International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 02 Issue: 08 | Nov-2015 www.irjet.net p-ISSN: 2395-0072

# CANCER PREDICTION SYSTEM USING DATAMINING TECHNIQUES

**K.Arutchelvan[1], Dr.R.Periyasamy[2]**

[1] *Programmer (SS), Department of Pharmacy, Annamalai University, Tamilnadu, India*
[2] *Associate Professor, Department of Computer Science, Nehru Memorial College, Tamilnadu, India*

---

**Abstract -** **Cancer is one of the major problem today, diagnosing cancer in earlier stage is still challenging for doctors. Identification of genetic and environmental factors is very important in developing novel methods to detect and prevent cancer. Therefore a novel multi layered method combining clustering and decision tree technique is used to build a cancer risk prediction system. The proposed system is predicts lung, breast, oral, cervix, stomach and blood cancers and it is user friendly and cost saving. This research uses data mining techniques such as classification, clustering and prediction to identify potential cancer patients. We have proposed this cancer prediction system based on data mining techniques. 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. The main aim of this model is to provide the earlier warning to the users and it is also cost efficient to the user. Finally a prediction system is developed to analyze risk levels which help in prognosis. This research helps in detection of a person's predisposition for cancer before going for clinical and lab tests which is cost and time consuming**.

***Key Words:*** Cancer, Data Mining, Clustering, Classification, Decision Tree.

## 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.

Data mining technique involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data set. These tools can include statistical models, mathematical algorithm and machine learning methods in early detection of cancer. In classification learning, the learning scheme is presented with a set of classified examples from which it is expected to learn a way of classifying unseen examples. In association learning, any association among features is sought, not just ones that predict a particular class value. In clustering, groups of examples that belong together are sought. In numeric prediction, the outcome to be predicted is not a discrete class but a numeric quantity. In this study, to classify the data and to mine frequent patterns in data set Decision Tree algorithm is used.

Data Mining techniques are implemented together to create a novel method to diagnose the existence of cancer for a particular patient. When beginning to work on a data mining problem, it is first necessary to bring all the data together into a set of instances. Integrating data from different sources usually presents many challenges. The data must be assembled, integrated, and cleaned up. Then only it can be used for processing through machine learning techniques. This developed system can be used by physicians and patients alike to easily know a person's cancer status and severity without screening them for testing cancer. Also it is useful to record and save large volumes of sensitive information which can be used to gain knowledge about the disease and its treatment.

## 2. REVIEW OF LITERATURE

Ritu Chauhan et al [1] focuses on clustering algorithm such as HAC and K-Means in which, HAC is applied on K-means to determine the number of clusters. The quality of cluster is improved, if HAC is applied on K-means.

Dechang Chen et al [2] algorithm EACCD developed which a two step clustering method. In the first step, a dissimilarity measure is learnt by using PAM, and in the second step, the learnt dissimilarity is used with a hierarchical clustering algorithm to obtain clusters of patients. These clusters of patients form a basis of a prognostic system.

S M Halawani et al [3] suggests that probabilistic clustering algorithms performed well than hierarchical clustering algorithms in which almost all data points were clustered into one cluster, may be due to inappropriate choice of distance measure.

Ada et al [4] made an attempt to detect the lung tumors from the cancer images and supportive tool is developed to check the normal and abnormal lungs and to predict survival rate and years of an abnormal patient so that cancer patients lives can be saved.

V.Krishnaiah et al [5] developed a prototype lung cancer disease prediction system using data mining classification techniques. The most effective model to predict patients with Lung cancer disease appears to be Naïve Bayes followed by IF-THEN rule, Decision Trees and Neural Network. For Diagnosis of Lung Cancer Disease Naïve Bayes observes better results and fared better than Decision Trees.

Charles Edeki et al [6] Suggests that none of the data mining and statistical learning algorithms applied to breast cancer, dataset outperformed the others in such way that it could be declared the optimal algorithm and none of the algorithm performed poorly as to be eliminated from future prediction model in breast cancer survivability tasks.

Zakaria Suliman zubi et al [7] used some data mining techniques such as neural networks for detection and classification of lung cancers in X-ray chest films to classify problems aiming at identifying the characteristics that indicate the group to which each case belongs.

Labeed K Abdulgafoor et al [8] wavelet transformation and K- means clustering algorithm have been used for intensity based segmentation.

Sahar A. Mokhtar et al [9] have analyzed three different classification models for the prediction of the severity of breast masses namely the decision tree, artificial neural network and support vector machine.

Rajashree Dash et al [10] a hybridized K-means algorithm has been proposed which combines the steps of dimensionality reduction through PCA, a novel initialization approach of cluster centers and the steps of assigning data points to appropriate clusters.

## 3. PROPOSED SYSTEM

**I .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.

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. It has four levels of risk like low level, intermediate level, high level and very high level. Based on the predicted risk values the range of risk will be assigned.

**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 4: The range of the risk is determined based on the predicted value.

Step 5: If the value is ≤18 the risk is considered as a low risk.

If the value is >18 and ≤ 21 the risk is considered as an intermediate risk.

If the value is >21 and ≤28 is considered as a high risk.

If the 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 obtained with the reference values of the data base.

**II. Rules for Decision Tree**

A decision tree is a flow chart like tree structure, where each internal node denotes a test on an attribute,

each branch represents an outcome of the test and each leaf node holds a class label.

The top most node is the root node. The attribute value of the data is tested against a decision tree. A path is traced from root to leaf node, which holds the class prediction for that data. Decision trees can be easily converted into classification rules. This decision tree is used to generate frequent patterns in the dataset.

The data and item sets that occur frequently in the data base are known as frequent patterns. The frequent patterns that is most significantly related to specific cancer types and are helpful in predicting the cancer and its type is known as Significant frequent pattern.

Using this significant patterns generated by decision tree the data set is clustered accordingly and risk scores are given.

If *symptoms* = none and *risk score x* = <35 then *result* = you may not have cancer, *tests* = do simple clinical tests to confirm.

If *symptom*= related to chest and shoulder and *risk score x* >= 40 then *result* = you may have cancer, *cancer type may be*= chest, tests = take CT scan of chest.

If *symptom*= related to head and throat and *risk score x* >= 40 then *result* = you may have cancer, *cancer type* = oral, *tests* = biopsy of tongue and inner mouth.

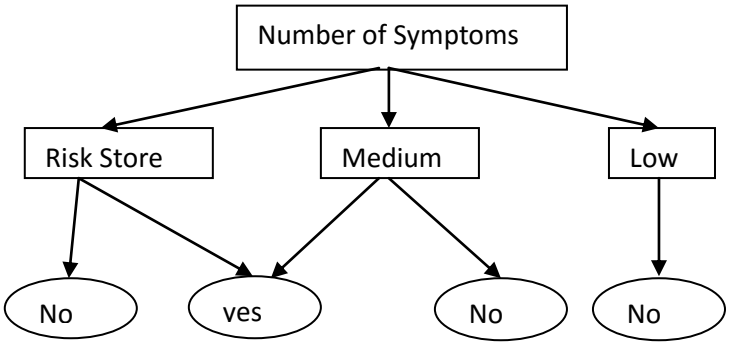
Else *symptom*= other symptoms and *risk score x* >= 40 then *result* = you may have cancer, *cancer type* = leukemia, *tests* = biopsy of bone marrow.

Else if *symptom*= related to stomach and *risk score x* >= 45 then *result* = you may have cancer, *cancer type* = stomach, *tests* = endoscopy of stomach

If *symptom*= related to breast and shoulder and *risk score x* >= 45 then *result* = you may have cancer, *cancer type* = breast, *tests*= mammogram and PET scan of breast

If *symptom*= related to pelvis and lower hip and *risk score x* >=55 then *result* = you may have cancer, *cancer type* = cervix, *tests* = do pap smear test

Based on the above mentioned rules and the calculated risk scores the severity of cancer is known as well as some tests were prescribed to confirm the presence of cancer.




## 4. 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 saving 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. The performance of the system is better than the existing system.

## 5. REFERENCES -

[1] Ritu Chauhan "Data clustering method for Discovering clusters in spatial cancer databases" International Journal of Computer Applications (0975-8887) Volume 10-No.6, November 2010.

[2] Dechang Chen "Developing Prognostic Systems of Cancer Patients by Ensemble Clustering" Hindawi publishing corporation, Journal of Biomedicine and Biotechnology Volume 2009, Article Id 632786.

[3] S M Halawani "A study of digital mammograms by using clustering algorithms" Journal of Scientific & Industrial Research Vol. 71, September 2012, pp. 594-600.

[4] Ada and Rajneet Kaur "Using Some Data Mining Techniques to Predict the Survival Year of Lung Cancer Patient" International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 2, Issue. 4, April 2013, pg.1 – 6, ISSN 2320–088X

[5] V.Krishnaiah "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques" International Journal of Computer Science and Information Technologies, Vol. 4 (1) 2013, 39 – 45 www.ijcsit.Com ISSN: 0975-9646

[6] Charles Edeki "Comparative Study of Data Mining and Statistical Learning Techniques for Prediction of Cancer Survivability" Mediterranean journal of Social Sciences Vol 3 (14) November 2012, ISSN: 2039-9340.

[7] Zakaria Suliman zubi "Improves Treatment Programs of Lung Cancer using Data Mining Techniques" Journal of Software Engineering and Applications, February 2014, 7, 69-77

[8] Labeed K Abdulgafoor "Detection of Brain Tumor using Modified K-Means Algorithm and SVM" International Journal of Computer Applications (0975 – 8887) National Conference on Recent Trends in Computer Applications NCRTCA 2013

[9] A. Sahar "Predicting the Serverity of Breast Masses with Data Mining Methods" International Journal of Computer Science Issues, Vol. 10, Issues 2, No 2, March 2013 ISSN (Print):1694-0814| ISSN (Online):1694-0784 www.IJCSI.org

[10] Rajashree Dash "A hybridized K-means clustering approach for high dimensional dataset" International Journal of Engineering, Science and Technology Vol. 2, No. 2, 2010, pp. 59-66.

[11] Alaa. M. Elsayad "Diagnosis of Breast Cancer using Decision Tree Models and SVM" International Journal of Computer Applications (0975 – 8887) Volume 83 – No 5, December 2013.

[12] Neelamadhab Padhy "The Survey of Data Mining Applications and Feature Scope" Asian Journal of Computer Science and Information Technology 2:4(2012) 68-77 ISSN 2249-5126.

[13] S. Santhosh Kumar "Development of an Efficient Clustering Technique for Colon Dataset" International Journal of Engineering and Innovative Technology" Volume 1, Issue 5, May 2012 ISSN: 2277-3754.

[14] Rafaqat Alam Khan "Classification and Regression Analysis of the Prognostic Breast Cancer using Generation Optimizing Algorithms" International Journal of Computer Applications (0975-8887) Volume 68- No.25, April 2013

[15] K.Kalaivani "Childhood Cancer-a Hospital based study using Decision Tree Techniques" Journal of Computer Science 7(12): 1819-1823, 2011 ISSN: 1549-3636

[16] Boris Milovic "Prediction and Decision Making in Health Care using Data Mining" International Journal of Public Health Science Vol. 1, No. 2, December 2012, pp. 69-78 ISSN: 2252-8806

[17] T.Revathi "A Survey on Data Mining Using Clustering Techniques" International Journal of Scientific & Engineering Research Http://Www.Ijser.Org, Volume 4, Issue 1, January-2013, Issn 2229-5518

[18] Shomona Gracia Jacob "Data Mining in Clinical Data Sets: A. Review" International Journals of Applied Information System (IJAIS) - ISSN: 2249-0868 Foundation of Computer Science FCS, New York, USA, Volume 4-No.6, December 2012-www.ijais.org

[19] G. Rajkumar " Intelligent Pattern Mining and Data Clustering for Pattern Cluster Analysis using Cancer Data" International journal of Engineering Science and Technology Vol. 2(12), 2010, ISSN: 7459-7469

[20] M. Durairaj "Data Mining Applications in Healthcare Sector: A Study" International journal of Scientific & Technology Research, Volume 2, Issue 10, October 2013, ISSN: 2277-8616

[21] Vikas Chaurasia "Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability" International journal of Computer Science and Mobile Computing (IJCSMC), Vol.3, Issue. 1, January 2014, pg.10-22, ISSN: 2320-088X

[22] T.Sridevi "An Intelligent Classifier for Breast Cancer Diagnosis based on K-Means Clustering and Rough Set" International Journal of Computer Applications (0975 – 8887) Volume 85 – No 11, January 2014

[23] Reeti Yadav "Chemotheraphy Prediction of Cancer Patient by Using Data Mining Techniques" International

Journal of Computer Applications (0975-8887), Volume 76-No.10, August 2013

**BIOGRAPHIES**

ARUTCHELVAN.K Working as a Programmer (SS) in the Department of Pharmacy, Annamalai University for the past 13 years and having efficient knowledge in Data mining as well as in Cancer.

Dr.R.Periyasamy Working as a Associate Professor in the Department of computer science, Nehru Memorial College for the past 28 years and having efficient knowledge in Data mining as well as in Cancer.