International Journal of Computer Science and Engineering Communications.
Vol.3, Issue 3, 2015, Page.1040-1046, ISSN: 2347–8586

# Mining Techniques for Clinical Expert System and Predicting and Treating Lung Cancer with Big Data

**N.Naveenkumar[1], G.Selvavinayagam[2]**
[1]PG Scholar, [2]Assistant Professor,
Department of Information Technology,
SNS College of Technology, Coimbatore, India.
***Email:*** *nknavi5@gmail.com[1]*

**Abstract:** Cancer is the very most important cause of death for both Men and Women. The early apprehension and treatment of cancer can be helpful in curing the disease completely. So the obligation of techniques to detect the state of cancer nodule in early stage increasing. A disease that is commonly misdiagnosed is Lung Cancer. Earlier diagnosis of Lung Cancer saves massive lives, flaw which may lead to other harsh problems causing sudden fatal end detection and diagnosis of the disease. Its treat rate and prediction depends mainly on the early detection and diagnosis sudden fatal end. Data mining have found numerous applications in business and methodical domain. Valuable information can be discovered from application of data mining techniques in healthcare system. In this study, we momentarily examine the potential use of classification based data mining techniques in healthcare system. Big data collects the huge amounts of healthcare data set and survey many data mining approaches for predicting the likelihood of patients getting a Lung Cancer disease. Aim of the paper is to suggest a model for early detection and correct diagnosis of the disease using which will help the doctor in saving the life of the patient.
***Keywords:*** *Lung Cancer, Data Mining, Big data, Healthcare*

## I.INRODUCTION

Lung cancer is cancer that begins in the Lungs. It is the second most familiar cancer in men and women, as well as the important cause of cancer death in both men and women. There are two main types of lung cancer. They are non-small cell lung cancer and small cell lung cancer. A doctor called a pathologist uses a microscope to look at the cancer cells collected during your biopsy to tell which type of cancer you have. These two types of lung cancer generate and spread in different ways. Therefore, they are treated differently. Non-small cell lung cancer usually grows and spreads more progressively than small cell lung cancer.

### 1.1 Non-small cell lung cancer (NSCLC)

About 85% to 90% of lung cancers are non-small cell. There are three main types of NSCLC. While there are slight differences between them, they tend to have a parallel prediction (outlook) and are usually treated the same way:

- Adenocarcinoma. This is the most regular kind of non-small cell lung cancer. And it's the most common type of lung tumor in nonsmokers. It tends to grow in the outer limits of your lungs, and usually grows more slowly than other types of lung cancer.
- Squamous cell carcinoma (epidermoid carcinoma). This type of non-small cell lung cancer develops more often in smokers. These cancers tend to found in the middle part of the lungs.
- Large cell. This is the least frequent of the three kinds of non-small cell lung cancer. It tends to grow and widen early to other organs, which can make it harder to treat.

### 1.2 Small cell lung cancer (SCLC)

Only about 10% to 15% of people diagnosed with lung cancer have small cell lung cancer. Small cell lung cancer is also called oat cell cancer. It grows and spreads more quickly than non-small cell lung cancer. It regularly spreads to other parts of the body at an early stage. This type of cancer is almost always linked with smokers. If you don't smoke, you aren't likely to get small cell lung cancer.

### 1.3 Risk Factors

Men between 60 and 65 and women about 70 are at greater risk of having lung cancer. Those who smoke have a risk of developing lung cancer that is 10 to 17 times higher than that of non-smokers. Women who smoke have a risk that is 5 to 10 times higher than that of women who do not smoke. The risk increases with the number of cigarettes smoked per day and the number of years the person has smoked. Some evidence suggests, however, that women and African-American men are more vulnerable. The risk of lung cancer for non-smokers who are exposed to smoke in the environment (known as second-hand, passive, or involuntary smoking), is as much as 30 percent higher than that of those who are not. The risk is even higher for exposure to side stream smoke (from the smoldering end of a cigarette) than for mainstream smoke (smoke that has been exhaled by the smoker). Industrial and atmospheric pollutants are responsible for a small percentage of lung cancer. For example, the risk of death from lung cancer is six to seven times greater for asbestos workers compared to the general population.

### 1.4 Signs & Symptoms

In many cases, symptoms do not appear until the cancer is quite advanced. However, by the time a tumor does cause changes within the lungs, the signs include:

- Difficulty breathing—stridor (a harsh sound with each breath), wheezing, labored breathing, shortness of breath (SOB);
- Coughing, possibly with blood in sputum;
- Recurring pneumonia or bronchitis;
- Chest, shoulder, or arm pain;
- Loss of appetite;
- Weight loss;
- Bone pain;
- Hoarseness;
- Headaches or seizures;
- Swelling of the face or neck;
- Fatigue.

### 1.5 Early Detection

There are no routine showing tests for lung cancer. Detection at an early stage is possible with an x-ray or sputum analysis and some doctors order these tests, particularly for people who smoke. However, there is no proof that such attempts at lung cancer screening have a positive impact on treatment or survival. If a doctor suspects lung cancer, an x-ray is the first step in diagnosis.

### 1.6 Lung Cancer Treatment

The two major types of lung cancers are in spirit two completely different diseases, each of which has its own suggested therapies. Non-small cell lung cancers (squamous, adenocarcinoma and large cell carcinoma) are potentially curable with surgery, but largely unresponsive to chemotherapy. Patients with distant metastases from non-small cell lung cancer can be treated palliatively with radiation. Conversely small cell lung cancers do respond to chemotherapy and radiation, but are normally too far advanced at diagnosis for a surgical cure.

Surgically resectable non-small cell lung cancers have the best cure rate because they are usually Stage I or II (localized) tumors. An substitute medically inoperable non-small cell lung cancer is curative radiation. More advanced disease (positive lymph nodes, or inoperable tumors) also respond to radiation therapy. Stage IIIB tumors (extensive primary or contralateral nodes) are best treated with radiation. Small cell carcinoma is extremely virulent, with a rapid clinical course if left untreated. However, because of its rapid growth rate (it tends to be widely disseminated at the time of diagnosis), it is also more responsive to chemotherapy and irradiation than non-small cell carcinoma. Surgery is not suggested for small cell carcinoma.

### 1.7 Big Data in Health Care

The purpose of the Transforming Health Care through Big Data plan is to help executives from hospitals, health systems and other contributor organizations appreciate models for innovative uses of data assets that can enable organizations to reduce costs, improve quality and provide safer care. Contributing executives will assess, identify and suggest a framework related to managing new sources, collection & storage, analytics, reporting and securing big data assets. The Transforming Health Care through Big Data Project is comprised of individuals from provider, health system, health information technology, academic, and health policy domains. This varied group is well-versed in data analysis, patient-centered care, health information technology, decision support systems, and the very important to transform health care delivery with innovative uses of health data. It will help recognize the latest and most proven strategies that leverage health care data enabling organizations to achieve high quality, cost effective care.

## II. MINING APPROACH FOR CLINICAL SYSTEM

Clinical repositories containing large amount of biological, clinical & administrative data are increasingly becoming available as health care systems integrate patients information for research and utilization objective. Data mining techniques applied on these databases discover interaction and pattern which are helpful in studying the evolution & the management of disease. Data mining refers to extracting or "mining" knowledge from large amounts of data. Knowledge discovery as a process consists of an iterative sequence of Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, Knowledge presentation. With the rapid advancement in information technology, many different data mining techniques and approaches have been applied to complementary medicine. Statistics provide an inspiring background to define and evaluate the result. Some of the mining approaches for Clinical System

- Decision Tree
- Artificial Neural Networks
- Naïve Bayes

## 2.1 DECISION TREE

Inductive inference is the process of moving from concrete examples to general models, where the goal is to learn how to categorize objects by analyzing a set of instances (already solved cases) whose classes are known. Instances are typically represented as attribute-value vectors. Learning input consists of a set of such vectors, each belonging to a known class, and the output consists of a mapping from attribute values to classes. This mapping should accurately classify both the given instances and other unseen instances. A decision tree [Quinlan, 1993] is formalism for expressing such mappings and consists of tests or attribute nodes linked to two or more sub-trees and leafs or decision nodes labeled with a class which means the decision. A test node computes some outcome based on the attribute values of an instance, where each possible outcome is associated with one of the subtrees. An instance is classified by starting at the root node of the tree. If this node is a test, the outcome for the instance is determined and the process continues using the appropriate subtree. When a leaf is eventually encountered, its label gives the predicted class of the instance. The finding of a solution with the help of decision trees starts by preparing a set of solved cases. The whole set is then divided into 1) a training set, which is used for the induction of a decision tree, and 2) a testing set, which is used to check the accuracy of an obtained solution. First, all attributes defining each case are described (input data) and among them one attribute is selected that represents a decision for the given problem (output data).

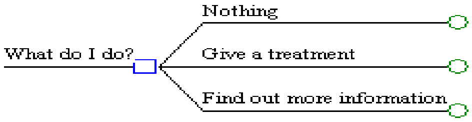

Figure1. The structure of a decision tree

- Square node
  - Decision node
  - Represents choice between actions
- Circle node
  - Chance node
  - Represents uncertainty
  - Potential outcomes of each decision

## 2.2 ARTIFICIAL NEURAL NETWORKS

Artificial neural network is a mathematical model that tries to simulate the structure and functionalities of biological neural networks. Basic building block of every structure artificial neural network is artificial neuron, that is, a simple mathematical model (function). Such a model has three simple sets of rules, multiplication, summation and activation. At the entrance of artificial neuron, the inputs are weighted, every input value is multiplied by individual weight in the middle section of artificial neuron is sum function that sums all weighted inputs and bias. At the exit of artificial neurons the sum of previously weighted inputs and bias is passing through activation function that is called also called transfer function. Although the working principles and simple set of rules of artificial neuron looks like nothing special the full potential and calculation power of these models come to life when we start to interconnect them into artificial neural networks. These artificial neural networks use simple fact that complexity can grow out of merely few basic and simple rules.

### A. Seedfill operation

Seed fill operation is an algorithm that determines the area connected to a given node in a multidimensional array. It performs the operation on background pixels of the binary image starting from the points specified in locations. It fills the holes in the binary image.

### B. Region of interest

A region of interest is a selected subset of samples within a dataset identified for a particular purpose. The concept of an ROI is commonly used in many application areas. In medical imaging, the boundaries of a tumor may be defined on an image or in a volume, for the purpose of measuring its size.

### C. Image segmentation

Segmentation is the process of partitioning an image into disjoint and homogenous this task can be equivalently achieved by finding the boundaries between the regions; these two strategies have been proven to be equivalent indeed. Regions of image segmentation should be uniform and homogeneous with respect to some characteristics such as gray tone or texture. Region interiors should be simple and without many small holes. Adjacent regions of segmentation should have significantly different values with respect to the characteristic on which they are uniform. Boundaries of each segment should be simple, not ragged, and must be spatially accurate." A more formal definition of segmentation can be given in the following way. Let I denote an image and let H define a certain homogeneity predicate; then the segmentation of I is a partition P of I into a set of N regions Rn, n = 1……. ∀ N, such that: 1) 1 N U R I =n n  = with R R ∅ ≠ In m  ; ≠n m ; 2) H(Rn) = true n∀ ; 3) H(Rn Y Rm) = false ∀ Rn and Rm adjacent. Condition 1) states that the partition has to cover the whole image; condition 2) states that each region has to be homogeneous with respect to the predicate H; and condition 3) states that the two adjacent region cannot be merged into a single region that satisfies the predicate H. Segmentation is an extremely important operation in several applications of image processing and computer vision, since it represents the very first step of low-level processing of imagery. As mentioned above, the essential goal of segmentation is to decompose an image into parts which should be meaningful for certain applications with color image segmentation which is becoming increasingly important in many applications. For instance, in digital libraries large collections of images and videos need to be catalogued, ordered, and stored in order to efficiently browse and retrieve visual information. Color and texture are the two most important low-level attributes used for content based retrieval of information in images and videos. Because of the complexity of the problem, segmentation with respect to both color and texture is often used for indexing and managing the data Texture feature extraction consists of finding the mean which is done by converting the size of an image into column matrix and adding each element of the matrix to find the sum which is divided by the product of rows and columns of the image. Entropy is calculated by using the formula E=sum(P*log(1/P)) Kurtosis is defined as measure of how outlier prone a distribution is. It is measure of whether the distribution is tall, skinny or short and squat compared to normal distribution of the same variance.

### D. Color representation

Several color representations are currently in use in color image processing. The most common is the RGB space where colors are represented by their red, green, and blue components color is better represented in terms of hue, saturation, and intensity. An example of such a kind of representation is the HSI space which can be obtained from RGB coordinates in various ways, e.g., by defining hue H= ) ) – – –G B R G B ( (arctan 3 ,2  saturation S=1-

min(R,G,B)/I, and intensity I= (R + G + B) /3, and by arranging them in a cylindrical coordinate system. The HSV space provides a description of color analogous to that of the HSI space, the hue H and the saturation S are similarly defined while the value V is defined as V =max(R, G,B).

### 2.3 NAÏVE BAYES

Naive Bayes classifier is a probabilistic classifier based on the Bayes theorem, considering Naive (Strong) independence assumption. Naive Bayes classifiers assume that the effect of a variable value on a given class is independent of the values of other variable. This assumption is called class conditional independence. Naive Bayes can often perform more sophisticated classification methods. It is mostly suited when the dimensionality of the inputs is high. When we want more competent output, as compared to other methods output we can use Naïve Bayes implementation. Naïve Bayesian is used to generate models with predictive capabilities.

#### 2.3.1 Bayes' Theorem:

Probility(B given A) = Probility(A and B)/Probility(A)

To calculate the probability of B given A, the algorithm counts the number of cases where A and B occur together and divides it by the number of cases where A occurs alone Let X be a data tuple. In, Bayesian terms, X is considered “evidence”. Let H be some hypothesis, such as that the data tuple X belongs class C. P(H|X) is the posterior probability, of H conditioned on X. In contrast, P(H) is the prior probability, of H. Bayes’ theorem is

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

$$\text{Posterior} = \frac{\text{Likelihood x Prior}}{\text{Evidence}}$$

Similarly, P(X|H) is the posterior probability of X conditioned on H. P(X) is the prior probability of X. In spite of their Naive design and apparently oversimplified assumptions, Naive Bayes classifiers often work much better in many complex real-world situations than one might expect. Recently, careful analysis of the Bayesian classification problem has shown that there are some theoretical reasons for the apparently unreasonable efficacy of Naive Bayes classifiers. An advantage of the Naive Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix. The Naive Bayesian classifier is fast and incremental can deal with discrete and continuous attributes, has excellent performance in real-life problems and can explain its decisions. as the sum of informational gains. However, its naivety may result in poor performance in domains with strong dependencies among attributes. In this paper, the algorithm of the Naive Bayesian classifier is applied successively enabling it to solve also non-linear problems while retaining all advantages of Naive Bayes. The comparison of performance in various domains confirms the advantages of successive learning and suggests its application to other learning algorithms.

## III. COMPARISION TABLE

| Techniques | Utility | Disease | Accuracy |
|---|---|---|---|
| Decision Tree | Decision Support | Diabetics | Medium |
| Neural Networks | Extracting Patterns, Detecting Trends | Liver | Medium |
| Naïve Bayes | Improving Classification Accuracy | Coronary Heart | High |

**Table.1 Comparison Table for Mining Techniques in Healthcare**

## IV. CONCLUSION

The purpose of this section is to provide an insight towards requirements of health domain and about suitable choice of available technique. Analyses show that most of the researches have been done in studying classic data mining algorithms such as Decision Trees, Naïve Bayes, and Artificial Neural Network, showing acceptable levels of accuracy. Comparing these three data mining techniques the Naïve Bayes technique are providing good accuracy result and it is the best one for detecting Lung Cancer in Clinical Expert System. We suggest it is the best model for early detection and correct diagnosis of the disease. It will help for the doctor in saving the life of the patient.

## REFERENCES

1. G.Parthiban, A.Rajesh, S.K.Srivasta, IJCA, Volume 24, No-3, June-2011, Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method.

2. Dhanashree S. Medhekar, Mayur P. Bote, Shruti D. Deshmukh, IJERSTE, Vol. 2 issue 3, March.-2013 Issn no: 2319-7463,Heart Disease Prediction System using Naive Bayes.

3. Divya Tomar and Sonali Agarwal, IJBSBT, Vol.5, No.5 (2013), pp. 241-266 , A survey on Data Mining approaches for Healthcare

4. Diana Dumitru, Annals of University of Craiova, Math. Comp. Sci. Ser, Volume 36(2), 2009, Pages 92-96 ISSN: 1223-6934, Prediction of recurrent events in breast cancer using the Naive Bayesian classification

5. Shweta Kharya, IJCSEIT, Vol.2, No.2, April 2012, Using Data Mining Techniques For Diagnosis And Prognosis Of Cancer Disease.

6. Divya Jain, Sumanlata Gautam, IJRDTM, Volume – 21, Issue 2, ISBN - 1-63102-446-9, June 2014,Study of data mining classification techniques in health care sector.

7. Sellappan Palaniappan, Rafiah Awang, IJCSNS, VOL.8 No.8, August 2008, Intelligent Heart Disease Prediction System Using Data Mining Techniques.

8. M. Durairaj, V. Ranjani, IJSTR, Volume 2, Issue 10, October 2013 Issn 2277-8616, Data Mining Applications In Healthcare Sector: A Study.

9. V.Krishnaiah,Dr.G.Narsimha,Dr.N.Subhash Chandra, IJCSIT , Vol. 4 (1) , 2013, 39 – 45, Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques.

10. Ms. Vibhavari Chavan, Prof. Rajesh. N. Phursule, IJCSIT, Vol. 5 (6) , 2014, 7932-7939, Survey Paper On Big Data.

11. N. Aditya Sundar, P. Pushpa Latha, M. Rama Chandra, IJESAT, Volume-2, Issue-3, 470 – 478, Performance Analysis Of Classification Data Mining Techniques Over Heart Disease Data Base.

12. N. Abirami, T. Kamalakannan, Dr. A. Muthukumaravel, IJETAE, Volume 3, Issue 7, July 2013, A Study on Analysis of Various Datamining Classification Techniques on Healthcare Data.