Early Prediction of Lung Cancer using Data Mining Classification Techniques

Ms. P. B. Hole[1] Ms. K.B.Mane[2] Ms. P.M.Murkute[3] Ms. M.M.Panchwagh[4] Prof. S.R.Vasekar[5]
[1,2,3,4,5]Department of Computer Science & Engineering
[1,2,3,4,5]SMSMPITR, Shankarnagar-Akluj, India

Abstract— Medical data mining is one of the big issues in this modern world. Medical problems are often in each and every human being. Cancer is one of the very dangerous diseases a human can ever had in Lung cancer is one of them. Lung cancer is a disease that occurs due to the uncontrolled cell growth in tissues of the lung. It is very hard to detect it in its early stages as its symptoms appear only in the advanced stages. Goal of this paper is to automate the classification process for the early prediction of Lung Cancer diseases. To justify this research, it includes classification algorithm i.e. Neural Network and for optimization GA (Genetic Algorithm) is used. Evaluation would be done on the basis of accurate classified sample data. For testing and training diacom images has been used.

Key words: Data Mining, Lung Cancer, Classification, Artificial Neural Networks, Back propagation Neural Networks, Genetic Algorithm

I. INTRODUCTION

Lung Cancer is a noteworthy reason for Mortality in the western world as exhibited by the striking factual numbers distributed consistently by the American Lung Cancer Society. They demonstrate that the 5-year linger rate for patients with lung malignancy can be enhanced from a normal of 14% up to 49% if the ailment is analyzed and treated at its initial stage. Medicinal images as a vital piece of therapeutic determination and treatment were focusing on these pictures for good. These pictures include success of concealed data that misused by doctors in settling on contemplated choices around a patient. Then again, removing this magnitude shrouded data is a basic first stride to their utilization. This reason inspires to utilize information digging systems abilities for productive learning extraction & find hidden lung.

Mining Medical images includes numerous procedures. Medicinal Data Mining is a promising zone of computational insight connected to a consequently break down patients records going for the disclosure of new information Valuable for restorative choice making. Affected knowledge is expected not just to increment exact determination and effective infection treatment, additionally to improve security by diminishing blunders. The systems in this paper arrange the advanced X-beam midsection movies in two classes: ordinary and strange. The ordinary ones are those portraying a solid patient. The erratic ones incorporate Type of lung tumor; we will utilize a typical arrangement technique specifically SVMs neural systems.

II. LITERATURE SURVEY

A. History

In the past, Lung Cancer is a disease of uncontrolled cell growth in tissues of the lung. Cancer disease is considered as the killer disease and it is on the rise. There are many kinds of cancer disease found out which can affect most of the parts of human body. The world statics report reveals that Lung cancer is in the top most places of cancer related deaths. Deaths due to Lung cancer are about 1.4 million per year worldwide. Earlier diagnosis of Lung Cancer saves enormous lives, failing which may lead to other severe problems causing sudden fatal end. Its cure rate and prediction depends mainly on the early detection and diagnosis of the disease. One of the most common forms of medical malpractices globally is an error in diagnosis. In general, a measure for early stage lung cancer diagnosis mainly includes those utilizing X-ray chest films, CT, MRI etc. Knowledge discovery and data mining have found numerous applications in business and scientific domain. Valuable knowledge can be discovered from application of data mining techniques in healthcare system. In this study, we shortly examine the potential use of classification based data mining techniques such as Rule based, Decision tree, Naïve Bayes and Artificial Neural Network to massive volume of healthcare data.

III. EXISTING SYSTEM

In the existing model is to classify only by using the x-ray, CT scan for detect lung cancer, mining a detection of lung cancer, survey of the lung cancer patients based on the countries, Predict the lung cancer disease and analysis the lung cancer disease by using the different data mining Techniques.

A. Disadvantages of Existing System:

- Time consuming process.
- In many parts of the world widespread screening by CT or MRI is not yet practical.

IV. PROPOSED SYSTEM

Some techniques are essential to the task of medical image mining, Lung Field Segmentation, Data Processing, Feature Extraction, Classification using neural network and SVMs. The methods used in this paper work states to classify digital X-ray chest films into two categories: normal and abnormal. Different learning experiments were performed on two different data sets, created by means of feature selection and SVMs trained with different parameters; the results are compared and reported.

A. Advantages of Proposed System:

- Utilization of time management.
- Fast process.

V. ALGORITHM USED

A. KNN (K-nearest neighbor) algorithm

1) Determining Parameter K=number of nearest neighbour (number of tumor).
2) Calculate the distance between the query instance and the entire training sample.
3) Sort the distance and determine nearest neighbor based on the k th minimum distance.
4) Gather the category of the nearest neighbour.
5) Use simple majority of the category of nearest neighbour as the prediction value of the query instance.

VI. FLOW OF SYSTEM

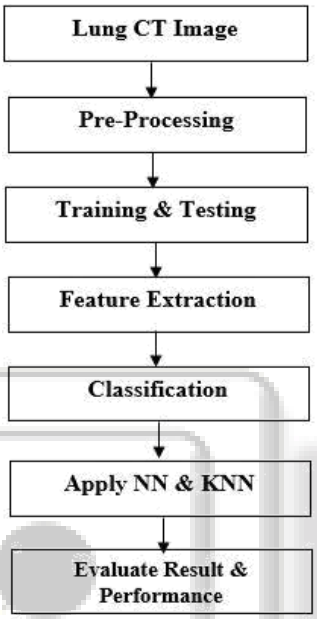

Fig. 6.1: flow of system

VII. MATHEMATICAL MODEL

System S as a whole can be defined with the following main components.

S= { I, P, T, F, BN, GA, Op}

S= System

I= Lung CT Image.

I= { I1, I2, I3, …, In}

P=Pre Processing- RGB image Converted to Gray Scale Image.

T=Training Samples & Testing Samples- Feed forward and feed forward back propagation neural networks are used for classification. The initial weights has to be chosen randomly and then training begins.

F= Feature Extraction- Attributes of an image are useful for knowledge extraction Geometrical features like Autocorrelation, contrast, cluster prominence, cluster shade, dissimilarity, energy, entropy, homogeneity, maximum probability, sum variance, sum entropy, difference variance, difference entropy and information measure.

A. BN= Backpropogation Neural Networks.

Backpropogation Algorithm:

1) Initialize weights randomly,
2) Present an input vector pattern to the network,
3) Evaluate the outputs of the network by propagating signals forwards,
4) For all output neurons calculate $\partial j = (yj-dj)$, where dj is the desired output of neuron j and yj is its current output: $yj= g(\sum_i w_{ij} x_i) = (1 + e^{-\sum_i w_{ij} x_i})^{-1}$, assuming a sigmoid activation function,
5) For all other neurons (from last hidden layer to first), compute $\partial j = \sum_k w_{jk} g'(x) \partial k$, where ∂k is the ∂j of the succeeding layer, and $g'(x) = y_k(1 - y_k)$,
6) Update the weights according to: $w_{ij}(t + 1) = w_{ij}(t) - \_ y_i y_j (1 - y_j) \partial j$, where, _ is a parameter called the learning rate.
7) Go to step 2 for a certain number of iterations, or until the error is less than a prespecified value.

B. GA= Genetic Algorithm.

Genetic Algorithm:

1) Initialize random population with time
2) Evaluate fitness function
3) Test for termination case
4) Initialize time counter
5) Select sub population
6) Select parents
7) Evaluate new fitness function
8) Agitate mated population
9) Select survivors from fitness function
10) End GA

Op= Evaluate .Results of Image Lung Cancer affected or not.

VIII. CONCLUSION

In this paper, we are going to use some data mining classification techniques such as neural network & SVMs for detection and classification of Lung Cancer in X-ray chest films. Due to high number of false positives extracted, a set of 160 features was calculated and a feature extraction technique was applied to select the best feature. We classify the digital X-ray films in two categories: normal and abnormal. The normal or negative ones are those characterizing a healthy patient. Abnormal or positive ones include types of lung cancer. We will use some procedures also Data Pre-processing, Feature Extraction etc. In this paper we well use classification methods in order to classify problems aim to identify the characteristics that indicate the group to which each case belongs.

ACKNOWLEDGMENTS

I would like to thank the anonymous refers for their helpful guidance that have improved the quality of this paper. Also I would like to thank our Guide Prof. S.R.Vasekar and Coordinator Prof. S. D. Pandhare for their valuable guidance.

REFERENCES

[1] Zakaria Suliman Zubi and Rema Asheibani Saad, "Uing Some Data Mining Techniques for Early Diagnosis of Lung Cancer," Recent Researches in Artificial Intelligence, Knowledge Engineering and Data Bases, Libya, 2007.
[2] Paola Campadelli, Elena Casiraghi, and Diana Artioli, "A Fully Automated Method for Lung Nodule Detection From Postero-Anterior Chest Radiographs,"

In Proc. of IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. 25, NO. 12, DECEMBER 2006.

[3] Jaba Sheela L and Dr.V.Shanthi, "An Approach for Discretization and Feature Selection Of Continuous-Valued Attributes in Medical Images for Classification Learning," International Journal of Computer Theory and Engineering, Vol. 1, No.2, June 2009.

[4] V.Krishnaiah, Dr.G.Narsimha, Dr.N.Subhash Chandra. 2013, "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques," International Journal of Computer Science and Information Technologies, Vol. 4 (1), 2013, 39 – 45.

