CANCER PREDICTION SYSTEM USING DATA MINING TECHNIQUES

Premalatha D[1], Niveditha N[2], Poornima Vasanth V[3], Priyanka G N[4], Niveditha K B[5]
[1]Assistant Professor, [2,3,4,5]Student,
Department of Computer Science and Engineering,
Dr. T Thimmaiah Institute of Technology, Kolar Gold Fields-563120, Karnataka, India
[1]premalatha51.d@gmail.com, [2]niveditha221997@gmail.com, [3]poornima91997@gmail.com,
[4]priyagn27@gmail.com, [5]nivigowda666@gmail.com

Abstract
Cancer is one of the leading causes of death worldwide. Early detection and prevention of cancer plays a very important role in reducing deaths caused by cancer. 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 techniques to build a cancer risk prediction system is proposed here which predicts lung, breast, oral, stomach and blood cancers and is also user friendly, time and cost saving. This project uses data mining technology such as classification, clustering and prediction to identify potential cancer patients. The gathered data is pre-processed, fed into the database and classified to yield significant patterns using decision tree algorithm. Then the data is clustered using K- means clustering algorithm to separate cancer and non-cancer patient data. Further the cancer cluster is subdivided into six clusters. Finally, a prediction system is developed to analyze risk levels which help in prognosis. This project helps in detection of a person's predisposition for cancer before going for clinical and lab tests which is cost and time consuming.
Keywords: Cancer, Data Mining, Decision Tree, K-Mean, Risk Levels.

I. INTRODUCTION

Cancer has been characterized as a heterogeneous disease consisting of many different subtypes. The early diagnosis and prognosis of a cancer type have become a necessity in cancer research, as it can facilitate the subsequent clinical management.
The importance of classifying cancer patients into high or low risk groups has led many research teams, from the biomedical and the bioinformatics field, to study the application of machine learning (ML) methods. Therefore, these techniques have been utilized as an aim to model the progression and treatment of cancerous conditions. In addition, the ability of ML tools to detect key features from complex datasets reveals their importance.

A variety of these techniques, including Artificial Neural Networks (ANNs), Bayesian Networks (BNs), Support Vector Machines (SVMs) and Decision Trees (DTs) have been widely applied in cancer research for the development of predictive models, resulting in effective and accurate decision making. Even though it is evident that the use of ML methods can improve our understanding of cancer progression, an appropriate level of validation is needed in order for these methods to be considered in the everyday clinical practice.

Over the past decades, a continuous evolution related to cancer research has been performed. Scientists applied different methods, such as screening in early stage, in order to find types of cancer before they cause symptoms. Moreover, they have developed new strategies for the early prediction of cancer treatment outcome. With

the advent of new technologies in the field of medicine, large amounts of cancer data have been collected and are available to the medical research community. However, the accurate prediction of a disease outcome is one of the most interesting and challenging tasks for physicians. As a result, ML methods have become a popular tool for medical researchers. These techniques can discover and identify patterns and relationships between them, from complex datasets, while they are able to effectively predict future outcomes of a cancer type.

II. 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 andto predict survival rate and years of an abnormal patient so that cancer patients lives can be saved.

Labeed K Abdulgafoor et al [5] wavelet transformation and K- means clustering algorithm have been used for intensity based segmentation.

III METHODOLOGY

Cancer has been characterized as a heterogeneous disease consisting of many different subtypes. The early diagnosis and prognosis of a cancer type have become a necessity in cancer research, as it can facilitate the subsequent clinical management of patients. In this project we deal with this problem of detecting and categorizing the cancer using machine learning techniques.

The overall flow of the methodology is shown below

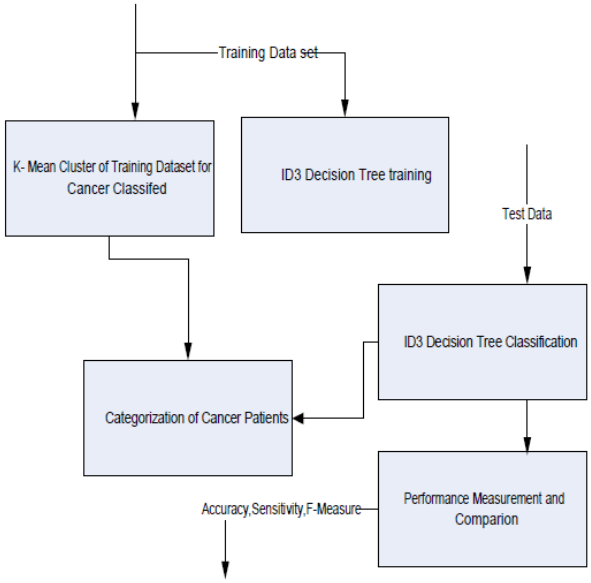

Fig 1: Proposed Work

Proposed methodology consists following subparts.

1. Using training data set to create a ID3 Decision tree classifier
2. Use the ID3 decision tree for classification of cancer or not.
3. Cluster the training data set for label of cancer and create categorizes of cancer. Using this categorize the cancer classified data.

IV SYSTEM ANALYSIS

System analysis is the method, by which we get some answers concerning the present issues, portrays things and necessities and evaluates the courses of action. It is the perspective about the affiliation and the issue it incorporates, a plan of advances those assistants in dealing with these issues. Feasibility study expects a key part in structure examination which gives the goal for layout and headway.

Feasibility Study

The datasets used for project testing is available in internet and the tools used as matured and easy to use. So, there is risk in implementing the project.

Economical Feasibility

Since all the tools and dataset are open source and easily downloaded without any cost, there is no economic risk in the project.

ISSN (PRINT): 2393-8374, (ONLINE): 2394-0697, VOLUME-6, ISSUE-6, 2019
DOI: 10.21276/ijcesr.2019.6.6.28

Technical Feasibility

The project coding is in Java and we estimate around 2000 lines of code for the project.

Social Feasibility

The project will be very useful at diagnostics centers and research institutions to verify their main stream diagnostic methods. So, it will be accepted by the community.

V SYSTEM DESIGN

System design is an inventive procedure; a great outline is the way to compelling framework. The framework "Outline" is characterized as "The procedure of applying different methods and standards with the end goal of characterizing a procedure or a framework in adequate point of interest to allow its physical acknowledgment". Different configuration components are taken after to build up the framework. The design detail portrays the elements of the framework, the parts or components of the framework and their appearance to end-clients.

Input Design

The input design is the procedure of changing over the client arranged inputs into the PC based organization. The objective of planning info information is to make the mechanization as simple and free from blunders as would be prudent.Giving a decent information configuration to the application simple information data and determination elements are embraced. The information plan necessities, for example, ease of use, reliable arrangement and intelligent dialog for giving the right message and help for the client at ideal time are additionally considered for the improvement of the undertaking.

Output Design

A quality yield is one, which meets the prerequisites of the end client and presents the data plainly. In any framework aftereffects of preparing are conveyed to the clients and to different frameworks through yields. It is most essential and direct source data to the client. Proficient and insightful yield enhances the frameworks association with source and destination machine. Yields from PCs are required essentially to get same parcel that the client has send rather than debased bundle and caricature parcels. They are likewise used to give to lasting duplicate of these outcomes for later counsel.

IV EXPERIMENTAL RESULTS

The results are separated into three parts. The first is the frequent and significant pattern discovery. The second is mapping the cancer to its cluster and the third is prediction by giving risk score as output. At the beginning all the input data is stored in the non cancer cluster further it gets classified and clustered by the model. A single user input data is fed into the system and gets classified according to the significant pattern to which it matches through decision tree, gets analyzed for its risk score merged with either one of the Non cancer and cancer clusters. This gives the result whether the patient has cancer or not. Again, the data is merged with any one of the subsequent cancer clusters to which its symptoms belong. The type of cancer the patient has is found out from this step. It is also compared with the entire database to find its exact or relevant match so that a data with severe cancer related symptoms gets a pair only in the cancer cluster and it cannot get merged with non cancer cluster even by mistake. With each new entry getting appended to the model the process becomes intelligent and ensures accurate results. This step ensures the accuracy of the model. The front end user interface is designed in a user friendly manner to help people use the system without any hassles

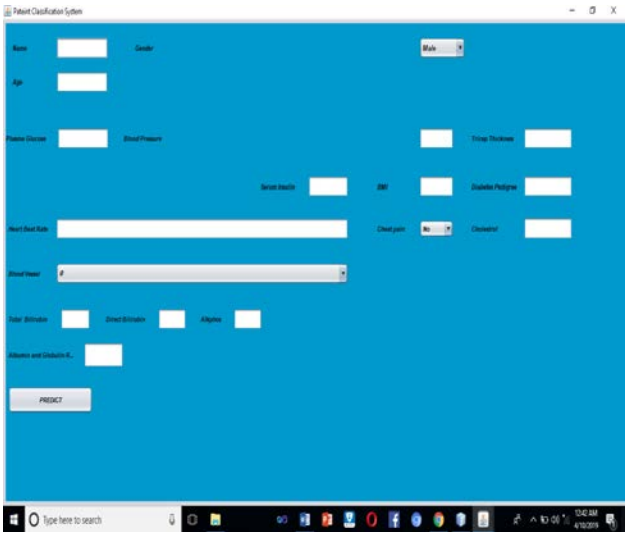

Fig 2: User Input Screen

ISSN (PRINT): 2393-8374, (ONLINE): 2394-0697, VOLUME-6, ISSUE-6, 2019
DOI: 10.21276/ijcesr.2019.6.6.28

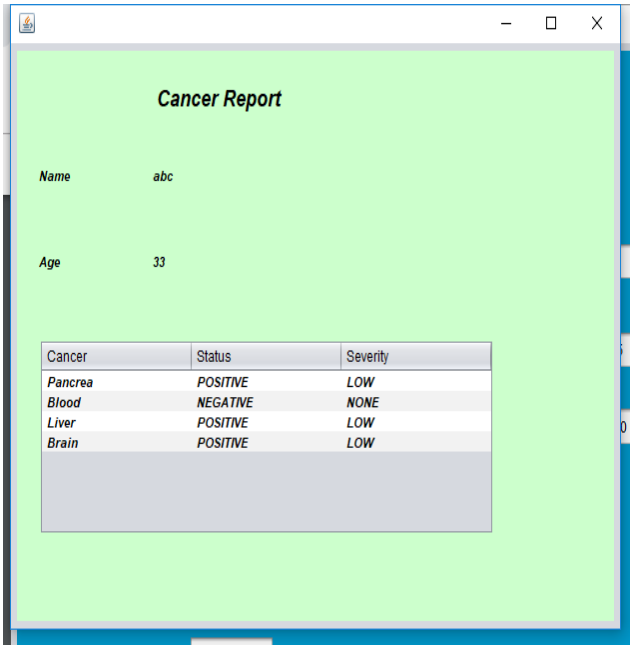

Fig 3: Report Screen with Prediction Results

The report shows the cancer status of a patient whether or not he has cancer by matching his data with the entire database, his risk score generated by the significant pattern mined by decision tree, the type of cancer he has which is given as a cluster output, whether his risk status is medium or severe and finally some recommended tests by medical experts to confirm the presence of cancer. This application is directly linked with the knowledge base and the back end model so that it could send the new raw data to the storage unit as well as the model to process it through analyzing the risk scores and also compares the data with existing cases in the knowledge base.

VII. Conclusion

We have developed a Cancer Detection tool to detect the type of cancer and the severity of cancer. The performance of the tool is tested against various dataset and the classifier is found to have accuracy of about 95%. Our future work will be on exploring other classification methods to still improve the accuracy.

REFERENCES

[1] Hanahan, R.A. Weinberg**Hallmarks of cancer: the next generation** Cell, 144 (2011), pp. 646-674.

[2] Polley, B. Freidlin, E.L. Korn, B.A. Conley, J.S. Abrams, L.M. McShane**Statistical and practical considerations for clinical evaluation of predictive biomarkers**

[3] J.A. Cruz, D.S. Wishart**Applications of machine learning in cancer prediction and prognosis** Cancer Informat, 2 (2006), p. 59

[4] Fortunato, M. Boeri, C. Verri, D. Conte, M. Mens ah, P. Suatoni, *et al.***Assessment of circulating microRNAs in plasma of lung cancer patients** Molecules, 19 (2014), pp. 3038-3054

[5] .M. Heneghan, N. Miller, M.J. Kerin**MiRNAs as biomarkers and therapeutic targets in cancer** Curr Opin Pharmacol, 10 (2010), pp. 543-550

[6] Madhavan, K. Cuk, B. Burwinkel, R. Yang**Cance r diagnosis and prognosis decoded by blood-based circulating microRNA signatures**
Front Genet, 4 (2013)

[7] K. Zen, C.Y. Zhang**Circulating microRNAs: a novel class of biomarkers to diagnose and monitor human cancers** Med Res Rev, 32 (2012), pp. 326-348

[8] S. Koscielny**Why most gene expression signatures of tumors have not been useful in the clinic**Sci Transl Med, 2 (2010)
[14 ps12-14 ps12]

[9] S. Michiels, S. Koscielny, C. Hill**Prediction of cancer outcome with microarrays: a multiple random validation strategy** Lancet, 365 (2005), pp. 488-492

[10] M. Bishop**Pattern recognition and machine learning** Springer, New York (2006)

[11] T.M. Mitchell**The discipline of machine learning: Carnegie Mellon University** Carnegie Mellon University, School of Computer Science, Machine Learning Department (2006)

[12] I.H. Witten, E. Frank**Data mining: practical machine learning tools and techniques** Morgan Kaufmann (2005)

[13] Petrovic **Introduction to computational intelligence techniques and areas of their applications in medicine** Med Appl Artif Intell, 51 (2013)