International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 30 Issue 2 – MARCH 2023.

RISK PREDICTION OF LUNG CANCER USING DATA ANALYTIC TECHNIQUES

R.VIMAL RAJA[1],S. GOPALAKRISHNAN[2]
DEPT. OF COMPUTER SCIENCE & ENGINEERING,
CK COLLEGE OF ENGINEERING & TECHNOLOGY,

Abstract— **According to the World Health Organization (WHO), Cancer is one of the most fatal diseases in the world. Early detection and prevention of cancer plays a very important role in reducing death caused by cancer. A disease that is commonly misdiagnosed is lung cancer. Our project is to prediction of lung cancer by analyzing a various collection of raw data from patients like occupational hazards, age, pollution, alcohol use, genetic risk, smoking, balanced diet, obesity, chest pain, coughing, fatigue, weight loss. The gathered raw data is pre- processed to data sets. We implement an effective prediction with the help of K-means algorithm and support vector machine. K-means clustering is used to cluster a datasets into data groups[11]. Support vector machine used to prediction of cancer levels like low, medium, high. Finally a prediction system is developed to analyze risk levels which help in prognosis[2]. If the lung cancer is detected and predicted in its early stages will increase survivalrate.**

***Keywords*-Lung Cancer, K-means algorithm, comma separated value, Support vector machine.**

I. INTRODUCTION

The riskiest and most lethal form of cancer is lung cancer. Smoking is the primary cause of lung cancer[1,2,3,4,5] and kills 85 out of every 100 people annually. Those who don't smoke have a lesser risk factor, but they could still be harmed by other smokers' smoke[3]. There are a lot of additional risk factors, including passive smoking, radiation exposure, and air pollution. A metallic chemical element called uranium breaks down over time to create radon gas, which spreads through the air and water and causes pollution as well as serious lung damage[4]. When a relative has lung cancer, the risk of developing lung cancer increases. This could be because of a shared environment, genetics, or both[4]. Lung cancer is also linked to a history of chronic pulmonary illnesses[4,5].

Data Analytics refers to the set of quantitative and qualitative approach in order to derive valuable insights from data. It involves many processes that include extracting data, categorizing it in order to analyze the various patterns, relations, connections and other such valuable insights from it. Data analytics is broken down into four basic types such as predictive analytics, prescriptive analytics, diagnostic analyticsand descriptive analytics in that we have chosen a predictive analytics. Predictive analytics technology uses data, statisticalalgorithmsandmachine-learningtechniquesto identify the likelihood of future outcomes based on historical data. We focus on lung cancer datasets. There are two types of attributes such as demographic attributes (gender, age, location) and diagnosisattributes (smoking, Alcohol usage). K-means clustering is to cluster a group of similar datasets and support vector machine is to classify a data sets for prediction of risk level. The identification of individuals at high risk will facilitate early diagnosis, reduce overall costs, and also improvethecurrentpoorsurvivalfromlungcancer.

I. LITERATURESURVEY

1. *Nooshin Hadavi[1], Md.Jan Nordin[2], Ali Shojaeipour*

Using CT scan image processing method to detect lung cancer cells. Computer aided diagnosis (CAD) is one of necessary part of diagnosis procedure for early detection[13]. A CAD system is created based on computer algorithm with medical science that uses the medical imagesasinputandperformssomeprocessonthem,and then the output of CAD helps doctors or radiologist to detect any disorder of tissues and makedecision.

2. *T. Sowmiya, M. Gopi, M. New Begin, L.ThomasRobinso*

In this paper they described Cancer as the most dangerous diseases in the world. Lung cancer is one of the most dangerous cancer types in the world. These diseases can spread worldwide by uncontrolled cell growth in the tissues of the lung. Early detection of the cancer can save the life and survivability of the patients who affected by this diseases. In this paper we survey several aspects of data mining procedures which are used for lung cancer prediction for the patients. Data mining concepts is useful in lung cancer classification. We also reviewed the aspects of ant colonyoptimization (ACO) technique in data mining[15]. Ant colony optimization helps in increasing or decreasing the disease prediction value of thediseases.

3. Ada[1], Rajneet Kaur[2] (2013)–

In this paper uses a computational procedure that sorts the images into groups according to their similarities.InthispaperHistogramEqualizationisused for pre-processing of the images and feature extraction process and neural network classifier to check the state of a patient in its early stage whether it is normal or abnormal. After that we predict the survival rate of a patient by extracted features. Experimental analysis is made with dataset to evaluate the performance of the differentclassifiers.

4. Vignesh M BalajiR

In this paper using a statistical analysis, A boxplot is a data analysis method used to find the output of the samples With the use of boxplot, we can easily compare the different datasets between upper quartile and lower quartile. Decision tree model is use toclassify aprediction.

II. PROPOSED SOLUTION:LCPS

In our proposed system we have a collection of raw data, then gathered data is pre-processed to produce required dataset and store in CSV (comma separated values) file format. We have to do analysis by using support vector machine algorithm. SVM classifies data sets into three types such as risk is low, medium and high. It produces significant accuracy result withless computation power. With the help of classified data we have to do clustering by using k-means algorithm. It performs certain number of iterations randomly which access the nearest observations into k, so as to attain the high speed time consumption and offers stability of the accurate result. This approach provides efficient and effective results in prediction.

III. SYSTEM ARCHITECTURE

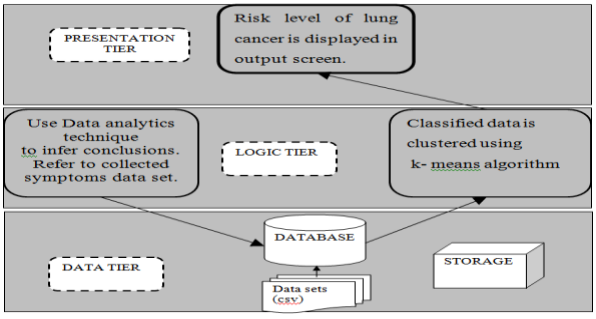

Fig 1: Architecture Diagram

IV.DATA MININGTOOLS

KEEL (Knowledge Extraction based on Evolutionary Learning), an open source data mining framework that is increasingly being used in healthcare. KEEL includes many algorithms for regression, classification, clustering, and more. It allows creating experiments using multiple datasets and algorithms, independently scripted from the user interface. We selected the open source WEKA framework, which has all the desired data processing and machine learning algorithms and is widely used for research and development. It has a great user interface and provides excellent visualization tools to understand the models. WEKA features hundreds of algorithms for data processing, feature selection, clustering, finding association rules, classification, etc. (Witten, 2016). LCPS integrates selected methods for removing outliers and irrelevant features, and predicting the patients' health status using classification rules, decision trees, instance based learning, probabilistic approach, and laterregressiontrees.Sincechoosingthe–best predictive algorithm for a given dataset requires a lot of computations, automation is needed. This in turns requires making assumptions, which is another reason why the current system supports databases for lung cancer only.

V. METHODOLOGY

A. Data Collection

1000 patient's data is obtained from different diagnostic center. There are male and female patients whose age between 20 to 80 years old. 23 risk factors were considered for Lung cancer assessment, which includes- Age, Gender, Air Pollution, Alcohol use, Dust Allergy, Occupational Hazards, Genetic Risk, Chronic LungDisease,BalancedDiet,Obesity,Smoking,passive smoker, chest pain, coughing of blood, Fatigue, weight loss, shortness of breath, wheezing, swallowing difficulty, clubbing of finger nails, Frequent Cold, Dry Cough, Snoring are taken to consider for predicting the lungcancer.

B. DataPre-Processing

Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, lacking in certain behaviours or trends, and is likely to contain many errors. Data pre- processing is a proven method of resolving such issues. The above risk factors have the equivalent numeric value based on the hazards of symptoms.

Fig 2: Data Preprocessing

C. Comma SeparatedValues(CSV)

Comma Separated Values (CSV) file is a delimited text file that uses a comma to separate values. A CSV file stores tabular data (numbersand text) in plain text. Each line of the file is a data record. Each record consists of one or more fields, separated by commas. They workboth with UNICODE and ASCII. A CSV file translates one character set to another. CSV files cannot represent object oriented database. Because CSV records expected to have same structures. CSV files are also called as flat files. All the risk factors such as Age, Gender, Air Pollution, Alcohol use, Dust Allergy, Occupational Hazards, Genetic Risk, and Chronic Lung Diseases etc. are separated by commas.

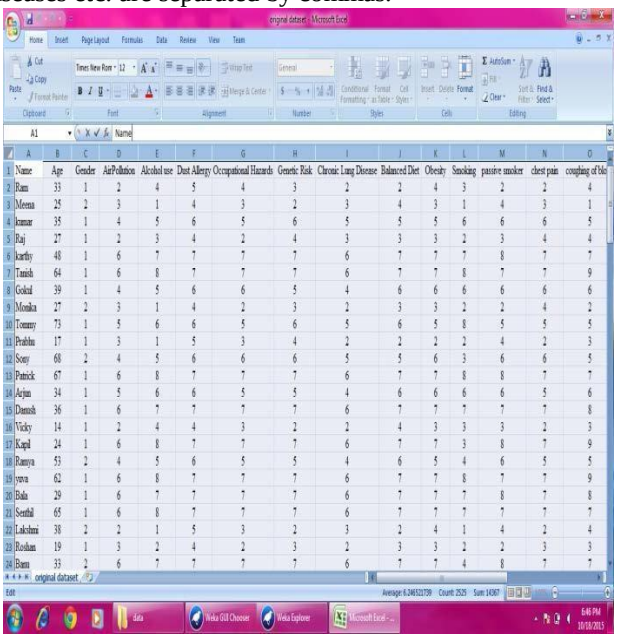

Fig 3: CSV File

D. Classification

Support vector machine is supervised learning models with associated learning algorithms that analyze data and recognize patterns used for classification. It employed in both classification and regression problems. The objective of the support vector machine algorithmis to find ahyperplaneinan N-dimensional space(N—thenumber of features) that distinctly classifies the data points. It can easily handle multiple continuous and categorical variables. SVM has an extra advantage of automatic model selection in the sense that both the optimal number and locations of the basis functions are automatically obtained during training. The SVM classifiers showed a great performance since it maps the features to a higher dimensional space. The SVM takes a set of input data and for each given input, predicts a risk level of Lung cancer which of three classes such as low, medium, high.

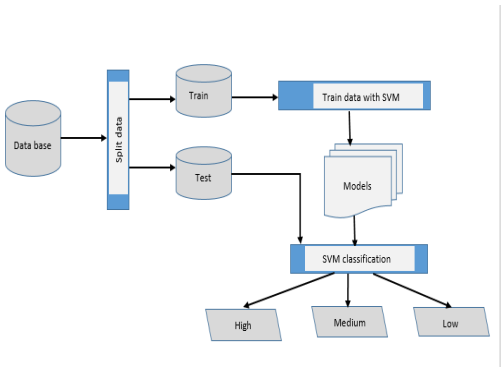

Fig 4: Support Vector Machine

E. Clustering

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K[11]. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity.

K means is an iterative clustering algorithm that aims to find local maxima in each iteration.

This algorithm works in these 5 steps :

1. Specify the desired number of clusters K : Let us choose k=3 for risk levels ofcancer.
2. Randomlyassigneachdatapointtoacluster.
3. Compute clustercentroids.
4. Re-assign each point to the closest cluster centroid.
5. Re-compute clustercentroids.
6. Repeat steps 4 and 5 until no improvementsare possible.

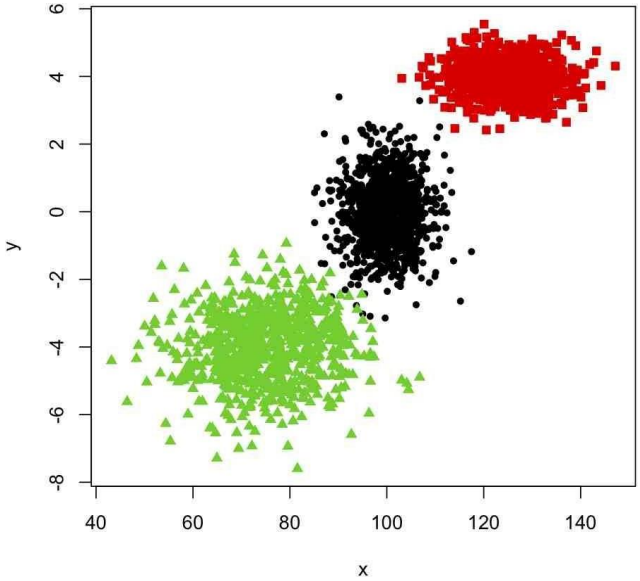

Fig 5: K-Means output

VII. RISK FACTORS OF LUNGCANCER

A. *Smoking*

Use of tobacco products and, in particular, cigarette smoking ,is responsible for the majority of lung cancer cases, with an estimated attributable risk of y 90% in males and 80% in females. Such attributable proportions are place specific and depend on the prevalence of tobacco smoking and of other exposures. The smoking-associated risks are dependent on the age of starting to smoke, the duration of smoking, and the level and pattern of smoking.

B. *Air pollution*

Air pollution comprises a large number of compounds that are usually correlated over relatively short time periods, but changes in emissions over longtime periods may result in substantial modification The dramatic increase in morbidity and mortality that occurred as a result of high air pollution levels.

VIII. FLOW CHART

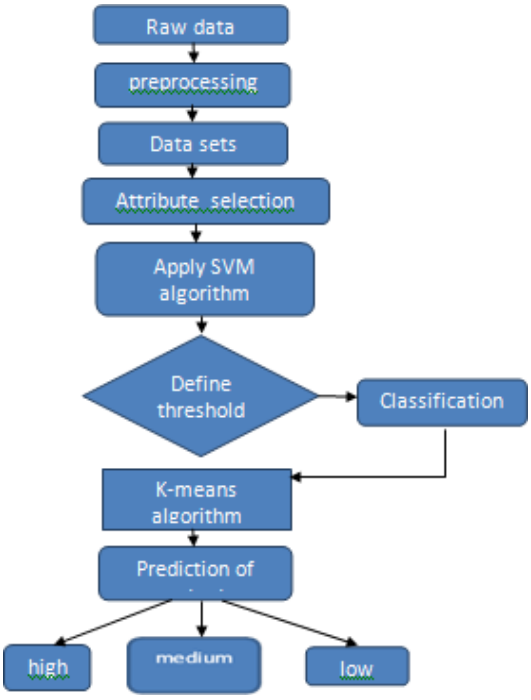

Fig 6: Flow chart

IX. CONCLUSION

In our project The prediction analysis is the technique in which user predicts the future on the basis of current situations. The prediction analysis consists of two steps. The first step is of classification, which will classify a risk levels as low ,medium, high. The second step consists of clustering, which will cluster the similar and dissimilar type of data clustering. K means clustering is used for the clustering. The SVM classifier is used to classify the dataset for predicting the complex data. The k-mean clustering consists of two steps. In the first step, the arithmetic mean of the loaded dataset is calculated which will be the centroid point. In the second step, Euclidean distance from the centroid point is calculated which defines similarity between the data points. The accuracy of clustering and classification is reduced. The proposed improvement leads to increase accuracy of classification. The proposed improvement and existing technique is being implemented in WEKA TOOL and it is being analyzed that accuracy is increased, execution time is reduced.

REFERENCES

[1] Chiefs of Ontario . Cancer Care Ontario and Institute for Clinical Evaluative Sciences; 2017. Lung Cancer in First Nations People in Ontario. Ontario. [Google Scholar]

[2] Ettinger D.S., Wood D.E., Aisner L.D., Akerley W., Bauman J., Bazhenova A.L. Non–small cell lung cancer, version 1.2017. J. Natl. Compr. Canc. Netw. 2016 October 14;2016 [Google Scholar]

[3] Kennedy M., Beddy P., Bruzzi J., Bruzzi J., Murray J., O'Regan K. sixteenth ed. Department of Health; Dublin: 2017. Diagnosis, Staging and Treatment of Lung Cancer (NCEC National Clinical Guideline.http://health.gov.ie/national-patient-safety-office/ncec/national-clinical-guidelines Available at: [Google Scholar]

[4] Shead D., Corrigan A., Kidney S., Hanisch L., Clarke R., Williams K. first ed. National Comprehensive Cancer Network; Washington: 2017. Lung Cancer Screening. [Google Scholar]

[5] Wood D.E., Kazerooni E.A., Baum S.L., Eapen G.A., Ettinger D.S., Hou L. Lung cancer screening, version 3.2018, NCCN clinical practice guidelines in oncology. J. Natl. Compr. Canc. Netw. 2018 Apr 1;16(4):412–441. [PMC free article] [PubMed] [Google Scholar]

[6] Lan-Wei Guo a,1 , Zhang-Yan Lyu b,1 , Qing-Cheng Meng c , Li-Yang Zheng a , Qiong Chen a , Yin Liu a , Hui-Fang Xu a , Rui-Hua Kang a , Lu-Yao Zhang a , Xiao-Qin Cao a , Shu-Zheng Liu a , Xi-Bin Sun a , Jian-Gong Zhang a , Shao-Kai Zhang a, - A risk prediction model for selecting high-risk population for computed tomography lung cancer inChina.https://doi.org/10.1016/j.lungcan.2021.11.015.Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license.

[7] Chiu, H.-Y.; Chao, H.-S.; Chen, Y.-M. Application of Artificial Intelligence in Lung Cancer. Cancers 2022, 14, 1370. https://doi.org/10.3390/cancers14061370

[8] Cancer. Available online: https://www.who.int/news-room/fact-sheets/detail/cancer (accessed on 29 November 2021).

[9] Luo, Y.H.; Chiu, C.H.; Scott Kuo, C.H.; Chou, T.Y.; Yeh, Y.C.; Hsu, H.S.; Yen, S.H.; Wu, Y.H.; Yang, J.C.; Liao, B.C.; et al. Lung Cancer in Republic of China. J. Thorac. Oncol. 2021, 16, 519–527. [CrossRef] [PubMed]

[10] Ankit Agrawal, SanchitMisra, RamanathanNarayanan, LalithPolepeddi, AlokChoudhary,–ALungCancerOutcome Calculator Using Ensemble Data Mining on SEERData"(2011).

[11] PurvashiMahajan,AbhishekSharma,–RoleOf K-means Algorithm in Disease Prediction‖, International Journal of Engineering And Computer Science, ISSN: 2319-7242,Vol.5, Issue 4, 2016,pp.16216-16217.

[12] Schilham A, Prokop M, van Ginneken B. Computer analysis of computed tomography scans of the lung : a survey. IEEE TransMedical Imaging. 2006; 25(4):385–405. doi:10.1109/TMI.2005.862753

[13] Dipali D, Pokale NB. Comprehensive survey on clustering algorithms and Similarity measures. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET). 2015 Jan; 4(1):239–42

[14] Ayyadurai.P, Kiruthiga.P, Valarmathi.S, Amritha.S , Respiratory Cancerous Cells Detection Using TRISS Model andAssociation Rule Mining, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 3, March2013.

[15] Sowmiya, M. Gopi, M. New Begin L.ThomasRobinson–OptimizationofLungCancerusing Modern data mining techniques.‖International Journal of Engineering Research ISSN:2319-6890