Rough Set Theory based feature selection and classification of lung nodules in CT images

Maria Jenifer.L[1], T.Sathiya[2], B.Sathiyabham a[3]
PG Scholar[1], Asst.Professor[2], Professor[3], Department of Computer Science and Engineering[1,2,3],
Sona College of Technology, Salem, India[1,2,3]
Email:alkafionaeden@gmail.com

Abstract— Medical Image processing techniques are extensively used in medical field for earlier detection of diseases. Computer Aided Diagnosis (CAD) systems is generated to collect the reluctant information to analyze and evaluate in short time period. In this system 2D lung CT images are taken as an input and processed using Image processing techniques like image acquisition, image preprocessing, image segmentation, feature extraction, feature selection and classification. For image acquisition stage 2D lung CT images are collected from LIDC/IDRI. Next stage is image preprocessing, in this stage the best low pass filtering technique is considered for removing the noise and for enhancing the image using median filter. Image segmentation and feature extraction is a precondition for image recognition and provide information for CAD. For segmentation, morphological open operation is performed based on the ROI. One of the important parts in the proposed method is to select the optimal features from the extracted features. For feature extraction, the static method called Gray Level Co-occurrence Matrix (GLCM) is used. Feature Reduction technique is done using optimization algorithm called Rough set. Result of the classifier gives, whether the Lung CT Image is a benign or malignant at early stage and to avoid serious I and II stages for lung cancer patients.

Keywords— Computer Aided Diagosis, Computed Tomography, canny, GLCM,Rough set, Region of Interest (ROI).

———————————— ◆ ————————————

1 Introduction

The aim of the project is to design a CAD for the lung cancer patients who are severely suffering from the deadliest diseases in world. Through this system we can help to increases the human survival rate. Main purpose of this system is to detect the disease in the earlier stage. According to the study on lung cancer, it tells that the growth of abnormal cell causes cancer, based on the abnormality of the cell appearances type of cancer can be identified. There are four stages of lung cancer, in which first two stages is difficult to detect. Only at last two stages it is been diagnosed. Lung cancer in mainly caused due to tobacco smoking and exposure to secondhand smoking. Cancer is being confined at the lung nodules, lung nodules are usually about 0.2 inch to 1.2 inches. A larger lung nodule, such as one that's 30 millimeters or larger is more likely to be cancerous. It appears as round, white shadows on the chest X-Ray or CT scans. Lung nodule itself does not often cause any symptoms. Nodules ate not large enough to interfere with breathing. For example, if a lung nodule is due to lung cancer, symptoms may include: shortness of breath, chest pain, coughing up blood, back pain and weight loss. The nodules will be shown clearly in the CT scan than X-Rays. After a lung nodule is discovered, the doctor will look at its size, shape and general appearance. Certain features can tell whether it's benign or malignant. There are many methods to diagnose at the later stages like biopsy, bronchoscopy and needle biopsy. There are few issues in giving such diagnoses. A small lung nodule can be difficult to biopsy and there are risks, such as bleeding or a collapsed lung. To solve this kind of problem we developed this CAD system to help the patients who are affected.

The further sections in this paper are as follows, Section 2 includes related work, Section 3 includes proposed work, and Section 4 contains conclusion and future work has been discussed.

2 Related Work

Taruna Aggarwal, Asna Furqan, Kunal Kalra,[9]. proposed a CAD system for detection and classification of lung nodules from chest CT scan images. For preprocessing they have used standard filtering method called median filter to remove noise and improve the contrast. Morphological closing operation is applied on the segmentation image to get the lung area template. In this paper statistical feature using GLCM and extracted only four features lie contrast, correlation, variance and homogeneity for a lung nodule. Through the LDA classifier based on the geometric, statistical and gray level characteristics they have shown 84% accuracy and 53.33% specificity.

Hong Shao, Li Cao, Yang Liu,[6] proposed a new detection for solitary pulmonary nodule based on CT images. For preprocessing adaptive wiener filter and lung nodules are segmented through the morphological method based on the ROI, then false nodules can be removed effectively through the compactness features. Furthermore utilizing features are extracted from obtained regions of interest (ROI). Then nodules are detected by Support Vector Machine (SVM) classifier. It has achieved a high 90.351% accuracy, 89.474% sensitivity,90.526% specificity.

- *Maria Jenifer.L is currently pursuing masters degree program in computer science and engineering in Anna University, India, PH-8056248108. E-mail: alkafionaeden@mail.com*

Mohan Allam, M.Nandhini,[11] a study on optimization techniques in feature selection for medical image analysis. In this paper is to compare a variety of algorithms for the selection of potential feature vectors from medical images based on evolutionary algorithm. They had given few selection methods with brief explanation for each algorithm and gave detailed explanation for the teaching learning based optimization.

Macedo Firmino, Antonio H Moris, Roberto M Mendoca, Marcel R Dantas, Helio R Hekis,[4] this paper presented a review of the existing literature on CAD systems for the lung cancer CT scans to identify challenges for future researches.

Ashis Kumar Dhara, Sudipta Mukhopadhyay, Anirvan Dutta, Mandeep Garg, Niranhan Khandelwal,[8] proposed a method to classify the lung nodules based on the shape and texture combinations. They have extracted 57 features like both 2D and 3D features. From the extracted features the few features are selected based on the ROCKIT and classified using the SVM. The main cause of proposed work is to improve the performance of the classification.

Kishore.R,[12] Describes the effective and efficient feature selection method for lung cancer detection. It gave brief details about the feature selection, image recognition, classification, retrieval.

Santosh Singh, Yogesh Singh, Ritu Vijay, [14] proposed a method to extract the features from the lung CT images. ROI is segmented with the combination of Thresholding and morphological operation. Extracted features like Area, Perimeter, Shape Complexity, Mean, Standard Deviation, and Circularity.

P.Mohanaiah, P.Sathyanarayana, L.GuruKumar,[15]. presents an application of gray level co-occurrence matrix (GLCM) to extract second order statistical texture features for motion estimation images .The Four features namely,Angular second Moment,Correlation ,Inverse Difference Moment and Entropy are computed using Xilinx FPGA.These texture features in this method have high discrimination accuracy,requires less computation time and hence efficiently used for real time pattern recognition. Third and higher order textures consider the relationships among three or more pixels .these are theoretically possible but not commonly implemented due to calculation time and interpretation difficulty.

Alex Sandro Aguiar Pessoa, Stephan Stephany,[1]Leila Maria Garcia Fonseca, Proposed a new feature selection technique and image classification using Rough Set theory(RST). In this paper RST is used in the classification stage, using the supervised training set with 9 ROI, three for each class are depicted. Rafael Bello et al. proposed a new hybrdization using ACO and rough set theory. The algorithm performance was compared according to the resulting reduct quantity and the average length of the resulting reducts set. In this paper they have considered features of the breast cancer, heart, LED, Lung cancer. For lung cancer 56 features are extracted from 56 features 33 is considered as the reducts. The comparison was made with other algorithms like Ant system model,aut system plus an elitist strategy. Ant colony system. The comparision of hybrid model with other methods to calculate reducts in RST shows that ant model based algorithm yield good results.

From the above literature survey we had different idea about the methods and techniques that has been used.

3 Proposed Method

Lung nodules are segmented using the segmentation technique. All the shape based, margin based, texture based features are computed from the segmented lung nodules. The block diagram for the proposed scheme is provided in Fig.1.

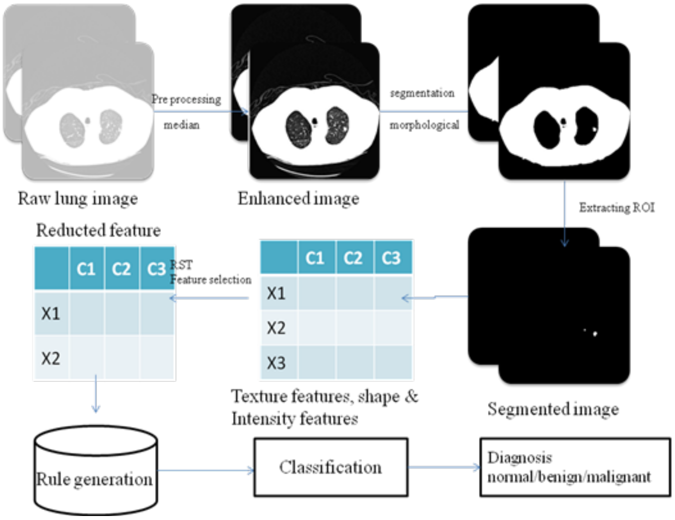

Fig 1: Architecture design

A. Data Acquisition

In this stage the input data is said to be an image in image processing. Lung CT image was acquired from the Lung Imaging Database Consortium (LIDC). We have taken 40 testing data and 60 training data to experiment the proposed algorithm technique. In fig 2, a sample lung CT image is shown.

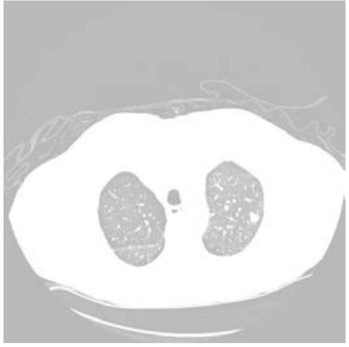

Fig 2: Lung CT Image

B. Image pre-processing

There are many techniques to filter the noise and remove the unwanted artifacts in the lung CT images. We have used the traditional method to filter the noise and artifacts called median filtering technique "Mokhled S.AL-Tarawneh"[13]. Median filter is the low pass filter. The median filter determines which pixel in the image has been affected by impulse noise. Median filter preserving useful detail in image. Median filter first sorts the entire pixel values from the surrounding neighborhood into numerical order and then replace it with the middle pixel value.

$$median[P(x) + Q(x)] \neq median[P(x)] + median[Q(x)] \quad (1)$$

Before going with the filtering technique we first imported the image to enhance it and later converted it into binary image. Then median filtering technique is applied. Fig 3(a)Before Enhancement, Fig 3(b) After Enhancement, Fig 3(c) Before preprocessing, Fig 3(d) After preprocessing

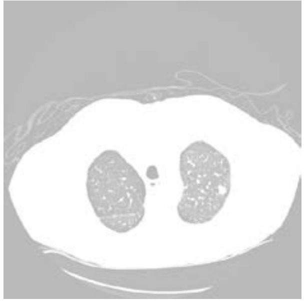

Fig 3(a): Before Enhancement

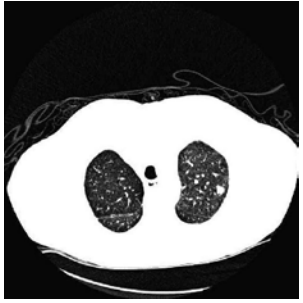

Fig3(b):AfterEnhancement

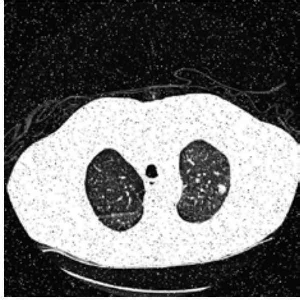

Fig 3(c): Before Preprocessing

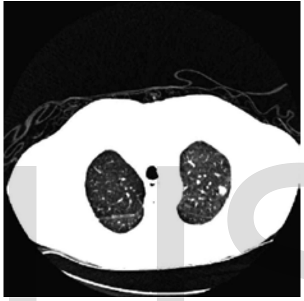

Fig 3(d): After Preprocessing

C. Lung volume segmentation

Before segmenting the lung nodule it is necessary to segment lung volume. For our research work, we used thresholding technique for lung segmentation "Hongping Lin, Haiquan Yao, Feng Peng"[5]. Thresholding is used to segment the image in to two regions based upon the difference of the pixel values between the object of interest in the image and the background "Emon Kumar Dey, Hossain Muhammad Muctadir"[7].

Let original image be A(x,y) and the binary image obtained after thresholding be B(x,y) where T is selected as threshold. The equation for calculating initial threshold is as follows:

$$f(x,y) = \left| \begin{matrix} 1 & A(x,y) \leq T \\ 0 & A(x,y) > T \end{matrix} \right| \quad (2)$$

We first extracted the intensity values of the pixels from the image histogram. Histogram threshold value usually ranges from 0-255. We extracted the 237-254 range of threshold value to segment the lung area. According to Fig 4 if we select the threshold range from 0-100 the opposite way of segmenting will be done so we go with 237-254. Later by means of Auto clustering and morphological opening operation we got the Malignant lung nodule by applying the radius 3 and N=0.Finally we cleared the borders and segmented the specific ROI. From that we got to extract the features in the next stage. Fig 4(a) histogram view to select threshold 4(b) Thresholding output 4(c) Morphological closing operation Output

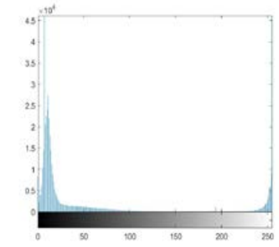

Fig 4(a) : Histogram view

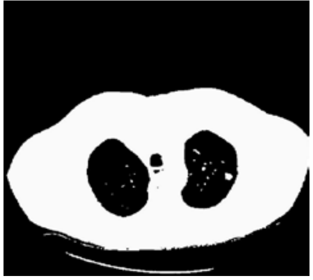

Fig4(b): Thresholding output

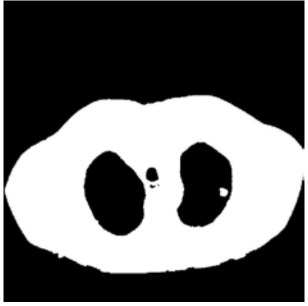

Fig 4(c): Morphological opening

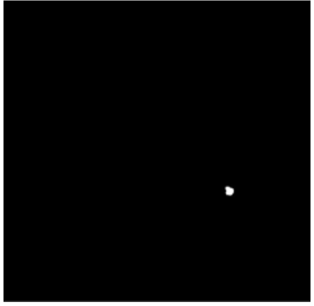

Fig 4(d): Segmented ROI

D. Feature extraction

Feature extraction plays a vital role in identifying the true-positive nodules accurately. For our research work, we extracted 8 2D shape intensity features for each nodule candidate separately "Wei-Li Zhang, Xi-Zhao Wang"[2]. They are area, perimeter, roundness, solidity, eccentricity, equivalent diameter, centroid and convex- area. Later we are going to extract 2DHaralick features like entropy, energy, Inverse difference moment, sum entropy, contrast, mean of HOG, variance of HOG, standard deviation HOG and local features like skew, kurtosis, Mean, variance & standard deviation "Khin Mya Mya Tun and Aung Soe Khaing"[3].

After segmentation is performed, the segmented lung nodules are used for feature extraction. A feature is a significant piece of information extracted from an image which provides more detailed understanding of the image. The features like shape intensity features, GLCM features and local features are extracted.

Shape measurements are physical dimensional measures that characterize the appearance of an object.

TEXTURAL FEATURES

The textural features are another feature set used in this work. Since the abnormality was widely spread in the image, the textuarl orientation of each class is different, which aid in better classification accuracy. Six features based on the first order histogram and seven features based on Gray Level Co-occurrence Matrix (GLCM) were used in this work.

Features based on First Order Histogram

The various features such as mean, standard deviation, skewness, kurtosis, energy and entropy based on the first order histogram are computed using the following equation.

The first order histogram estimate of $p(b)$is simply

$$p(b) = \frac{N(b)}{M} \qquad (6)$$

Where, b a gray level in the image

M total number of pixels in a neighborhood window centered about an expected pixel

$N(b)$ the number of pixels of gray value b in the same window that $0 \leq b \leq L-1$.

Features based on Gray Level Co-Occurrence Matrix

The image properties related to second-order statistics is estimated by the GLCM. Several researches suggested the use of Gray Level Co-occurrence matrices (GLCM) which have become one of the most well-known and widely used texture features. GLCM $\{P_{(d,\theta)}(i,j)\}$ represents the probability of occurrence of a pair of gray-levels(i,j) separate by a given distance d at angle θ. The commonly used unit pixel distances and the angles are $0°, 45°, 90°$ and $135°$ .The detailed algorithm of calculation of GLCM $\{P_{(d,\theta)}(i,j)\}$ is available in the literature.

$$p_x(i) = \sum_{i=1}^{N_g} p(i,j); p_y(j) = \sum_{i=1}^{N_g} p(i,j) \qquad (12)$$

$$p_{x+y}(k) = \sum_{i=1}^{N_g} \sum_{i=1}^{N_g} p(i,j), k = 2,3 \ldots \ldots 2N_g ; i+j = k \quad (13)$$

$$p_{x-y}(k) = \sum_{i=1}^{N_g} \sum_{i=1}^{N_g} p(i,j), k = 0,1 \ldots \ldots . N_g - 1; |i-j| = k$$
(14)

$p(i,j) =$ Gray level co-occurrence matrix

The feature such as contrast, inverse difference moment, correlation, variance, cluster shade, cluster prominence and homogeneity are calculated using the equations.

Feature extracted	Nodule candidate			
	1	2	3	4
Area	162	140	29	100
Perimeter	50.2843	44.0416	19.3137	43.7990
Circularity	1.2421	1.1025	1.0236	1.5266
Equivalent diameter	14.3619	13.3512	6.0765	11.2838
Roundness	0.8051	0.9070	0.9770	0.6551
Mean	6.1798e-04	5.3406e-04	7.5531e-04	0.0013
Variance	6.1760e-04	5.3377e-04	7.5474e-04	0.0010
Standard Deviation	0.0249	0.0231	0.0275	0.0319
skewness	40.1892	43.2372	36.3450	31.2860
Kurtosis	1.6162e+03	1.8705e+03	1.3220e+03	979.8138
Energy	0.1370	0.1632	0.1643	0.2115
Entropy	6.8426	6.6791	6.7858	6.3318
Contrast	0.5242	0.3405	0.4091	0.4343
Inverse difference moment	112.8971	97.7412	77.8236	97.4512
Correlation	0.8749	0.9603	0.9047	0.8347
Homogeneity	0.8497	0.8726	0.8945	0.8877
Variance	5.1141e+009	5.1174e+009	5.1169e+009	5.1251e+009
Cluster shade	-2.6682e+022	-2.7125e+022	-2.8782e+022	-1.1426e+022
Cluster prominence	2.6358e+028	2.6945e+028	2.9161e+028	8.5083e+027

Pulmonary nodules are circular in shape and blood vessels are slender-like structure, so we computed roundness of each nodule according to the formula

$$R = \frac{4\pi A}{P^2} \qquad (3)$$

Where, A is the area of the region and P is the perimeter, when the volume of R is close to 1, then ROI is close to the shape of a circle and it's more circular than the slender shape like blood vessels.

After calculating all the features from the above mentioned, we are going to select only few features to find whether it's benign or malignant and send through classifier for classification.

E. Feature Selection

The purpose of our research work is to improve performance of CAD system by selecting the optimal set of features. Feature selection removes the irrelevant and redundant features and enhances the performance of CAD system. In order to perform feature selection and to build up classification model using rough set theory we employed this method. Rough set theory is good for modeling, representing uncertainty and vagueness. Application of rough set theory can be seen in data reduction, data relationship, dependencies, similarity and differences. Hence it is good in pattern recognition. Where we know pattern recognition is the process of classifying input data into objects or classes based on the key features.

RST:

The reduced data will be done using Indiscernibility Relation through the Rough Set Method "G.Suseendran, M.Manivannan" [10]. Indiscernibility is of two type approximation they are Lower approximation and upper approximation. Lower approximation completely belongs to the class and upper approximation does not belong to the class. In this case we are going to select lower approximation for our research work.

In RST, data are stored in a tabular format, being a pair $S = (U, M)$called an information system, where U is the non-empty set of elements/objects called universe and A is a finite set of conditional attribute such that $U \to V_m$ contains the possible values of attribute m . In order to perform a supervised learning, the addition of a decision attribute is required. The resulting information system is $S = (U, M \cup \{d\})$, where d is the decision attribute.

$$IND_m(n) = \{(y, y' \in U) | \forall m \in N, m(y) \neq m(y')\} \qquad (4)$$

In RST, the size of the data set can be reduces either by representing whole class given by Indiscernibility relation or by eliminating the attribute that don't contribute to classification. Given original set M and set N is reducts(RED) when $IND_m(N) = IND_n(M)$ with $N \subseteq M$. Then if $|N| < |M|$, where $|.|$ denote cardinality of the set, this implies in a attribute selection, but preserving the information contained in the data that is relevant for classification. It may improve the classifier's accuracy and speed can be boosted.

SUPERVISED QUICK REDUCT (SQR)

The quick Reduct (QR) algorithm given in Algorithm 1 attempts to calculate a reduct without exhaustively generating all possible subsets. It starts off with an empty set and adds in turn, one at a time, those attributes that results in the greatest increase in the rough set dependency metric, until this produces its maximum possible value for the dataset. According to the algorithm, the dependency of each attribute is calculated and the best candidate is chosen

Algorithm 1:

SQR(C, D)

C, the set of all conditional features;

D, the set of decision features.

```
(1)$R \leftarrow \{\}$
(2) do
(3)$T \leftarrow R$
(4)$\forall x \in (C - R)$
(5) $\gamma_R \cup \{x\}^D = \frac{|POS_{R\cup\{x\}}(D)|}{|U|}$
(6) if $\gamma_{R\cup\{x\}}(D) > \gamma_T(D)$
(7)$T \leftarrow R \cup \{x\}$
(8)$R \leftarrow T$
(9) until $\gamma_R(D) = \gamma_c(D)$
(10) return R
```

F. Classification

The performance of the proposed Neighborhood rough set based classification algorithm is compared with traditional Pawlak's rough set (RS), K-nearest neighbor algorithm (KNN), Back propagation algorithm (BPN), Multilayer perceptron (MLP) and support vector machine (SVM). The obtained results of above classification algorithms are validated based on classification validation accuracy measures. Validation is important for the classification of medical data sets because accurate classification and decision making system is very important in data mining and medical diagnosis. There are many validation methods available for evaluating the accuracy of classification algorithm.

4 Conclusion

The proposed approach is more focused on the optimization of feature selection. We have carried out computer experiments on gray-level images in MATLAB tool. Lung CT images are taken from LIDC to experiment and analyze, it is being filtered using Median filter then the edge is being detected Thresholding and morphological open operations. The proposed method gives good results to obtain the optimal set of features. For feature extraction extracting 27 features like statistical, texture and structural features. From the extracted features Feature were selected using Rough Set Theory Method. Finally the specificity, Accuracy and sensitivity are being calculated using SVM to prove that this CAD system will give us optimal set of features to detect the cancer easily. We are working on the possibilities of applying a hybrid model (ant model +rough sets) for feature selection. This research includes, as an important issue, the setting of ant parameter.

References

[1] Alex sandro Aguiar Pessoa, Stephan Stephany, Leila Maria Garcia Fonseca, "Feature selection and Image classification using Rough Sets Theory"2011 IEEE,pg:2904-2907.

[2] Wei-Li Zhang, Xi-Zhao Wang, "Feature extraction and classification for human brain CT images", proceeding of the sixth international conference on machine learning and cybernetics,Hong Kong, 19-22 August 2007

[3] Khin Mya Mya Tun and Aung Soe Khaing, "Feature extraction and classification of lung cancer nodules using image processing Techniques", International journal of engineering Research & Technology, 2014

[4] Macedo Firmino, Antonio H Morais, Roberto M Mendoca, Marcel R Dantas, Helio R Hekis, "computer- aided detection system for lung cancer in computed tomography scans: review and future prospects".

[5] Hongping Lin, Haiquan Yao, Feng Peng, "CT image morphology features of pulmonary sclerosing hemangiomas", Chinese-German Journal of clinical Oncology, jan 2011.

[6] Hong Shao, Li Cao, Yang Liu, " A Detection approach for solitary pulmonary nodules based on CT images", second international conference on computer science and network technology, 2012,pg:1253-1257.

[7] Emon Kumar Dey, Hossain Muhammad Muctadir, "Chest X-Ray Analysis to detect mass tissue in Lung", third international conference of informatics, electronics and vision, 2014.

[8] Ashis Kumar Dhara, Sudipta Mukhopadhay, Anirvan Dutta, Mandeep Garg, Niranjan Khandelwal, "A combination of shape and texture features for classification of pulmunary nodules in lung CT images", society of imaging and informatics in medicine,6th january 2016.

[9] Tarun aggarwal, Asna Furqan, Kunal Kalra,"Feature Extraction and LDA based classification of lung nodules in chest CT scan images" International conference on advances in computing, communication and informatics(ICACCI),pg:1189-1193

[10] G.Suseendran, M.Manivannan, "Lung cancer image segmentation using rough set theory", Indian Journal of medicine and Healthcare Vol 4(6), Nov,2015.

[11] Mohan Allam, Dr M.Nandhini, " A Study on optimization Techniques in Feature Selection for medical Image Analysis", International journal on Computer Science and Engineering,vol 9 no.3 Mar 2017,pg:75-82

[12] Mr.R.Kishore, "An Effective and Efficient Feature selection method for lung cancer detection", International journal of computer science & information technology(IJCSIT),vol 7,No 4, August 2015,pg:135-

141.

[13] Mokhled S.AL-Tarawneh, “Lung cancer Detection using image Processing Techniques” ,Leonardo electronic journal of practices and technology,2012

[14] Santosh Singh, yogesh Singh, Ritu Vijay, “ An evaluation of feature extraction from lung CT image for the classification stage of malignancy”, IOSR journal of computer engineering, 2016,.

[15] P.Mohanaiah, P.Sathyanarayanan, L. GuruKumar, “Image Texture Feature Extraction using GLCM approach”, international journal of scientific and research publication, volume 3, issue 5, may 2013 pg:1-5.