Supervised Learning Processing Techniques for Pre-Diagnosis of Lung Cancer Disease

K.Balachandran
Research Scholar, Professor
Christ University, Hosur Road,
Bangalore- 560 029

Dr. R.Anitha
Director / MCA
K. S. Rangasamy College of Technology,
Tiruchengode - 637 215.

ABSTRACT
Lung cancer disease is one of the dreaded disease is leading cause of death among men in developed and developing countries. Its cure rate and prognosis depends mainly on the early detection and diagnosis of the disease. Creating awareness among the general public about the disease and screening probable impact group requires lot of painstaking effort. This paper mainly focuses on selectively screening susceptible people for pre-diagnosis of Lung cancer disease. The approach adopted here is, conceptualizing artificial neural network model, based on statistical parameters based on cancer registry, symptoms and Risk factors. Supervisory delta learning approach is used to train the model. The model is developed using multi layer perceptron network and trained by established Lung cancer data. This model is then used for the test data. Tested data is again compared with the clinical diagnosed report and the model is reconfigured by including the current information and new training weights are computed.

Categories and Subject Descriptors
I.2 ARTIFICIAL INTELLIGENCE :I.2.1 Applications and Expert Systems (H.4, J)

General Terms
Design.

Keywords
Lung cancer, Non-small cell, Small cell, Perceptron, neural network, Supervisory learning, Delta Learning, Reinforcement learning

1. INTRODUCTION

1.1 Medical Background
Lung cancer is the one of the leading cause of cancer deaths in both women and men. Its cure rate and prognosis depends mainly on the early detection and diagnosis of the disease. Manifestation of Lung cancer in the body of the patient reveals through early symptoms in most of the cases. [1].

Characteristics of the tumor that affect and predict the survival outcome of patients with cancer are prognostic markers for cancer.[25] According to Indian Council of Medical Research Report "The process of carcinogenicity presents a major challenge to scientists and provides limited tools for its control. Indian health services are also not adequately equipped with facilities and expertise for management of cancers. Mortality and morbidity due to tobacco use is very high. In view of the national priorities, the focus of research in the field of cancer has been on the a etiology with identification of preventable risk factors, understand the mechanism of carcinogenesis and on operational research for control of tobacco use and common cancers through existing infrastructures. The multidisciplinary research involved clinical, epidemiological as well as basic sciences including modern molecular techniques". [4]

Usually lung cancer, develops within the wall or epithelium of the bronchial tree. But it can start anywhere in the lungs and affect any part of the respiratory system. Lung cancer mostly affects people between the ages of 55 and 65 and often takes many years to develop. [2]

There are two major types of lung cancer. They are Non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) or oat cell cancer. Each type of lung cancer grows and spreads in different ways, and is treated differently. If the cancer has features of both types, it is called mixed small cell/large cell cancer.

Non-small cell lung cancer is more common than SCLC and it generally grows and spreads more slowly. SCLC is almost related with smoking and grows more quickly and form large tumors that can spread widely through the body. They often start in the bronchi near the center of the chest. Lung cancer death rate is related to total amount of cigarette smoked.[3]

Smoking cessation, diet modification, and chemoprevention are primary prevention activities. Screening is a form of secondary prevention. Our method of finding the possible Lung cancer patients is based on the systematic study of symptoms and risk factors. Non-clinical symptoms and risk factors are some of the generic indicators of the cancer diseases. Environmental factors have an important role in human cancer. Many carcinogens are present in the air we breathe, the food we eat, and the water we drink. The constant and sometimes unavoidable exposure to environmental carcinogens complicates the investigation of cancer causes in human beings. The complexity of human cancer causes is especially challenging for cancers with long latency, which are associated with exposure to ubiquitous environmental carcinogens.

1.2 Pre-diagnosis techniques
Pre-diagnosis helps to identify or narrow down the possibility of screening for lung cancer disease. Symptoms and risk factors (smoking, alcohol consumption, obesity, and insulin resistance) had a statistically significant effect in pre-diagnosis stage.[4].Some of the approaches that has been used in a similar medical domain diagnostic/ pre-diagnostic environments are: Expert systems[5], computational intelligent methods for rule based understanding, Artificial Neural network based Learning

techniques[7], Genetic Algorithms[8],Fuzzy systems own material.

2. METHODOLOGY

The approach that is being followed here for the pre-diagnostic technique is based on systematic study of the statistical factors, symptoms and risk factors associated with Lung cancer. Initially the parameters for the pre-diagnosis are collected by interacting with the pathological, clinical and medical oncologists(Domain experts). Major problems that are faced in the early detection are: A large number of parameters, the variable and sometime unpredictable dependencies of these parameters in the development and growth of the Lung cancer. In order to develop a model there is a need to reduce the dimensionality. Hence preprocessing and filtering the data set is necessary.

After transforming and reducing the factors , a model is developed for training and testing. The model that has been developed is based on the Multi-layer Neural Network model. 1)ConfirmedLung cancer patient data and 2)confirmed lung cancer-negative data are taken as input and output parameters. As both input and output data is available and the input variables are continuous supervised learning approach is chosen to train the model. Training data is then applied to the developed model to get the various weight factors, under supervisory learning approach. This memory associative net neural network is trained to associate a set of input vectors with a corresponding set of output vectors. The process is repeated till consistent weight factors are obtained. Once the network is trained the model is used for the test data.

2.1 Statistical incidence factors:

i. Age-adjusted rate (ARR)
ii. Primary histology
iii. Area-related incidence chance
iv. Crude incidence rate

2.2 Lung cancer symptoms

The following are the generic lung cancer symptoms[3]

i. A cough that does not go away and gets worse over time
ii. Coughing up blood (heamoptysis) or bloody mucus.
iii. Chest, shoulder, or back pain that doesn't go away and often is made worse by deep Hoarseness
iv. Weight loss and loss of appetite
v. Increase in volume of sputum
vi. Wheezing
vii. Shortness of breath
viii. Repeated respiratory infections, such as bronchitis or pneumonia
ix. Repeated problems with pneumonia or bronchitis
x. Fatigue and weakness
xi. New onset of wheezing
xii. Swelling of the neck and face
xiii. Clubbing of the fingers and toes. The nails appear to bulge out more than normal.
xiv. Paraneoplastic syndromes which are caused by biologically active substances that are secreted by the tumor.
xv. Fever
xvi. Hoarseness of voice
xvii. Puffiness of face
xviii. Loss of appetite
xix. Nausea and vomiting

2.3 Lung cancer risk factors

a. Smoking:
   i. Beedi
   ii. Cigarette
   iii. Hukka
b. Second-hand smoke
c. High dose of ionizing radiation
d. Radon exposure
e. Occupational exposure to mustard gas, chloromethyl ether, inorganic arsenic, chromium, nickel, vinyl chloride, radon asbestos
f. Air pollution
g. Insufficient consumption of fruits & vegetables
h. Suffering with other types of malignancy

2.4 Categorization

Categorizing and finding the impact of these factors on the development of the cancer is a herculean task. Few of the factors could be expressed in binary logic format(presence or absence), whereas remaining factors cannot be directly expressed in Binary logic. Those factors which cannot be expressed in Binary logic could be represented in the form of fuzzy logic(like smoking risk factor is converted to number of smoke years and normalized). Fuzzy logic can take continuous values from 0 to 1 (after normalization). Fuzzy input can be converted to crisp output using following processes as mentioned in Figure 1.

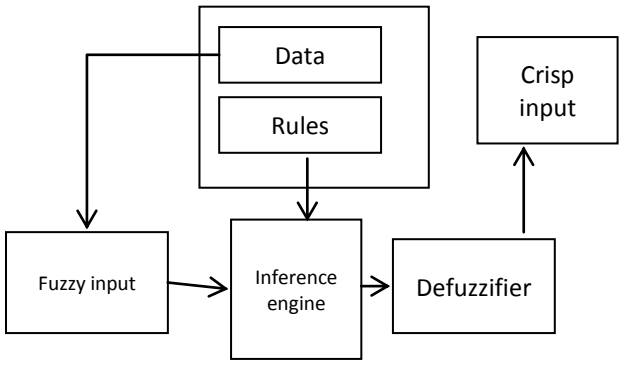

Figure 1- Fuzzy input to Crisp output

2.5 Developing a Model

Artificial Neural Network(ANN) these are essentially simple mathematical models defining a function f : X → Y . Each type of ANN model corresponds to a class of such functions. A widely used type of composition is the nonlinear weighted sum, where

$$f(x) = K\left[\sum_{i}^{n} Wi\ Gi\ (x)\right]$$

where K is some predefined function, such as the hyperbolic tangent. It will be convenient for the following to refer to a collection of functions Gi as simply a vector $G=(g_1,g_2 \ldots g_n)$.

Various processes of conceptualizing the model is represented in the following figure 2.

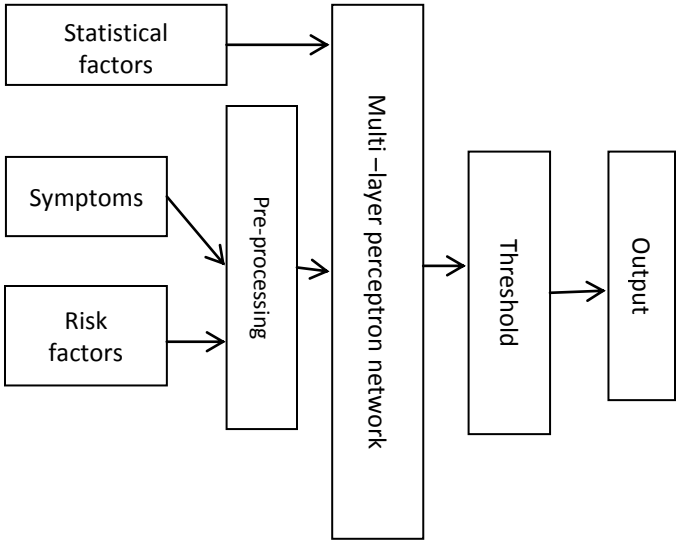

Figure 2 Process Model

Based on the above criteria multilayer neural network model was constructed. The schematic diagram of the same is shown in the figure 3. I, S and R are input layers, h1,h2..are hidden layers A, B, and D are output layers.

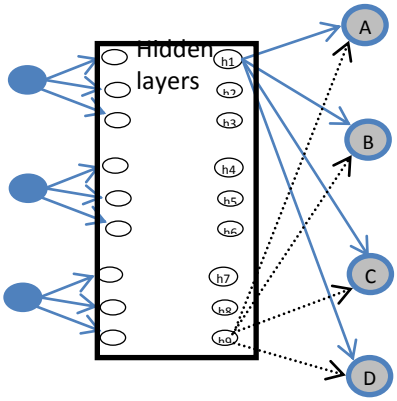

Figure 3 Multi-layer Perceptron Neural network

2.6 Supervised Learning approach

After constructing the model the next step is to frame the proper learning approach. Here the approach adopted is Delta learning rule as this rule supports for the continuous variable and it follows supervisory learning approach. For a single layer network with an output unit with a linear activation function, the output is, simply given by:

$$y = \sum_{i=1}^{n} w_i x_i + \theta$$

Such a simple network is able to represent a linear relationship between the value of the output unit and the value of the input units. By thresholding the output value, a classifier can be constructed , but here we focus on the linear relationship and use the network for a function approximation task. In high dimensional input spaces the network represents a (hyper)plane and multiple output units may be defined. Suppose we want to train the network such that a hyper plane, it is possible to train by giving set of samples consisting of input values x_p and desired (or target) output values d_p. For every given input sample, the output of the network differs from the target value d_p by (d_p-y_p)where y_p is the actual output for this pattern. The delta-rule now uses a cost- or-error-function based on these differences to adjust the weights. The error function, as indicated by the name least mean square(LMS), is the summed squared error. That is, the total error E is defined to be

$$E = \sum_{p} E^{p} = \frac{1}{2}\sum_{o=1}^{No}\left(d_o^p - y_o^p\right)^2$$

Where the index p ranges over the set of input patterns and Ep represents the error on pattern p. The LMS procedure finds the values of all the weights that minimize the error function by a method called gradient descent. The idea is to make a change in the weight proportional to the negative of the derivative of the error as measured on the current pattern with respect to each weight:

$$\Delta_p w_j = -\gamma \frac{\partial E^p}{\partial w_j}$$

Where γ is a constant of proportionality. The derivative is

$$\frac{\partial E^p}{\partial w_j} = \frac{\partial E^p}{\partial y^p}\frac{\partial y^p}{\partial w_j}$$

The delta rule modifies weight appropriately for target and actual outputs of either polarity

and for both continuous and binary input and output units.

The activation is a differentiable function of the total input, given by

$$y_k^p = \mathcal{F}(S_k^p)$$

To get the correct generalization of the delta rule we set

$$\Delta_p w_{jk} = -\gamma \frac{\partial E^p}{\partial w_{jk}}$$

The error measure Ep is defined as the total quadratic error for pattern p at the output units:$E^p = \frac{1}{2}\sum_{o=1}^{No}\left(d_o^p -\right.$

$y_o^p)^2$ where $d^p o$ is the desired output for unit o when pattern p is clamped.

This procedure constitutes the generalized delta rule for a feed-forward network of non-linearUnits. The system is then fed with training data in which both input and Output are known. The system runs through several epochs in order to stabilize the weight factors. Once the system is stabilized with non-variant weight factors, the system can be used for the test data.

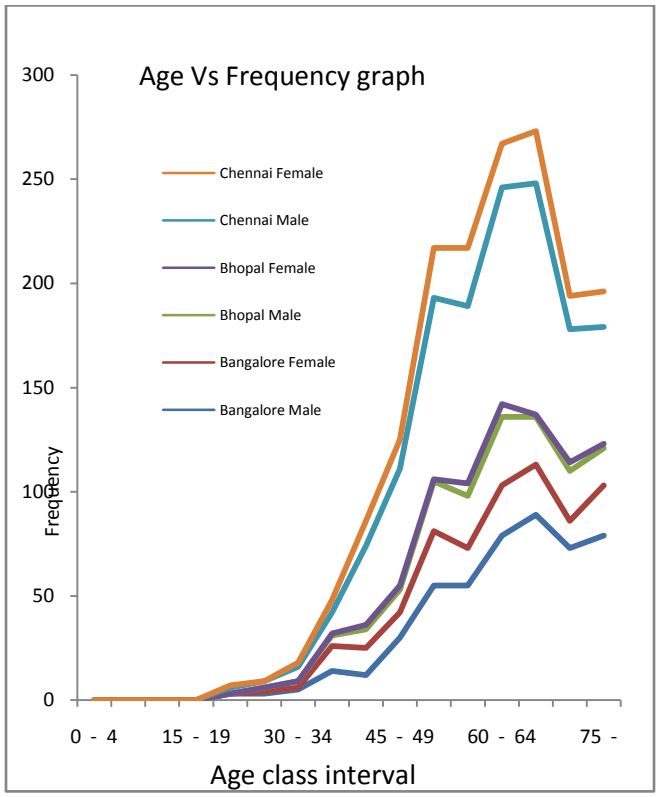

Fig.4 –Age-frequency related incident rate

The incident rate of lung cancer based on age group class interval of Bangalore, Bhopal and Chennai shown separately for male and female based on cancer registry between 2001-03 is shown in the plot (fig. 4). The above figure shown to indicate the need for adjusting the probability factor based on age, place and gender related variations when these factors are to be considered for network model.

3. RESULT AND ANALYSIS

This system is developed using MATLAB Neural network tool and it is based on the confirmed cases of Lung cancer patients of different stages. Output is mapped into four categories of threshold levels, more likely, likely, unpredictable, less likely. Training epochs is based on the consistency of the weight factors. This model works well for i) confirmed cases of Lung cancer patients of different stages and ii) confirmed non-lung cancer data. In the case of preliminary stages of Lung cancer the symptoms may not be present or noticeable, does not contribute much for the prediction. Hence the system has to be trained properly with proper mix of data. Many times the symptoms of the disease are revealed during the advanced stage of the cancer. Treatment and survival rate are very much depend on the stage of detection. The prediction rate, based on this system, in the established cases is high. However, in the absence of advanced clinical tools this method will enable to bring down the possible screening cases to minimum. Supervisory learning method may not be applicable in some circumstances when the output is not clearly known. It is planned to use reinforcement training model to overcome this problem. Reinforcement learning attempts to learn the input-output mapping through trial and error with a view to maximize a performance index called reinforcement signal. The system then knows whether the output is correct or not, but does not know the correct output.

4. CONCLUSION

Early detection of the cancer disease is crucial in diagnosing and treating the patient. The cure, metastasis (spreading of cancer disease to completely new location), recurrence (relapse), Remission(absence of all evidence of a cancer after treatment) and survival rate (Percentage of people who survive for a given period of time after treatment all directly attributable to the phase of detection of the cancer disease. Hence it is very essential that common man who has some symptoms and risk factors are better to undergo medical examination by a specialist at the earliest. Prevalence of Lung cancer disease is high in India, especially in rural India, did not get noticed at the early stage, because of the lack of awareness. Also it is not possible for the voluntary agencies to carry out the screening for all the people. The emphasis of this work is to find the target group of people who needs further screening for Lung cancer disease, so that the prevalence and mortality rate could be brought down.

5. ACKNOWLEDGMENTS

Our thanks to Naveen.B, Nithin.B, Annapurani, MV.Krishnan, Dr. Devuru, Surgical Oncologist, Dr. K. Ramachandra Reddy, Prof. & Head, epidemiology, Kidwai Memorial Institute of Oncology, faculty members of Post Graduate department of computer science, Christ university for their valuable suggestions.

6. REFERENCES

[1] 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

[2] Yongqian Qiang, Youmin Guo, Xue Li, Qiuping Wang, Hao Chen, & Duwu Cuic 2007 .The Diagnostic Rules of Peripheral Lung cancer Preliminary study based on Data Mining Technique. Journal of Nanjing Medical University, 21(3): 190-195

[3] Murat Karabhatak, M.Cevdet Ince 2008. Expert system for detection of breast cancer based on association rules and neural network. Journal: Expert systems with Applications

[4] ICMR Report 2006. Cancer Research in ICMR Achievements in Nineties

[5] Ta-Cheng Chen, Tung-Chow Hsu 2006. A GAs based approach for mining breast cancer pattern. Journal: Expert systems with Applications 30(2006) 674-681

[6] Petra Perner 1992. Mining Knowledge in X-ray images for Lung cancer diagnosis. Journal: Computer vision and applied computer Sciences

[7] W.Z.Liu, A.P.White, M.T.Hallissey, J.W.LFielding 1995. Machine learning techniques in early screening for gastric and esophageal cancer. Artificial Intelligence in Medicine 8(1996) 327-341

[8] Edward H.Shortliffe, A.Carlisle Scott, Miriam B.Bischoff, A.Bruce Campbell, William vanMelle, Charlotte D Jacobs ~1982 .Oncocin: An Expert system for Oncology protocol Management. Proceedings of the 7th IJCAI 1981

[9] Edward H.Shortliffe, A.Carlisle Scott, Miriam B.Bischoff, A.Bruce Campbell, William vanMelle, Charlotte D Jacobs 1981. An Expert system for Oncology protocol Management. Proceedings of the 7th IJCAI 1981 chapter 35 653-665

[10] James S.Gordon2008 .Mind-body, Medicine and Cancer. Journal: Hematology Oncology clinic N.America 22(2008) 683-708

[11] Kemal Polat, Salih Gunes 2008. Computer aided medical diagnosis system based on Principal Component analysis and artificial immune recognition system classifier algorithm. Expert systems with Applications 34(2008) 773-779

[12] Ira J. Kalet, Mrk Whipple. Silvia Pessah, Jerry Barker. Mary M. Austin Seymour, Linda G.Shapiro2002. A Rule based model for Local and Regional Tumor Spread. Journal: Artificial Intelligence in Medicine

[13] Astrid Pozet, Virginie Westeel, Pascal Berion, Arlettte Danzon, Didier Dabieuvre, Jean-Luc Beton, Alain Monnier, Jean Lahourcade, Jean-Charles Daphin, Mriette Mercier 2008.Rurality and survival differences in Lung cancer: A large population-based Multivariate analysis. Journal: Lung cancer (2008) 59, 291-300

[14] Ahmed Besaratinia, Gerd P Pfeifer 2008. Second-hand smoke and human lung cancer. Review article http://oncology.thelancet.com Vol 9 Jul-2008

[15] Maria Jose de Paula Castnha, Laecio Carvalho de Barros, Akebo Yamakami, Laercio Luis Vendite 2008. Fuzzy Expert system: An example in prostate cancer. Applied Mathematics and Computation 202(2008) 78-85

[16] Marko Bohanec, Blaz Zupan, Vladislav Rajkovic 2000. Applications of qualitative multi attribute decision models in health care. International Journal of medical Informatics 58-59 (2000) 191-205

[17] Maciej A.Mazrowski, Pior A Habas, Jacek Zurada, George D Tourassi 2008. Decision optimization of case based computer aided decision systems using genetic algorithms with application to mammography. Physics in Medicine and Biology 53(2008) 895-908

[18] Curtis P.Langlotz, Lawrence M. Fagan, Samson W.Tu, Branimir I.Sikic Edward H.Shortliffe 1987.A Therapy Planning Architecture that combines Decision Theory and Artificial Intelligence Techniques. Computers and Biomedical Research 20, 279-303 (1987)

[19] Maciej A.Mazrowski, Pior A Habas, Jacek Zurada, Joseph Y.Lo, Jay A.Baker, George D Tourassi 2008. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classifier performance. Neural networks 21(2008)427-436, Special Issue

[20] 20. WLodzizlaw Duch, Rudy Setiona, Jacek M.Zurada 2004. Computational Intelligence Methods for Rule-based data Understanding. Proceedings of the IEEE Vol-92 No. 5 , 771-805 May-2004

[21] Yanfeng Hou, Jacek M.Zurada, Waldemar Karwowski, William S.Marras, Kermit Davis 2007. Identification of key variables using Fuzzy average with Fuzzy cluster distribution. IEEE transactions on fuzzy systems vol.15 No. 4 Aug-2007 673-685

[22] Maciej A.Mazrowski, Jacek Zurada, George D Tourassi 2008. Selection of samples in case based computer aided decision systems. Physics in Medicine and Biology 53(2008) 6079-6088

[23] Alex L. Tay , Jacek M.Zurada, Lai Ping Wong and Jian Xu 2007. The Hierarchical fast learning artificial Neural Network (HieFLANN) - An autonomous platform for Hierarchical Neural Network construction. IEEE Transactions on Neural Networks Vol. 18 No. 6 Nov-2007 1645-1657

[24] Maciej Majewski, Jacek M.Zurada 2008. Sentence recognition using artificial neural networks. Knowledge based systems 21(2008) 629-635

[25] C-Q Zhu, W Shih, C-H Ling, M-S Tsao 2006. Immunohistochemical markers of prognosis in non-small cell lung cancer: a review and proposal for a multiphase approach to marker evaluation. Journal of Clinical Pathology 2006;59:790-800

[26] Consolidated report of population based cancer registries 2001-04, Incidence and Distribution of cancer, ICMR report, Bangalore.