

# EARLY DETECTION OF LUNG CANCER USING DATA ANALYTIC TECHNIQUES

S.Rajasekar<sup>#1</sup> and R.Vimal Raja<sup>\*2</sup>

<sup>#</sup> Dept. of CSE, CK College of Engineering & Technology, Cuddalore, Tamil Nadu, India

<sup>\*</sup> Dept. of CSE, CK College of Engineering & Technology, Cuddalore, Tamil Nadu, India

**Abstract**— According to the World Health Organization include extracting data, categorizing it in order to analyze the various patterns, relations, connections and (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. 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. If the lung cancer is detected and predicted in its early stages will increase survival rate.

**Index Terms**— Cancer, K-means algorithm, comma separated value, Support vector machine.

## I. INTRODUCTION

Lung cancer is the leading cause of cancer-related death worldwide, with mortality rate exceeding that of prostate, breast, and colon cancer combined. It is one of the poorest survival

outcomes of any cancer because a vast majority of patients are diagnosed at an advanced stage when surgical resection or other treatment options are less effective. The success of any lung cancer program will depend on the successful identification of individuals at high risk. Risk prediction models have been recognized as a method of identifying individuals at high risk of developing lung cancer. Data analytics is used for prediction of Lung cancer. Data Analytics refers to the set of quantitative and qualitative approach in order to derive valuable insights from data. It involves many processes that other such valuable insights from it. Data analytics is broken down into four basic types such as predictive analytics,

prescriptive analytics, diagnostic analytics and descriptive analytics in that we have chosen a predictive analytics. Predictive analytics technology uses data, statistical algorithms and machine-learning techniques to 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 diagnosis attributes (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 improve the current poor survival from lung cancer.

## II. LITERATURE SURVEY

1.Nooshin Hadavi<sup>1</sup>, Md.Jan Nordin<sup>2</sup>, 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. A CAD system is created based on computer algorithm with medical science that uses the medical images as input and performs some process on them, and then the output of CAD helps doctors or radiologist to detect any disorder of tissues and make decision.

2. T. Sowmiya, M. Gopi, M. New Begin, L.Thomas Robinso

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 colony optimization (ACO) technique in data mining. Ant colony optimization helps in increasing or decreasing the disease prediction value of the diseases.

3. Ada<sup>1</sup>, Rajneet Kaur<sup>2</sup> (2013) –

In this paper uses a computational procedure that sorts the images into groups according to their similarities. In this paper Histogram Equalization is used 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 different classifiers.

4. Vignesh M Balaji R

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 to classify a prediction.

### III. 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 with less

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.

### IV. SYSTEM ARCHITECTURE

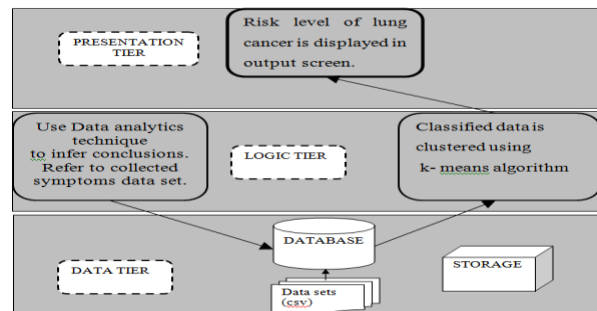


Fig 1: Architecture Diagram

### V. DATA MINING TOOLS

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 later regression trees. Since choosing the -bestl 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.

VI. 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 Lung Disease, Balanced Diet, 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 lung cancer.

B. Data Pre-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 Separated Values(CSV)

Comma Separated Values (CSV) file is a delimited text file that uses a comma to separate values. A CSV file stores tabular data (numbers and 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 work both

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.

Name	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	Occupational Hazards	Genetic Risk	Chronic Lung Disease	Balanced Diet	Obesity	Smoking	passive smoker	chest pain	coughing of blood
Ram	33	1	2	4	5	4	3	2	2	4	3	2	2	4
Menna	25	2	3	1	4	3	2	3	4	3	1	4	3	1
Kumar	35	1	4	5	6	5	6	5	5	5	6	6	6	5
Raj	27	1	2	3	4	2	4	3	3	3	2	3	4	4
Karthi	48	1	6	7	7	7	7	6	7	7	7	8	7	7
Tanish	64	1	6	8	7	7	7	6	7	7	8	7	7	9
Gokul	39	1	4	5	6	6	5	4	6	6	6	6	6	6
Mrunika	27	2	3	1	4	2	3	2	3	3	2	2	4	2
Toomay	73	1	5	6	6	5	6	5	6	5	8	5	5	5
Prabhu	17	1	3	1	5	3	4	2	2	2	2	4	2	3
Soory	68	2	4	5	6	6	6	5	5	6	3	6	6	5
Patrick	67	1	6	8	7	7	7	6	7	7	8	8	7	7
Arijan	34	1	5	6	6	5	5	4	6	6	6	6	5	6
Danush	36	1	6	7	7	7	7	6	7	7	7	7	7	8
Vicky	14	1	2	4	4	3	2	2	4	3	3	3	2	3
Kapil	24	1	6	8	7	7	7	6	7	7	3	8	7	9
Ramya	53	2	4	5	6	5	5	4	6	5	4	6	5	5
Yvona	62	1	6	8	7	7	7	6	7	7	8	7	7	9
Bala	29	1	6	7	7	7	7	6	7	7	7	8	7	8
Senthil	65	1	6	8	7	7	7	6	7	7	7	7	7	7
Lakshmi	38	2	2	1	5	3	2	3	2	4	1	4	2	4
Roshan	19	1	3	2	4	2	3	2	3	3	2	2	3	3
Bhanu	33	2	6	7	7	7	7	6	7	7	4	8	7	7

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 algorithm is to find a hyper plane in an N-dimensional space(N—the number 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.

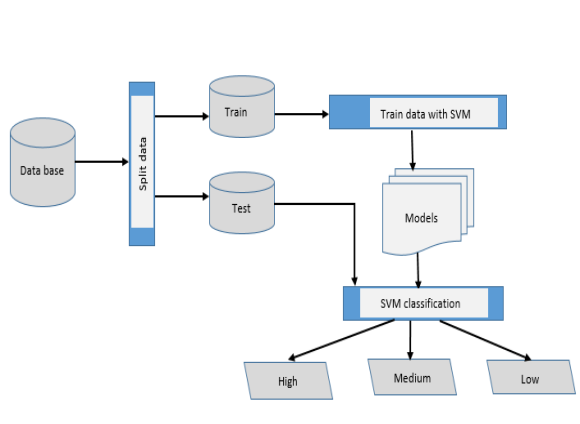


Fig 4: Support Vector Machine

### 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. 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 of cancer.
2. Randomly assign each data point to a cluster.
3. Compute cluster centroids.
4. Re-assign each point to the closest cluster centroid.
5. Re-compute cluster centroids.
6. Repeat steps 4 and 5 until no improvements are possible.

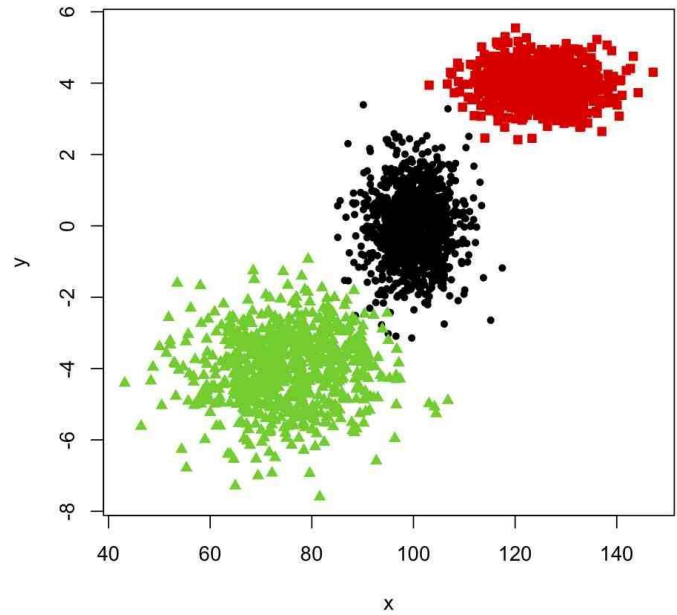


Fig 5: K-Means output

## VII. RISK FACTORS OF LUNG CANCER

### 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 long time 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

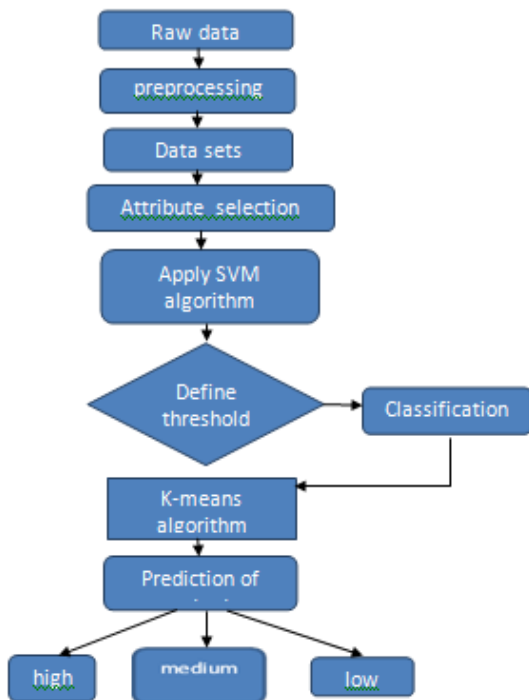


Fig 6: Flow chart

## 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] Ankit Agrawal, Sanchit Misra, RamanathanNarayanan, Lalith Polepeddi, Alok Choudhary, —A Lung Cancer Outcome Calculator Using Ensemble Data Mining on SEER Data"(2011).
- [2] Purvashi Mahajan, Abhishek Sharma, —Role Of K-means Algorithm in Disease Prediction], International Journal of Engineering And Computer Science, ISSN: 2319-7242,Vol.5, Issue 4, 2016, pp.16216-16217.
- [3] Roozgard A, Cheng S, Liu H. Malignant Nodule Detection on Lung CT Scan Images with Kernel RX Algorithm. Proceedings of the

- IEEE-EMBS International Conference on Biomedical and Health Informatics; 2012. p.499–502.
- [4] 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
- [5] 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
- [6] Ayyadurai.P, Kiruthiga.P, Valarmathi.S, Amritha.S , Respiratory Cancerous Cells Detection Using TRISS Model and Association Rule Mining, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 3, March 2013.
- [7] Sowmiya, M. Gopi, M. New Begin L.Thomas Robinson —Optimization of Lung Cancer using Modern data mining techniques. International Journal of Engineering Research ISSN:2319-6890(online),2347-5013(print)VolumeNo.3,Issue No.5, pp : 309-3149(2014).
- [8] Wingo PA, Ries LA, Giovino GA, et al. Annual report to the nation on the status of cancer, 1973– 1996, with a special section on lung cancer and tobacco smoking. J Natl Cancer Inst 1999; 91: 765–690.
- [9] Sellappan Palaniappan, Rafiah Awang —Intelligent Heart Disease Prediction System Using Data Mining Techniques|Department of Information Technology Malaysia University of Science
- [10] Technology Block C, Kelana Square, Jalan SS7/26 Kelana Jaya, 47301 Petaling Jaya, Selangor, Malaysia.
- [11] Sang Min Park, Min Kyung Lim, Soon Ae Shin & Young Ho Yun 2006. Impact of prediagnosis smoking, Alcohol, Obesity and Insulin resistance on survival in Male cancer Patients: National Health Insurance corporation study. Journal of clinical Oncology, Vol 24 Number 31 November 2006.