A project report on

SWE3004 SOFTWARE DESIGN AND DEVELOPMENT PROJECT

A MACHINE LEARNING FRAMEWORK FOR PREDICTION OF BREAST CANCER

Submitted in partial fulfillment for the award of the degree of

M.Tech

In

Software Engineering

by

G BHARGAV (15MIS0115)

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

SCHOOL OF INFORMATION TECHNOLOGY AND ENGINEERING

DEPARTMENT OF SOFTWARE AND SYSTEMS ENGINEERING

NOVEMBER 2019

A MACHINE LEARNING FRAMEWORK FOR PREDICTION OF BREAST CANCER

Submitted in partial fulfillment for the award of the degree of

M.Tech

In

Software Engineering

by

G BHARGAV (15MIS0115)

SCHOOL OF INFORMATION TECHNOLOGY AND ENGINEERING

DEPARTMENT OF SOFTWARE AND SYSTEMS ENGINEERING

NOVEMBER 2019

DECLARATION

I hereby declare that the thesis entitled "**A MACHINE LEARNING FRAMEWORK FOR PREDICTION OF BREAST CANCER"** submitted by me, for the award of the degree of **M Tech (Software Engineering)** is a record of bonafide work carried out by me under the supervision of **Prof. BRINDHA K**

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any degree or diploma in this institute or any other institute or university.

Place: Vellore

Date: **Signature of the Candidate**

CERTIFICATE

This is to certify that the thesis entitled **"A MACHINE LEARNING FRAMEWORK FOR PREDICTION OF BREAST CANCER"** submitted by **G BHARGAV (15MIS0115)** School of Information Technology and Engineering VIT, for the award of the degree of **M Tech Software Engineering** is a record of bonafide work carried out by him under my supervision.

The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university. The Project report fulfils the requirements and regulations of VIT and in my opinion meets the necessary standards for submission.

Signature of the Guide **Signature of the HOD**

Internal Examiner **External Examiner**

ABSTRACT

A breast cancer is a malignant tumor that the starts from cells of the breast. A malignant tumor is a group cells that may grow into (invade) surrounding tissues or spread (metastasize) to distant areas of the body. Breast cancer occurs mainly in omen, but men can get it, too. Many people do not realize that men have breast tissue and that they can develop breast cancer. Male breast cancer is a relatively rare cancer in men that originates from the breast. As it presents a similar pathology as female breast cancer. Male breast cancer remains under diagnosed and, due to delays in diagnosis, is often also undertreated. The investigation and management of male breast cancer are based on studies on female patients. At present there is a need for further research into male breast cancer. The symptoms, diagnosis and treatment for male breast cancer are all similar to female breast cancer.

Cancer of the male breast cancer accounts for about 1% of all malignancies in men and 1% of all breast cancer. Poor levels of awareness often results in late presentation and delayed diagnosis in our environment. It is estimated that more than 90% of male breast cancers are estrogen receptor – positive. Male breast cancer tissue may also be positive for androgen receptors.

Breast cancer metastasis accounts for the majority of deaths from breast cancer. Detection for breast cancer metastasis at the earliest stage is important for the management ad prediction of breast cancer progression. Emerging Techniques using the analysis of circulating tumor cells promising results in predicting and identifying the early stage of breast cancer metastasis in patients.

ACKNOWLEDGEMENT

It is my pleasure to express with deep sense of gratitude to **Prof. BRINDHA K, Assistant Professor Selection Grade,** School of Information Technology and Engineering, Vellore Institute of Technology, for his constant guidance, continual encouragement, understanding; more than all, he taught me patience in my endeavor. My association with him is not confined to academics only, but it is a great opportunity on my part of work an intellectual and expert in the field of **Machine Learning.**

I would like to express my gratitude to **Dr. G. Viswanathan,** Chancellor, and VIT University. I am highly grateful to our Vice President, **Shri Sekar Viswanathan,** Vice Chancellor **Dr. Anand A. Samuel,** Pro-Vice Chancellor **Dr. S. Narayana,** and Dean **Prof. Balakrushna Tripathy,** School of Information Technology and Engineering, for providing with an environment to work in and for his inspiration during the tenure of the course.

In jubilant mood I express ingeniously my whole-hearted thanks to **Prof. Sree Dharinya S,** HOD of Software and Systems Engineering, SITE all teaching staff and members working as limbs of our university for their not-self-centered enthusiasm coupled with timely encouragements showered on me with zeal, which prompted the acquirement of the requisite knowledge to finalize my course study successfully. I would like to thank my parents for their support.

It is indeed a pleasure to thank my friends who persuaded and encouraged me to take up and complete this task. At last nut not least, I express my gratitude and appreciation to all those who have helped me directly or indirectly toward the successful completion of this project.

Place: Vellore Name of the student

Date: 22-11-2019 **G BHARGAV**

TABLE OF CONTENTS

CONTENTS.....iii

LIST OF FIGURES……vi

LIST OF TABLES.....vii

LIST OF ACRONYMS……viii

CHAPTER 1

INTRODUCTION **1 - 5**

1.1 INTORDUCTION…….1

1.1.1 BACKGROUND……2

1.2 PROBLEM STATEMENT……2

1.3 EXISTING SYSTEM……3

1.4 SCOPE……...3

1.5 OBJECTIVE……..4

1.6 ORGANIZATION OF PROJECT…….5

CHAPTER 2 **6 - 27**

SYSTEM OVERVIEW

2.1 PROPOSED SYSTEM……..6

2.1.1 ADVANTAGES OF PROPOSED SYSTEM…….6-7

2.2 LITERATURE SURVEY…….7-19

2.3 FUNCTIONAL REQUIREMENT……..20

2.4 NON FUNCTIONAL REQUIREMENT……...21-22

2.5 SYSTEM OVERVIEW…….23-27

CHAPTER 3 **28 - 37**

SYSTEM DESIGN

3.1 MODULE DESCRIPTION…….....28-30

3.2 SYSTEM ARCHITECTURE……..31
3.3 SYSTEM CONFIGURATION……32
3.4 UML DIAGRAMS…….33-37
3.4.1 USE CASE DIAGRAM……33
3.4.2 CLASS DIAGRAM……..34
3.4.3 SEQUENCE DIAGRAM……..35
3.4.4 ACTIVITY DIAGRAM……36
3.4.5 DATA FLOW DIAGRAM……...37
CHAPTER 4 **38 - 45**
IMPLEMENTATION AND TESTING
4.1 IMPLEMENTATION……..38
4.2 TESTING…….39
4.2.1 UNIT TESTING……...39
4.2.2 INTEGRATION TESTING……..40
4.2.2.1 TOP DOWN INTEGRATION……..40
4.2.2.2 BOTTOM UP INTEGRATION……40
4.2.3 USER ACCEPTING TESTING……...41
4.2.4 OUTPUT TESTING…….41
4.2.5 FUNCTIONAL TESTING……...41
4.2.6 SYSTEM TESTING……41-42
4.2.7 WHITE BOX TESTING……..42
4.2.8 BLACK BOX TESTING……42-43
4.2.9 VALIDATION TESTING……...43
4.3 TEST CASES……..44

4.3.1 TESTING STRATEGY……45
4.3.2 USER TRAINING……45
4.3.3 MAINTAINANCE……...45
CHAPTER 5 **46 - 62**
SAMPLE CODE
SAMPLE CODE……..46-62
CHAPTER 6 **63 - 71**
RESULTS
6.1 DATA……63
6.2 PRE-PROCESSING DATA…….....64
6.3 NUMBER OF TRAINING SAMPLES……65
6.4 ACCURACY PARAMETER……66
6.5 ACCURACY WITH REGRESSION……67
6.6 PREDICTED LABEL……68
6.7 CLASSIFICATION ALGORITHM……69
6.8 SCATTER MATRIX……...70
6.9 ACCURACY……...71
CHAPTER 7 **72**
NOVELTY
7 NOVELTY…….....72
CHAPTER 8 **73**
CONCLUSION AND FUTURE WORK
8 CONCLUSION AND FUTURE WORK……...73
REFERENCES **74**

LIST OF FIGURES

3.2 SYSTEM ARCHITECTURE……..31
3.3 SYSTEM CONFIGURATION……32
3.4 UML DIAGRAMS…….33-37
3.4.1 USE CASE DIAGRAM……33
3.4.2 CLASS DIAGRAM……..34
3.4.3 SEQUENCE DIAGRAM……..35
3.4.4 ACTIVITY DIAGRAM……36
3.4.5 DATA FLOW DIAGRAM……...37
6.1 DATA…….63
6.2 PRE-PROCESSING DATA…….....64
6.3 NUMBER OF TRAINING SAMPLES…….65
6.4 ACCURACY PARAMETER…….66
6.5 ACCURACY WITH REGRESSION…….67
6.6 PREDICTED LABEL…….68
6.7 CLASSIFICATION ALGORITHM……69
6.8 SCATTER MATRIX……...70
6.9 ACCURACY……...71

LIST OF TABLES

4.3.1 TEST CASES 1……44

4.3.2 TEST CASES 2……44

LIST OF ACRONYMS

NN	Neural Networks
SVM	Support Vector Machine
NB	Naïve Bayes
DT	Decision Tree
RF	Random Forest
URL	Uniform Resource Locator
CSV	Comma Separated Values
TDM	Text Data Mining
FIG	Figure
TAB	Table

Chapter 1

INTORDUCTION

1.1 INTRODUCTION

Breast cancer is a malignant cell growth in the breast. If left untreated, the cancer spreads to other areas of the body. Excluding skin cancer, breast cancer is the most common type of cancer in women in India, accounting for one of every three cancer diagnoses.

The incidence of breast cancer rises after age 40. The highest incidence (approximately 80% of invasive cases) occurs in women over age 50.

Breast cancer is the most common cancer in women, affecting about 10% of all women at some stages of their life. In recent years, the incidence rate keeps increasing and data show that the survival rate is 88% after five years from diagnosis and 80% after 10 years from diagnosis. Early prediction of breast cancer is one of the most crucial works in the follow-up process.

It is the prime reason for demise of women. It is the second dangerous cancer after lung cancer. In the year according to the statistics provided by World Cancer Research Fund it is estimated that over 2 million new cases were recorded out of which 6 lakhs death were approximated. Of all the cancers, breast cancer constitutes of 11.6% in new cancer cases and come up with 24.2% of cancers among women. In case of any sign or symptom, usually people visit doctor immediately, who refer to on oncologist, if required. The oncologist can diagnose breast cancer by: Undertaking through medical history.

Breast cancer – extremely heterogeneous disease caused by interactions of both inherited and environment risk factors. Progressive accumulation of genetic and epigenetic changes in breast cancer cells. Tumors with similar clinical and pathological presentations may have different behaviors.

Breast cancer is a kind of cancer that develops from breast cells. Breast cancer usually starts off in the inner lining of milk ducts or the lobules that supply them with milk. A malignant tumor can spread to the different parts of the body.

Breast cancer can also occur in men, but it’s far less common.

1.1.1 BACKGROUND

A personal history of invasive breast cancer, ductal carcinoma in situ, or lobular carcinoma in situ. A personal history of benign (non-cancer) breast disease. A family history of breast cancer in a first-degree relative (mother, daughter, or sister). Inherited changes in the genes or in the other genes that increase the risk of breast cancer. Breast tissue that is dense on a mammogram. Exposure of breast tissue to estrogen made by the body. This may be cause by: At an early age. Older age at first birth or never having given birth. Starting at a later age. Taking hormones such as estrogen combined with progestin for symptoms of menopause. Treatment with radiation therapy to the breast/chest. It may caused due to drinking alcohol. And also it may caused due to Obesity.

1.2 PROBLEM STATEMENT

This project focuses in investigating the probability of predicting the type of breast cancer (malignant or benign) from the given characteristics of breast cancer computed from digitized images. The cases provided, are cases diagnosed with some type of tumor, but only some of them (approximately 37%) are malignant. This project will examine the data available and attempt to predict the possibility that a breast cancer diagnosis is malignant or benign on the attributes collected from the breast cancer. Discriminate benign and malignant in an unknown sample from fine needle aspirates taken from patient breasts. Comparing the performance of neural network classification with linear programming classification. Neural Networks and Support Vector Machine Classification can do better job classifying the data. In this project I am using high frequency intraday returns so as to reduce the noise preset in the input data. Using Neural Network which can calculate the number of hidden nodes needed for predicting volatility at runtime so as to reduce the problems involved in using more hidden nodes or less. Prediction of stock volatility is also tested using multilayer feed forward network. In this case sigmoidal function is used as activation function.

1.3 EXISTNG SYSTEM

Rule based approach is used for classification which gives static range value for different classes. Therefore we will not able dynamic images or outlier behavior images. Multi classification problem is not optimized and ignore the class imbalance problem. Therefore in learning all classes is not input in learning phases so it became a biased learning. Features set is not normalize. Therefore different features show different outputs and show different representation during training phase of classifiers.

The importance of classifying cancer patients into high or low risk groups has led many research teams, from the biomedical and the bioinformatics field, to study the application of machine learning methods. Therefore, these techniques have been utilized as an aim to model the progression and treatment of cancerous condition. In addition, the ability of ML tools to detect key features from complex datasets reveals their importance.

1.4 SCOPE

This educational activity has been developed to meet the educational needs of multidisciplinary health care professionals who provide care to postmenopausal women.

Upon completion of this activity, participants will be able to:

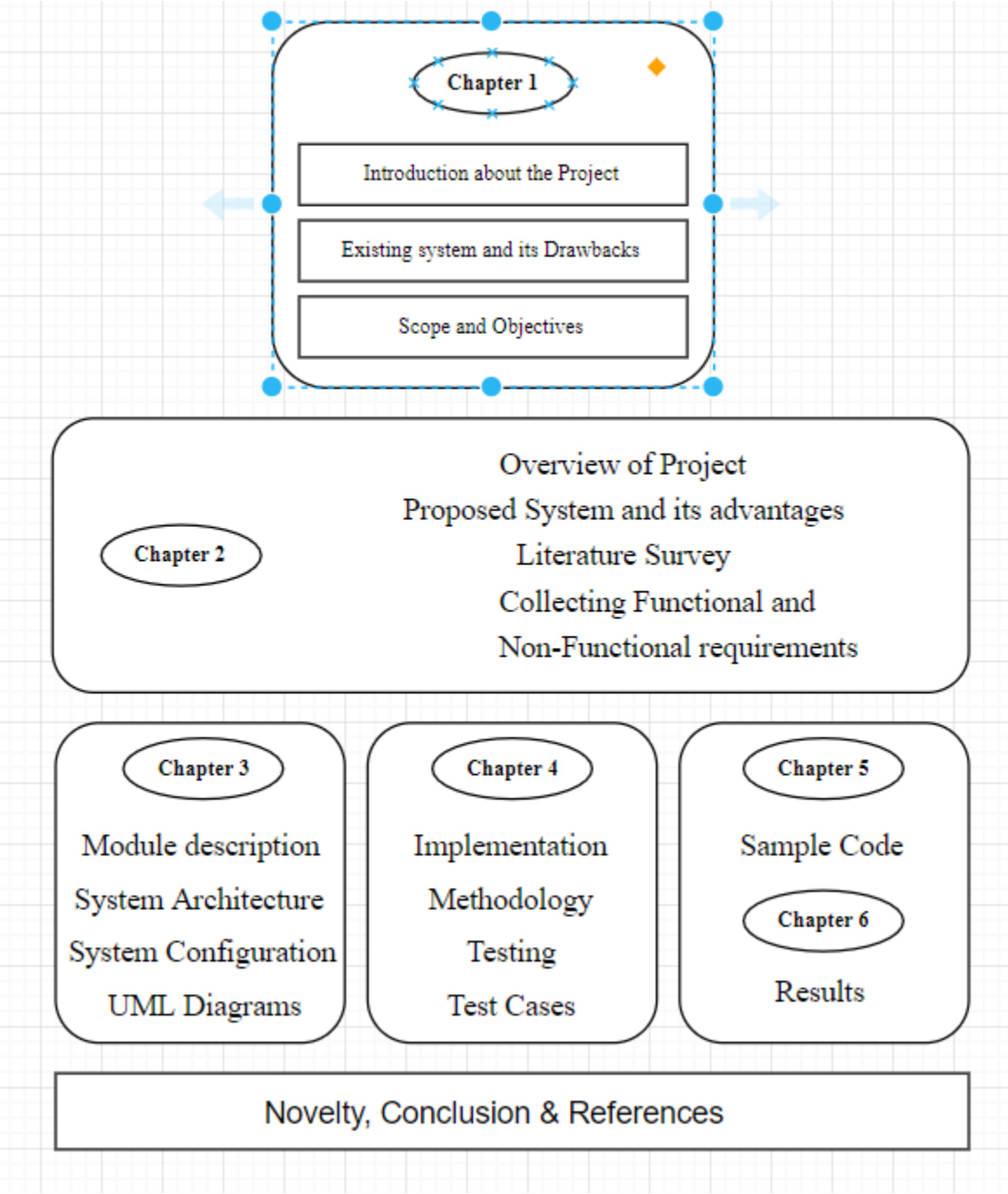
- Describe risk factors for breast cancer and identify populations in North America at increased risk of breast cancer.
- List the appropriate lifestyle modifications for patients
- Select the appropriate lifestyle modifications for patients
- Weight the relative risks and benefits of the primary chemoprevention options for breast cancer – including tamoxifen and raloxifen – when used in postmenopausal women.
- Recognize the importance of counselling prior to and during breast cancer treatment.

1.5 OBJECTIVE

The Objective of this paper is to compare and identify an accurate model to predict the incidence of breast cancer based on various patients clinical records. Three data mining models are applied i.e., Support Vector Machine (SVM), Artificial Neural Network (ANN), Naïve Bayes etc., Feature space is highly discussed due to its high influence on the efficiency and effectiveness of the learning process. To test the influence of feature space reduction, a hybrid between principal component analysis and related data mining models is proposed, which applies a principle component analysis method to reduce the feature space. To evaluate the performance of these models, two widely used test data sets are used, Breast Cancer Database 10-fold cross-validation method is implemented to estimate the test error of each model. The results performed by this analysis demonstrate a comprehensive trade-off between these strategies and also provides detailed evaluation on the models. It is expected that I real application, physicians and patients can benefit from the feature recognition outcome to prevent breast cancer.

- Improve the coordination and multidisciplinary management of breast cancer cases in breast units
- Exploiting novel sources of information
- Exploiting the rich information contained in routine imaging examinations
- Developing tools for the visual assessment of the possible aesthetic outcome of Breast Conservative therapy
- Testing new screening strategies based on individualized risk profiles
- Assess the use of imaging for the measurement of extent of disease and for monitoring and mid-therapy adaption of treatment protocols.
- Reduce the morbidity of breast cancer therapy through non-invasive imaging-guided mechanisms.

1.6 ORGANIZATION OF PROJECT



Chapter 2

SYSTEM OVERVIEW

2.1 PROPOSED SYSTEM

In our proposed system, mammogram image can be enhancement using Gaussian filter. Second the segmentation is done using Fuzzy C means for partitioning the mammogram image into multiple segments to identify the mass easily and features are extracted using HWT (Haar Wavelet Features). Further tumor has been analyzed and classified using Multi – SVM (Support Vector Machine) classifier.

SVMs deliver a unique solution, since the optimality problem is convex. This is an advantage compared to Neural Networks, which have multiple solutions associated with local minima and for this reason may not be robust over different samples.

Breast cancer starts to grow in the human body when cells in the breast are growing most in an unexpected manner. After these cells grow, it can be seen by x-ray. Basically, there are two types of breast cancer, cancer that spread into another area and cancer that can't spread into another area. Among the world women breast cancer is the first and the most leading of death of women and the accurate diagnosis have lots of advantage to prevent and detection of the disease. As breast cancer recurrence is high, good diagnosis is important. Many studies have been conducted to analyze Breast cancer Data. This research is going to be implemented by different data mining and Decision tree. So to get a more accurate value about the recurrence of breast cancer I am going to use data sets which were taken from the UCI machine learning repository and data was kept in CSV file.

2.1.1 ADVATAGES OF PROPOSED SYSTEM

- Reduces the risk of dying from breast cancer of 1k women who have a mammogram every 2 years for 20 years, 7 deaths are prevented.
- Reduces the risk of having to undergo chemotherapy Screening often allows for the detection of cancers at an early stage of development. Treatment is then possible without

chemotherapy

- Allows women to know the health of their breasts The vast majority of women (nearly 98%) will not have breast cancer if their mammograms and additional examinations do not reveal cancers.

2.2 LITERATURE SURVEY

In **[1]** author Ginsburg at 2011 have diagnosed, For a large number of women newly diagnosed in the world, it has been ascertain that, breast cancer is a neglected disease in terms of other numerically more frequent health problems. It has also been described as an orphan disease, in the sense that the very detailed knowledge about tumor characteristics and the necessary host biology capable of providing basic care is absent. Current international cancer policy and planning initiatives are irrelevant to breast cancer, with the exception of nutritional recommendation.

Breast cancer is the most prevalent cancer in the world (4.4 million survivors up to 5 years following diagnosis) and the second most common cause of cancer related mortality in women wide world (Parkin et at 2005). It also accounts for 23% (1.38 million) of the total new cancer cases and 14% (458,400) of the total cancer deaths in 2008 and ranks second most common cancer overall (10.9% of all cancers) but ranks fifth as cause of death. 1.15 million new breast cancer cases were recorded in 2004 and over 500,000 deaths reported around the world and more than half of all cases occurred in industrialized countries.

Ferlay at 2010, Breast cancer incidence rates vary from 19.3 per 100,000 women in Eastern Africa to 89.7 per 100,000 women in Western Europe. They are normally high in developed regions of the world (except Japan) and low in most of the developing regions. Due to more favorable survival of breast cancer in developed regions, the range of mortality rates is very much less, approximately 6-19 per 100,000. Notwithstanding, it is still the most frequent cause of cancer death in women in both developing (269 000 deaths, 12.7% of total) and developed regions, where the estimated 189 000 deaths is almost equal to the estimated number of deaths from lung cancer. For some time now, there have been some encouraging in both breast cancer incidence and mortality trends with the incidence of new cases stabilizing as well as death rates falling in some high income or developed countries. However, this appears to be vice versa in

developing countries (Kanavos 2006). Notably, breast cancer incidence rates have leveled off since 1990, with a decrease of 3.5%/year from 2001 to 2004 (Li et at 2003)

In **[2]** authors proposed a differences in the population Breast cancer is common in women both in the developed and the developing countries, comprising 16% of all female cancers. Although it is thought to be a common cancer in the developed countries, a majority (69%) of all breast cancer deaths occurs in developing world. Indeed, increase life expectancy, increase urbanization and adoption of western lifestyles have increased the incidence of breast cancer in the developing countries (Kanavos, 2006). Even though it is now the most common cancer both in developed and developing regions with around 690 000 new cases estimated in each region, much of the burden of incidence, morbidity, and mortality will occur in the developing world with population ratio of 1:4 (Ferlay et al., 2010). As developing countries succeed in achieving lifestyles similar to those in advanced economies, they will also encounter much higher cancer rates, particularly cancers of the breast. This forms part of a larger epidemiological transition in which the burden of chronic, non-communicable disease once limited to industrialized nations, is now increasing in less developed countries (Kanavos, 2006).
A report by Stewart et al (Stewart and Kleigues, 2003), mentioned that many of the new cancer cases are now occurring among women from low and middle income countries, where the incidence is increasing by as much as 5% per each year and there are about three fourths of breast cancer deaths occurring worldwide.

In **[3]** authors compared the different types Breast cancer variation among population, or the regional differences in the types have been attributed to the following: prevalence of major risk factors, availability and use of medical practices such as cancer screening, availability and quality of treatment, completeness of reporting, and age structure. However, geographic areas, and counties and parishes within countries also determine the frequency of the most commonly diagnosed cases or deaths (Garcia M et al., 2007). The highly penetrant but rare susceptibility genes, BRCA1 and BRCA2 (Fackenthal et al., 2007) and more prevalent, but lower penetrance genes, CHEK2 and FGFR (Easton et al., 2007) have been indicated to be the key inter-individual and inter-group differences in the distribution of reproductive risk factors. Countries with

massive economic development over the past 50 years, such Japan, Singapore, and urban areas of China have experience an increase in breast cancer incidence (Horn-Ross et al., 2000). Age-standardized incidence rates for breast cancer 1998–2002 were 110 (nonHispanic Caucasians, California), 82.3 (Ontario, Canada), 41.3 (Hong Kong) and 14.7 (Jiashan, China) (Curado et al., 2007). Reports on migration studies reveal that the incidence of breast cancer changes significantly over one to two generations to more closely reflect the breast cancer risk in the adopted country (Ziegler et al., 1993), which seems to occur in parallel with dynamics in diet and certain indicators of acculturation. Notably, evaluation of differences in risk factors and natural history of all tumor types, would permit for comparisons based on geographical regions, socioeconomic status and levels of industrialization (Ginsburg et al, 2011).

In **[4]** authors shown that the majority of breast tumors from Asian women are estrogen receptor (ER) negative. Also it has been indicated that both pre-and postmenopausal Asian women with breast cancer, are likely to have ER positive tumors as Caucasians (Uy et al., 2007). In addition, greater proportion of ER+ tumors in a Vietnamese cohort, has been found in a studies on ER positivity among premenopausal breast cancer cases as compared with the comparison group of Caucasian women in Australia. (Tran and Lawson, 2004).

Considerably, variation in the gene profiles of tumors from populations of different genetic/ethnic backgrounds have also been reported. About 15% of sporadic breast cancer, which are BRCA1 origin in Caucasian women appears to have the basal phenotype. On the other hand, other studies have also suggested that breast cancer in women of African ancestry may have a higher proportion of basal phenotype (Carey et al., 2006). In similar manner among Nigerians, a high frequency of basal like tumors was observed, where 87 of 148 (59%) breast cancer cases were both ER- and HER2- (Olopade et al., 2004)

1. Molecular changes during extended neoadjuvant letrozole treatment of breast cancer: distinguishing acquired resistance from dormant tumors.
2. This is the first patient-matched gene expression study investigating long-term aromatase inhibitor-induced dormancy and acquired resistance in breast cancer.
3. Dormant tumors continue to change during treatment whereas acquired resistant tumors more closely resemble their diagnostic samples.

4. Cancer begins when healthy cells in the breast change and grow out of control, forming a mass or sheet of cells called a tumor. A tumor can be cancerous or benign. A cancerous tumor is malignant, meaning it can be grow and spread to other parts of the body.
5. A benign tumor means the tumor can grow but will not spread.

In **[5]** authors researched the lymph system is very important in breast cancer research in that, it is one way breast cancers can spread and has several parts. Lymph nodes are small, bean-shaped collections of immune system cells that are connected by lymphatic vessels. These vessels are like small veins, except that they carry a clear fluid called lymph in place of blood away from the breast. They also contain Lymph tissue fluid and waste products, in 7 addition to immune system cells. Breast cancer cells can enter lymphatic vessels and begin to grow in lymph nodes. Most lymphatic vessels in the breast connect to lymph nodes under the arm (axillary nodes). Some lymphatic vessels that connect to lymph nodes inside the chest are called internal lymph nodes, and those either above or below the collarbone are called supraclavicular or infra clavicular nodes (American Cancer Society booklet).
There are several types of breast cancer, but some of them are quite rare. Currently, majority of all breast cancers worldwide are the ductal and lobular sub types .However, the ductal subtype accounting accounts for the majority of the diagnosed cases, constituting for about 40–75% (Rakha et al., 2006). There are two models for the ductal subtype The first 'ductal' model, reported by Lerwill (Lerwill, 2008) are as follow: First of all, it recognizes flat epithelial atypia (FEA), to atypical ductal hyperplasia (ADH) and then ductal carcinoma in situ (DCIS) as the non-obligate precursors of the advanced invasive and metastatic ductal carcinoma. In the second model, usual epithelial ductal hyperplasia (UDH) was proposed as an intermediate stage of progression between FEA and DCIS (Page et al., 1985). In the case of lobular subtype, atypical lobular hyperplasia (ALH) and lobular carcinoma in situ (LCIS) was also proposed as the non-obligate precursor lesions to invasive lobular carcinoma. (Boecker et al., 2002).

In **[6]** authors proposed a method of producing manipulated mouse mammary gland, In a mammary cancer model induced by the expression of polyoma middle T oncoprotein, ADAM-12 has been shown to increase stromal cell apoptosis and decrease tumor cell apoptosis (Kveiborg

et al, 2005) ADAM-10 knock-out (KO) embryos suffer from cell growth arrest and apoptosis associated with an overexpression of full-length E-cadherin (Maretzky et al., 2005).

1. Gene transduction in the mammary gland can be easily and quickly achieved, and gene expression can be controlled by Dox administration.
2. This system for genetic manipulation could be useful for analyzing genes involved in breast cancer.

In **[7]** authors Sue Hudson proposed a Statistical Methods and Linear Regression model, Adjusting for Breast Cancer in analyses of volumetric mammographic density and breast cancer risk.

Prevalence of major risk factors, availability and use of medical practices such as cancer screening, availability and quality of treatment, completeness of reporting, and age structure. However, geographic areas, and counties and parishes within countries also determine the frequency of the most commonly diagnosed cases or deaths (Garcia M et al., 2007). The highly penetrant but rare susceptibility genes, BRCA1 and BRCA2 (Fackenthal et al., 2007) and more prevalent, but lower penetrance genes, CHEK2 and FGFR (Easton et al., 2007) have been indicated to be the key inter-individual and inter-group differences in the distribution of reproductive risk factors. Countries with massive economic development over the past 50 years, such Japan, Singapore, and urban areas of China have experience an increase in breast cancer incidence (Horn-Ross et al., 2000)

1. Fully automated assessment of mammographic density (MD), a biomarker of breast cancer risk.
2. When volumetric MD-breast cancer risk associations are investigated, NDV can be used when BMI data are unavailable

In **[8]** authors Carolyn Nickson proposed a Statistical analyses and Gail Model, Prospective validation of the Breast Cancer Risk Assessment Tool. Adjuvant systemic therapy in women with early stage disease is guided by prognostic and predictive factors, including stage, grade, estrogen receptor (ER) and progesterone receptor (PR) status. These parameters help physicians to select adjuvant systemic therapy. However, these remain imperfect tools, in that some patients

receive systemic chemotherapy even though they can be cured by surgery alone. If these parameters are used alone in selecting treatment as recommended by 1998 and 2001 St Gallen consensus statement, up to 90% of node negative breast cancer patients will be candidate for adjuvant chemotherapy, although only about 30% of them will relapse and thus need adjuvant chemotherapy. In terms of tumor grade, it is certainly important in that it is predictive of risk over time, but it lacks standardization.

1. The findings from this study indicate that the Gail model, or a simplified version of this model, is an effective tool for active breast cancer screening participants aged 50-69 years to groups according to risk.

In **[9]** authors, Wendy Yi-Ying proposed a Statistical Methods and multi-state model, Over diagnosis in the population-based organized breast cancer screening program estimated by a non-homogenous. Breast cancer variation among population, or the regional differences in the types have been attributed to the following: prevalence of major risk factors, availability and use of medical practices such as cancer screening, availability and quality of treatment, completeness of reporting, and age structure. However, geographic areas, and counties and parishes within countries also determine the frequency of the most commonly diagnosed cases or deaths (Garcia M et al., 2007).

1. We estimated the frequency of over diagnosis of breast cancer due to screening in women 50-69 years old by using individual screening data from the population based organized screening program.
2. The frequency of over diagnosis in the prevalent screening round was higher than that in subsequent rounds.

In **[10]** authors, Vasiliki Pelekanou proposed a Classification Method, Co-localization in breast tumor micro environment predicts survival differently in ER-positive and negative cancers. Estrogen receptors (ER) and/or progesterone receptors (PR) presence is currently a component of routine evaluation of breast cancer specimens. ER was the first analyzed in breast cancer in the late 1950s, and was the first molecular marker evaluated for prognosis and therapy response for breast cancer. The status of ER has been shown to have significant predictive value on tumor

response to hormone therapy in metastatic disease as well as for adjuvant therapy after local excision (Dowsett M et al., 2008). On the other hand, the role of PR status in predicting tumor response to therapy is still unclear, although it has shown promise. From our current study, most of the cases were positive for both ER and PR, corresponding to 88% and 82% respectively which have been confirmed with other studies.

1. Tumor-associated macrophages display high molecular and functional complexity.
2. Macrophage activity markers correlate with survival differently in ER+ and ER- cancers.

In **[11]** authors, Tuong L.Nruyen proposed a Statistical Methods and Decision Tree Model, Predicting interval and screen-detected breast cancers from mammographic density defined by different brightness thresholds.

1. Measurement of mammographic density.
2. We have found the more risk-predicting information can be obtained from the conventional concept of breast density.

In **[12]** authors Garnet Anderson, proposed an model to resolve the Risk, Risk prediction for estrogen receptor-specific breast cancers in two large prospective cohorts. Adjuvant systemic therapy in women with early stage disease is guided by prognostic and predictive factors, including stage, grade, estrogen receptor (ER) and progesterone receptor (PR) status. These parameters help physicians to select adjuvant systemic therapy. However, these remain imperfect tools, in that some patients receive systemic chemotherapy even though they can be cured by surgery alone. In terms of tumor grade, it is certainly important in that it is predictive of risk over time, but it lacks standardization. In case of undifferentiated cancers (grade 3), patients are truly at high risk and may benefit from chemotherapy.

1. We found that modelling heterogeneous risk associations of epidemiological factors yields little improvement in BC risk prediction.
2. Country specific distributions of the risk factors among the EPIC women based in their Age and Years of follow-up

In **[13]** authors Sri Hari Krishna Vellanki, used a Quantitative polymerase chain reaction analysis Cleavage of the extracellular domain of junctional adhesion molecule-A is associated with resistance to anti-HER2 therapies in breast cancer settings. Cleavage activity and due to its overexpression during pregnancy. The mechanism employed is the cleavage of the pro domain from the rest of the protein. Its involvement in major contemporary pathologies like cancer, inflammatory and vascular diseases seem to be connected to its cleavage abilities and this is due to the large variety of substrates it is able to cut. Certain growth factors and receptors can be activated and inactivated. These proteolytic activities normally form part of cleavage cascades referred to as Regulated Intramembrane Proteolysis, leading to intracellular signaling. The cleavage has been reported to control the function of a given substrate leading to activation, inactivation or modulation of the activity. (Arribas and Esselens ,2009). Receptors cleavage offers another way to regulate the response of the cell to growth factors and cytokines.

1. Breast cancer cell culture and drug treatments. Immunohistochemistry on material of patients with breast cancer.
2. Cleavage is augmented in cells with resistance to trastuzumad and lapatinib therapy.

In **[14]** authors, found the Data analysis and Literature search and data collection. The tissue samples were collected as quickly as possible after the removal of the organ from the subjects (patients) and immediately fixed in 10% formalin. This was to preserve the cells and cellular constituents in a state as close as possible to the living cells and also maintains the antigenicity to allow them to be processed without change. Each sample was stratified according to the following: Age, size, weight, anatomical location, tumor relation to surgical margin and nodal status. They were then sectioned and placed on embedding cassette for tissue processing. Molecular patterns of cancer colonization in lymph nodes of breast cancer patients. The lymph nodes are functional units of the immune system that act as immunological hubs supporting the complex interactions between T cells, B cells. The LN is a dynamic organ capable of undergoing dramatic remodeling, in terms of both architecture and function.

In **[15]** authors, analyzed the Sentimental analysis on, No association between low-dose aspirin use and breast cancer outcomes overall: a Swedish population based study. Breast cancer is the

most common malignancy among women in high-income countries. Additional cost-effective therapies are still needed. Several studies have indicated that low-dose aspirin use around the time of a breast cancer diagnosis. 50 Cases of cancer patients who have undergone radical mastectomy with no prior treatment, with ages ranging from 24 - 89 were obtained from Saint Anthony General Hospital, Porto Portugal together with an informed consent form of participation in the project. Patients involved were pathologically confirmed of the following node-negative and node-positive breast cancer: Invasive ductal carcinoma, In situ ductal carcinoma and Invasive lobular carcinoma with histological grades 1 to 3. There were 40 node-negative cases and 10 node-positive cases with tumor size ranging from 0.6 cm to 4.7 cm.

In **[16]** authors, proposed the Baseline questionnaires and SVM model, Age-specific breast cancer risk by body mass index and familial risk: prospective family study cohort. Another strong marker of breast cancer risk is the degree of mammographic density. It has been indicated that, the risk in women with more dense breast is four to six times higher than those with less dense breast (Boyd et al., 1995). Evidence suggest that, the etiology of mammographic density may be due to the exposure to steroid hormone, since it decreases with age (Boyd et al., 2002) as well as in women. Time in the study began 2 months after the age of completion of the baseline questionnaire and ended at age of 80 years. We plotted the predicted age-specific absolute.

In **[17]** authors, proposed a Statistical analysis and Random Forest Model, Disrupted circadian clocks and altered tissue mechanics in primary human breast tumors. Most of the patients have smaller size of tumor with ages above 50 years. Grade 2 patients appeared to outnumber both grade 1 and 3, which call for proper attention on those patients. For instance, using grade as a marker, it is very easy to classify grade 1 as low stage of the disease and grade 3 as high and clinicians will find it quiet easy selecting patients for treatment. The most difficult population will be grade 2 since they are heterogeneous which need other markers to differentiate this population into low and high risk, which was one of the motivation behind this study.
In dealing with grade 2 cancer patients for instance, other markers that are able to differentiate them are urgently needed and some of these markers that has been validated include Estrogen Receptor (ER), Progesterone Receptor (PR) status. The cellular micro environment of normal

and breast cancer tissue based on patient ID and age.

In **[18]** authors V. Suzanne Klimberg, improved the methods and cosmesis for bilateral total skin sparing mastectomy, even for salvage treatment after radiation. Prevention of local recurrence or a new occurrence, particularly of invasive cancer, which, in principle, could metastasize, is a laudable goal. Many times, the physician concentrates on survival while the patients live with the fear of local recurrence every day, which is exacerbated by the side effects of treatment and follow-up screenings. Despite the much higher risk of local recurrence with lumpectomy and radiation compared with mastectomy, multiple retrospective studies and prospective randomized trials have established that survival is the same; however, this may not be so for the individual patient, depending on the type of recurrence. In favorable tumors as an alternative to radiation, which adds an additional 1-cm sterilization of the tumor bed to the lumpectomy procedure (similar to radiation), while maintaining a pristine breast for salvage procedures, if necessary. Despite the much higher risk of local recurrence with lumpectomy and radiation compared with mastectomy, multiple retrospective studies and prospective randomized trials have established that survival is the same; however, this may not be so for the individual patient, depending on the type of recurrence.

In **[19]** authors Heather McArthur, An ongoing Clinical Challenge and Opportunity for Innovation. In this issue of ONCOLOGY, Dr. Zagar and colleagues review the epidemiology and biology of breast cancer brain metastases and provide an overview of the diagnosis, symptom management, and treatment issues for this entity. Specifically, the evidence for surgery, stereotactic radiosurgery (SRS), whole-brain radiotherapy (WBRT), and various systemic therapy options for breast cancer brain metastases is reviewed, and the need for a multidisciplinary approach is emphasized. This overview also highlights several unique challenges and opportunities germane to the management of breast cancer brain metastases. First, the growing burden of this devastating complication of breast cancer across all subtypes warrants emphasis. Specifically, it is estimated that as many as 24% to 34% of women with stage IV breast cancer now develop brain metastases.[2] Furthermore, women appear to be living longer after a diagnosis of breast cancer brain metastases, regardless of subtype.

For example, according to a retrospective, single-institution study of women with breast cancer brain metastases between 2003 and 2009, overall survival after systemic treatment and WBRT was 13 months for women with human epidermal growth factor receptor 2 (HER2)-positive breast cancer but only 4 months for women with triple-negative breast cancer (TNBC).
Brain metastasis remains a relatively common and particularly devastating complication of breast cancer and has proven a particularly challenging area for therapeutic innovation.

In **[20]** authors Danielle N. Sin and Julia A. Smith tested the Clinical Hereditary Cancer Syndromes and Gene Panel Testing. Hereditary cancer risk assessment has grown from a limited concentration on a few well-described tumors and genetic mutations to an ever-widening understanding of cancer syndromes encompassing multiple tumor types across multiple generations. The identification of an increasing number of genetic abnormalities and the advances in our understanding of the role of genetics in the risk of developing a cancer have created new opportunities to impact patient care. Primary prevention (the possibility of changing outcomes associated with risk) and secondary prevention (the ability to identify cancers at earlier stages, with the goal of intervening at a time at which outcomes can be changed in a meaningful way) are becoming more of a reality as it becomes possible to recognize hereditary cancer syndromes.
An example is the targeting of breast cancer patients with *BRCA* mutations for treatment with poly (ADP-ribose) polymerase (PARP) inhibitors. This is the tip of the iceberg. Targeted cancer therapy, determined by an understanding of the genetics of the particular tumors involved, is on the horizon. Moreover, it is not too far in the future that whole-genome testing will be a reality. For some time, genetic testing has been predictive and prognostic. It is now assuming a therapeutic role as well. An example is the targeting of breast cancer patients with BRCA mutations for treatment with PARP inhibitors.

In **[21]** authors Simon B. Zeichner and Christine Stanislaw and Jane L. Meisel, found the Prevention and Screening in Hereditary Breast and Ovarian Cancer. According to guidelines issued by the National Comprehensive Cancer Network (NCCN) and supported by other medical societies, genetic testing for hereditary breast and/or ovarian cancer syndromes is currently

recommended for the following patients:

- Those with breast cancer diagnosed at or before age 45 years.
- Those with triple-negative breast cancer diagnosed at or before age 60 years.
- Patients who have two or more primary breast cancers.
- Patients with invasive ovarian, fallopian tube, or primary peritoneal cancer.
- Male patients with breast cancer.
- Patients who have breast cancer and a first-, second-, or third-degree relative with breast cancer diagnosed before age 50 years.
- Patients with breast cancer who have two or more relatives with breast cancer diagnosed at any age.
- Patients with breast cancer who have a first-, second-, or third-degree relative with ovarian cancer.
- Patients with breast cancer who have a relative with male breast cancer.

In **[22]** authors Vijaya Krishna K. Gadi and Julie R. Gralow, Expansion of precision medicines approaches will play a key role in optimizing care and improving outcomes for breast cancer patients in 2017. The breakneck pace of genomic and targeted therapy discoveries of the last decade will continue to pay dividends in the near future in the form of expected reporting of several practice-changing clinical trials and new US Food and Drug Administration (FDA) drug approvals. Here we highlight some of the studies and approvals we are most anxious to see in the coming year. In metastatic breast cancer, the addition of the anti–human epidermal growth factor receptor 2 (HER2) monoclonal antibody pertuzumab has led to substantial improvements in progression-free and overall survival. For patients with advanced HER2-overexpressed breast cancer, a number of new pipeline agents will have further results reported in 2017. For triple-negative breast cancer (TNBC; ER-negative, progesterone receptor–negative, HER2-negative), the promising Trop-2–targeted antibody-drug conjugate sacituzumab govitecan (IMMU-132) has received FDA breakthrough and fast track designations. Although a meta-analysis of adjuvant bisphosphonates in early breast cancer reported a reduction in relapses and deaths in postmenopausal women, routine clinical uptake has not yet occurred in the United States.

In **[23]** author Robert E. Coleman, has created an Impact of Bone-Targeted Treatments on Skeletal Morbidity and Survival in Breast Cancer. Healthy bone is in a constant state of remodeling, with bone-derived osteoblasts and osteoclasts working together to preserve structural integrity and minimize the risk of fragility fractures. The causes of bone loss in cancer patients and the subsequent functional consequences are multifactorial. They include both the side effects of specific anticancer therapies and preexisting clinical risk factors for fracture, such as advanced age, use of glucocorticoids, a history of smoking, low body mass index, a family or personal history of fragility fracture, and low bone mineral density (BMD). Bone metastases are common in advanced breast cancer, and may be associated with serious morbidity, including fractures, pain, nerve compression, and hypercalcemia. Through optimum multidisciplinary management and the use of bone-targeted treatments, patients with advanced breast cancer have experienced a major reduction in skeletal complications, less bone pain, and an improved quality of life.

In **[24]** authors Triste S. Park and E. Shelley Hwang, has summarize the current trends in the Management of Ductal Carcinoma in Situ. Ductal carcinoma in situ (DCIS) was once an uncommon breast lesion, but now comprises 20% to 30% of new breast cancer diagnoses detected on mammography. DCIS is traditionally considered to be a precursor to invasive disease; however, this paradigm has been challenged by the fact that although the incidence and treatment of DCIS have risen, there has not been a concomitant decline in the incidence of invasive breast cancer. This review will summarize the current trends in the diagnosis and management of DCIS and will highlight ongoing trials that are shaping future management of this entity.

In **[25]** authors Timothy M. Zagar and AmandaE. D. VanSwearingen and Orit Kaidar – Person and Matthew G. Ewend and Carey K. Anders, has summarizes the most up-to-date approach to the multidisciplinary management of patients with breast cancer brain metastases. Brain metastases are a challenging consequence of advanced cancer.. Brain metastases should be suspected in a patient with a history of breast cancer of any stage who presents with unexplained neurologic symptoms.

2.3 FUNCTIONAL REQUIREMENTS

TEST DATA

For the extracted data to find accuracy and intercept we need to find the test data. Like it can contain all types of id, diagnosis, radius, texture, perimeter, area etc.,

PRE-PROCESSING

The data that it may contains some null values. So, the data that is collected in the test data is to processed and find the null values and removing the values.

TRAIN DATA

For the extracted data to find the accuracy an intercept we need to find the train data. Like it can contain all types of id, diagnosis, radius, texture, perimeter, area etc.,. The data is given by means of 1.0002 float values denote it is positive or negative values.

TRAIN PRE-PROCESSING DATA

The data that is collected from data set has lots of repeated words. So, all the forms of noisy data are removed. Finally, I have achieved a dataset without any repetition.

DATA EXTRACTION

The data that is needed to be extracted its name is to be entered. So, we get the polarity of the values by comparing it with the dataset we have trained.

2.4 NON-FUNCTIONAL REQUIREMENTS

FLEXIBILITY & SCALABILITY

High level language are flexible, which means they are able to run across all major hardware and software platforms with few or no change in source code. Python is flexible can be used on Linux, Windows. R Studio is flexible and also can be used on Linux, Windows and many more

FRAGMENTATION

Python gave the same environment which is open; the entire IDE's which is open to all the devices which reduces fragmentation. If you develop an python programs, it will run on all the devices.

OPEN SOURCE

Python is publicly available open source software; anyone can use source code that doesn't cost anything.

SCALABILITY

The system can be extended to integrate the modifications done in the present program to improve the quality of the product. This is meant for the future works that is to be done on the program.

RELIABLITY

The program is being developed through python, the most famous, efficient and reliable language, so it is reliable aspect until and unless there is an error in the programming side.

PORTABILITY

This System must be intuitive enough such that user with average background in using system can quickly experiment with the system and learn how to use the project. The system has user friendly interface.

FEASIBILITY STUDY

A key part of the preliminary investigation that reviews anticipated costs and benefits and recommends a course of action based on operational, technical, economic and time factors.

TECHNICAL FEASIBILITY

To develop this program, a high-speed internet connection and software are required. The current project is technically feasible as the program was successfully deployed on LENOVO having Windows 10 OS and also Intel I5 Processor.

BEHAVIOURAL FEASIBILITY

The program is behavioural feasible since it requires no technical guidance, all the modules are user friendly and execute in a manner they were designed to.

2.5 SYSTEM OVERVIEW

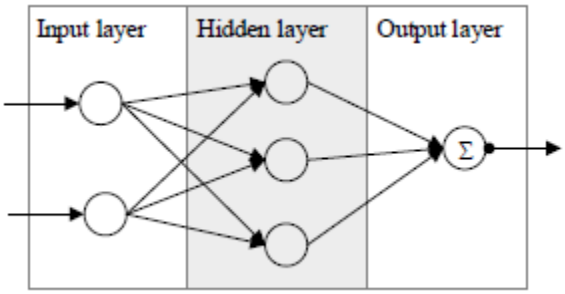
DEEP LEARNING TECHNIQUES

Deep learning (also known as deep structured learning or hierarchical learning) is part of a broader family of machine learning methods based on artificial neural networks. Learning can be supervised, semi-supervised or unsupervised. Deep learning is a class of machine learning algorithms that uses multiple layers to progressively extract higher level features from the raw input. For example, in image processing, lower layers may identify edges, while higher layers may identify the concepts relevant to a human such as digits or letters or faces. Deep learning architectures can be constructed with a greedy layer-by-layer method. Deep learning helps to disentangle these abstractions and pick out which features improve performance.
For supervised learning tasks, deep learning methods eliminate feature engineering, by translating the data into compact intermediate representations akin to principal components, and derive layered structures that remove redundancy in representation. Most modern deep learning models are based on artificial neural networks, specifically, Convolutional Neural Networks (CNN)s, although they can also include propositional formulas or latent variables organized layer-wise in deep generative models such as the nodes in deep belief networks and deep Boltzmann machines. Deep learning architectures such as deep neural networks, deep belief networks, recurrent neural networks and convolutional neural networks have been applied to fields including computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, bioinformatics, drug design, medical image analysis, material inspection and board game programs, where they have produced results comparable to and in some cases superior to human experts. Artificial Neural Networks (ANNs) were inspired by information processing and distributed communication nodes in biological systems. ANNs have various differences from biological brains. Specifically, neural networks tend to be static and symbolic, while the biological brain of most living organisms is dynamic (plastic) and analog.

NEURAL NETWORK ARCHITECTURE

Artificial neural networks (ANN) or connectionist systems are computing systems that are inspired by, but not identical to, biological neural networks that constitute animal brains. Such

systems "learn" to perform tasks by considering examples, generally without being programmed with task-specific rules. For example, in image recognition, they might learn to identify images that contain cats by analyzing example images that have been manually labeled as "cat" or "no cat" and using the results to identify cats in other images. They do this without any prior knowledge of cats, for example, that they have fur, tails, whiskers and cat-like faces. Instead, they automatically generate identifying characteristics from the examples that they process. In ANN implementations, the "signal" at a connection is a real number, and the output of each neuron is computed by some non-linear function of the sum of its inputs. The connections are called edges. Neurons and edges typically have a weight that adjusts as learning proceeds.



CONVOLUTIONAL NEURAL NETWORKS

Deep learning is a subfield of machine learning that is inspired by artificial neural networks, which in turn are inspired by biological neural networks. A specific kind of such a deep neural network is the convolutional network, which is commonly referred to as CNN or Convolutional Neural Network. It's a deep, feed-forward artificial neural network. Remember that feed-forward neural networks are also called multi-layer perceptrons (MLPs), which are the quintessential deep learning models. The models are called "feed-forward" because information flows right through the model. There are no feedback connections in which outputs of the model are fed back into itself.

SUPPORT VECTOR MACHINE

Support vector machine (SVM) was proposed by Vapnik in the 1970s by maximizing the margin between two hyperplanes of two clusters. SVM is widely applied in many areas, such as digit recognition, handwritten recognition, face detection, cancer classification, time series forecasting, etc.

Considering binary classification, the training set is classified as:

$$T = \{(x_i, y_i), i = 1, \ldots\ldots, N, x_i \epsilon R^M, y_i \epsilon \{1, -1\}\}$$

Where x_i denotes a M dimension feature vector of the i^{th} case, and y_i is class identifier.

Based on maximal margin of two classifier hyperplanes, SVM model is given as:

$$\min \frac{1}{2} w^T w + C \sum_{i=1}^{N} \varepsilon_i \qquad (1)$$

$$\text{s.t. } y_i(w^T x_i + b) \geq 1 - \varepsilon_i, \ i = 1, \ldots N \qquad (2)$$

$$\varepsilon_i > 0, \ I = 1, \ldots\ldots N \qquad (3)$$

where w and b indicate two parameters for separating hyperplanes, εi presents a slack variable to accommodate outliers and noise, C is used to balance the importance of maximizing the margin and satisfying the margin constraint for each data point.

When the hyperplane is linear, Equation (1) can be solved as quadratic optimization problem, but when the hyperplane is nonlinear, constraint (2) becomes nonlinear, thus, Equation (1) becomes more complex. For Equation (1), the dual form can be obtained, based on Lagrange duality and KKT condition, which is given as

$$\max_{\alpha} D(\alpha) = -\frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j) + \sum_{i=1}^{n} \alpha_i \qquad (4)$$

$$s.t. \sum_{i=1}^{n} \alpha_i y_i = 0, \qquad i = 1, \ldots, N, \qquad (5)$$

$$\alpha_i \geq 0, \qquad i = 1, \ldots, N, \qquad (6)$$

where $\propto_i$ denotes Lagrange multipliers, k(x_i, x_j) is kernel matrix which can map feature space into higher dimension space in order to transform nonlinear separable space into linear separable space. Kernel function is used to perform the feature mapping, among which commonly includes linear kernel function, quadratic kernel function, polynomial kernel function and Gaussian kernel function.

NAÏVE BAYES CLASSIFIER

Naive Bayes classifier is a type of probabilistic graphical model. The classifier is based on Bayes' Theorem (Equation (7)) which offers an explanation of the probability of associating certain classes at certain instances [12]. The Naive Bayes classifier model has an assumption that

features are conditionally independent. This model is given by

$$P(c_j|x_i) = \frac{P(x_i|c_j)P(c_j)}{\sum_k P(x_i|c_k)P(c_k)}, \quad (7)$$

$$P(x_i|c_j) = \prod_v P(x_{iv}|c_j), \quad (8)$$

where x_{iv} represents the v^{th} feature of the i^{th} instance, c_j is the identifier of j^{th} class. Then the class of instance x is classified by calculating:

AdaBoost

Boosting is a general method to improve the performance of learning process. Boosting is designed under the idea of gathering a set of "weak" learning algorithms into a "committee". A continuous adjustment of weight importance for each weak classifier model based on the training process leads to a continuous improvement in the overall learning algorithm performance. For a binary problem, given training set T, the adaptive boosting algorithm can be described as

Initialize: $D_1(i) = \frac{1}{m}$ for $i = 1, \dots, m$

For $t = 1, \dots, T$

Train decision tree using distribution D_t

Select h_t to minimalize the weighted error $\epsilon_t = P(h_t(x_i \neq y_i)$

Choose $\alpha_t = \frac{1}{2}\ln(\frac{1-\epsilon_t}{\epsilon_t})$

Update for $i = 1, \dots, m$

$$D_{t+1}(i) = \frac{D_t(i)}{z_t}\begin{cases} e^{-\alpha_t} & if\ h_t(x_i) = y_i \\ e^{\alpha_t} & if\ h_t(x_i) \neq y_i \end{cases}$$

$$= \frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{z_t}\begin{cases} e^{-\alpha_t} & if\ h_t(x_i) = y_i \\ e^{\alpha_t} & if\ h_t(x_i) \neq y_i \end{cases}$$

Get final classifier

$$H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$$

K – FOLD CROSS – VALIDATION

For a classification problem, test error is the most important measure to evaluate classification performance. However, when a data set is given, all the data are expected to be used to train the model, which can enhance the model to obtain a more reliable learning result. After producing the model, it is expected to use a very large designated test set to directly estimate the test error. In the absence of a very large test set, a number of statistic techniques are proposed to estimate

test error. K-fold cross-validation is applied in this research to measure the performance of each model. This method starts by randomly separating the data set into k groups with equal size. During the process, each fold is considered as a testing set and the remaining groups are considered as the training set. For each iteration, the test error is calculated

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} Err_i, \tag{10}$$

$$Err_i = I(y_i \neq \hat{y}_i). \tag{11}$$

Chapter 3

SYSTEM DESIGN

PRE – PROCESSING

The pre-processing is the most important step in the mammogram analysis due to poor captured mammogram image quality. Pre-processing is very important to correct and adjust the mammogram image for further study and processing. There are Different types of filtering techniques are available for pre-processing. This filters used to improve image quality, remove the noise, preserves the edges within an image, enhance and smoothen the image. In this paper, we have performed various filters namely, average filter, adaptive median filter, average or mean filter, and wiener filter.

The steps to be taken in:

- Mean Filter or Average Filter
- Median Filtering
- Adaptive Median Filter

FUTURE SELECTION

It has been an attempt over last many years that how to detect and cure cancer. Cancer which can be of various types like breast cancer, lung cancer, throat cancer, blood cancer etc. is known to be the deadliest disease which still now have not got any cure

There are several levels of cancer from 1 to 6. However, on the good side if cancer is detected when it is in level 1 or 2 or at the very initial stage, there is a significant probability that it will get cured within a period of time. With the advent of new technologies in the field of medicine, we get new ideas of curing the disease with methods like machine learning. The problem for the cancer can be broadly classified into three types.

1. **Firstly,** the problem is to predict whether a person in a particular stage of cancer has the chance to survive or not.

2. **Secondly,** the problem is to predict whether a person who has already encountered the disease in the past and got cured, has the probability of having the same disease in future.
3. **Thirdly,** the domain in which we are working includes the detection of cancer at the earliest stage.

CLASSIFICATION

Classification of benign tumors can help the patients avoid undertaking needless treatments. Most of them show that classification techniques give a good accuracy in prediction of the type of tumor. Our methodology involves use of classification techniques like Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), Logistic Regression, with Dimensionality Reduction technique i.e. Principal Component Analysis (PCA) .
A classification problem is when the result is a category like filtering emails “spam” or “not spam”. Unsupervised Learning: Unsupervised learning is giving away information to the machine that is neither classified nor labeled and allowing the algorithm to analyze the given information without providing any directions. In unsupervised learning algorithm the machine is trained from the data which is not labeled or classified making the algorithm to work without proper instructions. In our dataset we have the outcome variable or Dependent variable i.e. Y having only two set of values, either M (Malign) or B (Benign). So Classification algorithm of supervised learning is applied on it. We have chosen three different types of classification algorithms in Machine Learning.

Classification Techniques:

- **Logistic Regression:**
  - Logistic Regression is a supervised machine learning technique, employed in classification jobs (for predictions based on training data).
  - Logistic Regression uses an equation similar to Linear Regression but the outcome of logistic regression is a categorical variable whereas it is a value for other regression models.

- Binary outcomes can be predicted from the independent variables. The outcome of dependent variable is discrete. Logistic Regression uses a simple equation which shows the linear relation between the independent variables.

- **K-NN (K-Nearest Neighbor):**
  - K-Nearest Neighbor is a supervised machine learning algorithm as the data given to it is labeled. It is a non-parametric method as the classification of test data point relies upon the nearest training data points rather than considering the dimensions (parameters) of the dataset.
  - In Classification technique, it classifies the objects based on the k closest training examples in the feature space.
  - It catches the idea of proximity based on mathematical formula called as Euclidean distance, calculation of distance between two points in a plane.
- **SVM (Support Vector Machine):**
  - Support Vector Machine is a supervised machine learning algorithm which is doing well in pattern recognition problems and it is used as a training algorithm for studying classification and regression rules from data.
  - SVM is most precisely used when the number of features and number of instances are high. A binary classifier is built by the SVM algorithm.
  - This binary classifier is constructed using a hyper plane where it is a line in more than 3 – dimensions.

Classification states:

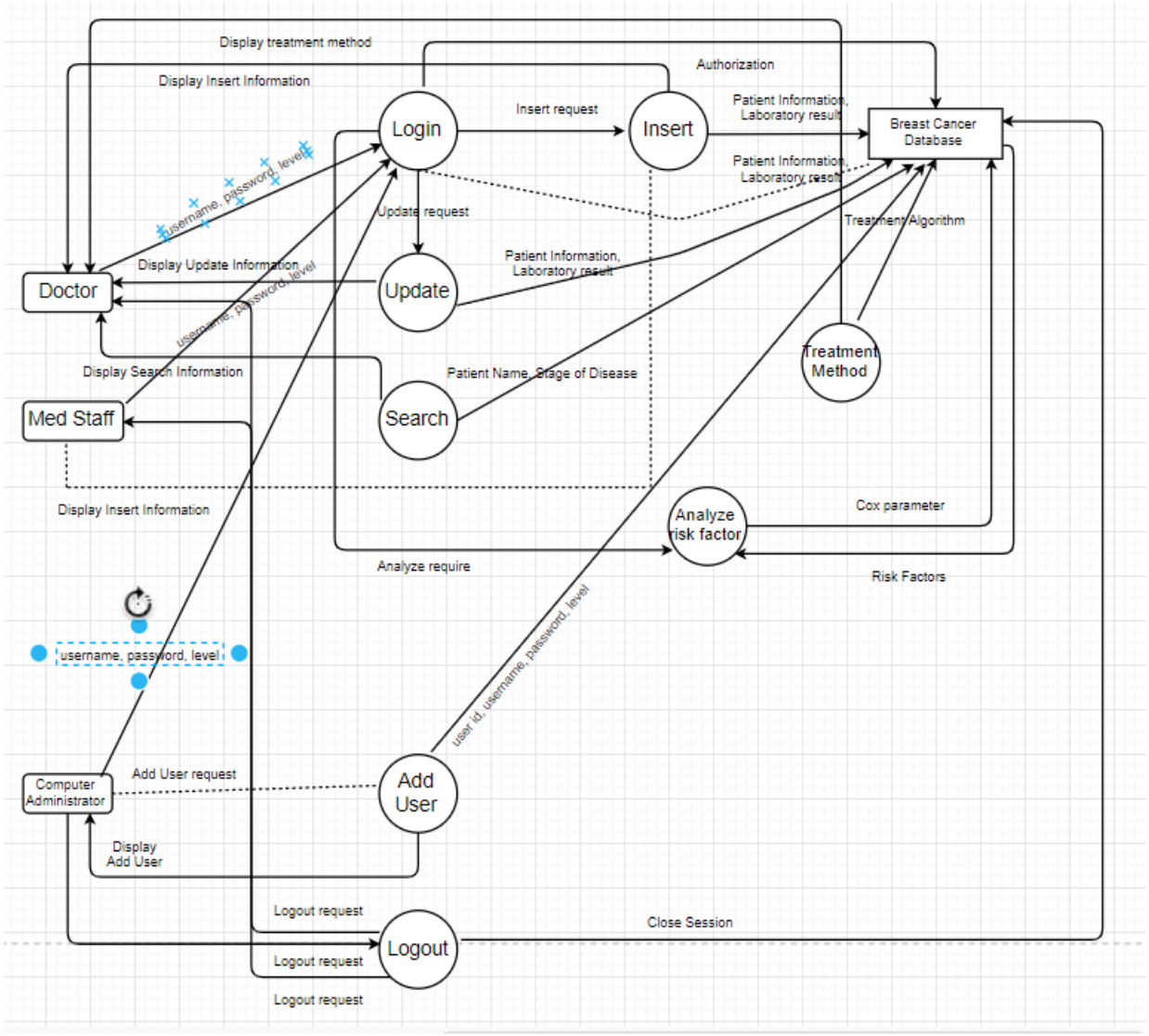
1. **Benign**

   Benign is a state of tumor or lump which does not spread across body. In simple terms these are non-cancerous type of tumor.

2. **Malignant**

   Malignant is a state of tumor or lump which spreads across the body. In simple terms these are cancerous type of tumor.

3.2 SYSTEM ARCHITECTURE



3.3 SYSTEM CONFIGURATION

- HARDWARE REQUIREMENTS

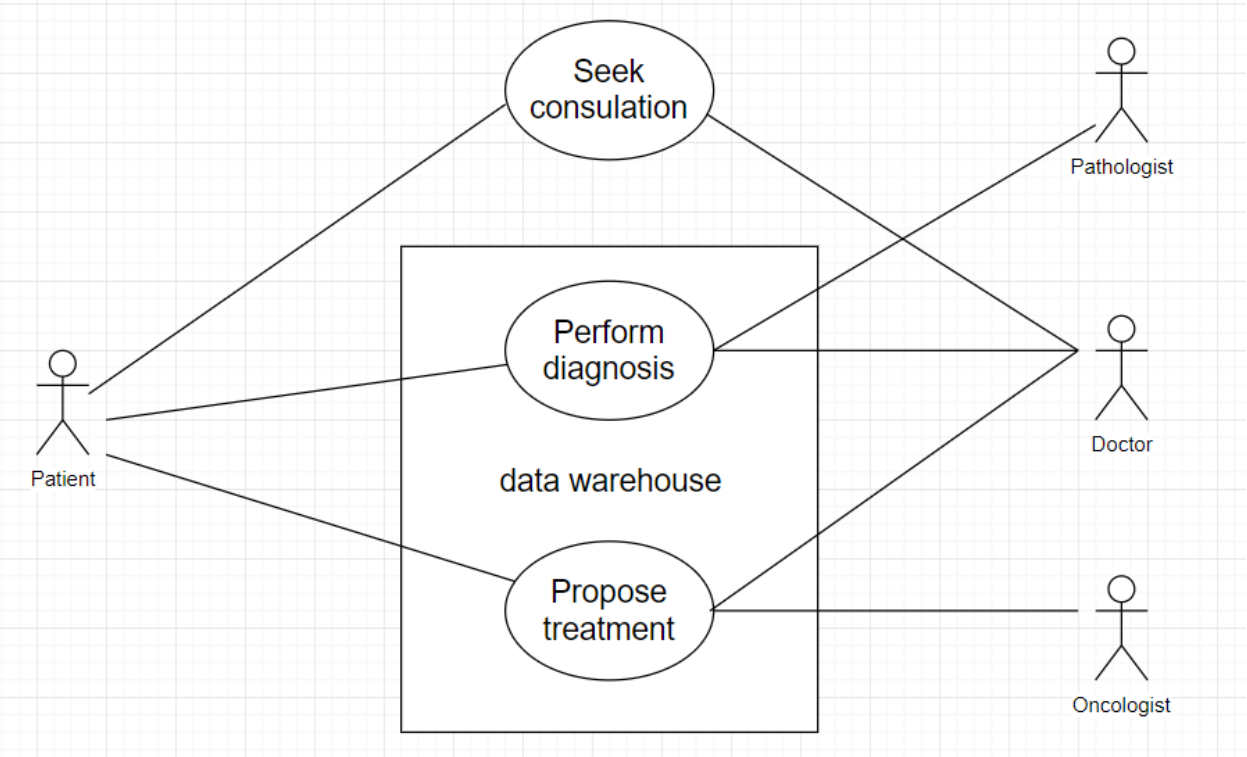
System	:	Pentium Dual Core
System type	:	64 bit
Processor (Gen)	:	Any Processor
Hard Disk	:	Min 100 GB
Ram	:	Min 4GB
Speed	:	1.1 GHz

- SOFTWARE REQUIREMENTS

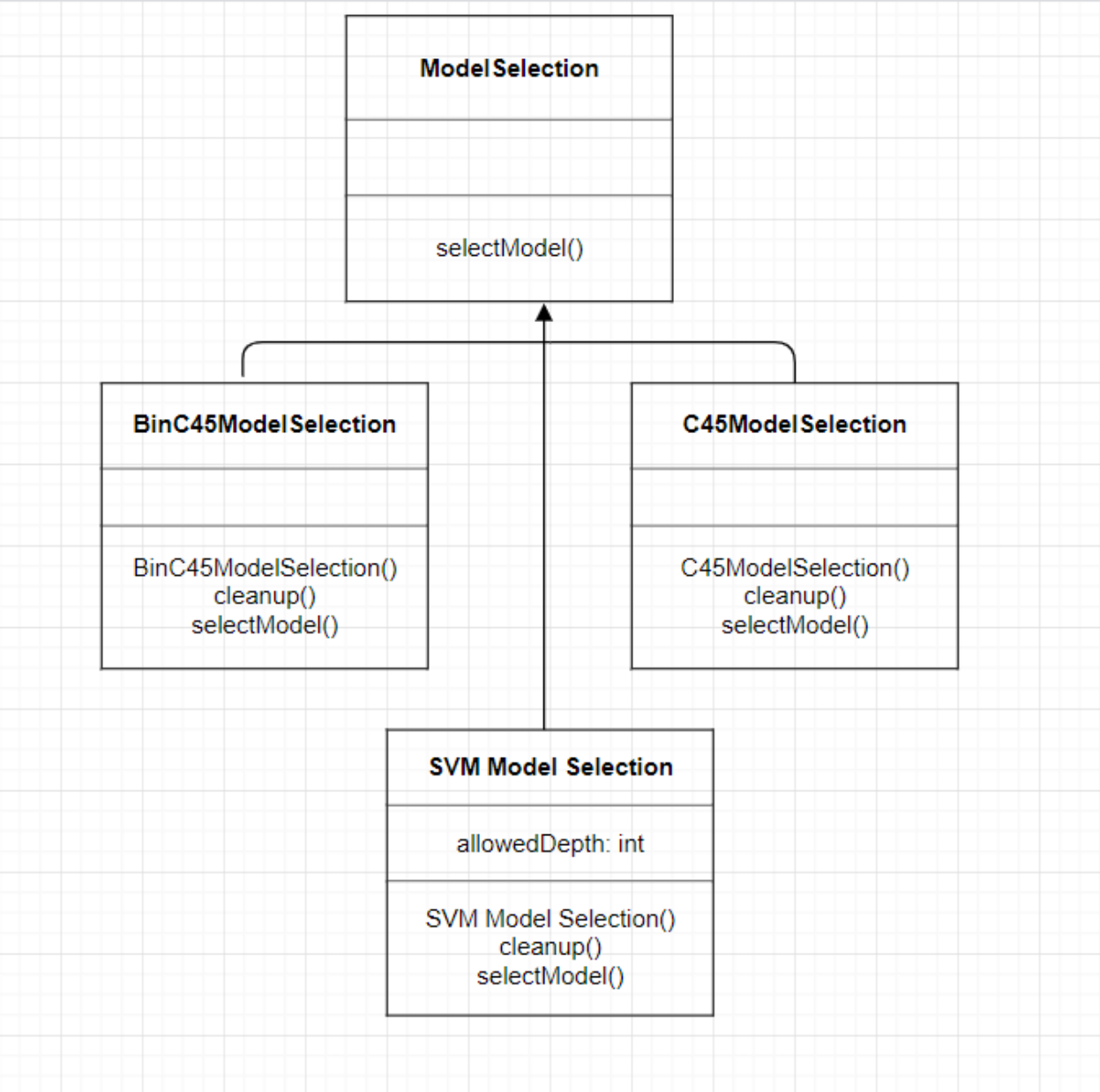
Operating system	:	Windows
Coding Language	:	PYTHON
UML's	:	StarUml
Python version	:	3.6
Other tools	:	RStudio

3.4 UML DIAGRAMS

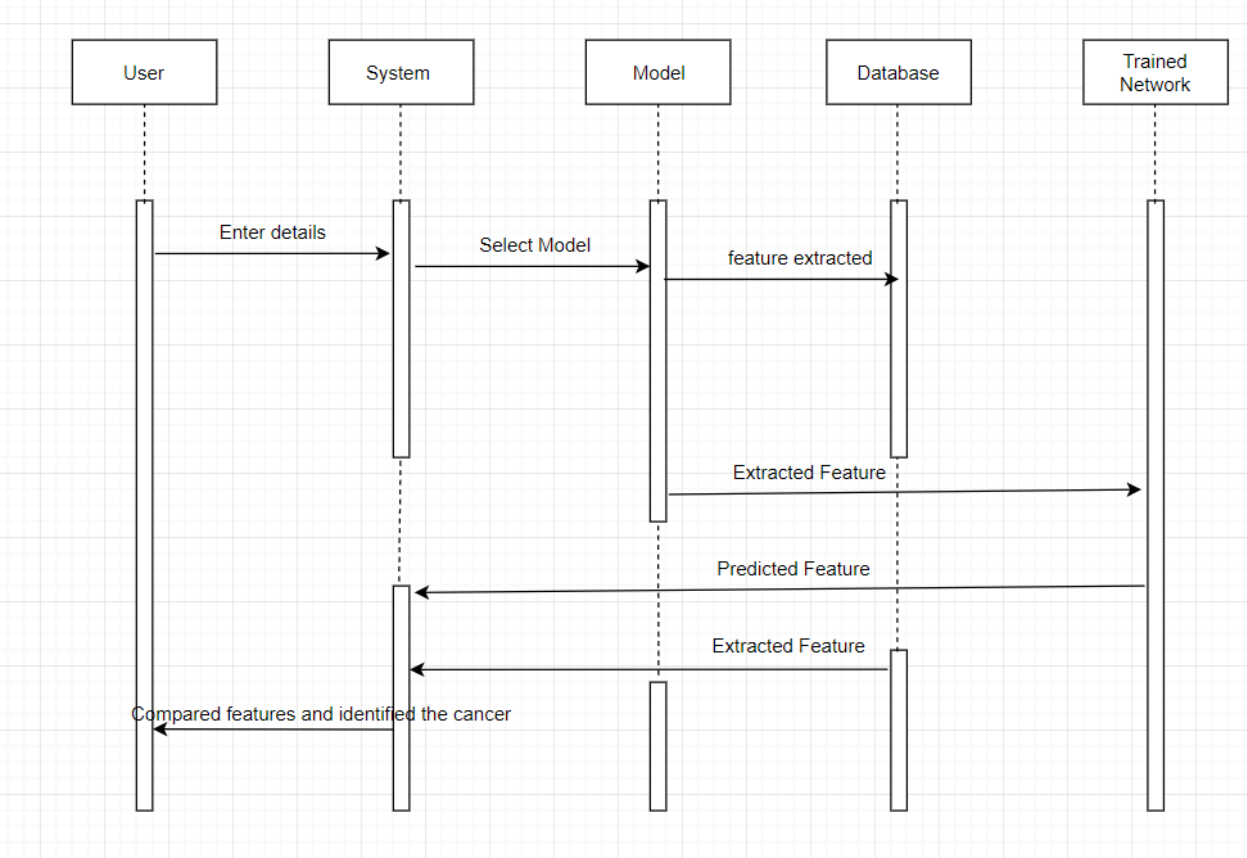
3.4.1 USE CASE DIAGRAM



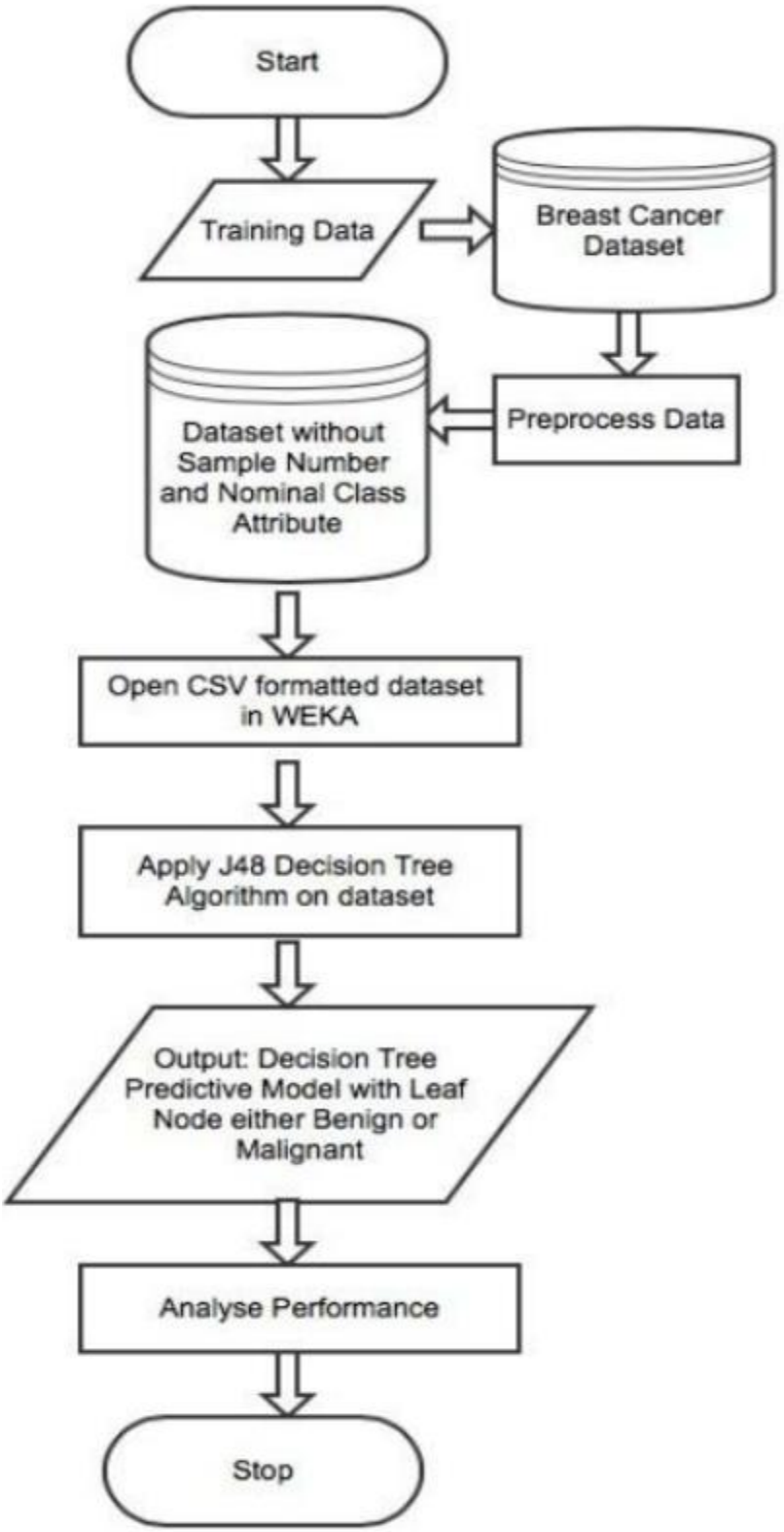
3.4.2 CLASS DIAGRAM



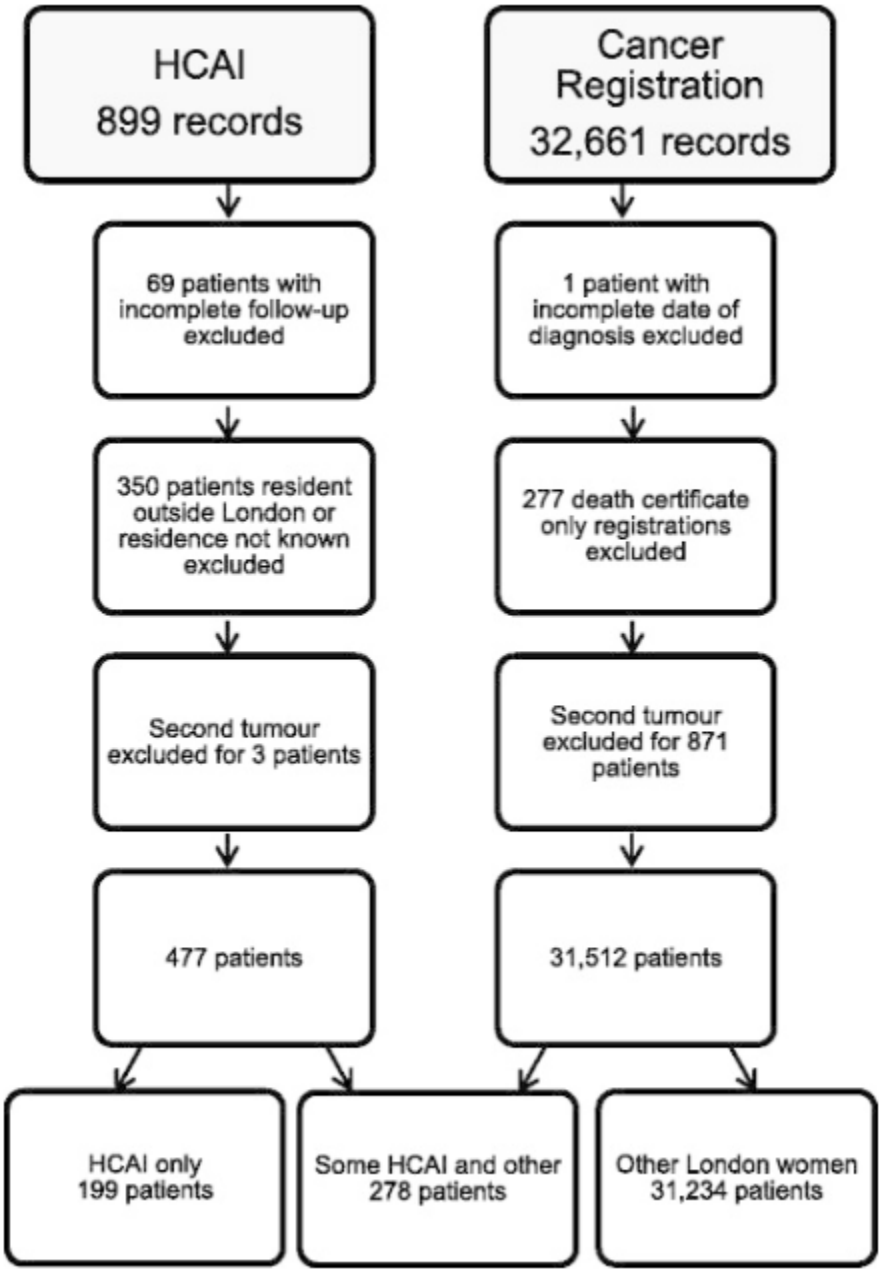
3.4.3 SEQUENCE DIAGRAM



3.4.4 ACTIVITY DIAGRAM



3.4.5 DATAFLOW DIAGRAM



Chapter 4

IMPLEMENTATION AND TESTING

4.1 IMPLEMENTATION

The Implementation in Phase where the data is downloaded. First of all I am importing the data of pandas as pd. Later I am reading the data from the resource folder and making the header values of 0 to n for rows and columns

Secondly, I am pre-processing the data it helps me to remove the null values with the help of importing the label of Label Encoder. And relocate the pre-processing data of the Decision tree variables of 1 and 2 for benign as 1 and malignant as 2.

Next, I am selecting the model using sklearn and train the data using splitting of Decision tree of benign and malignant. After pre-processing the data by using the import statement of decomposition of linear model finding the Test Accuracy of Logistic Regression.

Later, Importing the numpy and matplotlib for plotting the graph in X and Y-axis by using any model and importing linear curve statement by verifying the data in linear curve design. I am training the linear curve graph using train sizes, train mean, color, marker, marker size of plot and fill between and Accuracy of the model. Differentiating the X and Y axis as Accuracy and Number of training samples. After that I am validating the Curve. Then find the linear curve in between Accuracy and Parameter C.

Then finding the Accuracy for all models similarly K-NN and SVM and Random Forest etc. After I am finding the predicted label of confusion matric for all models

Finally, you will get a graph and plot and histogram and indicating the accuracy levels in between support vector machine and naïve ayes algorithm. Therefore, personalized recommendation would lose the confidence and support of the users. Infact users concerns plays a major barrier for the development of the project on personalized recommendation.

4.2 TESTING

In general, testing is finding out how well something works. In terms of human beings, testing tells what level of knowledge or skill has been acquired. In computer hardware and software development, testing is used at key checkpoints in the overall process to determine whether objectives are being met. There are various types of test. Each test type addresses a specific testing requirement.

The following are the Types of Testing:

- Unit Testing
- Integration Testing
- User Acceptance Testing
- Output Testing
- Functional Testing
- System Testing
- White Box Testing
- Black Box Testing
- Validation Testing

4.2.1 UNIT TESTING

Testing of an individual software component or module is termed as Unit Testing. It is typically done by the programmer and not by testers, as it requires detailed knowledge of the internal program design and code. It may also require developing test driver modules or test harnesses. Unit tests are typically automated tests written and run by software developers to ensure that a section of an application (known as the "unit") meets its design and behaves as intended.

During development, a software developer may code criteria, or results that are known to be good, into the test to verify the unit's correctness. During test case execution, frameworks log tests that fail any criterion and report them in a summary.

4.2.2 INTEGRATION TESTING

Testing of all integrated modules to verify the combined functionality after integration is termed as Integration Testing. Modules are typically code modules, individual applications, client and server applications on a network, etc. This type of testing is especially relevant to client/server and distributed systems. It is the phase in software testing in which individual software modules are combined and tested as a group. Integration testing is conducted to evaluate the compliance of a system or component with specified functional requirements.
In the big-bang approach, most of the developed modules are coupled together to form a complete software system or major part of the system and then used for integration testing. This method is very effective for saving time in the integration testing process. However, if the test cases and their results are not recorded properly, the entire integration process will be more complicated and may prevent the testing team from achieving the goal of integration testing.

4.2.2.1 TOP DOWN INTEGRATION

Top-down integration testing is an integration testing technique used in order to simulate the behaviour of the lower-level modules that are not yet integrated. Stubs are the modules that act as temporary replacement for a called module and give the same output as that of the actual product.

The replacement for the 'called' modules is known as 'Stubs' and is also used when the software needs to interact with an external system

4.2.2.2 BOTTOM-UP INTEGRATION

Each component at lower hierarchy is tested individually and then the components that rely upon these components are tested.

4.2.3 USER ACCEPTANCE TESTING

An Acceptance Test is performed by the client and verifies whether the end to end the flow of the system is as per the business requirements or not and if it is as per the needs of the end-user. Client accepts the software only when all the features and functionalities work as expected. It is the last phase of the testing, after which the software goes into production. This is also called User Acceptance Testing. Acceptance testing is a test conducted to determine if the requirements of a specification or contract are met. It may involve chemical tests, physical tests, or performance tests.

Formal testing with respect to user needs, requirements, and business processes conducted to determine whether a system satisfies the acceptance criteria and to enable the user, customers or other authorized entity to determine whether to accept the system. User Acceptance testing is also known as end-user testing, operational acceptance testing, acceptance test driven development, field testing.

4.2.4 OUTPUT TESTING

It is a type of software testing whereby the system is tested against the functional requirements/specifications. Functions (or features) are tested by feeding them input and examining the output. Determine the output based on the function's specifications. Execute the test case.

4.2.5 FUNCTIONAL TESTING

Functional Testing is a testing technique that is used to test the features/functionality of the system or Software, should cover all the scenarios including failure paths and boundary cases. Functional testing types includes of all these types of testing from Unit Testing to Beta testing

4.2.6 SYSTEM TESTING

System Testing is a type of software testing that is performed on a complete integrated system to evaluate the compliance of the system with the corresponding requirements.

In system testing, integration testing passed components are taken as input. The goal of

integration testing is to detect any irregularity between the units that are integrated together. System testing detects defects within both the integrated units and the whole system. The result of system testing is the observed behavior of a component or a system when it is tested.

System Testing is carried out on the whole system in the context of either system requirement specifications or functional requirement specifications or in the context of both. System testing tests the design and behavior of the system and also the expectations of the customer. It is performed to test the system beyond the bounds mentioned in the software requirements specification (SRS).

System Testing is basically performed by a testing team that is independent of the development team that helps to test the quality of the system impartial. It has both functional and non-functional testing. System testing is a black-box testing.

System testing is performed after the integration testing and before the acceptance testing

4.2.7 WHITE BOX TESTING

White box testing techniques analyze the internal structures the used data structures, internal design, code structure and the working of the software rather than just the functionality as in black box testing. It is also called glass box testing or clear box testing or structural testing.

Working process of white box testing:

Input: Requirements, Functional specifications, design documents, source code.

Processing: Performing risk analysis for guiding through the entire process.

Proper test planning: Designing test cases so as to cover entire code. Execute rinse-repeat until error-free software is reached. Also, the results are communicated.

Output: Preparing final report of the entire testing process.

4.2.8 BLACK BOX TESTING

Black box testing is a type of software testing in which the functionality of the software is not known. The testing is done without the internal knowledge of the products.

Black box testing can be done in following ways:

1. Syntax Driven Testing: This type of testing is applied to systems that can be syntactically represented by some language. For example- compilers, language that can be represented by context free grammar. In this, the test cases are generated so that each grammar rule is used at least once.

2. Equivalence partitioning: It is often seen that many type of inputs work similarly so instead of giving all of them separately we can group them together and test only one input of each group. The idea is to partition the input domain of the system into a number of equivalence classes such that each member of class works in a similar way, i.e., if a test case in one class results in some error, other members of class would also result into same error.

4.2.9 VALIDATION TESTING

Validation checks are performed on the following fields.

Validation testing is the process of ensuring if the tested and developed software satisfies the client /user needs. The business requirement logic or scenarios have to be tested in detail. All the critical functionalities of an application must be tested here.

As a tester, it is always important to know how to verify the business logic or scenarios that are given to you. One such method that helps in detail evaluation of the functionalities is the Validation Process.

Whenever you are asked to perform a validation test, it takes a great responsibility as you need to test all the critical business requirements based on the user needs. There should not be even a single miss on the requirements asked by the user. Hence a keen knowledge on validation testing is much important.

4.3 TEST CASES

TEST CASE 1	
Test Case Name	Empty extract count fields testing
Description	In the output screen if the count asked for the number of users
Output	The output fails showing during handling of the above exception another exception occurred i.e., failed to establish a new connection.

TEST CASE 2	
Test Case Name	Empty extract fields testing
Description	In the output screen if the user extract field is empty
Output	User extraction fails showing that get address information has extracted

4.3.1 TESTING STRATEGY

A strategy for system Testing is a type of software testing that is performed on a complete integrated system to evaluate the compliance of the system with the corresponding requirements.

In system testing, integration testing passed components are taken as input. The goal of integration testing is to detect any irregularity between the units that are integrated together. System testing detects defects within both the integrated units and the whole system. The result of system testing is the observed behavior of a component or a system when it is tested.

4.3.2 USER TRAINING

For this purpose, the normal working of the project was demonstrated to the prospective users. Its working is easily understandable and since the expected users are people who have good knowledge of computers, the use of this system is very easy. Whenever a new system is developed, user training is required to educate them about the working of the system so that it can be put to efficient use by those for whom the system has been primarily designed.

4.3.3 MAINTAINANCE

Maintenance therapy is the treatment of cancer with medication, typically following an initial round of treatment. Maintenance treatment may include chemotherapy, hormonal therapy, or targeted therapy. To prevent or delay the cancer's return if the cancer is in complete remission after the initial treatment. "Complete remission" means that the doctors cannot find cancer and you have no symptoms. To slow the growth of advanced cancer after the initial treatment. This can help shrink the cancer, which is called a partial remission. In this situation, maintenance therapy is not used to cure the cancer, but it can lengthen a person's life.

Chapter 5

SAMPLE CODE

PARAMETERS CODE

```
import os
import numpy as np
import pandas as pd
import seaborn as sns
import datetime as dt
import matplatlib.pyplot as plt
from sklearn import preprocessing
from sklearn.model selection import train_test_split
from sklearn.metrics import confusion_matrix
```

READING THE DATA

```
data = pd.read_csv('C:/Users/Gurramkonda Bhargav/Desktop/breastcancer.csv')
```

DATA FRAME

```
print("\n \t The data frame has {0[0]} rows and {0[1]} columns. \n".format(data.shape))
data.info()
```

CHECKING THE FIRST FEW ROWS OF THE DATA

```
data.head()
```

TRANSFORMS THE DATA

```
from sklearn.preprocessing import LabelEncoder
X = df.loc[:, 2:].values
y = df.loc[:, 1].values
le = LabelEncoder()
y = le.fit_transform(y)


le.transform(['M', 'B'])
```

SELECTING THE MODEL

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test =  train_test_split(X, y, test_size=0.20, random_state=1)
```

SCIKIT LEARN TRANSFORMERS

```
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
#estimator
from sklearn.linear_model import LogisticRegression
from sklearn.pipeline import Pipeline
pipe_lr = Pipeline([('scl', StandardScaler()),
                    ('pca', PCA(n_components=2)),
                    ('clf', LogisticRegression(random_state=1))])
pipe_lr.fit(X_train, y_train)
print('Test Accuracy: %.3f' % pipe_lr.score(X_test, y_test))
```

MATPLOT THE GRAPH

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import learning_curve

pipe_lr = Pipeline([('scl', StandardScaler()),
                    ('clf', LogisticRegression(penalty='l2', random_state=0))])
train_sizes, train_scores, test_scores =learning_curve(estimator=pipe_lr,
                     X=X_train,
                     y=y_train,
                     train_sizes=np.linspace(0.1, 1.0, 10),
                     cv=10,
                      n_jobs=1)
train_mean = np.mean(train_scores, axis=1)
train_std = np.std(train_scores, axis=1)
```

```
test_mean = np.mean(test_scores, axis=1)
test_std = np.std(test_scores, axis=1)
import matplotlib.pyplot as plt
plt.rcParams['figure.figsize'] = (8, 5)
plt.rcParams['axes.titlesize'] = 'large'
plt.plot(train_sizes, train_mean,
        color='blue', marker='o',
        markersize=5,
         label='training accuracy')
plt.fill_between(train_sizes,
                 train_mean + train_std,
                 train_mean - train_std,
                 alpha=0.15, color='blue')
plt.plot(train_sizes, test_mean,
         color='green', linestyle='--',
         marker='s', markersize=5,
        label='validation accuracy')
plt.fill_between(train_sizes,
                 test_mean + test_std,
                  test_mean - test_std,
                  alpha=0.15, color='green')
plt.grid()
plt.xlabel('Number of training samples')
plt.ylabel('Accuracy')
plt.legend(loc='lower right')
plt.ylim([0.9, 1.0])
plt.show()
```

VALIDATION OF CURVE

```
from sklearn.learning_curve import validation_curve
param_range = [0.001, 0.01, 0.1, 1.0, 10.0, 100.0]
train_scores, test_scores = validation_curve(
             estimator=pipe_lr,
              X=X_train,
             y=y_train,
             param_name='clf__C',
            param_range=param_range,
             cv=10)
train_mean = np.mean(train_scores, axis=1)
train_std = np.std(train_scores, axis=1)
test_mean = np.mean(test_scores, axis=1)
test_std = np.std(test_scores, axis=1)
plt.plot(param_range, train_mean,
        color='blue', marker='o',
        markersize=5,
        label='training accuracy')
plt.fill_between(param_range, train_mean + train_std,
                 train_mean - train_std, alpha=0.15,
                color='blue')
plt.plot(param_range, test_mean,
         color='green', linestyle='--',
         marker='s', markersize=5,
        label='validation accuracy')
plt.fill_between(param_range,
                 test_mean + test_std,
                test_mean - test_std,
                alpha=0.15, color='green')
```

```
plt.grid()
plt.xscale('log')
plt.legend(loc='lower right')
plt.xlabel('Parameter C')
plt.ylabel('Accuracy')
plt.ylim([0.92, 1.0])
plt.show()
from sklearn.model_selection import GridSearchCV
from sklearn.svm import SVC

pipe_svc = Pipeline([('scl', StandardScaler()),
                     ('clf', SVC(random_state=1))])

param_range = [0.0001, 0.001, 0.01, 0.1, 1.0, 10.0, 100.0, 1000.0]
param_grid = [{'clf__C': param_range,
               'clf__kernel': ['linear']},
              {'clf__C': param_range,
               'clf__gamma': param_range,
               'clf__kernel': ['rbf']}]

gs = GridSearchCV(estimator=pipe_svc,
                  param_grid=param_grid,
                  scoring='accuracy',
                  cv=10,
                  n_jobs=-1)
gs = gs.fit(X_train, y_train)
print("Best: %f using %s" % (gs.best_score_, gs.best_params_))
```

ESTIMATE THE PERFORMANCE OF THE BEST SELECTED MODE

```
clf = gs.best_estimator_
```

```
clf.fit(X_train, y_train)
print('Test accuracy: %.3f' % clf.score(X_test, y_test))
```

ACCURACY FOR CROSS VALIDATION

```
from sklearn.model_selection import cross_val_score
from sklearn.tree import DecisionTreeClassifier
gs = GridSearchCV( estimator=DecisionTreeClassifier(random_state=0),
      param_grid=[ {'max_depth': [1, 2, 3, 4, 5, 6, 7, None]}],
                scoring='accuracy',
                cv=5)


scores = cross_val_score(gs,X_train,  y_train, scoring='accuracy',cv=2)
print('CV accuracy: %.3f +/- %.3f' % (np.mean(scores), np.std(scores)))
from sklearn.metrics import confusion_matrix
pipe_svc.fit(X_train, y_train)
y_pred = pipe_svc.predict(X_test)
confmat = confusion_matrix(y_true=y_test, y_pred=y_pred)
print(confmat)
```

PREDICTED LABEL OF CONFUSION MATRIX FOR CROSS VALIDATION

```
fig, ax = plt.subplots(figsize=(5.5, 5.5))
ax.matshow(confmat, cmap=plt.cm.Blues, alpha=0.3)
for i in range(confmat.shape[0]):
   for j in range(confmat.shape[1]):
      ax.text(x=j, y=i,
         s=confmat[i, j],
         va='center', ha='center')
plt.xlabel('predicted label')
plt.ylabel('true label')
plt.show()
```

```
from sklearn.metrics import classification_report
from sklearn.metrics import accuracy_score
from sklearn.metrics import precision_score
from sklearn.metrics import recall_score, f1_score
print(accuracy_score(y_test, y_pred))
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
pipe_lr = Pipeline([('scl', StandardScaler()),
                    ('pca', PCA(n_components=2)),
                    ('clf', LogisticRegression(penalty='l2',
                                               random_state=0,
                                               C=100.0))])

X_train2 = X_train[:, [4, 14]]
cv = list(StratifiedKFold(n_splits=3, random_state=1).split(X_train, y_train))
pipe_lr = pipe_lr.fit(X_train2, y_train)
y_pred2 = pipe_lr.predict(X_test[:, [4, 14]])

from sklearn.metrics import roc_auc_score
from sklearn.metrics import accuracy_score
print('ROC AUC: %.3f' % roc_auc_score( y_true=y_test, y_score=y_pred2))
print('Accuracy: %.3f' % accuracy_score( y_true=y_test, y_pred=y_pred2))
```

CLASSIFICATION TEHNIQUE:

```
diagnosis_all = list(data.shape)[0]
diagnosis_categories = list(data['diagnosis'].value_counts())
print("\n \t The data has {} diagnosis, {} malignant and {} benign.".format(diagnosis_all,
                                                                            diagnosis_categories[0],
                                                                            diagnosis_categories[1]))
features_mean= list(data.columns[1:11])
```

```
plt.figure(figsize=(10,10))
sns.heatmap(data[features_mean].corr(), annot=True, square=True, cmap='coolwarm')
plt.show()
```

SCATTER MATRIX

```
color_dic = {'M':'red', 'B':'blue'}
colors = data['diagnosis'].map(lambda x: color_dic.get(x))
sm = pd.scatter_matrix(data[features_mean], c=colors, alpha=0.4, figsize=((15,15)));
plt.show()
```

TIGHT LAYOUT

```
bins = 12
plt.figure(figsize=(15,15))
for i, feature in enumerate(features_mean):
    rows = int(len(features_mean)/2)
    plt.subplot(rows, 2, i+1)
    sns.distplot(data[data['diagnosis']=='M'][feature], bins=bins, color='red', label='M');
    sns.distplot(data[data['diagnosis']=='B'][feature], bins=bins, color='blue', label='B');
    plt.legend(loc='upper right')
plt.tight_layout()
plt.show()
```

ACCURACY

```
from sklearn.linear_model import SGDClassifier

start = time.time()

clf = SGDClassifier()
clf.fit(X_train, y_train)
prediction = clf.predict(X_test)
scores = cross_val_score(clf, X, y, cv=5)
```

```
end = time.time()

accuracy_all.append(accuracy_score(prediction, y_test))

cvs_all.append(np.mean(scores))

print("SGD Classifier Accuracy: {0:.2%}".format(accuracy_score(prediction, y_test)))

print("Cross validation score: {0:.2%} (+/- {1:.2%})".format(np.mean(scores),
np.std(scores)*2))

print("Execution time: {0:.5} seconds \n".format(end-start))

from sklearn.svm import SVC, NuSVC, LinearSVC

start = time.time()

clf = SVC()

clf.fit(X_train, y_train)

prediction = clf.predict(X_test)

scores = cross_val_score(clf, X, y, cv=5)

end = time.time()

accuracy_all.append(accuracy_score(prediction, y_test))

cvs_all.append(np.mean(scores))

print("SVC Accuracy: {0:.2%}".format(accuracy_score(prediction, y_test)))

print("Cross validation score: {0:.2%} (+/- {1:.2%})".format(np.mean(scores),
np.std(scores)*2))

print("Execution time: {0:.5} seconds \n".format(end-start))

start = time.time()

clf = NuSVC()

clf.fit(X_train, y_train)

prediciton = clf.predict(X_test)

scores = cross_val_score(clf, X, y, cv=5)

end = time.time()

accuracy_all.append(accuracy_score(prediction, y_test))

cvs_all.append(np.mean(scores))

print("NuSVC Accuracy: {0:.2%}".format(accuracy_score(prediction, y_test)))
```

```
print("Cross validation score: {0:.2%} (+/- {1:.2%})".format(np.mean(scores),
np.std(scores)*2))
print("Execution time: {0:.5} seconds \n".format(end-start))
start = time.time()
clf = LinearSVC()
clf.fit(X_train, y_train)
prediction = clf.predict(X_test)
scores = cross_val_score(clf, X, y, cv=5)
end = time.time()
accuracy_all.append(accuracy_score(prediction, y_test))
cvs_all.append(np.mean(scores))
print("LinearSVC Accuracy: {0:.2%}".format(accuracy_score(prediction, y_test)))
print("Cross validation score: {0:.2%} (+/- {1:.2%})".format(np.mean(scores),
np.std(scores)*2))
print("Execution time: {0:.5} seconds \n".format(end-start))
from sklearn.neighbors import KNeighborsClassifier
start = time.time()
clf = KNeighborsClassifier()
clf.fit(X_train, y_train)
prediction = clf.predict(X_test)
scores = cross_val_score(clf, X, y, cv=5)
end = time.time()
accuracy_all.append(accuracy_score(prediction, y_test))
cvs_all.append(np.mean(scores))
print("Accuracy: {0:.2%}".format(accuracy_score(prediction, y_test)))
print("Cross validation score: {0:.2%} (+/- {1:.2%})".format(np.mean(scores),
np.std(scores)*2))
print("Execution time: {0:.5} seconds \n".format(end-start))
```

```
from sklearn.naive_bayes import GaussianNB

start = time.time()

clf = GaussianNB()

clf.fit(X_train, y_train)

prediction = clf.predict(X_test)

scores = cross_val_score(clf, X, y, cv=5)

end = time.time()

accuracy_all.append(accuracy_score(prediction, y_test))

cvs_all.append(np.mean(scores))

print("Accuracy: {0:.2%}".format(accuracy_score(prediction, y_test)))

print("Cross validation score: {0:.2%} (+/- {1:.2%})".format(np.mean(scores),
np.std(scores)*2))

print("Execution time: {0:.5} seconds \n".format(end-start))

from sklearn.ensemble import RandomForestClassifier

from sklearn.ensemble import ExtraTreesClassifier

from sklearn.tree import DecisionTreeClassifier

start = time.time()

clf = RandomForestClassifier()

clf.fit(X_train, y_train)

prediction = clf.predict(X_test)

scores = cross_val_score(clf, X, y, cv=5)

end = time.time()

accuracy_all.append(accuracy_score(prediction, y_test))

cvs_all.append(np.mean(scores))

print("Random Forest Accuracy: {0:.2%}".format(accuracy_score(prediction, y_test)))

print("Cross validation score: {0:.2%} (+/- {1:.2%})".format(np.mean(scores),
np.std(scores)*2))

print("Execution time: {0:.5} seconds \n".format(end-start))
```

```
start = time.time()
clf = ExtraTreesClassifier()
clf.fit(X_train, y_train)
prediction = clf.predict(X_test)
scores = cross_val_score(clf, X, y, cv=5)
end = time.time()
accuracy_all.append(accuracy_score(prediction, y_test))
cvs_all.append(np.mean(scores))
print("Extra Trees Accuracy: {0:.2%}".format(accuracy_score(prediction, y_test)))
print("Cross validation score: {0:.2%} (+/- {1:.2%})".format(np.mean(scores),
np.std(scores)*2))
print("Execution time: {0:.5} seconds \n".format(end-start))
start = time.time()
clf = DecisionTreeClassifier()
clf.fit(X_train, y_train)
prediction = clf.predict(X_test)
scores = cross_val_score(clf, X, y, cv=5)
end = time.time()
accuracy_all.append(accuracy_score(prediction, y_test))
cvs_all.append(np.mean(scores))
print("Dedicion Tree Accuracy: {0:.2%}".format(accuracy_score(prediction, y_test)))
print("Cross validation score: {0:.2%} (+/- {1:.2%})".format(np.mean(scores),
np.std(scores)*2))
print("Execution time: {0:.5} seconds \n".format(end-start))
X = data.loc[:,features_selection]
y = data.loc[:, 'diagnosis']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 42)
accuracy_selection = []
cvs_selection = []
```

```
from sklearn.linear_model import SGDClassifier

start = time.time()

clf = SGDClassifier()

clf.fit(X_train, y_train)

prediction = clf.predict(X_test)

scores = cross_val_score(clf, X, y, cv=5)

end = time.time()

accuracy_selection.append(accuracy_score(prediction, y_test))

cvs_selection.append(np.mean(scores))

print("SGD Classifier Accuracy: {0:.2%}".format(accuracy_score(prediction, y_test)))

print("Cross validation score: {0:.2%} (+/- {1:.2%})".format(np.mean(scores),
np.std(scores)*2))

print("Execution time: %s seconds \n" % "{0:.5}".format(end-start))

from sklearn.svm import SVC, NuSVC, LinearSVC

start = time.time()

clf = SVC()

clf.fit(X_train, y_train)

prediction = clf.predict(X_test)

scores = cross_val_score(clf, X, y, cv=5)

end = time.time()

accuracy_selection.append(accuracy_score(prediction, y_test))

cvs_selection.append(np.mean(scores))

print("SVC Accuracy: {0:.2%}".format(accuracy_score(prediction, y_test)))

print("Cross validation score: {0:.2%} (+/- {1:.2%})".format(np.mean(scores),
np.std(scores)*2))

print("Execution time: %s seconds \n" % "{0:.5}".format(end-start))

start = time.time()

clf = NuSVC()
```

```
clf.fit(X_train, y_train)

prediciton = clf.predict(X_test)

scores = cross_val_score(clf, X, y, cv=5)

end = time.time()

accuracy_selection.append(accuracy_score(prediction, y_test))

cvs_selection.append(np.mean(scores))

print("NuSVC Accuracy: {0:.2%}".format(accuracy_score(prediction, y_test)))

print("Cross validation score: {0:.2%} (+/- {1:.2%})".format(np.mean(scores),
np.std(scores)*2))

print("Execution time: %s seconds \n" % "{0:.5}".format(end-start))

start = time.time()

clf = LinearSVC()

clf.fit(X_train, y_train)

prediction = clf.predict(X_test)

scores = cross_val_score(clf, X, y, cv=5)

end = time.time()

accuracy_selection.append(accuracy_score(prediction, y_test))

cvs_selection.append(np.mean(scores))

print("LinearSVC Accuracy: {0:.2%}".format(accuracy_score(prediction, y_test)))

print("Cross validation score: {0:.2%} (+/- {1:.2%})".format(np.mean(scores),
np.std(scores)*2))

print("Execution time: %s seconds \n" % "{0:.5}".format(end-start))

from sklearn.neighbors import KNeighborsClassifier

start = time.time()

clf = KNeighborsClassifier()

clf.fit(X_train, y_train)

prediction = clf.predict(X_test)

scores = cross_val_score(clf, X, y, cv=5)

end = time.time()
```

```
accuracy_selection.append(accuracy_score(prediction, y_test))

cvs_selection.append(np.mean(scores))

print("Accuracy: {0:.2%}".format(accuracy_score(prediction, y_test)))

print("Cross validation score: {0:.2%} (+/- {1:.2%})".format(np.mean(scores),
np.std(scores)*2))

print("Execution time: %s seconds \n" % "{0:.5}".format(end-start))

from sklearn.naive_bayes import GaussianNB

start = time.time()

clf = GaussianNB()

clf.fit(X_train, y_train)

prediction = clf.predict(X_test)

scores = cross_val_score(clf, X, y, cv=5)

end = time.time()

accuracy_selection.append(accuracy_score(prediction, y_test))

cvs_selection.append(np.mean(scores))

print("Accuracy: {0:.2%}".format(accuracy_score(prediction, y_test)))

print("Cross validation score: {0:.2%} (+/- {1:.2%})".format(np.mean(scores),
np.std(scores)*2))

print("Execution time: %s seconds \n" % "{0:.5}".format(end-start))

from sklearn.ensemble import RandomForestClassifier

from sklearn.ensemble import ExtraTreesClassifier

from sklearn.tree import DecisionTreeClassifier

start = time.time()

clf = RandomForestClassifier()

clf.fit(X_train, y_train)

prediction = clf.predict(X_test)

scores = cross_val_score(clf, X, y, cv=5)

end = time.time()

accuracy_selection.append(accuracy_score(prediction, y_test))
```

```
cvs_selection.append(np.mean(scores))

print("Random Forest Accuracy: {0:.2%}".format(accuracy_score(prediction, y_test)))

print("Cross validation score: {0:.2%} (+/- {1:.2%})".format(np.mean(scores),
np.std(scores)*2))

print("Execution time: %s seconds \n" % "{0:.5}".format(end-start))

start = time.time()

clf = ExtraTreesClassifier()

clf.fit(X_train, y_train)

prediction = clf.predict(X_test)

scores = cross_val_score(clf, X, y, cv=5)

end = time.time()

accuracy_selection.append(accuracy_score(prediction, y_test))

cvs_selection.append(np.mean(scores))

print("Extra Trees Accuracy: {0:.2%}".format(accuracy_score(prediction, y_test)))

print("Cross validation score: {0:.2%} (+/- {1:.2%})".format(np.mean(scores),
np.std(scores)*2))

print("Execution time: %s seconds \n" % "{0:.5}".format(end-start))

start = time.time()

clf = DecisionTreeClassifier()

clf.fit(X_train, y_train)

prediction = clf.predict(X_test)

scores = cross_val_score(clf, X, y, cv=5)

end = time.time()

accuracy_selection.append(accuracy_score(prediction, y_test))

cvs_selection.append(np.mean(scores))

print("Dedicion Tree Accuracy: {0:.2%}".format(accuracy_score(prediction, y_test)))

print("Cross validation score: {0:.2%} (+/- {1:.2%})".format(np.mean(scores),
np.std(scores)*2))

print("Execution time: %s seconds \n" % "{0:.5}".format(end-start))
```

```
diff_accuracy = list(np.array(accuracy_selection) - np.array(accuracy_all))

diff_cvs = list(np.array(cvs_selection) - np.array(cvs_all))

d = {'accuracy_all':accuracy_all, 'accuracy_selection':accuracy_selection,
'diff_accuracy':diff_accuracy,

    'cvs_all':cvs_all, 'cvs_selection':cvs_selection, 'diff_cvs':diff_cvs,}

index = ['SGD', 'SVC', 'NuSVC', 'LinearSVC', 'KNeighbors', 'GaussianNB', 'RandomForest',
'ExtraTrees', 'DecisionTree']

df = pd.DataFrame(d, index=index)
```

Chapter 6

RESULTS

6.1 DATA

id	diagnosis	radius_me	texture_m	perimeter_	area_mea	smoothne	compactn	concavity_	concave p	symmetry_	fractal_dir
842302	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871
842517	M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667
84300903	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999
84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744
84358402	M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883
843786	M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613
844359	M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742
84458202	M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451
844981	M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389
84501001	M	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243
845636	M	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697
84610002	M	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.06082
846226	M	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078
846381	M	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.05338
84667401	M	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.07682
84799002	M	14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303	0.07077
848406	M	14.68	20.13	94.74	684.5	0.09867	0.072	0.07395	0.05259	0.1586	0.05922
84862001	M	16.13	20.68	108.1	798.8	0.117	0.2022	0.1722	0.1028	0.2164	0.07356
849014	M	19.81	22.15	130	1260	0.09831	0.1027	0.1479	0.09498	0.1582	0.05395
8510426	B	13.54	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	0.1885	0.05766
8510653	B	13.08	15.71	85.63	520	0.1075	0.127	0.04568	0.0311	0.1967	0.06811
8510824	B	9.504	12.44	60.34	273.9	0.1024	0.06492	0.02956	0.02076	0.1815	0.06905
8511133	M	15.34	14.26	102.5	704.4	0.1073	0.2135	0.2077	0.09756	0.2521	0.07032
851509	M	21.16	23.04	137.2	1404	0.09428	0.1022	0.1097	0.08632	0.1769	0.05278

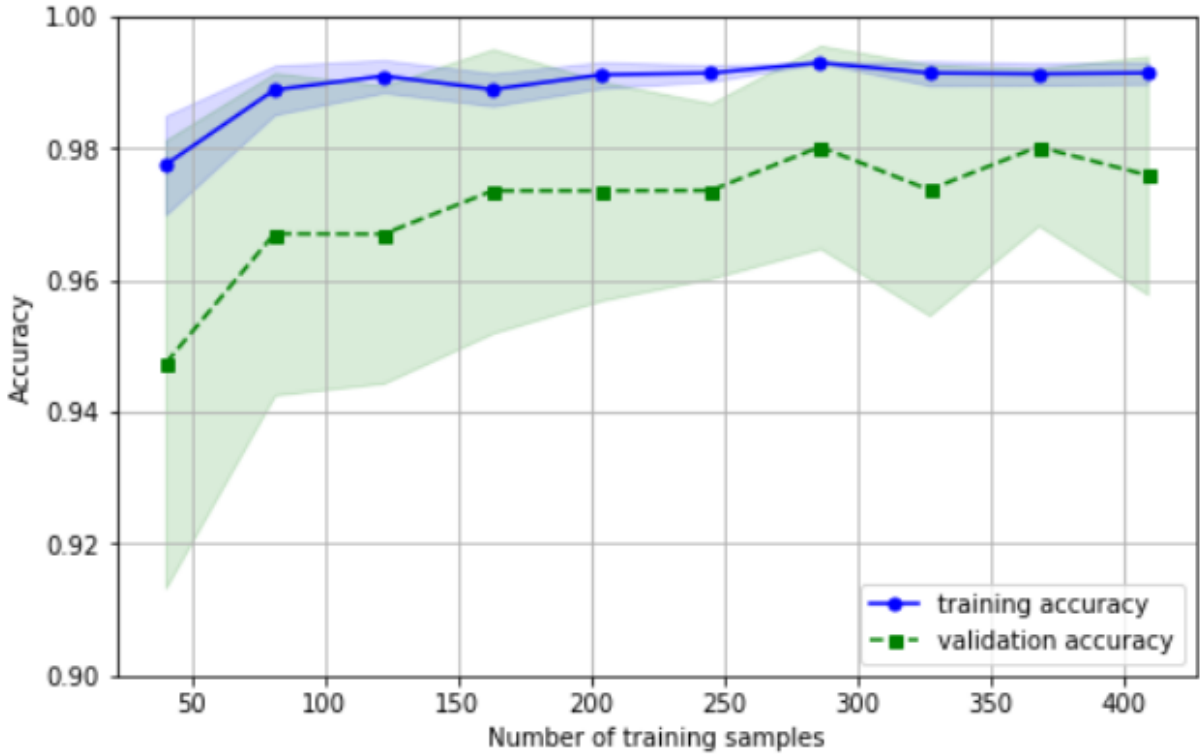
6.2 PRE-PROCESSED DATA

Out[1]:

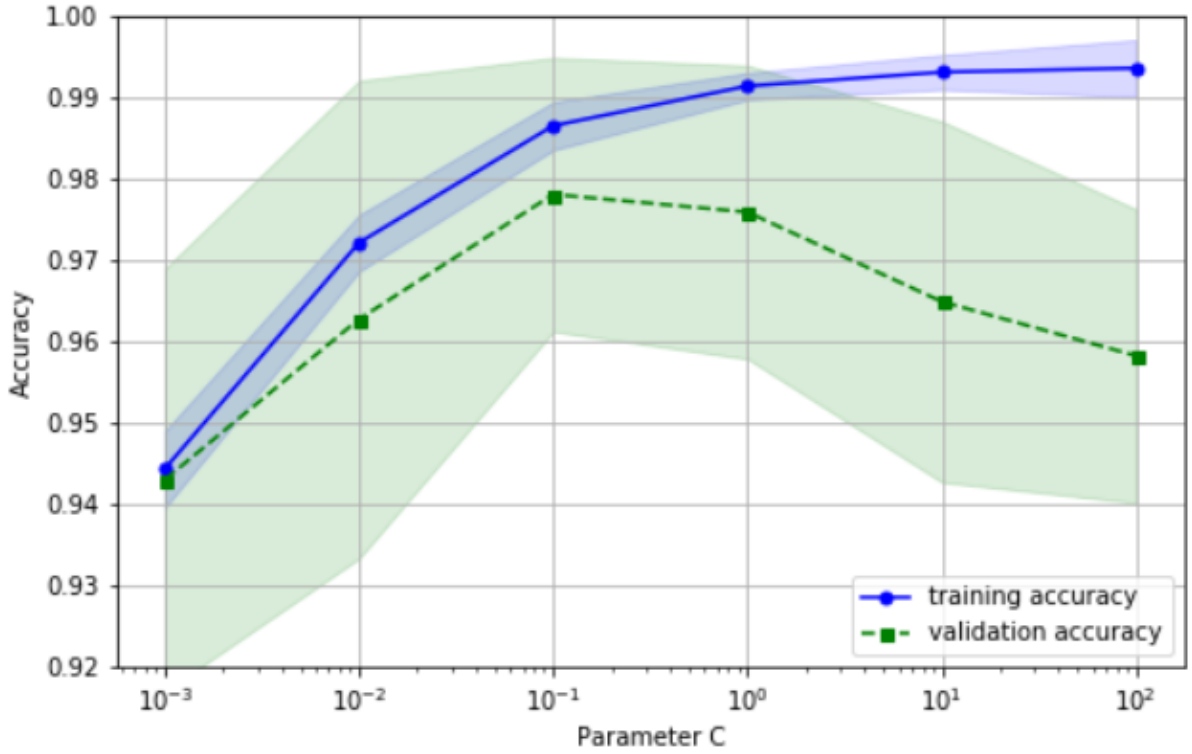
0	1	2	3	4	5	6	7	8	9	...	22	23	24	25	26	27	28	29	30	31
842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...	25.38	17.33	184.60	2019.0	0.1622	0.6656	0.7119	0.2654	0.4601	0.11890
842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...	24.99	23.41	158.80	1956.0	0.1238	0.1866	0.2416	0.1860	0.2750	0.08902
84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...	23.57	25.53	152.50	1709.0	0.1444	0.4245	0.4504	0.2430	0.3613	0.08758
84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...	14.91	26.50	98.87	567.7	0.2098	0.8663	0.6869	0.2575	0.6638	0.17300
84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...	22.54	16.67	152.20	1575.0	0.1374	0.2050	0.4000	0.1625	0.2364	0.07678

'ows × 32 columns

6.3 NUMBER OF TRAINING SAMPLES



6.4 ACCURACY PARAMETER C:



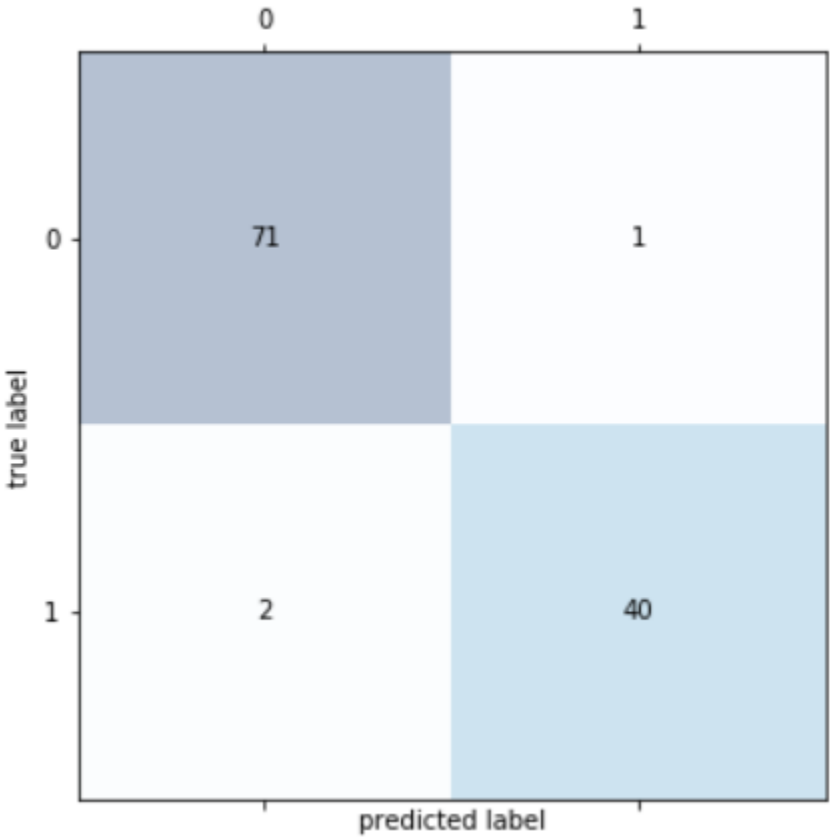
6.5 ACCURACY WITH REGRESSION:

```
Best: 0.978022 using {'clf__C': 0.1, 'clf__kernel': 'linear'}
```

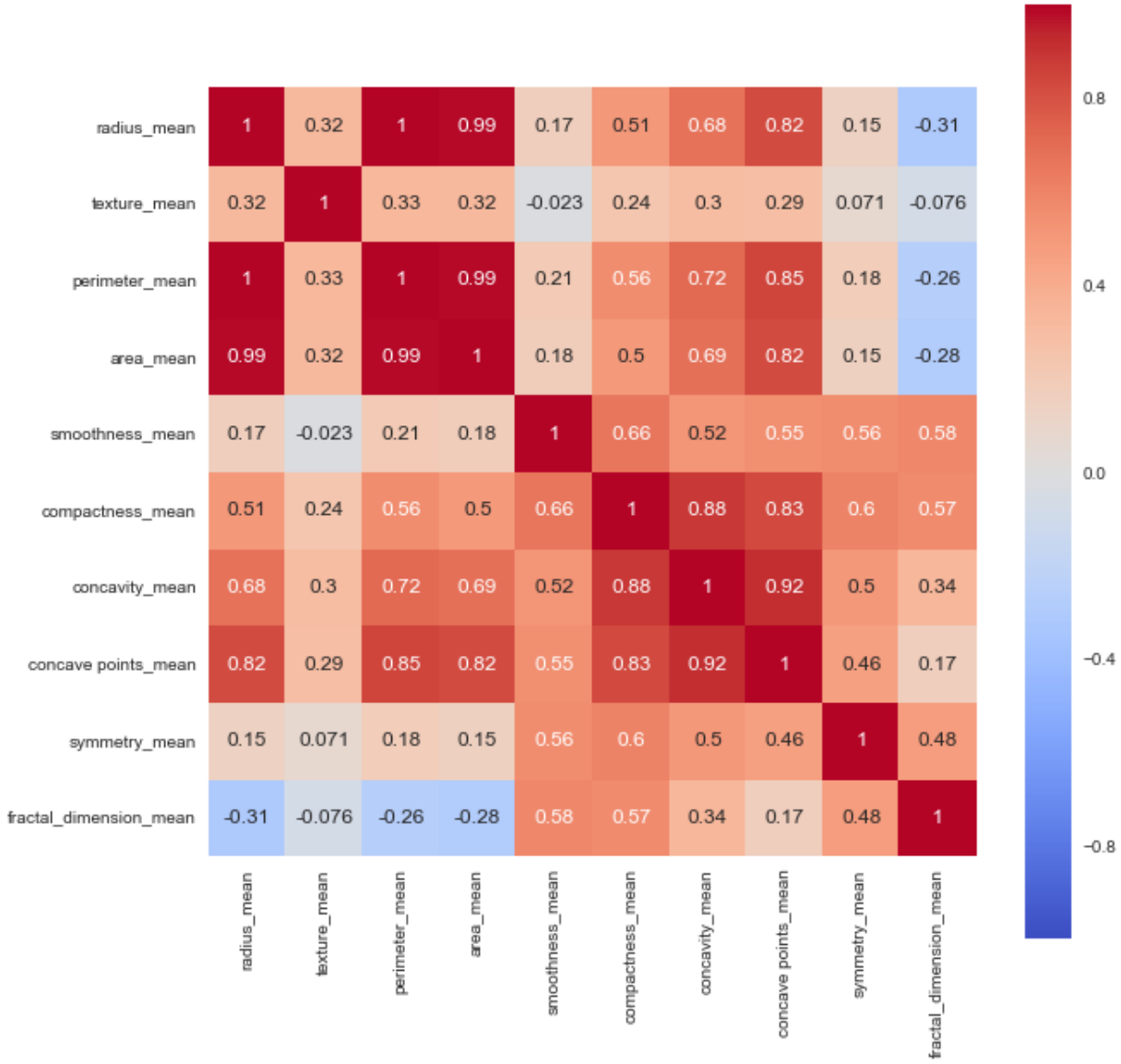
6.5.1 TEST ACCURACY:

```
Test accuracy: 0.965
```

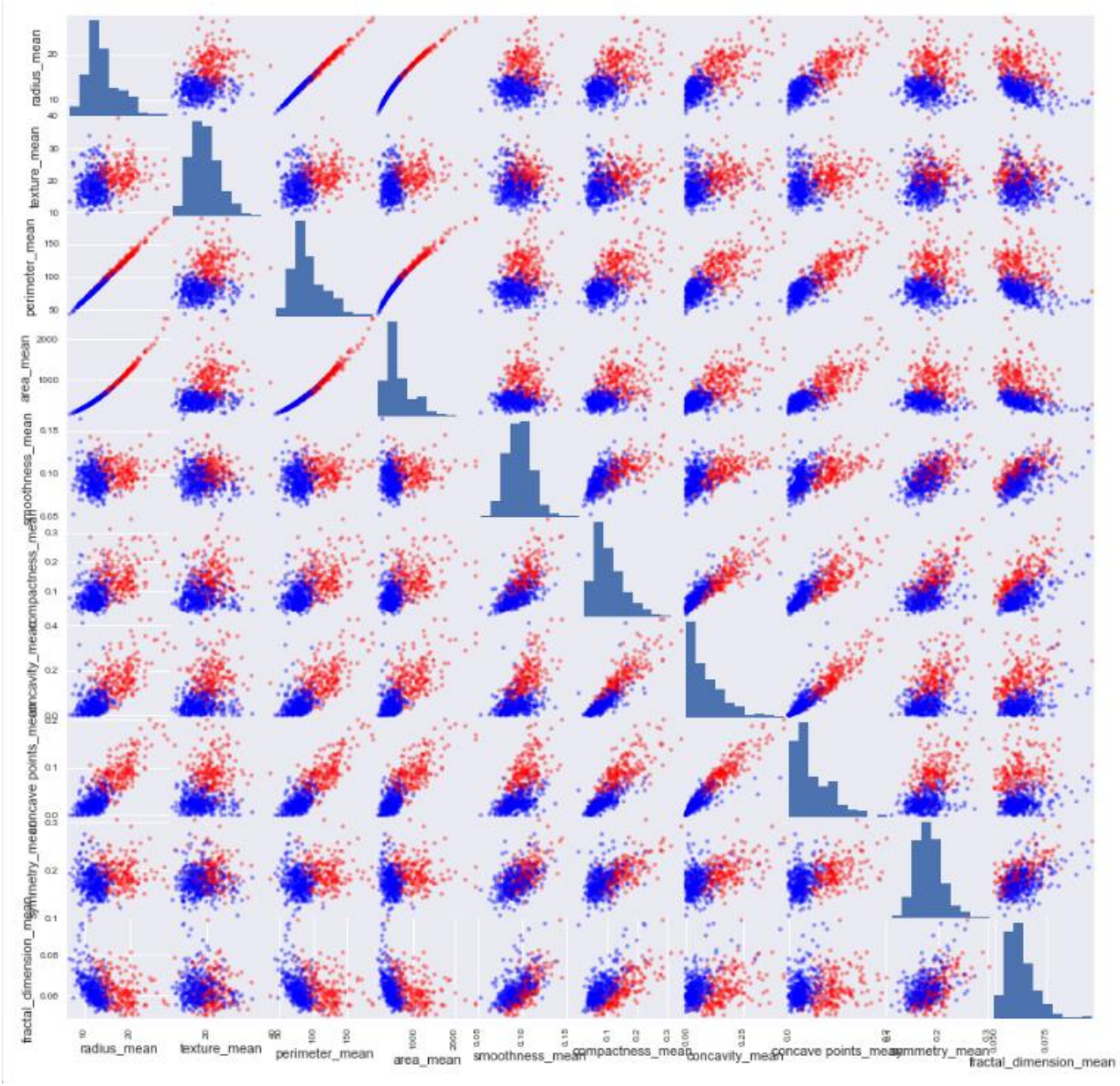
6.6 PREDICTED LABEL



6.7 CLASSIFICATION ALGORITHM:



6.8 SCATTER MATRIX



6.9 ACCURACY:

Out[28]:

	accuracy_all	accuracy_selection	cvs_all	cvs_selection	diff_accuracy	diff_cvs
SGD	0.842105	0.622807	0.718753	0.783409	-0.219298	0.064656
SVC	0.692982	0.745614	0.717045	0.782008	0.052632	0.064963
NuSVC	0.692982	0.745614	0.718815	0.804925	0.052632	0.086110
LinearSVC	0.824561	0.771930	0.744671	0.746010	-0.052632	0.001339
KNeighbors	0.938596	0.921053	0.886002	0.882493	-0.017544	-0.003509
GaussianNB	0.947368	0.947368	0.914013	0.908765	0.000000	-0.005248
RandomForest	0.929825	0.929825	0.935129	0.917707	0.000000	-0.017422
ExtraTrees	0.964912	0.947368	0.938607	0.919292	-0.017544	-0.019315
DecisionTree	0.947368	0.894737	0.919261	0.906933	-0.052632	-0.012328

Chapter 7

NOVELTY

EXISTING SYSTEM:

- Rule based approach is used for classification which gives static range value for different classes. Therefore we will not able dynamic images or outlier behaviour images.
- Multi classification problem is not optimized and ignore the class imbalance problem. Therefore in learning all classes is not input in learning phase so it became a biased learning.
- Features set is not normalize. Therefore different features show different outputs and show different representation during training phase of classifiers.

PROPOSED SYSTEM:

In our proposed system, mammogram image can be enhancement using Gaussian filter. Second the segmentation is done using Fuzzy C means for partitioning the mammogram image into multiple segments to identify the mass easily and features are extracted using HWT (Haar Wavelet Features). Further tumor has been analyzed and classified using Multi –SVM (Support Vector Machine) classifier.

SVMs deliver a unique solution, since the optimality problem is convex. This is an advantage compared to Neural Networks, which have multiple solutions associated with local minima and for this reason may not be robust over different samples.

ADVANTAGES:

- More Accuracy
- Reduced time consumption

Chapter 8

CONCLUSION AND FUTURE WORK

A plan for the diagnosis and treatment of cancer is a key component of any overall cancer control plan. Its main goal is to cure cancer patients or prolong their life considerably, ensuring a good quality of life. In order for a diagnosis and treatment programme to be effective, it must never be developed in isolation. It needs to be linked to an early detection programme so that cases are detected at an early stage, when treatment is more effective and there is a greater chance of cure. It also needs to be integrated with a palliative care programme, so that patients with advanced cancers, who can no longer benefit from treatment, will get adequate relief from their physical, psychosocial and spiritual suffering. Furthermore, programmes should include a awareness-raising component, to educate patients, family and community members about the cancer risk factors and the need for taking preventive measures to avoid developing cancer.

Where resources are limited, diagnosis and treatment services should initially target all patients presenting with curable cancers, such as breast, cervical and oral cancers that can be detected early. They could also include childhood acute lymphatic leukaemia, which has a high potential for cure although it cannot be detected early. Above all, services need to be provided in an equitable and sustainable manner. As and when more resources become available, the programme can be extended to include other curable cancers as well as cancers for which treatment can prolong survival considerably.

REFERENCES

[1] Sarvestan Soltani A. , Safavi A. A., Parandeh M. N. and Salehi M., 2010, “Predicting Breast Cancer Survivability using data mining techniques,” Software Technology and Engineering (ICSTE), 2nd International Conference, v2, 227-231.

[2] Gupta, S., Kumar, D., and Sharma, A., 2011, “Data Mining Classification Techniques Applied for Breast Cancer Diagnosis and Prognosis,” Indian Journal of Computer Science and Engineering (IJCSE), 2(2), 188-195.

[3] Meyer, D., Leisch, F., and Hornik, K., 2003, “The Support Vector Machine under Test,” Neurocomputing, 55(1), 169-186.

[4] Kleinbaum, D., Kupper, L., Nizam, A., and Rosenberg, E., 2013, Applied Regression Analysis and Other Multivariable Methods, Cengage Learning.

[5] Maimon, O., and Rokach, L., 2005, Data Mining and Knowledge Discovery Handbook (Vol. 2), Springer, New York

[6] Wang, W., Sun, X. W., Li, C. F., Lv, L., Li, Y. F., Chen, Y. B., and Zhou, Z. W., 2011, “Comparison of the 6th and 7th Editions of the UICC TNM Staging System for Gastric Cancer: Results of a Chinese Single-institution Study of 1,503 Patients,” Annals of Surgical Oncology, 18(4), 1060-1067.

[7] Psychogios, G., Waldfahrer, F., Bozzato, A., and Iro, H., 2010, “Evaluation of the Revised TNM Classification in Advanced Laryngeal Cancer,” European Archives of Oto-Rhino-Laryngology, 267(1), 117-121.

[8] Bellaachia, A. and Guven E., 2006, "Predicting Breast Cancer Survivability Using Data Mining Techniques," Age, 58(13), 10-110.

[9] Choi, J. P., Han, T. H. and Park, R. W, 2009, "A Hybrid Bayesian Network Model for Predicting Breast Cancer Prognosis," Journal of Korean Society of Medical Informatics, 15(1), 49-57.

[10] Delen, D., Walker, G. and Kadam, A., 2005, "Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods," Artificial Intelligence in Medicine, 34(2), 113-127.