



International Journal for Innovative Engineering and Management Research

A Peer Reviewed Open Access International Journal

www.ijiemr.org

COPY RIGHT



ELSEVIER
SSRN

2019 IJIEMR. Personal use of this material is permitted. Permission from IJIEMR must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. No Reprint should be done to this paper, all copy right is authenticated to Paper Authors

IJIEMR Transactions, online available on 23rd Dec 2019. Link

[:http://www.ijiemr.org/downloads.php?vol=Volume-08&issue=ISSUE-12](http://www.ijiemr.org/downloads.php?vol=Volume-08&issue=ISSUE-12)

Title: **DETECTION OF LUNG NODULES IN EARLIER STAGE BY USING SUPPORT VECTOR MACHINE (SVM) MACHINE LEARNING CLASSIFIER ALGORITHMS**

Volume 08, Issue 12, Pages: 194–198.

Paper Authors

V.SURENDRA REDDY, T.RAJYALAKSHMI



USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per **UGC Guidelines** We Are Providing A Electronic Bar Code

DETECTION OF LUNG NODULES IN EARLIER STAGE BY USING SUPPORT VECTOR MACHINE (SVM) MACHINE LEARNING CLASSIFIER ALGORITHMS

V.SURENDRA REDDY¹, T.RAJYALAKSHMI²

1. Professor, Dept of CSE, Sri Venkateswara College of Engineering & Technology, A.P, India

2. Assistant professor, Dept of CSE, University College of Narasaraopet (JNTUN), A.P, India

ABSTRACT

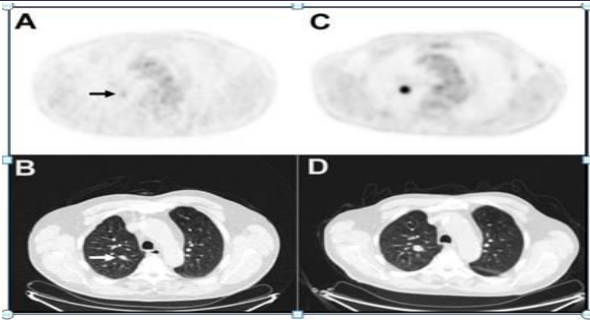
Now a day the lung cancer has been the most dangers death list disease. According to the recent survey of the medical research the lung cancer has highest rate of death compare to the all other cancers. We can reduce the death rate of lung cancer by early detection. By this earlier detection of the lung cancer we can improve the chances to survival. We can identify the lung cancer by using nodules which are attached to the walls of the lungs. These nodules will play key role to identify the lung cancers. The main aim of the project is to building of machine learning technique which is capable of classifying the lung nodule as benign or malignant. We took a data set from the LUNG_{xCT} 2015. These data sets consist of 83 lung nodule images. That ct images are cropped in and divide in to the nodule images and non- nodule images. And that nodule images are convert in to local binary pattern and co –occurrence matrix and we extracted the values .in this project we used support-vector machine algorithm for classification with 8 number of folds.

Index terms: nodules, CT, support vector machine, lungx2015.

I. INTRODUCTION

Cancer is nothing but growing of abnormal cells in body which they spread to other parts of body. It also defined as the group of diseases, which harm full to the human body and life also. in present days we have over 100 types of cancers which effect human lung nodules can be come under cancer but most of the lung nodules are not come under cancer in (benign).every human body have small masse of tissue in lungs this things are quite common. Nodules appear as round, white shadows on chest will see in x-ray and ct scanning.According to recent medical survey the lung nodule cancer is second most cancer in the world which causes the death to the men and women. The survival rate of the lung nodule cancer is near5% for over a period of the last five years we can reduce the death rate of the lung

cancer by detecting the cancer in the earlier stage .according to the survey only 20% of lung cancer cases are diagnosed in their early stage. The early detection we play the key roles in detection and accurate classification between the metastatic and non-metastatic types These diagnose will be done by using computed tomography. The computed tomography will give the image analysis to identify the lung cancer to radiologists. Depending anomalous globular tissue radiologist can classify the cancer either it is benign or malignant by its characteristics. But coming to very early stage of cancer it is too difficult to identify the nodule cancer. Because of very small tumors the radiologist will get mental burden due to examining many images over the course of the day.



Nodule in CT scanning

Due to the high competition world the computers are playing key role in assisting radiologist to diagnosis the diseases. Now a day we are using CAD machine to identify the nodule cancer in very earlier stage and improve the accuracy. The CAD will give automated approach with the machine learning .it analyzes the huge amount of data and it will assist the radiologist to predict the disease from this. The radiologist will spend more time to communicate with the patient. The CAD system is more efficient and it will work faster than the human eye, and when we combine with the machine learning it will give great potential to the radiologists.

Related work

The CAD (computer aided diagnostic system) is used for the detection of the lung cancer in the very early stage by using the computed tomography scans. The CAD is not an authorized device to identify but it will do with the association of machine learning technique. It will assist the location of the suspicious elements to increase the general sensitivity of the system.

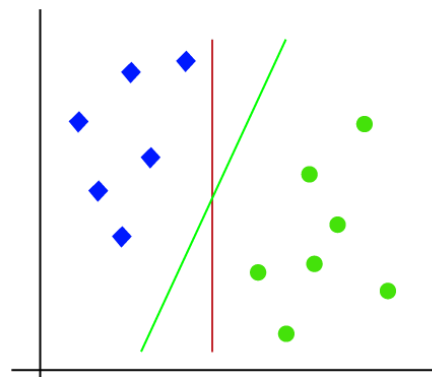
Computerized tomography images

CAD can be further improved for pulmonary nodule detection and assertive and quantitative evaluation through vessels characterized by the use of CT. Most cancers are an unusual cause of

cancer death worldwide. The subsequent duration of a computerized tomography experiment of tens of hundreds and many humans can be a fundamental task for radiologists to interpret. To fill this gap, computer-assisted detection algorithms can also be the most promising solution

Proposed methodology

In this proposal system we used SVM- support vector machine as classifier algorithm. The main function of SVM algorithm is to classification of the entities on base of their nature and characterization. In our algorithm we have two types of classifiers one is probabilistic classifier and second one is non-probabilistic classifier. Navies'bayes classifier comes under probabilistic classifier and svm classifier will come under non probability classifier. The main function of svm classifier is to separate the data across a decision boundary determined by a small subset of the data frame, data will be in two forms feature vectors. The data divide by the plane is known as hyper plane the sub set which will support the decision boundary is called as the support vectors.



svm model visualization

Support vector machine will divide the data in to two classes. That data object should have the two classes one is features variables $\{x_1, x_2, \dots, x_n\}$

and second is class variable Y_i . The svm will consider as each data point as a object in feature space and compare the one class to other class. The classifier compare the feature with the one class if it is ok then it add in to one group $Y_i=1$.if the feature not match with the class then it will go with another group $Y_i=-1$. the data will defined as the

$$\text{Data}=\{(x_i,y_i)|x_i \in \mathbb{R}^p,y_i \in (-1,+1)\}_{i=1}^n$$

In the support vector machine we used K-fold cross validation for separating of classes. And the main parameters of classifier are calculated as below.

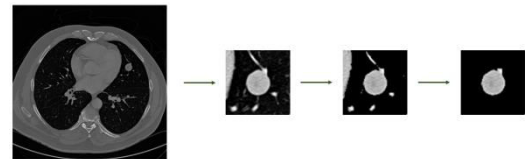
Sensitivity: $TP = \frac{TP}{(TP+FN)}$

Specification: $TN = \frac{TN}{(TN+FP)}$

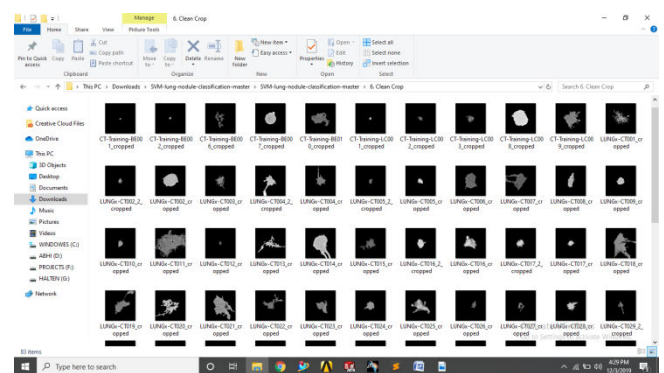
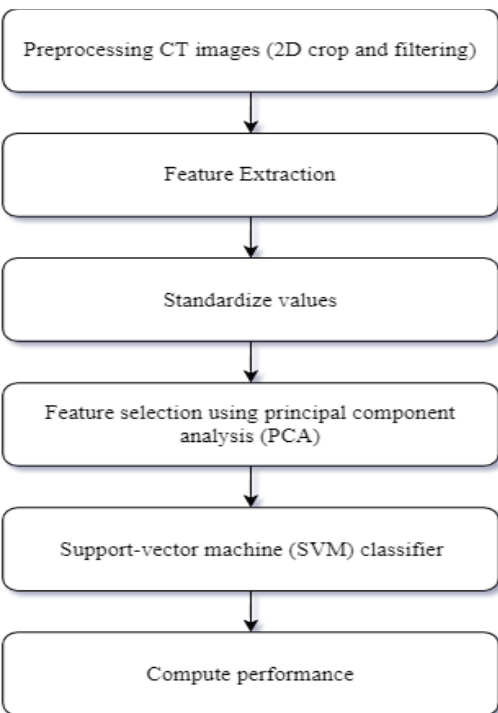
Accuracy= $\frac{(TP+TN)}{(TP+FP+TN+FN)}$

a) Processing of CT Image.

Processing of CT images are first step in the project in the step we got the data of ct images from the LUNG_xCT data set then we cropped the image by using cropping python program. we crop the image into 64X64 pixel individually. After that we separate the data part of the nodule by using the manual thresholding python script GIMP, an open-source image editing software was used to completely remove any remaining non-nodule bodies. Any pixel that did not contain the lung nodule were assigned a value of 0 (black). The preprocessing step is illustrated in



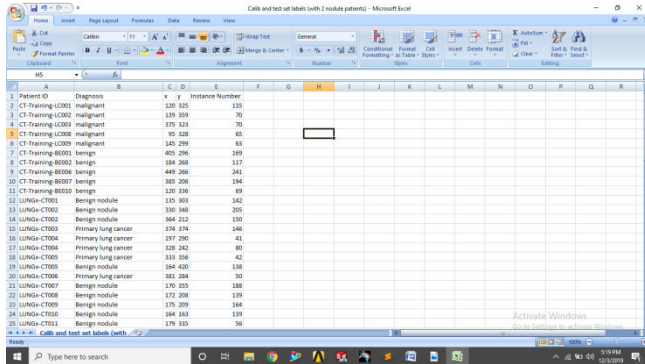
Isolating the lung nodule from other lung structures.



Clean cropped images

b) Feature extraction after competition of the ct image processing we entered into the feature extraction. The feature are extracted from the images in to data frames .every row of the data frame represents each nodule and each

column represents the extracted features. From each image six fixtures were extracted. And we removed the mean, and we can reduce the bias in the data by scaling the unit variance.



Patient ID	Diagnosis	Instance Number
1	CT-Training-LC001 malignant	120 325
2	CT-Training-LC002 malignant	130 309
3	CT-Training-LC003 malignant	170 323
4	CT-Training-LC008 malignant	50 328
5	CT-Training-LC005 malignant	160 299
6	CT-Training-BC001 benign	405 296
7	CT-Training-BC002 benign	164 268
8	CT-Training-BC006 benign	408 266
9	CT-Training-BC007 benign	380 206
10	CT-Training-BC008 benign	120 316
11	LU-Nodule-CT001 benign nodule	110 303
12	LU-Nodule-CT002 benign nodule	160 349
13	LU-Nodule-CT003 benign nodule	368 212
14	LU-Nodule-CT004 benign nodule	374 374
15	LU-Nodule-CT006 Primary lung cancer	100 290
16	LU-Nodule-CT004 Primary lung cancer	138 242
17	LU-Nodule-CT005 Primary lung cancer	164 200
18	LU-Nodule-CT006 Primary lung cancer	164 200
19	LU-Nodule-CT007 benign nodule	120 325
20	LU-Nodule-CT008 benign nodule	172 208
21	LU-Nodule-CT009 benign nodule	170 208
22	LU-Nodule-CT010 benign nodule	164 183
23	LU-Nodule-CT011 benign nodule	178 335

Image data in excel sheet as data frame

c) Principal component analysis(PCA)

After getting the data frame we can go with the principle component analysis to reduce the issue of the over fitting as well as to increase the ability to classify the new data. It will be done by the feature selection process. The principle component analysis which projects the data in to linearity and un-correlative vectors called principal components. The principal component analysis the first component vector has the maximum variance as well as maximum possible variance, but the component is to perpendicular to the last component vector. By using the principal component analysis module in python we can perform the operation

SVM classifier algorithm

the svm algorithm was developed by the Vladimir- vapnik in the year of 1999's.it is a supervised algorithm which is used to the classification of two classes in this project in the first step we got images from the ct scanning that ct scanning image data processed to the feature extraction the we did the principal component

analysis. After competition of the PCA process we get in to the support vector machine.

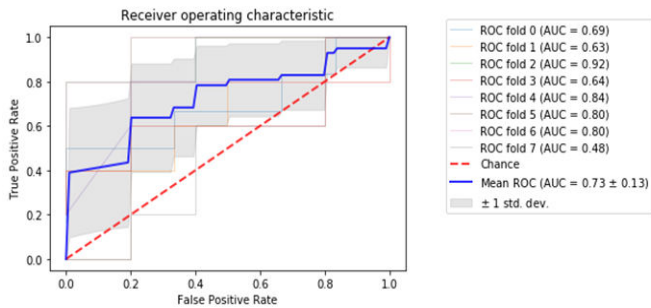
Svm takes the data which may not be linearly separable at the onset and projects the data into a higher dimensional space. The data will be classified by the hyper plane.

the data frame will shown with the knowledge of truth table of the nodules false is denote with the 0 that shows that benign, truth denote with the 1 that show the malignant, and the data set is split in to the two set one is training set testing set And we split the data in to k-folds by using the k-fold cross Validation .The main function of the k –fold cross is the it validate the data and it will split the data. next the svm will be trains on the k value of 1 in training fold and test the remaining folds. then the data of the testing fold will move to end of the each and every data until the data get the chance to be represents the testing fold .We can calculate the accuracy of the each iteration as an output. The range of the k will varies in between the 1-20 to see the results at highest accuracy .to obtain the average mean accuracy rate we fixed the k value as 8 for the svm classifier .to study the and implement the svm classifier we used sklearn module in python.

Results

in our project we did stratified k-fold cross validation to build the support vector machine algorithm with the 8 number of folds i.e.=8 and we took 12 percent of an data set as testing and remaining 88 percentage as a training set i.e. 10lung nodules as testing and 73 nodule as training. The below fig show the ROC and AUC for the each fold .the mean AUC for the 8 fold was 73%(13). To demonstrate how stable the model is, 1000 ROCs were generated, and their

mean values were averaged together, resulting in 72.3% as the mean.



K-fold validation results

Conclusion:

the support vector machine (SVM) was able to beat all the methodologies which was mentioned in the paper by armato III(2016).it will be use to make the CAD systems as the more accuracy system to identify the lung nodule cancer in very beginning stage then we can increase the survival rate of the lung nodule cancer patients. The publicly provided data lung nodule data hosted by The Cancer Imaging Archive remains to be a highly regarded tool for researchers in the radiomics field.

REFERENCE

[1] Key Statistics for Lung Cancer. (2018). Retrieved from <https://www.cancer.org/cancer/non-small-cell-lung-cancer/about/key-statistics.html>

[2] Lung Cancer Fact Sheet. (n.d.). Retrieved from <http://www.lung.org/lung-health-and-diseases/lung-disease-lookup/lung-cancer/resource-library/lung-cancer-fact-sheet.html>

[3] Gillies, R. J., PhD, Kinahan, P. E., PhD, &Hricak, H., MD, PhD, Dr(hc). (2015). Radiology, 563-577. doi:10.1148/radiol.2015151169

[4] Cancer Imaging Archive Wiki. (n.d.). Retrieved from [https://wiki.cancerimagingarchive.net/display/Public/LUNGx SPIE-AAPM-NCI Lung Nodule Classification Challenge](https://wiki.cancerimagingarchive.net/display/Public/LUNGx+SPIE-AAPM-NCI+Lung+Nodule+Classification+Challenge)

[5] Armato, S. G., Drukker, K., Li, F., Hadjiiski, L., Tourassi, G. D., Engelmann, R. M., ... Clarke, L. P. (2016). LUNGx Challenge for computerized lung nodule classification. *Journal of Medical Imaging*, 3(4), 044506. doi:10.1117/1.jmi.3.4.044506