International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

EFFICIENT BREAST CANCER PREDICTION USING MLTECHNIQUES

MR.B.ARJUN BALAJI[1], Dr.M.PARVATHY[2],
MR.S.ARUN INIGO[3], MR.M.SENTHILKUMAR[4], Mr.P.ANAND[5]
[1] PG Student, Department of CSE,
[2] Professor and Head, Department of CSE,
[3, 4, 5] Assistant Professor, Department of CSD& AI&DS,
Sethu Institute of Technology, Kariapatti-626115

ABSTRACT

- During their life, among 8% of women are diagnosed with Breast cancer (BC), after lung cancer, BC is the second popular cause of death in both developed and undeveloped worlds. BC is characterized by the mutation of genes, constant pain, changes in the size, color(redness), skin texture of breasts.
- Classification of breast cancer leads pathologists to find a systematic and objective prognostic, generally the most frequent classification is binary (benign cancer/malign cancer).
- Today, Machine Learning (ML) techniques are being broadly used in the breast cancer classification problem. They provide high classification accuracy and effective diagnostic capabilities.
- In this system, we present two different classifiers: Decision tree and logistic regression for breast cancer classification. We propose a comparison between the two new implementations and evaluate their accuracy using cross validation.

CHAPTER 1 INTRODUCTION

1.1 General Introduction:

Breast cancer is one of the most lethal and heterogeneous disease in this present era that causes the death of enormous number of women all over the world. It is the second largest disease that is responsible of women death. There are various machine learning and data mining algorithms that are being used for the prediction of breast cancer. Breast cancer is originated through malignant tumors, when the growth of the cell got out of control. The cancer cells spread throughout the tumors that cause different stages of cancer. There are different types of breast cancer which occurs when affected cells and tissues spread throughout the body. Breast cancer is a type of cancer that occurs mostly in females and is the leading cause of women's deaths. These deaths can be reduced by early detection of the cancerous cells. Cancerous cells are detected by performing various tests like MRI, mammogram, ultrasound and biopsy. The dataset used in this project contains features that are computed from a digitized image of a fine needle aspiration (FNA) biopsy of a breast mass. They describe characteristics of the cell nuclei present in the image. Diagnosis of breast cancer is done by classifying the tumour. Tumours can be either benign or malignant. Malignant tumours are more harmful than the benign. Unfortunately, not all physicians are expert in distinguishing between the benign and malignant tumours and the classification of tumour cells may take up to 2 days. Machine learning algorithms are used to predict the type of cancerous cells efficiently and accurately. Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn

for themselves. The different algorithms used are: Support Vector Machine (SVM), Decision Tree (DT), Naive Bayes (NB), Logistic regression (LR), and k Nearest Neighbours (k-NN). KNN makes predictions using the training dataset directly. Predictions are made for a new instance (x) by searching through the entire training set for the K most similar instances (the neighbours) and summarizing the output variable for those K instances. For regression this might be the mean output variable, in classification this might be the mode (or most common) class value. To determine which of the K instances in the training dataset are most similar to a new input a distance measure is used. For real-valued input variables, the most popular distance measure is Euclidean distance. Euclidean distance is calculated as the square root of the sum of the squared differences between a new point (x) and an existing point (xi) across all input attributes j. Euclidean Distance(x, xi) = sqrt(sum((xj – xij)^2)) The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples. In the classification phase, k is a user-defined constant, and an unlabelled vector (a query or test point) is classified by assigning the label which is most frequent among the k training samples nearest to that query point. Nowadays, computers have made significant improvements to technology that lead to the creation of huge volumes of data. In addition, advances in medical database management systems are creating a large number of medical databases. Knowledge creation and the management of large amounts of heterogeneous data has become a major research area, namely data mining. Data mining is a process of identifying new, potentially useful, valid and ultimately understandable models in data. Data mining techniques can be classified into supervised and unsupervised learning techniques. The unsupervised learning technique is not guided by variables and does not create hypotheses before analysis. Based on the results, a model will be constructed. A common unsupervised technique is clustering. The supervised learning technique requires the construction of a model that is used in the analysis of past performance. The supervised learning techniques used in medical and clinical research are classification, statistical regression and association rules. Since classification is the most commonly used data mining technique and uses a set of pre-classified examples to develop a model that can classify the document population in general. The main objective of the classification technique is to accurately predict the target class for each case in the data. This research uses classification techniques in medical science. It first classifies the data set and then determines the best algorithm for the diagnosis and prediction of breast cancer. Prediction begins with identifying symptoms in patients, then identifying sick patients from a large number of sick and healthy patients. Thus, the primary objective of this paper is to analyze data from a breast cancer data set using a classification technique to accurately predict the class in each case. Many authors have used the WEKA tool in their work to compare the performance of different classifiers applied to different datasets. But none of the authors worked on predicting the accuracy of the breast cancer data set. Here, we considered four type of classifiers to study their performance according to various parameters obtained by applying them in the data set. People facing cancer are naturally concerned about what the future holds. Understanding cancer and what to expect can help patients and their loved ones plan for treatment options, think about lifestyle changes, and make decisions about their quality of life and finances. Many people with cancer want to know their prognosis. Research in the field of cancer detection helps to give an idea of the likely course and outcome of this disease. Complementing these biological and clinical studies, data mining, as a powerful tool that can be used to discover patterns in medical data repositories, is finding its way into the analytics driven medical research arena. In this study, we used three popular data mining techniques. Healthcare related data mining is one of the most rewarding and challenging areas of application in data mining and knowledge discovery. The challenges are due to the data sets which are large, complex, heterogeneous, hierarchical, time series and of varying quality. The available healthcare data sets are fragmented and distributed in nature, thereby making the process of data integration a highly challenging task. Other issues related to the use of highly sensitive healthcare data that the data miner has to tackle are ethical,

legal and social aspects. Due to the lack of domain knowledge on the analysts' behalfit becomes necessary for an active collaboration between the domain specialist and thedata miner (Cios & Moore, 2002). Various data mining techniques are used in the areaof clinical decision support and, in particular, cancer diagnostics and prognostics. Explanatory and confirmative techniques are the most commonly used in medical datamining. The major issues related to data mining in the medical field faced by data miners. Supervised learning is fairly common in classification problems because the goal is often to get the computer to learn a classification system that we have created. Digit recognition, once again, is a common example of classification learning. More generally, classification learning is appropriate for any problem where deducing a classification is useful and the classification is easy to determine. In some cases, it might not even be necessary to give predetermined classifications to every instance ofa problem if the agent can work out the classifications for itself. Unsupervised learning4 4 seems much harder: the goal is to have the computer learn how to do something that we don't tell it how to do! There are actually two approaches to unsupervised learning. The first approach is to teach the agent not by giving explicit categorizations, but by using some sort of reward system to indicate success. Note thatthis type of training will generally fit into the decision problem framework because the goal is not to produce a classification but to make decisions that maximize rewards.Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n- dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well. Support Vectors are simply the co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes (hyper-plane/ line). Initially SVMs mapthe input vector into a feature space of higher dimensionality and identify the hyperplane that separates the data points into two classes. The marginal distance between the decision hyperplane and the instances that are closest to boundary is maximized. The resulting classifier achieves considerable generalizability and can therefore be used for the reliable classification of new samples. It is worth noting that probabilistic outputs can also be obtained for SVMs figure below illustrates how an SVM might work in order to classify tumours among benign and malignant based on their size and patients' age. The identified hyper plane can be thought as a decision boundary between the two clusters. Obviously, the existence of a decision boundary allows for the detection of any misclassification produced by the method.

Objectives:

The main objective of our project is,

- To predict or to classify the breast cancer effectively.
- To implement the combination of machine learning algorithms.
- To enhance the overall performance for classification algorithms.

CHAPTER 2 SYSTEM PROPOSAL

EXISTING SYSTEM:

In existing system, the comparative analysis of machine learning, deep learning and data mining techniques being used for the prediction of breast cancer. Many researchers have put their efforts on breast cancer diagnoses and prognoses, every technique has different accuracy rate and it varies for different situations, tools and datasets being used. Our main focus is to comparatively analyze different existing Machine Learning and Data Mining techniques in order to find out the most appropriate method that will support the large dataset with good accuracy of prediction. The main purpose of this review is to highlight all the previous studies of machine learning algorithms that are being used for breast cancer prediction and this article provides the all necessary information to the beginners who want to analyze themachine learning algorithms to gain the base of deep learning.

DISADVANTAGES:

- It is not efficient for large number of datasets.
- It doesn't implement the hybrid machine learning algorithms.
- The prediction results is not efficient.
- Time consumption is high.

PROPOSED SYSTEM:

In proposed system, the Wisconsin Breast Cancer dataset was taken as input. The input data was collected from dataset repository. Then, we have to implement the data pre- processing step. In this step, we have to handle the missing values for avoid wrong prediction, and to encode the label for input data. Then, we have to implement the feature selection for select the best features from the pre-processed data such as chi square. Then, we have to split the feature selected data into test and train. Training portion is used to evaluate the model and testing portion is used to predicting the model. After that, we have to implement the two different machine learning algorithms for predicting the breast cancer such as logistic regression and decision tree.

ADVANTAGES:

- It is efficient for large number of datasets.
- The experimental result is high when compared with existing system.
- To predict the breast cancer effectively.
- Time consumption is low.

LITERATURE SURVEY:

Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis,2020

Author*:* NOREEN FATIMA 1, LI LIU 1, SHA HONG1, AND HAROON AHMED

Methodology:

Breast cancer is type of tumor that occurs in the tissues of the breast. It is most common type of cancer found in women around the world and it is among the leading causes of deaths in women. This article presents the comparative analysis of machinelearning, deep learning and data mining techniques being used for the prediction of breast cancer. Many researchers have put their efforts on breast cancer diagnoses and prognoses, every technique has different accuracy rate and it varies for different situations, tools and datasets being used. Our main focus is to comparatively analyze different existing Machine Learning and Data Mining techniques in order to find out the most appropriate method that will support the large dataset with good accuracy of prediction. The main purpose of this review is to highlight all the previous studies of machine learning algorithms that are being used for breast cancer prediction and this article provides the all necessary information to the beginners who want to analyze themachine learning algorithms to gain the base of deep learning.

Advantage:

- Naives bayes provides the highest accuracy while calculating the probabilities of noisy data that is used as an input

Using Machine learning algorithms for breast cancer risk prediction anddiagnosis, 2018

Author: Anusha Bharat, Pooja N, R Anishka Reddy

Methodology:

Machine learning is frequently used in medical applications such as detection of the type of cancerous cells. Breast cancer represents one of the diseases that causes a highnumber of deaths every year. It is the most common type of cancer and the main causeof women's deaths worldwide. The cancerous cells are classified as Benign (B) or Malignant (M). There are many algorithms for classification and prediction of breast cancer: Support Vector Machine (SVM), Decision Tree (CART), Naive Bayes (NB) and k Nearest Neighbours (kNN). In this project, Support Vector Machine (SVM) onthe Wisconsin Breast Cancer dataset is used. The dataset is also trained with the other algorithms: KNN, Naives Bayes and CART and the accuracy of prediction for each algorithm is compared.

Advantage*:*

- k-fold cross validation is each testing subsample is used exactly once

Applying Best Machine Learning Algorithms for Breast Cancer PredictionandClassification

Author: Youness Khourdifi, Mohamed Bahaj

Methodology:

Breast cancer is one of the most common cancers among women in the world, accounting for the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem in today's society. Inthis paper, we will present an overview of the evolution of large data in the health system, and apply four learning algorithms to a breast cancer data set. The aim of thisresearch work is to predict breast cancer, which is the second leading cause of death among women worldwide, and with early detection and prevention can dramatically reduce the risk of death, using several machine-learning algorithms that are Random

Forest, Naïve Bayes, Support Vector Machines SVM, and K-Nearest Neighbors K- NN, and chose the most effective. The experimental results show that SVM gives thehighest accuracy 97.9%. The finding will help to select the best classification machine-learning algorithm for breast cancer prediction.

Advantage*:*

- RF and NB has the highest error rate

Disadvantage:

- Training time is high.

Analysis of cancer data: a data mining approach,2019Author: Dursun Delen

Methodology:

Even though cancer research has traditionally been clinical and biological in nature, in recent years data driven analytic studies have become a common complement. In medical domains where data and analytics driven research is successfully applied, new and novel research directions are identified to further advance the clinical and biological studies. In this research, we used three popular data mining techniques (decision trees, artificial neural networks and support vector machines) along with themost commonly used statistical analysis technique logistic regression to develop prediction models for prostate cancer survivability. The data set contained around 120000 records and 77 variables. A k-fold cross-validation methodology was used in model building, evaluation and comparison. The results showed that support vector machines are the most accurate predictor (with a test set accuracy of 92.85%) for thisdomain, followed by artificial neural networks and decision trees.

Advantage*:*

- They have a simple geometric interpretation and give a sparse solution.
- The computational complexity of SVMs does not depend on the dimensionality of the input space.

Disadvantage:

- Time consumption is high.

Predicting factors for survival of breast cancer patients using machinelearning techniques, 2019

Author: Mogana Darshini Ganggayah1 , Nur Aishah Taib2 , Yip Cheng Har2 , Pietro Lio3 and Sarinder Kaur Dhillon

Methodology:

A large hospital-based breast cancer dataset retrieved from the University Malaya Medical Centre, Kuala Lumpur, Malaysia (n = 8066) with diagnosis information between 1993 and 2016 was used in this study. The dataset contained 23 predictor variables and one dependent variable, which referred to the survival status of the patients (alive or dead). In determining the significant prognostic factors of breast cancer survival rate, prediction models were built using decision tree, random forest, neural networks, extreme boost, logistic regression, and support vector machine. Next, the dataset was clustered based on the receptor status of breast cancer patients identified via immunohistochemistry to perform advanced modelling using random forest. Subsequently, the important variables were ranked via variable selectionmethods in random forest. Finally, decision trees were built

and validation was performed using survival analysis.
Advantage*:*

- The type of biopsy may indicate the biology of cancer, whether or not completeremoval of the tumour has survival advantage from a needle core or FNAC.

Disadvantage:

- Less effective.

Title: Classifying Breast Cancer Types Based on Fine Needle Aspiration Biopsy Data Using Random Forest Classifier
Author: Farzana Kabir Ahmad, Nooraini YusoffMethodology:
This study aims to classify breast cancer lesions which have been obtained from fine needle aspiration (FNA) procedure using random forest. Random forest is a classifierbuilt based on the combination of decision trees and has been identified to perform well in comparison to other machine learning techniques.
Advantages:

1. The process is implemented with removing unwanted data.

Disadvantages:

1. Accuracy is low.
2. Prediction is not accurate

Title: Breast Cancer Prediction: A Comparative Study Using MachineLearning Techniques
Author: Md. Milon Islam1 · Md. Rezwanul Haque1 · Hasib Iqbal1 · Md. Munirul Hasan2 · Mahmudul Hasan3 ·Muhammad Nomani Kabir
Methodology:

The Wisconsin Breast Cancer dataset is obtained from a prominent machine learning database named UCI machine learning database. The performance of the study is measured with respect to accuracy, sensitivity, specificity, precision, negative predictive value, false-negative rate, false-positive rate, F1score, and Matthews Correlation Coefficient. Additionally, these techniques were appraised on precision– recall area under Curve.
Advantages:

1. Accuracy is high.

Disadvantages:

1. Training time is low while using SVM algorithm.

Title: Machine Learning techniques in breast cancer prognosis prediction:Aprimary evaluation
Author: Carlo Boeri | Corrado Chiappa | Federica Galli | Valentina DeBerardinis |Laura Bardelli | Giulio Carcano | Francesca Rovera
Methodology:

Three outcomes were chosen: cancer recurrence (both loco-regional and systemic) and death from the disease within 32 months. We developed two types of ML models for every outcome (Artificial Neural Network and Support Vector Machine). Each ML algorithm was tested in accuracy (=95.29%-96.86%), sensitivity (=0.35-0.64), specificity (=0.97-0.99), and AUC (=0.804-0.916).

Advantages:

Therefore, the authors did not undergo yet an ethics committee consultation that would be highly recommended before using ML algorithms in the clinical practice.

Disadvantages:

Low performance

Title: Breast Cancer Prediction using varying Parameters of MachineLearning Models

Author: Puja a Guptaa, Shruti GargMethodology:

Malignancy of tumour has caused major number of deaths among women. Machine learning tools with proper hyper parametric can help in identifying tumours efficiently. This paper presents six supervised machine learning algorithms such as k-Nearest Neighbourhood, Logistic Regression, Decision Tree, Random Forest, Support VectorMachine with radial basis function kernel.

Advantages:

SVM is adaptable for complex and higher dimension datasets. They can also be applied for linear and nonlinear data.

Disadvantages:

Training time is low.

Title: Breast Cancer Type Classification Using Machine LearningAuthor: Jiande Wu and Chindo Hicks

Methodology:

Among the four ML algorithms evaluated, the Support Vector Machine algorithm was able to classify breast cancer more accurately into triple negative and non-triple negative breast cancer and had less misclassification errors than the other three algorithms evaluated. Conclusions: The prediction results show that ML algorithms are efficient and can be used for classification of breast cancer into triple negative andnon-triple negative breast cancer types.

Advantages:

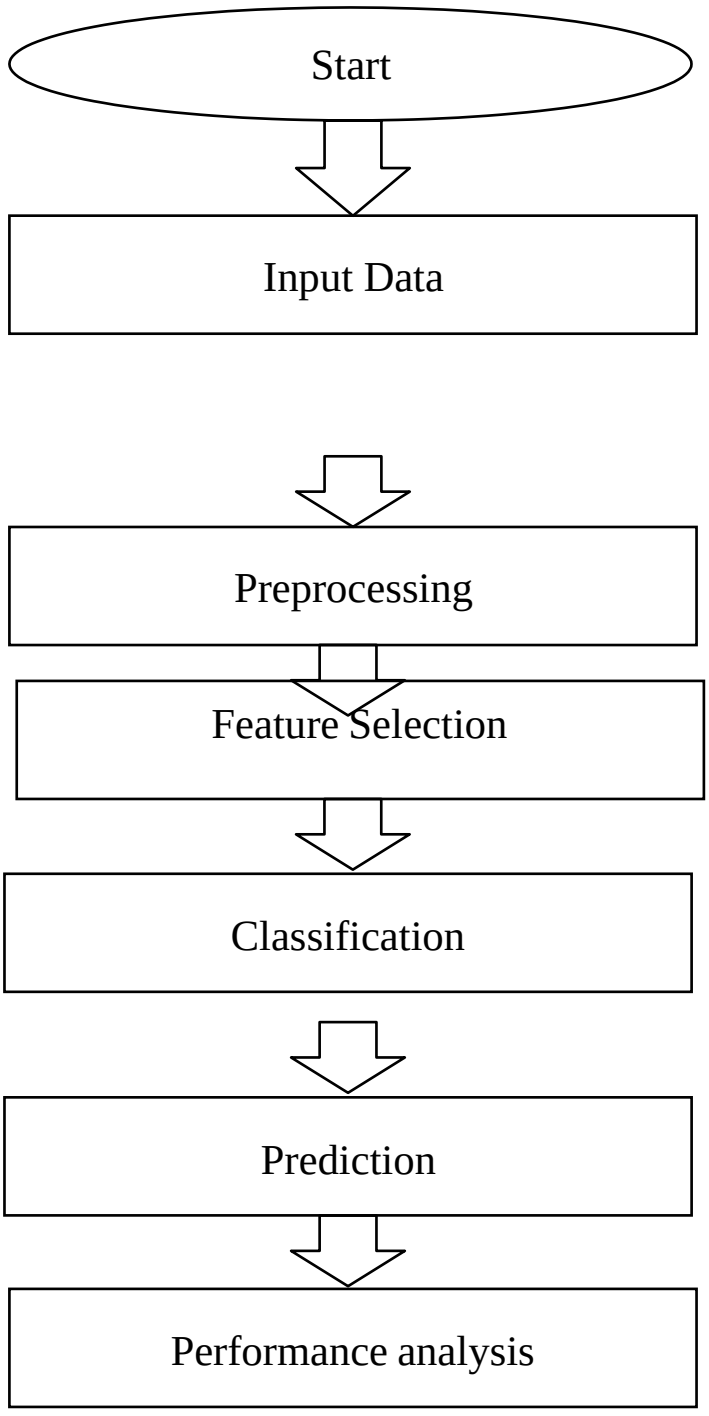
Processing speed is high.

Disadvantages:

Error rate is high.

CHAPTER 3 SYSTEM DIAGRAMS

SYSTEM ARCHITECTURE:



FLOW DIAGRAM

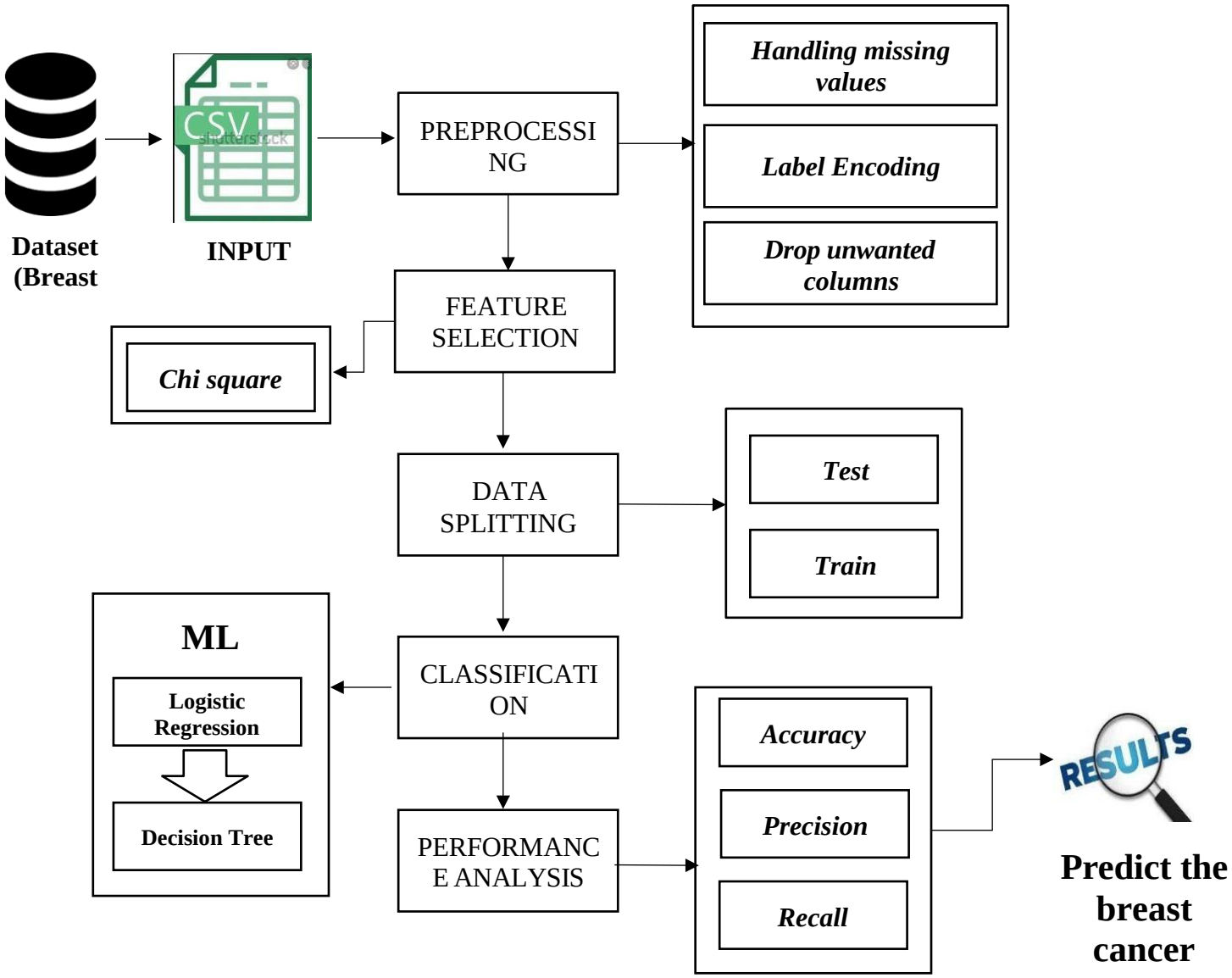


FIGURE 3.2: FLOW DIAGRAM

UML DIAGRAMS:

USE CASE DIAGRAM:

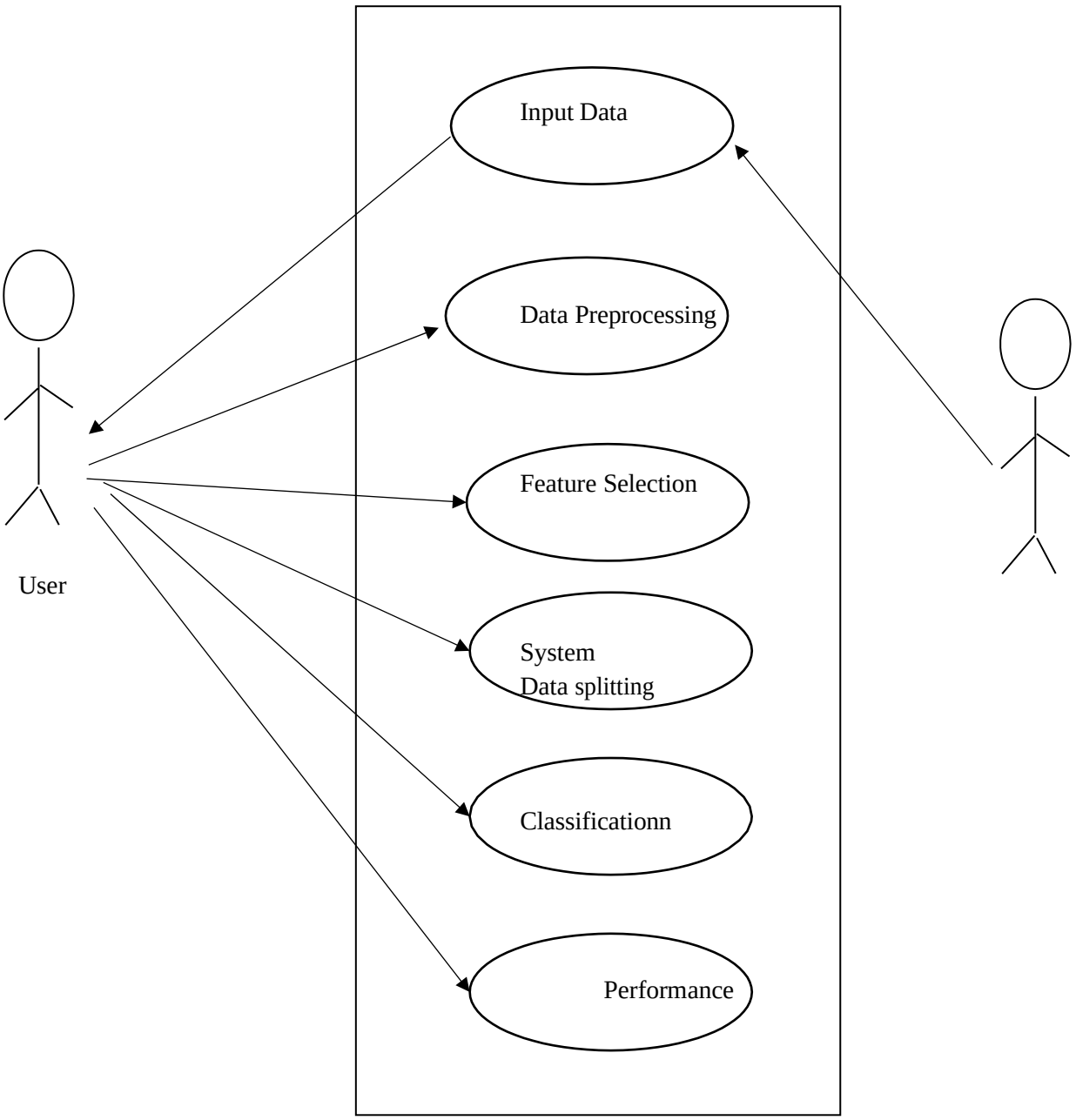


FIGURE 3.3.1: USE CASE DIAGRAM

ACTIVITY DIAGRAM:

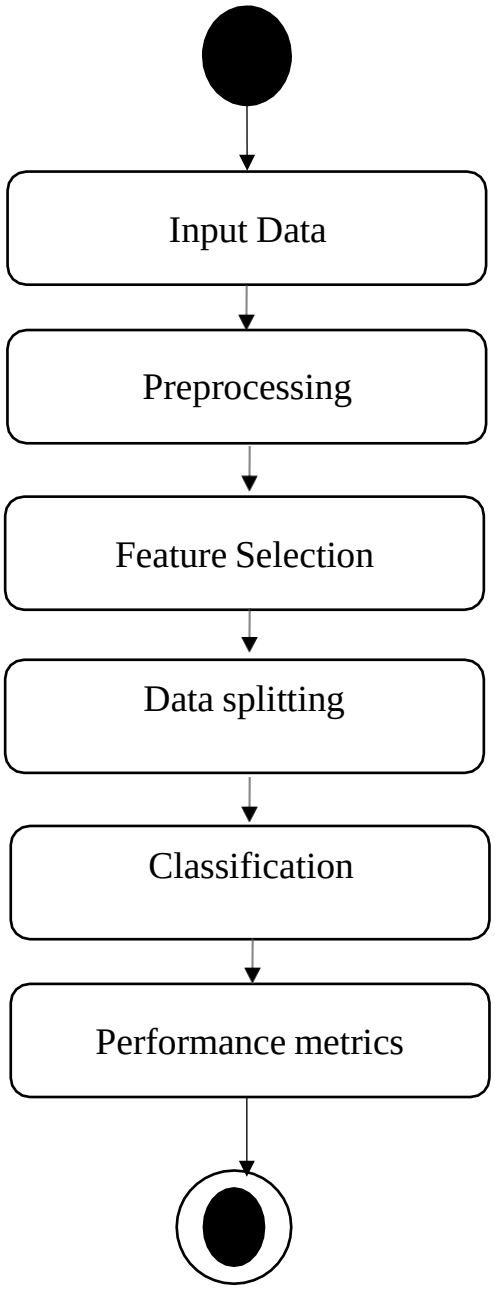


FIGURE 3.3.2: ACTIVITY DIAGRAM

SEQUENCE DIAGRAM:

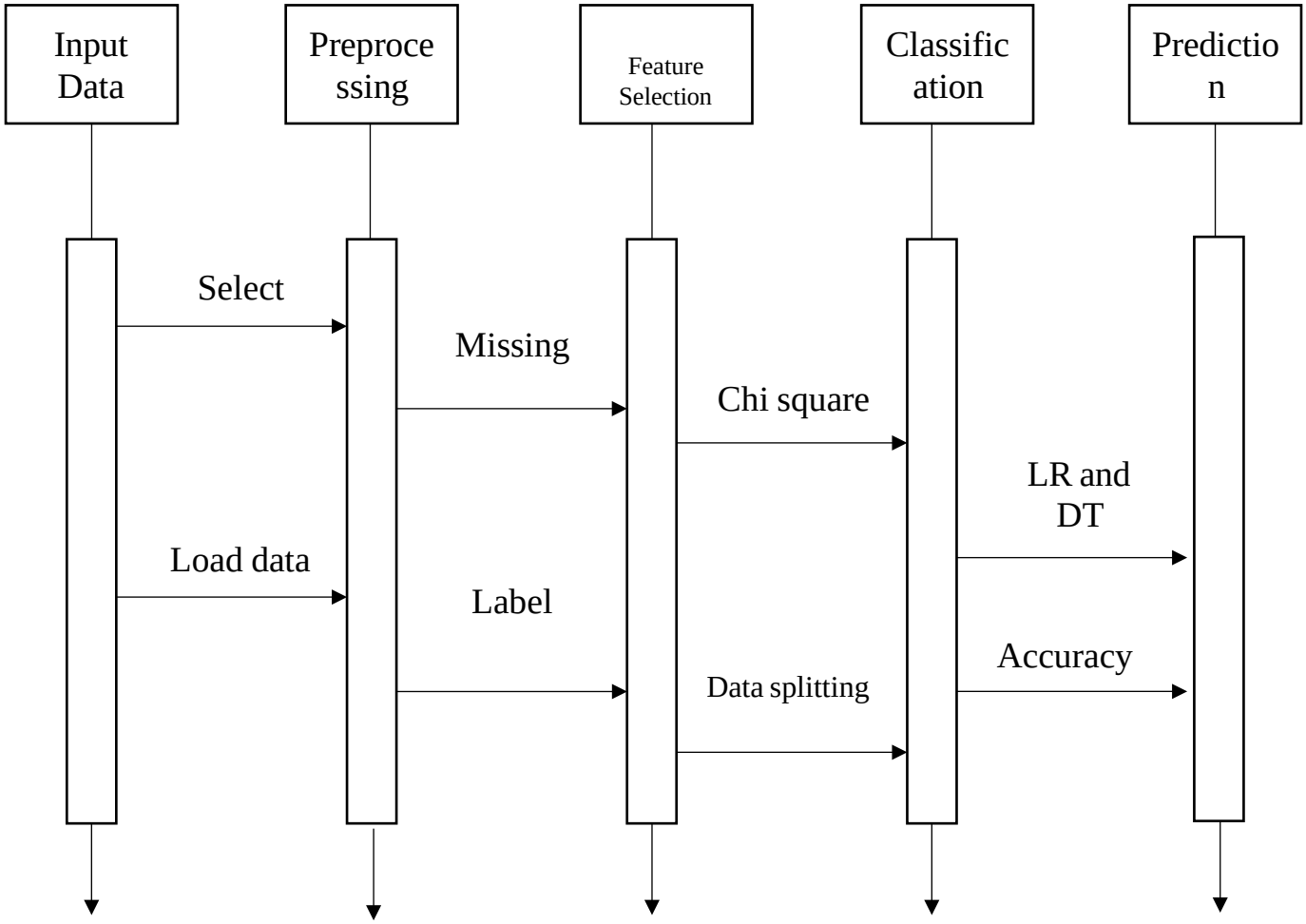


FIGURE 3.3.3: SEQUENCE DIAGRAM

ER DIAGRAM:

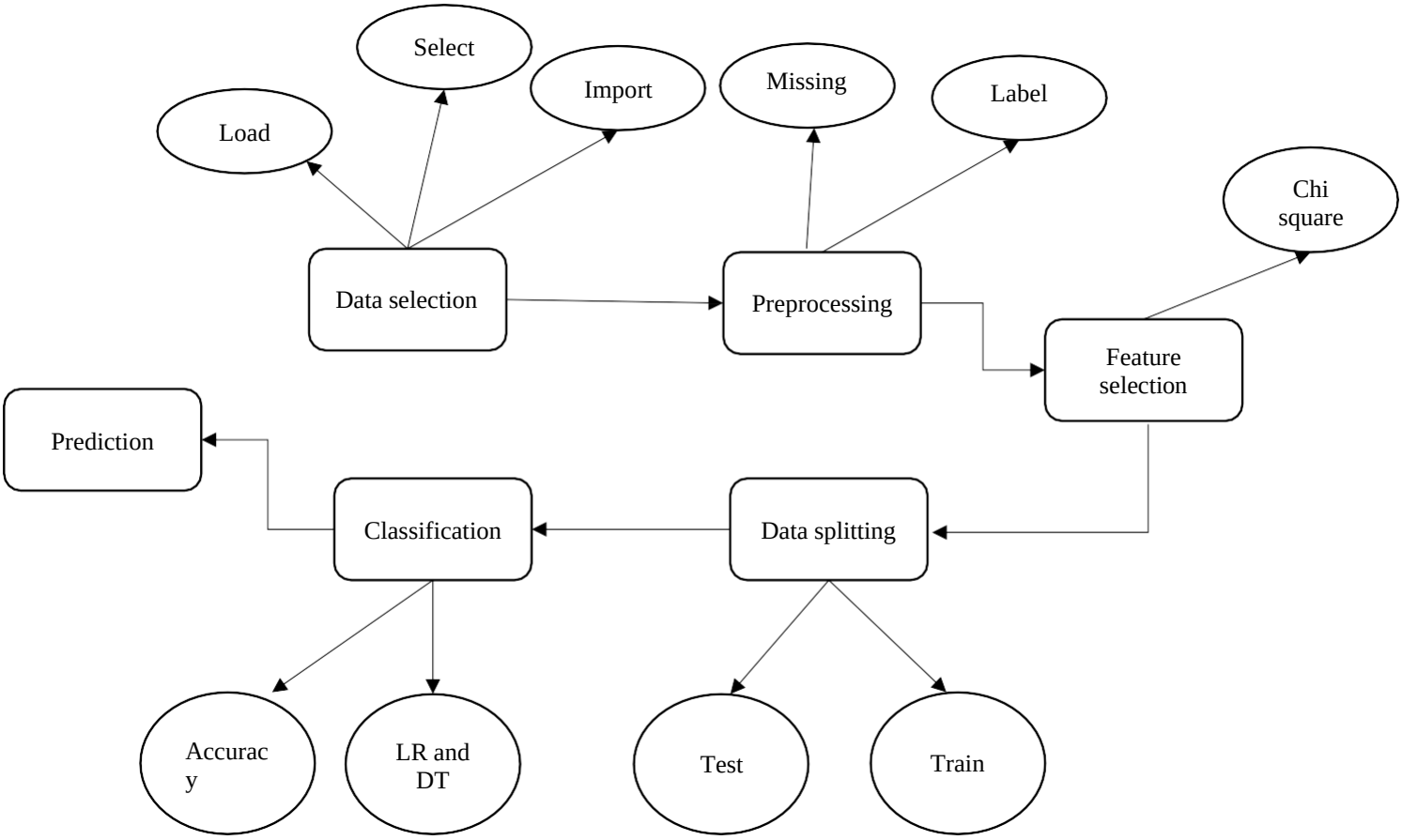


FIGURE 3.3.4: ER DIAGRAM

CLASS DIAGRAM:

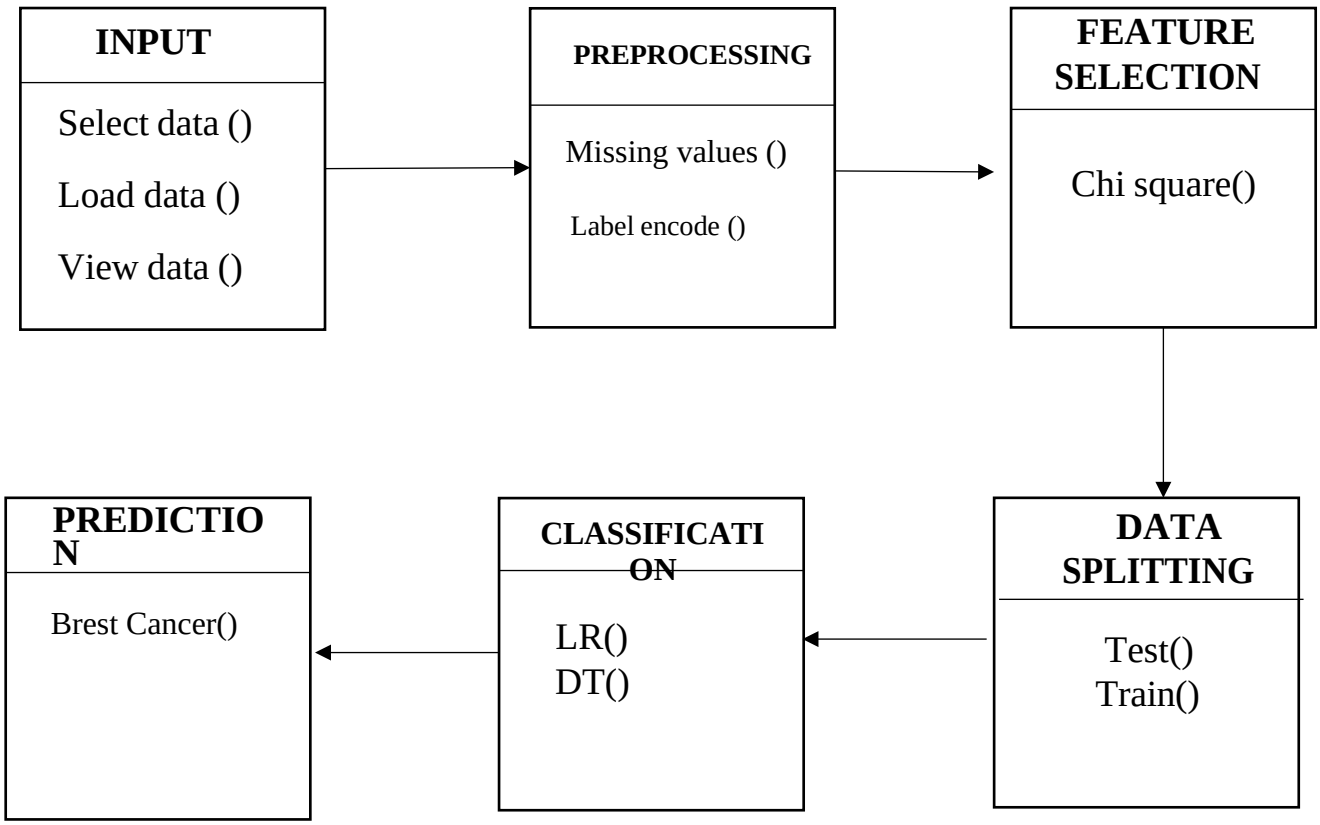


FIGURE 3.3.5: CLASS DIAGRAM

CHAPTER 4 IMPLEMENTATION

MODULES:

- Data selection
- Data preprocessing
- Feature selection
- Data Splitting
- Classification
- Prediction
- Performance Analysis

MODULES DESCRIPTION:

: DATA SELECTION:

- The data selection is the process of selecting the data for predict the cancer.
- In this project, we have to use the Wisconsin Breast Cancer dataset
- The dataset which contains the information about the Benign (B) or Malignant(M).
- In python, we have to read the dataset by using the panda's packages.
- Our dataset, is in the form of '.csv' file extension.

: DATA PREPROCESSING:

- Data pre-processing is the process of removing the unwanted data from thedataset.
- Pre-processing data transformation operations are used to transform the datasetinto a structure suitable for machine learning.
- This step also includes cleaning the dataset by removing irrelevant or corrupted data that can affect the accuracy of the dataset, which makes it more efficient.
- Missing data removal
- Encoding Categorical data
- Missing data removal: In this process, the null values such as missing valuesand Nan values are replaced by 0.
- Missing and duplicate values were removed and data was cleaned of anyabnormalities.
- Encoding Categorical data: That categorical data is defined as variables with a finite set of

label values.

- That most machine learning algorithms require numerical input and outputvariables.

FEATURE SELECTION:

- In our process, we have to implement the feature selection for selecting the bestfeatures such as chi square.
- A chi-square test is used in statistics to test the independence of two events. Given the data of two variables, we can get observed count O and expected count E.
- Chi-Square measures how expected count E and observed count O deviates each other.

DATA SPLITTING:

- During the machine learning process, data are needed so that learning can takeplace.
- In addition to the data required for training, test data are needed to evaluate theperformance of the algorithm in order to see how well it works.
- In our process, we considered 70% of the input dataset to be the training data and the remaining 30% to be the testing data.
- Data splitting is the act of partitioning available data into two portions, usually for cross-validator purposes.
- One Portion of the data is used to develop a predictive model and the other to evaluate the model's performance.
- Separating data into training and testing sets is an important part of evaluating data mining models.
- Typically, when you separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing.

CLASSIFICATION:

- In our process, we have to implement the two different machine learning algorithm such logistic regression and decision tree.
- **Logistic regression:** is **a** statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression)is estimating the parameters of a logistic model (a form of binary regression).
- A **decision tree** is a flowchart-like structure in which each internal node represents a test on a feature (e.g. whether a coin flip comes up heads or tails) ,

each leaf node represents a class label (decision taken after computing all features) and branches represent conjunctions of features that lead to those class.

RESULT GENERATION:

The Final Result will get generated based on the overall classification and prediction. The performance of this proposed approach is evaluated using some measures like,

- **Accuracy**

Accuracy of classifier refers to the ability of classifier. It predicts the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data.

AC= (TP+TN)/ (TP+TN+FP+FN)

- **Precision**

Precision is defined as the number of true positives divided by the number of truepositives plus the number of false positives.
Precision=TP/ (TP+FP)

- **Recall**

Recall is the number of correct results divided by the number of results that should have been returned. In binary classification, recall is called sensitivity. It can be viewed as the probability that a relevant document is retrieved by the query.
Recall=TP/ (TP+FN)

CHAPTER 5 SYSTEM REQUIREMENTS

HARDWARE REQUIREMENTS:

- System : Pentium IV 2.4 GHz
- Hard Disk : 200 GB
- Mouse : Logitech.
- Keyboard : 110 keys enhanced
- Ram : 4GB

SOFTWARE REQUIREMENTS:

- O/S : Windows 7.
- Language : Python
- Front End : Anaconda Navigator – Spyder

SOFTWARE DESCRIPTION:

Python

Python is one of those rare languages which can claim to be both *simple* and powerful. You will find yourself pleasantly surprised to see how easy it is to concentrate on the solution to the problem rather than the syntax and structure of the language you are programming in. The official introduction to Python is Python is aneasy to learn, powerful programming language. It has efficient high-level data structures and a simple but effective approach to object-oriented programming. Python's elegant syntax and dynamic typing, together with its interpreted nature, makeit an ideal language for scripting and rapid application development in many areas onmost platforms. I will discuss most of these features in more detail in the next section.

Features of Python

- **Simple**

Python is a simple and minimalistic language. Reading a good Python programfeels almost like reading English, although very strict English! This pseudo-code
nature of Python is one of its greatest strengths. It allows you to concentrate on thesolution to the problem rather than the language itself.

- **Easy to Learn**

As you will see, Python is extremely easy to get started with. Python has an extraordinarily simple syntax, as already mentioned.

- **Free and Open Source**

Python is an example of a *FLOSS* (Free/Libré and Open Source Software). In simple terms, you can freely distribute copies of this software, read its source code, make changes to it, and use pieces of it in new free programs. FLOSS is based on theconcept of a community which shares knowledge. This is one of the reasons why Python is so good - it has been created and is constantly improved by a community who just want to see a better Python.

- **High-level Language**

When you write programs in Python, you never need to bother about the low- level details such as managing the memory used by your program, etc.

- **Portable**

Due to its open-source nature, Python has been ported to (i.e. changed to makeit work on) many platforms. All your Python programs can work on any of these platforms without requiring any changes at all if you are careful enough to avoid any system-dependent features.

You can use Python on GNU/Linux, Windows, FreeBSD, Macintosh, Solaris, OS/2, Amiga, AROS, AS/400, BeOS, OS/390, z/OS, Palm OS, QNX, VMS, Psion,
Acorn RISC OS, VxWorks, PlayStation, Sharp Zaurus, Windows CE and PocketPC!
You can even use a platform like Kivy to create games for your computer *and*
for iPhone, iPad, and Android.

- **Interpreted**

This requires a bit of explanation.

A program written in a compiled language like C or C++ is converted from thesource language i.e. C or C++ into a language that is spoken by your computer (binarycode i.e. 0s and 1s) using a compiler with various flags and options. When you run theprogram, the linker/loader software copies the program from hard disk to memory and starts running it.

Python, on the other hand, does not need compilation to binary. You just *run* the program directly from the source code. Internally, Python converts the source code into an intermediate form called bytecodes and then translates this into the native language of your computer and then runs it. All this, actually, makes using Python much easier since you don't have to worry about compiling the program, making sure that the proper libraries are linked and loaded, etc. This also makes your Python programs much more portable, since you can just copy your Python program onto another computer and it just works!

- **Object Oriented**

Python supports procedure-oriented programming as well as object-oriented programming. In *procedure-oriented* languages, the program is built around procedures or functions which are nothing but reusable pieces of programs. In *object- oriented* languages, the program is built around objects which combine data and functionality. Python has a very powerful but simplistic way of doing OOP, especially when compared to big languages like C++ or Java.

- **Extensible**

If you need a critical piece of code to run very fast or want to have some piece of algorithm not to be open, you can code that part of your program in C or C++ and then use it from your Python program.

- **Embeddable**

You can embed Python within your C/C++ programs to give *scripting* capabilities for your program's users.

- **Extensive Libraries**

The Python Standard Library is huge indeed. It can help you do various things involving regular expressions, documentation generation, unit testing, threading, databases, web browsers, CGI, FTP, email, XML, XML-RPC, HTML, WAV files, cryptography, GUI (graphical user interfaces), and other system-dependent stuff. Remember, all this is always available wherever Python is installed. This is called the*Batteries Included* philosophy of Python.

Besides the standard library, there are various other high-quality libraries whichyou can find at the Python Package Index.

TESTING PRODUCTS:

System testing is the stage of implementation, which aimed at ensuring that system works accurately and efficiently before the live operation commence. Testing is the process of executing a program with the intent of finding an error. A good test case is one that has a high probability of finding an error. A successful test is one that answers a yet undiscovered error.
Testing is vital to the success of the system. System testing makes a logical assumption that if all parts of the system are correct, the goal will be successfully achieved. . A series of tests are performed before the system is ready for the user acceptance testing. Any engineered product can be tested in one of the following ways. Knowing the specified function that a product has been designed to from, test can be conducted to demonstrate each function is fully operational. Knowing the internal working of a product, tests can be conducted to ensure that “al gears mesh”, that is the internal operation of the product performs according to the specification and all internal components have been adequately exercised.

UNIT TESTING:

Unit testing is the testing of each module and the integration of the overall system is done. Unit testing becomes verification efforts on the smallest unit of software design in the module. This is also known as 'module testing'.
The modules of the system are tested separately. This testing is carried out during the programming itself. In this testing step, each model is found to be working satisfactorily as regard to the expected output from the module. There are some validation checks for the fields. For example, the validation check is done for verifying the data given by the user where both format and validity of the data entered is included. It is very easy to find error and debug the system.

INTEGRATION TESTING:

Data can be lost across an interface, one module can have an adverse effect on the other sub function, when combined, may not produce the desired major function. Integrated testing is systematic testing that can be done with sample data.The need for the integrated test is to find the overall system performance. There are two types of integration testing. They are:

i) Top-down integration testing.

ii) Bottom-up integration testing.

TESTING T E C H N I Q U E S /STRATEGIES:

- **WHITE BOX TESTING:**

White Box testing is a test case design method that uses the control structure of the procedural design to drive cases. Using the white box testing methods, we
Derived test cases that guarantee that all independent paths within a module havebeen exercised at least once.

- **BLACK BOX TESTING:**

1. Black box testing is done to find incorrect or missing function
2. Interface error
3. Errors in external database access
4. Performance errors.
5. Initialization and termination errors

In 'functional testing', is performed to validate an application conforms to its specifications of correctly performs all its required functions. So this testing is also called 'black box testing'. It tests the external behaviour of the system. Here the engineered product can be tested knowing the specified function that a product has been designed to perform, tests can be conducted to demonstrate that each function is fully operational.

SOFTWARE TESTING STRATEGIES

VALIDATION TESTING:

After the culmination of black box testing, software is completed assembly as a package, interfacing errors have been uncovered and corrected and final series of software validation tests begin validation testing can be defined as many, But a single definition is that validation succeeds when the software functions in a manner that can be reasonably expected by the customer

USER ACCEPTANCE TESTING:

User acceptance of the system is the key factor for the success of the system. The system under consideration is tested for user acceptance by constantly keeping in touch with prospective system at the time of developing changes whenever required.

OUTPUT TESTING:

After performing the validation testing, the next step is output asking the user about the format required testing of the proposed system, since no system could be useful if it does not produce the required output in the specific format.
The output displayed or generated by the system under consideration. Here the output format is considered in two ways. One is screen and the other is printed format. The output format on the screen is found to be correct as the format was designed in the system phase according to the user needs. For the hard copy also output comes out as the specified requirements by the user. Hence the output testing does not result in any connection in the system.

CHAPTER 6 CONCLUSION

We conclude that, the breast cancer dataset was collected from dataset repository as input. The input dataset was mentioned in our research paper.We are implemented or hybrid the two different classification algorithms (i.e.) machine learning algorithm. Then, machine learning algorithms such as logistic regression and decision tree. Finally, the result shows that the accuracy for above mentionedalgorithm and predict the breast cancer.

CHAPTER 7 FUTURE ENHANCEMENT

- In the future, we should like to hybrid the two different machine and deep learning algorithms.
- In future, it is possible to provide extensions or modifications to the proposed clustering and classification algorithms to achieve further increased performance.
- Apart from the experimented combination of data mining techniques, further combinations and other clustering algorithms can be used to improve the detection accuracy.

CHAPTER 8 SAMPLE CODING

```
//========================== Import Libraries
//==============================



import pandas as pd

from sklearn import preprocessing

from sklearn.feature_selection import SelectKBestfrom sklearn.feature_selection import chi2
from sklearn.model_selection import train_test_split from sklearn.linear_model import LogisticRegression from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import VotingClassifierfrom sklearn import metrics
#========================== Data Selection
==============================


data=pd.read_csv("data_1.csv") print()
print("_____Data Selection_____")
print() print(data.head(15))
#========================== Preprocessing
==============================

#==== checking missing values ====print()
print("----------  Checking Missing Values ----------")

print() print(data.isnull().sum())


#=== Label Encoding =====



label_encoder=preprocessing.LabelEncoder()



print()

print("----------  Before Label Encoding---------- ")

print() print(data['diagnosis'].head(15))
```

```
print()

print("----------------- After Label Encoding----------------------------------")

print() data['diagnosis']=label_encoder.fit_transform(data['diagnosis']) print(data['diagnosis'].head(15))
#============================ Feature Selection
==============================

x=data.drop("diagnosis",axis=1) y=data["diagnosis"]
print()

print("_____Feature selection_____")

print()

chi2_features = SelectKBest(chi2, k = 5) x_kbest= chi2_features.fit_transform(x, y)


print("No of original Features:",x.shape[1])
print()

print("No of reduced Features:",x_kbest.shape[1])


#============================ Data Splitting
==============================
X_train, X_test, Y_train, Y_test = train_test_split(x_kbest,y,test_size=0.3,random_state=40)


model1 = DecisionTreeClassifier(criterion = "gini", random_state =10,max_depth=100, min_samples_leaf=1)
model1.fit(X_train, Y_train)

y_pred1 = model1.predict(X_test)

#============================ Performance Analysis
==========================
print()

print("----------------- Performance Analysis for DT----------------------------------------")

print()


from sklearn import metrics

cm=metrics.confusion_matrix(y_pred1,Y_test)

#find the performance metricsTP = cm[0][0]
```

```
FP = cm[0][1]

FN = cm[1][0]

TN = cm[1][1]

#Total TP,TN,FP,FN Total=TP+FP+FN+TN
print()

print("1. Confusion Matrix :",cm)print()
#Accuracy Calculation accuracy1=(((TP+TN+FP)/Total)) *100
print("2.Accuracy :",accuracy1,'%')print()
#Precision Calculation precision=TP/(TP+FP)*100 print("3.Precision :",precision,'%')print()
#Sensitivity Calculation Sensitivity=TP/(TP+FN)*100 print("4.Sensitivity :",Sensitivity,'%')print()
#specificity Calculation specificity = (TN / (TN+FP))*100
print("5.specificity :",specificity,'%')print()
model2 = LogisticRegression(penalty = 'l2', random_state = 0)

model2.fit(X_train, Y_train)
y_pred = model2.predict(X_test)

#========================== Performance Analysis
==========================

print()

print("---------- Performance Analysis for LR ---------- ")

print()


from sklearn import metrics

cm=metrics.confusion_matrix(y_pred,Y_test)
#find the performance metricsTP = cm[0][0]
FP = cm[0][1]

FN = cm[1][0]

TN = cm[1][1]

#Total TP,TN,FP,FN
Total=TP+FP+FN+TN

print()

print("1. Confusion Matrix :",cm)print()
#Accuracy Calculation accuracy1=(((FN+TN)/Total)) *100 print("2.Accuracy :",accuracy1,'%')print()
```

```
#Precision Calculation precision=TP/(TP+FP)*100 print("3.Precision      :",precision,'%')print()
#Sensitivity Calculation Sensitivity=TP/(TP+FN)*100 print("4.Sensitivity :",Sensitivity,'%')print()
#specificity Calculation specificity = (TN / (TN+FP))*100
print("5.specificity :",specificity,'%')print()
#============================ Prediction
================================

print()

print("__________________Prediction__________________")

print()

y_predd=model2.predict([[926125,1347,118.8,179.1,1819]])

predicted_value=int(input("Enter the predicted value:"))print()
print()

print("                The predicted result is:                ")print()
if y_pred[predicted_value]==0: print("========================")
print() print("Benign/Normal") print()
print(" The patient is normal not affected by cancer ")print("========================")
else:

print("========================")

print()

print("Malignant Cancerous Cell")print()
print(" The patient was affected by breast cance. The affected cancerous isMalignant cancerous")
print("========================")
```

CHAPTER 9 SCREENSHOTS

```
--------------------  Data Selection  ----------------------

         id diagnosis  ...  symmetry_worst  fractal_dimension_worst
0    842302         M  ...          0.4601                  0.11890
1    842517         M  ...          0.2750                  0.08902
2   84300903        M  ...          0.3613                  0.08758
3   84348301        M  ...          0.6638                  0.17300
4   84358402        M  ...          0.2364                  0.07678
5    843786         M  ...          0.3985                  0.12440
6    844359         M  ...          0.3063                  0.08368
7   84458202        M  ...          0.3196                  0.11510
8    844981         M  ...          0.4378                  0.10720
9   84501001        M  ...          0.4366                  0.20750
10   845636         M  ...          0.2948                  0.08452
11  84610002        M  ...          0.3792                  0.10480
12   846226         M  ...          0.3176                  0.10230
13   846381         M  ...          0.2809                  0.06287
14  84667401        M  ...          0.3596                  0.14310

[15 rows x 32 columns]
```

```
---------------   Checking Missing Values  ------------------

id                        0
diagnosis                 0
radius_mean               0
texture_mean              0
perimeter_mean            0
area_mean                 0
smoothness_mean           0
compactness_mean          0
concavity_mean            0
concave points_mean       0
symmetry_mean             0
fractal_dimension_mean    0
radius_se                 0
texture_se                0
perimeter_se              0
area_se                   0
smoothness_se             0
compactness_se            0
concavity_se              0
concave points_se         0
symmetry_se               0
fractal_dimension_se      0
radius_worst              0
texture_worst             0
perimeter_worst           0
```

```
-----------------   Before Label Encoding  ------------------

0     M
1     M
2     M
3     M
4     M
5     M
6     M
7     M
8     M
9     M
10    M
11    M
12    M
13    M
14    M
Name: diagnosis, dtype: object
```

```
-----------------    After Label Encoding  --------------------

0     1
1     1
2     1
3     1
4     1
5     1
6     1
7     1
8     1
9     1
10    1
11    1
12    1
13    1
14    1
Name: diagnosis, dtype: int32
```

```
------------------    Feature selection  ------------------

No of original Features: 31

No of reduced Features: 5
```

```
----------------- Performance Analysis for DT ----------------


1. Confusion Matrix : [[103   6]
 [  6  71]]

2.Accuracy    : 96.7741935483871 %

3.Precision   : 94.4954128440367 %

4.Sensitivity  : 94.4954128440367 %

5.specificity  : 92.20779220779221 %
```

```
-----------------   Prediction  ------------------

Enter the predicted value:5

             The predicted result is:

========================

Benign/Normal

 The patient is normal not affected by cancer
========================
```

CHAPTER 10REFERENCES

1. Y.-S. Sun, Z. Zhao, Z.-N. Yang, F. Xu, H.-J. Lu, Z.-Y. Zhu, W. Shi, J. Jiang, P.-P. Yao, and H.-P. Zhu, ‘‘Risk factors and preventions of breast cancer,’’Int. J. Biol. Sci., vol. 13, no. 11, p. 1387, 2017.

2. Y. Khourdifi and M. Bahaj, ‘‘Applying best machine learning algorithms for breast cancer prediction and classification,’’ in Proc. Int. Conf. Electron., Control, Optim. Comput. Sci. (ICECOCS), Dec. 2018, pp. 1–5.

3. Y. Lu, J.-Y. Li, Y.-T. Su, and A.-A. Liu, ‘‘A review of breast cancer detection in medical images,’’ in Proc. IEEE Vis. Commun. Image Process. (VCIP), Dec. 2018, pp. 1–4.

4. F. K. Ahmad and N. Yusoff, ‘‘Classifying breast cancer types based on fine needle aspiration biopsy data using random forest classifier,’’ in Proc. 13th Int. Conf. Intellient Syst. Design Appl., Dec. 2013, pp. 121–125.

5. R. Hou, M. A. Mazurowski, L. J. Grimm, J. R. Marks, L. M. King, C. C. Maley, E.-S.-S. Hwang, and J. Y. Lo, ‘‘Prediction of upstaged ductal carcinoma in situ using forced labeling and domain adaptation,’’ IEEE Trans. Biomed. Eng., vol. 67, no. 6, pp. 1565–1572, Jun. 2020.

6. A. R. Chaudhury, R. Iyer, K. K. Iychettira, and A. Sreedevi, ‘‘Diagnosis of invasive ductal carcinoma using image processing techniques,’’ in Proc. Int. Conf. Image Inf. Process., Nov. 2011, pp. 1–6.

7. T. Evgeniou and M. Pontil, ‘‘Support vector machines: Theory and applications,’’ in Advanced Course on Artificial Intelligence. Berlin, Germany: Springer, 2005, pp. 249–257.

8. M. D. Ganggayah, N. A. Taib, Y. C. Har, P. Lio, and S. K. Dhillon, ‘‘Predicting factors for survival of breast cancer patients using machine learning techniques,’’ BMC Med. Inform. Decis. Making, vol. 19, no. 1, 2019, Art. no. 48.

9. Ponnuraja CC, Lakshmanan B, Srinivasan V, Prasanth BK. Decision Tree Classification and Model Evaluation for Breast Cancer Survivability: A DataMining Approach. Biomed Pharmacol J. 2017;10:281–9.

10. Malehi AS. Diagnostic classification scheme in Iranian breast cancer patientsusing a decision tree. Asian Pac J Cancer Prev. 2014;15:5593–6.

11. Shrivastava SS, Sant A, Aharwal RP. An overview on data mining approach on breast Cancer data. Int J Adv Comput Res. 2013;3(4):256–62.

12. Islam T, Bhoo-Pathy N, Su TT, Majid HA, Nahar AM, Ng CG, et al. The Malaysian breast Cancer survivorship cohort (MyBCC): a study protocol. BMJ Open Br Med J Publ Group. 2015;5:e008643.

13. Taib NA, Akmal M, Mohamed I, Yip C-H. Improvement in survival of breastcancer patients - trends over two time periods in a single institution in an Asia Pacific country, Malaysia. Asian Pac J Cancer Prev. 2011;12:345–9.

14. Leong SPL, Shen ZZ, Liu TJ, Agarwal G, Tajima T, Paik NS, et al. Is breast Cancer the same disease in Asian and Western countries? World J Surg. 2010;34:2308–24.

15. Bhoo-Pathy N, Verkooijen HM, Tan E-Y, Miao H, Taib NAM, Brand JS, et al. Trends in presentation, management and survival of patients with de novometastatic breast cancer in a southeast Asian setting. Sci Rep. 2015;5:16252.