Implementation of breast Cancer Prediction System (BCPS): To Assist in Prediction of breast Cancer-Based on Data Mining Techniques

Alaa El Din M El Ghazali
Department of Computer
and Information Systems, Sadat Academy
for Management Sciences

Ahmed Ali Obaya
Lecturer of clinical oncology Faculty
of medicine Zagazig University

Amr Ibrahim Moubarak
Demonstrator at El Shrouk Academy,
Management Information System

1. Introduction

Data mining is usually defined as the extraction of before unknown and possibly valuable information from a database. With the rising volumes of electronic patient records, data mining has become general to excerpt hidden patterns inpatient data for healthier understanding of relationships within the data. Data mining in medical domain is single from that in other domains due to the special characteristics of medical datasets [1].

Breast cancer rates are increasing in developing countries, including Egypt, and are largely attributed to the aging of the population, delay in time of first pregnancy, decrease in number of children and in breast feeding, and a move toward high-calorie Western diets [2]. Although breast cancer incidence rates in Egypt are substantially lower than the rates in the United States and other developed countries, breast cancer is the most commonly cancer among women in Egypt. Furthermore, the current demographic trends favor the likelihood that breast cancer will become an even greater public health concern in Egypt in the future [3].

In Egypt, breast cancer is estimated to be the most commonly cancer among females accounting for 37.7% of their total with 12,621 new cases in 2008. It is also the leading cause of cancer related mortality accounting for 29.1% of their total with 6546 deaths. These estimates are confirmed in many regional Egyptian cancer registries [4].

In this paper, data mining is used to offer unique opportunities to predict the survival of patients with fatal diseases and predict treatment outcomes. This thesis presents an approach for diagnosis of breast cancer disease

Breast cancer is one of the fatal diseases in the world nowadays. It is caused by some genetic and non-genetic factors. It is primary cause of death in developed and developing countries. Early detection of cancer is the perfect way to reduce it, so that it may be curable. The target is to reduce mortality rates by diagnosis of breast cancer earlier in Egypt and thereforee having a better chance of surviving from this disease. The goal of breast cancer prediction system and diagnosis is to help female and doctors to discover the disease before the appearance of symptoms and help millions of lives. The system is a breast cancer prediction application which will be an easy way for normal prediction of breast cancer and according to some steps will be followed by the female whatever if she is affected or not to be as early prediction. This study is intended to enhance the breast cancer diagnosis by focusing on using the data mining techniques. Classification is a major technique in data mining and widely used in various fields. It's a data mining function that assigns items in a collection to target categories or classes, and accurately predict the target class for each case. Final results through WEKA which use a decision tree that shows that applying C4.5 builds a more efficient tree which gives high prediction accuracies, faster and better results. As a result, the tree generated by C4.5 algorithm was much higher than the rest. As seen from results, data mining techniques can contribute to breast cancer diagnosis performance by predicting the breast cancer.

Keywords: Breast cancer in Egypt, Medical Data Mining, Classification, Decision Tree, WEKA, RandomTree, C4.5, REPTree, SimpleCart.

using data mining techniques. The predictive System has a mission which is Women's health promotion through breast cancer awareness spreading, education and early detection of this a fatal disease by providing a test method which is affordable and accessible to anyone.

Early detection is key in the treatment of breast cancer. The earlier patients can detect the signs and factors of the disease, the more opportunity of recovery increases. There are steps you can take to detect breast cancer early when it is most treatable, since the early stages of breast cancer are much higher in recovery possibility than late diagnosed stages of the disease. As time passes on the disease without detecting it, the surviving rates grossly decrease.

In the near future, the software aims to spread awareness of breast cancer symptoms, the consequences of late and early detection of the disease, and eradicate the backward idea of stigma attached to having breast cancer or to announce it.

2. Related Work

In the first study, data mining techniques are used to predict the cancer disease at an earlier stage. Different researchers have proposed different techniques to predict the cancer disorder and different kinds of accuracy level as per used techniques. These techniques help to minimize the irrelevant data of patient's data from the databases in medical center. Algorithms such as decision tree, ANN, Support Vector Machine, Naïve bays, Back propagation and Regression are considered for the study. These algorithms gave thevarious results based on speed, accuracy, performance and cost. Also these effective classification data helps to find the treatment to the patient [5].

In the second study, used data mining techniques for diagnosis and prognosis of cancer. Cancer is the most important cause of death for both men and women. The early detection of cancer can be helpful in curing the disease completely and presented a system for diagnosis and prognosis of cancer using Classification and Association approach in Data Mining to predict the possibility of cancer in context to age. The result showed the Decision tree is found to be best predictor [6].

In the third study, two data mining techniques are used to predict breast cancer risks in Nigerian patients by using the naïve Bayes algorithm and the J48 decision tree algorithms. The performance of both classification techniques was evaluated in order to determine the most efficient and effective model. The J48 decision trees showed a higher accuracy with lower error rates compared to that of the naïve Bayes algorithm method while the evaluation criteria proved the J48 decision trees to be a more effective and efficient classification techniques for the prediction of breast cancer risks among patients of the study location [7].

In the fourth study, an overview of the current research is carried out by using the data mining techniques to enhance the prognosis of lung cancer. Aim of the study is to propose a model for early detection and correct diagnosis of the disease which will help the doctor in saving the life of the patient [8].

In the fifth study, a performance comparison between different machine learning algorithms: Support Vector Machine (SVM), Decision Tree (C4.5), Naive Bayes (NB) and k Nearest Neighbors (k-NN) on the Wisconsin Breast Cancer (original) datasets is conducted. The main objective is to assess the correctness in classifying data with respect to efficiency and effectiveness of each algorithm in terms of accuracy, precision, sensitivity and specificity. Experimental results show that SVM gives the highest accuracy with lowest error rate [9].

3. Data Mining Concepts

There are many definitions of data mining [10]

Data mining is the process of searching for specific and useful information through large amounts of data. It is also the process of analysis and exploration of large amounts of data by using automatic or semi-automatic means in order to find and discover meaningful patterns and rules.

Data mining is to discover and obtain knowledge of patterns associations, changes, anomalies and significant structures through large amounts of data stored in database, data warehouses or stored in other information stores.

Data mining consists of five major elements [11]:

1. Extract, transform, and load transaction data onto the data warehouse system.
2. Store and manage the data in a multidimensional database system.
3. Provide data access to business analysts and information technology professionals.
4. Analyze the data by application software.
5. Present the data in a useful format, such as a graph or table.

3.1. Data Mining Techniques

Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. While data mining and knowledge discovery in database are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The sequences of steps identified in extracting knowledge from data are as represented in knowledge discovery. Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc. are used for knowledge discovery from databases [12]. The different classification algorithms mentioned below in Fig. 1 are used to predict or to analyze various diseases.

Fig. 1 Different Techniques in Data mining

3.1. Selection Data Mining Techniques

Objective of this paper is to build the classification model using a decision tree algorithm. The decision tree is a very good method because it is relatively fast, it can be converted to a simple rule, and accuracy level is high. The decision tree function in this research is focusing on classifying the data, to reach our data mining goal, which will be used to find the relationship between specific features. By using the RandomTree, C4.5, REPTree and SimpleCart algorithm .In this research RandomTree, C4.5, REPTree and SimpleCart is applied because they commonly used algorithms.

1. Proposed System Architecture

Predictive system helps the user to predict the risk of breast cancer. This is made of steps that include the usage of data mining and mathematical probability to prognostic outcomes. This system consists of hardware, software and database which is a number of predictive factors, these factors are also called attributes that are likely to affect the output of this system. A sample of data which is related to the output is collected, and the predictive attributes are assigned after the data clean and selection stages, then a mathematical method is formulated to retrieve the output of the method. This system depends on the model which can be shown in Fig. 2. The model is working on recognizing patterns and information from historical data and building a decision function to make predictions based on the factors extracted from these historical data. In Fig. 3 which shows that, once the user enters into the breast cancer prediction system (BCPS), she needs to answer the questions, related to genetic and non-genetic factors. Then the BCPS assigns the risk value to each question based on the user answers. Once the risk value is predicted, the range of the risk can be determined by the prediction system and this is the system feedback. The system can be used to predict anything from meteorology and weather forecasting to medical fields.

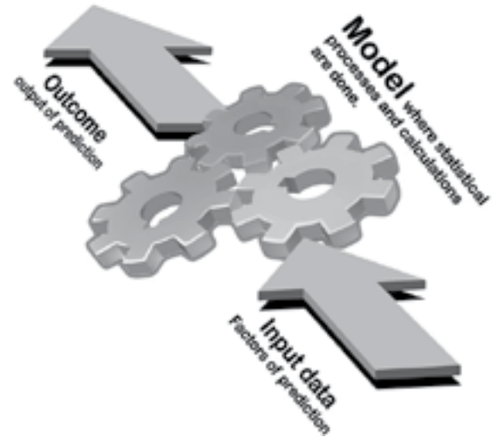


Fig. 2 Cycle of Predictive Modeling

The proposed system Architecture includes stages as follows see: (Fig. 3)

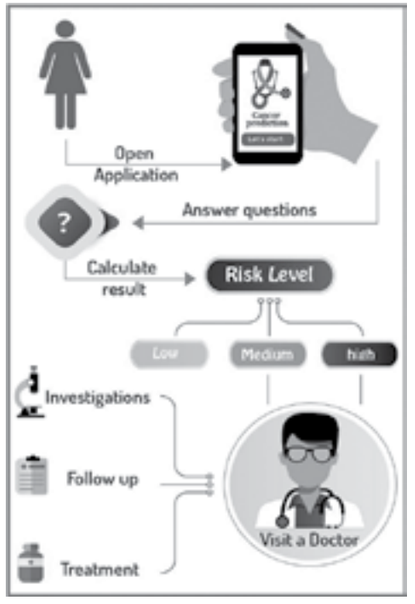


Fig. 3 Architecture of breast Cancer Prediction System

1. Implementation of breast Cancer Prediction System

The BCPS is a stand–alone system it is an application which could be downloaded to any Mobile and the test is done through some questions.

5.1. Software Requirements Specification (SRS)

A software requirements specification (SRS) is a detailed description of software to be developed with its functional and non-functional requirements. It may include the use cases of how user is going to interact with system. To develop the system we should have a clear understanding of system. To achieve this we need to continuous communication with doctors to gather all requirements. Functional requirements are given to show the system features and expected user interaction.

Non-functional requirements are outlined for later verification. They are classified into:

5.1.1. Nonfunctional Requirements

Nonfunctional Requirements define system attributes such as security, reliability, performance, maintainability, scalability, and usability. They serve as constraints on the design of the system across the different backlogs. Nonfunctional requirements may include:

5.1.1.1. Operational Requirements

Usability: 80% of users will not need to read the user Help to be able to use the system, as it was designed to be easy and simple to use.

5.1.1.2. Performance Requirements

Performance requirements specific to the metrics and parameters that describe the product's capacity. Performance requirements usually cover the following concerns:

1. Response Time

The system response times are clearly identified as part of a business case. The system is reacting instantaneously, meaning that no special feedback is necessary except to display the result. All the calculations and predictive process in the background are instantaneously.

2. Workload

The business case or existing process should be the start of the workload definition. The workload is often described as the scenarios that the users are likely to execute. The system can take too much load of transactions at any particular moment. The system is built and designed to receive thousands of users and the underlying infrastructure is able to support the high traffic.

3. Scalability

In one respect scalability is simply specified as the increase in the system's workload that the system should be able to process. The scalability required is often driven by the lifespan and the maturity of the system.

The breast cancer prediction system is not expected to face any sudden increase in workload, since it will smoothly become popular with people.

4. Platform

A system platform is defined as the underlying hardware and software (operating system and software utilities) which will house the system.

The breast cancer risk prediction system will include a mobile application designed to work on Android and iPhone Operating System (IOS) platforms.

5.1.1.3. Security Requirements

Breast Cancer Prediction System (BCPS) account security is provided by secure login to the Account. Login information input via BCPS application will not be stored. All users' information are safe and protected against any hack threatens. User approaches as anonymous, and no one can access any private information or any kind of data belongs to the system user.

5.1.1.4. Documentation and Training

The system will be delivered to users as a download program without documentation or training. User guide and system documentation will be provided to stakeholders.

5.1.2. Functional Requirements

A functional requirement defines a function of a system or its component, where a function is described as a specification of behavior between outputs and inputs. And it's a basic functionality or desired behavior documented clearly and quantitatively. Sequence diagram is one of the functional requirements main collections:

5.1.2.1. Unified Modeling Language (UML) Sequence Diagrams

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.

A sequence diagram shows, as parallel vertical lines (lifelines), different processes or objects that live simultaneously, and, as horizontal arrows, the messages exchanged between them, in the order in which they occur. This allows the specification of simple runtime scenarios in a graphical manner. They are shown in the following two figures.

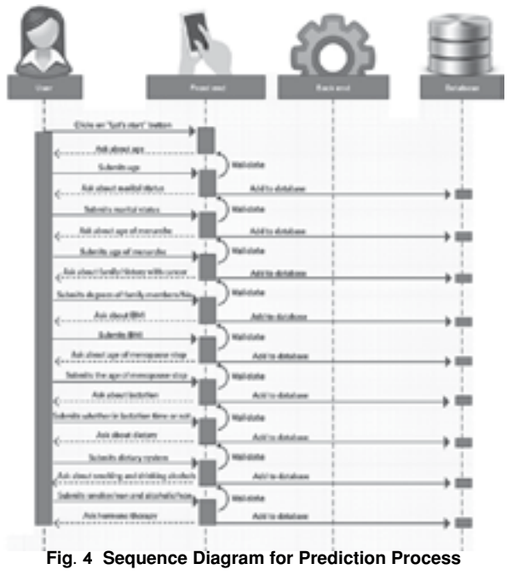


Fig. 4 Sequence Diagram for Prediction Process

The first part of the sequence diagram shows the repeated approach of collecting predictive factors and validate them first, then record them in database.

After collecting all needed data, classification method in the backend starts the prediction process. A ratio represents the level of risk is outputted to the user.

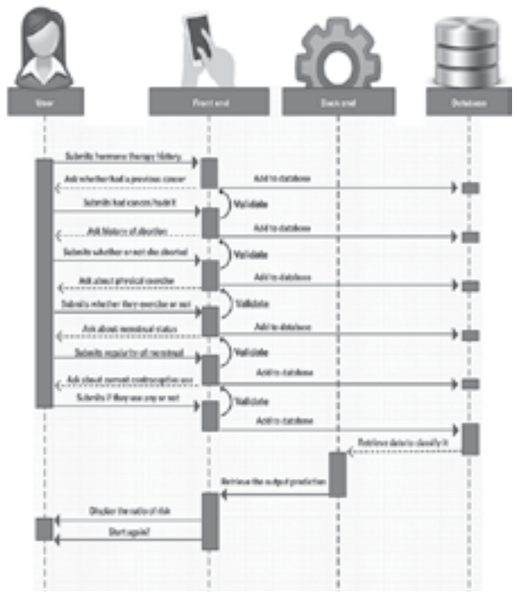


Fig. 5 Sequence Diagram for Prediction Process and Final Result

5.1.1.1. Stakeholders

The term stakeholder is used to refer to any person or group who will be affected by the system or affected by its outcome, or can influence its outcome, directly or indirectly. Stakeholders include end-users who interact with the system and everyone else in an organization that may be affected by its installation. Other system stakeholders may be engineers who are developing or maintaining related systems, business managers, domain experts, and trade union representatives. The stakeholder analysis improves the identification and understanding of stakeholders in the system. System stakeholders are categorized as follows:

1. System users: daily female system users are the most important segment in stakeholders. The system guarantees an easy and simple design and smooth to use which satisfies all users' preferences. The system also aims to target all ages of women.

2. Employees: A software team to develop the system and to work on maintaining the system, analysts and designers to guarantee the best service to the system users.

3. Organizations: Medical and statistical organization may offer a collaboration with the system in seek of improving the diagnosis of the disease.

5.2. Context Model of breast Cancer

A context model defines how context data are structured and maintained it aims to produce a formal or semi-formal description of the context information that is present in the system. In other words, the context is the surrounding element for the system, and a model provides the mathematical interface and a behavioral description of the surrounding environment. The system deals with three main entities from its surrounding context:

1. Women society: which is the largest group of the users base, the data transaction starts when user submit the predictive attributes on the system, then the system displays back his risk level of breast cancer diagnosis.

2. Employees: that maintains and develops the system regularly.

3. Medical organizations: that guarantees the validity of the predictive health conditions and factors which the system builds its prediction on, and may benefit from the collected data and use them in fighting disease research.

The context diagram shows how data travel among those main objects and how the system collects its data. Thit can be represented in the following Figure

Fig. 6 Context Model

5.1. User Interface

The general personality of the interface should convey a conservative, professional, authoritative or fun attitude.

The interface will be intuitive. As a mobile app it will be streamlined and simple to use. No training will be provided and it is expected that 90% of users will be able to use the app without any training. Thit can be represented in Fig. 7a and Fig. 7b

Fig. 7 Sample Design Applications

1. Output Result of breast Cancer Prediction System

The breast cancer prediction system has an outcome of an integer percentage value represents the level of risk. There are three levels of risk which are low, intermediate and high risk. Based on the predicted risk values the range of risk can be estimated. Thit can be represented in Fig. 8

Fig. 8 Result of breast Cancer Prediction System (BCPS)

1. Analysis of breast cancer prediction system

There are four data mining techniques were chosen to find the effectiveness of BCPS which are RandomTree, C4.5, REPTree and SimpleCart.

7.1 Classification

In this paper, the classification approaches are used to predict breast cancer diagnosis and there are two cases Benign and Malignant and how other attributes affect them. Four classification techniques are used which are decision tree (RandomTree, C4.5, REPTree and SimpleCart). A decision tree classifier extracts a set of rules that show relationships between attributes of the data set and the class label.

It uses a set of IF-THEN rules for classification. Rules are easier for humans to understand, given a class, from training data and then allow the use of these probabilities to classify new entities.

7.2 Experimental Results and Discussions

There are some accuracy and inaccuracy measures on the basis of the performance of classifiers to be evaluated. Along with them there are some error measures which are used to find out and then the predicted value is from actual known value.

In this test we can use systematic method to increment sample by 5%.

Classification accuracy is our starting point. It is the number of correct predictions made divided by the total number of predictions made, multiplied by 100 to turn it into a percentage.

7.2.1 Accuracy Classified

The Accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier. Thit can be shown in the following Table 1

Table 1 Accuracy Classified

Training Set (→) / Algorithm (↓)	35%	40%	45%	50%	55%	60%	65%	70%
Random Tree	90%	93.84%	90.41%	90.84%	92.96%	92.98%	94.97%	92.98%
C4.5	90%	94.72%	93.61%	94.36%	93.75%	95.61%	90.95%	92.98%
REPTree	91.35%	94.42%	91.05%	93.66%	93.75%	92.10%	94.47%	94.73%
SimpleCart	93.24%	93.54%	93.92%	94.01%	93.75%	92.54%	92.96%	92.98%

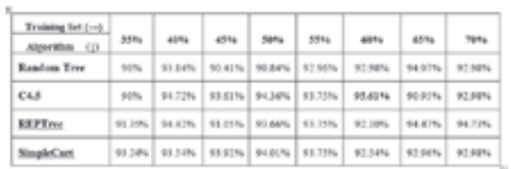
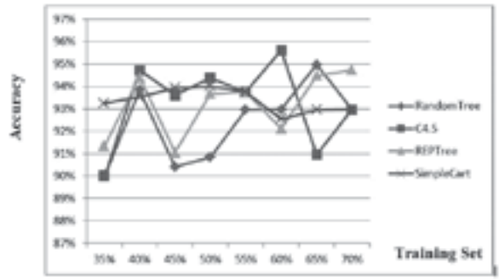


Fig. 9 Accuracy Classified



In Fig. 9, accuracy was calculated for RandomTree, C4.5, REPTree and SimpleCart by using WEKA through eight tests of the sample they are (35%, 40%, 45%, 50%, 55%, 60%, 65% and 70%). It was found that the best percentage of Training set is 60% and Testing Set 40% at C4.5 algorithm because it gave us the highest accuracy (95.61%).

Training Set (→) / Algorithm (↓)	35%	40%	45%	50%	55%	60%	65%	70%
Random Tree	90%	93.84%	90.41%	90.84%	92.96%	92.98%	94.97%	92.98%
C4.5	90%	94.72%	93.61%	94.36%	93.75%	95.61%	90.95%	92.98%
REPTree	91.35%	94.42%	91.05%	93.66%	93.75%	92.10%	94.47%	94.73%
SimpleCart	93.24%	93.54%	93.92%	94.01%	93.75%	92.54%	92.96%	92.98%

7.2.2 Incorrectly Classified

The incorrectly shows the percentage of error in data during analysis, so that the incorrectly classified can be represented in the following table 2

Table 2 Incorrectly Classified

Classification (→) / Algorithm (↓)	Correctly Classified Instances	Incorrectly Classified Instances
Random Tree	212	16
C4.5	218	10
REPTree	210	18
SimpleCart	213	17

Fig 10 Incorrectly Classified

In Fig. 10, it was calculated incorrectly for RandomTree, C4.5, REPTree and SimpleCart by using WEKA through eight tests of the sample they are (35%, 40%, 45%, 50%, 55%, 60%, 65% and 70%). The best percentage of Training set is 60% and Testing Set 40% at C4.5 algorithm because it gave us the lowest Incorrectly (4.38%).

2. Breast Cancer Evaluation of Dataset

Total numbers of instances in breast cancer evaluation dataset are 569 instances. Tenfold cross validation technique is used to obtain the number of classified and in classified number of instances in breast cancer evaluation datasets.

Training Set (→) / Algorithm (↓)	35%	40%	45%	50%	55%	60%	65%	70%
RandomTree	5.0%	4.15%	9.3%	9.13%	7.69%	7.01%	5.82%	7.01%
C4.5	5.0%	5.27%	6.38%	5.63%	6.25%	**4.38%**	9.04%	7.01%
REPTree	8.64%	5.57%	8.94%	8.33%	6.25%	7.89%	5.32%	5.26%
SimpleCart	6.75%	6.45%	6.97%	5.9%	6.25%	7.45%	7.80%	7.01%

Table 3 shows the number of correctly classified and incorrectly classified instances for four classification algorithms. C4.5 algorithm with the highest number of 218 classified instances and REPTree algorithm with greatest number of incorrectly classified instances 18. REPTree 18 incorrect instances are highest as compared to number of incorrectly classified instances of other three studied algorithms.

Fig. 11 Number of Classified Instances for breast cancer Evaluation Data Set

From the Table 3 and fig. 11 it is evident that from breast cancer evaluation dataset C4.5 has the highest number of correctly classified instances followed by RandomTree algorithm. All the two algorithms perform well in classifying the instances. RandomTree algorithm shows average performance. Whereas SimpleCart and REPTree algorithm have the lowest performance in terms of correctly classification of instances.

Table 3 shows the number of correctly classified an incorrectly classified instances for four classification

algorithms. C4.5 algorithm with the highest number of 218 classified instances and REPTree algorithm with greatest number of incorrectly classified instances 18. REPTree 18 incorrect instances are highest as compared to number of incorrectly classified instances of other three studied algorithms.

Fig. 11 Number of Classified Instances for breast cancer Evaluation Data Set

From the Table 3 and fig. 11 it is evident that from breast cancer evaluation dataset C4.5 has the highest number of correctly classified instances followed by RandomTree algorithm. All the two algorithms perform well in classifying the instances. RandomTree algorithm shows average performance. Whereas SimpleCart and REPTree algorithm have the lowest performance in terms of correctly classification of instances.

1. CONCLUSION

Breast cancer is a fatal disease. Breast cancer represents a major threat to female. Detection of breast cancer is challenging for the doctors till now. Even now the reason and complete treatment of breast cancer is not invented. If the breast cancer is detected early it may be curable. In this work we developed a system called data mining based BCPS. The main goal of this system is to provide early warning to female, and it is also free and not time consuming. There are three specific breast cancer risks which are low, intermediate and high risk. Breast cancer prediction system can estimate the risk of breast cancer by examination of a number of user-provided genetic and non-genetic factors. This system compares its predicted results with the patient's previous medical records and then it's analyzed by using WEKA system. This system is a mobile application over the internet so the female can use it easily and check her risk and take appropriate action based on her risk status.

Based on this analysis, we can confirm that the best algorithm is C4.5 because it gave us the highest accuracy (95.61%) and gave the lowest incorrectly (4.38%) compared with RandomTree, REPTree and SimpleCart.

2. REFERENCES

1. P. Ponmu thuramalingam, M . Yasodha,"An Approach On Sequential Data Mining Algorithm For Breast Cancer Diseases Management", International Journal of Advanced Research in Computer Science and Software Engineering, 2013.

2. Zakaria Suliman Zubi and Rema Asheibani Saad, "Improves Treatment Programs of Lung Cancer Using DataMining Techniques»,Journal of Software Engineering and Applications, 2014.

3. V.Krishnaiah , G.Narsimha and N.Subhash Chandra,» Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques»,International Journal of Computer Science and Information Technologies,2013.

4. Tawfik R. Elkhodary , Mohamed A. Ebrahim , Elsayed E. Hatata ,Nermeen A. Niazy, "Prognostic Value of Lymph Node Ratio in Node-Positive breast cancer in Egyptian patients ",Journal of the Egyptian National Cancer Institute,2014.

5. P. Saranya and B. Satheeskumar , "A Survey on Feature Selection of Cancer Disease Using Data Mining Techniques ", International Journal of Computer Science and Mobile Computing, May- 2016

6. Jaimini Majali, Rishikesh Niranjan, Vinamra Phatak and Omkar Tadakhe, "Data Mining Techniques for Diagnosis Jaimini Majali, Rishikesh Niranjan, Vinamra Phatak and Omkar Tadakhe, "Data Mining Techniques for Diagnosis and Prognosis of Cancer", International Journal of Advanced Research in Computer and Communication Engineering, March 2015.

7. Peter Adebayo Idowu, Kehinde Oladipo Williams, Jeremiah Ademola Balogun and Adeniran Ishola Oluwaranti, " Breast Cancer Risk Prediction Using Data Mining Classification Techniques", «http://dx.doi.org/10.14738/tnc.32.662», April 2015.

8. Supreet Kaur1 and Amanjot Kaur Grewal, " A Review Paper On Data Mining Classification Techniques For Detection Of Lung Cancer", International Research Journal of Engineering and Technology (IRJET), Nov -2016.

9. Hiba Asria, Hajar Mousannifb, Hassan Al Moatassimec and Thomas Noeld, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis", https://www.sciencedirect.com/ , 2016.

10. Neha Rathee, Sarika Choudhary , "Study Of Different Data Mining System & Platform", International Research Journal of Engineering and Technology (IRJET),2015.

11. Kawasar Ahmed, Tanuba Jesmin, Md.Zamilur Rahman "Early Prevention and Detection of Skin Cancer using Data mining", International Journal of Computer Application, 2013 Volume 62-No.4.

12. Parvez Ahmad, Saqib Qamar and Syed Qasim Afser Rizv « Techniques of Data Mining In Healthcare « International Journal of Computer Applications (0975 - 8887) Volume 120 - No.15, June 2015