

**A  
Project Report  
on  
BREAST CANCER CLASSIFICATION USING MACHINE  
LEARNING**

**Submitted in partial fulfillment of the requirements  
for the award of the degree of**

**Bachelor of Technology  
in  
Information Technology**

**By  
MAYANK PATHAK  
(1709713058)  
LAKSHAY VARDHAN  
(1709713056)  
NIKHIL KUMAR  
(1709713066)**

**Under the Supervision of  
MR K. RAJKUMAR  
Professor, Department of I.T.**



**Galgotia's College of Engineering & Technology  
Greater Noida, Uttar Pradesh  
India-201306  
Affiliated to**



**Dr. A.P.J. Abdul Kalam Technical University  
Lucknow, Uttar Pradesh  
July 2021**



**GALGOTIAS COLLEGE OF ENGINEERING & TECHNOLOGY**  
**GREATER NOIDA, UTTAR PRADESH, INDIA- 201306.**

## **CERTIFICATE**

This is to certify that the project report entitled “**BREAST CANCER CLASSIFICATION USING MACHINE LEARNING**” submitted by **Mr. MAYANK PATHAK (1709713058), Mr. Lakshay Vardhan (1709713056), Mr. NIKHIL KUMAR (1709713066)** to the Galgotias College of Engineering & Technology, Greater Noida, Uttar Pradesh, Affiliated to Dr. A.P.J. Abdul Kalam Technical University Lucknow, Uttar Pradesh in partial fulfillment for the award of Degree of Bachelor of Technology in Information Technology is a bonafide record of the project work carried out by them under my supervision during the year 2020-2021.

**MR K RAJKUMAR**

**ASSISTANT PROFESSOR  
DEPT. OF IT**

**DR. SANJEEV KUMAR SINGH**

**PROFESSOR AND HEAD  
DEPT. OF IT**



**GALGOTIAS COLLEGE OF ENGINEERING &  
TECHNOLOGY**  
**GREATER NOIDA, UTTER PRADESH, INDIA- 201306.**

## **ACKNOWLEDGEMENT**

We have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. We would like to extend my sincere thanks to all of them. We are highly indebted to **MR K. RAJKUMAR** for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project. We are extremely indebted to Dr. SANJEEV KUMAR SINGH, the HOD of Department of Information Technology, GCET and Dr. Javed Miya, Project Coordinator, Department of Information Technology, GCET for his valuable suggestions and constant support throughout my project tenure. We would like to express our thanks to all faculty and Staff members of Department of Information Technology, GCET for their support in completing this project on time.

We also express gratitude towards our parents for their kind co-operation and encouragement which help me in completion of this project. Our thanks and appreciations also go to our friends in developing the project and people who have willingly helped me out with their abilities.

MAYANK PATHAK (1709713058)

LAKSHAY VARDHAN (1709713056)

NIKHIL KUMAR (1709713066)

## ABSTRACT

Breast cancer may be a heterogeneous disease, commonly known to begin as neighborhood lesion within the breast, and so spread gradually. The second leading cause of death among women is cancer. In 2020 within the US there'll be 2,79,100 new cases and 42,690 estimated deaths. Together with 48,530 new cases of non-invasive (in-situ) cancer approximately every one-fifth of the affected women will die because of this disease. This (breast) dual organ inside a woman makes significant changes in size, shape and function in conjunction with pregnancy, lactation, puberty and postpartum so that a protective pregnancy should take place before the age of 30. When the breast grow abnormally it leads to breast cancer. These cells form a lump or mass because they divide quicker than healthy cells do and accumulate.

The lymph nodes and other part of the body can also be spreaded by the cells of the breast which is dangerous. Cells can often be seen from X-Ray or felt as a lump. Great information is obtainable so analysing this huge amount of knowledge to extract the novel and usable information or knowledge is incredibly complicated and time-consuming task the best technique for this is data mining. There is class imbalance within the data. Since the probability of having the disease is much more than not having the disease. The goal of this research paper is to differentiate Malignant and Benign patients. The dataset which we have used here is Wisconsin Dataset (WBC).

**Keywords:** *Machine Learning; Breast Cancer; Benign; Malignant.*

## TABLE OF CONTENT

CONTENTS	PAGE
CERTIFICATE.....	2
ACKNOWLEDGEMENT.....	3
ABSTRACT.....	4
TABLE.....	5
LIST OF TABLES .....	7
LIST OF FIGURES .....	8
LIST OF ABBREVIATIONS .....	9
CHAPTER-1.....	10
INTRODUCTION.....	10
1.2 HOW AI WORKS? .....	12
1.3 CLASSIFICATION ALGORITHMS .....	13
1.3.1 K-Nearest Neighbors: .....	13
1.3.2 Logistic Regression: .....	14
1.3.3 Support Vector Machine: .....	14
1.3.4 Decision tree: .....	14
1.3.5 Random Forest: .....	15
1.3.6 Naive Bayes: .....	16
1.3.7 Neural Network: .....	17
CHAPTER 2 .....	20
LITERATURE REVIEW .....	20
2.1 Breast Cancer Statistic .....	20
2.2 Breast Cancer Detection and Classification: .....	21
2.3 Breast Cancer Detection in Digital Mammograms: .....	21

BRIEF LITERATURE SURVEY: .....	22
<b>CHAPTER 3.....</b>	<b>24</b>
<b>PROBLEM FORMULATION .....</b>	<b>24</b>
3.1 DATA DIFFICULTIES:.....	24
3.2 BREAST CANCER CLASSIFICATION TECHNIQUES/PROBLEMS/METHO ....	25
3.3 REPRESENTATION MODELS: .....	25
<b>CHAPTER 4.....</b>	<b>26</b>
<b>PROPOSED WORK.....</b>	<b>26</b>
4.1 IMPLEMENTATION PROCESS .....	26
4.2 CLEANING OF DATA:.....	26
4.3 EXPLORATORY DATAANALYSIS (EDA): .....	28
4.1.3. FEATURE SELECTION: .....	28
5.2 REALITY OF AI FOR MEDICAL DIAGNOSIS .....	29
5.2.1 ONCOLOGY .....	30
5.2.2 PATHOLOGY .....	30
5.2.3 DERMATOLOGY .....	31
5.2.4. GENETICS AND GENEMICS: .....	31
5.2.5 MENTAL HEALTH .....	32
5.2.6 CRITICAL CARE.....	32
5.2.7 EYECARE.....	33
5.2.8 DIABETES: .....	33
<b>CHAPTER 6.....</b>	<b>36</b>
<b>IMPLEMENTATION .....</b>	<b>36</b>
6.1 Breast Cancer Classification .....	36
6.2 Machine learning algorithms .....	37
PROGRAM.....	40
<b>CHAPTER 7 .....</b>	<b>42</b>
<b>RESULT ANALYSIS .....</b>	<b>42</b>
<b>CHAPTER 8.....</b>	<b>44</b>
<b>CONCLUSION AND FUTURE SCOPE .....</b>	<b>44</b>
8.1 CONCLUSION .....	44

8.2 FUTURE SCOPE .....	45
------------------------	----

/

## LIST OF TABLES

	<b>Table Title</b>	<b>Page</b>
1	Brief Literature Survey	7
2	Feature selection of dataset	28

## LIST OF FIGURES

<b>Figure Title</b>	<b>Page</b>
Figure 1.3.4 Decision Tree Work flow	15
Figure 1.3.5 Random forest Workflow	16
Figure 1.3.6 Naive Bayes formula	17
Figure 1.3.7 Perceptron Neural Network	18
Figure 4.2 Breast cancer classification mechanism	27
Figure 5.1 Breast cancer classification using KNN and SVM	29
Figure 5.2 Breast Cancer detection using KNN and SVM	35
Figure 6.2.1 SVM Mechanism	38
Figure 6.2.2 KNN Process flow chart	39



## **LIST OF ABBREVIATIONS**

IBC	Inflammatory Breast Cancer
KDD	Knowledge Discovery in Databases
IE	Information Extraction
SVM	Support Vector Machine
KNN	K-Nearest Neighbors
Tf-Idf	Term Frequency–Inverse Document Frequency
WSD	Word Sense Disambiguation
PoS	Part of Speech
IR	Information Retrieval
HMM	Hidden Markov Model
CRF	Conditional Random Field
MBL	Memory Based Learning
TBL	Transformation Based Learning
VSM	Vector Space Model
BOW	Bag-of-Words

# CHAPTER-1

## INTRODUCTION

### ***Description:***

*This chapter is an introduction which tells about what kind of content is present or used in this entire report and explains all the basic concepts required to accomplish the goal i.e., to Classification of Breast Cancer.*

---

Breast cancer classification divides breast cancer into categories according to different schemes criteria and serving a different purpose. The major categories are the histopathological type, the grade of the tumor, the stage of the tumor, and the expression of proteins and genes. As knowledge of cancer cell biology develops these classifications are updated.

The purpose of classification is to select the best treatment. The effectiveness of a specific treatment is demonstrated for a specific breast cancer (usually by randomized, controlled trials). That treatment may not be effective in a different breast cancer. Some breast cancers are aggressive and life-threatening, and must be treated with aggressive treatments that have major adverse effects. Other breast cancers are less aggressive and can be treated with less aggressive treatments, such as lumpectomy.

Treatment algorithms rely on breast cancer classification to define specific subgroups that are each treated according to the best evidence available. Classification aspects must be carefully tested and validated, such that confounding effects are minimized, making them either true prognostic factors, which estimate disease outcomes such as disease-free or overall survival in the absence of therapy, or true predictive factors, which estimate the likelihood of response or lack of response to a specific treatment.

Machine learning classifiers are very popular for detecting Cancer. Several research works have been done in this area. Here a classifier algorithm named “Support Vector Machine”

has been used to detect the malignancy or benignancy of the tumorous cell more accurately. Cancer arises from the transformation of normal cells into tumour cells in a multi-stage process that generally progresses from a precancerous lesion to a malignant tumour. These changes are the result of the interaction between a person's genetic factors and three categories of external agents, including:

1.1.1 physical carcinogens, such as ultraviolet and ionizing radiation;

1.1.2 chemical carcinogens, such as asbestos, components of tobacco smoke, aflatoxin (a food contaminant), and arsenic (a drinking water contaminant);

1.1.3 biological carcinogens, such as infections from certain viruses, bacteria, or parasites.

Classification of breast cancer is usually, but not always, primarily based on the histological appearance of tissue in the tumor. A variant from this approach, defined on the basis of physical exam findings, is that inflammatory breast cancer (IBC), a form of ductal carcinoma or malignant cancer in the ducts, is distinguished from other carcinomas by the inflamed appearance of the affected breast, which correlates with increased cancer aggressivity

1.1.4 Histopathology

1.1.5 Grade.

1.1.6 Stage.

Breast cancer is cancer that develops from breast tissue. Signs of breast cancer may include a lump in the breast, a change in breast shape, dimpling of the skin, fluid coming from the nipple, a newly inverted nipple, or a red or scaly patch of skin. In those with distant spread of the disease, there may be bone pain, swollen lymph nodes, shortness of breath, or yellow skin.

Risk factors for developing breast cancer include being female, obesity, a lack of physical exercise,

alcoholism, hormone replacement therapy during menopause, ionizing radiation, an early age at first menstruation, having children late in life or not at all, older age, having a prior history of breast cancer, and a family history of breast cancer. About 5–10% of cases are the result of a genetic predisposition inherited from a person's parents, including BRCA1 and BRCA2 among others. Breast cancer most commonly develops in cells from the lining of milk ducts and the lobules that supply these ducts with milk. Cancers developing from the ducts are known as ductal carcinomas, while those developing from lobules are known as lobular carcinomas. There are more than 18 other sub-types of breast cancer. Some, such as ductal carcinoma in situ, develop from pre-invasive lesions. The diagnosis of breast cancer is confirmed by taking a biopsy of the concerning tissue. Once the diagnosis is made, further tests are done to determine if the cancer has spread beyond the breast and which treatments are most likely to be effective.

Machine learning algorithms improve the more data they are exposed to. If there is one thing the healthcare system has in abundance, it's data. Due to different storage systems, ownership and privacy concerns, and no established process that allows people to easily share data with each other, there is a major amount of analysis that's not currently being done that could glean tremendous results for patients, doctors and healthcare organizations.

## 1.2 HOW AI WORKS?

Building an AI system is a careful process of reverse-engineering human traits and capabilities in a machine, and using its computational prowess to surpass what we are capable of.

To understand How Artificial Intelligence actually works, one needs to deep dive into the various sub domains of Artificial Intelligence and understand how those domains could be applied into the various fields of the industry. You can also take up an artificial intelligence course that will help you gain a comprehensive understanding.

**1.2.1 Machine Learning:** ML teaches a machine how to make inferences and decisions based on past experience. It identifies patterns, analyses past data to 3 infer the meaning of these data points to reach a possible conclusion without having

to involve human experience. This automation to reach conclusions by evaluating data, saves a human time for businesses and helps them make a better decision.

**1.2.2 Deep Learning:** Deep Learning is an ML technique. It teaches a machine to process inputs through layers in order to classify, infer and predict the outcome.

**1.2.3 Neural Networks:** Neural Networks work on the similar principles as Human Neural cells. They are a series of algorithms that captures the relationship between various underlying variables and processes the data as a human brain does.

**1.2.4 Natural Language Processing:** NLP is a science of reading, understanding, interpreting a language by a machine. Once a machine understands what the user intends to communicate, it responds accordingly.

**1.2.5 Computer Vision:** Computer vision algorithms try to understand an image by breaking down an image and studying different parts of the objects. This helps the machine classify and learn from a set of images, to make a better output decision based on previous observations.

**1.2.6 Cognitive Computing:** Cognitive computing algorithms try to mimic a human brain by analyzing text/speech/images/objects in a manner that a human does and tries to give the desired output.

## **1.3 CLASSIFICATION ALGORITHMS**

In machine learning, classification refers to a predictive modeling problem where a class label is predicted for a given example of input data. Types of classification algorithms are:

### **1.3.1 K-Nearest Neighbors:**

K nearest neighbors is a supervised machine learning algorithm often used in classification problems. It works on the simple assumption that “The apple doesn't fall far from the tree” meaning similar things are always in close proximity. This algorithm works by classifying the data points based on how the neighbors are classified. Any new case is classified based on a similarity measure of all the available cases. Technically, the algorithm classifies an unknown item by looking at k of its already -classified, nearest neighbor items by finding

out majority votes from nearest neighbors that have similar attributes as those used to map the items.

### **1.3.2 Logistic Regression:**

It's a classification algorithm that is used where the response variable is categorical. The idea of Logistic Regression is to find a relationship between features and probability of particular outcome. E.g., When we have to predict if a student passes or fails in an exam when the number of hours spent studying is given as a feature, the response variable has two values, pass and fail. This type of a problem is referred to as Binomial Logistic Regression, where the response variable has two values 0 and 1 or pass and fail or true and false. Multinomial Logistic Regression deals with situations where the response variable can have three or more possible values.

### **1.3.3 Support Vector Machine:**

Support vector machines (SVMs) are a popular linear classifier, the current version of which was developed by Vladimir Vapnik and Corinna Cortes. SVMs are supervised learning models, meaning sample data must be labeled, that can be applied to almost any type of data. They are especially effective at classification, numeral prediction, and pattern recognition tasks. SVMs find a line in between different classes of data such that the distance on either side of that line or hyperplane to the next-closest data points is maximized. In other words, support vector machines calculate a maximum-margin boundary that leads to a homogeneous partition of all data points. This classifies an SVM as a maximum margin classifier.

### **1.3.4 Decision tree:**

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It

is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

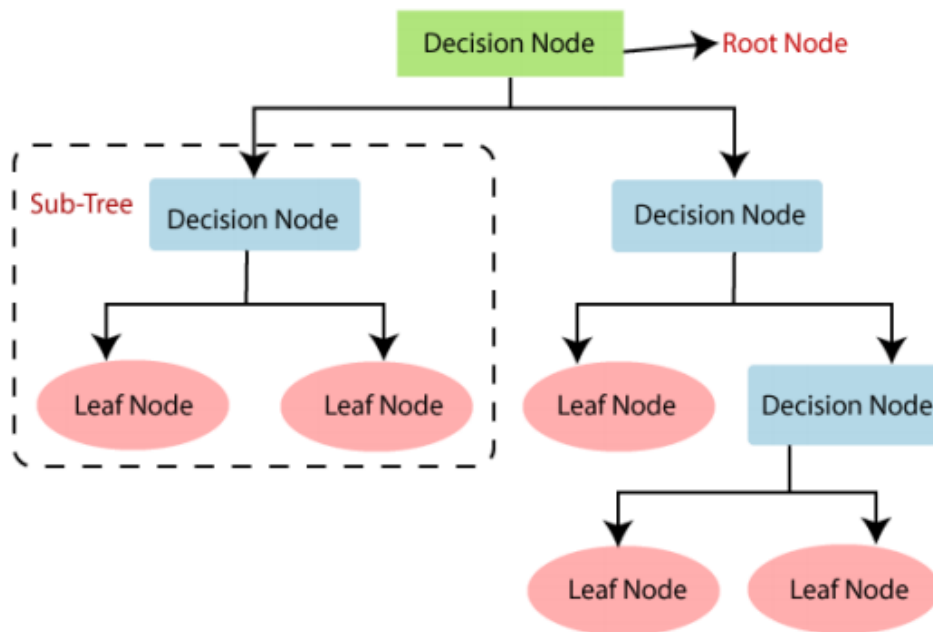
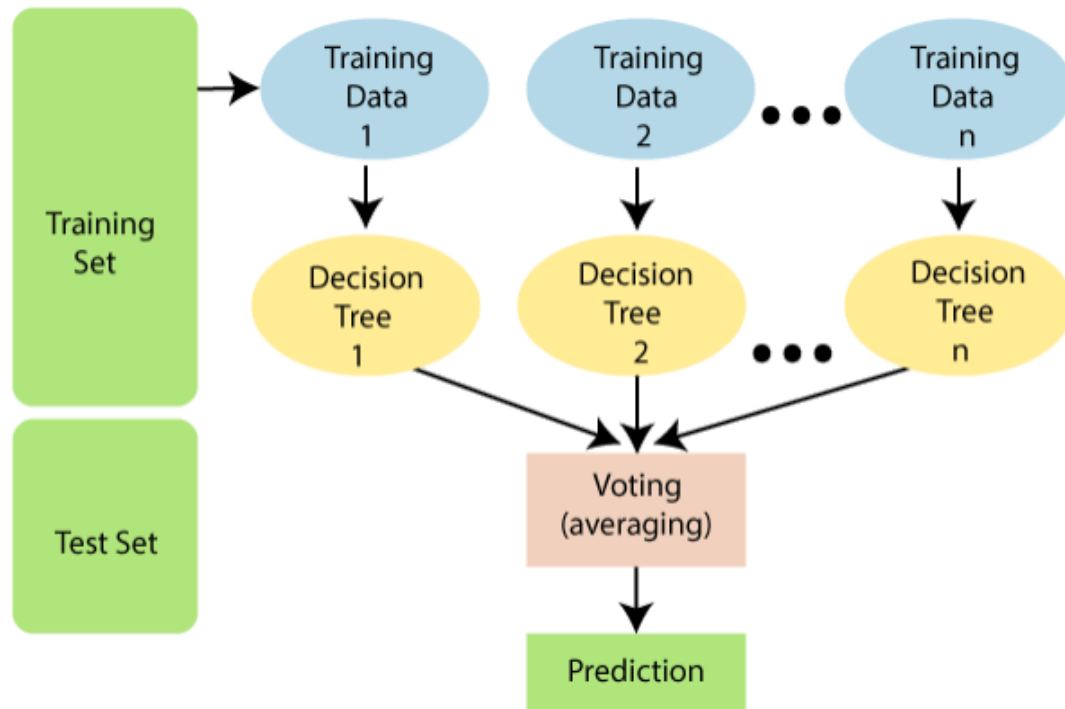


Figure 1.3.4 Decision Tree Work flow

### 1.3.5 Random Forest:

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output



*Figure 1.3.5 Random forest Workflow*

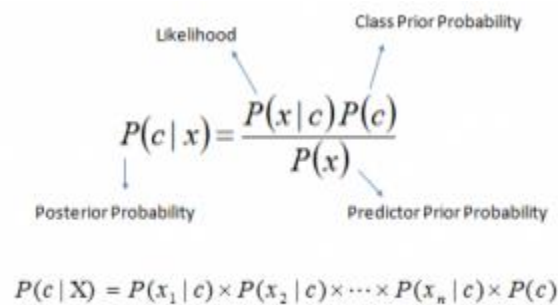
### **1.3.6 Naive Bayes:**

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.



The Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Bayes theorem provides a way of calculating posterior probability  $P(c|x)$  from  $P(c)$ ,  $P(x)$  and  $P(x|c)$ . Look at the equation below:



The diagram shows the Naive Bayes formula with labels for each term. The formula is:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Labels and arrows:

- $P(c|x)$  is labeled "Posterior Probability" with a downward arrow.
- $P(x|c)$  is labeled "Likelihood" with a leftward arrow.
- $P(c)$  is labeled "Class Prior Probability" with an upward arrow.
- $P(x)$  is labeled "Predictor Prior Probability" with a downward arrow.

Below the main formula, the joint probability formula is given:

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

*Figure 1.3.6 Naive Bayes formula*

Above,

- $P(c|x)$  is the posterior probability of class (c, target) given predictor (x, attributes).
- $P(c)$  is the prior probability of class.
- $P(x|c)$  is the likelihood which is the probability of the predictor given the class.
- $P(x)$  is the prior probability of the predictor.

### 1.3.7 Neural Network:

Neural networks are a set of algorithms, modeled loosely after the human brain, that are designed to recognize patterns. They interpret sensory data through a kind of machine perception, labeling or clustering raw input. The patterns they recognize are numerical, contained in vectors, into which all real-world data, be it images, sound, text or time series, must be translated.

Neural networks help us cluster and classify. You can think of them as a clustering and classification layer on top of the data you store and manage. They help to group unlabeled data according to similarities among the example inputs, and they classify data when they have a labeled dataset to train on. (Neural networks can also extract features that are fed to other algorithms for clustering and classification; so, you can think of deep neural networks as components of larger machine-learning applications involving algorithms for reinforcement learning, classification and regression.

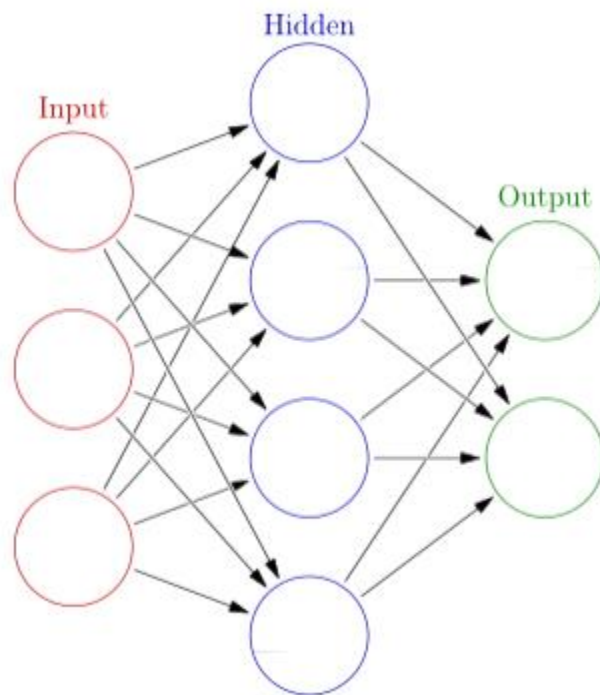


Figure 1.3.7 Perceptron Neural Network

**Applications of Neural Networks** Neural networks can be applied to a broad range of problems and can assess many different types of input, including images, videos, files, databases, and more. They also do not require explicit programming to interpret the content of those inputs. Because of the generalized approach to problem solving that neural networks offer, there is virtually no limit to the areas that this technique can be applied.

Some common 10 applications of neural networks today include image/pattern recognition, self driving vehicle trajectory prediction, facial recognition, data mining, email spam filtering, medical diagnosis, and cancer research. There are many more ways that neural nets are used today, and adoption is increasing rapidly.

## CHAPTER 2

### LITERATURE REVIEW

#### *Description*

*This chapter gives the brief history of detection of cancer in breast by classifying breast cells, and improvement within the upcoming year.*

---

#### **2.1 Breast Cancer Statistic**

Breast cancer may be a heterogeneous disease, commonly known to begin as neighborhood lesion within the breast, and so spread gradually. The second leading cause of death among women is cancer. In 2020 within the US there'll be 2,79,100 new cases and 42,690 estimated deaths. Together with 48,530 new cases of non-invasive (in-situ) cancer approximately every one-fifth of the affected women will die because of this disease. This (breast) dual organ inside a woman makes significant changes in size, shape and function in conjunction with pregnancy, lactation, puberty and postpartum so that a protective pregnancy should take place before the age of 30. When the breast grow abnormally it leads to breast cancer. These cells form a lump or mass because they divide quicker than healthy cells do and accumulate.

The lymph nodes and other part of the body can also be spreaded by the cells of the breast which is dangerous. Cells can often be seen from X-Ray or felt as a lump. Great information is obtainable so analysing this huge amount of knowledge to extract the novel and usable information or knowledge is incredibly complicated and time-consuming task the best technique for this is data mining. There is class imbalance within the data. Since the probability of having the disease is much more than not having the disease. The goal of this research paper is to differentiate Malignant and Benign patients. The dataset which we have used here is Wisconsin Dataset (WBC).

## **2.2 Breast Cancer Detection and Classification:**

This literature survey mostly focuses on identification of breast cancer and the region of breast affected by it. Mammography screening images CC and MLO is widely used in the diagnosis process. In this paper, preprocessing operation is done on input mammogram image and then the tumor region segmented from the image using MLO and results in a highlighted view on mammogram image. And they have used a Random forest classifier whose processing time is 6.25s and accuracy is 95% for images. GLCM, entropy and mean are utilized to examine the texture of the images . For testing on mammogram images, it has taken 3.16sec. Proper and fitting features are needed during the feature extraction to attain the best accuracy training database.

## **2.3 Breast Cancer Detection in Digital Mammograms:**

This paper basically discussed an automatic approach for detection of abnormalities in mammogram images. Unsharp masking is applied for enhancement of mammograms . Image preprocessing techniques are applied there to find out the suspicious region of interest. Median filtering has been used for noise removal. . Discrete wavelet transform has been applied on filtered images to get the accurate result prior to segmentation. Suspicious ROI has been segmented using the fuzzy-C-means with thresholding technique. Tamura features, shape based features and moment invariants are extracted from the segmented ROI to detect the abnormalities in the mammograms. Proposed algorithm has been validated on the 12 Mini-MIAS data set. This proposed method helps the algorithm to enhance ,segment and classify the abnormalities in mammograms. According to which , subtraction of enhanced preprocessed and enhanced inverted preprocessed images improves the detection of suspicious regions in mammogram images. Accuracy of the segmentation is improved using the FCM algorithm. Moment based features give better results than tamura and region based features. Accuracy of the proposed algorithm can be increased by feature extraction on suspicious regions.

## BRIEF LITERATURE SURVEY:

There have been several researches done in the area of data mining. Now a days data mining is becoming one of the emerging technologies. Based on study of some papers we have compiled our own literature survey as given below

*Table 2.1: Brief Literature Survey*

COMPARATIVE STUDY OF EXISTING CLASSIFICATION AND TECHNIQUES				
S.NO	PAPER TITLE	ALGORITHM	DATASET	RESULT
1.	COMPARATIVE STUDY ON DIFFERENT CLASSIFICATION TECHNIQUES FOR BREAST CANCER DATASET [6]	J48, MLP Rough Set	BREAST CANCER DATASET	J48: - 79.97% MLP: -73.35% ROUGH SET: - 71.36%
2.	INTEGRATION OF DATA MINING CLASSIFICATION AND ESEMBLE LEARNING FOR PREDICTION THE TYPE OF BREAST CANCER RECURRENCE [7].	NB, SVM, GRNN and J48	UCI BREAST CANCER DATASET	GRNN & J48 accuracy: 91% NB & SVM: 89%
3.	BREAST CANCER PREDICTION VIA MACHINE LEARNING [8]	KNN, SVM, RANDOM FOREST, GRADIENT BOOSTING	BREAST CANCER WISCONSIN(DIAGNOSTIC) DATA	KNN, SVM, RANDOM FOREST, GRADIENT BOOSTING accuracy: 70%
4.	BREAST CANCER DETECTION IN DIGITAL MAMMOGRAMS [9]	MOMENT IN VARIATION+FRACTAL DIMENSION, REGION BASED, TAMURA FEATURES	MINI-MIAS DATA SET	MOMENT IN VARIATION+FRACTAL DIMENSION accuracy:96.92  REGION BASED accuracy:91%  TAMURA FEATURES accuracy:78.6

5.	BREAST CANCER CLASSIFICATION AND DETECTION [10]	RANDOM FOREST	CT SCAN DATASET	RANDOM FOREST accuracy: 95%
6.	BREAST CANCER CLASSIFICATION USING DEEP LEARNING [11]	MULTILAYER PERCEPTRON ALGORITHM	BREAST CANCER DATASET	MULTILAYER PERCEPTRON ALGORITHM accuracy: 96.5%
7.	COMPARISON OF MACHINE LEARNING METHODS FOR BREAST CANCER DIAGNOSIS [12]	SUPPORT VECTOR MACHINE, ANN	WISCONSIN BREAST CANCER DATASET	SUPPORT VECTOR MACHINE accuracy: 96.9% ANN accuracy: 95.4%
8.	CLASSIFICATION OF BREAST CANCER DATA USING MACHINE LEARNING ALGORITHM [13]	KNN, SVM	WISCONSIN BREAST CANCER DATASET	KNN, SVM accuracy: 96%

## CHAPTER 3

### PROBLEM FORMULATION

#### *Description*

*This chapter explains the importance of Breast Cancer classification. This chapter also states the final goal of this project along with the objectives which are needed to be accomplished.*

---

Nowadays, cancer is one of the leading causes of morbidity and mortality around the world. Approximately 14.5 million people have died due to cancer, and it is estimated that this number will be above 28 million by 2030. According to a study by the American Cancer Society (ACS), in the USA the estimated deaths due to breast cancer account for approximately 14% of all cancer deaths (a total of 41,000 in 2017) which is in the second-leading cause of cancer death in women after lung and bronchus cancer. Additionally, breast cancer accounts for 30% of all newly discovered cancer cases. So, we are working on early detection for the cure of it.

#### **3.1 DATA DIFFICULTIES:**

Huge amount of data is available over internet and we have to extract the useful data from internet and for this we have used data mining technique. Selecting the dataset is very important because it is only used for training the machine. We have used the dataset that is Wisconsin Breast Cancer Dataset from UCL Machine Learning repository. The data set which we have used has 699 instances. Our one of the main aims is to remove the class imbalance and missing values as the data we have seen or tooked can have null or empty values which can overfit the model. If the values are missing, we replace it with higher negative value so ac to maintain accuracy and efficiency

For machine learning models to understand how to perform various actions, training datasets must first be fed into the machine learning algorithm, followed by validation



datasets (or testing datasets) to ensure that the model is interpreting this data accurately. Once you feed these training and validation sets into the system, subsequent datasets can then be used to sculpt your machine learning model going forward. The more data you provide to the ML system, the faster that model can learn and improve.

### **3.2 BREAST CANCER CLASSIFICATION TECHNIQUES/PROBLEMS**

There are many approaches for classifying breast cancer. We can also go manually for the testing of breast cancer but we have to find a relation between the different machine learning algorithm we do this by making the heat map of the available entities and relation. After comparing all the algorithm, it is very tedious for us to say which is the best system. Alternatively, we can also use ANN which is developed on the basis of human brain. The problem we have faced is in classifying the dataset into benign and malignant cases because it will tell us about the cancer and in this, we cannot expect error

### **3.3 REPRESENTATION MODELS:**

The most widely used algorithm is Support Vector Machine (SVM) and KNN algorithm. The data is splitted into training and testing dataset. There are many models we can select but the most widely and efficient according to our project we have selected. Training and testing the data again and again can increase the efficiency and this will help the model in prediction

## **CHAPTER 4**

### **PROPOSED WORK**

#### ***Description***

*This chapter explains the work we are going to in this project. This chapter also states the implementation process which are needed to be accomplished.*

---

In order to achieve the objective of the problem statement, we are using different machine learning algorithms.

#### **4.1 IMPLEMENTATION PROCESS**

The implementation process involves five steps:

#### **4.2 CLEANING OF DATA:**

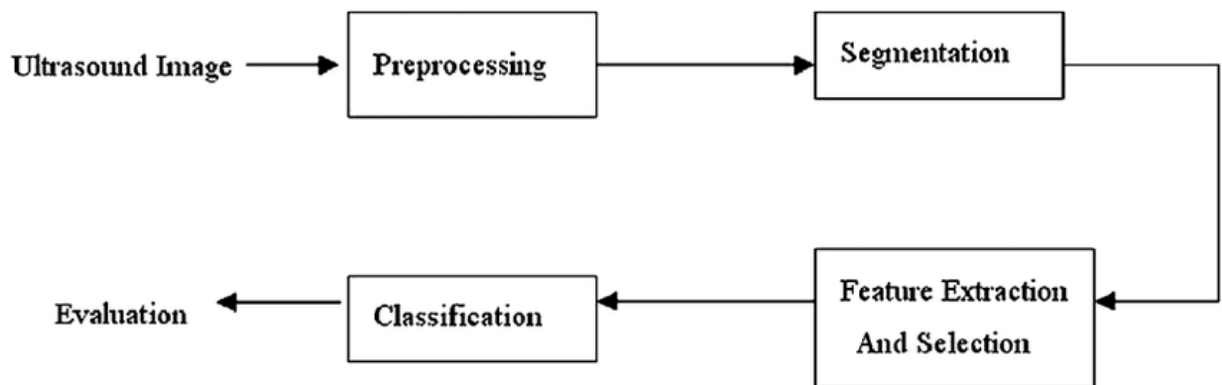
Initially the data that we collected from the UCI Machine learning library had a total of 699 records and 2 of them have some missing values. The data type of two features was also unknown i.e., having object data type. Therefore, data cleaning plays a crucial role before heading to the data exploration part.

This proposed system presents a comparison of machine learning (ML) algorithms: Support machine vector (SVM), Nearest Neighbor (NN) search. The data-set used is obtained from the Wisconsin datasets. For the implementation of the ML algorithms, the dataset was partitioned into the training set and testing set. A comparison between all the two algorithms will be made.

The algorithm that gives the best results will be supplied as a model to the website. The website will be made from a python framework, called flask. And it will host the database on Xampp or Firebase or inbuilt Python and flask libraries.

This data set is available on the UCI Machine Learning Repository. It consists of 11 real world attributes which are multivariate. The total number of instances is 699 and there are no missing values in this data set. The process of the proposed system is as follows,

1. The patient books an appointment through our website.
2. The patient will then meet the doctor offline for the respective appointment.
3. The doctor will first check the patient manually, then perform a breast mammogram or an ultrasound. That ultra sound will show an image of the breast consisting the lumps or not.
4. If the lumps are detected, a biopsy will be performed. The digitized image of the Fine Needle Aspirate (FNA) is what forms the features of the dataset.
5. Those numbers will be provided to the system by the doctor and the model will detect if it's a benign or a malignant cancer.
6. The report will be then forwarded to the patient on their respective account.



*Figure 4.2 Breast cancer classification mechanism*

#### 4.3 EXPLORATORY DATA ANALYSIS (EDA):

Data exploration has its own importance as it helps in finding the insights or the relation between different features of a cell. A count plot is drawn between Mitosis vs Class in below Fig. 5.1 which shows directly proportional relationship, where 2 represents Benign cell and 4 represents Malignant cell.

##### 4.1.3. FEATURE SELECTION:

There are many algorithms or methodologies present for feature selection like Principle Component Analysis , Logistic Regression etc. So Logistic regression is used to select the features. Logistic regression gives us the weights/coefficients for every feature which determine the dependency of each feature in classifying the cells. After applying this algorithm, the following data is received shown in the following

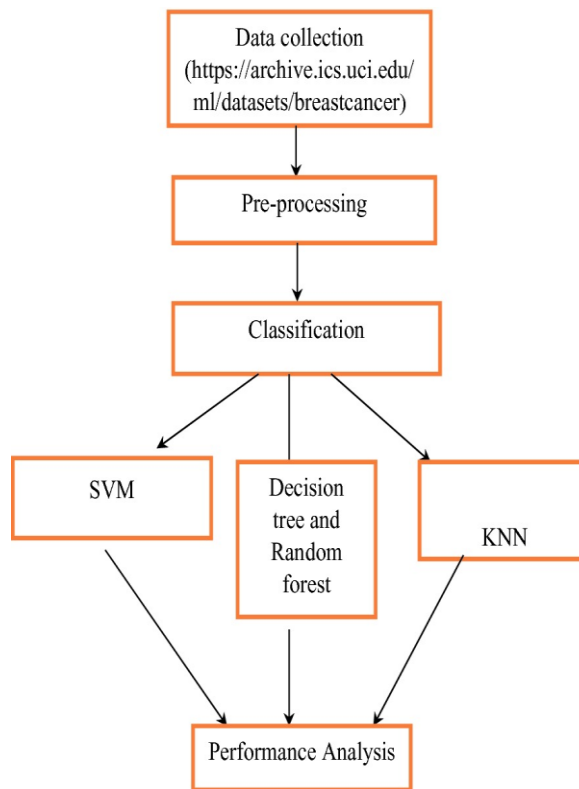
**Table 4.1 Features of the dataset**

	Features						
	Mitosis	Uniform cell shape	Bare Nuclei	Clump Thickness	Uniform Size	Marginal Adhesion	Bland Chromatin
Weight / Coeff.	0.4606	0.5028	0.3789	0.3745	0.2182	0.2547	0.2980

## CHAPTER 5

### SYSTEM DESIGN AND APPLICATION

#### 5.1 Introduction



*Figure 5.1 Breast cancer classification using KNN and SVM*

#### 5.2 REALITY OF AI FOR MEDICAL DIAGNOSIS

Throughout the last decade, artificial intelligence has penetrated healthcare at a growing pace. The projections for this technology's growth in the next five years are promising, as well — in fact, researchers project a 44.9% growth rate for medical AI.

How do healthcare professionals make use of artificial intelligence, machine learning, and the other powerful tools at their disposal? Here's our review of the top 10 applications related to machine learning for medical diagnosis.

### **5.2.1 ONCOLOGY**

In oncology, the importance of detecting a malignant tumor on time is vital. This is why the accuracy and precision of the diagnosis are crucial in this field. Machine learning helps oncologists detect the disease at its earliest stages. With the help of tools like DeepGene, medical professionals can detect somatic mutations easily (a somatic mutation is an acquired change in a genetic code of one or more cells). Artificial intelligence pinpoints mutation markers faster and with higher accuracy than humans do. In addition to pinpointing the tumor, machine learning can accurately determine if it's malignant or benign in milliseconds. Although computer-based predictions aren't error-free, the accuracy of classification is impressive at 88%.

### **5.2.2 PATHOLOGY**

Given the worldwide shortage of pathologists, there's a considerable need for adopting machine learning to make progress in this field. The need to process large datasets also makes Pathology extremely lucrative for artificial intelligence implementation. Here are the most promising ways of using machine learning to indicate medical diagnosis:

1. Improving the precision of blood and culture analysis using automated tissue and cell quantification.
2. Mapping disease cells and flagging areas of interest on a medical slide.
3. Creating tumor staging paradigms.
4. Improving healthcare professionals' productivity by increasing the speed of profile scanning.

### **5.2.3 DERMATOLOGY**

In dermatology, artificial intelligence is used to improve clinical decision-making and ensure the accuracy of skin disease diagnoses. Physicians hope that machine learning implementation in this field will reduce the number of unnecessary biopsies dermatologists have to put patients through. There are plenty of functional machine learning implementations in Dermatology, namely:

1. An algorithm that separates melanomas from benign skin lesions with higher precision than that of a human.
2. Tools that track the development and changes in skin moles, helping detect pathological conditions at the earliest stages.
3. Algorithms that pinpoint biological markers for acne, nail fungus, and seborrheic dermatitis.

### **5.2.4. GENETICS AND GENEMICS:**

Recently, artificial intelligence has helped geneticists progress significantly in the transcription of human genes. Although the Human Genome Project is the poster case of healthcare and technology joining forces for potentially revolutionary research, it's not the only way to use machine learning in medical diagnosis. Machine learning and AI technologies are key players in preventive genetics. Scientists increasingly rely on algorithms to determine how drugs, chemicals, and environmental factors influence the human genome. Last but not least, geneticists are hopeful that they will be able to improve the efficiency of gene editing, changing DNA fragments to protect a fetus from the impact of a mutation or reverse its effect. Since the scientific community has repeatedly expressed ethical concerns regarding gene editing, its use in genetics is limited to fighting diseases that are considered incurable. At the moment, gene editing researchers are focused on fighting Cystic Fibrosis and Huntington Disease.

### **5.2.5 MENTAL HEALTH**

Statistically, mental health disorders are one of the costliest conditions to manage in the United States. Research shows that 1 in 5 adult Americans is affected by a mental disorder. The impact of leaving these conditions untreated or misdiagnosed is disastrous: low productivity, increased health spending, and lower overall life quality. Artificial intelligence can have a groundbreaking impact on mental health research and the efficiency of medical diagnosis through machine learning. The top applications of innovative technologies in the field are:

1. Personalized cognitive behavior therapy (CBT) fueled by chatbots and virtual therapists.
2. Mental disease prevention by creating machine learning tools that help high-risk groups avoid social isolation.
3. Identifying groups with a high risk of suicide and providing them with support and assistance.
4. Early detection of mental disorders using machine learning and data science: diagnosing clinical depression, bipolar disorder, anxiety, and more.

### **5.2.6 CRITICAL CARE**

Artificial intelligence has the potential to reduce the length of an average ICU stay by predicting early-onset sepsis and adjusting ventilator and other equipment settings according to a patient's conditions. Using artificial intelligence helps doctors avoid poor judgment calls — premature extubation or prolonged intubation — that have a strong link to raising ICU mortality rates. In addition, machine learning in ICU can help physicians identify high-risk patients to make sure no early deterioration sign is left unnoticed. Innovative technologies can provide physicians with insights into patients' well-being inside the ICU. For example, through the use of technology, intensive care physicians discovered that delirious patients are more sensitive to light than to noise.



### **5.2.7 EYECARE**

The diagnosis of Ophthalmology conditions has a lot of room for machine learning optimization. Some of the latest innovations that these healthcare centers have adopted are:

1. AI-driven vision screening programs that help provide a point-of-care medical diagnosis based on machine learning for Ophthalmological conditions.
2. Identifying Diabetic Retinopathy and providing physicians with treatment insights by analyzing patient data (in 2018, the FDA approved the first among these machine learning scanners for clinical use).
3. Early-stage diagnosis of Macular Degeneration with the help of deep learning algorithms.
4. High-precision glaucoma and cataract screening.

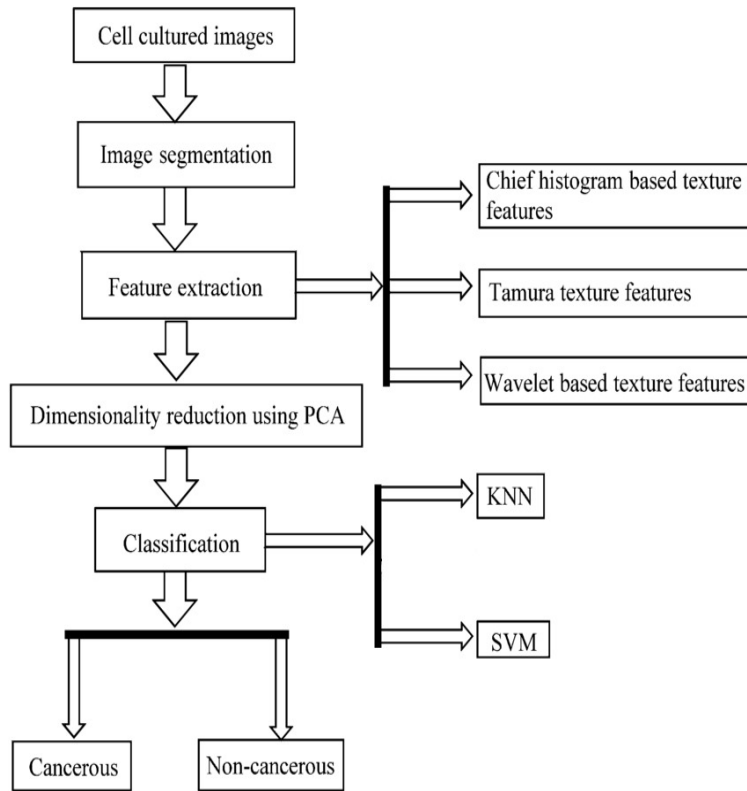
### **5.2.8 DIABETES:**

1. Since Type I and Type II diabetes are highly widespread conditions, the amount of data researchers have at their disposal is tremendous.
2. To advance research in this field, scientists need to focus on consolidating these insights and putting them into a single framework. This is one of the primary machine learning goals in the field.
3. Over the last decade, the range of machine learning application examples for diagnosing and treating diabetes has grown exponentially:
4. Using vector machine modeling and building neural networks for pre-diabetes screening.

5. Creating tools for managing personalized insulin delivery, as well as artificial pancreas systems. Predicting treatable complications in diabetes patients to improve the quality of their lives.
6. Identifying genetic and other biomarkers for diabetes.

#### **5.1.9 PUBLIC HEALTH**

1. Artificial intelligence allows healthcare professionals to increase the scale of medical diagnoses using machine learning and shift from analyzing individual cases to monitoring communities and predicting disease outbreaks.
2. Artificial intelligence and data science help Epidemiologists, public officials, and healthcare facility managers aggregate social media, clinical, and other patient data to determine healthcare trends and address the pain points of community residents.
3. In some countries, AI-based prediction monitoring tools are adopted at a national level
4. Some other Applications can be seen in below figure:



*Figure 5.2 Breast Cancer detection using KNN and SVM*

So, from the above method we found out in our case the KNN algorithm is giving us higher accuracy more than that of SVM. But when we train the dataset again and again then the SVM can be better in such cases we can reduce the dimensionality by using the PCA algorithm.

## **CHAPTER 6**

### **IMPLEMENTATION**

#### **6.1 Breast Cancer Classification**

Breast cancer classification divides carcinoma into categories depending on how they have spread or if they have spread at all. Classification algorithms predict one or more discrete variables, supported the opposite attributes within the dataset. data processing software is required to run the classification algorithms. the aim of classification is to pick the simplest treatment. Classification is vital because it allows scientists to spot, group, and properly name organisms via a uniform system. Classification and clustering are two widely used methods in data processing.

\Clustering methods aim to extract information from a knowledge set to get groups or clusters and describe the info set itself. Classification, also referred to as supervised learning in machine learning, aims to classify unknown situations supported learning existing patterns and categories from the info set and subsequently predict future situations. The training set, which is employed to create the classifying structure, and therefore the test set, which tends to assess the classifier, are commonly mentioned in classification tasks. Classification may be a quite complex optimizationl problem.

Many ML techniques are applied by researchers in solving this classification problem. The most famous algorithm that is used for breast cancer classification or prediction is an artificial neural network, random forest, support vector machine, etc. Scientists strive to seek out the simplest algorithm to realise the foremost accurate classification result, however, data of variable quality also will influence the classification result. Further, the rarity of knowledge will influence the number of algorithm applications also. If the carcinoma is found early, there are more treatment options and a far better chance for

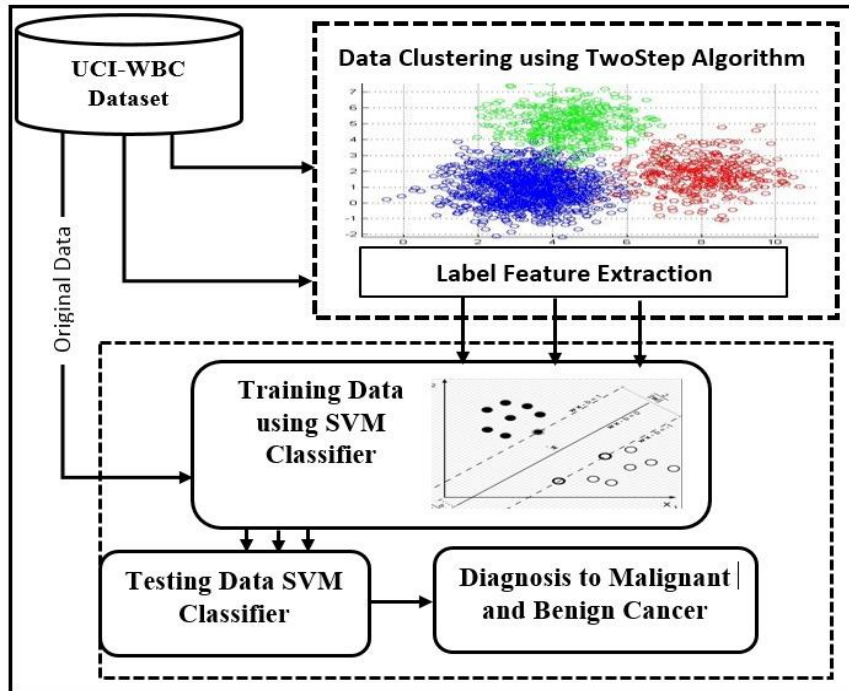
survival. Women whose carcinoma is detected at an early stage have a 93 percent or higher survival rate within the first five years. Getting checked regularly can put your mind comfortable. Finding cancer early can also save your life.

## **6.2 Machine learning algorithms**

Machine learning is an application of AI (AI) that gives systems the power to automatically learn and improve from experience without being manually programmed. Machine learning focuses and depends on the event of computer programs that will access the data provided and use it to learn for themselves. The method of learning begins with data or datasets, examples, experiences, or instructions, so they can then figure out a pattern and or improve them in the near future, if necessary.

### **1.Support vector machine:**

In machine learning, support vector machines are supervised models. A support vector machine creates a hyperplane when classifying the objects. A hyperplane is a line on a plane that distinguishes the two classes. Given a group of coaching examples, each marked as belonging to at least one or the opposite of two categories, an SVM training algorithm builds a model that assigns new examples to at least one category or the opposite, making it a non-probabilistic binary linear classifier (although methods like Platt scaling exist to use SVM during a probabilistic classification setting). New examples are then mapped into that very same space and predicted to belong to a category supported the side of the gap on which they fall.



*Figure 6.2.1 SVM Mechanism*

**1. Hyperplane** – As we can see in the above diagram, it is a decision plane or space which is divided between a set of objects having different classes.

**2. Margin** – It may be defined as the gap between two lines on the closet data points of different classes. It can be calculated as the perpendicular distance from the line to the support vectors. Large margin is considered as a good margin and small margin is considered as a bad margin.

## **2. K nearest neighbors:**

KNN (K- Nearest Neighbors) is one among many supervised learning algorithms utilised in data processing and machine learning, it's a classifier algorithm where the training is predicated "how similar" may be a data from other. It is a lazy algorithm. KNN works by finding the distances between a point and all the examples within the data, selecting the required number examples (K) closest to the point, then votes for the leading frequent label.

# KNN Classifier Algorithm

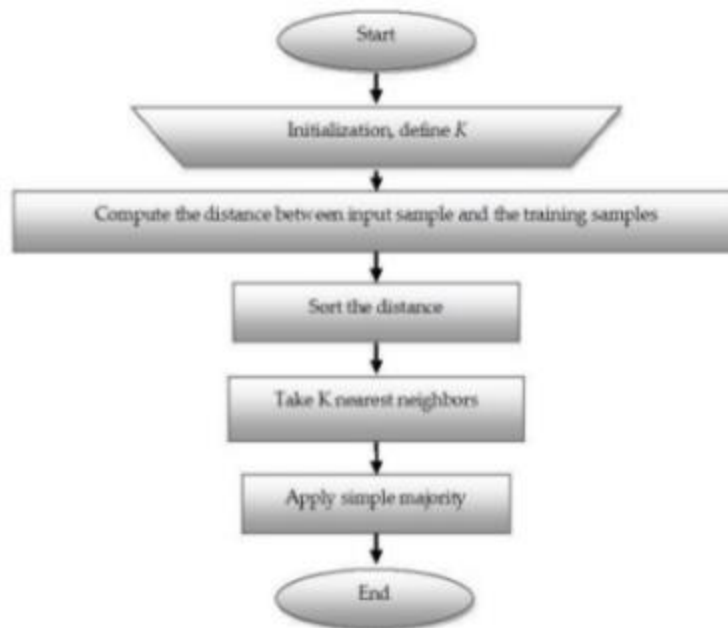
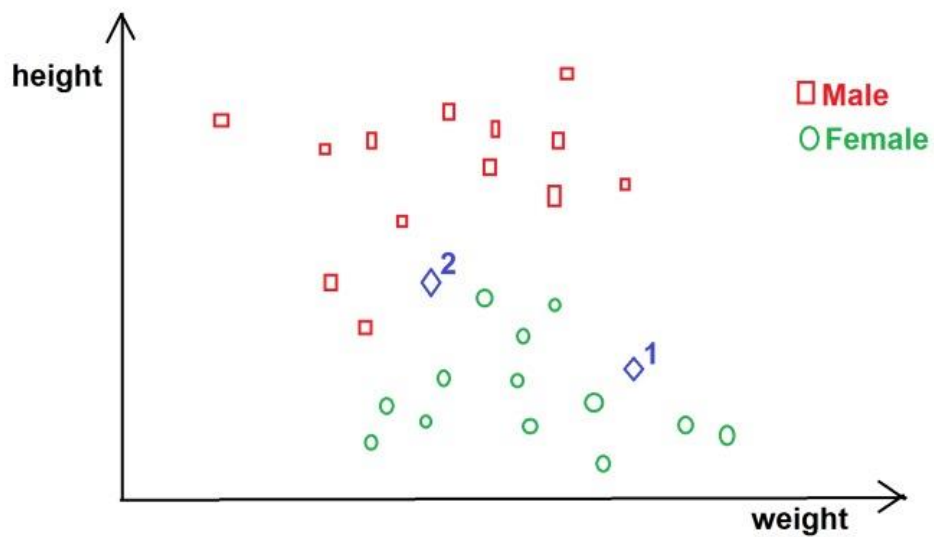


Figure 6.2.2 KNN Process flow chart



### **3. Artificial neural network:**

Artificial neural networks (ANN) or neural network systems are computing systems that mimic the functioning of a human brain. The main aim of the algorithm is to provide a faster result with more accuracy than an old or traditional system. If the algorithm has been given the data or an image about a particular object then the algorithm will quickly be able to identify or categorize images that do not contain the said object.

### **PROGRAM**

#### **#Imported Libraries**

```
from sklearn import model_selection
from sklearn.metrics import classification_report, accuracy_score
from pandas.plotting import scatter_matrix
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
from sklearn.model_selection import train_test_split
```

#### **#Classifier Library**

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
```

#### **#Replacing the missing values with high negative to remove overfitting**

```
df.replace('?', -9999, inplace=True)
print(df.axes)
df.drop(['id'], 1, inplace=True)
```



### **#Splitting Dataset**

```
X_train, X_test, Y_train, Y_test = train_test_split(X,Y,test_size=0.25 , random_state=0)
```

### **# We are doing cross validation in 10 folds to get the best result and then printing the accuracy**

```
for name,model in models:
```

```
kfold = model_selection.KFold(n_splits=10, random_state=seed) #run 10 times and select the best results
```

```
cv_results =model_selection.cross_val_score(model,X_train,y_train,cv=kfold,scoring=scoring)
```

### **#Fitting a model and computing the score**

```
results.append(cv_results)
```

```
names.append(name)
```

```
msg="%s: %f (%f)" % (name,cv_results.mean(),cv_results.std())
```

```
print(msg)
```

### **#Accuracy only on training data**

```
print(results)
```

### **#After selecting best result verifying it with passing the values in it**

```
clf.fit(X_train,y_train)
```

```
accuracy = clf.score(X_test,y_test)
```

```
print(accuracy)
```

```
example = np.array([[4,2,1,1,1,2,3,2,5]])
```

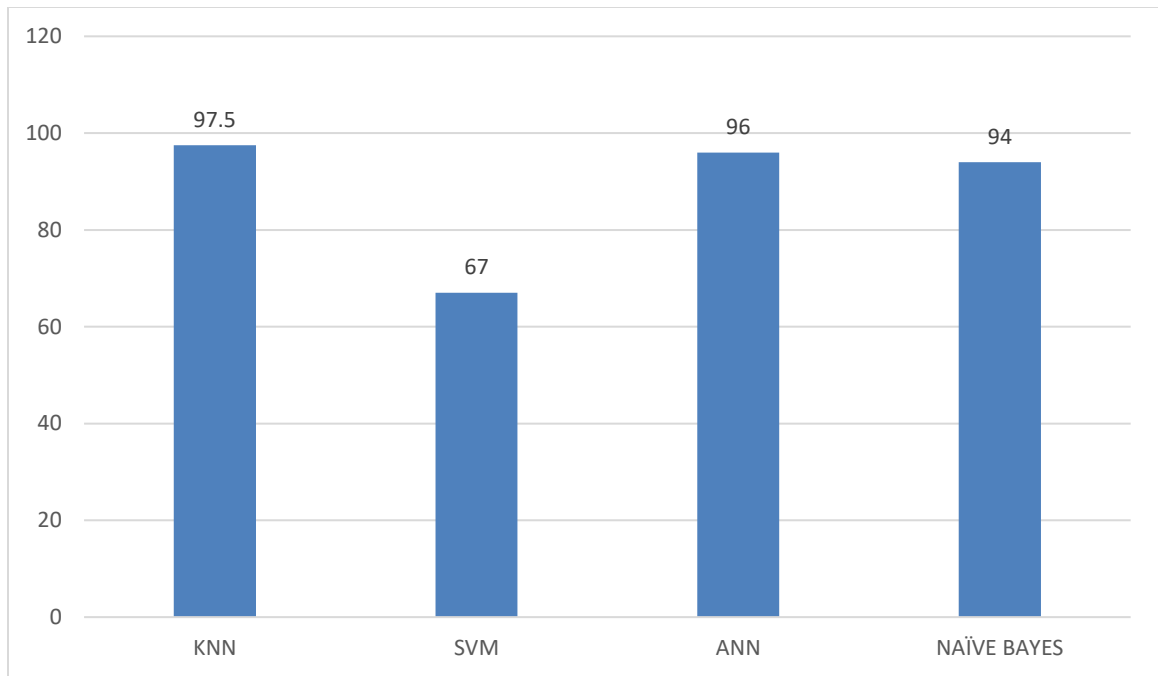
## **CHAPTER 7**

### **RESULT ANALYSIS**

Breast cancer is one among the foremost widely recognized kind of cancer among feminine population in the entire world. It is still challenging task to detect and classify the cancer tumor precisely. Mammography is considered as a standout amongst the most conclusive and dependable method for proper identification and classification of the breast cancer.

Here, in this paper we are proposing a system based on machine learning for classification of breast cancer (BC) along with the comparative study of two machine learning (ML) classifier. The idea is to select the region of interest (ROI) at very first from the mammograms. At that point important features have been extracted using GLCM (grey level co-occurrence matrix).

Thereafter, extracted features are then utilized to train our classifiers SVM and KNN individually. The mammogram are then characterized either into benign or malignant using the trained classifier. Proposed system is implemented on standard MIAS databases. Classification performance of both classifiers are contrasted in terms of accuracy, recall, precision, specificity and F1 score. We found that SVM achieved higher accuracy of 94% than KNN with better recall and F1 score.



## **CHAPTER 8**

### **CONCLUSION AND FUTURE SCOPE**

#### **8.1 CONCLUSION**

Breast cancer if found at an early stage will help save lives of thousands of women or even men. These projects help the real-world patients and doctors to gather as much information as they can. The research on nine papers has helped us gather the data for the project proposed by us. By using machine learning algorithms, we will be able to classify and predict the cancer into being or malignant. Machine learning algorithms can be used for medical oriented research, it advances the system, reduces human errors and lowers manual mistakes. Women worldwide are suffering from breast cancer on large scale and it is the second largest in numbers of death among women.

This paper introduces us using two algorithms Support vector Machines and KNN Algorithm in breast cancer classification. We are getting best accuracy in KNN method but both the systems are giving excellent accuracy. The accuracy in KNN method is 97.5%. If we increase the size of data set than the running time will also increase and we can achieve higher accuracy. Shuffling the training data again can fluctuate the accuracy very sharply.

For future research we can use CAdE and CAdx technologies which take high resolution images and neural network can be used for detecting it faster. Later we will be working on detection side to give more help in detection by just giving the data and directly getting result about cancer.

## **8.2 FUTURE SCOPE**

The proposed model can be used in future for the help of medical staff and also in early detection of breast cancer. The manual error will be reduced and the chance of not detecting virus will be very less. Creating mobile applications and making it more user friendly for early detection. Machine learning technology from DeepMind Health and the AI research team at Google will be applied to approximately 7,500 mammograms provided by the Cancer Research UK-funded OPTIMAM database at the Royal Surrey County Hospital NHS Foundation Trust. The team plans to evaluate the possibility of training the computer algorithm to analyze the images for signs of cancer and alert radiologists more accurately than is possible with current technology.

## REFERENCES

- [1] May, R. M. 1997. The Scientific Wealth of Nations, *Science*, vol. 275, no. 5301, pp. 793-796.
- [2] Torres, R. McNee, S. M. Abel, M. Konstan, J. A. and Riedl, J. 2004. Enhancing Digital Libraries with TechLens, *Proceedings of JCDL'04*, pp. 228-236.
- [3] Pennock, D. M. Horvitz, E. Lawrence, S. and Giles, L. C. 2000. Collaborative Filtering by Personality Diagnosis: A Hybrid Memory- and Model-Based Approach, in *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*.
- [4] Fano, R. M. 1956. Information theory and the retrieval of recorded information, in *Documentation in Action*, Shera, J. H. Kent, A. Perry, J. W. (Edts), New York: Reinhold Publ. Co., pp.238–244.
- [5] Small, H. 1973. Co-citation in the scientific literature: a new measure of the relationship between two documents, *Journal of the American Society for Information Science*, vol. 24, pp. 265–269
- [6] Giles, C. L. Bollacker, K. D. And Lawrence, S. 1998. Cite Seer: an automatic citation indexing system, In *Digital Libraries 98 -The Third ACM Conference on Digital Libraries*, pp. 89-98.
- [7] Garfield, E. and Welljams-Dor, A. 1992. Citation data: their use as quantitative indicators for science and technology evaluation and policy-making, *Science & Public Policy*, vol. 19, no. 5, pp. 321-327.
- [8] <https://www.breastcancerindia.net/statistics/trends.html>
- [9] Sayeth Saabith, Elankovan Sundararajan, Azuraliza Abu Bakar,” *Comparitive Study on Different Classification techniques for Breast Cancer Dataset, IJCSMC*, Vol. 3, Issue. 10, October 2014, pg.185 – 191

- [10] Jesús Silva, Omar Bonerge Pineda Lezama, Noel Varela, Luz Adriana Borrero, "Integration of Data Mining Classification and Ensemble Learning For Prediction The Type Of Breast Cancer Recurrence", International Conference on Green, Pervasive, and Cloud Computing GPC 2019: Green, Pervasive, and Cloud Computing pp 18-30
- [11] Mamatha Sai Yarabarla, Lakshmi Kavya Ravi, Dr. A. Sivasangari, "Breast Cancer Prediction via Machine Learning", Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019) IEEE Xplore Part Number: CFP19J32-ART
- [12] Ebru Aydınođ Bayrak, Pınar Kırıcı, Tolga Ensari, "Comparison of Machine Learning Methods for Breast Cancer Diagnosis", 978-1-7281-1013-4/19/\$31.00 ©2019 IEEE
- [13] Burak Akbugday, "Classification of Breast Cancer Data Using Machine Learning Algorithms", 978-1-7281-2420-9/19/\$31.00 ©2019 IEEE
- [14] <https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html>
- [15].<https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/1097->
- [16] <https://www.breastcancerindia.net/statistics/trends.html>
- [17]. [https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+\(original\)](https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original))
- [18] <https://searchenterpriseai.techtarget.com/definition/machine-learning-ML>

