IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

CANCER PREDICTION USING DATAMINING TECHNIQUES

Ayush Kalra[1], Vineet[2], Vishal Malik[3] , Bindu Garg[4]

[1,2,3]**Student, Department of Computer Engineering**

(AICT, NBA, NAAC A+ grade)

BharatiVidyapeeth(Deemed to be) University, College Of Engineering, Pune

(AICT, NAAC A+ grade)

Pune, India

ABSTRACT

Cancer is a major cause of death. An estimated 1,735,350 new cases of cancer were diagnosed in US alone in the year 2018 and 609,640 people died due to the illness. There are various kinds of cancer like melanoma of skin, lung bronchus cancer, breast cancer, prostate cancer, colon and rectum cancer, bladder cancer, kidney and renal pelvis cancer etc. With these many different types of cancer present and large number of affected population, cancer has gained popularity in the research field. Till date, novel methods for prevention and diagnosis of cancer are being continuously researched on. We have aimed to develop a robust and viable system for cancer detection using data mining techniques. This Cancer Disease Prediction application is an end user support and online consultation project. It makes use of multi-layer data mining techniques such as clustering and decision tree technique to build a novel application. The proposed model, is an intelligent online system that relies on the various details of cancer disease which are pre-fed to the application. Based on these details an instant guidance is offered on the progression of the disease. The users are allowed to input their specific health related information, obtaining the data, the application uses pre-fed information to predict the potential cause of the symptoms using data mining techniques. Thus, the developed prediction system warns the users for potential diseases based on their current medical information and is also cost-effective, hence, offering dual benefits.

Keywords- *Cancer, data mining techniques, clustering, decision tree technique*.

I.INTRODUCTION

Estimations suggest that every 6th death in the world is due to cancer. Various forms of cancer has killed over 9.6 million people in the year 2017. Cancer is the second major reason of death falling behind cardiovascular disease such as myocardium infraction. Reports by NCRP, ICMR, Bangalore, India, suggest that almost 1300 people die of cancer every day in India. The mortality rate of cancer is very high in India and a rise of 6% was observed from year 2012 to 2014. Out of 2,820,179 patients

491,598 people died of the disease. One of the leading risk factors of cancer is age. 87% of deaths occurred in patients that were above 50 years of age. Almost 46% of the accounted deaths from cancer in 2017 were reported in patients of age 70 years or above. Another 41% constituted of patients in the age group of 50-69 years. Other risk factors of cancer include Tobacco use, alcohol use, lower or higher body mass index, unhealthy diet, poor workout schedule, other chronic infections such as Hepatitis B and C etc.

Cancer is caused by the loss of a property called “contact inhibition” in normal cells. The loss of this property result in abnormal cell growth which further results in the deprivation of essential nutrients and oxygen for normal cells. The abnormal growth results in the formation of a mass of cells called tumor. These tumor cells can be malignant or benign in nature. Malignancy is a property of cancer cells which allows them to migrate to other parts of bodies through blood and other sources to form tumor cells elsewhere, thus resulting in other type of cancer. There are almost 200 different types of cancer depending on the part in which their growth occurs. The most common types of cancers, with number of cases and deaths are provided in the table below.

TYPES OF CANCER	NUMBER OF CASES	DEATHS
Lung Cancer	2.09 million	1.76 million
Breast Cancer	2.09 million	627,000
Colorectal Cancer	1.80 million	862,000
Prostate Cancer	1.28 million	358,989
Skin Cancer (non-melanoma)	1.04 million	(not estimated)
Stomach Cancer	1.03 million	783,000

Table 1- Types of cancer, number of cases and deaths

Causes of cancer are genetic and environmental. The environmental factors can be divided into three classes defined in the table below.

S. No.	Types	Examples
1	Physical carcinogens	ionizing radiations, ultraviolet radiations etc.
2	Chemical carcinogens	asbestos, components of tobacco smoke, aflatoxin (a food contaminant), and arsenic (a drinking water contaminant)
3	Biological carcinogens	Infections from certain viruses, parasites, bacteria or fungus.

Table 2- Environmental factors and their examples

Patterns and relationships between large dataset is determined through many sophisticated data mining techniques. These tools are essentially statistical, mathematical or machine learning based models that help in the prediction of the disease such as cancer. It is a multi-stage process that are used in a strict combination to study giant database to predict cancer. The first step is to assemble the data followed by integration of data from different sources. After the data has been successfully assembled and integrated, it needs to be thoroughly cleaned up to only include relevant data and clean up the junk. After these steps, working on data mining problem can be started. Thus, assembly, integration and clean up are three important introductory steps of taking up a data mining problem.

The said application can be used by patients and medical professionals alike to check the status of cancer in patients. In this project, classification, clustering and decision tree algorithm are mainly used to build up the application for the successful prediction.

II. LITERATURE SURVEY

In one of the models, dimensionality reduction steps are combined through PCA. This is done with the help of hybridized K-means algorithm. This approach successfully approaches various cluster centers and assigns data points. *(Rajashree Dash, et al.)*In dimensionality reduction, the number of random variables under consideration are reduced by acquiring a collection of key variables is in math, machine learning and knowledge theory. Approaches can be broken down into collection of features and extraction of features. The data required for a function or a problem that consist of many attributes or parameters, but not all of these attributes may influence performance equally. A large number of attributes, or features, that affect the complexity of the computation and may even result in over-fitting that results in poor results.

Data is constantly being generated in large amounts. However, the data available is not minded. Hence, One Dependency Augmented Naïve Bayes classifier (ODANB) and naive creedal classifier 2 (NCC2) are used to process the data. These systems deliver dependable classifications with data even if the data is small and incomplete. These are essentially an extension of the regular Naïve Bayes classifier with additional benefits. Using these data mining techniques, lung cancer has been predicted in patients by forming formerly unknown patterns and relationships. *(VasundaraKrishnaiah, et al.)*

In yet another paper on breast cancer, the author suggests that data mining and statistical learning algorithms could not be successfully applied and the dataset could not be dealt with any optimal algorithm. Yet, none of these models were so poor to be declared as incompatible and to be eliminated from the prediction system. *(Charles Edeki et al.)*

For breast cancer prediction, a combination of FP growth algorithm (Frequent pattern mining) and Decision tree algorithm. Decision Tree algorithm is majorly used to predict whether the tumor is malignant or benign.*(JaiminiMajali et al.)*

Another research on breast cancer used three different data mining techniques for determining the severity of breast cancer. The model uses decision tree algorithm, support vector machine and artificial neural network. *(Sahar A. Mokhtar)*

Supportive tools are used along with cancer images to assess lung tumors and to even predict survival and mortality rate of the patients. *(Ada et al.)*

Using clustering algorithm to determine the number of clusters. The said project used two types of cluster algorithm for improvisation of the clusters. Two types of cluster algorithm used are:

- HAC
- K-means *(Ritu Chauhan, et al.)*

K-means cluster algorithm is also used in combination with wavelet transformation for intensity based segmentation by another project. *(Labeed K Abdulgafoor et al.)*

Another project utilizes clustering algorithms and drew a comparison between clustering algorithms and hierarchical clustering algorithms. Clustering into one cluster point of all data points was the drawback of the later. *(S M Haawani et al.)*

III. PROPOSED SYSTEM

This project's prime aim was to build architecture of a cancer prediction system that makes use of data mining techniques to predict cancer and send a warning to patients. This system is designed to be compatible for patient's use. The data mining technique used in the said project for classification is Decision tree algorithm.The application utilizes different learning methods to create an impactful, error free and reliable platform for prediction. The different methods exploited are:

- Classification learning- This learning method is used to classify unseen, unpresented and unrecorded cases from a set of defined and recorded examples. Thus, various examples are employed which are pre classified to learn to classify into reliable sets.
- Association learning- In association learning method, an association or relation is established between pre given values and new values that are
  provided. Hence, relation establishment is sought out in the data sets.
- Cluster learning- As the name suggest, in cluster learning, entries that share maximum features with each other are clustered together into one class.
- Numeric prediction- In this, outcome comes as a numeric value instead of a discrete class.

3.1 Overview: The mechanism of working of this project is based on the genetic and non-genetic factors that affect the onset and growth of cancer. The values of these factors are inputted by the users. The responses are taken in and measured and then risk values are assigned to each of the responses. Risk values represent four stages i.e. Low risk values, Intermediate risk values, High risk values and Very high risk values. Based on the risk value, risk ranges are further assigned by the system.

3.2 Decision Tree- Decision tree is one among the many predictive modeling approaches. It is actively used in data mining, statistics and machine learning. It derives conclusion about the entry by looking at its observations which are represented as the branches of a tree, thus the name.
Representation in decision tree algorithm:

1.	Internal node	Attribute
2.	Branch	Outcome of test
3.	Leaf node	Class label
4.	Top most node	Root node

Table 3- Representations

The attributes or the topmost values are tested in the decision tree. The class prediction is done from across the path from root to leaf node. Thus patterns are generated using decision tree. Frequent patterns are created from the most frequent readings, these further help in the prediction of cancer by acting as reference. The types are commonly known as Significant frequent patterns. The clustering is, henceforth, done and risk factors are assigned.

1	symptoms = none and risk score x = <35	Do simple clinical tests to confirm.
2	symptom= related to chest and shoulder and risk score x >= 40	You may have cancer, cancer type may be- chest, tests - take CT scan of chest.
3	symptom= related to head and throat and risk score x >= 40	You may have cancer, cancer type - leukemia, tests - biopsy of bone marrow.
4	symptom= other symptoms and risk score x >= 40	You may have cancer, cancer type - leukemia, tests - biopsy of bone marrow.
5	symptom= related to stomach and risk score x >= 45	You may have cancer, cancer type - stomach, tests - endoscopy of stomach
6	symptom= related to breast and shoulder and risk score x >= 45	You may have cancer, cancer type - breast, tests - mammogram and PET scan of breast
7	symptom= related to pelvis and lower hip and risk score x >=55	You may have cancer, cancer type - cervix, tests - do pap smear test

Table 4- Values and their indication

3.3 Algorithm- The algorithm of the application can be successfully described in the following steps.

Step 1- Users answer the set of questions presented to them
Step 2- The conditions are checked thoroughly by the prediction system
Step 3- Values are assigned by the system to the answers.
Step 4- Risk values are assigned to each input.
Step 5- Risk values based on the numeric values are assigned as follows

S. No.	RISK SET	VALUES
1	Low risk	Lower than or equal to 18
2	Intermediate risk	Greater than 18 and lower than or equal to 21
3	High risk	Greater than 21 and lower than or equal to 28
4	Very high risk	Greater than 28

Table 5- Risk values classification

Step 6- A database is created to store the data Step 7- Results are displayed.

IV. CONCLUSION

Detection of cancer in earlier stages can give a better survival chances to patients. Often, the diagnosis of cancer takes too much time and when it's diagnosed it's too late. It is a fatal disease that requires immediate medical attention. Even though, good care is given, the chances of survival are not high because of unavailability of proper treatment. Cancer prediction system is absolutely essential and more focus on it should be diverted. Further improvements can be done in the system by making use of more data mining techniques.

V. REFERENCES

1. Max Roser and Hannah Ritchie, "CANCER"; Our World In Data; 2020
2. Plummer M, Vignat J, Ferlay J, Franceschi S, de Martel C, Bray F.
3. International Agency for Research on Cancer, 2014, Stewart BW, Wild CP
4. Global Initiative for Cancer Registry Development. International Agency for Research on Cancer
5. VasundaraKrishnaiah, GugulothuNarsimha, NaithaniSubhash Chandra; "Diagnosis of lung cancer prediction using data mining techniques and classification"; 2013.
6. JaiminiMajali, R. K. Niranjan; "Data mining techniques for diagnosis and prognosis of breast cancer"; 2014
7. Ritu Chauhan; "Data clustering method for discovering clusters in spatial cancer database"; International Journal of computer application; 2010; Volume 10
8. Dechang Chen, Hindawi publishing co.; "Developing prognostic systems of cancer patients by Ensemble Clustering"; Journalof biomedicine and Biotechnology; 2009
9. SM Halwani; " A study of digital mammograms by using clustering algorithms"; Journal of Scientific & industrial research; 2012; Volume 71
10. Charles Edeki; "Comparitive study of data mining and statistical learning techniques for prediction of cancer survivability"; Mediterranean journal of Social sciences; 2012; Volume 3
11. A Sahar; "Predicting the severity of breast masses with data mining methods"; International Journal of computer science issue; 2013; Volume 10; issue 2
12. ZakariaSulimanzubi; "Improves treatement programs of lung cancer using data mining techniques; Journal of software engineering and applications; 2014
13. Rajashree Dash; "A hybridized K-means clustering approach for high dimensional dataset"; International Journal of Engineering, science, and technology; 2010; Volume 2
14. Labeed K Abdulgafoor; "Detection of Brain Tumor using modified K-means Algotithm and SVM"; International Journal of computer applications, NCRTCA; 2013
15. AdaandRajneet Kaur; "Using some data mining techniques to predict the survival year of lung cancer patient"; International Journal of computer science and mobile computing; 2013; Volume 2