A Survey on Lung Cancer Prediction Using Ml Algorithms

Dr.Vijaya Ravichandran[1] Sathish R[2] Nandhisha S[3] Nivetha M[4] Sudha P[5]
[1,2,3,4,5]Department of Information Technology
[1,2,3,4,5]KGiSL Institute of Technology, Coimbatore, India

Abstract— Cancer which can be defined as a disease in which an abnormal cells divide uncontrollably and destroy body tissue. Lung cancer is the most common killer and plays an important role in mortality of vast amount of people. Lung cancer is the second most common cancer among others. Major reason for lung cancer is due to smoking. To prevent lung cancer deaths, high risk individuals are being screened with low-dose CT scans, because early prediction doubles the survival rate of lung cancer patients. Automatically identifying cancerous lesions in CT scans will save radiologists a lot of time. It will make diagnosing more affordable and hence will save many more lives. Currently lung cancer is predicting by using MRI scan and CT scans. We proposed a new system to predict lung cancer using textual data. In particular, we investigated sex, variables related to smoking history and addiction to nicotine, personal medical history, family history of lung cancer etc. In this work, we use supervised learning algorithms namely logistic regression, k-nearest neighbour etc., to predict lung cancer. Aim of the paper is to propose a model for early prediction and correct diagnosis of the disease which will help the doctor in saving the life of the patient.

Key words: Lung Cancer, CT Scans, MRI Scans, Textual Data, Supervised Learning Algorithm, Logistic Regression, K-Nearest Neighbour

I. INTRODUCTION

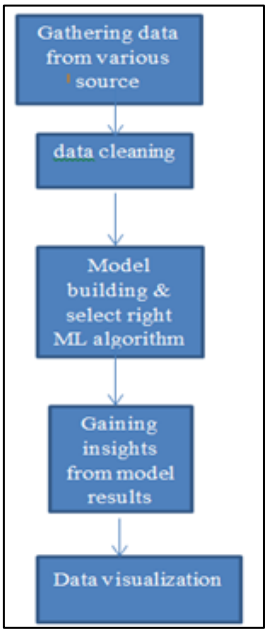
Lung cancer is the most common cause of death worldwide. Lung cancer is the uncontrolled growth of abnormal cells that start off in one or both Lung. The earlier detection of cancer is not easier process but if it is detected, it is curable[3]. A person who never smoked has lesser risk as compared to a person who smokes one pack daily. Nonsmokers are also causing lung cancer, however ratio is less than smokers. In order to predict lung cancer machine learning algorithms are used. Machine Learning is a method of data analysis that automates analytical model building. It is an application of artificial intelligence that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly. Machine Learning enables analysis of massive quantities of data. It may also require additional time and resource to train it properly. Mostly classification and clustering techniques are used in medical science field.Some of the machine learning methods are supervised learning, Unsupervised learning, Reinforcement learning.

1) Supervised machine learning: can apply what has been learned in the past to new data using labelled examples to predict future events. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.
2) Unsupervised machine learning: are used when the information used to train is either classified or labelled. The system doesn't figure out the right output , but it explores the data and can draw inferences from datasets to describe hidden structures from unlabelled data.
3) *Reinforcement learning:* is a learning method that interacts with its environment by producing actions and discovers errors and rewards. Trial and error search are the relevant characteristic of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behaviour within a specific context in order to maximize its performance.

Supervised Learning	Unsupervised Learning
Input data -Uses known and labeled data as input	Uses unknown data as input
Computational complexity -Very complex	Less complex
Accuracy -Accurate and reliable	Moderate accurate and reliable
Definition -The machine is already fed with required feature set into classify between inputs	The machine needs to figure out the output on its own by identifying patterns in the raw data provided to it.
Training data -Requires labeled training data	Does not requires training data

Table 1: Comparison between Supervised and Unsupervised Learning

II. MACHINE LEARNING PROCESS



III. METHODS

A. Data Preprocessing:

Data preprocessing is one of the most import feature. A huge unstructured data is available, but it is difficult to extract valuable information. While data preprocessing, duplicate values were deleted, there was no missing value that's why missing values technique was not used, steps of data preprocessing are shown in Figure 1.

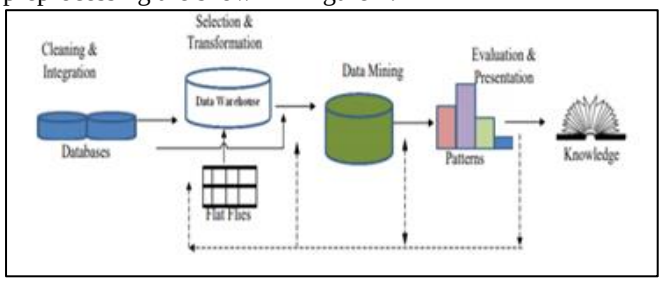


Fig. 1: Data Mining Steps

B. Data Analysis

To get the useful information from the data, data analysis is applied for describe and summaries irrespective of qualitative or quantitative data also identified the relationship/difference between the variables and comparison between the variables.

C. Cleaning and Integration

In this stage, irrelevant and unnecessary data was removed from the huge dataset. A clean dataset can give accurate results. Data cleaning phase was conducted before integration in data preprocessing. In data integration data from the multiple sources is combined in a common data source after performing data cleaning steps.

D. Selection and Transformation

In this phase, only relevant data was retrieved by applying feature selection technique and get relevant and decided data. There were large number of attributes in the dataset but keeping in mind the valuable data from the system only interesting attributes were extracted.

IV. ALGORITHMS

A. Decision Tree Algorithm

Decision tree is an algorithm used as a support tool for making decisions. [2]It uses a tree-like graph or structure of decisions and their possible outcomes that include the possibilities of an event, resource costs and utility. In a decision tree that has a flowchart-like structure, each internal node is called as a "test" on an attribute (e.g. where a coin flip possible outcomes are head or tail). Each branch refers to the outcome of the test and each leaf node refers to a class label (decision taken after computing all attributes). The path from root to the leaf is called as the classification rules.

B. Support Vector Machine (SVM)

Support Vector Machine (SVM) is mainly used for the classification process.[2] They are built on the idea that it defines the conclusion bordered between groups of instances. A decision plane of SVM is used to separate a set of items from different groups and also distinct a few support vectors in the training set.

C. Naive Bayes:

Naive Bayes classifiers are a collection of classification algorithms based on **Bayes' Theorem**. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

ALGORITHMS	ACCURACY
Naïve Bayes	65.5%
Decision Tree	65.2%
SVM	82.1%

Table 2: Comparison between algorithms:

V. DATA SET

Following attributes are used in this paper. [2] The attributes with their description and type, these attributes are used in this research which are used for early detection of lung cancer. These attributes are shown in following Table 2:

ATTRIBUTE	DESCRIPTION	TYPE
Patient Id	Patients ID	Numerical
Gender	Sex	Numerical
Age	Age in years	Numerical
Smoking	Does patient is smoker	1=yes, 2=no
Yellow fingers	Does patient has yellow finger	1= yes, 2=no
Anxiety	Does patient have anxiety	1= yes, 2=no
Peer Pressure	Does patient have peer pressure	1= yes, 2=no
Chronic disease	Does patient have chronic disease	1= yes, 2=no
Fatigue	Does patient feel tired mentally or physically	1= yes, 2=no
Allergy	Does patient have allergy	1= yes, 2=no
Wheezing	Does patient have wheezing problem	1= yes, 2=no
Alcohol	Does patient consume alcohol	1= yes, 2=no
Coughing	Does patient have cough problem	1= yes, 2=no
Shortness	Does patient have any difficulty in breathing	1= yes, 2=no

Swallow	Does patient have any swallow problem	1= yes, 2=no
Chest Pain	Does patient have chest pain	1= yes, 2=no

VI. EXISTING MODEL

Lung cancer is one of the major cancer types in the men as well as in women. Smoking is the main causes for the development of Lung Cancer. Overall 20% and 80% lung cancer occurs in Non-Smokers and Smokers respectively. [2]Currently, lung cancer is detected by using Chest X-Ray, Magnetic Resonance Imaging (MRI) scan and Computed Tomography (CT) scans - they can be used to: diagnose conditions – including damage to bones, injuries to internal organs, problems with blood flow, stroke, and cancer, PET CT (Positron Emission Tomography/Computed Tomography) -

This scan can sometimes detect disease before it shows up on other imaging tests and Bronchoscopy etc. by the health professional.

VII. PROPOSED METHODOLOGY:

Lung cancer will be normally predicted by using MRI scan and Computed Tomography (CT) scans. We proposed a system that will predict lung cancer by using textual data such as age, gender, smoking, alcohol, consumption, breathing problem, swallow problem, chest pain etc., We use clustering algorithm to group the data and supervised algorithms to predict the lung cancer using the attributes mentioned in data set table. We are going to implement a new algorithm which predict a person has lung cancer or not.

VIII. CONCLUSION

Lung Cancer is menacing cancer in the world with high mortality rate. It helps the society to change the lifestyle of the human being to avoid such malignant disease. We have applied data preprocessing techniques on our dataset for removing of noise and dirty data, dirty data is a common term in data mining. Data is cleaned by applying different techniques of data preprocessing in data mining. Classification technique is applied on cleaned data. As shown in Figure 2 that lung cancer ratio is greater in male as compare to female. [2] As shown Figure 3 that age was divided into 5 groups (Group one consists age from 1 to 18, Group Two consists from age 19 to 30, Group Three consists age from 31 to 45, Group Four consists age from 46 to 60 and Group Five consists age 61 to 100. This paper mainly focuses on predicting the lung cancer based on survey using machine learning algorithms. This paper compares various techniques based on efficiency in classification in order to predict lung cancer.

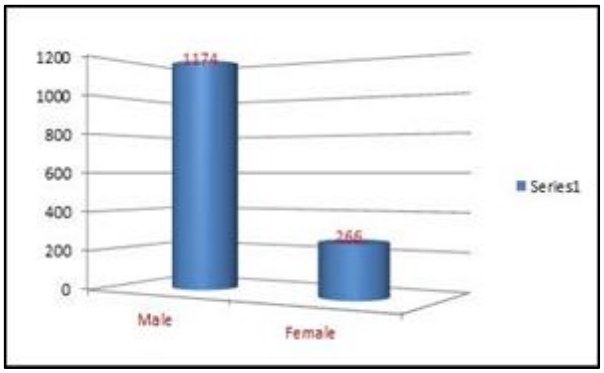


Fig 2 Gender wise lung cancer statistics

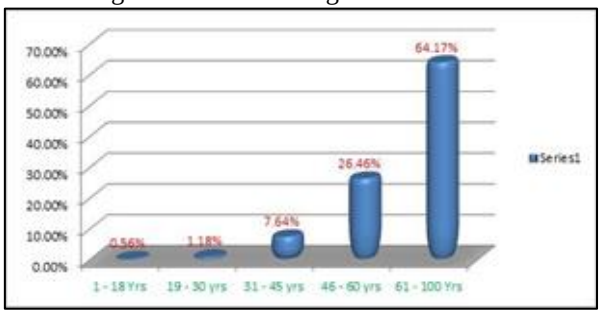


Fig 3 Age wise lung cancer statistics

REFERENCES

[1] [Survey on Lung Cancer Diagnosis Using Novel Methods by K. Kavitha and Dr.K.Rohini published in International journal of pure and applied mathematics
[2] Detection of lung cancer in Smokers and Non-smokers by applying Data Mining Techniques by Roy Qaiser Hussain and Abdul Azia published in Indian journal of science and technology.
[3] Study of classification Algorithm for lung cancer Prediction by Dr. T Christopher and J. Jamera Banu published in International Journal of Innovative science, Engineering and technology
[4] [Classification of multi-class microarray cancer data using ensemble learning by B.H Shekar and Guesh Dagnew
[5] Sex and smoking status effects on the early detection of early lung cancer in high-risk smokers by Annette MC Williams, Parmida Beigi , Akhila Srinidhi, Stephen Lam.
[6] Predicting Lung Cancer Survivability using Ensemble Learning Methods by Ali Safiyari and Reza Javidan
[7] Small-Cell Lung Cancer Detection Using a Supervised Machine Learning Algorithm by Quin Wu and Wenbing Zhao
[8] Real Time Data Collection and Analytics of Social Media Sites Using Netlytic by R. Sathish and M. Ambika

