ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

A Survey on Early Diagnosis of Lung Cancer using Classification Data Mining Technique

Pricilla.S, Pebila Shani.S

PG Scholar, Dept. of Computer Technology, Sri Krishna Arts and Science College, Coimbatore, Tamil Nadu, India

ABSTRACT: Cancer is one of the most common diseases in the world that results in majority of death caused by uncontrolled growth of cells in any of the tissues or parts of the body. For early prevention and detection of the breast cancer patients, data mining algorithm is used. Detection of Lung Cancer in its early stage is the key of its cure. Lung Cancer is a major cause of death in the world as established by the striking statistical numbers published every year Medical Data Mining is a hopeful area of computational intelligence applied to detect lung cancer in its early stage Induced knowledge of lung cancer is predictable not only to increase accurate diagnosis and successful treatment, but also to improve safety by reducing errors. The classification data mining technique is a key for the diagnosis of lung cancer. To diagnosis in early stage takes two classification data mining techniques.

KEYWORDS: Data mining, Classification, Artificial neural networks, Support Vector Machine.

I.INTRODUCTION

Cancer is a leading cause of death worldwide it is caused by unrestrained growth of cells in any of the tissues or parts of the body. Early detection of cancer to a great extent increases the chances for successful treatment lung cancer is the second most widespread cancer in both men and Women. Two types of lung cancers are NSCLC and SCLC, the malignant tumour develops when cells in the lung tissue divide and grow without the normal controls on cell death and cell division. This method used in this paper work states to classify digital X-ray chest films into two categories: normal and abnormal. The prediction condition is based on image mining related to the lung cancer. the three main key to the task of medical image mining, Lung Field Segmentation, Data Processing, Feature Extraction, Classification using neural network and SVMs. Data Mining techniques are implemented together to create a novel method to diagnose the existence of cancer for a particular patient

II.RELATED WORK

In [1] proposed work states to classify digital X-ray chest pictures into two kinds. Diverse learning experiments were performed on two different data sets, produced by means of feature selection and SVMs trained with different parameters. In [2] the future work presents a fully auto-mated system processing digital chest radio-graphs that starts by producing an accurate segmentation of the lung field area. In [3] the work provides a Computer Aided Diagnosis System (CAD) for early detection of lung cancer nodules from the Chest Computer Tomography (CT) images. In [4] focuses on SVM can be used for classifying the medical data because of its simplicity. Real time lung images are taken for the study. Lung images are segmented to retrieve the region of interest and these regions or nodules are used for classification. In [5] the proposed work is based on a combination of a segmentation method and an analytical method and aims to improve these two methods in order to develop an interface that can assist dermatologists in the diagnostic phase. In [6] a study is presented in which an analysis is provided for the diagnosis and prediction of lung cancer at initial stage using Image Processing and Data Mining techniques.

III. PROBLEM STATEMENT

Lung Cancer is the second leading cancer being developed globally. Purely 16% of lung cancer cases are diagnosed at an early stage. The five-year endurance rate is only 4 percent, statistically, the 5-year survival rate for patients can be

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

improved from an average of 4% up to 49%. Near the beginning detection increases the chances for winning treatment. Hence, apart from remedial solutions some Data mining solution needs to be incorporated for resolving the death causing issue.

IV. DATA MINING TECHNIQUES

Data mining technique involves the use of sophisticated data study tools to make out formerly unknown, convincing patterns and contact in large data set. Support Vector Machine (SVM) is a differentiate classifier properly defined by an unravelling hyper plane. In other words, given labelled training data (supervised learning) artificial neuron network (ANN) is a computational model based on the structure and functions of biological neural networks. Information that flows all the way through the network affects the structure of the ANN for the reason that a neural network changes - or learns, in a sense - based on that input and output

V. METHODS AND TECHNIQUES

A. *Lung Field Segmentation:*

Segmentation methods embrace the concealed parts in the lung area and avoid assumptions regarding chest position, size and orientation. It works with images where the chest is not always positioned in the central part of the images may be skewed and may have structural abnormalities. This algorithm detects the most perceptible lung edges by means of the first derivatives of Gaussian Filters taken at 4 different point of reference. The edges detected provide an initial outline of the lung borders.
Plung is the initial point for an edge-tracking practice that works on 3 images in place of the chest at 3 different levels. These methods generate segmentation Mask where concealed lung areas are excluded. Once the segmentation mask has been distinct, a further method has been urbanized to find the separation between the hidden and the visible lung areas

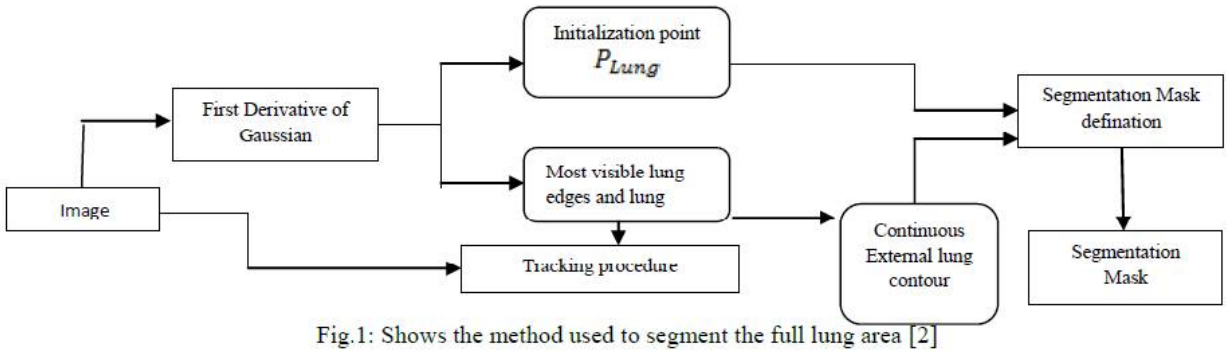


Fig.1: Shows the method used to segment the full lung area [2]

B. *Data Preprocessing:*

Pre-processing part of the images is necessary to advance the quality of the images and formulate the feature extraction phase more consistent. This stage consists of some processes. These processes contain Data Normalization, Data Preparation, Data Transformation, Data Cleaning and Data Formatting. Normalization techniques are vital to combine the different image formats to a regular format. Data Preparation modifies images to present them in a suitable format for transformation techniques. The image will transform in turn to obtain a squashed presentation. Segmentation completed to identify regions of interest (ROI) for the mining task regularly achieved using classifier systems

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

C. *Feature Extraction:*

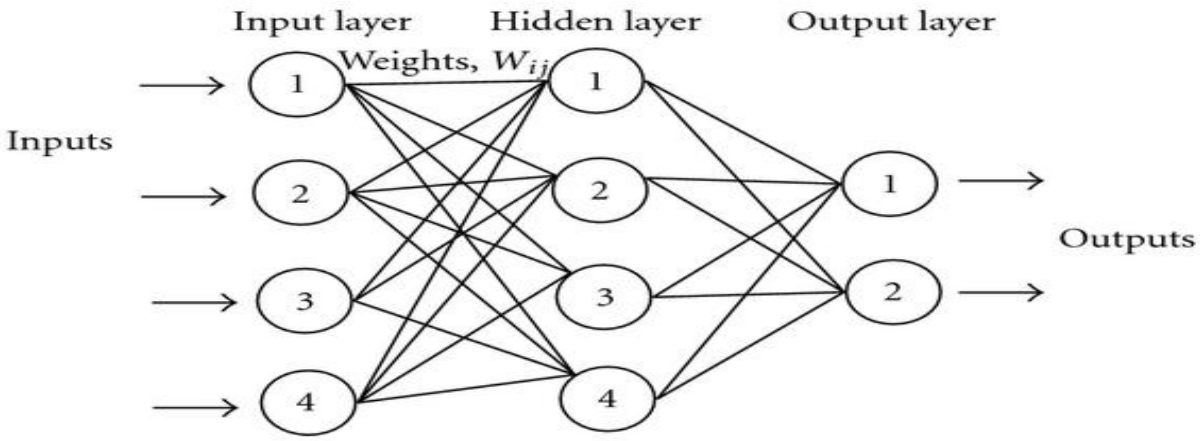
Images usually have a massive number of aspects. It is important to identify and extract remarkable features for an challenging task in order to decrease the difficulty of processing. Not all the attributes of an image are constructive for knowledge extraction. Image processing algorithm routinely extracts image attributes such as local colour, global colour, texture, structure. The extraction of the features from an image can done using a diversity of image processing techniques. Based on this, the image is processed to look for a capacity that helps in selecting the pixels that communicate to the centres of the nodule. We concentrate on the extraction process to very small regions in order to make sure that we capture all areas.

D. *Classification:*

In newest days, many complicated classification approaches, such as neural networks, expert system and SVM have been broadly applied for image classification. In most cases, image classification approaches grouped as supervised & unsupervised machine learning approaches or parametric and non-parametric or hard and soft classification. The most used non-parametric classification approaches are neural networks, support vector machines & expert systems. Parametric classifier are sturdiness and easy to contact for any image-processing software.

VI. ARTIFICIAL NEURAL NETWORK

An artificial neural network is a mathematical illustration based on organic neural networks. It consists of an steady group of artificial neurons and processes information using a connectionist move in the direction of computation. Neurons are prepared into layers. The input layer consists purely of the inventive data, even as the output layer nodes exemplify the classes. Then, there may be copious hidden layers. A key emphasize of neural networks is an iterative learning process in which data samples are unfilled to the network one at a illustration, and the weights are proverbial in order to forecast the correct class label. Recompense of neural networks includes their high indulgence to noisy data, as well as their propensity to classify patterns on which they have not been skilled. A review of advantages and disadvantages of neural networks in the perspective of microarray analysis exists , the planning of the neural network has three layers such as input layer, hidden layer and output layer. The nodes in the input layer connected with a number of nodes in the hidden layer. Each input node attached to each node in the hidden layer. The nodes in the hidden layer possibly will fix to nodes in another hidden layer, or to an output layer. The output layer consists of one or more answer variables.A main alarm of the training phase is to focus on the core weights of the neural network which familiar according to the connections used in the learning process. This perception drives us to transform the interior weights while trained neural network worn to classify new images.



ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

VII. SUPPORT VECTOR MACHINE (SVM)

SVM is introduced by Cortes is habitually used for categorization purpose. SVMs area unit inexpensive learning approaches for guidance classifiers supported many functions like polynomial functions, radial basis functions, neural networks etc. it's reflection about as a supervised learning move toward that produces input -output mapping functions from a labelled training dataset. SVM has imperative psychological capacity and thus is roughly dialect applied in pattern recognition. SVMs area unit collective approximates that rely upon the functional math and optimizing theory. The SVM is chiefly placing the biological investigation and competent to handle noise, massive dataset and enormous input areas.
The essential plan of SVM may be represented as follows:
A. primarily, the inputs area unit developed as feature vectors.
B. next, by victimization the kernel perform, these feature vectors area unit mapped into a feature house.
C. lastly, a division is computed within the feature house to separate the categories of training vectors

SVM classifier

For dual classification SVM determines a finest Separating Hyper plane (OSH) that produces a most fringe between 2 classes of knowledge. To form an OSH, SVM maps facts into a better dimensional quality house and carries out this nonlinear mapping with the backing of a kernel operate. Then, SVM builds a linear OSH among 2 categories
of knowledge within the higher feature house. Knowledge vectors that four-sided figure measure quicker to the OSH within the higher feature house four-sided figure measure referred to as Support Vectors (SVs) and squeeze all knowledge necessary for cataloguing. A kernel operates and also the parameters ought to be influential for constructing the support vector machine classifier. Here, 3 kernel functions four-sided figure measure adapted construct SVM classifiers:
a. Linear kernel operate
b. Polynomial kernel operate
c. Radial basis operate
The most worn kernel operate for SVM is Radial Basis operate (RBF) owing to their restricted and finite responses transversely the complete vary of real coordinate axis. The classification correctness of RBF kernel was high; in addition, the bias price and also the error rate of RBF kernel were little compared to dissimilar kernels.

VIII. CROSS-VALIDATION

Cross-validation could be a system for associate degree analysing however the consequences of a functional math analysis can take a broad view to an freelance information set. It‘s used in substance; wherever the target is prediction, and to approximate however a closely prognosticative model can act upon apply. One globular of cross-validation includes dividing or partitioning a trial of information into harmonizing subsets, in performance the analysis on one set (training set), and guarantee the analysis on the opposite set (validation set or testing set). Multiple rounds of cross-validation are performed persecution totally different partitions to scale back variability, and also the validation results are averaged over the rounds

Applying the feature selection technique, we preferred 36 features to figure the data set & its subset i.e. 18 features dataset. Since the numeral of true positives extracted (151) is much subordinate than the number of false positives (18916), both the 36 features and the 18 features Data Sets are very disturbed ; this shows that the positive-enriched data sets that were built for training and testing the SVMs. For all data set, we discretely considered the true positive and the false positive (negative) examples, and we arbitrarily split the existing positive data into 136 examples for training and 15 examples for testing, according to a train/test ratio of 9/1. From the set of negative data, we extracted devoid of replacement a number of negative examples equivalent to 30 times the number of positive data obtaining, correspondingly, 136 × 30=4080 negative examples for the training set and 15 × 30=450 for the test data. To test system recital when the number of positive examples used for training decreases, we ran additional tests via a train/test

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

ratio equal to 7/3, i.e., 106 examples for training and the residual 45 for testing. In this case, we had 106 × 30=3180 in the training set & 45 × 30=1350 in the test set. This method was recurring 10 times, obtaining 10 pairs of training and test sets. By means of the two data sets (the 36 features data set and the 18 features data sets), and the two ratios, 9/1 and 7/3, for splitting the positive examples into the train/test sets, we ran four experiments for all SVM.

S.no	Data Items (Features)	Type of Data
1	19	shape and position of the region.
2	16	grey level distribution of the region pixels
3	6	radius value for candidate region
4	11	Compute a: 11 different scales
5	108	Compute by Gaussian filters

IX. CONCLUSION AND FUTURE WORK

In this paper, we are departing to use a quantity of data mining classification techniques such as neural network & SVMs for detection and classification of Lung Cancer in X-ray chest films. Due to immense number of false positives extracted, a set of 160 features was deliberate and a feature extraction technique was functional to select the preeminent feature. We grade the digital X-ray films in two categories: normal and abnormal. The normal or negative ones are those characterizing in good physical shape patient. Abnormal or positive ones embrace types of lung cancer. We will use various procedures like Data Pre-processing, Feature Extraction etc. In this paper we well use classification methods in sort to classify problems endeavour to identify the uniqueness that specify the group to which every case belongs. Still a lot of bare space is available for future development and new techniques are generated and further techniques are required to recognize a notorious feature of the X-ray which is really important.

REFERENCES

[1] Zakaria Suliman Zubi and Rema Asheibani Saad, "Using 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]Ms. Swati P. Tidke, Prof. Vrishali A. Chakkarawar ―Classification of Lung Tumour Using SVM‖ (IJCER) Vol. 2 Issue.5.
[4]M.Gomathi, Dr. P. Thangaraj―An Effective Classification of Bening anf Malignant Nodules Using Support Vector Machine‖ (JGRCS)Vol 3 , No. 7, July 2012
[5]. Nadia Smaoui, Souhir Bessassi," A developed system for melanoma diagnosis",2013, International Journal of Computer Vision and Signal Processing, 3(1), 10-17(2013).
[6]. Teresa Mendonca, Pedro M. Ferreira,Jorge S. Marques, Andre R. S. Marcal, Jorge Rozeira," PH2 – dermoscopic image database for research and benchmarking",2013, 35th Annual International Conference of the IEEE EMBS Osaka,Japan, 3 - 7 July, 2013.