iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 4 **Issue:** XI **Month of publication:** November 2016

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Computer Aided Analysis System for Detection of Lung Cancer with Content Based Image Retrieval

K. Lakshmi Vadivoo[1], D. Jenifer[2], T. Ratha Jeya Lakshmi[3]
Department of Computer Applications, Sri Sarada College for Women, Tirunelveli, Tamilnadu, India

Abstract ***Lung Cancer is the unrestrained expansion of Abnormal cells and it is the main reason for casualty of many people .So there is a need for detection of Lung Cancer at the early stage .This paper is based on detection of Lung cancer using Content Based Image Retrieval(CBIR) .It is urban on human lung cancer region .The presence of lung nodule is detected at the early stage of occurrence. Computed Tomography or Magnetic Resonance Images are given as input to the CAD (Computer Aided Diagnosis) .A CAD is a system that is designed by the integration of medical science and computers. CBIR is based on the features of an image. There are various phase describe in the proposed CAD system. These are withdrawal of lung section from chest computer tomography (CT) images, pre-processing, and segmentation of the lung region using graph-cut method, feature extraction from the segmented region, and the classification of occurrence and non-occurrence of nodules in the Lung. This paper describes the on hand literature and the technique used for the detection of lung cancer.***
Keywords – CAD system; similarity index; feature extraction; image retrieval

I. INTRODUCTION

In medical field the objective of Content based image retrieval is to permit radiologist to retrieve images of similar features that direct to similar analysis as the input image. This is different from other field where the objective is to find the adjacent image from the same group of an image. Therefore CBMIR: Content Based Image Retrieval Medical Images such technique cannot straight be applied in the medical field. In this paper, the image rescues provide a flexible means of sharp an image based on the description of the desired image.

Lung Nodule detection using Content based medical image retrieval: The lung cancer is considered as the notable cancer because it claims more than a million deaths every day. This lead to the condition of lung nodule detection in chest Computer Tomography (CT) images in advance. Thus the Computer Aided Diagnosis (CAD). This system is very necessary for early detection of lung disease. Early finding of the disease is serious but the truth remains that only 20% of cases are detected in the first phase. Radiologists can let pass up to 30% of lung nodules (which may develop into cancer) in chest radiographs due to the locale anatomy of the lungs which can hide the nodules. CAD helps radiologists by performing arts preprocessing of the images and suggesting the most likely location for nodules. [4] Detection of lung nodules proceeds through technique for suppresses the background structures in lungs which include the blood vessels, ribs and the images. The images obtained will afford better chest construction which make good regions for nodule and can be further classified depending on characteristics like size, contrast and shapes. Simple rule based classifications on such features tend to produce a lot of false positives. To overcome these harms, this proposed a Computer Aided Diagnosing (CAD) system for detection of lung nodules. This study initially apply the different image processing techniques such as Bit-Plane Slicing, Erosion, Median Filter, Dilation, Outlining, Lung Border Extraction and Flood-Fill algorithms for extraction of lung region. Then for segmentation algorithm is used and for knowledge and classification Support Vector Machine (SVM) is used. Medical Image databases and collections can be enormous in size, containing hundreds, thousands of images. The unadventurous method of medical image repossession is searching for a keyword that would match the descriptive keyword assigned to the image by a human categorizer [6]. Currently under development, even though several systems exist, is the retrieval of medical images based on their content, called Content Based medical Image Retrieval, CBMIR. While computationally exclusive, the results are far more accurate than usual image indexing. Hence, there exists a tradeoff between accuracy and computational cost. This tradeoff decrease as more capable algorithms are utilized and increased computational power becomes cheap. Computer Tomography (CT) has been considered as the most sensitive imaging technique for early detection of lung cancer.

There is a requirement for automated methodology to make use of large amount of data obtained CT images. Computer Aided Diagnosis (CAD) can be used powerfully for early detection of Lung Cancer. The usage of existing CAD system for early detection of lung cancer with the help of CT images has been unacceptable because of its low sensitivity and False Positive Rates (FPR). This

study presents a CAD system which can manually detect the lung cancer nodules with reduction in false positive rates. In this study, different image processing techniques are applied initially in order to obtain the lung region from the CT scan chest images. Then the segmentation is carried with the help of clustering algorithm. Finally for automatic detection of cancer nodules, Support Vector Machine (SVM) is used which helps in better categorization of cancer nodules. The testing is conducted for the proposed technique by CT images. In several articles, content based access to medical images for supporting clinical managerial has been proposed that would ease the management of clinical data.

II. RELATED WORK

Confined feature extraction methods have good results. Indexing design should be efficient for the searching technique, in the CBIR model. A content based retrieval system is a computer system for browsing and searching, and retrieving the images from a large database of digital images in image retrieval [1]. A bulk of work has been done in this area previously The only way of searching the images was by indexing or simply browsing. Now digital images databases opened the way to content-based searching described [2]. An Image indexing is the part of image retrieval system that Groups the similar images in a single cluster such that any query regarding the image matching, all the images can be retrieved with minimum delay. The rapid progress of multimedia computing and application has brought about a fiery growth of digital images in computer systems and networks [3]. This development has remarkably increased the need for image retrieval systems that are able to effectively directory a large amount of images and to efficiently retrieve them based on their visual contents. In developing a visual content-based image recovery system, the first critical decision to be made is to determine what image feature, or combination of image features are to be used for image indexing and retrieval [4]. By comparing results of an image analysis with ground truth, researchers are able to determine how exact their algorithms are. Currently there are very few tools to help researchers in assembly and analyzing this ground truth [5]. The value of each histogram bin is a feature, and all features jointly form a feature vector that we input to our classifier. Through empirical experiments, we found that using 32 bins for each descriptor gives us good realistic results [6]. Another important idea in the proposed method is that the deformable model is constrained by both population-based and patient-specific shape statistics. The shape statistics collect from the segmentation results of a population [7].

III. PROPOSED SYSTEM

The proposed CBMIR support is shown in the following figure (Figure.1). The database, where the images are kept is call Image database. In the pre-processing technique, the images are enhanced, segmented, and subdivided in order to make flexible work environment for further processing works. The proposed model is a mixture of feature extraction methods namely texture and gray scale motion. Then this combined form of feature set is stored as a single feature vector in the feature database. When the user submit a query image, the same process steps (such as pre-processing, feature extraction steps) are carried out as in the offline image database process in order to get the feature vector value for the query image. Then this query image feature vector value will be compare with feature vector value of the feature database. Based on the comparison, images that are directly similar to the query image are retrieving from the databases and displayed.

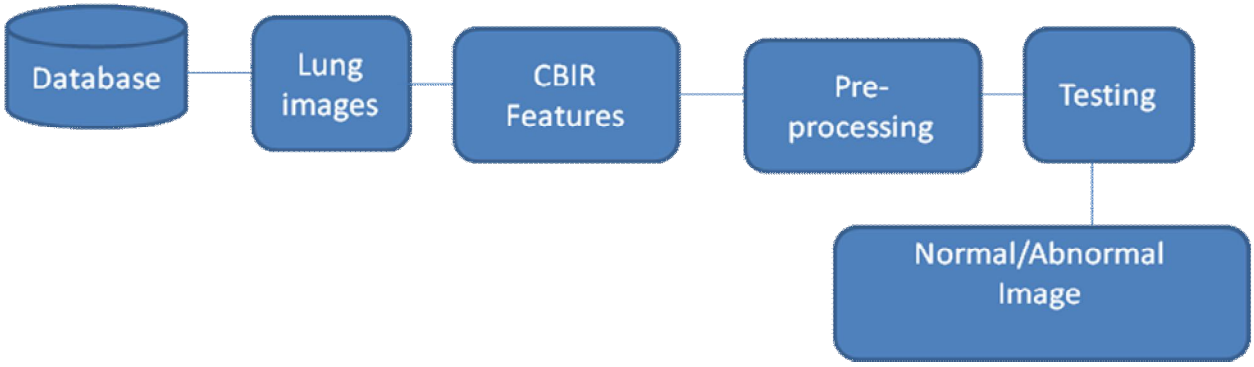


Figure1: Content Based Image Retrieval System As A Diagnosis Aid.

A. *Lung region extraction*

The initial stage of the proposed Computer Aided Diagnosing (CAD) (Wiemker et al., 2003; Wiemker et al., 2002) techniques is the extraction of lung region from the CT scan image. The basic image processing techniques are utilized for this purpose.

B. Pre-Processing

Image Pre-Processing is done for the enhancement of the image and it eliminates some distortion. The unwanted signal acquired during image acquisition is removed by using median filter. All the corrupted and noisy pixels in the Lung image is enhanced. This procedure helps to maintain accuracy in future processing of the lung region.

Figure 2 : (a) Before Pre-processing (b)After Pre-processing

C. Region Based segmentation Approach-Region Growing

Image segmentation is the process of unraveling or grouping an image into different parts. These parts normally communicate to something that humans can easily separate and view as individual objects. The segmentation process is based on various features found in the image. The goal of image segmentation is to cluster pixels into salient image regions, where the regions corresponding to individual surfaces, objects, or natural parts of objects. Segmentation could be used for object recognition, occlusion boundary, estimation within motion or stereo systems, image compression, image editing, or image database look-up. Segmentation is an important procedure in medical image analysis. The segmentation process is carried out as pre-processing in the process. This method is used to separate the particular region in the image, since there are typically some clearly defined areas within the image.

The first region-growing method was the seeded region growing method. This method takes a set of seeds as input along with the image. The seeds mark each of the objects to be segmented. The regions are iteratively grown by comparing all unallocated neighboring pixels to the regions. The difference between a pixel's intensity value and the region's mean, δ, is used as a measure of similarity. The pixel with the smallest difference measured this way is allocated to the respective region. This process continues until all pixels are allocated to a region. Seeded region growing requires seeds as additional input. The segmentation results are dependent on the choice of seeds. Noise in the image can cause the seeds to be poorly placed. Unseeded region growing is a modified algorithm that doesn't require explicit seeds. It starts off with a single region A1 – the pixel chosen here does not significantly influence final segmentation. At each iteration it considers the bordering pixels in the same way as seeded region growing. It differs from seeded region growing in that if the minimum δ is less than a predefined threshold T then it is added to the respective region Aj. If not, then the pixel is considered significantly different from all current regions Ai and a new region An + 1 is created with this pixel.

Initially its picks an arbitrary(r, c) pixel from the domain of image to be segmented.

This pixel is called as seed pixel.

Now examine the nearest neighbor of one by one and the neighboring pixel is accepted to belong to the same region, if they together satisfy the homogeneity property of a region.

Once a new pixel is accepted as a member of a current region, the nearest neighbor of this new pixel are examined.

This process goes on recursively until no more pixel is accepted.

All the pixels of current region are marked.

Then another seed pixel is picked up and the same process is repeated.

D. Graph cut method

In Graph-cut the images are represented as graph and the pixels are represented as nodes. There is the connection between each and every pair of nodes. The image is segmented into two parts one is the source and another one is the sink. The source is the area that is segmented and sink is the background. Graph cuts method is employed for efficiently solving low-level computer vision problems.

Volume 4 Issue XI, November 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

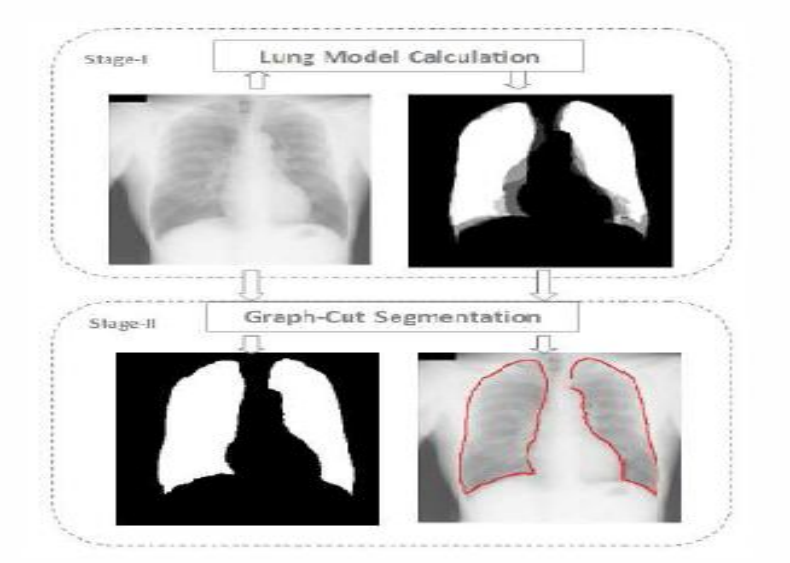


Figure 3: Graph Cut Differentiation

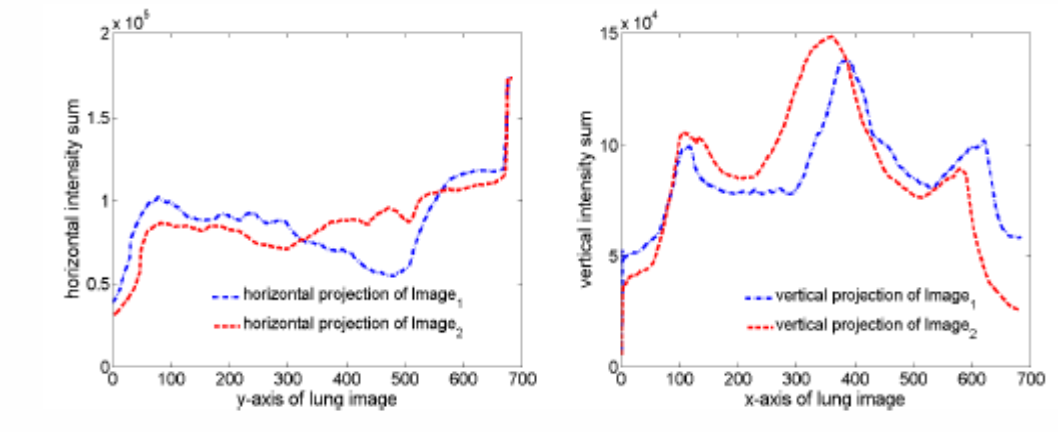


Figure 4: Comparison Graph

Sensitivity of BPNN		Sensitivity of KNN		Sensitivity of GONN		
Mean	SD	Mean	SD	Mean	SD	Accuracy
90.30	0.125	92.40	0.6327	98.17	0.589	91%
89.50	0.850	91.50	0.6589	94.25	0.256	52%
87.56	0.250	98.60	0.5231	82.15	0.369	96.5%
93.58	0.189	93.25	0.2314	74.15	0.278	98.13%
97.20	0.178	94.20	0.8569	98.26	0.458	89.35%
88.89	0.125	96.32	0.2310	93.25	0.256	99.17%

Table 1: Sensitivity and Accuracy Calculation

Techniques	FALSE POSITIVE	FALSE NAEGATIVE
KNN Accuracy	20	13.5
SVM Accuracy	0.25	0.75
Linear Kernel Accuracy	11.5	19.5
Non-Linear Kernel Accuracy	2	10

Table 2: False Positive And False Negative Range

Volume 4 Issue XI, November 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

E. Feature extraction

After the segmentation is performed on the lung region, the features can be obtained from it for determining the diagnosis for detecting the cancer nodules in the lung region perfectly. The features that are used in this study are Texture features using Co-occurrence matrix representation.

F. GLCM

Gray level Co-occurrence matrix (GLCM) based texture feature extraction has been considered as the powerful technique and still now has been used in many applications of remote sensing for texture analysis. GLCM method comes under the statistical approach of texture analysis which describes texture as a set of statistical by R.M.Haralick measures based on the spatial distribution of gray levels within the band of the remotely sensed imagery. GLCM matrix is computed from a relative displacement vector (d,) which is formed based on the relative frequencies of gray level pairs of pixels separated by a distance d in direction. Haralick suggests texture statistical measures based on GLCM matrix and the most popularly used texture measures are as follows:

1) *Energy*: Most cancers, including lung tumors, are made of cells that divide more rapidly than those in normal lung tissue, holding out the hope that the tumor can be eliminated without damaging surrounding normal tissues. Radiotherapy acts by attacking the genetic material or DNA within tumor cells, making it impossible for them to grow and create more cancer cells. Normal body cells may also be damaged—though less markedly but they are able to repair themselves and function properly once again. The key strategy is to give daily doses of radiation large enough to kill a high percentage of the rapidly dividing cancer cells, while at the same time minimizing damage to the more slowly dividing normal tissue cells in the same area.

$$\text{Energy} = \sum_{i} \sum_{j} p_d^{\,2}(i,j)$$

2) *Entropy*: The cancer stem cell hypothesis, that small populations of tumour cells are responsible for tumor genesis and cancer progression is becoming widely accepted and recent evidence has suggested a prognostic and predictive role for such cells. Intra-tumor heterogeneity, the diversity of the cancer cell population within the tumor of an individual patient, is related to cancer stem cells and is also considered a potential prognostic indicator in oncology. The measurement of cancer stem cell abundance and intra-tumour heterogeneity in a clinically relevant manner however, currently presents a challenge. Here we propose signalling entropy, a measure of signalling pathway promiscuity derived from a sample's genome-wide gene expression profile, as an estimate of the stemness of a tumour sample.

$$\text{Entropy} = \sum_{i} \sum_{j} p_d^{\,2}(i,j) \log p_d(i,j)$$

3) *Correlation:* A correlation is a statistical measure of the degree to which changes to the value of one variable predict change to the value of another. In positively correlated variables, the value increases or decreases in tandem. In negatively correlated variables, the value of one increases as the value of the other decreases. Correlation are expressed as values between +1 and -1. A coefficient of +1 indicates a perfect positive correlation. A change in the value of one variable will predict a change in the same direction in the second variable. A coefficient of -1 indicates a perfect negative correlation. A change in the value of one variable predicts a change in the opposite direction in the second variable. Lesser degrees of correlation are expressed as non-zero decimals. A coefficient of zero indicates there is no discernable relationship between fluctuations of the variables.

$$\text{Correlation} = \sum_{i} \sum_{j} (i-\mu)(j-\mu) p_d(i,j) \,/\, \sigma_x \sigma_y$$

4) *Homogeneity*: Equal group sizes may be defined by the ratio of the largest to smallest group being less than 1.5. If group sizes are vastly unequal and homogeneity of variance is violated, then the F statistic is considered liberal when large sample variances are associated with small group sizes. When this occurs, the alpha value is greater than the level of significance. This indicates that the null hypothesis is being falsely rejected. On the other hand, the F statistic is considered too conservative if large variances are associated with large group sizes. This would mean that the actual alpha value is less than the level of significance. This does not cause the same problems as falsely rejecting the null hypothesis however; it can cause a decrease in the power of the study.

$$\text{Homogenity} = \sum_i \sum_j p_d(i,j) / 1+(i-j)^2$$

Where
pd refers the probability matrix obtained through GLCM
μ is the mean of pd.
GLCM measures are calculated and depicted for each pixel in the image.

5) *Algorithm for GLCM:* The steps for extracting texture features of image using GLCM can be given as below.
Separate the R,G,B planes of image.
Repeat steps 3-6 for each plane.
Compute four GLCM matrices (directions for
δ=00, δ=450, δ=900, δ= 1350) as given by eq.
For each GLCM matrix compute the statistical features Energy (Angular second moment), Entropy (ENT), Correlation (COR), Contrast (CON).
Compute the feature vector using the means and variances of all the parameters.

G. Classification of Occurrence and Non Occurrence of Cancer in the Lung

The final phase in the proposed CAD system the classification of occurrence and non occurrence of cancer nodule for supplied lung image. The classifiers used in this paper are support vector machine. Support Vector Machine (SVM). SVM is usually used for classification tasks introduced by Cortes. For binary classification SVM is used to find an Optimal Separating Hyper plane (OSH) which generates a maximum margin between two categories of data. To construct an OSH, SVM maps data into a higher dimensional feature space. SVM performs this nonlinear mapping by using a kernel function.

Then, SVM constructs a linear OSH between two categories of data in the higher feature space. Data vectors which are nearest to the OSH in the higher feature space are called Support Vectors (SVs) and contain all information required for classification of the particular phenomenon. In brief, the theory of SVM works as follows:

1) *SVM kernel functions:* The classification ability of feature combinations in gait applications is obtained with first attempt work of SVM kernel function. The three main kernel functions are used for our study here. Partial kernel function, influence to data near test points. The above mentioned kernel functions are briefly explained in this chapter. The most used kernel function for SVM is Radial Basis Function (RBF). Radial basis function kernel: The B-Spline kernel is defined on the interval [-1, 1]. It is given by the recursive formula:

k(x, y)= B2p+1 (x -y)

$$p \in N \text{ with } B_{i+1} := B_i \otimes B_0$$

Where
In the study by Bart Hamers it is given by:

$$k(x,y) = \prod_{p=1}^{d} B_{2n+1}(X_p - y_p)$$

Alternatively, Bn can be computed using the explicit expression:

$$B_n(x) = \frac{1}{n!}\sum_{k=0}^{n+1}\binom{n+1}{k}(-1)^k\left(x + \frac{n+1}{2} - k\right)^n$$

Where x+ is defined as the truncated power function:

$$x_+^d = \begin{cases} x^d, \text{if } x > 0 \\ 0, \text{otherwise} \end{cases}$$

Linear kernel: The Linear kernel is the simplest kernel function. It is given by the inner product <x,y> in addition with an optional constant c. Kernel algorithms using a linear kernel are often equivalent to their non-kernel counterparts:

$$k(x, y) = zty + c \text{ [4]}$$

Polynomial kernel: The Polynomial kernel is a nonstationary kernel. Polynomial kernels are apt for problems where all the training data is normalized:

Volume 4 Issue XI, November 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

$$k(x,y) = (\infty xty + c)d \text{ [4]}$$

Modifiable parameters are the slope alpha, the constant term c and the polynomial degree d.

After the learning process is completed by providing several conditions, the proposed technique is able to detect the cancer occurrence in the lung region automatically

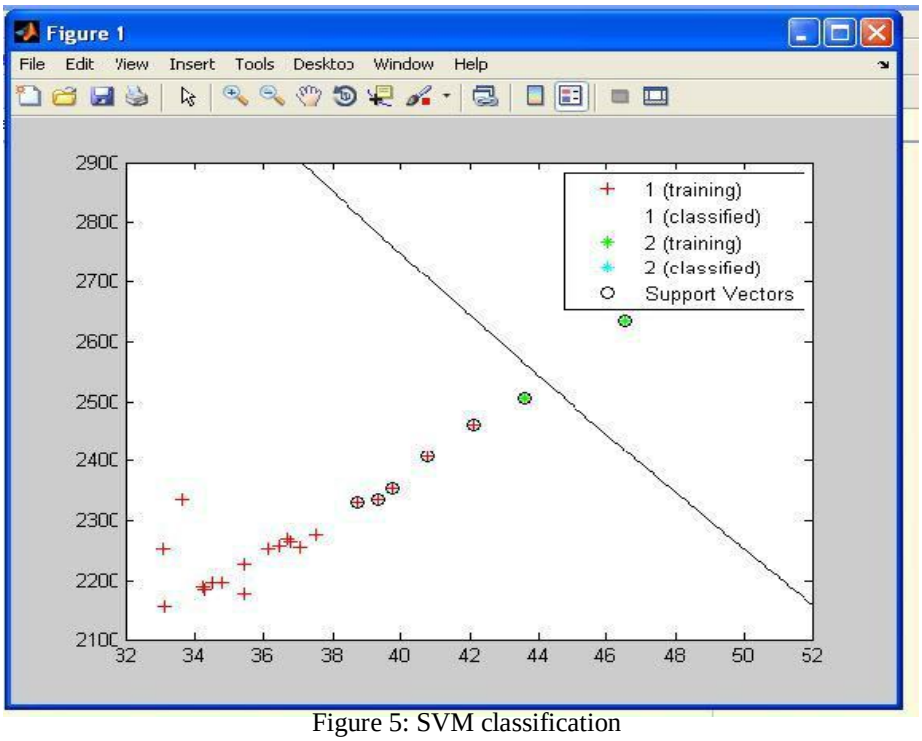


Figure 5: SVM classification

IV. EXPERIMENTAL RESULT

The experiments are conducted on the proposed computer-aided diagnosis systems with the help of lung images obtained from the website. This experimentation data consists of 55 lung images. Those 55 lung images are passed to the proposed CAD system. Out of these 35 images taken as training set and 20 images taken as test set.The diagnosis rules are then generated from those images and these rules are passed to the Support Vector Machine (SVM) for the learning process. After learning, a lung image is passed to the proposed CAD system. Then the proposed system will process through its processing steps and finally it will detect whether the supplied lung image is affected by cancer or not. The results shows that there are few mis-detections but overall efficiency of vision based efficiency measure is more than 80%.

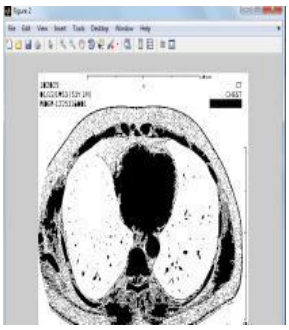


Figure 6: (a) Original CT Scan Image (b) Lung Nodule detected after Segmentation.

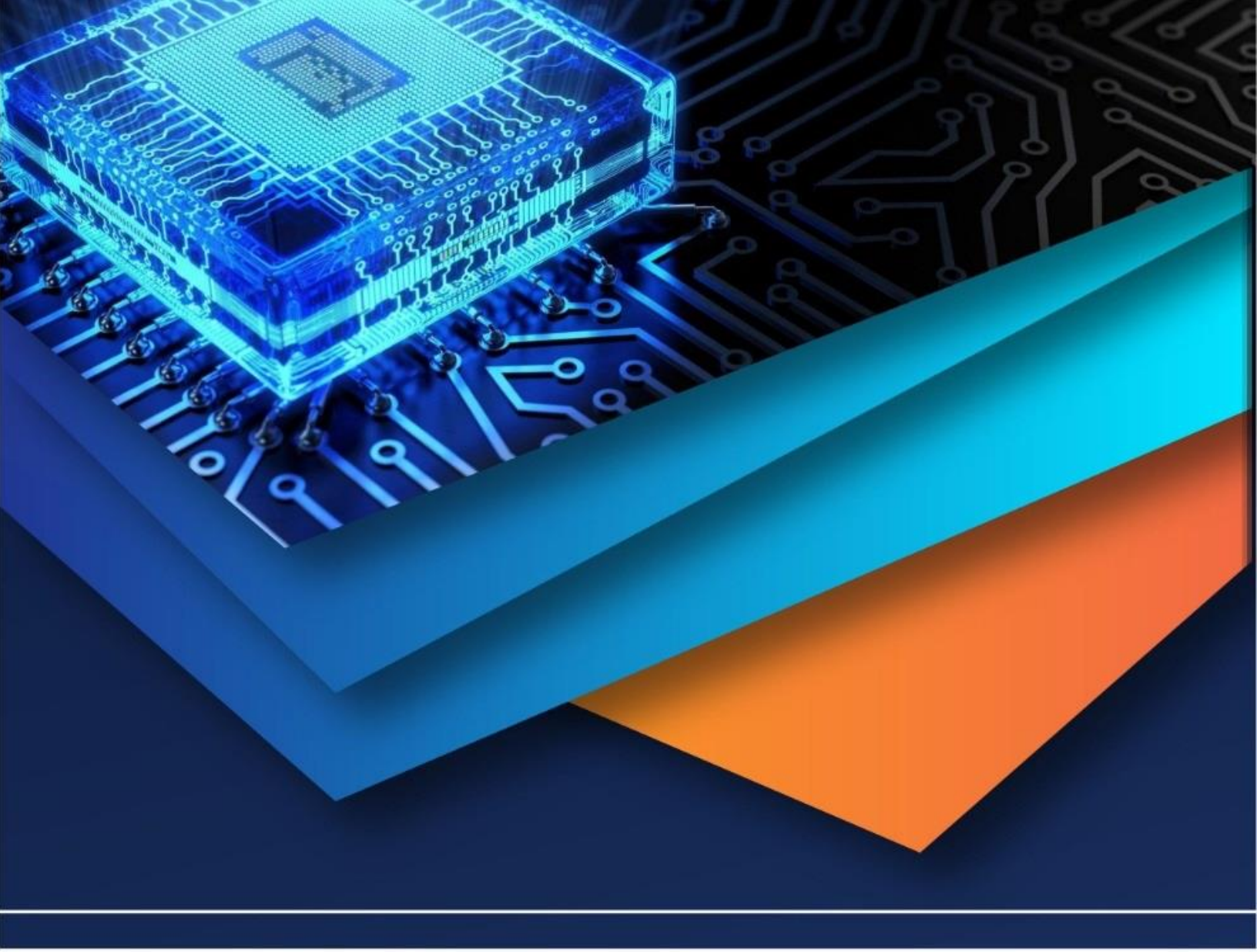
V. CONCLUSION

The difficulty in the early detection of lung cancer nodules is overcome in this paper. This paper provides a computer aided

diagnosis system for early detection of lung cancer. The chest computer tomography image is used in this paper. In the first phase of the proposed technique, the lung region is extracted from the chest tomography image. The different basic image processing techniques are used for this purpose. In the second phase, extracted lung is segmented with the help of region based approch. The next phase is extraction of features for diagnosis from the segmented image. Finally, the classification is performed to detect the occurrence of cancer nodules. The experimental result reveals the advantage of the proposed CAD system for detecting lung cancer. Thus normal images and abnormal images can be distinguished and assist in diagnosis. The visual characteristics of a disease carry diagnostic information and oftentimes visually similar images correspond to the same disease category. By consulting the output of a CBMIR system, the physician can gain more confidence in his/her decision or even consider other possibilities. The results shows that there are few mis-detections but overall efficiency of vision based efficiency measure is more than 80%.

REFERENCES

[1] B. Ramamurthy, K. R. Chandran, S. Aishwarya, CBMIR: Content Based Image Retrieval using Invariant Moments, GLCM and Grayscale Resolution for Medical Images(2011).
[2] Dr. H.B.Kekre, Sudeep D. Thepade, Tanuja K. Sarode and Vashali Suryawanshi, Image Retrieval using Texture Features extracted from GLCM, LBG and KPE (2010)
[3] M. Gomathi, P. Thangaraj A Computer Aided Diagnosis System for Lung Cancer Detection using Machine Learning Technique, (2011).
[4] Michael Lama, Tim Disneyb, Mailan Phamc, Daniela Raicud, Jacob Furstd, Ruchaneewan Susomboond aJames, Content-Based Image Retrieval for Pulmonary Computed Tomography Nodule Images, Madison University, (2006)
[5] "DIGITAL IMAGE PROCESSING", by Rafael C.Gonzalez and Richard E.Woods
[6] Marwa N. Muhammada, Daniela S. Raicub, Jacob D. Furstb, Ekarin Varutbangkulb, Texture versus Shape Analysis for Lung Nodule Similarity in Computed Tomography Studies,(2008)
[7] Li Guohui, Liu Wei, Cao Lihua animage retrieval method based on color perceived feature Journal of Image and Graphics, (1999).
[8] Kawata, Y.; Niki, N.; Ohmatsu, H.; Kusumoto, M.; Kakinuma, R.; Mori, K.; Nishiyama, H.; Eguchi, K.; Kaneko, M.; Moriyama, N., "Searching similar images for classification of pulmonary nodules in three-dimensional CT images," Biomedical Imaging, 2002. Proceedings. 2002 IEEE International Symposium on , vol., no., pp.189,192, 2002
[9] R. M. Haralick, K. Shanmugam, and I. Dinstein,"Textural Features of Image Classification", IEEE Transactions on Systems, Man and Cybernetics, vol. SMC-3(6), Nov. 1973.
[10] N.G. Yadav, "Detection of lung nodule using Content Based Medical Image Retrieval" International Journal of Electrical, Electronics and Data Communication, ISSN (p): 2320-2084, Volume-1, Issue-2, April-2013



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY