IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified

Vol. 6, Issue 2, February 2017

Healthcare Analysis using Olex Genetic Algorithm

Ms. Priyanka Shivagunde[1], Dr. Ms. Anita R Kulkarni[2]

PG Student, Dept. of CSE, WIT, Solapur, Maharashtra, India[1]

Professor, Dept. of CSE, WIT, Solapur, Maharashtra, India[2]

Abstract: Health is the most valuable factor affecting our life. People are highly focused on the health care with high preference. Now-a-days there are so many fatal diseases occurring in individuals, Cancer is one of those fatal diseases which is the cause of death of the several peoples in a year and in those people, breast cancer is a major cause of women death. According to the recent survey, about 1 in 8 women (about 12 percent) have breast cancer. Diagnosis of disease is usually done in the last stage and hence cannot be cured by treatment. [6]Early diagnosis of this disease is essential and which needs regular check-up should be done by women above 40 years of age. This paper provides a methodology for automatic diagnosis of disease which is more feasible to be used by the individual person, caretaker, friends and family members which is more feasible. A system analyses patient's biomedical data and find out an 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. The genetic Olex algorithm is a text-based classification algorithm. The visual report of diagnosis is generated which is easy to be understood by an individual from non-medical background. Also, one more feature of this system is the adaptation. New symptoms ant tests are saved in a database and those will train manually. This approach helps the patients, doctors and family members to find out.

Keywords: Text based classification, Data Mining, Supervised learning, Olex GA.

I. INTRODUCTION

Breast Cancer In the human body, a cell is the unit of building structure. The tissue is set of similar types of cells. A lifetime of cells is limited, when a new cell produces, an old cell is destroyed that place is taken by the new cell. This happens regularly in a human body. But if old cells do not get destroyed or new cells are generated irregularly then there is lump or cyst or tumour occurs in the human body. These cells may be carcinoma cells. In the case of breast cancer most, of the time carcinoma cells are developed from milk cells. There are different causes in women for an occurrence of breast cancer. Some threat factors are as follows:
Female
Less breast feeding
Late pregnancy after 30 years old
Hormonal therapy
Menopause
Radiation therapy is done on organs near breast e.g. on chest etc.
Nearest blood relatives have breast cancer

There are two types of tumours in human body
1. Benign tumour: This type of tumour is not harmful and do not spread any other organ over body. Benign tumour rarely grow near the organ and can remove by surgery also do not grow back.
2. Malignant tumour: this type of is harmful and may spread any other organ over body. It always grow in an organ. Malignant tumour can remove by surgery but sometimes grow back.

[11]In breast cancer there are three stages T-stage, N-stage, and M-stage. In T-stage breast cells are irregular and the tumour exists and tumour tissue consists of carcinoma cells. In this stage patiently is not serious, it can cure by surgery. In N-stage carcinoma cells got spread to lymph nodes near the breast. Lymph nodes contain cancerous tumour. In this stage also by chemotherapy, this can be cured. In M-stage the carcinoma cells got spread from the breast to other organs like hand, brain, lung, uterus, cervix, liver, throat, leg, chest etc. M-stage is the last stage of breast cancer and in this stage survivability of patient is less, by chemotherapy treatment can be done but the patient may not give a response to treatment, carcinoma cells getting spread 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 the first stage is important. Survivability of patient is high in the first and second stage. The solution of this problem is making diagnosis easy, such that patient can check her health status early. 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.

Importance of health care services

Now-a-days health is a most precious thing in human's life, Due to modern lifestyle and bad environment or some genetic factors, many diseases are occurring in human and

DOI 10.17148/IJARCCE.2017.62115

the average lifetime of the 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 uses 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 that is useful for people.

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

Techniques used in healthcare services
[2]In healthcare services there are human is classified as a healthy and unhealthy person, for this classification different [4] data mining 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-

Supervised learning technique
Supervised learning is a machine learning task of inferring a function from labelled 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 the desired output value. This type of learning algorithm analyses 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 are described as bellow:
Initially it determines the types of training examples and after that gathers the training data and determines the input feature representation of the learned function. Further determines the structure of the learned function and corresponding learning algorithm and completes the design and finally evaluate the accuracy of learned function.

[3]The wide range of supervised learning algorithms are available, each with its strength and weaknesses. The supervised learning classification algorithms are the wide range, some of these are used in a field of medical diagnosis. These are as follows,

C4.5 Decision tree algorithm:
Used to generate a decision tree. [8]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 the 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 a 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 a sample. The drawback of this algorithm is the complexity of building decision tree is high.

K-Nearest Neighbouralgorithm: 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 the 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 the computational cost very high because we need to compute the 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 a 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 classification of images. Handwritten characters are also recognized by SVM.SVM can produce accurate and robust classification result.

Neural Networks:
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 a neural network are the good performance in nonlinear statistical modelling and provide logistic classification. Disadvantage includes its black box nature and greater computational Burdon.

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

DOI 10.17148/IJARCCE.2017.62115

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 the global structure, it did not work well.

II. RELATED WORK

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. Some researcher are used supervised classification technique and some are used unsupervised classification technique.

Andrew Kusiak[12] have used pre-processing of data, a transformation of data and a data mining approach to elicit knowledge about the interaction between parameters measured and survival of the 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 a cost of selecting patients.

Sadikkara[13] had concentrated on a diagnostic research of the neural disease using pattern electrographs signals by using an artificial neural network. The final result was classified in healthy and diseased. The final result has shown with effective interpretation.
Abhishek[14] 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. The aim of the author was proposing the best tool for diagnosis e.g. identification of kidney stone, by reducing the requirement of time for diagnosis and accuracy as well as efficiency got improved.
Ashfaq Ahmad k[15] 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 analyse for the purpose of prediction
Basmabackend [16] presented the big data evaluation in a healthcare system and they applied a learning algorithm on a set of medical data. The aim of an author is predicting chronically kidney disease by using C 4.5 decision tree algorithm is used to improve the performance of prediction of results in terms of minimum execution time and accuracy.

This 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.

III. PROPOSED METHODOLOGY

Architecture:

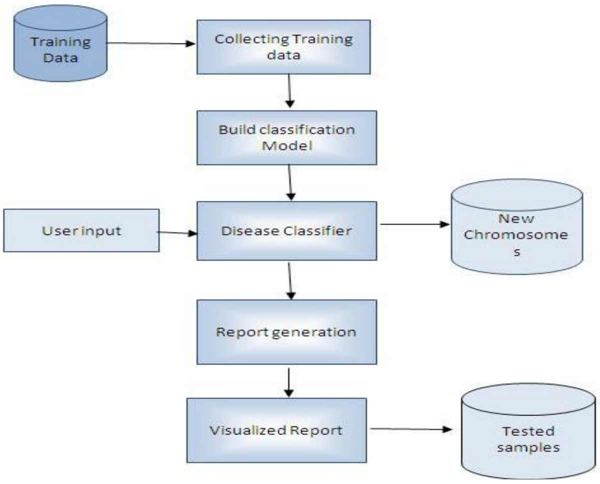

Figure 1 System Architecture

Training Data Collector:
Data is necessary to solve the problem and this data is 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 center, 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. These patients may have Breast cancer and they have done diagnosis in a hospital.
Datasets are stored in the different location as per their stage such as if T, which is the first stage of Breast cancer, if N which is the second's stage of Breast Cancer and M which is the last stage of Breast Cancer. This class Training Data Collector checks the stage of disease of the patient and stored in an appropriate location.

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 the new document.

Steps to build classification model
i. For classification model the first number of categories is determined, as Breast Cancer have three stages, a total

DOI 10.17148/IJARCCE.2017.62115

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 analysed for a 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 a number of documents are less then there have to generate more generations and if a 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. [9]At this step for each category Hc (Pos, Neg) over training data built. Where, Pos(t1,t2,….,in) and Neg(tn+1,tn+2,….,tm) 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 Hc (Pos, Neg) is returned

Disease Classifier
This is the testing phase. In this class document d is classified 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 …. Or am belongs to d) holds. The disease classifier consists following steps

New patient enters their biomedical data to system, which contains many factors name, age, height, weight, BP, HR, symptoms, tests, test reports. Above entered data contains set of chromosomes called document d. now the aim of this class is to classify d into the corresponding category
[17]$c \ = (t1\ \varepsilon d\ \lor\ldots\lor tn\varepsilon d) \land \neg(t(n+1)\varepsilon d\ \lor..\lor tn + m\varepsilon d)$
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.
iii. In d if any new chromosome detects then that will store in another database "New". And that will train after.

Generate Report
In the 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 the medical person or non-medical person, all should understand the report. In the report, there is the relation between the location where carcinoma cells spread and current stage of a patient is shown so the patient will understand her health status. We provided final report which has following factors: Name, Disease, Factors of input data patient due to which patient is classified into particular stage, Stage detected by Olex GA, Stage detected by C4.5, visualized Template
The template shown in the report contains two bars.
· First bar indicates that in which part carcinoma cells got spread, this bar has three parts skin, lymph node, and another organ. The carcinoma spread part is indicated by red colour. And the where carcinoma cells are absent is indicated by green colour
· Second bar indicates that in which stage of breast cancer the patient have currently. This bar has three parts T, N and M. If a patient is in T stage then only T part is red and remaining parts are green. If a patient is in N-stage then T and N parts are red and M part in green. And finally, if a 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 a relation between parts where carcinoma cell are a present and current stage of breast cancer the patient have, Due to this patient will understand that why she have that particular stage of breast cancer.

TABLE I. Stages in breast cancer

Carcinoma Cells in Patient/Cancerous tumour in patient	Stage
Skin of breast, Breast	T
Lymph nodes	N
Other organs (brain, lung, chest, liver, kidney, uterus, etc.)	M

Algorithm
The Genetic Algorithm:
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.[7] 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 Olex GA:
The Olex GA is the text-based classification algorithm in which patient is classified into one of the three categories of breast cancer by analysing her biomedical data. The current stage is decided depending on the negative texts present in patient's biomedical data. The set of positive and negative texts are decided in the stage of training data. The problem of inducing propositional text classifiers of the form

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

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.

IV. RESULT AND ANALYSIS

Dataset

Data is necessary to solve the problem and this data is 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.

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.

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,

TABLE II Metrices of comparison

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)

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.

TABLE III Results of test cases set1

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

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.

TABLE IV Results of test cases set2

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

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.

TABLE V Results of test cases set3

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 (av2)	90	80
Percentage of error(e2)	10	20

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.

DOI 10.17148/IJARCCE.2017.62115

TABLE VI Results of test cases set4

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

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.

TABLE VII Results of test cases set5

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

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.

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	Percentage of error
C 4.5	82	18
OlexGA	94	6

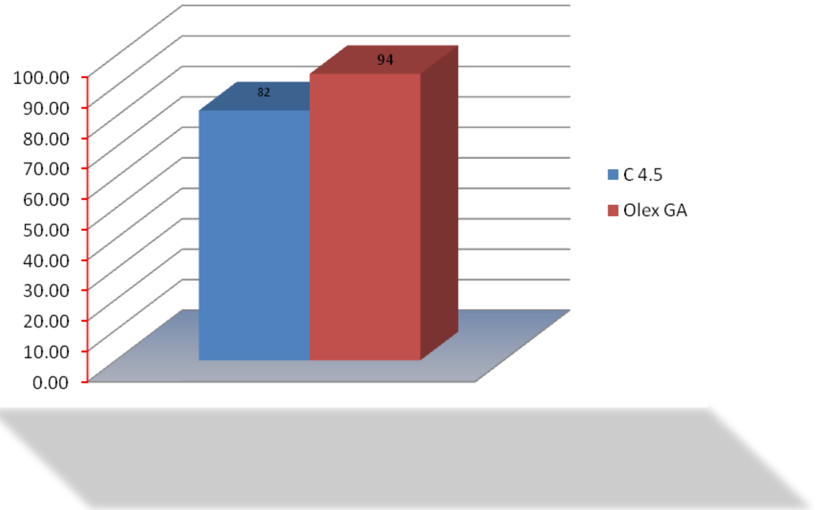

Figure 2 Comparison of accuracies of C4.5 and Olex GA

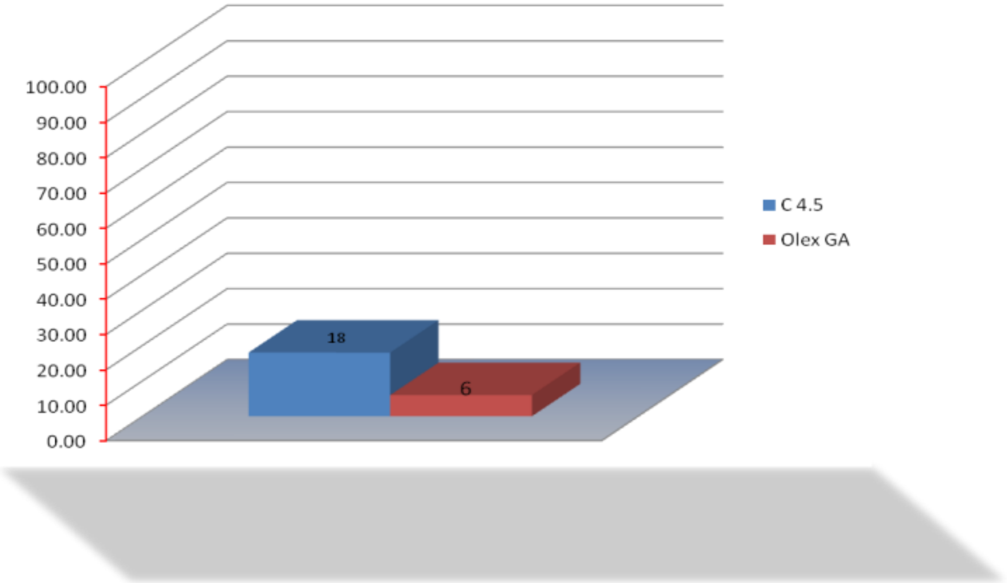

Figure 3 Comparison of errors of C4.5 and Olex GA

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

V. CONCLUSION AND FUTURE SCOPE

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. 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.

REFERENCES

[1] Xiaoli Li and Bing Liu “Learning to Classify Texts Using Positive and Unlabeled Data” in MIT Allience 2016

[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 Cancerdata” in International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)Volume-3 Number-4 Issue-13 December-2013

[4] ShwetaKharya “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] ZehraKarapinarSenturk and Resul Kara “BREAST CANCER DIAGNOSIS VIA DATA MINING: PERFORMANCEANALYSIS OF SEVEN DIFFERENTALGORITHMS” in Computer Science & engineering: An International Journal(CSEIJ), Vol. 4, No. 1, February 2014

[6] AbdelghaniBellaachia, ErhanGuven “Predicting Breast Cancer Survivability Using Data Mining Techniques”

[7] Shelly gupta, dharminderkumar and anand Sharma “data mining classification techniques applied forbreastcancerdiagnosis and prognosis” in Indian Journal of Computer Science and Engineering (IJCSE) Vol. 2 No. 2 Apr-May 2011ISSN : 0976-5166

[8] RonakSumbaly, N. Vishnusri, and S. Jeyalatha “Diagnosis of Breast Cancer using Decision Tree Data MiningTechnique” 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 TextClassification Rule Induction” in W. Daelemans et al. (Eds.): ECML PKDD 2008, Part II, LNAI 5212, pp. 188–203, 2008 cSpringer-Verlag Berlin Heidelberg 2008.

[10] LiqiangNie, Member, IEEE, Yi-Liang Zhao, Mohammad Akbari, JialieShen, Member, IEEE, and Tat-SengChua, 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] VikasChourasia 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.

DOI 10.17148/IJARCCE.2017.62115

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

[13] Sadik Kara, AysegulGuvenb, AyseOztUrkOnerc, (2006) "Utilization of artificial neural networks in thediagnosis of optic nerve diseases", Elsevier Publication, Computers in Biology and Medicine, Vol. 36, pp 428–437

[14] Abhishek, GourSundarMitra Thakur, Dolly Gupta, (2012) "Proposing Efficient Neural Network Training Modelfor Kidney Stone Diagnosis", International Journal of Computer Science and Information Technologies, Vol. 3 (3), pp3900-3904

[15] Basma Boukenze1, HajarMousannif and AbdelkrimHaqiq "Predictive Pnalytics In Healthcare System usingData mining techniques" International Journal of Computer Science and Information Technologies

[16] Ashfaq Ahmed K, Sultan Aljahdali and Syed NaimatullahHussain, (2013) "Comparative Prediction Performancewith Support Vector Machine and Random Forest Classification Techniques", International Journal of ComputerApplications Vol. 69, No.11, pp 12-16

[17] Adriana Pietramala, University of Calabria "A Genetic Algorithm for Text Classification

[18] Rule Induction" on http://videolectures.net/ecmlpkdd08_pietramala_agaf/

DOI 10.17148/IJARCCE.2017.62115