The Role of Data Mining-Based Cancer Prediction system (DMBCPS) in Cancer Awareness

A.PRIYANGA
M.Phil (CS) Research Scholar
SCSVMV University
Kanchipuram, India
priyaa.madhu88@gmail.com

Dr.S.PRAKASAM
Assistant Professor
SCSVMV University
Kanchipuram, India
prakasam_sp@yahoo.com

ABSTRACT: Cancer is one of the major problem today, diagnosing cancer in earlier stage is still challenging for doctors. Breast cancer is one of the major death causing diseases of the women today all over the world. Every year more than million women are diagnosed with breast cancer worldwide over half of them will die because of the late diagnosing of the disease. So many researches have undergone for detecting the cancer based on data mining technology each approach has its own limitations. This makes us to take up this problem and to implement the Data mining based cancer prediction System (DMBCPS). We have proposed this cancer prediction system based on data mining technology. This system estimates the risk of the breast cancer in the earlier stage. This system is validated by comparing its predicted results with patient's prior medical information and it was analyzed by using weka system. The main aim of this model is to provide the earlier warning to the users, and it is also cost efficient to the user**.**

INTRODUCTION: The body is made up of trillions of living cells. Normal body cells grow, divide into new cells, and die in an orderly way. Cancer begins when cells in a part of the body start to grow out of control. Cancer cell growth is different from normal cell growth. Instead of dying, cancer cells keep on growing and form new cancer cells. Breast cancer is a malignant tumor that starts in the cells of the breast. Breast cancer is characterised by the uncontrolled growth of abnormal cells in the milk producing glands of the breast. There is no sure way to prevent from the breast cancer. But there are things all women can do that might reduce their breast cancer risk. Breast cancer not only found in women, men also have less chance of getting this cancer. Various tests are available for predicting breast cancer, but detecting cancer in earlier stage is difficult, but earlier detection of cancer is curable. In the following sections, previous researches are discussed. We have proposed the cancer prediction system based on data mining. Cancer prediction system estimates the risk of the breast cancer at the earlier stage. This system was validated by comparing its predicted results with patient's prior medical information and analyzed through weka tool.

*Prior Studies of Cancer Prediction***:**

K. Rama Lakshmi et al [2013] this research paper analyzes how data mining techniques are used for predicting different types of major life threatening diseases. It reviewed the research papers which mainly concentrated on predicting heart disease, Diabetes, Breast cancer, HIV/AIDS and Tuberculosis. Ankit Agarwal et al [2011] collected a data from SEER dataset and develop the accurate survival prediction model for lung cancer using data mining techniques. They were used several classification techniques for preprocessing the data and they used tree classifiers for best prediction. They have developed the online lung cancer outcome calculator for estimating risk of morality after 6 months 9 months, 1 year, 2 years, and 5 years of diagnosis. Seyyid Ahmed Medjahed et al [2013] worked on K-NN method. It is one of the popular methods used to diagonise breast cancer. The quality of the results depends largely on the distance and the value of the parameter "k" which represent the number of the nearest neighbors.

Paper Type: Research
Corresponding Author: Priyanga.A, email: priyaa.madhu88@gmail.com
scientistlink.com

In this paper, they study and evaluate the performance of different distances that can be used in the K-NN algorithm. Also, they analyze this distance by using different values of the parameter "k" and by using several rules of classification. This work will be performed on the WBCD database (Wisconsin Breast Cancer Database) obtained by the university of Wisconsin Hospital. Shwetha kharya [2013] collected a data from SEER dataset and various data mining techniques were used to predict the breast cancer. Among the various data mining techniques she found that decision tree is the best classifier with greater accuracy.

Architecture of Data Mining - Based Cancer Prediction System

Detecting cancer is still challenging for the doctors in the field of medicine. Even now the actual reason and complete cure of cancer is not invented. Various tests are available for predicting cancer, but detecting cancer in earlier stage is difficult, but earlier detection of cancer is curable. With the help of data mining we try to predict the risk of cancer in earlier stage. We develop a system called the cancer prediction tool which predicts three specific cancer risks. Specifically, Cancer prediction tool estimates the risk of the breast, skin, and lung cancers by examining a number of user-provided genetic and non-genetic factors. The main aim of this model to provide the earlier warning to the users, to make a precaution based on their risk status.

The Architecture of Data mining – Based Cancer Prediction System

In this work, architecture is designed and implemented using decision tree algorithm (Data mining technique). Decision tree is one of the easier data structure to understand data mining. Rules from the training dataset are first extracted to form the decision tree which is then used for classification of the testing dataset. A decision tree is necessarily a tree with an arbitrary degree that classifies instances. When the user login into the cancer prediction system the Home screen will provide the information about cancer. It will give the details about the cancer which can be predicted by the cancer prediction system. It also shows characteristic features which are considered to be the increasing risk factor for causing cancer.

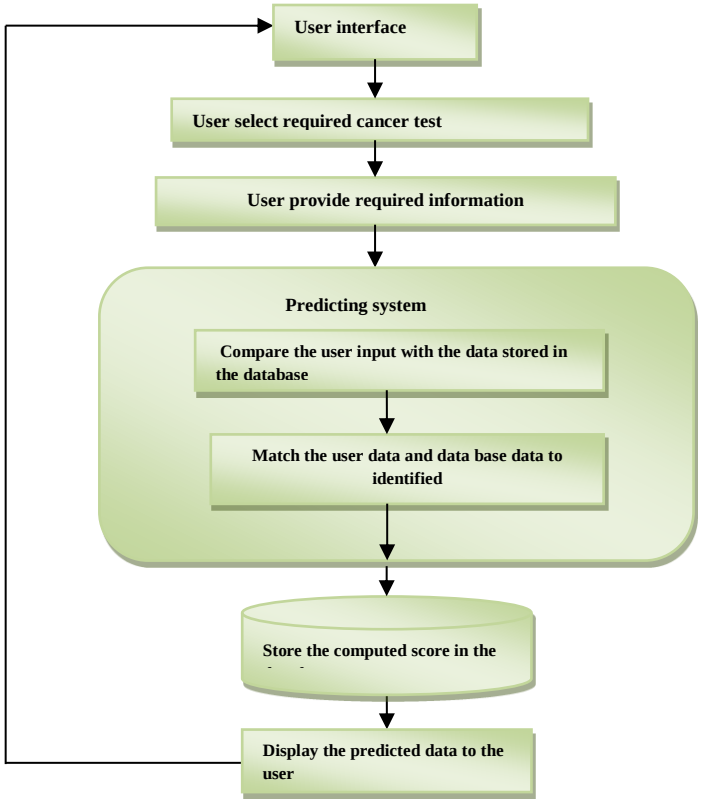

Figure.1 Architecture of Data mining Based Cancer Prediction System

It provides some basic symptoms of cancer which will help the user to consult the doctor for medical advice. The doctor will analyze the symptom and the treatment will be given in the early stage if it is predicted as cancer. When user enters into the cancer prediction test page, there will be a list of questions in the screen the user need to answer all the questions that is given in the list. Each question has some value. The value was given by the researchers after consultation with the doctors and previous research. Based on the answer provided by the user, the cancer prediction system will assign the value for each answer. The final value will be compared with the predefined risk value to assign cancer risk. Generally prediction system have four levels of risk like low level, medium level, high level, very high level. Once the risk is assigned the data given by the user is stored in the data base. The result will be shown to the user through the database.

Algorithm

Step 1: Enter the text

Step 2: Predicting system will checks for the condition.

Step 3: System predicts the values based on the user answers.

Step 5: The range of the risk is determined based on the predicted value.

Step 6: If the value is ≤18 the risk is considered as a low risk.

If the value is 18< risk value ≤ 21 the risk is considered as a intermediate risk

If the risk value is 21< risk value ≤23 is considered as a high risk.

If the risk value is > 28 is considered as a very high risk.

Step 6: The user data is stored in data base.

Step 7: The result is shown to the user through data base.

The implementation of Data mining based Cancer Prediction System

This work constructed an expert system called the cancer prediction system which predicts breast cancer risk. It helps the user to predict cancer risk level. It can save costs and time. It helps the user to predict their risk and take the necessary steps based on their risk status. This system was implemented using vb.net and sql.

This prediction system consists of various functional units listed below:

- Administrator
  - Report
- New user
- User Page
  - Prediction test
    - ✓ Breast cancer
- Feedback

PERFORMANCE EVALUATION OF CANCER PREDICTION SYSTEM

The data mining based cancer prediction system has been developed and implemented for predicting cancer. A study has been conducted to measure the effectiveness of Data Mining Based Cancer Prediction System among users.

The purpose of the study is twofold.

- Effectiveness of Data Mining Based Cancer Prediction System through feedback.
- Cancer prediction system through WEKA tool.

Cancer prediction system – population and sample

To find the effectiveness of data mining based cancer prediction system, this system has implemented on web. Around 496 responses have been collected during September to October 2013.Details of the responses given in the table 4.1

Gender	No of respondents
Male	**379**
Female	**117**
Total	**496**

Table.1. No of respondent based on Gender

Objectives

- To find out the performance of the Data Mining Based Cancer Prediction System among the users based on cancer prediction.
- To find the user opinion about the newly developed Data Mining Based Cancer Prediction System based on gender.

Data Analysis

Instrumentation

The feedback form was designed to find out the performance of the Data Mining Based Cancer Prediction System.

Questionnaire design

It was decided to prepare a questionnaire following the guideline given by Liket (1932). Considering variables under study, a scale was constructed and standardized by using psychometric techniques such as item analysis, reliability etc., and it was administered on the sample of the study. The researcher was very careful to phrase questions clearly and unambiguously, so that respondent is in no doubt which answer to give.

Procedure of data collection

The feedback form was provided in their prediction system software itself. The user filled the form after the completion of risk prediction.

Performance Analysis

The effectiveness of cancer prediction system is analyzed in two ways, one is getting feedback from the user after the completion of risk prediction using Data Mining Based Cancer Prediction System and another one is analysis of cancer prediction system through weka tool. We have used classification techniques (data mining technique) to know the efficiency of Data Mining Based Cancer Prediction System through weka tool. Classification is a technique that predicts categorical class labels. It classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data. Classification is a two – step process consisting of Model Construction and Model Usage. Model Construction is defined as a process of describing a set of predetermined classes whereas Model Usage is helpful for classifying future or unknown objects. We have used patient's prior medical data, healthy person data set as a training data set and data obtained from the cancer prediction system used as a test data. We have used two classification techniques decision tree and navie bayes to know the efficiency of Data Mining Based Cancer Prediction System. The experiments run on a smaller dataset.

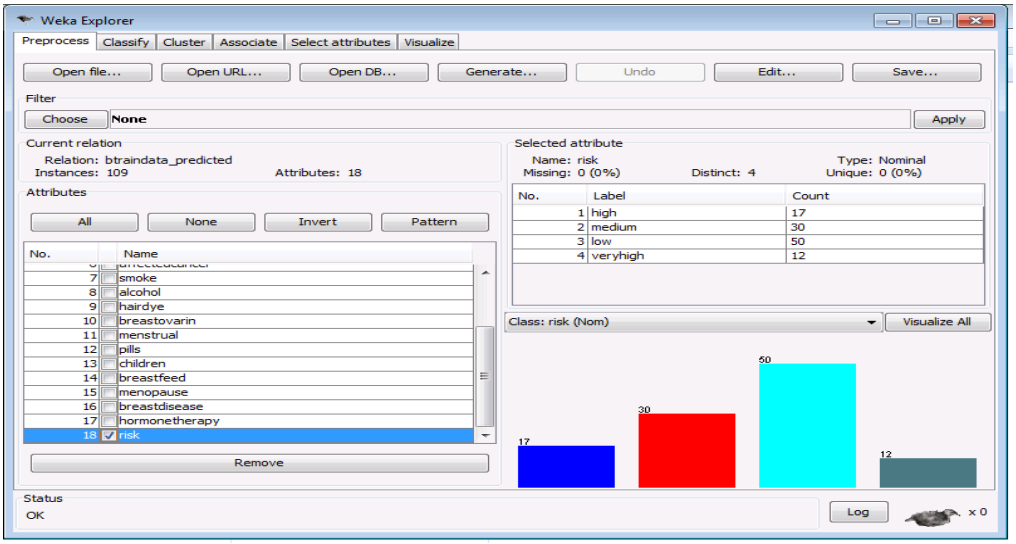

Fig.2. J48 risk prediction for breast cancer using WEKA

Experiments using WEKA for Breast Cancer

A. Decision Tree J48

The decision tree approach is more powerful for classification problems. There are two steps in this techniques building a tree & applying the tree to the dataset. The decision tree used in WEKA is termed as J 4.8 which is a modification of the C4.5 algorithm. J48 algorithm uses pruning method to build a tree. Pruning is a technique that reduces size of tree by removing over fitting data. The J48 algorithm recursively classifies data until it has been categorized as perfectly as possible. This technique gives maximum accuracy on training data. The overall concept is to build a tree that provides balance of flexibility & accuracy. In this we have used J48 algorithm to know the efficiency of prediction system. The front screen of the WEKA software is shown in the following figure. All the attributes in this database are displayed in row format in the left half of the screen and on the right side of the screen the bar graphs represent the distributions of the different attributes that are considered for data mining.

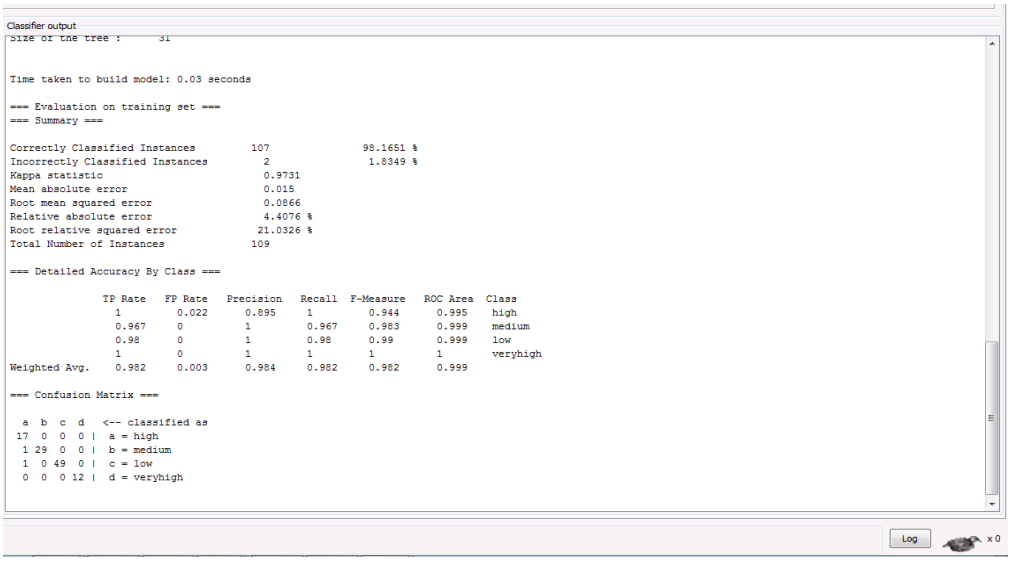

Fig.3. Result of Breast cancer prediction using J48 in WEKA

In the graph ash color bar represents the very high cancer risk, blue color represents high risk, red color represents the intermediate cancer risk, cyan represent the low risk. The decision tree to be created, rules are required to be extracted from the training data. Once the rules are extracted, the decision tree is created based on the rules and the association between the attributes. The decision tree with respect to breast cancer research is shown in the following figure. Classification on the test data is done based on the decision tree that is created. The confusion matrix is displayed in the classifier output screen as shown in the below fig4.4. A confusion matrix is a matrix showing the predicted and actual classifications. Suppose we have m attributes then the confusion matrix is of size m x m.

B. ID3

ID3 builds a decision tree from a fixed set of samples. The resulting tree is used to classify future dataset. The leaf nodes of the decision tree contain the class name whereas a non-leaf node is a decision node. In the graph (fig 4.5) ash color bar represents the very high cancer risk, blue color represents high risk, red color represents the intermediate cancer risk, cyan represent the low risk.

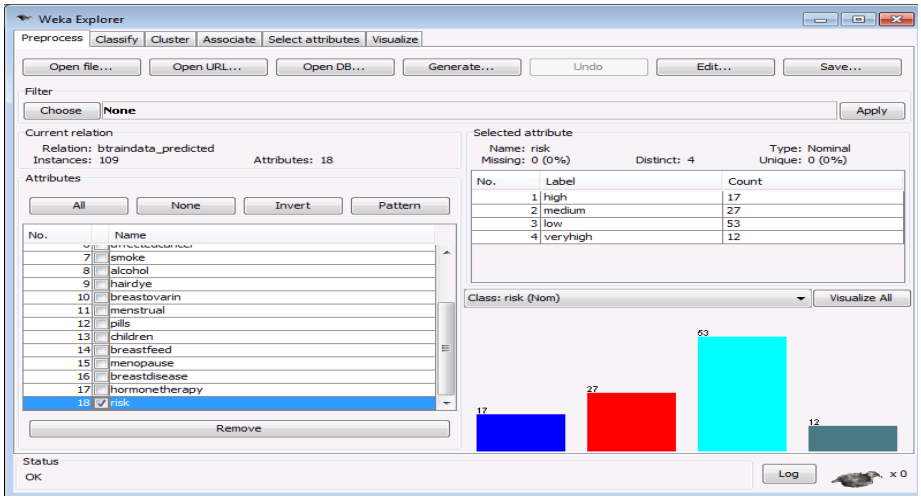

Figure 4. ID3 Risk Prediction for breast cancer using WEKA

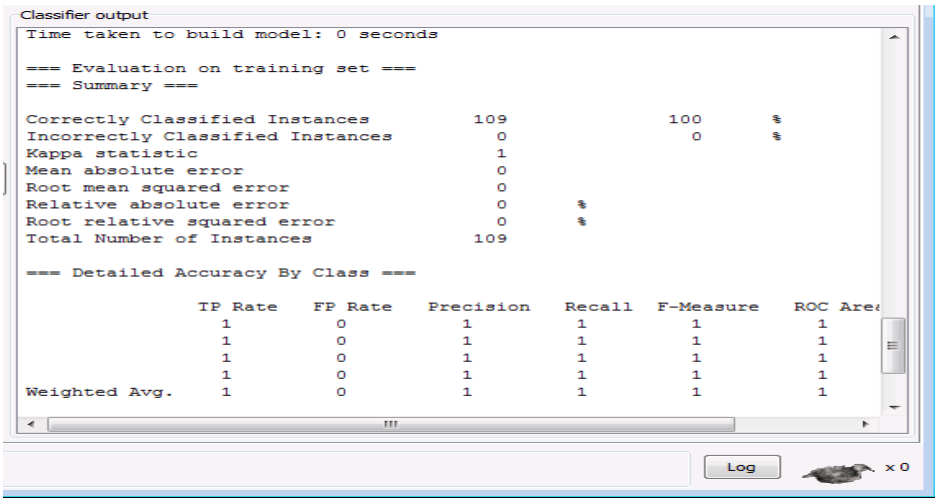

Fig 5. Result of Breast cancer prediction using ID3 in WEKA

B. Navie Bayes

Naïve Bayes is a statistical classifier which assumes no dependency between attributes. It attempts to maximize the posterior probability in determining the class. All the attributes in this database are displayed in row format in the left half of the screen and on the right side of the screen the bar graphs represent the distributions of the different attributes that are considered for data mining**.**

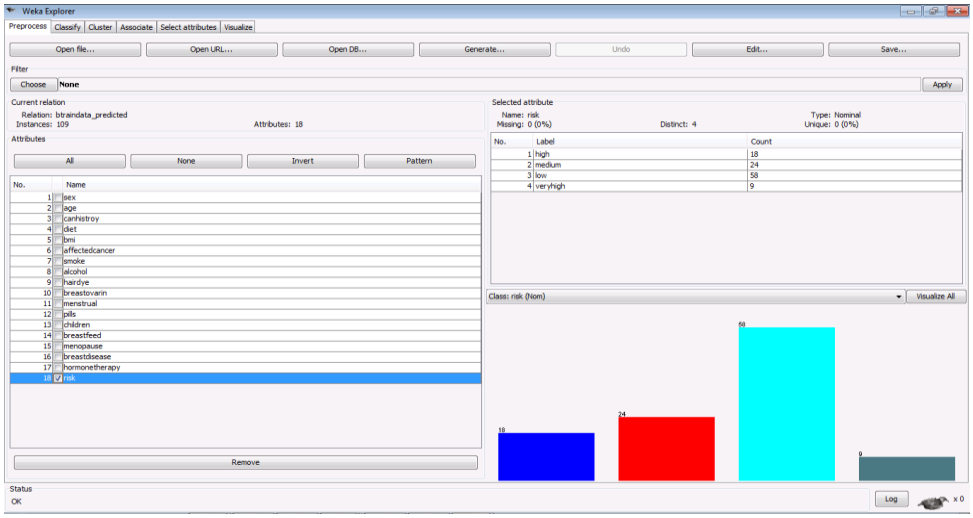

Fig 6. navie bayes risk prediction for breast cancer using WEKA

In the graph ash color bar represents the very high cancer risk, blue color represents high risk, red color represents the intermediate cancer risk, cyan represent the low risk.

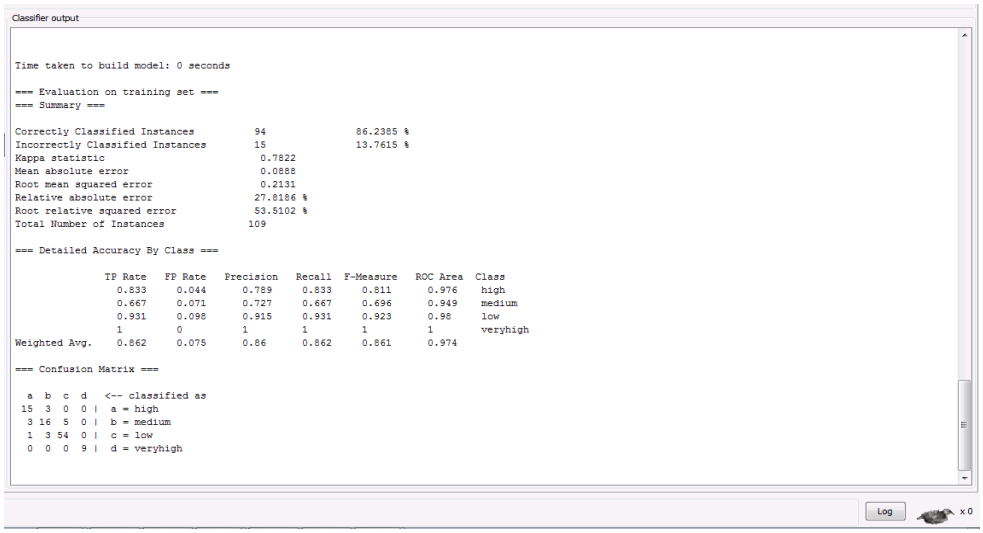

Fig 7. Result of Breast cancer prediction using Navie bayes in WEKA

The performance of the newly developed system is analyzed based on the feedback obtained from the users. We have used decision tree, navie bayes algorithms to find the effectiveness of DMBCPS through weka tool. The proposed method is efficiently calculating the risks of breast cancer. It helps the user to predict their risk and take the necessary steps based on their risk status. The results of this tests shows that ID3 algorithm provides better performance on DMBCPS.

CONCLUSION

In this work we have developed a system called data mining based cancer prediction system, which predicts three specific cancer risks. Specifically, Cancer prediction tool estimates the risk of the breast cancer by examining a number of user-provided genetic and non-genetic factors. An architecture of this data mining technique based prediction system, combining the prediction system with mining technology. In this model we have used one of the classification algorithms called decision tree. This tool is validated by comparing its predicted results with the patient's prior medical record, and also this is analyzed using weka tool. Once the user enters into the cancer prediction system, they need to answer the queries, related to genetic and non genetic factors. Then the prediction system assigns the risk value to each question based on the user responses. Once the risk value is predicted, the range of the risk can be determined by the prediction system. We have four levels of risk low level, intermediate, high level and very high level. Based on the predicted risk values, the range of risk will be assigned. The result can be shown to the user through data base. The above mentioned technique can be successfully applied to the data sets breast cancer as it was successfully verified on the breast cancer. Finally this prediction system is validated is through a weka tool, it provides the better accuracy compare to the existing system. The main aim of this model to provide the earlier warning to the users, and it is also cost and time benefit to the user.

REFERENCES

[1] N.Revathy, Dr.R.Amalraj(2011) Accurate Cancer Classification using Expression of Very few Genes. International Journal of Computer Application Volume 14 – No.4.

[2] Tasnuba Jesmin, Kaswar Ahmed, Md. Badrul Alam Miah (2013) Brain Cancer Risk Prediction System Using Data mining. International Journal of Computer Applications,Volume 61- No.

[3] Seyyid Ahmed Medjahed, Tamazouzt Ait Saadi, Abdelkader Benyettou (2013) Breast Cancer Diagnosis using K-Nearest Neibhor with Different Distances And Classification Rules. International Journal of Computer Applications, Volume 62-N0.1.

[4] Wafa Mokharrak, Nedhal Al Khalaf, Tom altman Application of Bioinformatics and Data mining in Cancer Prediction.

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

[6] Abdelghani Bellachia, Erhan Guven Predicting Breast Cancer Survivablity Using Data mining Techniques.

[7] S. Jothi, S.Anitha (2012) Data mining Classification Techniques Applied Fo Cancer Disease – A case Study Using Xlminer. International Journal of Engineering Research & Technolgy, Vol 1 Issue 8.

[8] V.Kroshnaiah, Dr.G.Narsimha, Dr.N.Subhash Chandra-2013 Diagnosis of Lung Cancer Prediction System Using Data mining Classification Techniques International Journal of Computer Science and Information Technologies, Vol 4, 39-45.

[9] Ada Ranjneet Kaur (2013) A Study of Lung Cancer Using Data mining Classification Techniques. International Journal of Advanced Research in Computer Science and Software Engineering, Vol 3, Issue 3.

[10] K.Rama Lakshmi and S.Prem Kumar (2013) Utilization of Data mining Techniques for Prediction and Diagnosis of Major Life Threatening Diseases Survivability- Review International Journal of Scientific & Engineering Research, Vol 4, Issue 6.

[11] Juliet R.Rajan, Jefrin J Prakash Early Diagnosis of Lung Cancer using a Mining System IJETTCS.

[12] E.Barati, M.Saraee, A.Mohammadi, N.Adibi and M.R.Ahamadzadeh (2011) A Survwy on Utilization of Data mining Approaches for Dermatological (Skin) Diseases Prediction Cyber Journals: Multidisciplinary Journals in Science and Technology, Journal of Selected Areas in Health Informatics.

[13] Charalampos Mavroforakis Data mining with WEKA a use-case to help you gets started.

[14] Jiawei Han and Micheline Kamber Data mining Concepts and Techniques,Second Edition.

[15] Calculate your risk.Australian Government . Available: http://canceraustralia.nbocc.org.au/risk/caculator.php 110-115.