning has the potential to eliminate these constraints and advance breast cancer care [2]. Additionally, doctors can utilise machine learning to develop individualised treatment plans based on the distinct genetic knowledge and medical background of each patient.

However, applying machine learning in the treatment of breast cancer has a number of downsides and restrictions. The precision and flexibility of machine learning algorithms, for instance, may be impacted by the quantity and quality of data [3]. Data privacy, prejudice, and responsibility are further ethical concerns raised by machine learning in healthcare.

Overall, a complete summary of the existing state of machine learning applications in breast cancer management is required [4]. The thesis "Revolutionising Breast Cancer Management: Harnessing the Power of Machine Learning in Diagnoses and Treatment" seeks to present such an overview while also identifying potential and problems connected with the use of machine learning into clinical practise.

## 2.2 Breast cancer

The greatest prevalent kind of cancer in women universally is breast 1 the difference between benign and malignant tumors is critical in medical fields and cancer research. Furthermore, having this information may assist doctors in determining the best strategy for controlling and treat cancer, particularly breast cancer [5]. In contrast to benign tumors, which develop gradually and do not poliferate, malignant tumors develop rapidly. Malignant tumors are cancerous, while benign tumors are not. Breast cancer identification is critical for successful therapy and better outcomes. Women should self-examine their breasts and have regular mammograms as prescribed by their healthcare physician. The risk factors for breast cancer include becoming significant, having a family antiquity of the disease, having definite genetic abnormalities, and being exposed to hormones.

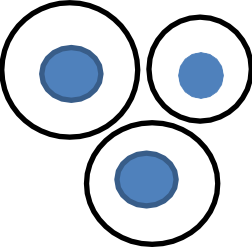
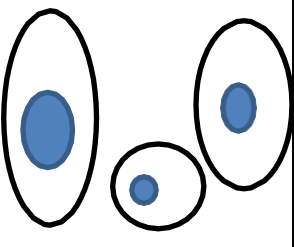
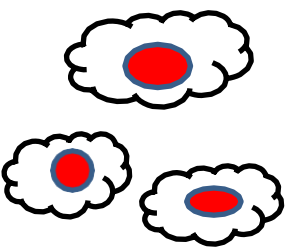
Cell Morphology			
Growth	Normal and regulated growth	Slow growth	Rapid Growth
Chromosome	Diploid	Diploid	Aneuploidy
Metastasis	Grow in one location	Do not spread	Spread in other organs
Cancer	No	Non-cancerous	Cancerous

Fig.1. Cell morphology during cancer

## 2.3 Symptoms

- A tumour or mass in the breast region or underarm area: The appearance of an abnormal mass or lump in the tissues of the breast or under the arm is the most prevalent symptom of breast cancer [6]. The lump could be firm, hard, or uneven in shape, and it could be painful or not.
- The breast area's size or shape could change as a result of breast cancer., such as swelling, thickness, or wrinkles of the skin.
- Breast cancer can produce nipple alterations such as inversion (when the nipple bends inward), discharge (which can be apparent bloody, or coloured), or a scaly, crusty, or inflammatory region surrounding the nipple.
- Breast pain or tenderness: Although uncommon, some women suffering breast cancer might notice tenderness or pain in the breast.
- Changes in the skin: Breast cancer can produce modifications to the outermost layer of the breast, including inflammation, warmth, or thickness

## 2.4 Stages of breast cancer

Even though breast cancer is localised to the breast, the stage can still be near. If nearly all of the tumours are detected in lymphatic nodes, which are commonly in the armpit, it is localized [7]. While the disease has migrated to different body areas, distant breast carcinoma is at this stage.

Breast cancer is divided into various stages based on the extent and dissemination of the tumour as well as the presence or absence of tumour cells in nearby lymph nodes or other bodily parts. After receiving the lab results from your breast surgery or biopsy, the stage can be changed. The T-N-M scale, which relates to tumour size [8], adjacent lymph node involvement, and whether the tumour has grown or metastasized (spread) outside of the breast to other regions of your body, is what your doctor will consider when assessing the phase of your cancer.

Tumor size (T): This reveals the primary tumor's size and whether it has banquet to neighbouring tissues. T stage consists of stage T0 to T4, with larger values denoting bigger tumours and greater levels of



invasion.

- TX: evaluation of the original tumour is possible.
- T0: The initial tumor's presence is not apparent.
- Tis: In situ (DCIS or Paget's disease of the breast without an associated tumour)
- T1: greater than or roughly 2 cm in original tumour diameter
- T2: The first tumor's diameter ranges from 2 to 5 cm.
- T3: Original tumor diameter is > 5 cm
- T4: The original tumour has attacked various organs, including the skin and chest wall.

Lymph node involvement (N): If neighboring lymph nodes have been exaggerated by cancer cell spread, this is referred to. A larger number indicates that more lymph nodes have been pretentious; it is represented by the initial N that follows a symbol or number [9]. The N group for classifying breast cancer has five subcategories. The N stage is divided into N0 through N3 categories. N0 denotes no contribution of the lymph nodes, while N1, N2, and N3 denote increasing levels of lymph node involvement.

- NX: Lymph nodes are incapable to be estimated
- N0: There is no proliferation to neighboring lymph nodes
- N1: Less than 3 lymph nodes underneath the arm or any number of internal breast lymph nodes around the sternum have been grandiose by cancer.
- N2: Between 4 and 9 underarm lymph nodes have been affected by cancer, as have the interior mammary lymph nodes.

- N3: The malignancy has expanded to more than 10 lymph nodes under the arm, lymph nodes under the collarbone, and lymph nodes immediately above the collarbone on the same side of the body, to at least 1 underarm lymph node and expanded breast bone lymph nodes [10], to at least 4 underarm lymph nodes and cancer has been recognised through biopsy to have spread to those lymph nodes.

Metastasis (M): This shows if the cancer has grown beyond of the immediate area of the body. The M division for screening breast cancer only has two classifications, either followed by a 0 or a 1. M1 denotes the existence of metastasis, while M0 denotes the absence of distant metastases.

- M0: Imaging scans did not reveal any spread in any remote parts of the body.
- M1: Imaging studies have revealed developed in one or more remote body parts, including the bones or brain, and a biopsy has been performed to confirm an excursion of 0.2 millimetres or more.

#### **2.4.1 Stage 0 (in situ carcinoma):**

Without penetrating neighbouring tissues, cancer cells are only found in the portion of cells enclosing milk ducts or lobules. This degree explains the non-invasive breast cancer subtype known as ductal tumour in situ (DCIS) [10]. It is a stage in which there are no signs of cancer cells forming in any breast tissue or encroaching on neighbouring healthy tissue. (WHO, 2014).

#### **2.4.2 Stage I:**

refers to a type of early-stage breast cancer where the tumour is small and only affects the breast or nearby lymph nodes [11]. The breast cancer tumour staging system helps to assess the condition's severity and guides treatment decisions.

In Stage 1 breast cancer, there are two subcategories:

1. Stage 1A: The tumour is only up to 2 cm in size and hasn't spread to the lymph nodes or any other part of the body.
2. Stage 1B: There are small clumps of cancer cells in the lymph nodes that are between 0.2 mm and 2

mm in size, and the tumour ranges from 0.2 cm and 2 cm in size.

Surgery, radiation therapy, and occasionally chemotherapy or targeted therapy are common options for treatment for Stage 1 breast cancer [11]. The characteristics of the tumour, the individual's overall well-being, and their treatment preferences will all influence the particular course of therapy.

### **2.4.3 Stage 2**

Compared to Stage 1, stage 2 breast cancer is a little more sophisticated stage of the disease. The tumour is larger than in Stage 1 breast cancer and may have blowout to surrounding lymph nodes [12], but it has not yet reached remote regions of the body in Stage 2.

1. Stage 2A: In this stage, one of the following situations applies:

- Even though the tumour is smaller than 2 cm, the cancer has already spread to the nearby axillary lymph nodes, that are located beneath the arm.
- The tumour is not spreading to the lymph nodes and ranges in size from 2 to 5 centimetres.

2. Stage 2B: In this stage, one of the following conditions applies

- The tumour has progressed to the adjacent lymph nodes under the arm and varies between 2 and centimetres in size.
- Despite being over 5 cm in size, the tumour has not yet metastasized to the lymph nodes.

### 2.4.4 Stage 3

Breast cancer is a stage of the disease where the tumour has progressed to larger parts of the body in addition to the breast and adjacent lymph nodes [13]. There are three subtypes of stage 3 breast cancer: 3, A, B, and C. The specific characteristics of each subcategory are as follows:

1. Stage 3A: In this stage, one of the following conditions applies:

- Although there isn't a tumour in the breast, interior mammary lymph nodes surrounding the breastbone do contain cancerous cells.
- The tumour, which is larger than 5 cm in size, has affected one to three axillary lymph nodes, or the tumour, regardless of size, has damaged four to 9 axillary lymph nodes.
- Although the tumour has not moved to distant areas of the body, it has grown into adjacent structures such as the skin or chest wall.

2. Stage 3B: In this stage, one of the following conditions applies:

- The skin or chest wall have been affected by the tumour [14], which might be of any size. Additionally, it could result in edoema other locations.
- Internal mammary lymph nodes, which are close to the breastbone, have seen the cancer's spread. The axillary lymph nodes may also be affected.

3. Stage 3C: Depending on the sum of lymph nodes associated and if the disease has progressed to other body parts, grade 3C breast cancer can be further separated into three subgroups. During this stage, a sizable tumour and numerous lymph nodes are frequently involved.

Breast cancer is routinely treated with a variety of therapies, including surgery (such as a mastectomy), chemotherapy, radiation therapy, targeted therapies [15], and hormone therapy. The treatment plan is influenced by the nature of the tumour, the involvement of the lymph nodes, the hormone receptor status, genetic factors, and the general health of the patient.

#### 2.4.5 Stage 4

Breast cancer with metastatic spread or advanced breast cancer is the most severe type of the disease. In addition to the breast and nearby lymph nodes at this stage, the cancer has spread to other distant organs or bodily tissues [16]. It might affect the liver, lungs, brain, bones, liver, or liver.

Although treatment seeks to manage the illness, control symptoms, and enhance quality of life despite the fact that breast cancer is thought to be incurable. Stage 4 breast cancer is often treated with a mix of medicines, some of which may include

1. Systemic therapies: These treatments are directed through the bloodstream to spread cancer cells all over the body. They include:

- Chemotherapy: Cancer cells can be killed or slowed down in growth with medications.
- Hormonal therapy: It seeks to inhibit or decrease the effects of hormones that promote the formation of hormone receptor-positive breast tumours.
- Targeted therapy: To prevent the growth and spread of cancer cells, it attacks particular traits of those cells.
- Immunotherapy: The immune system is prompted to find and target cancer cells as a result.

2. Localized treatments: These treatments concentrate on symptom management and cancer control in particular body regions. They include:

- Radiation therapy: It can be applied to treat symptoms and reduce tumour size in particular locations, including bone metastases or brain metastases.
- Surgery: Surgery may be necessary in some situations to remove a small tumour or to treat particular issues, including spinal cord compression.

The extent of metastases, the presence or absence of hormone receptors, the presence or absence of HER2, the patient's overall health, and individual inclination all play a part in the therapy strategy for breast cancer at stage 4 [17]. Working closely with a medical expert or oncologist with advanced cancer care is critical for those with Stage 4 breast cancer. Based on the most recent scientific developments and unique situations, they can offer specialised options for therapy, support, and direction. Palliative care may also be used to control symptoms, offer emotional support, and enhance general quality of life.

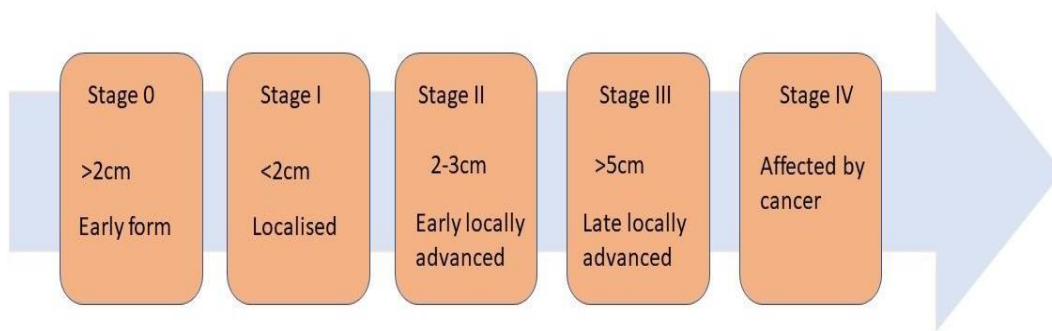


Fig.2. Different stages of Breast Cancer

## 2.5 Traditional methods of breast cancer management

The modern medical environment is distinguished by the acquisition of massive amounts of patient, a medical facility, and administrative data, making traditional ways to studying this data individually less suitable.

Mammography, biopsy, and histopathology are traditional procedures in breast cancer care. Mammography is a type of screening that employs X-rays to identify abnormal growths or mutations in breast tissue [18]. A biopsy is a procedure that comprises of extracting a sample of breast tissue and studying it under a microscope to dictate whether it is malignant. Cell abnormalities and cancer extent are determined by histopathology by inspecting tissue samples. These approaches, while effective in the identification and diagnosis of breast cancer, have limits in terms of precision and efficacy [19]. Mammography, for example, has a comparatively high false-positive rate, which might result in unneeded biopsies and distress for patients. Biopsies are also uncomfortable and invasive, and they do not always offer a conclusive diagnosis.

Machine learning (ML), on the other hand, can incorporate big and complicated data sets. These sources of information have the ability to improve patient outcomes and care. Increases in omics-data are inextricably linked to a personalised medical strategy [20]. DNA sequence databases, for example, double in size every two years. Indeed, breakthroughs in computer processing, along with rapid cost reductions in genetic sequencing, have overtaken the rate of computing hardware advancement.

In addition, some patients may not benefit from standard breast cancer management strategies. Some women, for example, may have thick breast tissue, which might make mammography abnormalities more challenging. Furthermore, traditional methods of detection may be ineffective for certain kinds of breast cancer, especially inflammatory breast cancer, which can present with symptoms that are not specific.

Detection and diagnosis of breast cancer can be made easier using traditional management methods. However, there is a need for better and more precise procedures [21]. Machine learning has become a viable method for improving breast cancer management by analyzing vast datasets and detecting patterns that human specialists may miss.

During the development of machine learning algorithms, a focus is placed on making algorithms that automatically identify patterns and relationships within data without prior programming. Machine learning algorithms are designed to find patterns and relationships in.

## **2.6 Risk Factors**

The precise cause of carcinogenesis has yet to be determined, however different factors that increase the risk of developing breast cancer are known. According to the epidemiological data presented above, a country's age, sex, and stage of development are some of the most important factors. The number of births, the age at which the first kid was born, and nursing are all important procreative factors [22]. Additionally significant are hormonal parameters that are mostly related to the time of oestrogen exposure. Genetic predispositions, the use of hormone replacement treatment, a poor diet, and concurrent obesity all have a significant role in the development of breast cancer.

As substantial risk factors for a breast cancer diagnosis, alcohol consumption, early ionising radiation exposure, and hormonal contraception are also mentioned.

### **2.6.1 Non modifiable factors**

1. Age - Nearly 80% of breast cancer patients today are over the age of 50, and more than 40% are over the age of 65. Age 40 brings an increase in risk of 1.5%, age 50 an increase of 3%, and age 70 an increase of more than 4%. Unexpectedly, a link between a patient's age and a specific molecular cancer variant was found [23]. Highly resistant triple-negative breast cancer variants are most frequently detected in patients under the age of 40, while luminescent breast cancer variants are detected in patients over forty years of age. Patients above 70 are typically where a subtype is discovered. The occurrence of tumours in old age is often not limited to breast cancer; with time, carcinogenesis increases due to the accumulation of numerous cellular mutations and susceptibility to potential carcinogens.

2. Family history – Family tradition of the disease is a substantial risk factor associated with an increased possibility of developing breast cancer. Between 13 and 19% of breast cancer sufferers have a first-degree relative who has the disease. Additionally, the number of first-degree relatives who develop breast cancer significantly increases the probability of developing it. The risk could be significantly higher for people under 50 who are affected [24]. No matter their age, everyone with a family history of breast cancer had significantly higher incidence rates overall.

3. Density of breast tissue - Even though a person's breast tissue density changes throughout the course of their life, many classifications, including low-density, high-density, and obese breasts, have been established in medical practise. Younger women with lower BMI who are either breastfeeding or becoming pregnant, as well as those who are on hormone replacement therapy, have greater breast density.

4. Reproductive history -In especially for oestrogen and progesterone, numerous studies have linked endogenous hormone exposure to an increased risk of breast cancers in females. To determine the likelihood of the start of carcinogenic activities in the breast microenvironment, it is critical to consider the timing, duration, and any associated hormonal imbalance of specific situations like pregnancy, breastfeeding, the first menstrual period, and menopause.

5. Density of breast - Though the thickness of breast tissue changes over the course of a person's life, there are a number of classifications, including low-density, high-density, and chubby breasts, that have



been established in clinical praxis [25]. Females who are turning mothers or nursing, as well as those who are receiving hormonal replacement therapy, tend to have greater breast density. Both premenstrual and postmenopausal females exhibit the general trend of increased breast tissue density being related with an increased risk of breast cancer.

## **2.6.2 Modifiable Factors**

1 Physical activity - Women who do not have close relatives with previous diagnoses of breast cancer can nevertheless benefit from physical activity. In contrast to the earlier study, Thune et al. reported more significant effects in females who had not yet achieved menopause. A number of theories have been put forth to explain how physical activity lowers the risk of breast cancer. These theories include the possibility that exercise may lessen the interaction with endogenous sex hormones, change immunological reactions, or increase levels of insulin-like growth factor-1.

2 Alcohol Intake - Although excessive alcohol use has been connected to an augmented risk of breast cancer, it has also been proven to cause an increased risk of gastrointestinal malignancies. The quantity of alcoholic beverages consumed, as opposed to the type of alcohol, has the greatest impact on the risk of cancer.

3 Smoking - Breast tissue is exposed to tobacco carcinogens, which increases the possibility that oncogenes and suppressor genes will change [26]. Smoking, both active and passive, thus, greatly contributes to the beginning of pro-carcinogenic processes. In addition, women who have a family tradition of breast cancer are more likely to smoke more frequently and before their first full-term pregnancy.

4 Chemical exposure - Chronic chemical exposure can quicken breast carcinogenesis by altering the tumour microenvironment, causing epigenetic changes, and activating pro-carcinogenic pathways.. Breast cancer risk is significantly increased in females who are exposed to chemicals over an extended period of time, and this risk is further modified by the duration of the exposure.

5 Deficiency of Vitamins - Although the exact mechanism of vitamins' anticancer benefits is unknown, they may help prevent a variety of malignancies, including breast cancer [27]. The impact of vitamin consumption (vitamin C, E, and folic acid) on the risk of breast cancer is often studied; however, the findings are still contradictory and inadequate to allow for comparisons and the drawing of meaningful conclusions.

### **2.3 Commonly used diagnostic methods**

1. **Mammography:** The breast tissue is examined via mammography, a kind of X-ray imaging. It has the ability to find minute anomalies, including lumps or calcifications, that could be signs of breast cancer. Women over 40 or those at elevated risk for breast cancer are often advised to get mammograms.
2. **Breast Ultrasound:** The pictures of the tissue in the breast are created by ultrasound using sound waves. It can aid in distinguishing between solid tumours and cysts that contain fluid. Ultrasound is frequently employed to further assess anomalies seen on a mammogram or to direct needle biopsies.
3. **Magnetic Resonance Imaging (MRI):** MRI produces finely intricate pictures of the breast using strong magnets and radio waves. It is particularly helpful in identifying cancer in high-risk women or determining the degree of malignancy in the breast and its adjacent tissues.
4. **Biopsy:** Breast cancer can only be accurately diagnosed through a biopsy. A biopsy involves the removal of an insignificant amount of breast tissue for microscopic examination. There are numerous kinds of biopsies, including surgical and needle-based biopsies (Fine-needle aspiration biopsy, incisional biopsy, and excisional biopsy). The kind, grade, and state of the cancer's hormone receptors can all be determined through biopsies.
5. **Breast Self-Examination (BSE):** Frequent breast examinations can help people become comfortable with their breasts and see any changes or abnormalities, even if it is not a diagnostic technique. It is crucial to speak with a healthcare practitioner for additional assessment if any suspicious tumours or changes are seen.

6. Clinical Breast Examination (CBE): During a medical breast assessment, a healthcare professional physically inspects the breasts and the areas around them to look for any anomalies, such as lumps or modifications to texture or shape.

## **CHAPTER 3**

### **3.1 Machine Learning**

According to Arthur Samuel, the field of study that enables computers to learn despite being explicitly taught is known as machine learning. Machine learning can be supervised or unsupervised. If you have fewer data and training data that is clearly indicated, go with supervised learning. For large data sets in general, unsupervised learning would offer higher performance and outcomes.

Numerous advances have been made in a wide range of scientific fields as a result of the improvement of data availability, expansion in processing capacity, and the introduction of novel learning techniques. This also pertains to scientific and medical research, where applications for molecular biology, picture data analysis, and clinical practise are only a few examples [28]. The concept of a computer constantly learning some abstract concepts, however, has existed at least since the 1950s, when the very first neural networks were created. Prior to that, similar concepts were applied to other techniques like Markov chains and Bayesian statistics. The pharmacometrics and medical pharmacology communities have several names for many of these techniques. If the dataset has been adequately trained and the model used has been thoroughly validated, machine learning methods can be trusted. There are a lot of prospects for ML, and as the education sector adapts to new realities, ML and other emerging technologies may become more important in areas like learning analytics, e-learning, evaluations of students, performance prediction, and monitoring the progress of teaching and learning. Several techniques, including multilayer perceptron (MLP) and CART regression tree models (CART), have diverse potentials that academic stakeholders are very interested in.

Machine learning uses mathematical techniques that are utilised as software programmes to find patterns

in huge datasets and iteratively get better at it as more data is added. The algorithms are frequently used in many different fields and applications, such as advertisements, insurance, banking, social media, and fraud detection, and they have access to a wide variety of data types that are gathered in real-time from many sources [29]. However, as patient data is frequently unavailable for public study, employing these methods to analyse disease outcomes may be challenging.

## 3.2 Types of Machine Learning

Machine learning can be classified into:

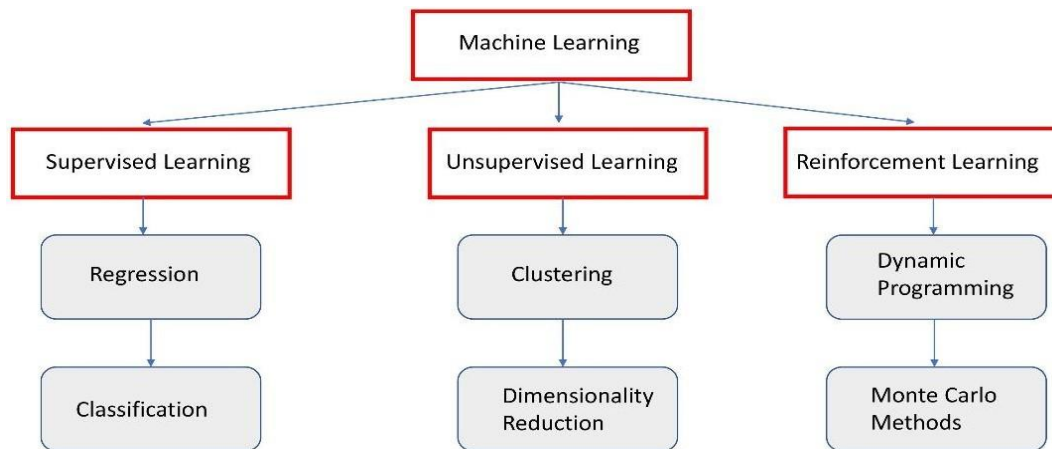


Fig.3. Classification of Machine Learning

### 3.2.1 Supervised learning

Using sample input-output sets, machine learning is frequently used to train an algorithm that transfers an instruction to an output. It employs labelled training data and a group of training samples to conjecture a function. The objective of supervise learning is to generate precise results with a particular set of inputs, which involves a task-driven strategy. "Classification" (information separation) and "regression" (information fitting) are the two most common supervised tasks. One illustration of supervised learning is determining the classification or interpretation of a textual item, such as an internet post or a product review.

In addition, supervised learning can be employed to project whether a patient will respond to certain

mental health interventions (such an obsessive-compulsive disorder programme or cognitive behavioural therapy for depression).

Based on high or low predicted risk/prognosis, risk and prognostic ratings may be especially useful in identifying those who need preventative actions or tailored/intensive therapy. Although the majority of research use predictors that are evaluated at one point in time (for example, pre-treatment) to create risk and prognosis scores.

### **3.2.2 Unsupervised learning**

Pathologists may be less familiar with unsupervised learning than supervised learning, which is a subfield of machine learning [30]. Convolutional neural networks can be used to categorize pathological images, which is another application of supervised learning that pathologists are likely to come across in a diagnostic environment. As with supervised learning, unsupervised techniques rely on patterns that can be detected in unlabeled data without labels, as opposed to supervised learning that requires labelled data.

Clustering, dimensionality reduction methods, autoencoders, and generative adversarial networks (GANs) are examples of unsupervised approaches that are often used. Autoencoders are a type of self-supervised learning, or unsupervised learning. Reinforcement Learning.

### **3.2.3 Reinforcement learning**

Reinforcement learning is a subset of machine learning that teaches a representative how to behave in certain circumstances in order to maximise the incentive signal. The agent develops a strategy that maps contends to actions that maximise its projected cumulative reward over time through interaction with the environment. In reinforcement learning, the agent receives recommendations in the shape of plunders or punishments for its actions.

The agent picks up new skills through trial and error, adapting its behaviour in response to input from the environment. Since input is usually delayed, it is crucial that the agent develop the ability to balance short-term advantages with long-term objectives. Robotics, gaming, and natural language.

### **3.1 ML in Healthcare**

In 1950, Alan Turing presented the use of machine learning. By utilising data and minimising human engagement, machine learning is utilised in healthcare to enhance the efficacy and overall character of care. Since handling of patients in acute healthcare facilities necessitates precise diagnosis and prompt implementation of validated treatments, this environment is probably one in which cognitive expansion of the clinician's delivery of care with artificial intelligence-based technology, such as machine learning (ML), will take place. ML has been applied in medical fields for a range of activities and has even been a component of key clinical workflows in some circumstances. Image categorization is a common application that is represented by haematology digital morphological analyzers (for example, CellaVision).

Based on prior training on massive picture data sets, machine learning algorithms can correctly allocate labels (class) to different cell types. With these models, enormous quantities of data can be analysed much more quickly than with manual differential counting. Because the bulk of ordinary samples will not be highlighted for manual examination, the haematology laboratory gains tremendous efficiency. Although these ML algorithms operate on big data sets and hence reduce superfluous human labour, manual verification continues to be regarded as the gold standard for morphologic diagnosis for specimens that meet criteria to fall within particular aberrant categories. This paradigm may reflect a conservative but essential role for machine learning algorithms, which is to operate as a prescreening assistance to bin cases that require human intervention or review. Combining ML models' ability to identify relevant features from big data sets with traditional human verification of discoveries may thus be the initial stage in ML application.<sup>6</sup> Alternatively, trained ML systems may one day outperform human decision-making in categorization performance.

Learning health systems with adequate resources are more likely to use ML technology to its utmost potential. To prevent increasing healthcare disparities, immediate attention must be given to how these clinical advantages might be made available to those suffering in healthcare systems that are neither well-equipped nor equipped with the appropriate data collection technology.

Healthcare organisations are using ML systems to track and anticipate future pandemic outbreaks around the world. In order to predict disease outbreaks, this digital system makes use of satellite data, immediate notifications on social networks, and other crucial information accessible online.

Third-world nations with sparse healthcare infrastructure may benefit especially from its implementation.

Long wait times, worries about excessive medical costs, problematic appointment scheduling, and trouble finding the right healthcare practitioner are some of the issues that machine learning and associated data-driven techniques aim to address. Some of the difficulties that machine learning and related data-driven techniques attempt to address include lengthy wait times, concerns about too much medical costs, difficult appointment scheduling, and difficulty selecting the correct healthcare provider.

Healthcare data needs to be organised for ML (machine learning) through a procedure called annotation, in which humans tag different aspects of the dataset. This process speeds up pattern identification and enables the drawing of useful conclusions. By analysing the data, developing new rules, and improving machine performance, medical professionals also play a vital role. But precise and pertinent annotations that capture crucial concepts in the right context are required if ML processes in healthcare are to learn quickly and efficiently.

Successful surgical outcomes demand accurate execution, flexibility to adjust to shifting conditions, and a consistent strategy over a lengthy period of time. Even though qualified surgeons already have these attributes, one potential use of ML for medical purposes entails using robots to carry out surgical procedures. Insights from previously obtained data on active ingredients in drugs and their impacts on the body can be used to recreate an active component that functions under comparable settings.



### 3.3.1 Benefits of ML in healthcare

#### 1. Improvement in diagnosis

By analysing medical data and photos with ML-enabled technologies, better diagnoses can be made in the field of healthcare. A machine learning algorithm, for instance, can forecast an illness based on training data from previous cases and perform better pattern recognition. Diagnostic algorithms have struggled to match the accuracy of clinicians in cases of differential diagnosis, where there are several potential explanations of a patient's symptoms, despite major academic efforts and increasing commercial interest. automating cell and tissue quantification for blood and culture research precision. locating illness cells on a diagnostic slide and marking regions of interest. developing paradigms for tumour staging.

#### 2 Cost Reduction

Machine learning can improve the total expense efficiencies of healthcare services by employing automation instead of manual labour. Focusing on the discovery and manufacturing of drugs that are affordable, efficient, safe, and have a small probability of side effects. For instance, the cost of needles and labels on items like operational sheets may be lowered by about 18%. ML can revolutionise the pharmaceutical business, increase drug efficacy and safety, and ultimately help patients by lowering the time and outflow required to bring novel therapies to marketplace.

#### 3 New treatments and medication

A model that utilizes deep learning could accelerate the development of novel medications for many diseases. In order to enhance patient care and safety, ML techniques can also be worn to more effectively analyse the massive volume of facts gathered during clinical trials. Maintaining medical records has become easier thanks to machine learning, which also saves time and money. Future smart medical data based on ML will also help with more accurate and improved clinical diagnosis and therapy recommendations.

Software can perform automated ML and pre-processed information via its machine learning platforms, boosting accuracy and obviating the need for time-consuming tasks that are typically performed by people in a variety of health sectors, involving biopharmaceuticals, accurate medicine and technology, hospitals, and health systems.

#### 4 Clinical Trails and Research

Research and clinical trials benefit greatly from machine learning since it allows for simultaneous access to many different data points. Additionally, it uses computerised record-keeping, real-time monitoring, and trial subjects' data access to lower data-based errors. Clinical inventories and extensive implementation work are sparked by preclinical research and observational studies, which then result in classical trials and trials with pragmatic components. The field of clinical research is vast. Clinical research is essential to enhancing healthcare and outcomes, but it is currently conducted in a complex, labor-intensive, expensive manner that may occasionally be vulnerable to unanticipated biases and errors, endangering its successful application, adoption, and acceptability.

#### 5 Data Collection

To get important insights and enhance patient care, machine learning algorithms examine enormous volumes of data from numerous sources, including digital health records, diagnostic imaging, wearable technology, and genetics. These algorithms can identify trends, categorise diseases, forecast results, and support the development of new drugs. Healthcare practitioners can improve public health surveillance, personalise treatments, and make better judgements by utilising machine learning. However, in this information-driven healthcare environment, privacy and security issues must be resolved to guarantee the confidentiality and safety of patient data.

# CHAPTER 4

## 4.1 Proposed Methodology

The report's methodology is described in the following section. The dataset utilised and its constraints are described in Section 3.1. The mathematical programmes of the ML techniques are presented in Section 3.2 random forest, artificial neural network, Naïve bayes and Decision Tree. This chapter describes the experimental techniques utilised to accomplish the thesis's objectives and respond to its research questions. Four machine learning methods that were tested and trained on the dataset are being compared in the experiment. The four machine learning algorithms are contrasted in terms of their accuracy and precision.

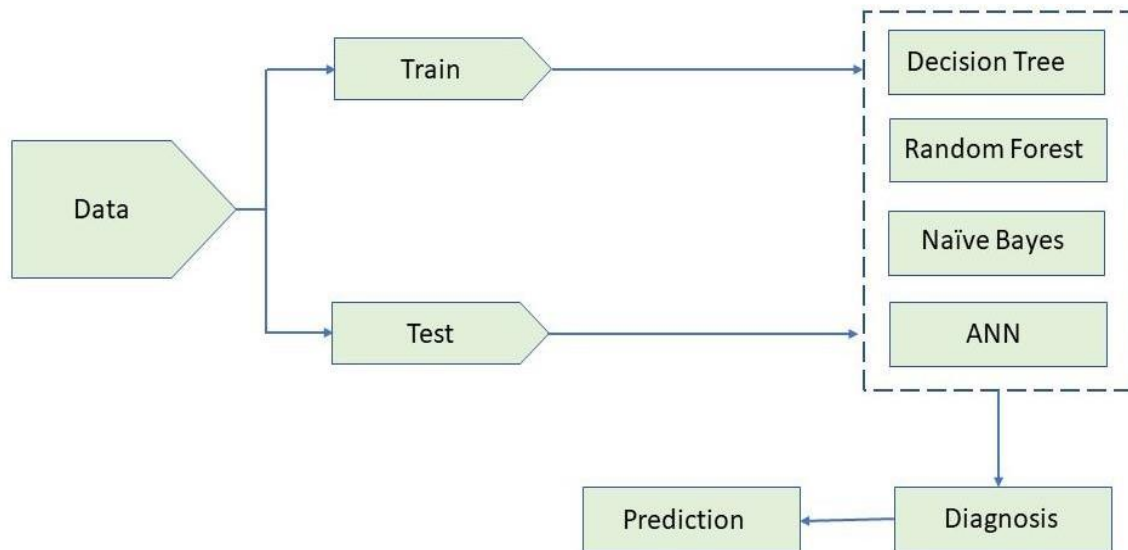


Fig.4. Flow chart of Experiment process

### 4.1.1 Data Collection

Data collection is the first stage in developing a machine learning recognition system. You might need a medical dataset, for instance. The greatest data science community in the world with a wealth of tools and resources, Kaggle, is where the data was gathered. The Wisconsin dataset, was analysed. The acclaimed UCI-Repository for Machine Learning hosts the dataset, which is openly available there. 10 features, including radius, texture, area, perimeter, compactness, smoothness, concave points, concavity, symmetry, and fractal dimension, were confirmed in our dataset of 570 samples.

Radius	The average distance between the centre and points on the outermost edge
Texture	Standard deviation of values in grayscale
Area	Number of pixels inside the snake, plus one-half more pixels around the outside
Perimeter	The nuclear perimeter is the sum of the distances between the snake points.
Compactness	$\text{Perimeter}^2 / \text{area}$
Smoothness	By comparing a radial line's length to the average length of the lines around it, one may quantify regional variations in radius length.
Concave points	The proportion of the contour's concave areas.
Concavity	severity of the contour's concave areas
Symmetry	the difference in length between lines that are parallel to both directions of the major axis and the cell boundary.
Fractal dimension	An approximate coast. A greater number indicates a less uniform contour and, hence, a higher likelihood of cancer.

TABLE 2. Description of Dataset

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	id	radius_me	texture_m	perimeter	area_me	smoothne	compactni	concavity	concave p	symmetry	fractal_dir	radius_se	texture_se	perimeter	area_se	smoothne	compactni	concavity	concave p	symmetry	fractal_dir	radius_wo	texture_w
2	842302	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095	0.9053	8.589	153.4	0.006399	0.04904	0.05373	0.01587	0.03003	0.006193	25.38	17.33
3	842517	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.7339	3.398	74.08	0.005225	0.01308	0.0186	0.0134	0.01389	0.003532	24.99	23.41
4	84300903	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	0.7456	0.7869	4.585	94.03	0.00615	0.04006	0.03832	0.02058	0.0225	0.004571	23.57	25.53
5	84348301	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	0.4956	1.156	3.445	27.23	0.00911	0.07458	0.05661	0.01867	0.05963	0.009208	14.91	26.5
6	84358402	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	0.7572	0.7813	5.438	94.44	0.01149	0.02461	0.05688	0.01885	0.01756	0.005115	22.54	16.67
7	843786	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	0.3345	0.8902	2.217	27.19	0.00751	0.03345	0.03672	0.01137	0.02165	0.005082	15.47	23.75
8	844359	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	0.4467	0.7732	3.18	53.91	0.004314	0.01382	0.02254	0.01039	0.01369	0.002179	22.88	27.66
9	84458202	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	0.5835	1.377	3.856	50.96	0.008805	0.03029	0.02488	0.01448	0.01486	0.005412	17.06	28.14
10	844981	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	0.3063	1.002	2.406	24.32	0.005731	0.03502	0.03553	0.01226	0.02143	0.003749	15.49	30.73
11	84501001	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243	0.2976	1.599	2.039	23.94	0.007149	0.07217	0.07743	0.01432	0.01789	0.01008	15.09	40.68
12	845636	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697	0.3795	1.187	2.466	40.51	0.004029	0.009269	0.01101	0.007591	0.0146	0.003042	19.19	33.88
13	84610002	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.06082	0.5058	0.9849	3.564	54.16	0.005771	0.04061	0.02791	0.01282	0.02008	0.004144	20.42	27.28
14	846226	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078	0.9555	3.568	11.07	116.2	0.003139	0.08297	0.0889	0.0409	0.04484	0.01284	20.96	29.94
15	846381	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.05338	0.4033	1.078	2.903	36.58	0.009769	0.03126	0.05051	0.01992	0.02981	0.003002	16.84	27.66
16	84667401	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.07682	0.2121	1.169	2.061	19.21	0.006429	0.05936	0.05501	0.01628	0.01961	0.008093	15.03	32.01
17	84799002	14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303	0.07077	0.37	1.033	2.879	32.55	0.005607	0.0424	0.04741	0.0109	0.01857	0.005466	17.46	37.13
18	848406	14.68	20.13	94.74	684.5	0.09867	0.072	0.07395	0.05259	0.1586	0.05922	0.4727	1.24	3.195	45.4	0.005718	0.01162	0.01998	0.01109	0.0141	0.002085	19.07	30.88
19	84862001	16.13	20.68	108.1	798.8	0.117	0.2022	0.1722	0.1028	0.2164	0.07356	0.5692	1.073	3.854	54.18	0.007026	0.02501	0.03188	0.01297	0.01689	0.004142	20.96	31.48
20	849014	19.81	22.15	130	1260	0.09831	0.1027	0.1479	0.09498	0.1582	0.05395	0.7582	1.017	5.865	112.4	0.006494	0.01893	0.03391	0.01521	0.01356	0.001997	27.32	30.88
21	8510426	13.54	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	0.1885	0.05766	0.2699	0.7886	2.058	23.56	0.008462	0.0146	0.02387	0.01315	0.0198	0.0023	15.11	19.26
22	8510653	13.08	15.71	85.63	520	0.1075	0.127	0.04568	0.0311	0.1967	0.06811	0.1852	0.7477	1.383	14.67	0.004097	0.01898	0.01698	0.00649	0.01678	0.002425	14.5	20.49
23	8510824	9.504	12.44	60.34	273.9	0.1024	0.06492	0.02956	0.02076	0.1815	0.06905	0.2773	0.9768	1.909	15.7	0.009606	0.01432	0.01985	0.01421	0.02027	0.002968	10.23	15.66
24	8511122	15.24	14.26	102.5	704.4	0.1072	0.2125	0.2077	0.00756	0.1531	0.07021	0.4200	0.7006	2.204	44.01	0.006700	0.05210	0.06416	0.02153	0.02673	0.004204	18.07	10.00

Fig.6. Overview of Dataset

#### 4.1.2 Data pre-processing

One of the key phases in machine learning is the pre-processing of data. The most crucial step in creating more accurate machine learning models, it is. The handling of missing values, the elimination of outliers, and any potential problems are all part of this process. Data that is category or ordinal must be transformed in some way into numeric features since, as we know, machine learning (ML) algorithms can only handle numerical features. Median values were used in place of missing values.

#### 4.1.3 Feature selection

The next step is to decide which pertinent features or variables will be utilised in training the model. By choosing traits that are pertinent to the health condition being detected, a process referred to as feature extraction is carried out. In order to find and rank the most crucial aspects in the dataset, activation processes and neurons are used at this stage.

#### 4.1.4 Model selection

The machine learning framework is chosen after the feature selection and data preparation. Numerous algorithms are used in machine learning, including decision trees, random forests, naive bayes, artificial neural networks, and more. For better performance on tasks like picture segmentation, classification, and disease classification, these models make use of deep learning methodology and optimisation approaches.

Python libraries are used to run the model.

#### 4.1.5 Training and Testing

An algorithm for machine learning operates in two stages as we process datasets. For testing and training phases, we typically divide the data by 20% to 80%. We present the model with brand-new data that we have labels for in order to assess how well the algorithms perform. This is often accomplished by dividing the labelled data that we have gathered into two sections using Technique `train_test_split`. The training data, also known as the training set, comprises 75% of the data that we utilised to create our machine learning model. Test data or test set refers to the 25% of the collected information that will be utilised for assessing how effectively the model performs. We evaluate the results after testing the algorithms to choose the algorithm that offers the highest accuracy and determine the most predictive model for the identification of breast cancer.



Fig.6. Split Raw Data into Training and Testing Dataset

# **CHAPTER 5**

## **5.1 Technical Approach**

### **5.1.1 Naïve Bayes**

Naive Bayes is a powerful statistical classification technique that has been utilised successfully in bioinformatics. Based on previous knowledge and present evidence, Bayes' theorem can be utilised to create predictions. The prediction changes as evidence accumulates. Although most input data are categorical or binary in nature, Naive Bayes can handle continuously input features by splitting them into categories. It can handle data gaps by either ignoring them or creating a new category for them. The prediction is the subsequent probability that the investigators have an interest in. Prior knowledge is referred to as prior probability, and it is the most likely bet about the result in the absence of more evidence. The current evidence is expressed as likelihood, which represents the likelihood of a predictor given a specific result.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

In naive Bayes classification, A represents categorical outcome events and B represents a set of predictors. The term "naive" denotes that the predictors are independent of one another if the outcome value is the same. As a result,  $P(b_1, b_2, b_3|A)$  may be represented as  $P(b_1|A) P(b_2|A) P(b_3|A)$ , making the computation procedure considerably easier. Because of its accessibility, speed, and accuracy, Naive Bayes is a popular method for classification applications. It has applications in many fields, including text classification, classification of images, medical evaluation, identifying fraud, and recommendation systems. By taking into account the quantity of phrases or words in the text, Naive Bayes is extensively used in text classification for filtering spam, sentiment assessment, and document categorization.

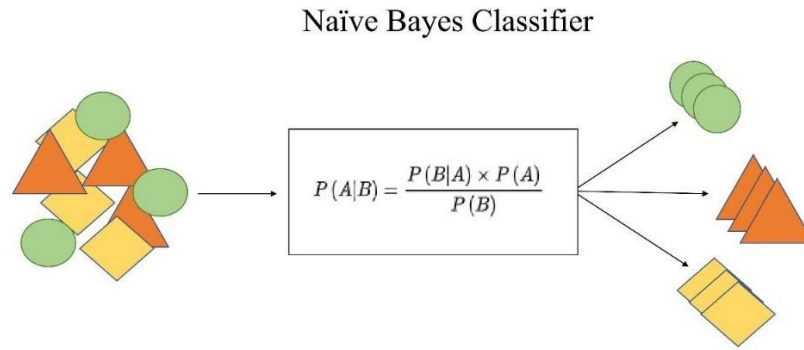


Fig.7. Diagram depicting how Naive Bayes Classifier works

### 5.1.2 Random Forest

Using machine learning methods, such as random forests, a random forest classifier was developed to predict AL occurrence. The concept of a random forest can be described as an algorithm for predicting trees using values derived from a random vector examined autonomously and with exactly the same dispersal for every tree in the forest. As the total number of trees in a forest grows enormous, the generalisation error approaches a limit. Forests of tree classifiers are characterised by their strong trees as well as their correlation, which determines their generalisation error.

It is advantageous to use a random forest structure in two ways. The first class is from a computational standpoint, and the second is from a statistical standpoint. Compared to other classification and regression algorithms, random forests can handle both challenges, which is advantageous in terms of computation. Because the train and prediction procedures in this classifier are executed at rapid speeds, the random forest is referred to as a fast classic classifier. In addition to its direct application to high-dimensional problems, random forests can be used as a form of optimization.

#### Features

- There's no comparison between it and the decision tree technique in relations of accuracy.
- It allows missing data to be handled effectively.
- There is no need to tune hyperparameters in order to get a rational estimation.
- Overfitting in decision trees is solved by this method.
- A subset of features at each node's splitting point is randomly selected for every random forest tree.



### 5.1.3 Decision Tree

One of the most effective methods for data mining, decision trees were created in the 1960s and are now frequently employed in a variety of fields due to their ease of use, lack of uncertainty, and resilience even in the presence of missing data. Both consecutive and discontinuous variables may be employed as preferred or independent variables. The decision tree approach is a popular data mining strategy for developing classification systems based on several parameters or predicting algorithms for an intended variable. With this method, a population as a whole is divided into divisions that resemble branches and create a tree inverted with root, internal, and leaf nodes. The approach handles huge, convoluted data sets without establishing a cumbersome parametric framework since it is non-parametric. The information gathered from the study can be divided into training and evaluation datasets when a study's sample is large enough. Further research has demonstrated that an ensemble of decision trees is typically more accurate than any single tree in many real-world applications, such as SKICAT.

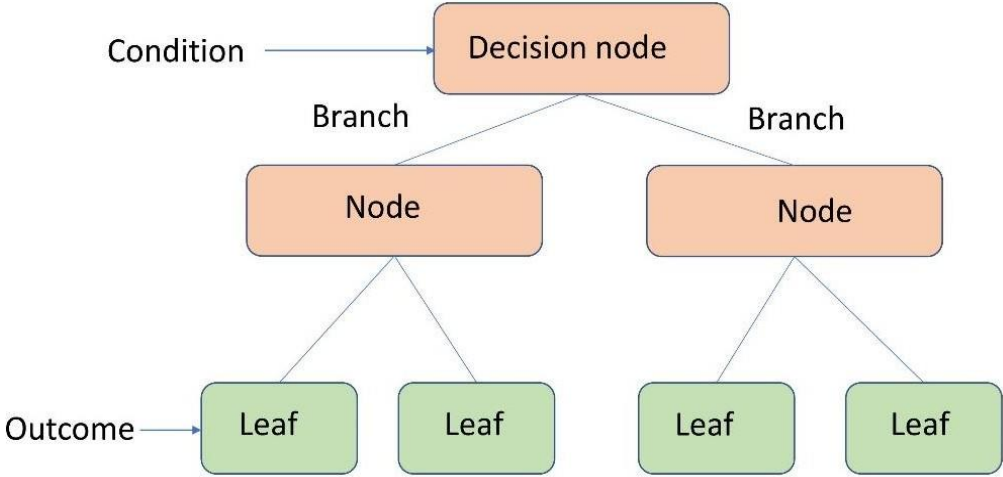


Fig.8. Decision tree

### 5.1.4 Artificial neural network

Software programmes called artificial neural networks (ANNs) are biologically inspired and replicate the way the human brain processes information. Instead than employing programming, ANNs understand (or are taught) by finding trends and correlations in data. They also learn via practise. The mathematical processes used by feature-extracting ANN, which translate multidimensional set of data onto two-dimensional domains, were developed utilising the gradient descent scheme. Nonadaptive Given warped or noisy data, ANN maps might be able to restore their patterns. Associating networks are often used in the pharmaceutical industry in place of conventional response surface approaches, feature-extracting systems in place of prime element analysis, and nonadaptive networks for image acknowledgement. Built on these characteristics, the ANN technique has a wide range of possible application areas in the pharmaceutical sciences, including biopharmacy, drug and medication design, and data interpretations.

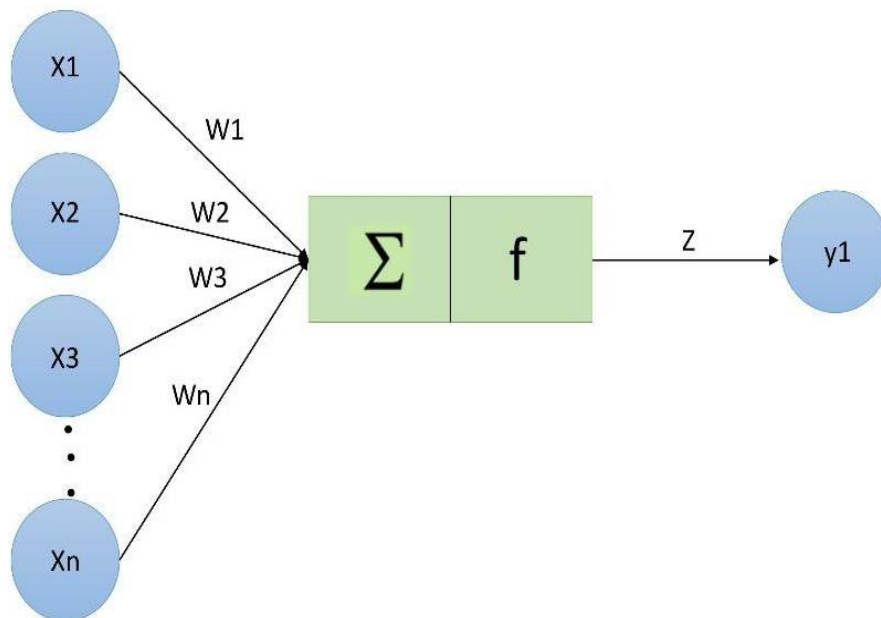


Fig.9. Artificial Neural Network Workflow

# CHAPTER 6

## 6.1 Algorithms and their results

### 6.1.1 Naive Bayes algorithm

```
+ Code + Text
```

```
# Import necessary libraries and modules
from sklearn.datasets import load_breast_cancer
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import accuracy_score, classification_report

# Load the Breast Cancer dataset from scikit-learn library
data = load_breast_cancer()

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(data.data, data.target, test_size=0.2, random_state=42)

# Initialize the Gaussian Naive Bayes model
nb = GaussianNB()

# Train the model on the training set
nb.fit(X_train, y_train)

# Make predictions on the testing set
y_pred = nb.predict(X_test)

# Evaluate the performance of the model
accuracy = accuracy_score(y_test, y_pred)
report = classification_report(y_test, y_pred)
print(f"Accuracy: {accuracy:.3f}")
print(f"Classification report:\n{report}")
```

Accuracy: 0.974  
Classification report:

	precision	recall	f1-score	support
0	1.00	0.93	0.96	43
1	0.96	1.00	0.98	71
accuracy			0.97	114
macro avg	0.98	0.97	0.97	114
weighted avg	0.97	0.97	0.97	114

Fig 10. Naïve Bayes result

## 6.1.2 Random Forest algorithm

```
Q
{x}
# Step 1: Import libraries and load dataset
import numpy as np
import pandas as pd
from sklearn.datasets import load_breast_cancer
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score

data = load_breast_cancer()
X = pd.DataFrame(data.data, columns=data.feature_names)
y = data.target

# Step 2: Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Step 3: Define hyperparameters for random forest model
n_estimators = 100
max_depth = None
min_samples_leaf = 1

# Step 4: Train random forest model
model = RandomForestClassifier(n_estimators=n_estimators, max_depth=max_depth, min_samples_leaf=min_samples_leaf)
model.fit(X_train, y_train)

# Step 5: Make predictions on testing data
y_pred = model.predict(X_test)

# Step 6: Evaluate model performance
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)

print("Accuracy:", accuracy)
print("Precision:", precision)
print("Recall:", recall)
print("F1 score:", f1)

# Step 7: Hyperparameter tuning (optional)

# Step 8: Repeat steps 4-7 until satisfactory performance is achieved

# Step 9: Use final model to make predictions on new, unseen data

Accuracy: 0.956140350877193
Precision: 0.9583333333333334
Recall: 0.971830985915493
F1 score: 0.965034965034965
```

Fig 11: Random Forest result

### 6.1.3 Artificial neural network algorithm

```
# Import necessary libraries
from keras.models import Sequential
from keras.layers import Dense
from sklearn.datasets import load_breast_cancer
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

# Load the dataset
data = load_breast_cancer()

# Preprocess the data
X = data.data
y = data.target
scaler = StandardScaler()
X = scaler.fit_transform(X)

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Build the model
model = Sequential()
model.add(Dense(units=16, activation='relu', input_dim=X.shape[1]))
model.add(Dense(units=8, activation='relu'))
model.add(Dense(units=1, activation='sigmoid'))

# Compile the model
model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])

# Train the model
model.fit(X_train, y_train, epochs=50, batch_size=32, validation_data=(X_test, y_test))

# Evaluate the model
loss, accuracy = model.evaluate(X_test, y_test)
print(f'Test loss: {loss}, Test accuracy: {accuracy}')
```

```
Epoch 1/50
15/15 [=====] - 1s 19ms/step - loss: 0.6822 - accuracy: 0.6835 - val_loss: 0.5948 - val_accuracy: 0.6842
Epoch 2/50
15/15 [=====] - 0s 5ms/step - loss: 0.5217 - accuracy: 0.7516 - val_loss: 0.4790 - val_accuracy: 0.7632
Epoch 3/50
15/15 [=====] - 0s 4ms/step - loss: 0.4360 - accuracy: 0.8176 - val_loss: 0.4032 - val_accuracy: 0.8509
Epoch 4/50
15/15 [=====] - 0s 6ms/step - loss: 0.3763 - accuracy: 0.8725 - val_loss: 0.3459 - val_accuracy: 0.8860
Epoch 5/50
15/15 [=====] - 0s 5ms/step - loss: 0.3262 - accuracy: 0.8989 - val_loss: 0.2959 - val_accuracy: 0.9123
Epoch 6/50
15/15 [=====] - 0s 5ms/step - loss: 0.2858 - accuracy: 0.9121 - val_loss: 0.2522 - val_accuracy: 0.9561
Epoch 7/50
15/15 [=====] - 0s 6ms/step - loss: 0.2504 - accuracy: 0.9187 - val_loss: 0.2175 - val_accuracy: 0.9649
```

```

Epoch 1/50
15/15 [=====] - 1s 19ms/step - loss: 0.6822 - accuracy: 0.6835 - val_loss: 0.5948 - val_accuracy: 0.6842
Epoch 2/50
15/15 [=====] - 0s 5ms/step - loss: 0.5217 - accuracy: 0.7516 - val_loss: 0.4790 - val_accuracy: 0.7632
Epoch 3/50
15/15 [=====] - 0s 4ms/step - loss: 0.4360 - accuracy: 0.8176 - val_loss: 0.4032 - val_accuracy: 0.8509
Epoch 4/50
15/15 [=====] - 0s 6ms/step - loss: 0.3763 - accuracy: 0.8725 - val_loss: 0.3459 - val_accuracy: 0.8860
Epoch 5/50
15/15 [=====] - 0s 5ms/step - loss: 0.3262 - accuracy: 0.8989 - val_loss: 0.2959 - val_accuracy: 0.9123
Epoch 6/50
15/15 [=====] - 0s 5ms/step - loss: 0.2858 - accuracy: 0.9121 - val_loss: 0.2522 - val_accuracy: 0.9561
Epoch 7/50
15/15 [=====] - 0s 6ms/step - loss: 0.2504 - accuracy: 0.9187 - val_loss: 0.2175 - val_accuracy: 0.9649
Epoch 8/50
15/15 [=====] - 0s 5ms/step - loss: 0.2212 - accuracy: 0.9385 - val_loss: 0.1919 - val_accuracy: 0.9649
Epoch 9/50
15/15 [=====] - 0s 5ms/step - loss: 0.1969 - accuracy: 0.9429 - val_loss: 0.1711 - val_accuracy: 0.9649
Epoch 10/50
15/15 [=====] - 0s 5ms/step - loss: 0.1765 - accuracy: 0.9429 - val_loss: 0.1541 - val_accuracy: 0.9649
Epoch 11/50
15/15 [=====] - 0s 5ms/step - loss: 0.1604 - accuracy: 0.9451 - val_loss: 0.1407 - val_accuracy: 0.9649
Epoch 12/50
15/15 [=====] - 0s 5ms/step - loss: 0.1464 - accuracy: 0.9516 - val_loss: 0.1303 - val_accuracy: 0.9649
Epoch 13/50
15/15 [=====] - 0s 4ms/step - loss: 0.1335 - accuracy: 0.9604 - val_loss: 0.1215 - val_accuracy: 0.9737
Epoch 14/50
15/15 [=====] - 0s 5ms/step - loss: 0.1232 - accuracy: 0.9648 - val_loss: 0.1152 - val_accuracy: 0.9737
Epoch 15/50
15/15 [=====] - 0s 5ms/step - loss: 0.1149 - accuracy: 0.9736 - val_loss: 0.1092 - val_accuracy: 0.9737
Epoch 16/50
15/15 [=====] - 0s 5ms/step - loss: 0.1074 - accuracy: 0.9780 - val_loss: 0.1047 - val_accuracy: 0.9737
Epoch 17/50
15/15 [=====] - 0s 4ms/step - loss: 0.1013 - accuracy: 0.9802 - val_loss: 0.1013 - val_accuracy: 0.9737
Epoch 18/50
15/15 [=====] - 0s 4ms/step - loss: 0.0957 - accuracy: 0.9802 - val_loss: 0.0982 - val_accuracy: 0.9649
Epoch 19/50
15/15 [=====] - 0s 4ms/step - loss: 0.0912 - accuracy: 0.9802 - val_loss: 0.0948 - val_accuracy: 0.9561
Epoch 20/50
15/15 [=====] - 0s 4ms/step - loss: 0.0873 - accuracy: 0.9824 - val_loss: 0.0935 - val_accuracy: 0.9649
Epoch 21/50
15/15 [=====] - 0s 5ms/step - loss: 0.0834 - accuracy: 0.9824 - val_loss: 0.0921 - val_accuracy: 0.9561
Epoch 22/50
15/15 [=====] - 0s 5ms/step - loss: 0.0799 - accuracy: 0.9824 - val_loss: 0.0907 - val_accuracy: 0.9561
Epoch 23/50
15/15 [=====] - 0s 5ms/step - loss: 0.0774 - accuracy: 0.9824 - val_loss: 0.0884 - val_accuracy: 0.9561
Epoch 24/50
15/15 [=====] - 0s 5ms/step - loss: 0.0744 - accuracy: 0.9824 - val_loss: 0.0877 - val_accuracy: 0.9561
Epoch 25/50
15/15 [=====] - 0s 5ms/step - loss: 0.0720 - accuracy: 0.9824 - val_loss: 0.0878 - val_accuracy: 0.9561
Epoch 26/50
15/15 [=====] - 0s 4ms/step - loss: 0.0696 - accuracy: 0.9824 - val_loss: 0.0870 - val_accuracy: 0.9561
Epoch 27/50
15/15 [=====] - 0s 5ms/step - loss: 0.0671 - accuracy: 0.9824 - val_loss: 0.0862 - val_accuracy: 0.9561

```

Fig 12: ANN result



### 3.1.4 Decision Tree Algorithm

```

+ Code + Text
Connect
import pandas as pd
from sklearn.datasets import load_breast_cancer

[ ] data = load_breast_cancer()
dataset = pd.DataFrame(data=data['data'], columns=data['feature_names'])
dataset

   mean radius  mean texture  mean perimeter  mean area  mean smoothness  mean compactness  mean concavity  mean concave points  mean symmetry  mean fractal dimension  ...  worst radius  worst texture  worst perimeter  worst area  worst smoothness  worst compactness  worst concavity  worst concave points  worst symmetry  worst fractal dimension
0    17.99    10.38    122.80    1001.0    0.11840    0.27760    0.30010    0.14710    0.2419    0.07871  ...    25.380    17.33    184.60    2019.0    0.16220    0.66560    0.7119    0.2654    0.4601    0.11890
1    20.57    17.77    132.90    1326.0    0.08474    0.07864    0.08690    0.07017    0.1812    0.05667  ...    24.990    23.41    158.80    1956.0    0.12380    0.18660    0.2416    0.1860    0.2750    0.08902
2    19.69    21.25    130.00    1203.0    0.10960    0.15990    0.19740    0.12790    0.2069    0.05999  ...    23.570    25.53    152.50    1709.0    0.14440    0.42450    0.4504    0.2430    0.3613    0.08758
3    11.42    20.38    77.58    386.1    0.14250    0.28390    0.24140    0.10520    0.2597    0.09744  ...    14.910    26.50    98.67    567.7    0.20980    0.86630    0.6869    0.2575    0.6638    0.17300
4    20.29    14.34    135.10    1297.0    0.10030    0.13280    0.19800    0.10430    0.1809    0.05883  ...    22.540    16.67    152.20    1575.0    0.13740    0.20500    0.4000    0.1625    0.2364    0.07678
...  ...  ...  ...  ...  ...  ...  ...  ...  ...  ...  ...  ...  ...  ...  ...  ...  ...  ...  ...  ...
564   21.56   22.39   142.00  1479.0   0.11100   0.11590   0.24390   0.13890   0.1726   0.05623  ...   25.450   26.40   166.10  2027.0   0.14100   0.21130   0.4107   0.2216   0.2060   0.07115
565   20.13   28.25   131.20  1261.0   0.09780   0.10340   0.14400   0.09791   0.1752   0.05533  ...   23.690   38.25   155.00  1731.0   0.11660   0.19220   0.3215   0.1628   0.2572   0.06637
566   16.60   28.08   108.30   858.1   0.08455   0.10230   0.09251   0.05302   0.1590   0.05648  ...   18.980   34.12   126.70  1124.0   0.11390   0.30940   0.3403   0.1418   0.2218   0.07820
567   20.60   29.33   140.10  1265.0   0.11780   0.27700   0.35140   0.15200   0.2397   0.07016  ...   25.740   39.42   184.60  1821.0   0.16500   0.86810   0.9387   0.2650   0.4087   0.12400
568    7.76   24.54    47.92   181.0   0.05263   0.04362   0.00000   0.00000   0.1587   0.05884  ...    9.456   30.37    59.16   288.6   0.08996   0.06444   0.0000   0.0000   0.2871   0.07039

569 rows x 30 columns

[ ] from sklearn.model_selection import train_test_split
X = dataset.copy()
y = data['target']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33)

[ ] from sklearn.tree import DecisionTreeClassifier
clf = DecisionTreeClassifier(ccp_alpha=0.01)
clf = clf.fit(X_train, y_train)

[ ] clf.get_params()

```

+ Code + Text

```
[ ] from sklearn.metrics import accuracy_score
accuracy_score(y_test, predictions)
```

```
0.898936170212766
```

```
[ ] from sklearn.metrics import confusion_matrix
confusion_matrix(y_test, predictions, labels=[0,1])
```

```
array([[ 58,   3],
       [ 16, 111]])
```

```
[ ] from sklearn.metrics import precision_score
precision_score(y_test, predictions)
```

```
0.9736842105263158
```

```
[ ] from sklearn.metrics import recall_score
recall_score(y_test, predictions)
```

```
0.8740157480314961
```

```
[ ] from sklearn.metrics import classification_report
print(classification_report(y_test, predictions, target_names=['malignant', 'benign']))
```

	precision	recall	f1-score	support
malignant	0.78	0.95	0.86	61
benign	0.97	0.87	0.92	127
accuracy			0.90	188
macro avg	0.88	0.91	0.89	188
weighted avg	0.91	0.90	0.90	188

```
▶ feature_names = X.columns
feature_names
```

```
☞ Index(['mean radius', 'mean texture', 'mean perimeter', 'mean area',
        'mean smoothness', 'mean compactness', 'mean concavity',
        'mean concave points', 'mean symmetry', 'mean fractal dimension',
        'radius error', 'texture error', 'perimeter error', 'area error',
        'smoothness error', 'compactness error', 'concavity error',
        'concave points error', 'symmetry error', 'fractal dimension error',
        'worst radius', 'worst texture', 'worst perimeter', 'worst area',
        'worst smoothness', 'worst compactness', 'worst concavity',
        'worst concave points', 'worst symmetry', 'worst fractal dimension'],
        dtype='object')
```



+ Code + Text

```
[ ] {'ccp_alpha': 0.01,
     'class_weight': None,
     'criterion': 'gini',
     'max_depth': None,
     'max_features': None,
     'max_leaf_nodes': None,
     'min_impurity_decrease': 0.0,
     'min_samples_leaf': 1,
     'min_samples_split': 2,
     'min_weight_fraction_leaf': 0.0,
     'random_state': None,
     'splitter': 'best'}
```

```
[ ] predictions = clf.predict(X_test)
predictions
```

```
array([[0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1,
        0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0,
        0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1,
        1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1, 1,
        0, 1, 0, 0, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1,
        1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0,
        0, 1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 0,
        0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0,
        0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1])
```

```
▶ clf.predict_proba(X_test)
```

```
[ ] [0.50475202, 0.01520710],
     [0.01415094, 0.98584906],
     [0.01415094, 0.98584906],
     [0.25      , 0.75      ],
     [0.01415094, 0.98584906],
     [0.01415094, 0.98584906],
     [0.01415094, 0.98584906],
     [0.01415094, 0.98584906],
     [0.01415094, 0.98584906],
     [0.01415094, 0.98584906],
     [0.98473282, 0.01526718],
     [0.01415094, 0.98584906],
     [0.98473282, 0.01526718],
     [0.01415094, 0.98584906],
     [0.01415094, 0.98584906],
     [0.01415094, 0.98584906],
     [0.01415094, 0.98584906],
     [0.98473282, 0.01526718],
     [0.01415094, 0.98584906],
     [0.01415094, 0.98584906],
     [0.01415094, 0.98584906],
     [0.01415094, 0.98584906],
     [0.01415094, 0.98584906],
     [0.98473282, 0.01526718],
     [0.          , 1.          ],
     [0.01415094, 0.98584906],
     [0.01415094, 0.98584906],
     [0.98473282, 0.01526718],
     [0.98473282, 0.01526718],
     [0.01415094, 0.98584906],
     [0.98473282, 0.01526718]
```

+ Code + Text

```
[ ] clf.feature_importances_
```

```
array([[0.          , 0.06052835, 0.          , 0.          , 0.          ,  
       0.          , 0.          , 0.          , 0.          , 0.          ,  
       0.          , 0.          , 0.          , 0.          , 0.          ,  
       0.          , 0.          , 0.          , 0.          , 0.          ,  
       0.          , 0.          , 0.80616997, 0.          , 0.          ,  
       0.          , 0.04307367, 0.01507494, 0.07515306, 0.          ]])
```

```
▶ feature_importance = pd.DataFrame(clf.feature_importances_, index = feature_names).sort_values(0, ascending=False)  
feature_importance
```

	0
worst perimeter	0.806170
worst symmetry	0.075153
mean texture	0.060528
worst concavity	0.043074
worst concave points	0.015075
mean radius	0.000000
worst compactness	0.000000
worst smoothness	0.000000
worst area	0.000000
worst texture	0.000000
worst radius	0.000000
fractal dimension error	0.000000
symmetry error	0.000000
concave points error	0.000000
concavity error	0.000000
compactness error	0.000000
smoothness error	0.000000
area error	0.000000
perimeter error	0.000000
texture error	0.000000
radius error	0.000000
mean fractal dimension	0.000000
mean symmetry	0.000000
mean concave points	0.000000
mean concavity	0.000000

+ Code + Text

```
[ ] clf.feature_importances_
```

```
array([[0.          , 0.06052835, 0.          , 0.          , 0.          ,  
       0.          , 0.          , 0.          , 0.          , 0.          ,  
       0.          , 0.          , 0.          , 0.          , 0.          ,  
       0.          , 0.          , 0.80616997, 0.          , 0.          ,  
       0.          , 0.04307367, 0.01507494, 0.07515306, 0.          ]])
```

```
feature_importance = pd.DataFrame(clf.feature_importances_, index = feature_names).sort_values(0, ascending=False)  
feature_importance
```

	0
worst perimeter	0.806170
worst symmetry	0.075153
mean texture	0.060528
worst concavity	0.043074
worst concave points	0.015075
mean radius	0.000000
worst compactness	0.000000
worst smoothness	0.000000
worst area	0.000000
worst texture	0.000000
worst radius	0.000000
fractal dimension error	0.000000
symmetry error	0.000000
concave points error	0.000000
concavity error	0.000000
compactness error	0.000000
smoothness error	0.000000
area error	0.000000
perimeter error	0.000000
texture error	0.000000
radius error	0.000000
mean fractal dimension	0.000000
mean symmetry	0.000000
mean concave points	0.000000
mean concavity	0.000000

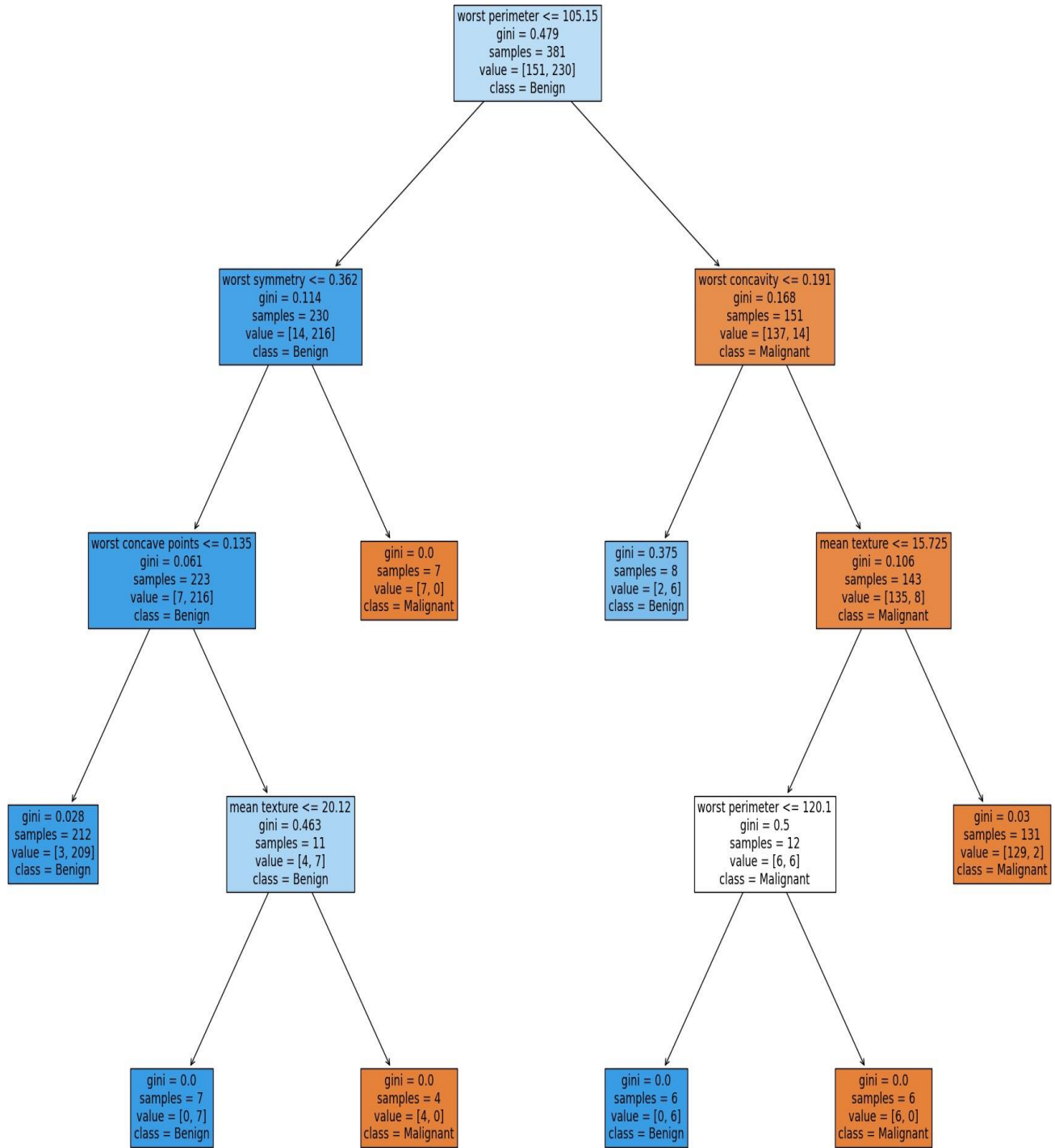


Fig 13: Decision Tree result

+ Code + Text

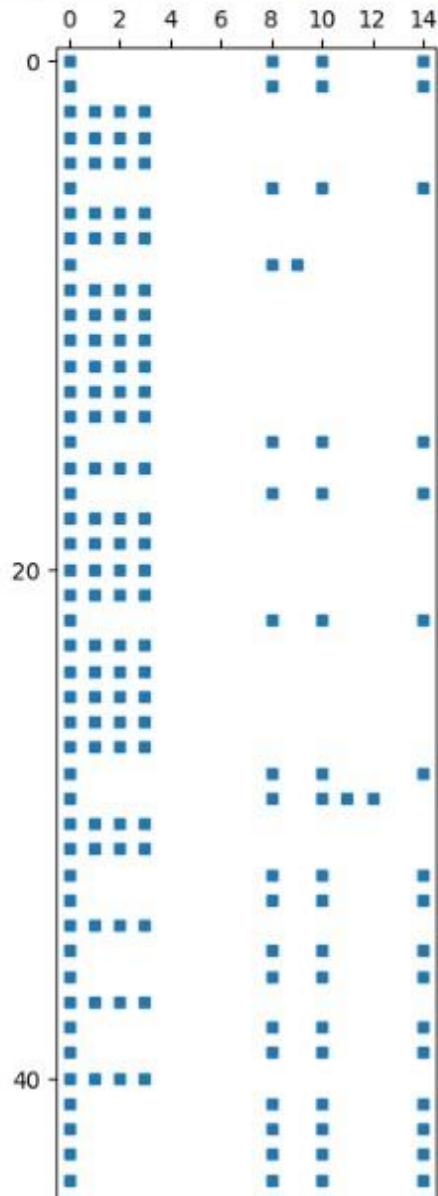
```
[ ] clf.decision_path(X_test)
```

```
<188x15 sparse matrix of type '<class 'numpy.int64''>'  
with 754 stored elements in Compressed Sparse Row format>
```

```
[ ] sparse = clf.decision_path(X_test).toarray()[:101]
```

```
▶ plt.figure(figsize=(20, 20))  
plt.spy(sparse, markersize=5)
```

```
↳ <matplotlib.lines.Line2D at 0x7fe3b61f47c0>
```



## CHAPTER 7

### 7.1 Conclusion

On the WBCD Wisconsin Breast Cancer Diagnostic dataset, we applied four primary algorithms: Naive Bayes, Random Forests, Decision Tree, and artificial neural networks to calculate, compare, and evaluate various results according to confusion accuracy, sensitivity, and precision in order to determine the best machine learning algorithm that is accurate, dependable, and finds the higher accuracy. All algorithms were created using the Python scikit-learn library and the Anaconda environment. Following a precise comparison of our models, we discovered that Nave Bayes performed better than all other algorithms, achieving higher efficiency, precision, and F-1 score of 97.4%, 98%, and 97% respectively. Finally, Nave Bayes has proven to be effective in predicting and diagnosing Breast Cancer and achieves the greatest efficiency in terms of precision and accuracy. The fact that all of the results are restricted to the WBCD database should be noted as a limitation of our work. As a result, it is important to consider applying the same algorithms and techniques to other databases in future work to validate the results obtained using this database. In addition, we intend to use our and additional machine learning algorithms with additional parameters on bigger sets of data with numerous disease classes to obtain.

Model	Dataset	Algorithm used	Accuracy	Best accuracy
Model Proposed	The Wisconsin Breast Cancer (Diagnostic) Data Set	Random Forest	95.61%	<b>Naïve Bayes</b>  <b>97.4%</b>
		Naïve Bayes	97.4%	
		ANN	97.36%	
		Decision Tree	89.89%	

TABLE 3. Comparison of different algorithms

### **7.1.1 Limitation**

**Data accuracy and bias:** For the purpose of making precise predictions, ML models largely rely on reliable, unbiased, and representative data. Inaccuracies or skewed outcomes could result from the frequent incompleteness, inconsistency, or bias of healthcare data.

**Transparency and comprehensibility:** Deep learning neural networks, which are one type of ML model, are frequently referred to as "black boxes" because they can be difficult to interpret and comprehend. This absence of interpretability might be a serious drawback in complex healthcare settings where professionals and patients must understand the logic underlying the model's predictions.

**generalising about various demographics:** Machine learning algorithms models that were developed using data from a single population might not translate well to data from other populations or demographics. This might exacerbate already-existing healthcare inequities by creating discrepancies and resulting in unequal performance across diverse populations.

**Privacy and ethical issues:** ML models rely on enormous volumes of patient data, which raises questions regarding patient confidentiality, safety of information, and the moral use of private medical data. There are many difficulties in protecting patient privacy while using data for ML.

**minimal human supervision:** In healthcare, machine learning (ML) models should support, not replace, human expertise. Decisions may be made incorrectly or may even be detrimental if there is an excessive reliance on machine learning algorithms without the proper oversight of people and clinical validation.

**Lack of ability to quickly adapt:** Since ML models are frequently created using historical data, they might not be able to quickly adapt to changing or novel healthcare scenarios. Models must be updated and modified frequently to be accurate and useful in dynamic healthcare situations.

## **REFERENCES**

- [1] Osareh, A., & Shadgar, B. (2010, April). Machine learning techniques to diagnose breast cancer. In 2010 5th international symposium on health informatics and bioinformatics (pp. 114-120). IEEE.
- [2] Yue, W., Wang, Z., Chen, H., Payne, A., & Liu, X. (2018). Machine learning with applications in breast cancer diagnosis and prognosis. *Designs*, 2(2), 13.
- [3] Fatima, N., Liu, L., Hong, S., & Ahmed, H. (2020). Prediction of breast cancer, comparative review of machine learning techniques, and their analysis. *IEEE Access*, 8, 150360-150376.
- [4] Vaka, A. R., Soni, B., & Reddy, S. (2020). Breast cancer detection by leveraging Machine Learning. *Ict Express*, 6(4), 320-324.
- [5] Cowin, P., Rowlands, T. M., & Hatsell, S. J. (2005). Cadherins and catenins in breast cancer. *Current opinion in cell biology*, 17(5), 499-508.
- [6] Sun, Y. S., Zhao, Z., Yang, Z. N., Xu, F., Lu, H. J., Zhu, Z. Y., ... & Zhu, H. P. (2017). Risk factors and preventions of breast cancer. *International journal of biological sciences*, 13(11), 1387.
- [7] Waks, A. G., & Winer, E. P. (2019). Breast cancer treatment: a review. *Jama*, 321(3), 288-300.
- [8] Hortobagyi, G. N. (1998). Treatment of breast cancer. *New England Journal of Medicine*, 339(14), 974-984.
- [9] Moulder, S., & Hortobagyi, G. N. (2008). Advances in the treatment of breast cancer. *Clinical Pharmacology & Therapeutics*, 83(1), 26-36.
- [10] Tong, C. W., Wu, M., Cho, W. C., & To, K. K. (2018). Recent advances in the treatment of breast cancer. *Frontiers in oncology*, 8, 227.
- [11] Sharma, G. N., Dave, R., Sanadya, J., Sharma, P., & Sharma, K. (2010). Various types and management of breast cancer: an overview. *Journal of advanced pharmaceutical technology & research*, 1(2), 109.
- [12] Assiri, A. S., Nazir, S., & Velastin, S. A. (2020). Breast tumor classification using an ensemble machine learning method. *Journal of Imaging*, 6(6), 39.
- [13] Heim, E., Valach, L., & Schaffner, L. (1997). Coping and psychosocial adaptation: Longitudinal effects over time and stages in breast cancer. *Psychosomatic medicine*, 59(4), 408-418.
- [14] Cui, Y., Whiteman, M. K., Flaws, J. A., Langenberg, P., Tkaczuk, K. H., & Bush, T. L. (2002). Body mass and stage of breast cancer at diagnosis. *International journal of cancer*, 98(2),



279-283.

- [15] Khanzode, S. S., Muddeshwar, M. G., Khanzode, S. D., & Dakhale, G. N. (2004). Antioxidant enzymes and lipid peroxidation in different stages of breast cancer. *Free radical research*, 38(1), 81-85.
- [16] Levine, M. N., Guyatt, G. H., Gent, M., De Pauw, S., Goodyear, M. D., Hryniuk, W. M., ... & Bramwell, V. H. (1988). Quality of life in stage II breast cancer: an instrument for clinical trials. *Journal of Clinical Oncology*, 6(12), 1798-1810.
- [17] Maughan, K. L., Lutterbie, M. A., & Ham, P. (2010). Treatment of breast cancer. *American family physician*, 81(11), 1339-1346.
- [18] Provenzano, E., Ulaner, G. A., & Chin, S. F. (2018). Molecular classification of breast cancer. *PET clinics*, 13(3), 325-338.
- [19] Gruver, A. M., Portier, B. P., & Tubbs, R. R. (2011). Molecular pathology of breast cancer: the journey from traditional practice toward embracing the complexity of a molecular classification. *Archives of pathology & laboratory medicine*, 135(5), 544-557.
- [20] Tamimi, R. M., Colditz, G. A., Hazra, A., Baer, H. J., Hankinson, S. E., Rosner, B., ... & Collins, L. C. (2012). Traditional breast cancer risk factors in relation to molecular subtypes of breast cancer. *Breast cancer research and treatment*, 131, 159-167.
- [21] Sharma, N., & Purkayastha, A. (2017). Factors affecting quality of life in breast cancer patients: a descriptive and cross-sectional study with review of literature. *Journal of mid-life health*, 8(2), 75.
- [22] McGuire, A., Brown, J. A., Malone, C., McLaughlin, R., & Kerin, M. J. (2015). Effects of age on the detection and management of breast cancer. *Cancers*, 7(2), 908-929.
- [23] Smith, S. G., Sestak, I., Forster, A., Partridge, A., Side, L., Wolf, M. S., ... & Cuzick, J. (2016). Factors affecting uptake and adherence to breast cancer chemoprevention: a systematic review and meta-analysis. *Annals of Oncology*, 27(4), 575-590.
- [24] Northouse, L. L., Dorris, G., & Charron-Moore, C. (1995). Factors affecting couples' adjustment to recurrent breast cancer. *Social Science & Medicine*, 41(1), 69-76.
- [25] Jung, S. Y., Shin, K. H., Kim, M., Chung, S. H., Lee, S., Kang, H. S., ... & Ro, J. (2014). Treatment factors affecting breast cancer-related lymphedema after systemic chemotherapy and radiotherapy in stage II/III breast cancer patients. *Breast cancer research and treatment*, 148, 91-98.
- [26] He, Z., Chen, Z., Tan, M., Elingarami, S., Liu, Y., Li, T., ... & Li, W. (2020). A review on methods for diagnosis of breast cancer cells and tissues. *Cell proliferation*, 53(7), e12822.

- [27] Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70, 245-317.
- [28] Singh, A., Thakur, N., & Sharma, A. (2016, March). A review of supervised machine learning algorithms. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 1310-1315). Ieee.
- [29] Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020). Supervised machine learning: a brief primer. *Behavior Therapy*, 51(5), 675-687.
- [30] Qayyum, A., Qadir, J., Bilal, M., & Al-Fuqaha, A. (2020). Secure and robust machine learning for healthcare: A survey. *IEEE Reviews in Biomedical Engineering*, 14, 156-180.
- [31] Bhardwaj, R., Nambiar, A. R., & Dutta, D. (2017, July). A study of machine learning in healthcare. In *2017 IEEE 41st annual computer software and applications conference (COMPSAC)* (Vol. 2, pp. 236-241). IEEE.
- [32] Wiens, J., & Shenoy, E. S. (2018). Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. *Clinical Infectious Diseases*, 66(1), 149-153.
- [33] Dua, S., Acharya, U. R., & Dua, P. (Eds.). (2014). *Machine learning in healthcare informatics* (Vol. 56). Berlin: Springer.
- [34] Gayathri, B. M., Sumathi, C. P., & Santhanam, T. (2013). Breast cancer diagnosis using machine learning algorithms-a survey. *International Journal of Distributed and Parallel Systems*, 4(3), 105.
- [35] Webb, G. I., Keogh, E., & Miikkulainen, R. (2010). Naïve Bayes. *Encyclopedia of machine learning*, 15, 713-714.
- [36] Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* (Vol. 3, No. 22, pp. 41-46).
- [37] Kaur, P., Kumar, R., & Kumar, M. (2019). A healthcare monitoring system using random forest and internet of things (IoT). *Multimedia Tools and Applications*, 78, 19905-19916.
- [38] Khalilia, M., Chakraborty, S., & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC medical informatics and decision making*, 11, 1-13.
- [39] Sibbritt, D., & Gibberd, R. (2004). The effective use of a summary table and decision tree methodology to analyze very large healthcare datasets. *Health care management science*, 7, 163-171.
- [40] Soleimani, F., Mohammadi, P., & Hakimi, P. (2012). Application of decision tree algorithm for data mining in healthcare operations: a case study. *Int J Comput Appl*, 52(6), 21-26

- [41] Ray, S. (2019, February). A quick review of machine learning algorithms. In 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon) (pp. 35-39). IEEE.
- [42] Mohammed, M., Khan, M. B., & Bashier, E. B. M. (2016). Machine learning: algorithms and applications. Crc Press.
- [43] Prabadevi, B., Deepa, N., Krithika, L. B., & Vinod, V. (2020, February). Analysis of machine learning algorithms on cancer dataset. In 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) (pp. 1-10). IEEE.
- [44] Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science*, 83, 1064-1069.
- [45] Alarabeyyat, A., & Alhanahnah, M. (2016, August). Breast cancer detection using k-nearest neighbor machine learning algorithm. In *2016 9th International Conference on Developments in eSystems Engineering (DeSE)* (pp. 35-39). IEEE.
- [46] Urgiriye, A., & Bhartiya, R. (2020). Review of Machine Learning Algorithm on Cancer Data Set. *International Journal of Scientific Research & Engineering Trends*, 6(6), 3259-3267.
- [47] Kumar, B. S., Daniya, T., & Ajayan, J. (2020). Breast cancer prediction using machine learning algorithms. *International Journal of Advanced Science and Technology*, 29(3).
- [48] Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, 2, 117693510600200030.
- [49] Tan, A. C., & Gilbert, D. (2003). Ensemble machine learning on gene expression data for cancer classification.
- [50] Wu, J., & Hicks, C. (2021). Breast cancer type classification using machine learning. *Journal of personalized medicine*, 11(2), 61.

# Behaviour Analysis Using Machine Learning Algorithms In Health Care Sector

Anukriti Yadav

Department of Biotechnology  
Engineering, Delhi Technological  
University, Shahbad Daultapur, Main  
Bawana Road, Delhi 110042, India

anukritiyadav454@gmail.com

Deepak Kumar

Department of Biotechnology  
Engineering, Delhi Technological  
University, Shahbad Daultapur, Main  
Bawana Road, Delhi 110042, India

deepak545@gmail.com

Yasha Hasija\*

Department of Biotechnology  
Engineering, Delhi Technological  
University, Shahbad Daultapur, Main  
Bawana Road, Delhi 110042, India

yashahasija06@gmail.com

\*Corresponding Author

traditional healthcare [1]. Through the development of smart healthcare, it is possible to analyze patient behavior more

**Abstract** - A behavioral analytics approach uses big data analytics in combination with machine learning (ML) to identify patterns, trends, aberrations, and other useful insights. The behavior of an individual can be analyzed by expressions, postures, and activity levels. Using ML algorithms could revolutionize the way clinicians make decisions in health care sector. Studies of human behavior have been conducted in a range of scientific disciplines (e.g sociology, psychology, computer science). ML algorithms have the potential to transform the way doctors and instructors make choices. This methodology has been slow to be adopted by behavior analysis experts to maximize its application to practical issues and to aid them in learning more about human behavior. ML algorithms are dominating the healthcare industry. Recent researches have indicated that these techniques can be used to anticipate disease based on health data. Our study examines several machine learning algorithms used in early disease detection and identifies key trends in their performance. The analysis suggests that human behavior may play a role in a variety of conditions, including diabetes, cancer, heart disease, autism, mental illness, Alzheimer's, and others. A number of daily habits are associated with this behavior, including food, respiration rate, blood pressure, voice output, social abnormalities, insomnia, and so on. A few examples of ML applications integrated into healthcare services are naive bayes (NB), support vector machines (SVM), random forest (RF), and convolutional neural networks (CNN). In a variety of cancer classification applications, these models are proved to be highly efficient in diagnosing various cancer types. This review includes a number of research investigations that employ ML to analyze behavioral data. As we gain further insights into the factors influencing organisms' behavior, we are able to create computational models which allow disease prediction and management to become more accurate.

**Keywords**— Behavioral analytics, Machine learning, Algorithm, accuracy, healthcare services

## I. INTRODUCTION

Patients with neurological diseases, head traumas, and mental illnesses benefit greatly from behavioral analysis in the health sector. It is helpful to determine the root cause of a disease by analyzing the patient's behavior. There are many challenges associated with patient behavioral analysis in

easily. Using ML for human behavior recognition has become a new topic of analysis due to the issues relating to potency and accuracy of conventional artificial feature-extraction behavior identification. **Figure 1** shows different type of machine learning algorithms that are used behavioral analysis

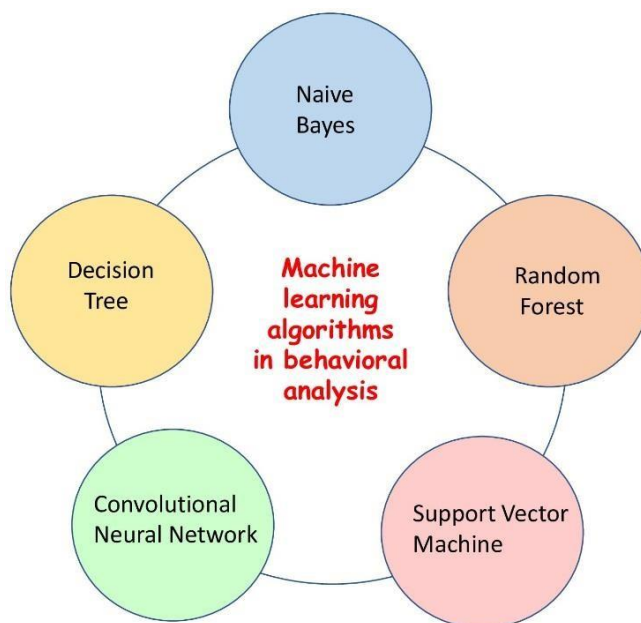


Fig. 1. Different type of machine learning algorithms in behavioral analysis

Research investigators used ML tools to detect behavioral patterns for different patient groups based on the experimental case study they conducted such as pattern analysis for anxiety and depression level [2], assessing patients who have been diagnosed with autism and those who have not [3].

In spite of extensive instrumental as well as scoring noise, ML is capable of detecting interactions that are complex, high-dimensional and, non-linear that may notify prognosis [4]. **Figure 2** depicts how behavioral analysis is carried out using machine learning algorithms. Many biology and

behavioral research laboratories, however, find it strenuous to implement these advanced analyses, which may explain why they have not yet been widely adopted. In this article, we

have reviewed about various machine learning techniques (like SVM, CNN, Decision tree, Random forest, Naive bayes) to study behavior analysis in healthcare sector.

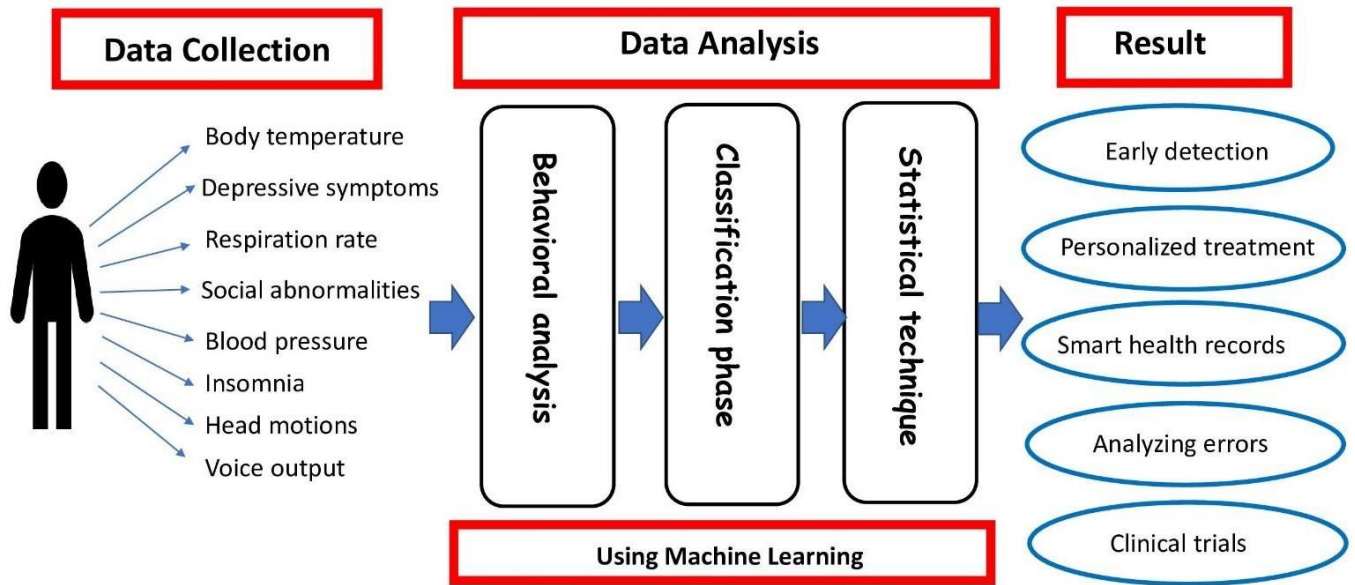


Fig. 2. Use of machine learning algorithm in behavioral analysis

in accuracy, precision, error and recall, measuring 0.87, 0.84, 0.13 and 0.79, respectively [7].

## II. ALGORITHMS IN DETECTING BEHAVIOURAL ANALYSIS

### A. Naive bayes

The Naive Bayes algorithm is a statistical classifier which predicts the likelihood that a given tuple belongs to a given class in accordance with Bayesian analysis theorem [5]. A naive bayes method is used in healthcare to assess a patient's behavior, such as their mental health, using a multiclass classification approach and probabilistic algorithm. Radwan Qasrawi et al used an empirical Bayesian approach to investigate the factors associated with depression and anxiety in school-aged children. A total of 3984 West Bank students from 5th to 9th grades, age ranging 10-15 years, were studied in refugee and community schools. During the academic year 2013-2014 data was assembled using the health behaviors of school children examination to identify risk variables related to student mental health symptoms. ML was then used to analyze the data. An analysis of 5 ML algorithms, including RF, neural networks, decision trees, SVMs, and NB were used and it was concluded that NB had the best accuracy in predicting depressive disorders (87.1) and anxiety (72.7) [6].

As a tool for predicting how a person's body will behave if she/he contracts Covid-19, Rabie et al. developed the Covid-19 Prudential Expectation Strategy (CPES). In this method there are three steps: Outlier Rejection Phase (ORP), Feature Selection Phase (FSP), and Classification Phase (CP). CPES makes use of a Statistical Naive Bayes (SNB) classifier, CP, to categorize people according to their body's reaction to Covid-19 infection. There were 2215 persons that filled out the form in total. Compared to current classification algorithms, Prudential Expectation Strategy performs better

### B. SVM (Support Vector Machine)

A support vector network is the most prevailing supervised learning model which uses deep learning algorithms to map data into a high-dimensional feature space for classifying and predicting data from two groups. SVM increases effectiveness and makes healthcare more convenient and personalized for patients in a healthcare institution. The algorithm is utilized in several healthcare practices to anticipate if a patient has a particular health issue. Its high classification accuracy, sensitivity, and specificity make it an excellent option for diagnosing diseases like heart disease, stress, and influenza.

Athira et al. [8] used SVM technique for development of multi-parameter for monitoring patients health. The researchers developed a multi parameter system based on IOT which had four parameters including heart rate, hotness and coldness, pulse rate, and oxygenation are observed using analogous sensors and in case of emergency an email is conveyed to patient's guardian. They achieved classification accuracy to 95 percent of the mpm system (Multi-Parameter Patient Monitor) by improving the algorithm of SVM.

Using behavioral risk as a predictor of cervical cancer, Degrimenci [9] investigated KNN, Random Forest, and SVM algorithms for their potential role in cervical cancer

prediction. The information utilized in this study was collected from the UCI Machine Learning Repository (a library of data on cervical cancer behavior risk). A total of 72 Indonesians were recruited to provide the samples. Results showed that 21 were in danger (positive) and the remaining 51 were not in danger (negative). The facts that were collected related to cervix cancer behavior which includes 19 characteristics, such as diet, hygiene practices, sexual risk, emotion, etc. Out of all the approaches that were tested, the SVM technique had the most appreciative accuracy (91.67% with sigmoid-SVM).

### C. Random Forest

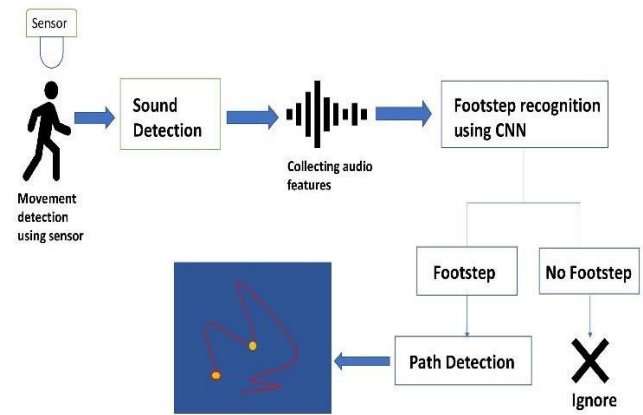
Random Forest (RF) algorithm is an ensemble classifier that uses a variety of decision tree models to improve presaging accuracy. Based on the set of training data, it generates several classification trees that are each trained using bootstrap sampling [10]. Random forests are effective at estimating variable significance with neural networks. Their ability to handle huge data sets with hundreds of variables makes them an excellent approach for dealing with missing data. Probabilities of contracting a disease can be forecast using a random forest model based on past diagnoses. The concept may be useful for managing risks, communicating customized health information, and assisting with healthcare decisions [11].

Khalilia et al. [12] trained random forest classifiers for predicting diseases using HCUP (Healthcare Cost and Utilization Project) data from National Inpatient Sample (NIS). To predict the threat of 8 chronic diseases, they evaluated the effectiveness of SVM, bagging, boosting, and RF ensemble learning hinge on the Area Under the ROC Curve (AUC). Additionally, RF calculates the significance of each and every variable in the classification procedures, which helped them overcome the class imbalance issue. An average AUC of 88.79% was calculated for eight disease types using the HCUP data set and found to be promising in disease prediction.

Kazuya et al. [13] conducted a study in Japan to show a link between stroke-related search behavior and stroke-related mortality. The regression analysis used a number of characteristics as predictors, including sex, lifespan, hospitalizations, progress, strokes, etc. They identified 9476 abstracts from Japanese literature relating to stroke symptoms and signs and a score of 89.94% was achieved using age-adjusted mortality from stroke in the RF analysis. It revealed that the query with high relevance score was stroke, and that it was associated with Japan's age-adjusted mortality rate.

### D. CNN (CONVOLUTED NEURAL NETWORK)

CNN is a kind of artificial neural network which uses mathematical operations that it is typically employed in image recognition, segmentation, classification and has one or more convolutional layers, and other correlated tasks. As CNN categorizes hundreds of pictures each minute, it could



be useful because new photos could be categorized instantly. Photos are sent for grading to physicians when a patient comes in for screening, but they are not appropriately rated. A trained CNN is capable of making a rapid diagnosis and responding to a patient immediately. In similar fashion, the network produced these results using only one image per eye.

Fig. 3. An illustration of how data filtering works

Oliveira et al. [14] developed a CNN method to detect the wandering movements of Alzheimer's patients based on data gathered from non-intrusive sensors around the house (**Figure 3**). 220 paths were generated in the dataset. This data was identified by CNN using visual features (such as loops or random movements). The data was compared with 60 min and 30 min datasets and the 30 min datasets had a precision difference of 55.57%, a recall difference of 20% and a F1 score difference of 17.86%.

Pratt et al. [15] carried out a CNN based model for the classification of 5 classes of diabetic retinopathy disease. They took 80000 images from the Kaggle through which state-of-the-art DR stage classification technique was put together by utilizing complex DR characteristics such as transude on the retina, MAS, and HEMS, they employed a CNN architecture with expansion of data to designate 5 degrees of diabetic retinopathy severity.

### D. Decision Tree

Decision tree is a type of supervised machine learning technique which predicts and processes data using classification and regression analogies based on real life [5]. Health systems can use decision trees to determine the initial course of treatment for behavior problems and to implement empirically validated treatment procedures [16].

In a study by Cohen et al. [17] behavioral profiles of children with Autism Spectrum Disorder are used to guide treatment decisions through CART decision trees. They compared the PDDBI with the ADOS-2 to determine its criterion-related validity. A total of 110 candidates were selected between 1.5



and 6.9 years age and grouped into 2 behavioral aspects: Receptive/Expressive Social Communication Abilities (REXSCAs) and Approach Withdrawal Problems (AWPs). Based on T-scores, various cut-off scores was evaluated in this study for domains such as sensory behaviors, traditions, repetition of language, offensive behaviour, and ability to express.

Using a decision tree algorithm, Batterham et al. analysed depression outcomes. This study showed that environmental factors are associated with various social factors that are

related to depression. After 4 weeks of follow-up, patients were allowed to begin taking new antidepressants; 25% of placebo patients and 7% of zuranolone patients received new antidepressants in the phase 2 study. Decision trees were discovered to have higher susceptibility and selectivity than logistic regressions when analogous predictors were used [18].

TABLE I. RECENT STUDIES CONDUCTED ON BEHAVIOR ANALYSIS WITH THE HELP OF ML FOR THE TREATMENT OF VARIOUS DISEASE

S.No	Technique	Paper	Worked on	Accuracy	Year
1	Naïve Bayes	Radwan Qasrawi et al. [6]	Depression and anxiety	87.1%(depression) 72.7%(anxiety)	2022
		Rabie et al. [7]	Covid	87%	2022
2	SVM	Degirmenci [9]	Cervical cancer	91.67%	2022
		Athira et al. [8]	Heart rate, temperature, respiration rate, oxygen and saturation	95%	2020
3	Random Forest	Khalilia et al. [12]	8 chronic diseases	88.79%	2011
		Kazuya et al. [13]	Stroke	89.94%	2022
4	CNN	Pratt et al. [15]	Daibetic retinopathy disease	-	2016
		Oliveira et al. [14]	Alzheimer disease	82.65%	2022
5	Decision Tree	Cohen et al. [17]	Autisum spectrum disorder	-	2019
		Batterham et al. [18]	Depression	-	2009

### III. CONCLUSION

As machine learning becomes more prevalent, it is being used for diagnosing diseases in multiple industries. A number of scientists have discussed the benefit of machine learning-based disease diagnostics (MLBDD) in terms of time and cost efficiency. Traditionally, diagnostic techniques are labor-intensive, expensive, and require human involvement frequently. According to WHO estimates, lifestyle changes are responsible for 30% of all deaths worldwide. These deaths can be avoided by correctly identifying the risk factors that go along with them and developing behavior modification strategies. Prevention of potentially fatal consequences requires changes in health-related behavior. Life expectancy will be increased if early diagnosis, prevention assistance, and appropriate treatment are provided as soon as possible. Machine learning algorithms will be implemented to further investigate methods like sensor based feature extraction, such

as Electrocardiogram (ECG), Electroencephalogram (EEG) etc, for the diagnosis of early-stage diseases from human behavior on various platforms. The idea behind automated

patient and disease monitoring activities are to conserve time and fill in when all doctors are busy, like during an emergency. The deployment of smart technology in this industry can help save lives during pandemics, such as the COVID-19 epidemic.

### REFERENCES

- [1] F. Yao, "Deep learning analysis of human behaviour recognition based on convolutional neural network analysis," *Behav. Inf. Technol.*, vol. 0, no. 0, pp. 1–9, 2020, doi: 10.1080/0144929X.2020.1716390.
- [2] T. Richter, B. Fishbain, A. Markus, G. Richter-Levin, and H. Okon-Singer, "Using machine learning-based analysis for

behavioral differentiation between anxiety and depression.” *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, 2020, doi: 10.1038/s41598-020-72289-9.

- [3] D. P. Wall, R. Dally, R. Luyster, J. Y. Jung, and T. F. DeLuca, “Use of artificial intelligence to shorten the behavioral diagnosis of autism,” *PLoS One*, vol. 7, no. 8, 2012, doi: 10.1371/journal.pone.0043855.
- [4] B. Kanchanatawan, S. Thika, S. Sirivichayakul, A. F. Carvalho, M. Geffard, and M. Maes, “In Schizophrenia, Depression, Anxiety, and Physiosomatic Symptoms Are Strongly Related to Psychotic Symptoms and Excitation, Impairments in Episodic Memory, and Increased Production of Neurotoxic Tryptophan Catabolites: a Multivariate and Machine Learning,” *Neurotox. Res.*, vol. 33, no. 3, pp. 641–655, 2018, doi: 10.1007/s12640-018-9868-4.
- [5] M. Srividya, S. Mohanavalli, and N. Bhalaji, “Behavioral Modeling for Mental Health using Machine Learning Algorithms,” *J. Med. Syst.*, vol. 42, no. 5, 2018, doi: 10.1007/s10916-018-0934-5.
- [6] R. Qasrawi, S. P. V. Polo, D. A. Al-Halawa, S. Hallaq, and Z. Abdeen, “Assessment and Prediction of Depression and Anxiety Risk Factors in Schoolchildren: Machine Learning Techniques Performance Analysis,” *JMIR Form. Res.*, vol. 6, no. 8, pp. 1–15, 2022, doi: 10.2196/32736.
- [7] A. H. Rabie, N. A. Mansour, A. I. Saleh, and A. E. Takieldeem, “Expecting individuals’ body reaction to Covid-19 based on statistical Naïve Bayes technique,” *Pattern Recognit.*, vol. 128, no. April, 2022, doi: 10.1016/j.patcog.2022.108693.
- [8] A. Athira, T. D. Devika, K. R. Varsha, and S. S. Bose, “Design and Development of IOT Based Multi-Parameter Patient Monitoring System,” *2020 6th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2020*, pp. 862–866, 2020, doi: 10.1109/ICACCS48705.2020.9074293.
- [9] A. Degirmenci, “Performance Comparison of kNN , Random Forest and SVM in the Prediction of Cervical Cancer from Behavioral Risk,” vol. 7, no. 10, pp. 71–79, 2022.
- [10] W. Mao and F.-Y. Wang, “Cultural Modeling for Behavior Analysis and Prediction,” *Adv. Intell. Secur. Informatics*, pp. 91–102, 2012, doi: 10.1016/b978-0-12-397200-2.00008-7.
- [11] G. Biau and E. Scomet, “A random forest guided tour,” *Test*, vol. 25, no. 2, pp. 197–227, 2016, doi: 10.1007/s11749-016-0481-7.
- [12] M. Khalilia, S. Chakraborty, and M. Popescu, “Predicting disease risks from highly imbalanced data using random forest,” *BMC Med. Inform. Decis. Mak.*, vol. 11, no. 1, 2011, doi: 10.1186/1472-6947-11-51.
- [13] K. Taira and S. Fujita, “Prediction of Age-Adjusted Mortality from Stroke in Japanese Prefectures: Ecological Study Using Search Engine Queries,” *JMIR Form. Res.*, vol. 6, no. 1, 2022, doi: 10.2196/27805.
- [14] R. Oliveira, R. Feres, F. Barreto, and R. Abreu, “CNN for Elderly Wandering Prediction in Indoor Scenarios,” *SN Comput. Sci.*, vol. 3, no. 3, pp. 1–11, 2022, doi: 10.1007/s42979-022-01091-3.
- [15] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng, “Convolutional Neural Networks for Diabetic Retinopathy,” *Procedia Comput. Sci.*, vol. 90, no. July, pp. 200–205, 2016, doi: 10.1016/j.procs.2016.07.014.
- [16] V. Podgorelec, P. Kokol, B. Stiglic, and I. Rozman, “Decision trees: An overview and their use in medicine,” *J. Med. Syst.*, vol. 26, no. 5, pp. 445–463, 2002, doi: 10.1023/A:1016409317640.
- [17] I. L. Cohen and M. J. Flory, “Autism Spectrum Disorder Decision Tree Subgroups Predict Adaptive Behavior and Autism Severity Trajectories in Children with ASD,” *J. Autism Dev. Disord.*, vol. 49, no. 4, pp. 1423–1437, 2019, doi: 10.1007/s10803-018-3830-4.
- [18] P. J. Batterham, H. Christensen, and A. J. Mackinnon, “Modifiable risk factors predicting major depressive disorder at four year follow-up: A decision tree approach,” *BMC Psychiatry*, vol. 9, pp. 4–11, 2009, doi: 10.1186/1471-244X-9-75.



Dear Yasha Hasija  
Delhi Technological University

Greetings from InCACCT-2023 ...!!!

Congratulations...!!!!!!

On behalf of the InCACCT-23 Program Committee, we are delighted to inform you that the submission of Paper ID: 1058 titled "Behaviour Analysis Using Machine Learning Algorithms In Health Care Sector " has been accepted for presentation at the InCACCT-23. The conference proceedings are approved by IEEE Xplore (Conference Record Number -#57535) and Accepted papers will be submitted for inclusion into IEEE Xplore subject to meeting IEEE Xplore's scope and quality requirements.

Please complete your registration by clicking on the following Link:  
<https://forms.gle/PUveRswyoweUFCj17> on or before 7th March 2023

Activities still to be carried out at your end are as follows:

1. E-copyright Transfer (corresponding author will get separate email from IEEE containing login credential and link for signing e-copyright transfer)
2. Presentation of Paper on the day of conference (via online or offline mode).
3. Providing any other desired information timely.

If you have any query regarding registration process or face any problem in making online payment, you can Contact @ 8708951544 (Call) / 9416345948 (Whatsapp) or write us at [icacct2021@cumail.in](mailto:icacct2021@cumail.in).

Regards:  
Organizing committee  
InCACCT - 23

Sr. No. 148



## Certificate of Participation

This is to certify that **Prof./Dr./Mr./Ms. ANUKRITI YADAV**  
of **Delhi Technological University (DTU)**

*participated/presented a paper titled*

**Behaviour analysis using machine learning algorithms in healthcare sector**

*in 1st International Conference on Advancement in Computation & Computer Technologies (InCACCT- 2023)  
organized by the Department of Computer Science & Engineering, with the technical sponsor IEEE Delhi Section (IEEE  
Conference Record No.: 57535X) held on 05th – 06th May 2023 at Chandigarh University,  
Gharuan, Mohali, Punjab, India.*

**Dr. Meenu Gupta**  
Convener & Conf. Organizing Chair  
Chandigarh University, Punjab, India

**Prof. (Dr.) Rakesh Kumar**  
Convener & Conf. Organizing Chair  
AD-CSE, Chandigarh University, Punjab, India

## PAPER NAME

**breastcancer writeup.docx**

---

## WORD COUNT

**8449 Words**

## CHARACTER COUNT

**47593 Characters**

## PAGE COUNT

**46 Pages**

## FILE SIZE

**1.5MB**

## SUBMISSION DATE

**May 26, 2023 7:54 PM GMT+5:30**

## REPORT DATE

**May 26, 2023 7:55 PM GMT+5:30**

---

**● 12% Overall Similarity**

The combined total of all matches, including overlapping sources, for each database.

- 2% Internet database
- 1% Publications database
- Crossref database
- Crossref Posted Content database
- 12% Submitted Works database

**● Excluded from Similarity Report**

- Bibliographic material

DELHI TECHNOLOGICAL UNIVERSITY  
(Formerly Delhi College of Engineering)  
Bhawana Road, Delhi-110042

**Certificate**

I hereby certify that the Project Dissertation titled “**Revolutionizing Breast Cancer Management: Harnessing the Power of Machine Learning in Diagnoses and Treatment**” which is submitted by **Anukriti Yadav (2K21/MSCBIO/06)** Department of Biotechnology, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Science is recorded for the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any degree or any diploma to this university or elsewhere.

Place: Delhi

Date:

**Prof. Yasha Hasija**

**(Supervisor)**

**Professor**

Department of Biotechnology

Delhi Technological University

**Prof. Pravir Kumar**

**Head of Department**

**Dean (International Affairs)**

Department of Biotechnology

Delhi Technological University

## **Acknowledgement**

I would like to express my gratitude towards my supervisor, Prof. Yasha Hasija, for giving me the opportunity to do research and providing invaluable guidance throughout this research. Her dynamism, vision, sincerity and motivation have deeply inspired me. She has motivated to carry out the research and to present my work works as clearly as possible. It was a great privilege and honor to work and study under her guidance. I am extremely grateful for what he has offered me. Her insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

I am extremely grateful to my parents for their love, prayers, caring and sacrifices for educating and preparing me for my future. I would also like to thank to Jaishree Mam for guiding me and helping me to complete mt thesis.

A special thanks to my lab mate Deepak Kumar for moral support, tolerance and help from the beginning to the end.

I would also like the institution Delhi Technological University, Delhi for giving me the opportunities throughout the tenure of study.

Finally, my thanks go to all the people who have supported me to complete the research work directly or indirectly.

Anukriti Yadav



**DELHI TECHNOLOGICAL UNIVERSITY**  
(Formerly Delhi College of Engineering)  
Bhawana Road, Delhi-110042

**Certificate**

I hereby certify that the Project Dissertation titled "**Revolutionizing Breast Cancer Management: Harnessing the Power of Machine Learning in Diagnoses and Treatment**" which is submitted by **Anukriti Yadav (2K21/MSCBIO/06)** Department of Biotechnology, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Science is recorded for the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any degree or any diploma to this university or elsewhere.

Place: Delhi

Date: 29.05.23

*Yasha Hasija*  
29.05.23

**Prof. Yasha Hasija**

**(Supervisor)**

**Professor**

Department of Biotechnology

Delhi Technological University

*Pravir Kumar*  
30/05/2023

**Prof. Pravir Kumar**

**Head of Department**

**Dean (International Affairs)**

Department of Biotechnology

Delhi Technological University