Effectiveness of Data Mining - based Cancer Prediction System (DMBCPS)

A.Priyanga
M.Phil (CS) Research Scholar
SCSVMV University
Enathur, Kanchipuram

S.Prakasam, Ph.D
Assistant Professor
Department of CSA, SCSVMV University
Enathur, Kanchipuram

ABSTRACT
Cancer is one of the deadly diseases in the world today. Cancer is caused because of some genetic factors and/or environmental factors and/or today's modern lifestyle. Cancer has become the primary reason of death in developed countries. The most effective way to reduce cancer death is to detect it earlier. The earlier detection of cancer is not easier process but if it is detected, it is curable. Many works have been done in predicting cancer; different data mining approaches and algorithms were adopted by different people. Each work has some limitations such as lack of intelligent prediction, and inefficient in structure that motivated to take up this problem and to implement the Data mining based cancer prediction System (DMBCPS). We have proposed the cancer prediction system based on data mining. This system estimates the risk of the breast, skin, and lung cancers. 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.

General Terms
Data mining, Cancer

Keywords
Breast cancer, Lung Cancer, Skin Cancer, J48, Id3, Navie bayes

1. INTRODUCTION
Cancer is a potentially fatal disease caused mainly by environmental factors that mutate genes encoding critical cell-regulatory proteins. The resultant aberrant cell behavior leads to expansive masses of abnormal cells that destroy surrounding normal tissue and can spread to vital organs resulting in disseminated disease, commonly a harbinger of imminent patient death. More significantly, globalization of unhealthy lifestyles, particularly cigarette smoking and the adoption of many features of the modern Western diet (high fat, low fiber content) will increase cancer incidence. (S. Jothi et al.,) 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. In the following sections, previous research, related literatures are discussed.

We have proposed the cancer prediction system based on data mining. Cancer prediction system estimates the risk of the breast, skin, and lung cancers. This system was validated by comparing its predicted results with patient's prior medical information and analyzed using weka system.

2. PRIOR STUDIES OF CANCER PREDICTION
Ample of work have been done to predict the risk of the cancer. There are different techniques proposed by different authors for the detection of cancer risk. Each and every method has its own advantages and some disadvantages.

Cancer prediction is certainly a very complex and nondeterministic endeavor so many tests are available for cancer prediction but it's of high cost(Wafa Mokharrak et al.,) Table 1.1 shows the Comparative study of existing Data mining based cancer prediction methods

Table 1.1 Comparative study of existing Data mining based cancer prediction methods

Authors	Paper	Technique used	Results	Limitation
Seyyid Ahmed Medjahed, Tamazouzt Ait Saadi, Abdelkader Benyettou [2013]	Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules.	1.K-Nearest Neighbors, 2. Distance, Classification Rule.	K-NN method is used to diagnose 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. 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	K-NN algorithm works with Euclidean distance and Manhattan. It gives better performance but they may consuming much amount of time.

Authors	Paper	Technique used	Results	Limitation
			different values of the parameter "k" and by using several rules of classificatio n.	
S.Jothi, S.Anita [2012]	Data Mining Classificati on Technique s Applied For Cancer Disease - A Case Study Using Xlminer.	1. classificati on 2. XLminer	They Collected the data from the people according to their Symptoms, and then they built the prediction model based on the prior cancer data set. This technique can be successfully applied to the cancer such as Bone Cancer, Bladder Cancer, Stomach Cancer, Kidney Cancer, and Uterus cancer. They used XLMiner tool, it helps to predict the cancer accurately.	They were not used any pre processing method.
JulietR Rajan, Jefrin Prakash	Early Diagnosis of Lung Cancer using a Mining System	1.Artificial Intelligenc e 2.Data mining	In this paper they predict the lung cancer at an early stage thereby increasing the survival rate of the patient by five years. This system works efficiently in pre-diagnosing lung cancer based at Stage 1.	They didn't implement any system in online for predicting cancer
Abdelghani Bellaachia, Erhan Guven	Predicting Breast Cancer Survivabili ty Using Data Mining Technique s	1.Naïve Bayes, 2.Backpro pagated 3.Neural network, 4. C4.5 decision tree algorithm.	They have analyzed the prediction of survivabilit y rate of breast cancer patients using data mining techniques. They have used three data mining techniques: Naïve Bayes, back-propagated neural network, and C4.5 decision tree algorithm, out of which C4.5 gives better performanc e	In this approach they did not include any methodolog y to handle missing records.

3. ARCHITECTURE OF CANCER PREDICTION SYSTEM

In this work, an architecture data mining technique based cancer prediction system combining the prediction system with mining technology was used. In this model we have used one of the classification algorithms called decision tree.

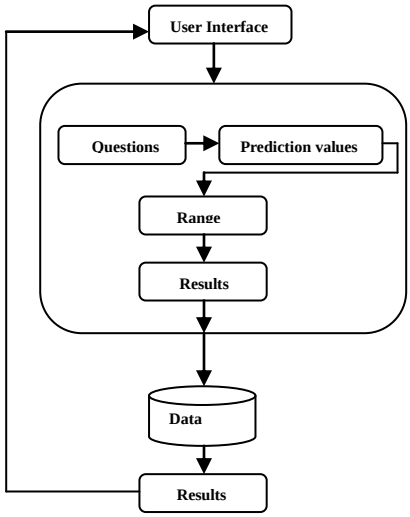


Fig 1.1 Architecture of Cancer prediction system

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.

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.

3.1 Implementing the Architecture of Cancer Prediction System

In this work we have constructed an expert system called the cancer prediction system which predicts three specific cancer (breast, lung, skin) risks*;* 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 prediction system consists of various functional units listed below:

- Administrator
  - ✓ Report
- New user
- User page
  - ✓ Prediction test
    - o Breast cancer
    - o Lung cancer
    - o Skin cancer
- Feedback

This prediction system was implemented by using VB.net and SQL.

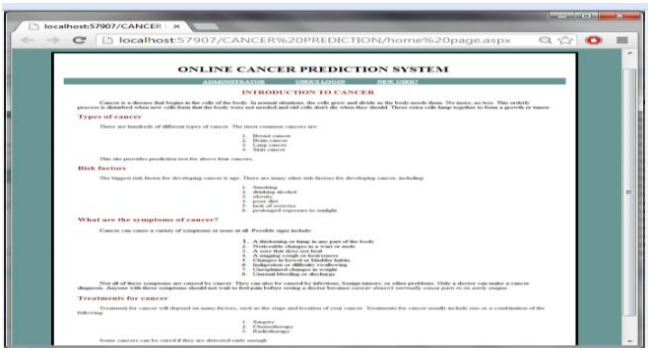


Fig 1.2 Cancer Prediction System

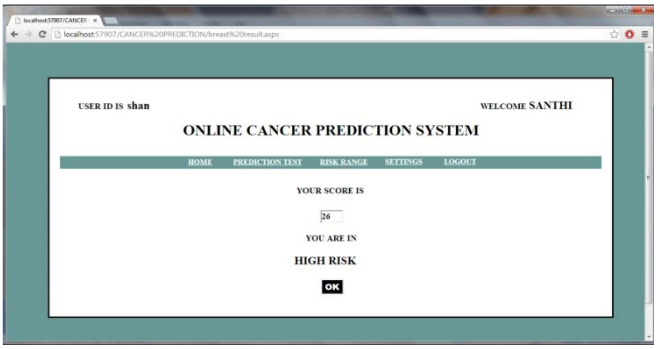


Fig 1.3 Predicting risk of cancer

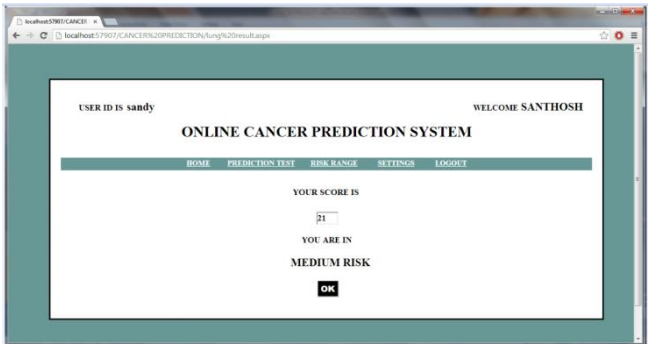


Fig 1.4 predicting risk of cancer

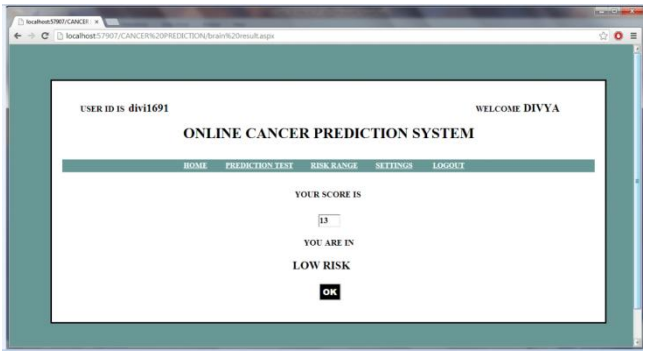


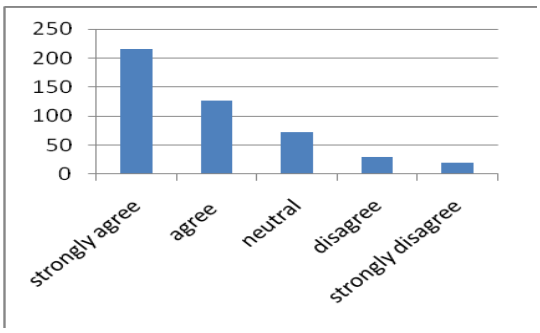
Fig 1.5 predicting risk of cancer

4. PERFORMANCE EVALUATION OF CANCER PREDICTION SYSTEM

The effectiveness of cancer prediction system is analyzed in two ways, one is cancer prediction system and another one is analysis of cancer prediction system through weka tool. Data mining based cancer prediction system is used to predict the cancer risks. This system helps the people to know their cancer risk with low cost and it also helps the people to take the appropriate decision based on their cancer risk status. 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 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. The result can be shown to the user through data base. The above mentioned technique can be successfully applied to the data sets for any cancer (such as breast cancer, lung cancer, skin cancer), as it was successfully verified on the breast, lung and skin cancer.

This system was implemented on the web during September and October 2013 people visited this site and we have got feedback from 463 people. The feedback table is shown below. This paper presents how the collected data were analyzed through weka system and the results of data analysis.

Table 1.1 Feedbacks from users



4.1 Analysis of Cancer Prediction system using Weka

WEKA, formally called Waikato Environment for Knowledge Learning, is a computer program. It supports many different standard data mining tasks such as data preprocessing, classification, clustering, regression, visualization and feature selection. The basic premise of the application is to utilize a computer application that can be trained to perform machine learning capabilities and derive useful information in the form of trends and patterns.

To find the effectiveness of Data mining-based Cancer prediction system author collected data from cancer institute make those data as trained data for cancer prediction system. Then after author collected the data from people during July 2013 to October 2013 and the data is trained through the cancer prediction system architecture.

Based on the previous work author selected 8 attributes as a general attributes and selected some specific attribute for specific cancers. The attributes are listed in the following table 1.2.

Table 1.2 General attributes

Attributes	Values
Sex	Male =2 Female =1
Age	Age ≤13 = 1 13< Age ≤24 = 2 24<Age ≤ 45 = 3 Age > 45 = 4
Affected family members	Yes = 2 No = 1
Consuming greens and vegetables	Yes = 1 No = 2
BMI	≤18 = 1 18 <BMI ≤ 24 = 2 ≥ 24 = 3
Affected by any cancer before	Yes = 3 No = 1
Are you a smoker	Yes = 3 No = 1
Are you consuming alcohol	Yes = 3 No = 1
Have you applying hair dye	Yes = 2 No = 1

Table 1.3 attributes of breast cancer

Attribute	Values
Multiple family members who have had breast, ovarian and/ or prostate cancer	Yes = 2 No = 1
Menstrual cycles starts before 12	Yes = 2 No = 1
Birth control pills	Yes = 2 No = 1
Gone through menopause	Yes before 55 = 3 Yes after 55 = 2 No = 2
Gone through menopause	Yes = 2 No = 1
Breast diseases	Yes = 3 No = 1
breast feed	Yes = 1 No = 2
Undergone estrogen and prostogen hormone therapy	Yes = 3 No =1

Table 1.4 Attributes of Lung cancer

Attributes	Values
Passive smoker	Yes = 2 No = 1
Lived in city	Yes = 2 No = 1
Working with chemicals	Yes without protection = 3 Yes with protection = 2 No = 1
Radiation therapy to chest area	Yes = 3 No = 1
Consumed tobacco	Yes = 3 No = 1
Have you suffered from any chronic diseases	Yes = 3 No = 1

Table 1.5 Attributes of Skin cancer

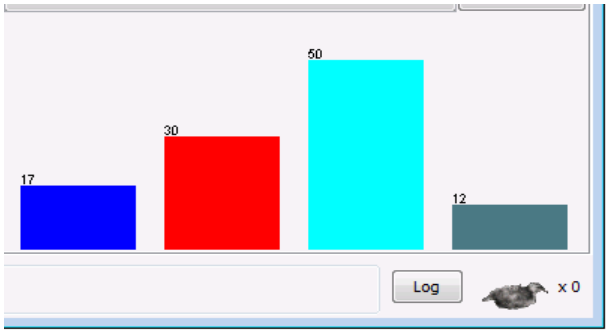
Attributes	Values
Outdoor activities	Frequent = 3 Medium = 2 No = 1
Color of skin	Black = 1 white =2
Working period in industries	More than 5 hours = 3 Less than 5 hours = 2 No = 1
Protect yourself from sun	Yes = 1 No = 2

4.2 Performance analysis of cancer prediction system

We have chosen two data mining techniques to find the effectiveness of cancer prediction system they are navie bayes, and decision tree.

4.2.1 Decision tree

The decision tree algorithm is one of the popular algorithm for classification problems. In decision tree, rules are extracted from the training dataset to form a tree structure, and this rule will be applied to the classification of testing data. There are many popular decision tree algorithms CART, J48, ID3, C4.5, and CHAID. In this paper we have chosen J48 and ID3 for performance analysis. The J48 algorithm recursively classifies data until it has been classified as perfectly. This technique gives maximum accuracy on training data. The experiments run on a smaller dataset.

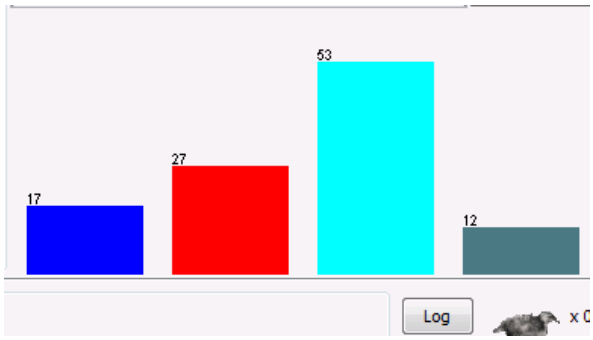


Graph 1.1 Results of breast cancer using J48

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.

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.



Graph 1.2 Results of breast cancer using ID3

Navie bayes

Navie Bayes model is a simple and well known method for performing supervised learning of a classification problem. The Navie Bayesian Classifier make the assumption of class conditional independence, i.e, given the class label of a tuple, the values of the attributes are assumed to be conditionally independent of one another.

Graph 1.3 Results of breast cancer using Navie bayes

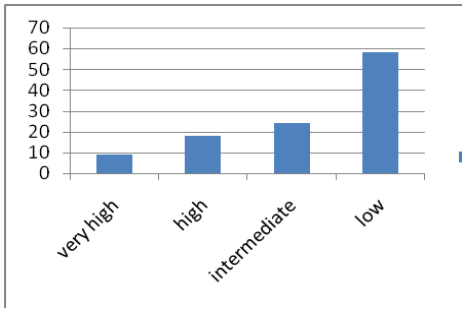
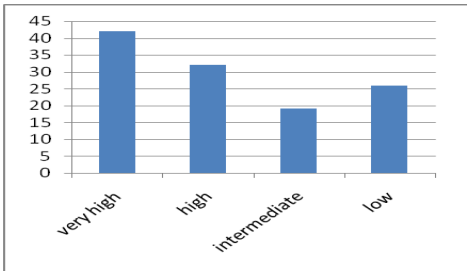


Table 1.6 Accuracy table for breast cancer

weka	J48	ID3	navie bayes
	98.16%	100%	86.23%

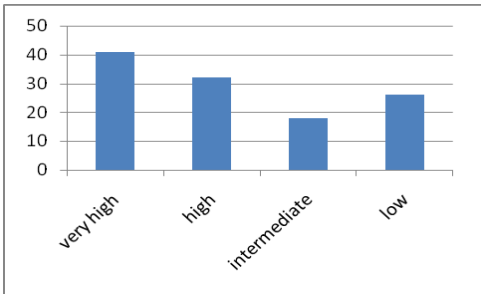
ID3 algorithm provides highest accuracy in breast cancer prediction.

J48



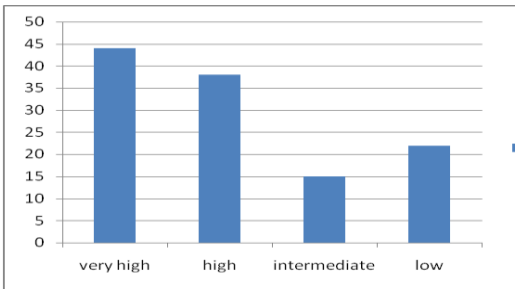
Graph 1.4 Results of lung cancer using J48

ID3



Graph 1.5 Results of lung cancer using Id3

Navie bayes



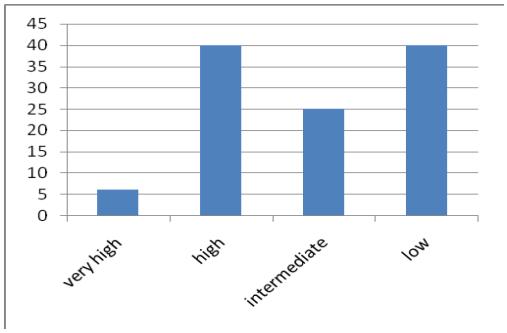
Graph 1.6 Results of lung cancer using Navie Bayes

Table 1.7 accuracy prediction for lung cancer

weka	J48	ID3	navie bayes
	98.31%	100%	89.03%

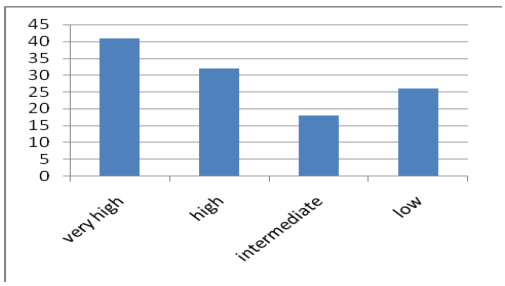
ID3 algorithm provides highest accuracy in lung cancer prediction.

J48



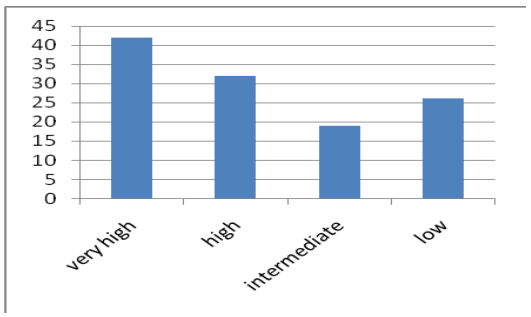
Graph 1.7 Results of skin cancer using J48

ID3



Graph 1.8 Results of skin cancer using Id3

Navie bayes



Graph 1.9 Results of skin cancer using Navie Bayes

Table 1.8 Accracy table for skin cancer

weka	J48	ID3	navie bayes
	80%	100%	78.3%

ID3 algorithm provides highest accuracy in skin cancer prediction.

5. CONCLUSION

Cancer is potentially fatal disease. 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.

Detection of cancer in earlier stage is curable. In this work we have developed a system called data mining based cancer prediction system. The main aim of this model is to provide the earlier warning to the users, and it is also cost and time benefit to the user. It predicts three specific cancer risks. Specifically, Cancer prediction system estimates the risk of

the breast, skin, and lung cancers by examining a number of user-provided genetic and non-genetic factors. This system is validated by comparing its predicted results with the patient's prior medical record, and also this is analyzed using weka system. This prediction system is available in online, people can easily check their risk and take appropriate action based on their risk status. This system performs well than the existing system.

6. 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(1), 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 [Oneline]. Available: http://canceraustralia.nbocc.org.au/risk/caculator.php 110-115.