A

DISSERTATION REPORT

ON

"Visual Healthcare Analytics using Adaptive data mining"

Submitted in partial fulfillment for the award of

Master of Engineering

In

Computer Science and Engineering

SolapurUniversity

By

Miss. Priyanka B. Shivagunde

Under the Guidance of

Dr. Ms. A. R. Kulkarni

Department of Computer Science & Engineering,

Walchand Institute of Technology

Solapur 413006.

(2016-17)

CERTIFICATE

This is to certify that the Dissertation work entitled

"Visual Healthcare Analytics using Adaptive data mining"

Has been satisfactorily completed and submitted by

Miss. Priyanka B. Shivagunde

In partial fulfillment of requirement for the award of the degree of Master of Engineering

In

Computer Science & Engineering

As per rules laid down by Solapur University, Solapur

This report is the record to her bona fide work carried out under my supervision and guidance

(Prof. Dr. Mrs. A. R. Kulkarni) Guide	(Prof. Mr. R. V. Argiddi) Head Dept. of Computer Sc. & Engg.	(Prof. Mr. R. V. Argiddi) M. E. Coordinator Dept. of Computer Sc. & Engg.

(Dr. S. A. Halkude)

Principal

Walchand Institute of Technology, Solapur

ACKOWLEDGEMENT

A project is always coordinated and scheduled effort, but it can never reach completion without proper guidance and encouragement. At the outset I would like to take this opportunity to express my deep gratitude to my guide Dr. Prof. Mrs. A. R. Kulkarni to help me decided on for the research problem and its feasibility study and concluding the problem statement. I would like to thank, my guide for being the source of inspiration and to have shown tremendous faith in me. Her guidance has been and shall be a source of huge encouragement in future. I am especially thankful for her patience in resolving the queries.

I would like to thank our Head of Department, Computer Science and Engineering, Prof. Mr. R. V. Argiddi for the support they have given using the form of infrastructure and facilities. Thank Dr. S. A. Halkude, Principal and the entire staff member for their whole hearted cooperation. I would also like to thank our college staff and software laboratory assistant for their valuable help in the laboratory.

I especially thanks to admin officer and staff of Shree Siddheshwar Cancer Haspital and research Center, Solapur for providing me records of cancer patients those I needed as training datasets.

Last but not least, the backbone of my success and confidence lies solely on blessings of my family members. I express my gratitude towards those who have directly and indirectly help towards the project.

Date: /0 /2017 Miss. Priyanka B. Shivagunde

Place: Solapur M. E. (Computer Sci. & Engg.)

TABLE OF CONTENTS

LIST OF FIGURE

LIST OF TABLE

Abstract

Health is most valuable factor affecting our life. People are highly focused on the healthcare with high preference. Now-a-days there are so many fetal diseases occurring in individuals, Cancer is one of those fetal disease which is cause of death of several peoples in year and from those breast cancer is major cause in women death. According to recent survey about 1 in 8 women (about 12 percent) have breast cancer. Diagnosis of disease is usually done in last stage and hence cannot be cured by treatment. Early diagnosis of such diseases is essential and a significant factor

For early detection regular check up should be done by women above 40 years of age. There is one solution to this problem that we can provide a system which automatically diagnose the disease that can check by individual person, caretaker, friend or family member which is more feasible. a system can be provided which can be automatically diagnose the disease by analyzing patient's biomedical data and find out existence of breast cancer in the patient. The genetic algorithm Olex GA used to classify patient in different stages as per her symptoms and test reports. Genetic Olex algorithm is a text based classification algorithm. The final report is generated which is visualized by bar graph. So due to this visualized report the non-medical individual also understand the result. Also one more feature of this system is adaptation. New symptoms ant tests are saved in database and those will train manually. This approach helps the patients, doctors and family members to find out.

Chapter 1

INTRODUCTION

1.1) Need of dissertation

1.1.1) Breast Cancer

In human body cell is unit of building structure. Tissue is set of similar type of cells. Lifetime of cells are limited, when a new cell produces, an old cell is destroyed that place is taken by new cell. This happens regularly in human body. But if old cells do not got destroyed or new cells are generated irregularly then there is lump or cyst or tumor occurs in human body. These cells may be carcinoma cells.

In case of breast cancer most of the time carcinoma cells are developed from milk cells. There are different causes in women for occurrence of breast cancer. Some threat factors are as follows:

- Female
- Less breastfeeding
- Late pregnancy after 30 years old
- Hormonal therapy
- Menopause
- Radiation therapy done on organs near breast e.g. on chest etc.
- Nearest blood relatives have breast cancer

There are two types of tumors in human body:benign and malignant tumor

Benign tumor is not harmful because it do not got spread to other organs near to that. By surgery if it is removed then there is less chances to come back. Malignant tumor is harmful because it gets spread to other organs and if that is removed by surgery then chances of grow back is higher as compared to benign tumor.

In breast cancer there are three stages T-stage, N-stage and M-stage. In T-stage breast cells are irregular and tumor is exist and tumor tissue consists of carcinoma cells. In this stage patient is not serious, it can cured by surgery. In

N-stage carcinoma cells are spreaded to lymph nodes near breast. Lymph nodes contain cancerous tumor. In this stage also by chemotherapy this can be cured. In M-stage the carcinoma cells are spreaded from breast to other organs like hand, brain, lung, uterus, cervix, liver, throat, leg, chest etc. M-stage is last stage of breast cancer and in this stage survivability of patient is very low, by chemotherapy treatment can be done but patient may not give response to treatment carcinoma cells getting spreading and organs may stop working this will cause of patient's death.

From above description we can conclude that early diagnosis of disease in the patient in first stage is essential. Survivability of patient is high in first and second stage. The solution of this problem is make diagnosis easy, such that patient can check her health status in home. Individual person, her caretaker, friends or family members can check that individual is suffered from breast cancer or not. So such automatic detection system is necessary. This Dissertation helps to patient for early detection.

1.1.2) Importance of health care services

Now-a-days health is most precious thing in human's life, Due to his lifestyle and environment or some genetic factors, many diseases are occurring in human and the average lifetime of human is got decreased. So people should strictly take care of his health.

Healthcare services have fast growing trends and in that automatic detection of health status is necessary. There are so many health care services are already exist, each service used different methodologies and so that they have some advantages and some disadvantage. The accurate result in these types of services is very important. The technique which gives the accurate health status and explanation of status is necessary.

1.1.3) **Importance of Visualization and Adaptation**-

In previous techniques the report is generated only in textual format which may not understand to the non-medical person. This visualized report also mention that exactly where the carcinoma cells are spreaded, which is cause of exact stage. This identifies relation between the spreading of carcinoma cells and current stage of breast cancer in patient. From the both content (area where carcinoma cells spreaded and stage) patient can understand her health status correctly

1.2) Techniques used in healthcare services

In healthcare services there are human is classified into healthy and unhealthy person, for this classification different classification algorithms are implemented until now. These algorithms have two phases one is training and other is testing. While training data there is classifiers are used by each algorithm. There are two main types of techniques for classification-

1. Supervised technique
2. Unsupervised technique

1.2.1 Supervised learning-

Supervised learning is a machine learning task of inferring a function from labeled training data. Training data is set of training examples.

In this type of learning, each example is a pair consisting of an input object and a desired output value.This type of learning algorithm analyzes the training data and produces inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable".

Working steps:

I. Determine the types of training examples

II. Gather the training data

III. Determine the input feature representation of the learned function.

IV. Determine the structure of the learned function and corresponding learning algorithm.

V. Complete the design

VI. Evaluate the accuracy of learned function.

The wide range of supervised learning algorithms are available, each with its strength and weaknesses.

The supervised learning classification algorithms are wide range, some of these are used in field of medical diagnosis. These are as follows,

- C4.5
- k-Nearest Neighbor algorithm
- Support Vector Machine
- Neural Network

C4.5 Decision tree algorithm:

Used to generate a decision tree. The decision tree generated by C4.5 can be used for classification. Training data contains samples of already classified examples. Each sample Si consists of p-dimensional vector, where each vector represents attribute values or features of the sample as well as the class in which the sample falls. At each node of tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched into one class or other.

Advantages and disadvantages: The main advantage of C4.5 is that possibility to select best among given set of classes for sample. Drawback of this algorithm is complexity of building decision tree is high.

K-Nearest Neighbor algorithm:

Nonparametric c method used for classification. Input is k-closest training examples and output is class membership. Training examples are vectors in multidimensional feature space. The training phase of algorithm consists of storing the feature vectors and class labels of training samples. In classification phase distance metric is used.

Advantages and disadvantages: It is robust to noisy training data. The main drawback of this algorithm is computational cost is very high because we need to compute distance of each query instance of all training samples.

Support Vector Machine:

SVM training algorithm builds a model that assigns new examples into one category or the other. An SVM model is representation of the examples as points in space mapped, so that the examples of the separate categories are divided by a clear gap that is as wide as possible.

Advantage and disadvantage: The main advantage of this algorithm is helpful in text and hypertext categorization. It also useful in in classification of images. Handwritten characters are also recognized by SVM.SVM can produce accurate and robust classification result.

Neural Network:

Artificial neural networks are a family of models inspired by biological neural networks, which are used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown. Artificial neural network are typically specified using three things

Architecture: Describe variable and their topologies

Activity rate: change in responses between neurons.

Learning rate: A way in which neural network works.

Advantage and disadvantages: Advantage of neural network are good performance in nonlinear statistical modeling and provide logistic classification. Disadvantage includes its black box nature and greater computational Burdon.

1.2.2Unsupervised Method

Unsupervised learning technique is the machine learning task of inferring function to describe hidden structure from unlabeled data.

k- Means Algorithm:

K-means algorithm used to partition n-observations into k-clusters, in which each observation belongs to the cluster with the nearest mean.

Advantage and disadvantage: Advantage of k-means algorithm is that it yet faster when variables are huge. The drawback of this algorithm is that difficult to predict k- value and with global structure it did not work well.

Anamoly detection:

Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior. These non-conforming patterns are often referred to as anomalies, outliers, discordant observations, exceptions, aberrations, surprises, peculiarities or contaminants in different application domains.

Advantage and disadvantage: high processing speed is an advantage of this method but drawback of using AUC might be that it is not ideal for unbalanced class problems and methods like area under precision-recall curve or Matthews correlation coefficient could possibly better emphasize small detection performance changes.

Hebbian learning:

in this method the analysis of their performance in classifying known patterns, a sensitivity analysis of the applied classification scheme, regarding some configuration parameters have taken place

1.3 Structure of report:

This dissertation focused on the project automatic detection of breast cancer using patient's biomedical data. In this first chapter is introduction, describes need of dissertation and information about breast cancer and its stages. Further algorithmused for the detection is described. Various detection techniques are explained with their advantages and disadvantages.

The second chapter in this dissertation is literature review in which implemented healthcare services using data mining are described with their advantages. Further objective and scope of project is mentioned. Literature review will end with problem statement of project.

Chapter 3 is methodology, deep explanation of Olex GA algorithm consists in thischapter with working technique and pseudo code.This chapter end with conceptual overview of system.

Further in chapter 4 Data flow diagrams of project is explained. Some UML diagrams are described in chapter 5, system architecture, use case diagram, class diagram, sequence diagram and activity diagram.

Implementation is explained in detail in chapter 6, which describes all javaclasses and webpages and their use in project. Snapshots of webpages are also included in this chapter.

Result and analysis of project is described in chapter 7 taking different cases. The comparison with results of C4.5 is analyzed by taking time and accuracy as metrc. Conclusion and future work is mentioned in chapter 8.

Chapter 2

LITERATURE REVIEW

Healthcare is an important factor of today's life. Healthcare services are coming with more facilities. Up till now many researchers are worked to implement data mining algorithm in healthcare services.

[12]Andrew kusiak have used preprocessing of data, transformation of data and a data mining approach to elicite knowledge about the interaction between parameters measured and survival of patient, for extraction of knowledge in decision rule, there are two different data mining algorithms used. Decision making algorithm uses those rules which predict survivability of new patients. They have introduced an approach in their work have been applied and testing is done using collected data. The presented approach reduces effort as well as cost of selecting patients.

[13]Sadik kara had concentrated on diagnostic research of the neural disease using pattern electrographs signals by using artificial neural network. The final result was classified in healthy and diseased. The final result has shown with effective interpretation.

[14]Abhishek proposed a system which uses double neural network technique, back propagation algorithm, support vector machine, radial basis function and their accuracy and efficiency were compared. WEKA 3.6.5 tool is used to implement the best technique among above three algorithms for diagnosis of kidney stone. Aim of author was proposing best tool for diagnosis e.g. identification of kidney stone, by reducing requirement of time for diagnosis and accuracy as well as efficiency got improved.

[15]Ashfaq ahmad k presented a thesis using machine learning technique like random forest and support vector machine. Results of above both algorithms

were compared for different datasets such as heart disease dataset, liver disease dataset and breast cancer disease dataset. By good learning technique efficient results can analyzed for the purpose of prediction

[16]Basma baukenze presented the big data evaluation in healthcare system and they applied a learning algorithm on a set of medical data. Aim of author is predicting chronically kidney disease by using C 4.5 decision tree algorithm is used to improve performance of prediction of results in terms of minimum execution time and accuracy.

[18]Abdul rahim proposed a model BDCaM facilitates the analysis of big data inside a cloud environment. It first mines the trends and patterns in the data of an individual patient with associated probabilities and utilizes that knowledge to learn proper abnormal conditions. The outcome of this learning method applied in context aware decision making process for the patient. the use case is implemented to illustrate the applicability of the framework that discovers knowledge of classification to identify the true abnormal condition. The accuracy and efficiency obtained for the implemented case study demonstrate the effectiveness of proposed model

[19]Liqing nie presented a novel scheme to code the medical records jointly utilized local mining and global learning approaches which are tightly linked and usually reinforced. Local mining attempts to code the individual's medical record by independently extracting the medical concepts from the medical record itself and then mapping them to authenticated terminology vocabulary is naturally constructed as a byproduct, which is used as a terminology space for global learning comprehensive experiment well validate the proposed scheme and each of its components.

[20] Student of department of computer Engineering of MITCOE, Pune presents the innovative wireless sensor network based Mobile Real-time Health care Monitoring (WMRHM) framework which has the capacity of giving health predictions online based on continuously monitored real time vital body signals.

[21]kung siau proposed new business strategy requires health care organizations to implement new technologies, such as Internet applications, enterprise systems, and mobile technologies in order to achieve their desired business changes. This article offers a conceptual model for implementing new information systems, integrating internal data, and linking suppliers and patients.

[22] Marek Laskowski proposed methodology uses the supervised classification technique; the Olex GA text based classification algorithm. Texts are parsed from patient's biomedical data and classifies patient in to particular stage of disease. Advantage of this technique is more accuracy because it concentrates on texts present in data. Visualization of result by graph is implemented in this research.

2.1 Objective and scope:

Objective:

Objective of this dissertation is to design diagnostic system of the disease breast cancer (Ca. Breast) more accurately and easily by using patient's biomedical data.

To provide the visualization of diagnosis report in a format that is easily understood by the individual from nonmedical background.

To make the system adaptive to the new symptoms and diagnostic tests by dynamically training the system accordingly.

To compare the performance of the implemented system with the existing system using accuracy and time as a matrices.

Scope:

This technique can use for diagnosis of other diseases like diabetes, heart diseases, kidney diseases, blood pressure, lung disease etc. by changing the training data, the system can be extended to automatically adopt to new symptom and test.

Input:

Datasets of patient's biomedical record including various test reports and symptoms.

Output:

Report of patient gives the diagnosis of disease and also the current stage of the disease.

Hardware requirement:

RAM 1GB and above.

Software requirement:

JDK 8

Netbeans IDE 8 and above

Operating system (Win XP/7/8)

Apache tomcat server

Weka 3.6

Database required

MySQL

2.2 Problem Statement

“Visual health care analytics using adaptive data mining”

Most of the work is done in automatic diagnosis of disease using different classifiers such as C4.5, k- Nearest neighbor, SVM, neural network etc.

The aim of this work is to study and implement automatic diagnosis of breast cancer using patient’s biomedical data. The technique used for this is Olex GA algorithm.

The final report has shown in visualized format by using java template, which gives the diagnosis of disease and also the current stage of the disease. In previous techniques the report were generated only in textual format which was not understood by the individual with the non-medical background. This visualized report also extend to which carcinoma cells has spread. Depending on this the decision is taken about current stage of breast cancer.

From the both contents (area where carcinoma cells present and current stage) patient can understand their health status clearly. When new symptom or diagnostic test is detected in patient system will train manually adopt to the new symptoms. The adaptation method have implemented in this project

Chapter 3

METHODOLOGY

3.1) Dataset:

Data is necessary to solve the problem and this data is collected from hospital. For this project the dataset required is records of patients who have done breast cancer diagnosis in hospital. This data is collected from 'Shri Sidhdeshwar cancer hospital and research center, Solpaur'. 100 records of patients have collected for training purpose. These patients may have breast cancer and they have done diagnosis in hospital. 50 records of patients have collected for testing purpose.

The training data required for the algorithm is collected. This data contain records of patients of breast cancer , such as:Name, Age, Height, Weight, Blood Pressure, Heart Rate, Symptoms, Tests, Corresponding test reports, Disease and Stage

The dataset has been collected from Shri. Siddheshwar Cancer Hospital and Research Center, Solapur.

We have collected 100 records of breast cancer patients from this hospital. These 100 records includes patients with different stages: T, N and M stages.

3.2) Genetic Olex algorithm

The Olex GA algorithm works in two phases : Training and Testing.

(A) Training Phase

The algorithm starts with retrieving datasets from training data All the datasets are retrieved which are already stored as per different stages of Cancer:

1. T-stage datasets
2. N-stage datasets
3. M-stage datasets

first all the texual data from the dataset are retrieved.

The population and generation are predefined. In this method if number of records in datasets is small then for more accuracy the generation should be high. If the number of records is high then number of generations should be less to achieve accuracy.

As per population is predefined those number of texts are selected and tries to predict feture of every texts whether positive or negative. This operation of prediction is repeated number of times as per predefined generation. Further there is cross checks are done by using redundant texts.

After cross checking, redundant texts are eliminated and set of chromosomes get build.

- ChromosomeT- set of positive and negative chromosomes in stage T.
- ChromosomeN- set of positive and negative chromosomes in stage N.
- ChromosomeM- set of positive and negative chromosomes in stage M.

Now the classification model got built.

The task of finding Pos and Neg which maximize the F measure when H_c(Pos,Neg) is applied to the training set. MAX F can be represented as a 0 1 combinatorial problem

(B) Testing Phase

In this stage the patient's biomedical data is collected.

- Name
- Age
- Height
- Weight
- Blood Pressure
- Heart rate
- Symptoms
- Tests
- Test Reports

Now Olex GA algorithm checks the texts present in patient's biomedical data, while testing this algorithm works as follows,

If all the texts from patient's biomedical data match with positive chromosome then the patient is healthy and she does not have breast cancer.

But if atleast on negative chromosome exist in patient's biomedical data then system detects the patient to breast cancer. Futher by using disease classifier stage of breast cancer is detected.

The logic used by disease classifier as follows,

Classify document d under category c if t1 belongs to d or t2 belongs to d or …or tn belongs to d and not (t_{n+1} belongs to d or … tm belogs to d) holds

Where each t_i is a term

Olex GA adopts an efficient approach that "several rules per individual" binary representation and uses F measure as a fitness function

Olex Genetic Algoithm is a text based classification algorithm. Text classification is a task of assigning natural language texts to one or more thematic categories on the basis of their contents. Genetic algorithm is a random probability distribution or pattern analysis search method inspired to the biological evaluation. The basic idea is that each individual encodes a candidate solution (i.e., a classification rule or a classifier), and that its fitness is evaluated in terms of predictive accuracy.

The problem of inducing propositional text classifiers of the form

$$c \leftarrow (t1 \in d \vee \cdots \vee tn \in d) \wedge \neg(tn{+}1 \in d \vee \cdots \vee tn{+}m \in d)$$

where c category

d document

each t_i – a term taken from given probability

c classifier H_c (Pos, Neg)

Pos(t1,t2,…,tn)

Neg(tn+1, tn+2,…,tn+m)

Positive terms in Pos are used to cover the training set of c, while negative terms in Neg are used to take precision under control

Olex GA:

Classify document d under category c if t1 belongs to d or t2 belongs to d or …or tn belongs to d and not (t_{n+1} belongs to d or … tm belongs to d) holds

Where each t_i is a term

Olex GA adopts an efficient approach that "several rules per individual" binary representation and uses F measure as a fitness function

Text classification is a task of assigning natural language texts to one or more thematic categories on the basis of their contents. Genetic algorithm is a random probability distribution or pattern analysis search method inspired to the biological evaluation. The basic idea is that each individual encodes a candidate solution (i.e., a classification rule or a classifier), and that its fitness is evaluated in terms of predictive accuracy.

The problem of inducing propositional text classifiers of the form

$c \leftarrow (t1 \in d \vee \cdots \vee tn \in d) \wedge \neg(tn+1 \in d \vee \cdots \vee tn+m \in d)$

where c category

d document

each t_i – a term taken from given probability

c classifier H_c (Pos, Neg)

Pos(t1,t2,…,tn)

Neg(tn+1, tn+2,…,tn+m)

Positive terms in Pos are used to cover the training set of c, while negative terms in Neg are used to take precision under control

The task of finding Pos and Neg which maximize the F measure when H_c(Pos,Neg) is applied to the training set. MAX F can be represented as a 0 1 combinatorial problem

Once the population of individuals has been suitably initialized, evaluation takes place by iterating elitism, selection, crossover and mutation until the predefined numbers of generations are created.

Algorithm Olex GA

Input: vocabulary V (f, k) over the training set *TS*; number n of generations;

Output: "best" classifier Hc(*Pos*,*Neg*) of c over *TS*;

begin

Evaluate the sets of candidate positive and negative terms from V (f, k);

Create the population *oldPop* and initialize each chromosome;

Repeat n times

Evaluate the fitness of each chromosome in *oldPop*;

newPop = ∅;

Copy in *NewPop* the best r chromosomes of *oldPop* (elitism r is

determined on the basis of the elitism percentage)

While *size*(*newPop*) < *size*(*oldPop*)

select *parent1* and *parent2* in *oldPop* via roulette wheel

generate *kid1*, *kid2* through *crossover*(*parent*1, *parent*2)

apply mutation, i.e., *kid*1 = *mut*(*kid*1) and *kid*2 = *mut*(*kid*2)

apply the repair operator ρ to both *kid*1 and *kid*2;

add *kid1* and *kid2* to *newPop*;

end while

oldPop = *newPop*;

end repeat;

Select the best chromosome *K* in *oldPop*;

Eliminate redundancies from *K*;

3.4) System Desi

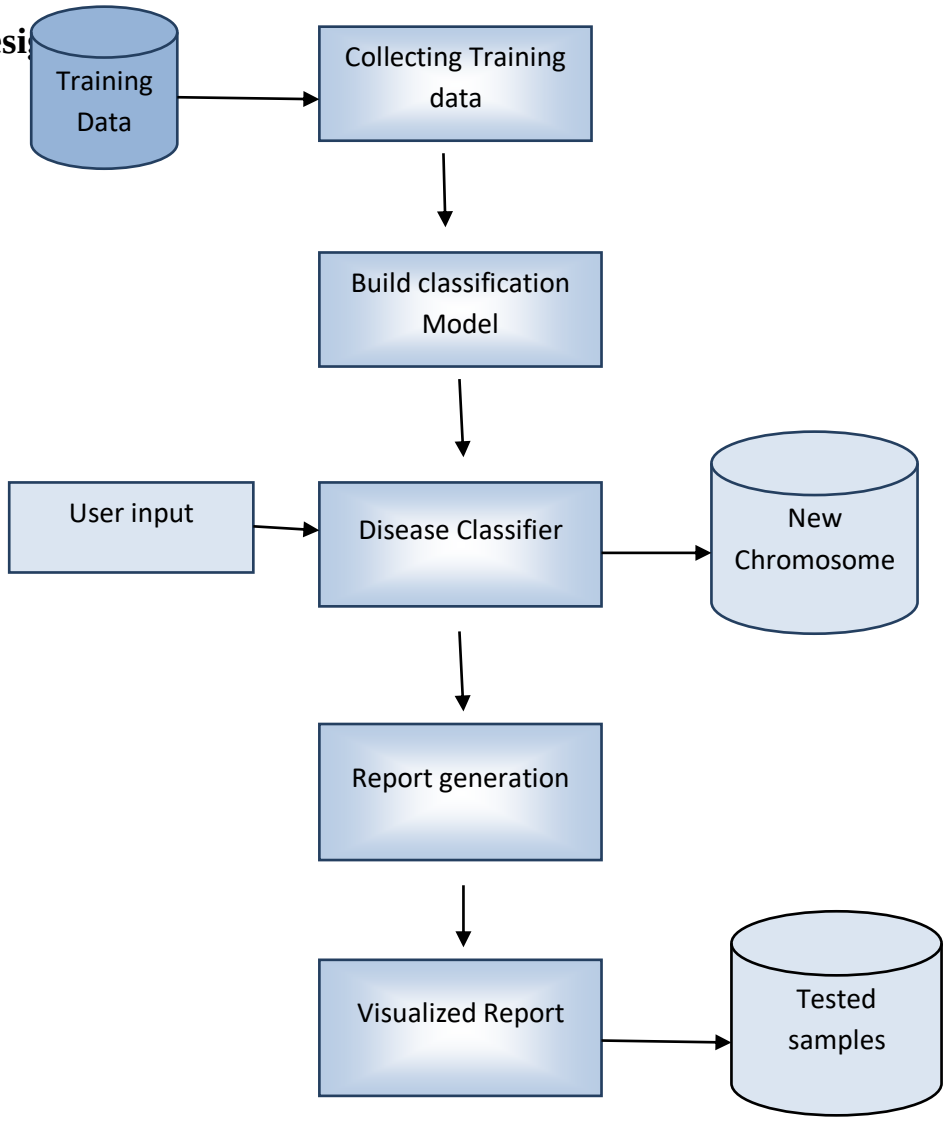


Fig 2 System Design

[1] **Collecting training data**:

Dataset are stored as per stage, T, N, and M. training data collector check the stage of disease of patient and store in appropriate location.

[2] build classification model:

steps of building classification model

i. for classification model first number of categories are determined, as breast cancer have three stages, total number of categories is three.

ii. Population and generations have predefined, for every category the populations and generations have same.

iii. Now iteratively for each category each chromosome is analysed for number times of generations. For each iteration there is decided that the chromosome is positive or negative and from its actual result accuracy is decided. Finally the maximum accuracy value is taken and decided whether the chromosome is positive or negative.

iv. For each category sets of positive and negative chromosomes found.

v. Finally redundancies are eliminated from set of chromosomes.

[3] Disease classifier:

In this class the document d is classified under category c if t1 belongs t d or t2 belofngs to d up to tn belongs to d and not tn+1 belongs to d or tm belongs to d holds. Disease classifier consists following steps:

i. The new patient's biomedical data is entered to system which consist name, age, height, weight BP, HR, symptoms, tests and test reports.

ii. Above entered data contains set of chromosomes called document d and disease classifier classify d in to category c.

iii. In d if any new chromosome is detected then that is store in database

[4] Report generator:

Category c means the current stage of breast cancer of patient. Using java template the result is visualized. First patient's name is mentioned in report after

that his disease and stage will mention, after textual format the visual format is shown.

[5] Visualized format:

Report generation is an important part of our project because one think has taken care that the report should be easy to understand to all person they may be medical person or non-medical person, all should understand the report. In report there is the relation between the location where carcinoma cells spreaded and current stage of patient is shown so patient will understand her health status.

We provided final report which have following factors

1. Name
2. Disease
3. Factors of input data patient due to which patient is classified into particular stage
4. Stage detected by Olex GA
5. Stage detected by C4.5
6. Template visualized

The template shown in report contains two bars.

- First bar indicates that in which part carcinoma cells got spreaded, this bar have three parts skin, lymph node and other organ. The carcinoma spreaded part is indicated by red color. And the where carcinoma cells are absent is indicated by green color

- Second bar indicates that in which stage of breast cancer the patient have currently. This bar have three parts T, N and M. If patient is in T stage then only T part is red and remaining parts are green. If patient is in N-stage then T and N parts are red and M part in green. And finally if patient is in last stage i.e. in M stage all the three parts are red.

By using two bars we are trying to indicate relation between parts where carcinoma cell are present and current stage of breast cancer the patient have, Due to this patient will understood that why she have that particular stage of breast cancer.

Carcinoma Cells Present/Cancerous tumor present	**Stage**
Skin of breast, Breast	T
Lymph nodes	N
Other organs(brain, lung, chest, liver, kidney, uterus, etc.)	M

Table 3.1: Presence of carcinoma cells and stages.

3.5) form design:

I. Input form:

This is the webpage contains parameters for patient's general information like name, age, height, weight, BP, heart rate. Further for diagnostic purpose symptom and test reports are asked to input on this form

II. Output form:

This is the webpage containing patients report. It contains both the formats textual and graphical. For the graphical report java template is used. It containing two bars first indicates where the carcinoma cells present and second bar indicates the current stage of patient. This shows the result of C4.5 and Olex GA algorithm.

3.6) **Conceptual Overview of proposed system-**

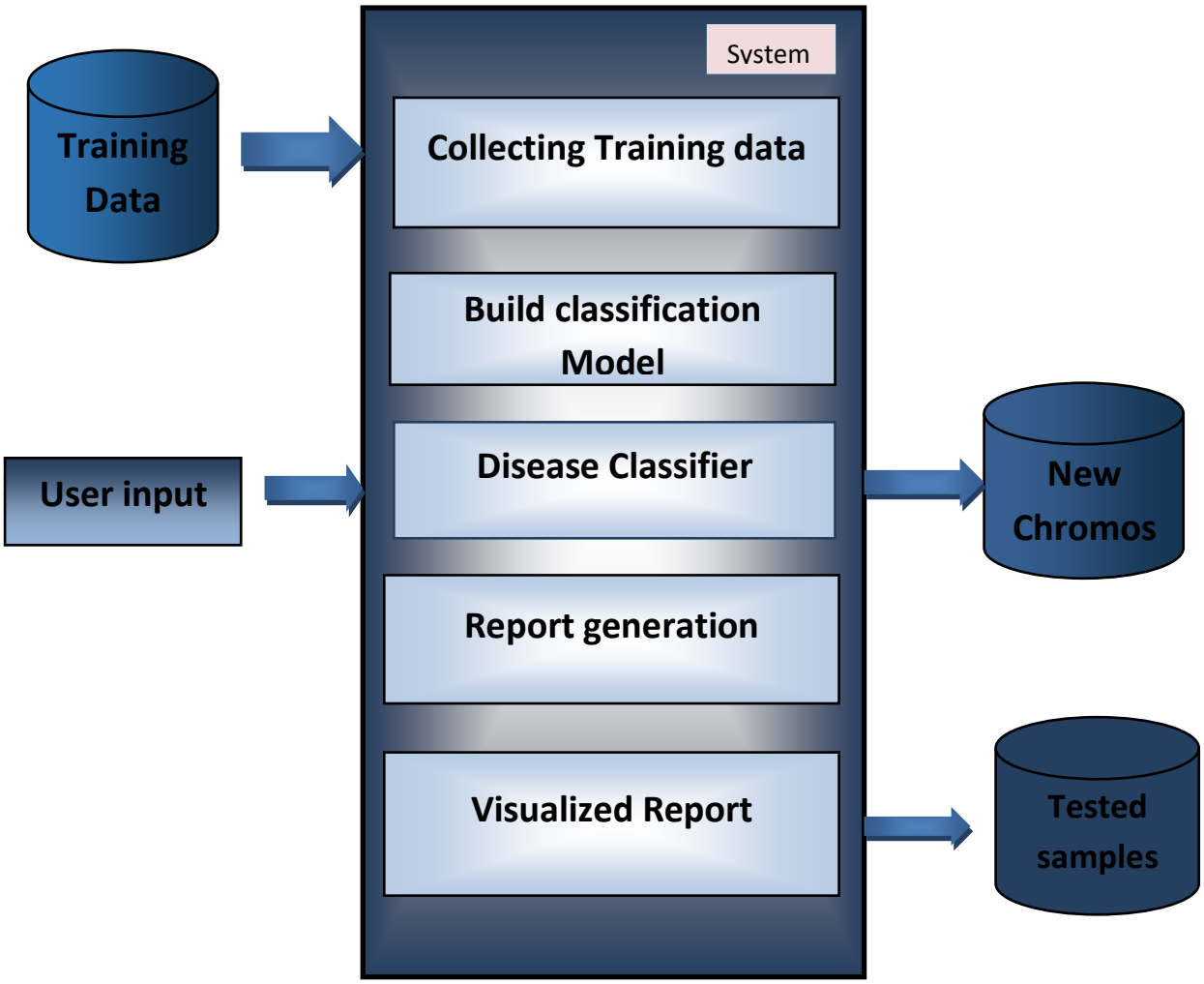


Fig 3.1: Conceptual Overview of Proposed System

In the first step training data is collected from source.

In the second step the set of positive and negative chromosomes are classified using Olex GA that by the end of this step we will get classification model i.e. set of positive and negative chromosomes for each stage T, N and M.

In third step we take input from patient and that patient is classified into T or N or M stage. using classification model i.e. checking chromosomes present in new patient's biomedical data. This step is testing phase, while testing if we found any unknown symptom or test then we will save that in database and that will trained manually. This feature of system is called adaptation.

In fourth step, system detected current stage of breast cancer already, for this status result is generated which is visualized. The visualized report mentions two factors first is the area where carcinoma cells present and second is the current stage of breast cancer and finally the new patient's biomedical data and his result will store in separate database.

Chapter 4

DATA FLOW DIAGRAMS

Data flow diagram is a graphical representation of the flow of data It shows the system's major processes, data flows and data stores at a high level of abstraction. Often they are preliminary step used to create an overview of the system which can later be elaborated. When the Context Diagram is expanded into DFDlevel-0, all the connections that flow into and out of process 0 needs to be retained

4.1 Level 0 DFD:

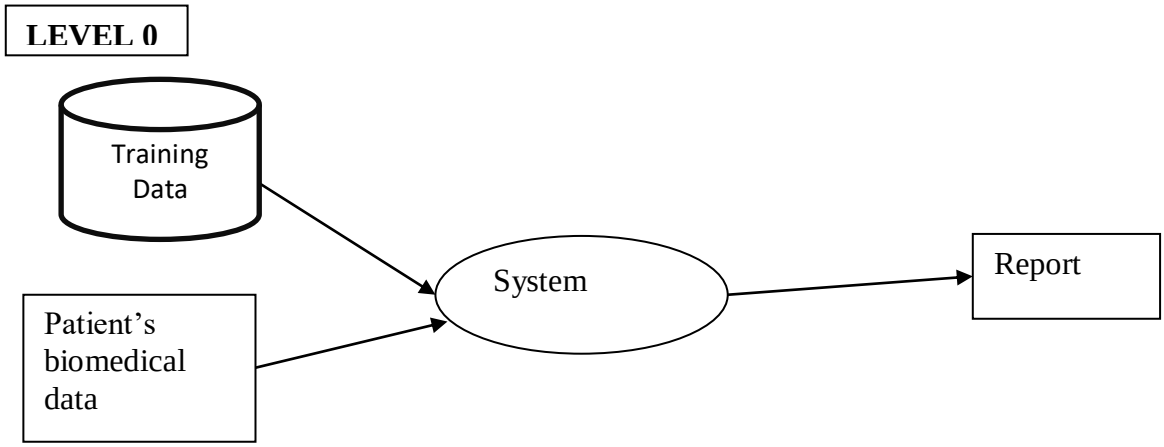


Fig. 4.1 Level 0 DFD

A data flow diagram is that which can be used to indicate the clear progress of a process. In the process of coming up with a data flow diagram, the level one provides an overview of the major

Functional areas of the undertaking

4.2 Level 1 DFD:

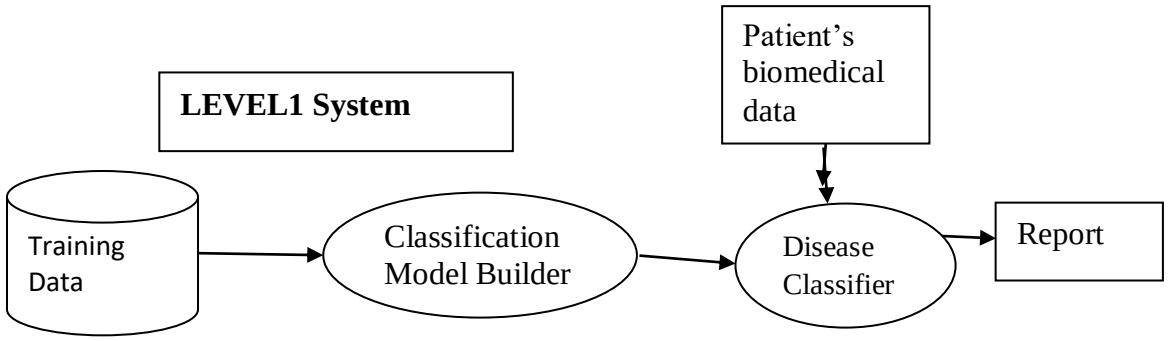


Fig 4.2: Level 1 DFD

4.3 Level 2 DFD:

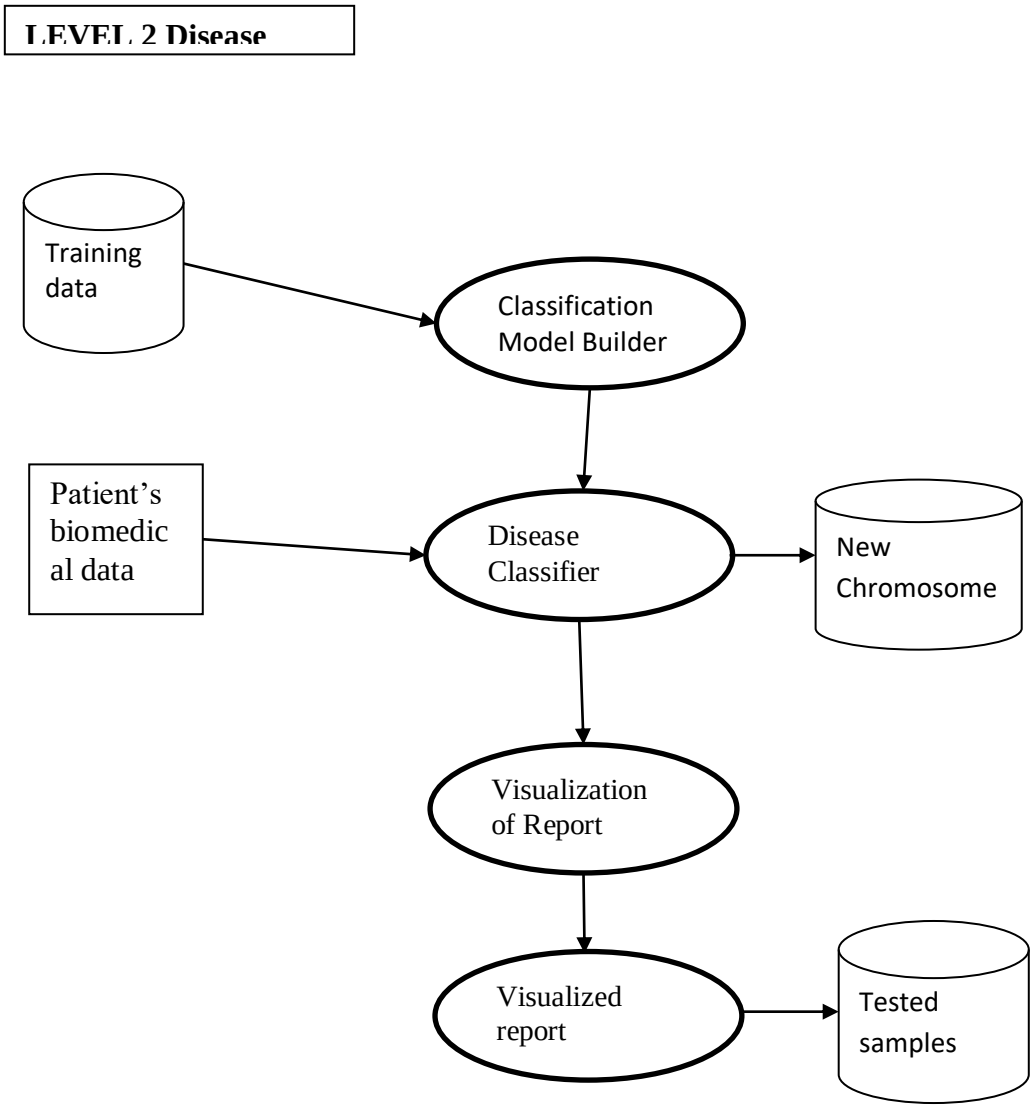


Fig 4.3: Level 2 DFD

Chapter 5

UML DIAGRAMS

A diagram is the graphical presentation of a set of elements, most often rendered as a connected graph of vertices (things) and arcs (relationship).

Diagrams are a projection into a system. For all but the most trivial system, diagram represents an elided view of elements that make up system. The same element may appear in all diagrams. The UML includes the following diagrams:

5.1 Use case Diagram

A use case diagram at its simplest is a representation of a user's interaction with the system and depicting the specifications of a use case. A use case diagram can portray the different types of users of a system and the various ways that they interact with the system. This type of diagram is typically used in conjunction with the textual use case and will often be accompanied by other types of diagrams as well.

Actors:

An actor is a direct external user of a system-an object or a set of objects that communicates directly with the system.

Use case:

A use case is a coherent piece of functionality that a system can provide by interacting with actors.

A use case diagram involves a set of use cases and a set of actors. Each use case represent a slice of the functionality the system provides. An actor represents one kind of objects for which the system can behavior. The UML notation for use case diagram is as shown in fig. a rectangle contains the use cases for the system with the actors listed outside the rectangle.

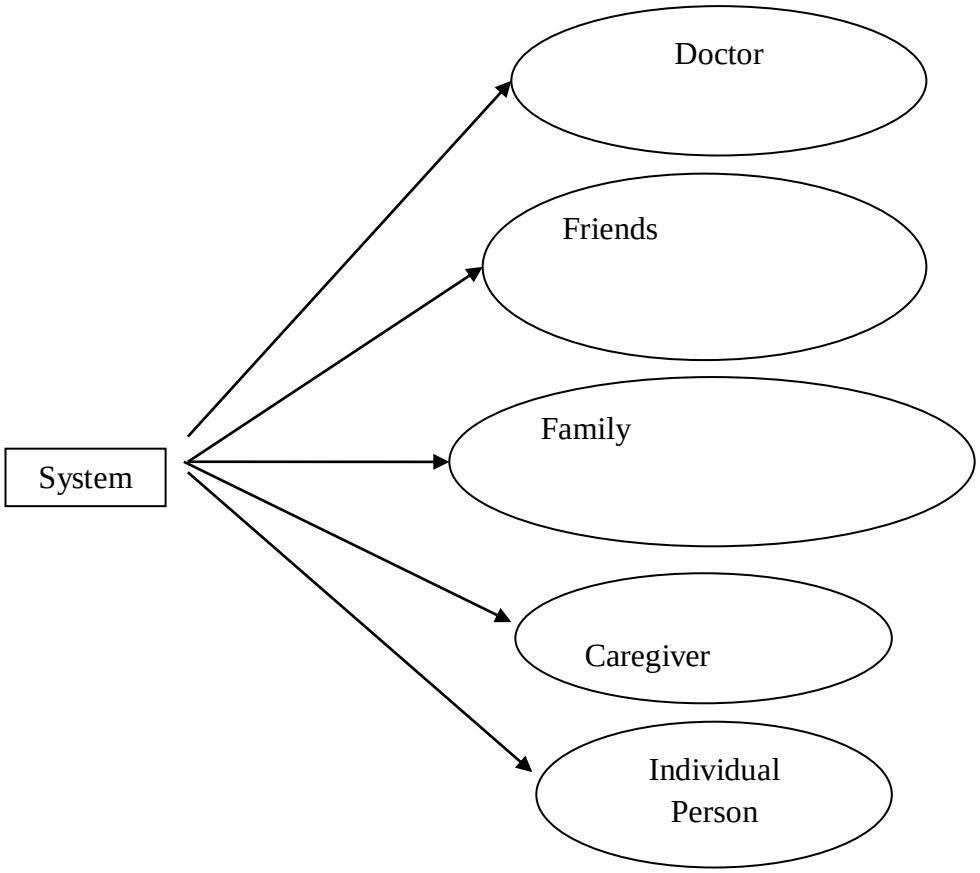


Fig.5.1 Use case Diagram

5.2. Sequence Diagram

A sequence diagram is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. A sequence diagram shows object interactions arranged in time sequence. It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario. Sequence diagrams are typically associated with use case realizations in the Logical View of the system under development. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagram

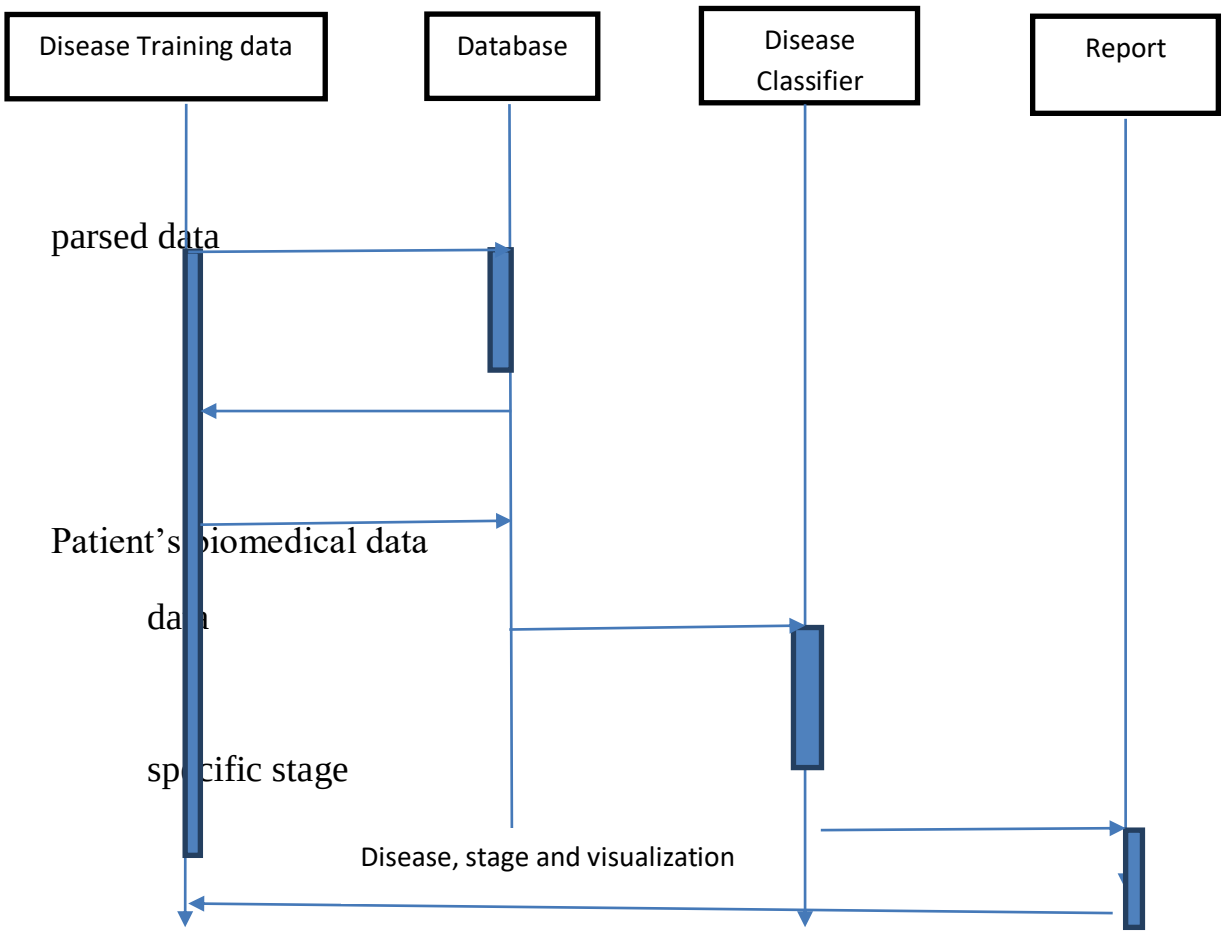


Fig.5.2 Sequence diagram

5.3. Collaborative Diagram

A collaboration diagram describes interactions among objects in terms of sequenced messages. Collaboration diagrams represent a combination of information taken from class, sequence, and use case diagrams describing both the static structure and dynamic behavior of a system.

The Collaboration diagram is same as sequence diagram which contains objects interacting by massages. The objects are placed randomly. Therefore transfer of massages is represented by using Numbering to messages.

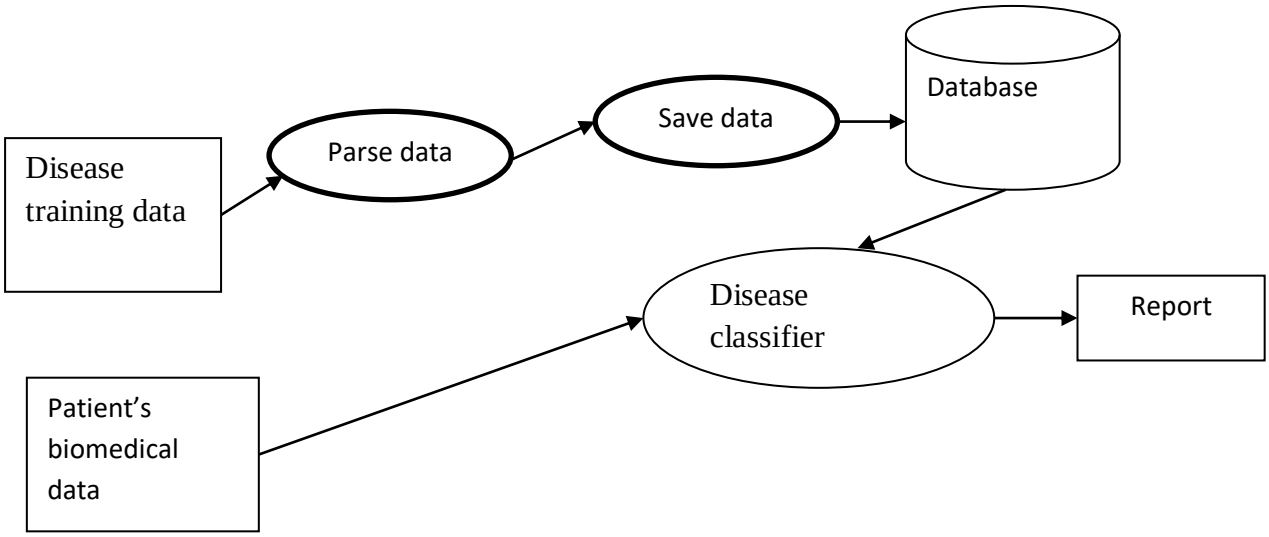


Fig.5.3 CollaborationDiagram

5.4. Class Diagram

The class diagram is the main building block of object oriented modeling. It is used both for general conceptual modeling of the systematic of the application, and for detailed modeling translating the models into programming code. Class diagrams can also be used for data modeling. The classes in a class diagram represent both the main objects, interactions in the application and the classes to be programmed.

A Class is with three sections. In the diagram, classes are represented with boxes which contain three parts:

- The upper part holds the name of the class
- The middle part contains the attributes of the class

The bottom part gives the methods or operations the class can take or undertake.

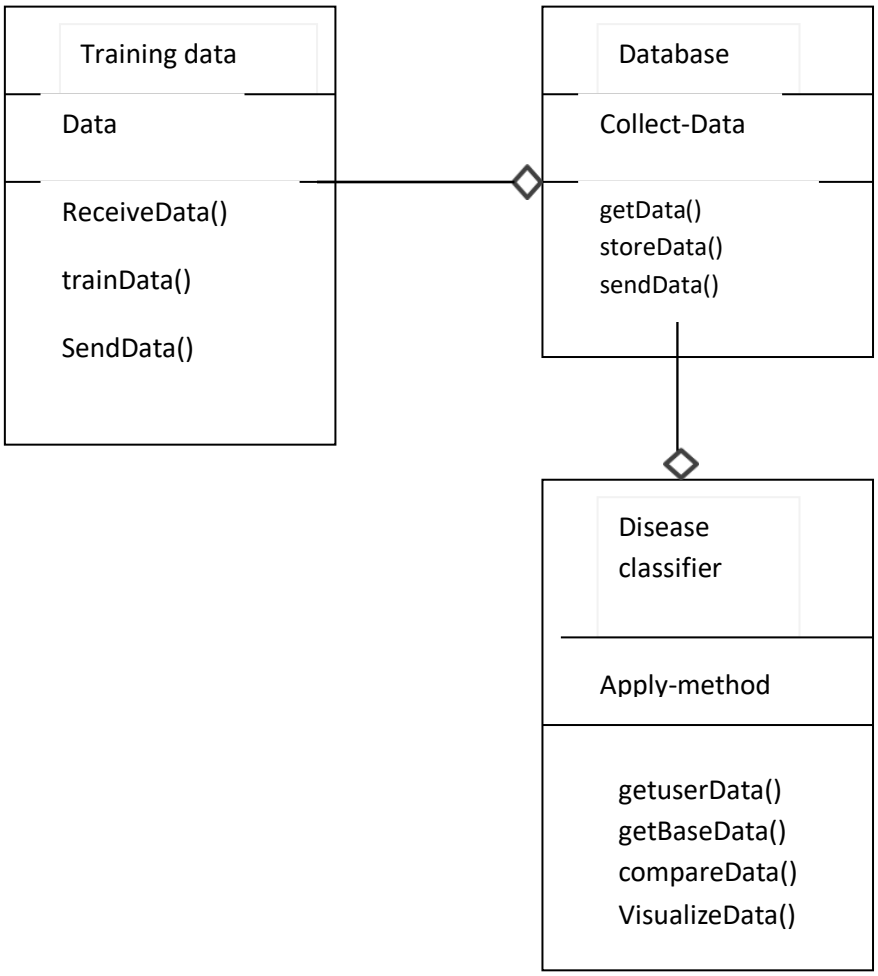


Fig.5.4 Class diagram

In above class diagram,four different classes are represented as, Twitter, Database, Summarizar . The Classes have different attributes and methods. The Database Class have method gettweet() Using this class collects data and store data in database..

5.5 Activity Diagram

Activity diagram is another important diagram in UML to describe dynamic aspects of the system. It is basically a flow chart to represent the flow form one activity to another activity. The activity can be described as an operation of the

system. So the control flow is drawn from one operation to another. This flow can be sequential, branched or concurrent.

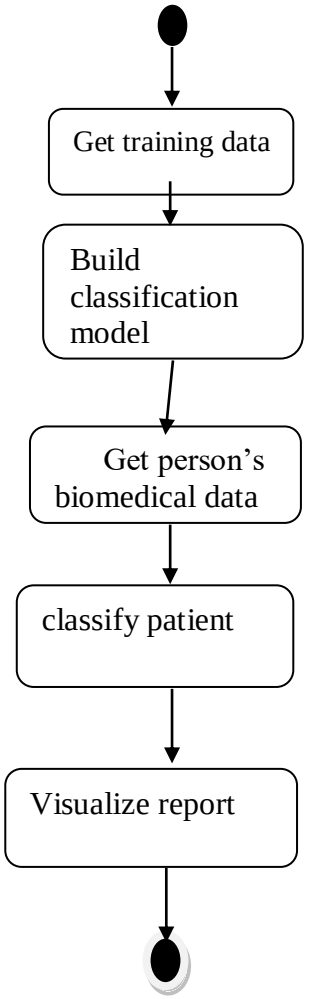


Fig.5.5 Activity diagram

5.6.Component Diagram

In the UnifiedModelingLanguage, a component diagram depicts how components are wired together to form larger components and or softwaresystems. They are used to illustrate the structure of arbitrarily complex systems. Components are wired together by using an *assembly connector* to connect the require interface of one component with the provided interface of another component.

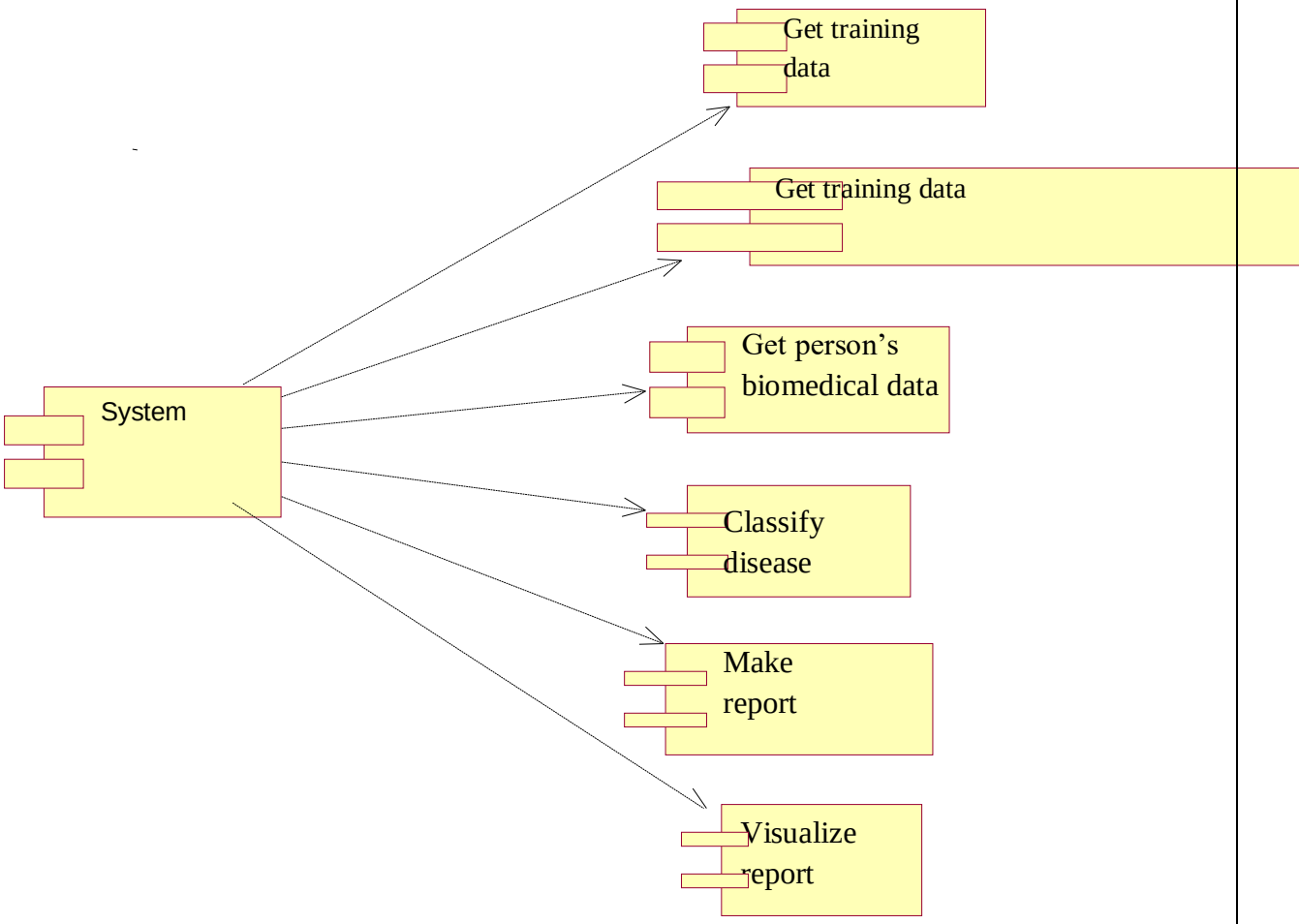


Fig. 5.6 Component Diagram

5.7. Deployment Diagram

A deployment diagram in theUnified Modeling Language models the *physical* deployment of artifacts on nodes. To describe a web site, for example, a deployment diagram would show what hardware components ("nodes") exist (e.g., a web server, an application server, and a database server), what software components ("artifacts") run on each node (e.g., web application, database), and how the different pieces are connected (e.g. JDBC, REST, RMI).

The nodes appear as boxes, and the artifacts allocated to each node appear as rectangles within the boxes. Nodes may have sub nodes, which appear as nested boxes. A single node in a deployment diagram may conceptually represent multiple physical nodes, such as a cluster of database servers.

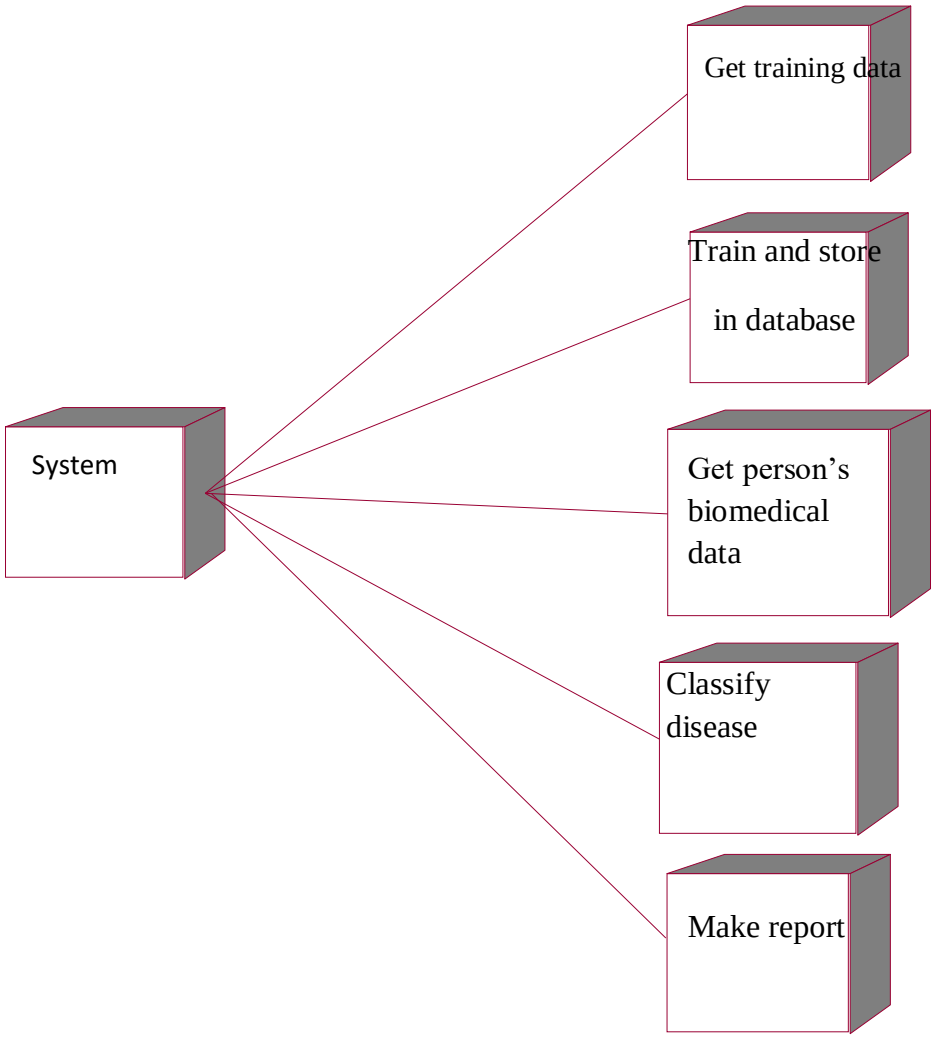


Fig. 5.7 Deployment Diagram

Chapter 6

Implementation

6.1) Classes:

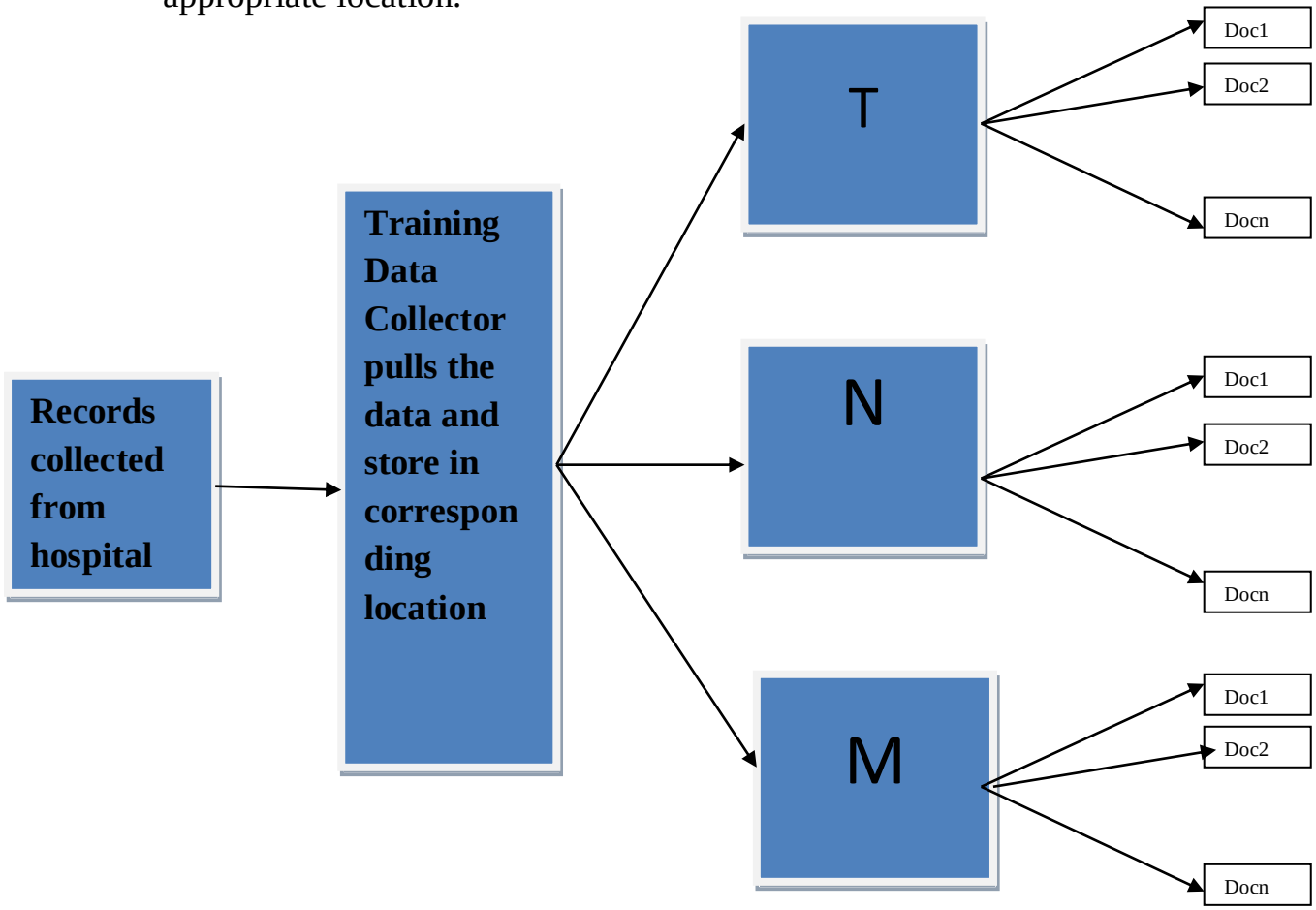
Following are the classes those using for implementation of project.

1) Training data collector
2) Classification model builder
3) Disease classifier
4) Report generator

6.1.1 Training Data Collector:

Data is necessary to solve the problem and this data is collected from hospital. For this project the dataset required is records of patients who have done breast cancer diagnosis in hospital. This data is collected from 'Shri Sidhdeshwar cancer hospital and research center, Solpaur'. 100 records of patients have collected for training purpose. These patients may have breast cancer and they have done diagnosis in hospital. 50 records of patients have collected for testing purpose. These patients may have Breast cancer and they have done diagnosis in hospital.

Datasets are stored in different location as per their stage such as if T, which is the first stage of Breast cancer , if N which is the seconds stage of Breast Cancer and M which is last stage of Breast Cancer. This class Training Data Collector checks the stage of disease of patient and store in appropriate location.



6.1.2) Build Classification Model:

This class trains data sets as per their categories. Olex GA is the supervised classification algorithm in which while training phase a model is built and that will further used to classify new document.

Steps to build classification model

i. For classification model first number of categories is determined, as Breast Cancer have three stages , total number of categories is three. Here categories are determined as three because datasets are located in three groups.

ii. Populations and generations are defined. For every category the population and generation are same

iii. Now iteratively for each category each chromosome is analyzed for number of time of generations.For each iteration there is decided that the chromosome is positive or negative and by comparing assumed value with actual value accuracy is decided at the end of each generation. If number of documents are less then there have to generate more generations and if number of documents are more then by assigning less number of generations there can achieve more accuracy.Finally the maximum accuracy value is taken and decided whether the chromosome is negative or positive.

iv. At this step for each category H_c (Pos, Neg) over training data built. Where,

$Pos(t_1,t2,\ldots,t_n)$ and

$Neg(t_{n+1},t_{n+2},\ldots,t_m)$

Positive terms in Pos used to cover the training set of c category, while negative terms in Negare used to take precision under control.

v. Further crossover made between vales assigned for chromosomes

vi. Finally redundancies are eliminated from set of chromosomes and H_c (Pos, Neg) is returned

6.1.3 Disease Classifier:

This is the testing phase. In this class the document d is classified under category c if t_1 belongs to d or t_2 belongs to d or ….or t_n belongs to d and not (t_{n+1} belongs to d or …. Or t_m belongs to d) holds.the disease classifer consists following steps

i. New patient enters their biomedical data to system, which contains following factors
   - Name
   - Age
   - Height
   - Weight
   - BP
   - HR
   - Symptoms
   - Tests
   - Test reports

ii. Above entered data contains set of chromosomes called document d. now aim of this class is to classify d into corresponding category

$$c \leftarrow (t1 \in d \vee \cdots \vee tn \in d) \wedge \neg(tn{+}1 \in d \vee \cdots \vee tn{+}m \in d)$$

where c category

d document

each t_i – a term taken from given probability

c classifier H_c (Pos, Neg)

Pos(t1,t2,…,tn)

Neg(tn+1, tn+2,…,tn+m)

iii. In d if any new chromosome detects then that will store in another database “New”. And that will train after.

6.1.4 Generate Report:

In previous class, Disease Classifier finds disease and current stage of disease of breast cancer in the patient. Using java template the result is visualized

Report generation is an important part of our project because one think has taken care that the report should be easy to understand to all person they may be medical person or non-medicalperson, all should understand the report. In report there is the relation between the location where carcinoma cells spreaded and current stage of patient is shown so patient will understand her health status.

We provided final report which have following factors

VII. Name
VIII. Disease
IX. Factors of input data patient due to which patient is classified into particular stage
X. Stage detected by Olex GA
XI. Stage detected by C4.5
XII. Template visualized

The template shown in report contains two bars.

- First bar indicates that in which part carcinoma cells got spreaded, this bar have three parts skin, lymph node and other organ. The carcinoma spreaded part is indicated by red color. And the where carcinoma cells are absent is indicated by green color

- Second bar indicates that in which stage of breast cancer the patient have currently. This bar have three parts T, N and M. If patient is in T stage then only T part is red and remaining parts are green. If patient is in N-stage then T and N parts are red and M part in green. And finally if patient is in last stage i.e. in M stage all the three parts are red.

By using two bars we are trying to indicate relation between parts where carcinoma cell are present and current stage of breast cancer the patient have, Due to this patient will understood that why she have that particular stage of breast cancer.

Carcinoma Cells Present/Cancerous tumor present	**Stage**
Skin of breast, Breast	T
Lymph nodes	N
Other organs(brain, lung, chest, liver, kidney, uterus, etc.)	M

Table 6.1 : Stages of Breast cancer.

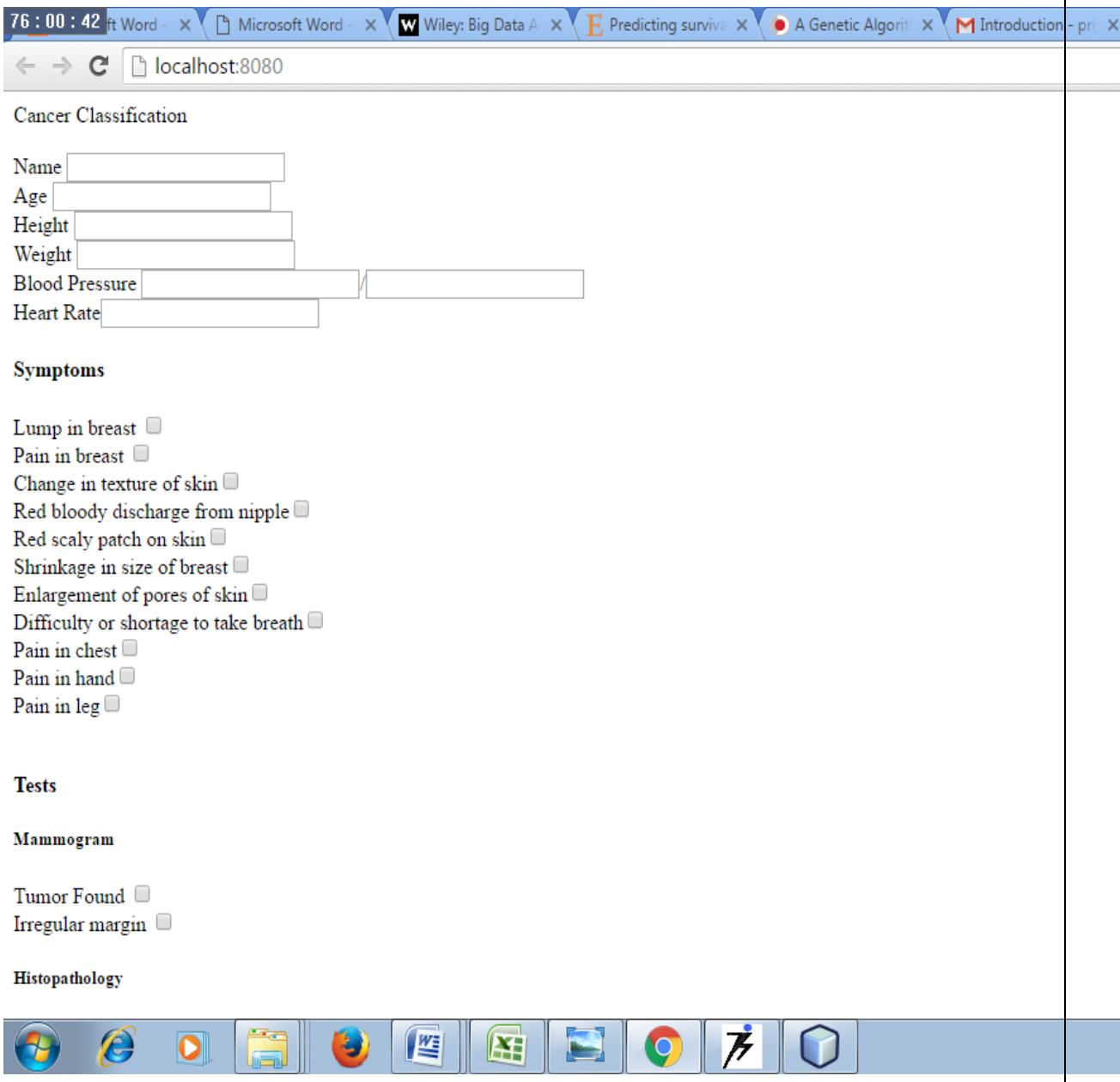
For implementation of this project following servers and software used

1	Apache tomcat server	Here Apache tomcat server is used for training data because of its relioable performance. Its performance is stable in any situation.
2	Weka 3.6	Weka is a software package for data mining. In this project weka is used to implement the classification algorithm C4.5. results of C4.5 are used to compare accuracy and time of Olex GA classifier.
3	MySQL	To store new symptoms and new tests, MySQL database is used in this project. After testing each testcase and its result also store in database.
4	Netbeans IDE 8.1	For implementation of project Netbeans IDE platform is used because final report is on webpage and that is supported by Netbeans 8.1.

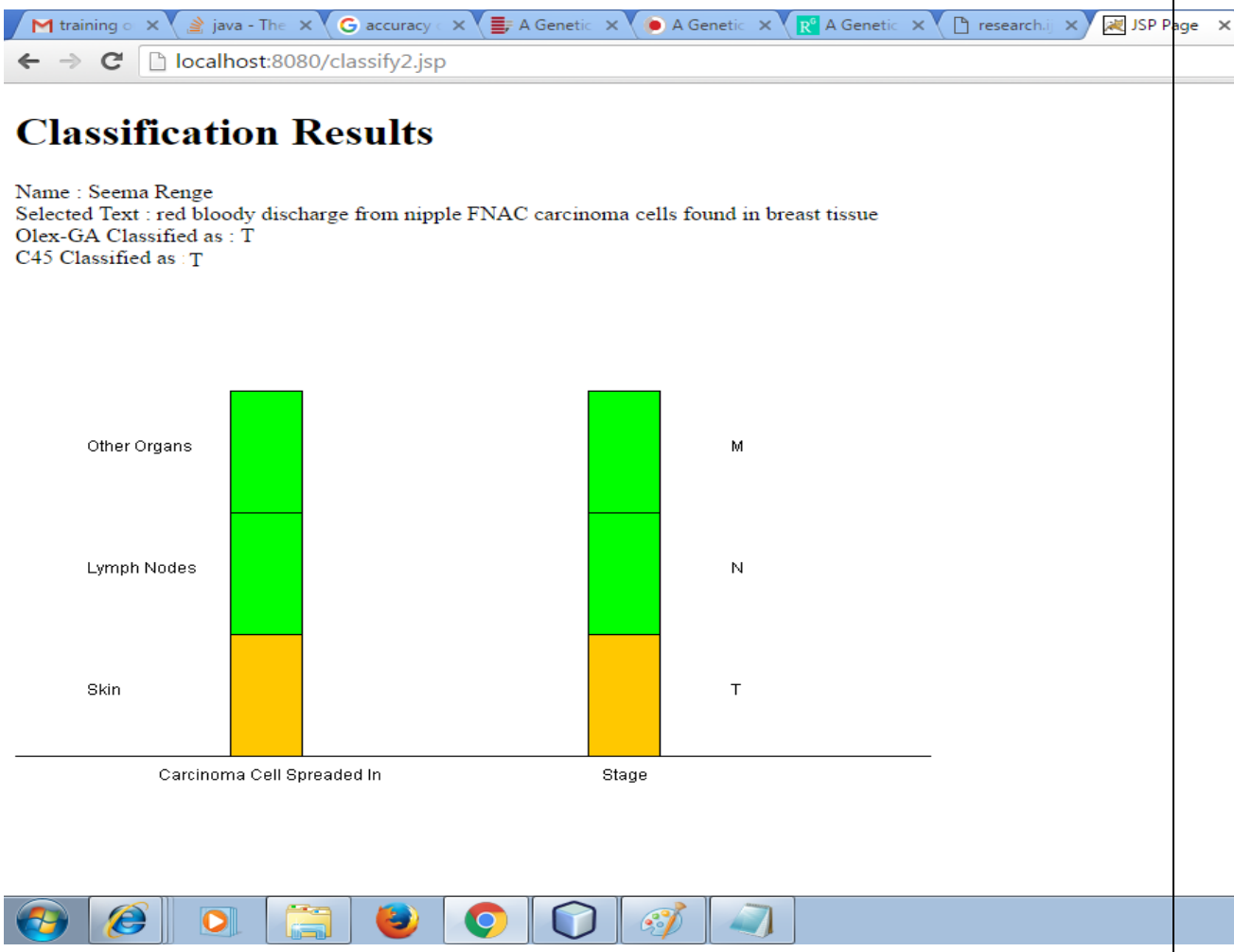
Table 6.2: Tools and softwares and their uses.

Screenshots:

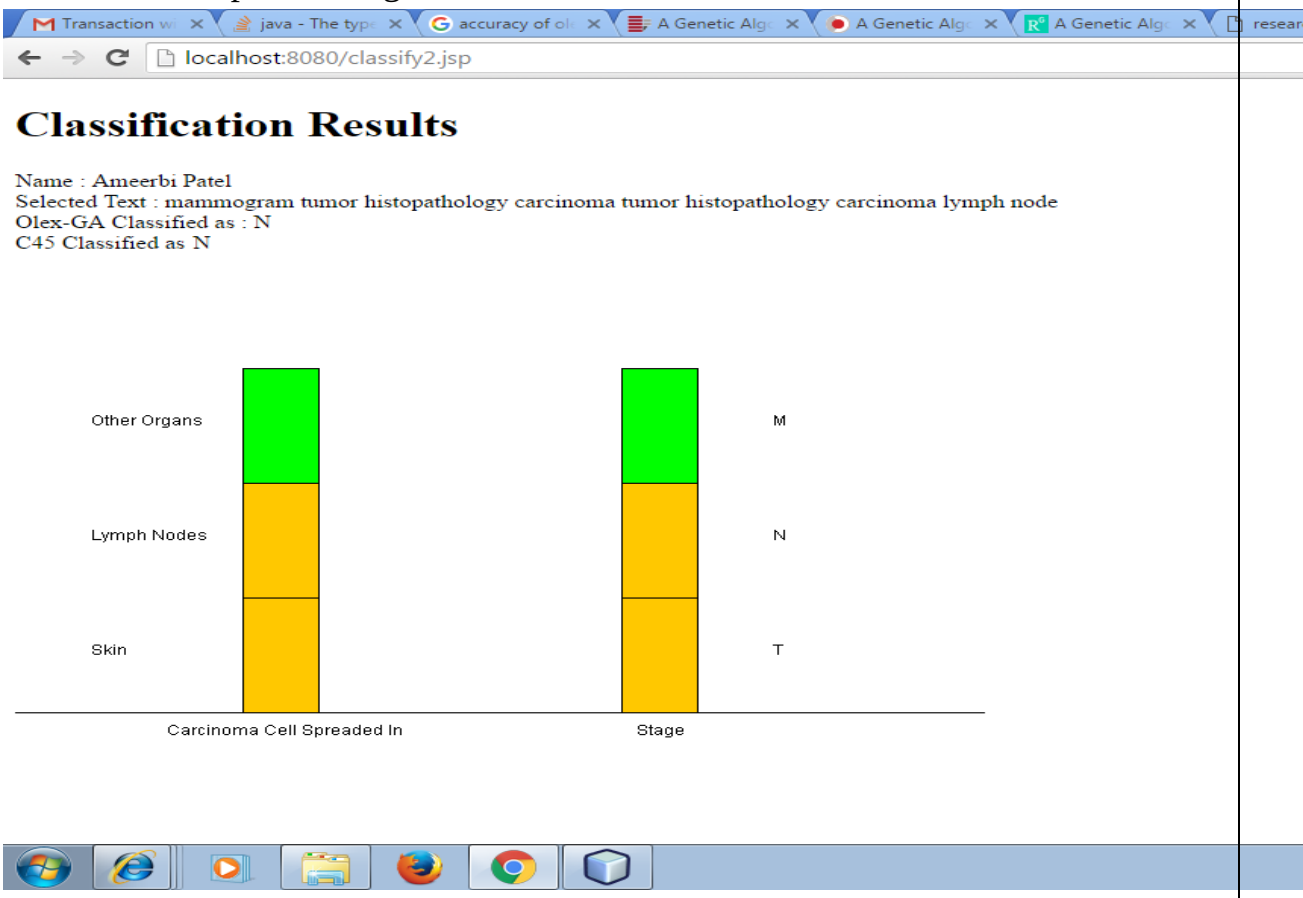
7.1 New patients input



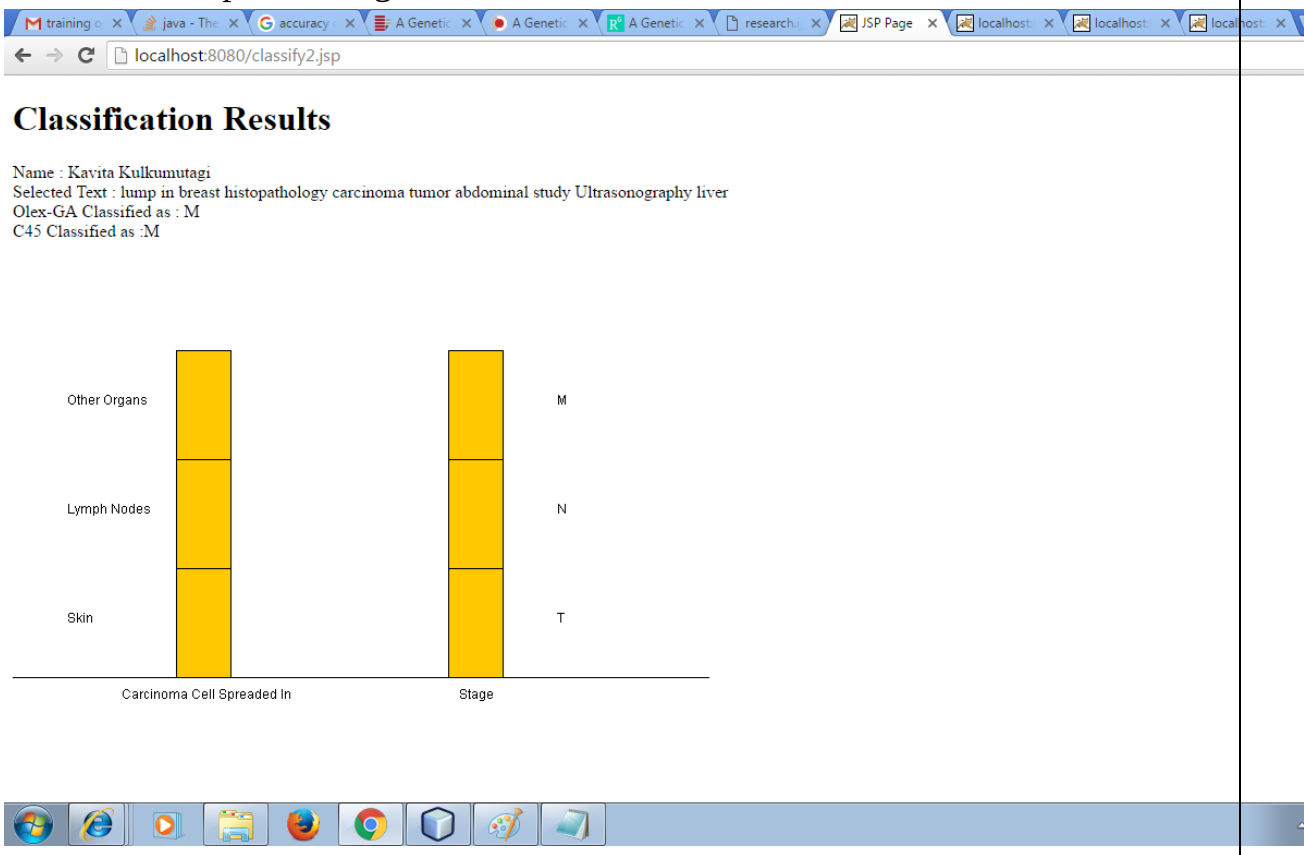
7.2 Visualized output for stage T



7.3 Visualized output for stage N



7.4 Visualized output for stage M



Chapter 7

Result and Analysis

The performances of our system is compared with existing techniques C 4.5 decision tree algorithm. In this project the Olex GA algorithm used to diagnosis of breast cancer disease. These achieved performances can compare with the existing technique C 4.5 decision tree algorithm. Mostly in healthcare services accuracy of result (correct detection) is the most important factor to measure the performance. The metrics used for performance measurementare percentage of accuracy and percentage of error.

Here tried to find out most suitable technique to diagnose the breast cancer. In order to have a fair measure of the performance of the classification algorithms, there is used 5 subgroups of dataset, each subgroup contains 10 records.

7.1 Dataset

Data is necessary to solve the problem and this data collected from a hospital. For this project, the dataset required is records of patients who have done breast cancer diagnosis in a hospital. This data is collected from ‘Shri Sidhdeshwar cancer hospital and research centre, Solapur'. 100 records of patients have

collected for training purpose. These patients may have breast cancer and they have done diagnosis in a hospital. 50 records of patients have collected for the testing purpose. The training data required for the algorithm is collected. This data contain records of patients of breast cancer which have attributes: Name, Age, Height, Weight, Blood Pressure, Heart Rate, Symptoms and Tests, Corresponding test reports, Disease, Stage.

Above described detailed records are collected from Shri. Siddheshwar Cancer Hospital and Research Centre, Solapur. We have collected 100 records of breast cancer from this hospital. These 100 records include patients with different stages, T, N and M stages.

7.2 Evaluation Methodology

[10]The algorithm starts with retrieving datasets from training data All the datasets are retrieved which are already stored as per the stage.

1. T-stage datasets

2. N-stage datasets

3. M-stage datasets

While training these datasets, first all the present texts are retrieved. The population and generation are predefined. In this method, if a number of records in datasets are small then for more accuracy the generation should be high and if the number of records is high then by giving less number of generation also occurs accurate result. As per population is given those number of texts are selected and tries to predict feature of every text whether positive or negative. [11]This operation of prediction is repeated number of times predefined as the generation. Further, there is cross checks are done by using redundant texts. After cross checking, redundant texts are eliminated and set of chromosomes get build.

· ChromosomeT- set of positive and negative chromosomes in stage T.

· ChromosomeN- set of positive and negative chromosomes in stage N.

· ChromosomeM- set of positive and negative chromosomes in stage M.

Now the classification model got built. The task of finding Pos and Neg which maximize the F-measure when Hc(Pos, Neg) is applied to the training set. MAX F can be represented as a 0 1 combinatorial problem. Testing Data. In this stage, the patient's biomedical data is collected. Now Olex GA algorithm checks the texts present in patient's biomedical data while testing this algorithm works as follows If all the texts present in patient's biomedical data then the patient is healthy and she has not breast cancer. But if at least on negative chromosome exist in patient's biomedical data then she has breast cancer, further by using disease classifier stage of breast cancer is detected. The logic used by disease classified as follows, Classify document d under category c if t1 belongs to d or t2 belongs to d or …or in belongs to d and not (tn+1 belongs to d or … to belongs to d) holds Where each ti is a term Olex GA adopts an efficient approach that "several rules per individual" binary representation and uses F-measure as a fitness function. Text classification is a task of assigning natural language texts to one or more thematic categories on the basis of their contents. A genetic algorithm is a random probability distribution or pattern analysis search method inspired to the biological evaluation. The basic idea is that each individual encodes a candidate solution (i.e., a classification rule or a classifier) and that its fitness is evaluated in terms of predictive accuracy. The problem of inducing propositional text classifiers of the form

$$c = (t1\ \varepsilon d \ \vee \ldots \vee tn\varepsilon d) \wedge \neg(t(n+1)\varepsilon d \ \vee .. \vee tn+m\varepsilon d)$$

where c is category, d is document and each ti is a term taken from given probability, c is classifier Hc (Pos, Neg), Pos(t1,t2,…,tn) is a set of positive terms which used to cover the training set of c and Neg(tn+1, tn+2,…,tn+m) is a set of negative terms which are used to take precision under control. The achieved detection performances are comparable to existing techniques. In this project, the Olex GA algorithm used to diagnosis of breast cancer disease. These achieved performances can compare with the existing technique C 4.5 decision tree algorithm.

Mostly in healthcare services, accuracy of a result (correct detection) is the most important factor, so for comparison, the metric used is accuracy here tried to find out the most suitable technique to diagnose the breast cancer. In order to have a fair measure of the performance of the classification algorithms, there is used 5 subgroups of the dataset, each subgroup contains 10 records.

7.3 Comparison

Olex GA:

[5]The implemented technique Olex GA is the text-based classification algorithm. This technique first builds the set of positive and negative texts for given number of categories. As breast cancer occurs in three stages, a number of category for classification is three. Depending on text present in patient's biomedical data Olex GA detects the current stage of breast cancer in the patient.

C4.5

[7]The existing technique C 4.5 is the decision tree based algorithm. C 4.5 first builds the decision tree, each leaf node of the decision tree are the possibilities of the result, further by trimming branches of tree one by one, the remaining single leaf is considered as the final result. C4.5 already exists and it is implemented by many authors for classification purpose. The techniques of calculating accuracies and error percent's are defined as follows,

Percentage of accuracy	(Number of correct results/total results)*100
Percentage of error	(Number of wrong results/total results)*100
Average accuracy	(sum of percentage of accuracy of all data groups/number of data groups)
Average error	(sum of percentage of error of all data groups/number of data groups)

Table 7.1: Metrices of comparison

For comparison of the accuracy of results of Olex GA and C 4.5 for the same test cases and the datasets used to train both algorithms are also same. Here for same training data and testing data the accuracies, of results are compared. Testing datasets

are divided into five subgroups. Each subgroup consists of 10 data records. These records are tested and the result is compared with the actual result of those records. For some cases of testing data sets, Olex GA shows correct result and C 4.5 shows the wrong result. For some cases of testing data sets, C 4.5 shows correct result and the Olex GA shows the wrong result. Sometimes both the algorithms show wrong results. To calculate the overall accuracy of algorithms average accuracy and average error percentages are calculated after separate calculation for each subgroup. The table shows results occurred of subgroup1 of Olex GA and C 4.5 for same test cases.

Actual Result	OlexGA	C 4.5
T	T	T
N	N	N
T	T	M
N	N	N
N	N	N
T	T	M
T	T	T
N	N	N
T	T	T
T	N	T
Percentage of accuracy(av1)	90	80
Percentage of error(e1)	10	20

Table 7.2: Results of test cases for dataset1

From above table for 10 test cases Olex GA shows 9 correct results and 1 wrong result, therefore its percentage accuracy is 90, and percentage of error is 10. For same test cases C4.5 shows 8 correct results and 2 wrong results, therefore its percentage accuracy is 80, and percentage of error is 20. The table shows results occurred of subgroup2 of Olex GA and C 4.5 for same test cases.

ActualResult	OlexGA	C 4.5
T	T	T
N	N	N

T	T	M
N	N	N
N	N	N
T	T	M
T	T	T
N	N	N
T	T	T
T	N	T
Percentage of accuracy (av1)	90	80
Percentage of error(e1)	10	20

Table 7.3: Results of test cases for dataset2

From above table for 10 test cases Olex GA shows 9 correct results and 1 wrong result, therefore its percentage accuracy is 90, and percentage of error is 10. For same test cases C4.5 shows 8 correct results and 2 wrong results, therefore its percentage accuracy is 80, and percentage of error is 20. The table shows results occurred of subgroup3 of Olex GA and C 4.5 for same test cases.

ActualResult	OlexGA	C 4.5
N	N	N
N	N	N
N	N	N
N	N	N
N	N	N
N	N	M
N	T	M
N	N	N
N	N	N
T	T	T
Percentage of accuracy	90	80
Percentage of error(e2)	10	20

Table 7.4: Results of test cases for dataset3

From above table for 10 test cases Olex GA shows 9 correct results and 1 wrong result, therefore its percentage accuracy is 90, and percentage of error is 10. For same test cases C4.5 shows 8 correct results and 2 wrong results, therefore its percentage accuracy is 80, and percentage of error is 20. The table shows results occurred of subgroup3 of Olex GA and C 4.5 for same test cases.

Actual Result	OlexGA	C 4.5
M	M	M
M	M	M
M	M	M
M	M	M
M	M	M
M	M	T
M	M	M
M	M	T
M	M	M
M	T	M
Percentage of accuracy (av3)	90	80
Percentage of error(e3)	10	20

Table 7.5: Results of test cases for dataset4

From above table for 10 test cases Olex GA shows all correct results, therefore its percentage accuracy is 100, and percentage of

Error is 0. For same test cases C4.5 shows 9 correct results and 1 wrong result, therefore its percentage accuracy is 90, and

Percentage of error is 10.

The table shows results occurred of subgroup5 of Olex GA and C 4.5 for same test cases.

Actual Result	OlexGA	C4.5
T	T	M
T	T	T

T	T	T
T	T	T
T	T	T
T	T	T
T	T	T
T	T	T
T	T	T
T	T	T
Percentage of accuracy(av4)	100	90
Percentage of error(e4)	0	10

Table 7.6: Results of test cases for dataset5

From above table for 10 test cases Olex GA shows all correct results, therefore its percentage accuracy is 100, and percentage of

Error is 0. For same test cases C4.5 shows 9 correct results and 1 wrong result, therefore its percentage accuracy is 90, and

Percentage of error is 10.

7.4 Analysis

The table shows results occurred of subgroup5 of Olex GA and C 4.5 for same test cases.

TABLE VIII

Results of test cases

Average accuracy algorithm Olex GA= sum of percentage of accuracy of all data groups/number of data groups) =(av1+av2+av3+av4+av5)/5=94%
Average accuracy algorithm Olex GA= sum of percentage of accuracy of all data groups/number of data groups) =(e1+e2+e3+e4+e5)/5=6%
Average accuracy algorithm C4.5= sum of percentage of accuracy of all data groups/number of data groups) =(av1+av2+av3+av4+av5)/5=82%

Average accuracy algorithm C4.5= sum of percentage of accuracy of all data groups/number of data groups) =(e1+e2+e3+e4+e5)/5=18%

The table shows the average accuracies and average error of Olex GA algorithms:

Algorithm	Percentage of accuracy	Percentageoferror
C 4.5	82	18
OlexGA	94	6

Table 7.7: Results of test cases

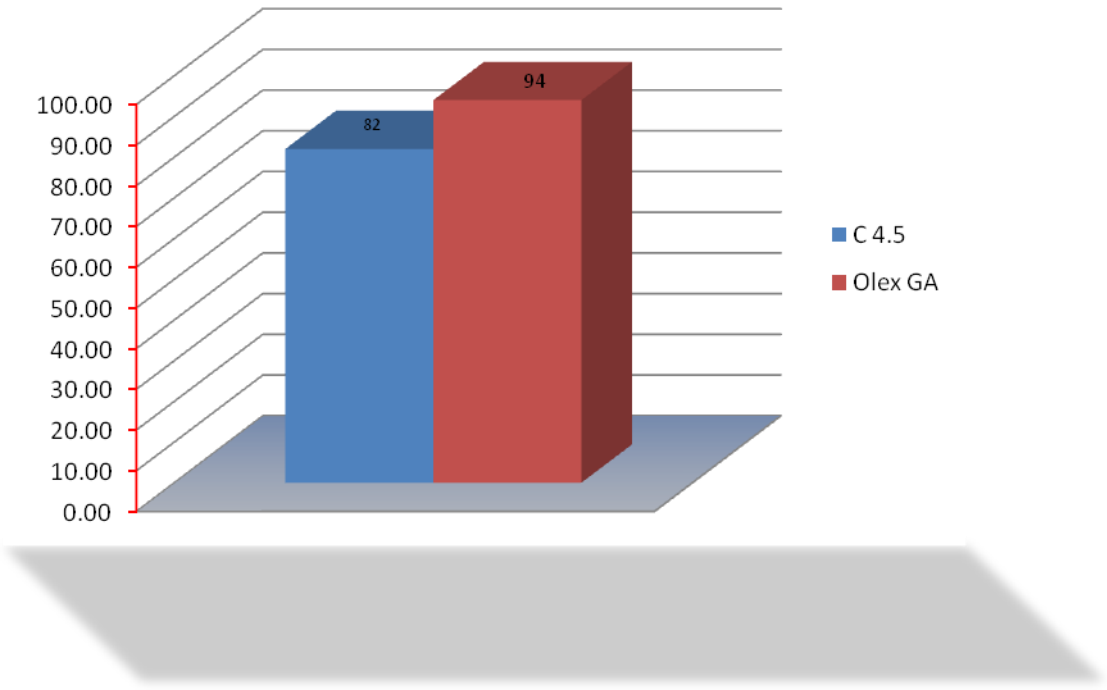


Figure 7.1 Comparison of accuracies of C4.5 and Olex GA

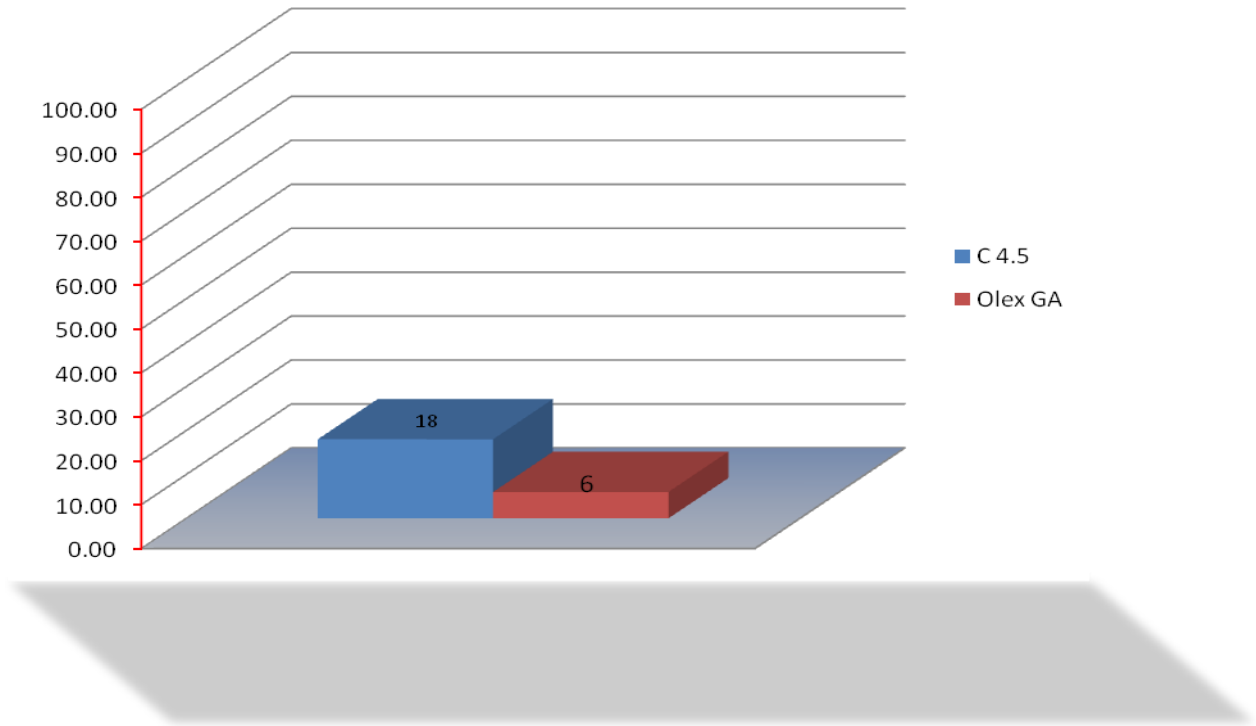


Figure 7.2 Comparison of errors of C4.5 and Olex GA

From above graphs, the dissertation conclude that the Olex GA gives more accurate result in diagnosis of breast cancer. The average percentage of error is more in C 4.5. therefore the more suitable technique for breast cancer diagnosis is Olex GA.

Chapter 8

Conclusion and Future Work

8.1 Conclusion

Olex GA data mining algorithm is used to diagnose the disease breast cancer. The performance of classification technique of this algorithm is based on percentage of accuracy and percentage of error. The performance is compared with the existing technique C4.5. In the area of healthcare different classification algorithms are implemented. These algorithms classify the patient into one the categories. An important challenge in data mining is to build an accurate and efficient classifier for medical application. The performance of Olex GA shows the highly accurate results. 50 test cases are tested on system the average accuracy of Olex GA is 94% and that of C4.5 is 82%. For same test cases average error rate of Olex GA is 6% and that of C 4.5 is 18%. The performance of Olex GA has high accuracy and low error rate. It is suggested as the most appropriate method for the diagnosis of breast cancer. The attributes used for this are symptoms and all test reports. Additional features provided with diagnosis are the visualization of result and adaptation to new diagnostic tests and symptoms.

The accuracy of classification techniques is evaluated based on the existed classifier algorithm. Olex GA data mining algorithm is used to diagnose the disease breast cancer. In the area of healthcare different classification algorithms are implemented. These algorithms classify the patient into one the categories. An important challenge in data mining is to build an accurate and efficient classifier for medical application. The performance of Olex GA shows the highly accurate results. Therefore Olex GA classifier is suggested for diagnosis of Breast Cancer disease based classification to get better results with accuracy, low error rate and performance. The attributes used for this are symptoms and all test reports. Additional features provided with the diagnosis are the visualization of result and adaptation to new diagnostic tests and symptoms.

8.2 Future Scope

This project diagnoses the current stage of breast cancer, the olex GA algorithm is used in this technique. The proposed technique can use to diagnose another disease such as Diabetes, lung diseases, heart diseases etc. The framework can put as it is, by changing its training datasets and checking parameters, this system can use for

diagnosis of other diseases. The checking parameters depend on the disease. Adaptation to new symptoms or tests is manually trained but this can do automatically, which is the future task of this project.

Adaptation to new symptoms or tests is manually trained but this can do automatically, which is the future task of this project.

REFERENCES

[1] Vikas Chourasia and Saurabh Pal, "A Novel Approach for Breast cancer detection Using Data mining techniques" in 'International Journal of Innovative Research in Computer and Communication Engineering' Vol. 2, Issue 1, January 2014.

[2] Miss Janhavi Joshi and Dr. Jigar Patel "Diagnosis and Prognosis Breast Cancer using Classification rules" in International Journal of Engineering Research and General Science Volume 2, Issue 6, October-November, 2014 ISSN 2091-2730

[3] Shiv Shakti Shrivastava, Anjali Sant, Ramesh Prasad Aharwal"An Overview on Data Mining Approach on Breast Cancer data" in International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970) Volume-3 Number-4 Issue-13 December-2013

[4] Shweta Kharya "using data mining techniques for diagnosis and prognosis of cancer disease" in International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012

[5] Zehra Karapinar Senturk and Resul Kara "BREAST CANCER DIAGNOSIS VIA DATA MINING:PERFORMANCE ANALYSIS OF SEVEN DIFFERENTALGORITHMS" in Computer Science & engineering: An International Journal (CSEIJ), Vol. 4, No. 1, February 2014

[6] Abdelghani Bellaachia, Erhan Guven "Predicting Breast Cancer Survivability Using Data Mining Techniques"

[7] Shelly gupta, dharminder kumar and anand Sharma "data mining classification techniques applied forbreast cancer diagnosis and prognosis" in Indian Journal of Computer Science and Engineering (IJCSE)Vol. 2 No. 2 Apr-May 2011ISSN : 0976-5166

[8] Ronak Sumbaly, N. Vishnusri, and S. Jeyalatha "Diagnosis of Breast Cancer using Decision Tree Data Mining Technique" in International Journal of Computer Applications (0975 – 8887) Volume 98– No.10, July 2014.

[9] Adriana Pietramala, Veronica L. Policicchio1, Pasquale Rullo,and Inderbir Sidhu "A Genetic Algorithm for Text Classification Rule Induction" in W. Daelemans et al. (Eds.): ECML PKDD 2008, Part II, LNAI 5212, pp. 188–203, 2008c Springer-Verlag Berlin Heidelberg 2008.

[10] Liqiang Nie, Member, IEEE, Yi-Liang Zhao, Mohammad Akbari, Jialie Shen, Member, IEEE, and Tat-Seng Chua, Member, IEEE "Bridging the Vocabulary Gap between Health Seekers and Healthcare Knowledge" in

IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 2, FEBRUARY 2015.

[11] Xiaoli Li and Bing Liu "Learning to Classify Texts Using Positive and Unlabeled Data" in MIT Allience 2016

[12] AndrewKusiak, Bradley Dixonb, Shital Shaha, (2005),'' Predicting survival time for kidney dialysis patients: a data mining approach'', Elsevier Publication, Computers in Biology and Medicine, Vol. 35, pp 311–327

[13] Sadik Kara, Aysegul Guvenb, Ayse OztUrk Onerc, (2006) "Utilization of artificial neural networks in the diagnosis of optic nerve diseases", Elsevier Publication, Computers in Biology and Medicine,Vol. 36, pp 428–437

[14] Abhishek, Gour Sundar Mitra Thakur, Dolly Gupta, (2012) "Proposing Efficient Neural Network Training Model for Kidney Stone Diagnosis", International Journal of Computer Science and Information Technologies, Vol. 3 (3), pp 3900-3904

[15] Basma Boukenze1, Hajar Mousannif and Abdelkrim Haqiq "Predictive Pnalytics In Healthcare System using Data mining techniques"

[16] Ashfaq Ahmed K, Sultan Aljahdali and Syed Naimatullah Hussain, (2013) "Comparative Prediction Performance with Support Vector Machine and Random Forest Classification Techniques", International Journal of Computer Applications Vol. 69, No.11, pp 12-16

[17] Adriana Pietramala, University of Calabria "A Genetic Algorithm for Text Classification Rule Induction" on http://videolectures.net/ecmlpkdd08_pietramala_agaf/

[18] Ronak Sumbaly, N. Vishnusri, and S. Jeyalatha "Diagnosis of Breast Cancer using Decision Tree Data Mining Technique" in International Journal of Computer Applications (0975 – 8887) Volume 98– No.10, July 2014.

[19] Adriana Pietramala, Veronica L. Policicchio1, Pasquale Rullo,and Inderbir Sidhu "A Genetic Algorithm for Text Classification Rule Induction" in W. Daelemans et al. (Eds.): ECML PKDD 2008, Part II, LNAI 5212, pp. 188–203, 2008c Springer-Verlag Berlin Heidelberg 2008.

[20] student of department of computer Engineering of MITCOE, Pune.]"ADAPTIVE REAL TIME DATA MINING METHODOLOGY FOR WIRELESS BODY AREA NETWORK BASED HEALTHCARE APPLICATIONS"

[21] by kung siau. This paper proposed new business " Health Care Informatics"

[22] by Marek Laskowski, Bryan C. P. Demianyk, Julia Witt, Shamir N. Mukhi, *Member, IEEE*, Marcia R. Friesen, and Robert D. McLeod, *Member, IEEE* This paper proposed useful system for emergency department. **"Agent-Based Modeling of the Spread of Influenza-Like Illness in an Emergency Department: A Simulation Study"**

10. PUBLICATIONS:

International Journal

1) Priyanka B. Shivagunde, Prof. Mrs. A. R. Kulkarni “Visual Healthcare analytics using adaptive data mining”, International Journal of Computer Applications (0975 – 8887) National Seminar on Recent Trends in Data Mining (RTDM 2016)
2)