iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** III **Month of publication:** March 2018

DOI: http://doi.org/10.22214/ijraset.2018.3490

www.ijraset.com

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

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue III, March 2018- Available at www.ijraset.com

The Diagnosis of Lung Cancer using Machine Learning

S. Lavanya[1], Dr. A. Tamilarasi[2]
[1]*Pg Scholar - Department of Computer Applications, Kongu Engineering College, Perundurai, Tamilnadu, India*
[2]*Head of the Department -Department of Computer Applications, Kongu Engineering College, Perundurai, Tamilnadu, India*

Abstract: ***Nowadays cancer has become huge threat in human life there are many types of cancer, Lung cancer is one of the common types causing very high mortality rate. The best way of protection from lung cancer is its early detection and prediction The detection of lung cancer in early stage is a challenging problem, due to the structure of the cancer cells, where utmost of the cells are overlapped with each other.It is a computational procedure that sort images into groups according to their similarities.In this Histogram Equalization is used for preprocessing of the images and feature extraction process and Support Vector Machine classifier to check the condition of a patient in its early stage whether it is normal or abnormal. The performance is based on the correct and incorrect classification of the classifier.***
Keywords: Lung cancer, CT scan, preprocessing Resize wiener filtering, Features Homogenty, constrast.

I. INTRODUCTION

Lung cancer is one of the most dangerous cancers in the world, with the least survival rate after the diagnosis, with increase in the number of deaths every year gradually. Still a exact treatment is not found Early detection of is Lung cancer important for a successful treatment. CT Images are considered to be the most widely used technique for the detection of lung cancer. It is a complex task to analyze these images as they are projected images. A medical expert has to make extensive knowledge of anatomy and imaging techniques.

A. Computer Tomography (CT) Images

Lung cancer is one of the most dangerous forms of cancer because it claims more than a million precious lives every year. So, lung nodule detection in chest Computed Tomography (CT) images becomes very necessary in the present clinical world. Thus the Computer Aided Diagnosis (CAD) system is very essential for early detection of lung cancer.CT uses special x-ray equipment to get image data from various angles around the human body, and then utilizes computer processing of the information to demonstrate a cross-section of tissues and organs.

CT imaging is very useful as it can display various types of tissues and organs with high clarity, when an intravenous contrast (x-ray dye) is utilized. Moreover, tissues like kidney or gallstones can be accurately detected with CT, and abnormal fluid or enlarged lymph nodes in the abdomen or pelvis can also be identified with great accuracy. Some organs like stomach are not that much accurately assessed by the CT, but it is used to indirectly diagnose these organs by detecting abnormalities in the adjacent soft tissues. CT Colonography (Virtual Colonoscopy) is a novel technique which allows primary assessment of the distended colon to detect polyps, the precursor to colon cancer, and is very effective in screening for this disease CT is becoming one of the most popular and effective methods for diagnosing many diseases including diverticulitis and appendicitis, and for visualizing the liver, spleen, pancreas and kidneys as it is a non-invasive procedure that provides detailed, cross-sectional views of all types of tissue. CT can quickly identify the source of pain in cases of acute abdominal distress. The speed, ease and accuracy of a CT examination can minimize the risk of serious complications, when pain is due to infection and inflammation. CT is widely used for diagnosing various types of cancer, including kidney and pancreatic cancer. CT is very effective as the image gives full information about the presence of a tumor and to measure its size, precise location, and the extent of the tumor involvement with other nearby tissue (staging the tumor). CT assessments of the lower Gastro Intestinal (GI) tract are very much useful to plan and properly administer radiation treatment for tumors. CT also plays a crucial role in abdominal trauma assessment, since it is very sensitive at picking up bleeding within and around the solid organs. CT is also very effective in the detection, diagnosis and treatment of vascular disorders through a new approach called CT Angiography.

1) *Spiral CT:* Spiral CT is also called as “volume scanning”. It is clearly different from usual CT and the tomographic technique is used in Spiral CT. Spiral CT uses a different scanning principle. The patient on the table is moved constantly

through the scan field in the z direction while the gantry performs multiple 360° rotations in the same direction. The X-ray draws a spiral rotation around the body and constructs a data volume. This volume is created from a multitude of three-dimensional picture components, i.e. voxels. The movement of the table in the 'z' direction will usually produce inconsistent sets of data, which causes every image reconstructed directly from a volume data set to be degraded by artifacts. Software applications Facilitate the use of spiral CT even for regions which are subject to involuntary movements.

2) *Production of CT Image:* A thin, needle-like beam linearly scans the object. A sort of shadow image is produced which is called as "attenuation profile" or "projection". It is recorded by the detector and the image processor. The tube and the detector are further rotated by a small angle. A second shadow image is produced by linearly scanning the object from another direction. This process is repeated until the object has been scanned for a 180° rotation.
3) *CT in General Clinical Use:* Spiral CT has greatly reduced the scan times compared to sequential CT. Spiral CT is very much useful when examining patients who are unable to cooperate. The complete volume in Spiral CT is scanned faster and without gaps so that motion artifacts due to various respiratory conditions during the acquisition are minimized. Multiple scans due to breathing, during the acquisition are not needed. So, patient dose is also minimized. Figure 1.1 shows an example of CT image.
4) *Advantages of spiral CT in clinical use*

a) Reduced scan times
b) Full coverage of organs in a single respiratory position
c) Added diagnostic information due to improved resolution and3Dvisualization in routine operation

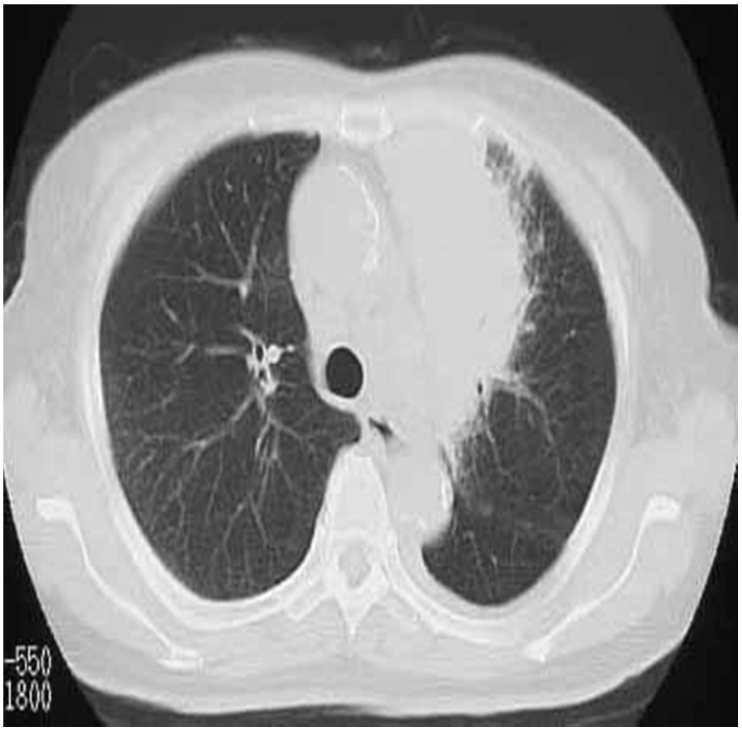


Figure 1.1 An Example of CT image

5) *Benefits of CT scan*

a) CT scan can easily detect the causes of abdominal pain with very high accuracy, enabling faster treatment
b) CTscanning provides thorough views of many types of tissues, including the lungs, bones, soft tissues and blood vessels.
c) CT scanning is painless, non-invasive and accurate.
d) CT examinations are very simple and rapid.
e) Detection and diagnostics done with CT eliminates the need for invasive exploratory surgery and surgical biopsy.
f) Scanning done using CT identifies both normal and abnormal structures, making it a useful tool to guide radiotherapy, needle biopsies etc.
g) CT is cost-effective imaging tool for a wide range of clinical problems.

Computed Tomography (CT) is a dominant tool which allows very quick creation of x-ray images of the body with high-resolution cross-sectional imaging. The quick, detailed result has made CT very valuable, especially in the emergency department (ED). High-quality, cross-sectional images are available in a quick time that helps to define the medical status of the patient.

II. LITERATURE REVIEW

Cancer is the most vicious disease, the cure of which must be the prime target through scientific investigation. The early detection of cancer can be helpful in curing the disease completely. There are several techniques available in the literature for the detection of cancer. Many researchers have contributed their ideas in the detection of cancer.

This chapter mainly discusses about the existing cancer detection techniques available in the literature. Several domains and concepts are used in the detection of cancer. The main domains used in this detection technique include neural networks, image processing, nanotechnology etc.

A. Neural Networks In Cancer Detection

Ginneken et al (2001) described a survey in which the authors classify the lung regions extraction approaches into two different categories; 1) Rule-based 2)Pixel classification based category

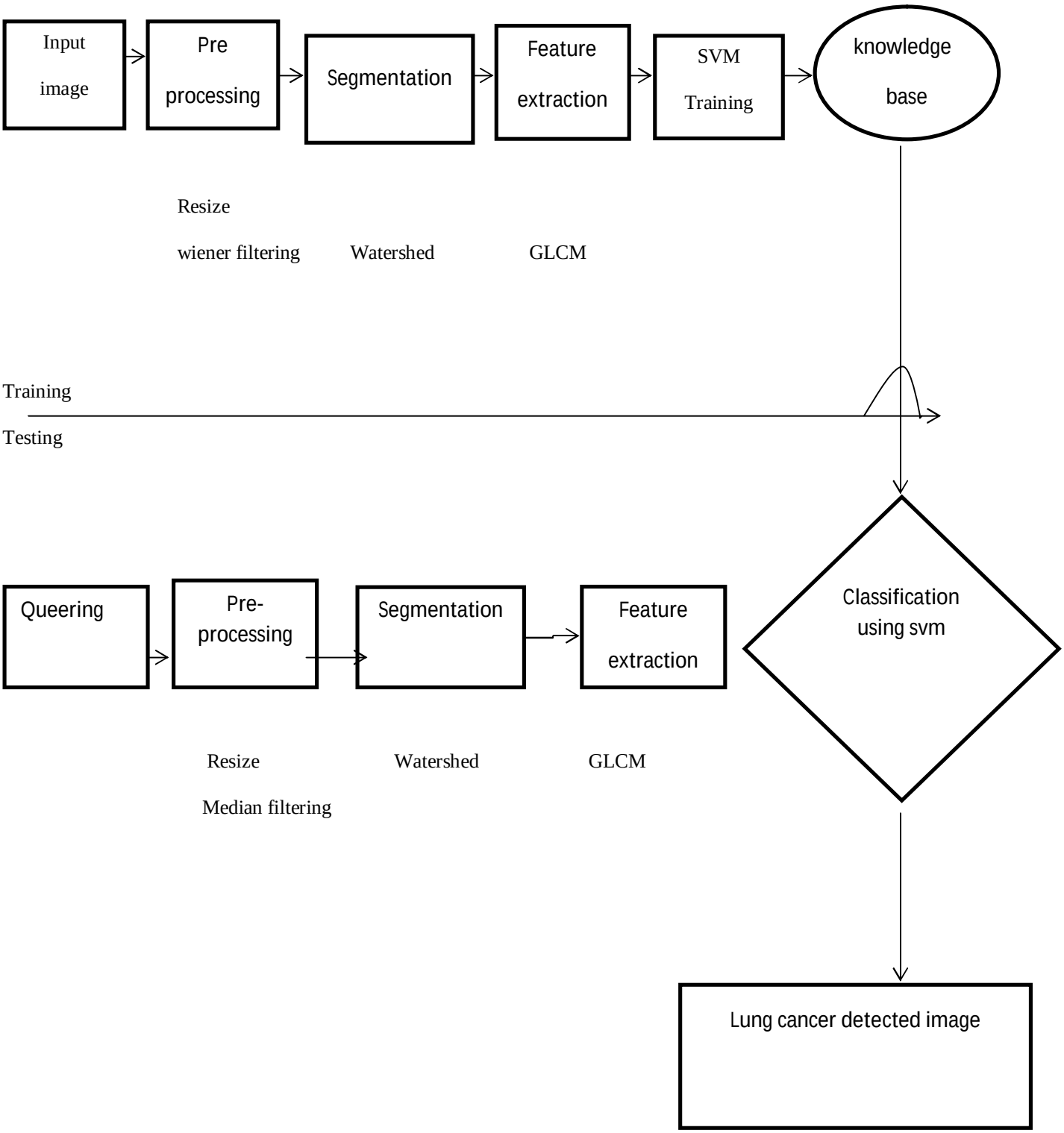
In the rule-based technique, sequence of steps, tests and rules are used in the extraction process and most of the proposed techniques belong to this category. Thresholding, region growing, edge detection, ridge detection morphological operations, and dynamic programming are some of the techniques used. Pixel classification is another approach used for the lung regions extraction process, where each pixel in the CT image is categorized into an anatomical class. Various types of neural networks, Markov random field modeling are the classifiers trained with a variety of local features including intensity, location, and texture measures.

Woten et al (2007) proposed a numerical investigation into the improvement of artificial neural network detection of breast cancer using a planar broadband antenna and a three-region breast approach. In this approach, Modified Four point antennas are used, which are capable of producing various wave polarizations. The effect of wave polarization on statistical detection is fully described in this approach by the author.

Mass spectrometry-based proteomics offers a capable technique for the accurate diagnosis of different diseases. But, there are certain problems in the mass spectral data such as huge volume, data complexity and the presence of noise which make the analysis of the proteomic pattern very difficult. In this proposed approach, a neural network-based system is proposed by Xu et al (2009) for proteomic pattern analysis for prostate cancer screening. The technique is mainly of three stages namely feature selection based on statistical significance test, classification by a Radial Basis Function Neural Network (RBFNN) a probabilistic neural network (PNN), and finally results in optimization through ROC analysis. The experimental observation shows that the proposed approach is very effective when compared with the existing approaches. The proposed approach has high sensitivity (97.1%) and specificity (96.8%) when combined with prostatic biopsy and is expected to help in early detection of prostate cancer.

III. METHODOLOGY

Our Proposed system includes two modules training and testing .In Training module the cropped CT scan image is input to the system, followed by Pre-Processing to enhance the image. In next step features are extracted and passed to SVM Classifier. In Testing module CT image is sent to Pre-Processing phase and the second part is image segmentation to extract the lung region and ROI .The third part is feature extraction and selection to extract the main features of the tumor. The last part is the classifier to discriminate the Detection of cancer or not a cancer .Below figure1 shows the block diagram of proposed lung cancer detection system.



A. Pre-Processing

Medical images are corrupted with noise and artifacts due to body movements Preprocessing is done to remove unwanted noise and it gives clarity to the images at this stage where filtering is done to remove noise.In our proposed system we are using resize and thus wiener s used to remove noise.

1) *Input Image:* Here, the input images are chest CT scan images in JPEG format that contain tumors. First image selected from the file specified by the string filename. The user has to select therequired lung CT scan image for further processing. Then each image is resized to 256*256.
2) *Wiener Filter:* The intention of the Wiener filter is to clear out noise that has corrupted a signal. It is based on a statistical approach.Natural filters are designed for a desired frequency response. The Wiener filter approaches filtering from different angle. One is assumed to have knowledge of the spectral properties of the common signal and the noise, and one seeks the LTI filter whose output would come as close to the original signal as possible.

Wiener filters are characterized by the following.

Assumption: signal and (additive) noise are stationary linear random processes with known spectral characteristics.

Requirement: the filter must be physically realizable, i.e. causal (this requirement can be dropped, resulting in a non-causal solution).

Performance criteria: minimum mean-square error.

IV. SEGMENTATION

A. Watershed Algorithms

1) *Read in the color image and convert it to grayscale:*

```
rgb = imread('pears.png');
I = rgb2gray(rgb);
imshow(I)
 text(732,501,'Image courtesy of Corel(R)',...
    'FontSize',7,'HorizontalAlignment','right')
```

2) *Use the gradient Magnitude as the segmentation function:*

Use the Sobel edge masks, imfilter, and some simple arithmetic to compute the gradient magnitude. The gradient is high at the borders of the objects and low (mostly) inside the objects.

```
hy = fspecial('sobel');
hx = hy';
Iy = imfilter(double(I), hy, 'replicate');
Ix = imfilter(double(I), hx, 'replicate');
gradmag = sqrt(Ix.^2 + Iy.^2);
Figure
imshow(gradmag,[]), title('Gradient magnitude (gradmag)')
L = watershed(gradmag);
Lrgb = label2rgb(L);
figure, imshow(Lrgb), title('Watershed transform of gradient magnitude (Lrgb)')
```

3) *Mark the foreground objects*

Opening is an erosion followed by a dilation, while opening-by-reconstruction is an erosion followed by a morphological reconstruction. Let's compare the two. First, compute the opening using imopen.

```
se = strel('disk', 20);
 Io = imopen(I, se);
figure imshow(Io),
title('Opening (Io)')
```

a) Next compute the opening-by-reconstruction using imerode and imreconstruct.

```
Ie = imerode(I, se);
Iobr = imreconstruct(Ie, I);
figureimshow(Iobr), title('Opening-by-reconstruction (Iobr)')
```

b) Following the opening with a closing can remove the dark spots and stem marks. Compare a regular morphological closing with a closing-by-reconstruction. First try imclose:

```
Ioc = imclose(Io, se);
figure
imshow(Ioc), title('Opening-closing (Ioc)')
```

c) Now use imdilate followed by imreconstruct. Notice you must complement the image inputs and output of imreconstruct.

```
Iobrd = imdilate(Iobr, se);
Iobrcbr = imreconstruct(imcomplement(Iobrd), imcomplement(Iobr));
Iobrcbr = imcomplement(Iobrcbr);
figure
imshow(Iobrcbr), title('Opening-closing by reconstruction (Iobrcbr)')
```

4) Compute background markers

Now you need to mark the background. In the cleaned-up image, Iobrcbr, the dark pixels belong to the background, so you could start with a thresholding operation.

```
bw = imbinarize(Iobrcbr);
figure
imshow(bw), title('Thresholded opening-closing by reconstruction (bw)')
```

5) Visualize the result

```
I4 = I;
I4(imdilate(L == 0, ones(3, 3)) | bgm | fgm4) = 255;
figure
imshow(I4) title('Markers and object boundaries superimposed on original image (I4)')
```

V. FEATURE EXTRACTION

In statistical texture analysis, texture features are computed from the statistical distribution of observed combinations of intensities at specified positions relative to each other in the image. According to the number of intensity points (pixels) in each combination, statistics are classified into first-order, second-order and higher-order statistics. The Gray Level Co-occurrence Matrix (GLCM) method is a way of extracting second order statistical texture features.

Statistical Features: A statistical feature is one of the early methods proposed in image processing. The gray level co-occurrence matrix (GLCM) of the ROI was used as suggested by Haralick. The following features are extracted from the GLCM of the ROI kidney images using MATLAB: Energy, Entropy, Contrast, Homogeneity, Maximum probability and correlation, etc.

1) Energy is a measure of local homogeneity and it is calculated using:

Where, *i* and *j* are the pixel values.

Energy=$\mu = (1/MN) * \sum_{i=1} \sum_{j=1} p(i,j)$

2) Entropy measures the average, global information content of an image in terms of average bits per pixel. As the magnitude of entropy increases, more information is associated with the ima*ge*.

Entropy=$f_3=\sum_{i=0}^{N_{g-1}} \sum_{j=0}^{N_{g-1}} p_{d,0}(i,j) \log(p_{d,0}(i,j))$

3) Contrast defines the difference between the lightest and darkest areas on an image.

Contrast=$f_2=\sum_{n=0}^{N_{g-1}} n^2 \left\{\sum_{i=0}^{N_{g-1}} \sum_{j=0}^{N_{g-1}} p_{d,0}(i,j)\right\}$, where $n = |i-j|$

4) Homogeneity is the state or quality of being homogeneous, biological or other similarities within a group.

Homogeneity=$\sum_{i,j=0}^{N-1} \frac{p_y}{1+(i-j)^2}$

5) Correlation is a measure of the strongest of the relationship between two variables.

Correlation= f3=$\sum_{i=0}^{N_{g-1}} \sum_{j=0}^{N_{g-1}} p_{d,0}(i,j) \frac{(i-\mu_x)(j-\mu_y)}{\sigma_x \sigma_y}$

VI. CLASSIFICATION

A. Support Vector Machine

Support Vector Machines (SVM) is a set of related, supervised learning methods used in classification and regression. Given a set of training examples, each marked as belonging to one of two categories, SVM training algorithm predicts whether a new example

falls into a specific category or not. In training datasets, only regions that could be labeled with high confidence are used whereas in testing datasets, all pixels are labeled because manual segmentation process is used for quantitative assessment of the accuracy of testing results. The SVM classification and manually segmented test data are compared so as to evaluate the performance of the SVM. An Accuracy score is computed through identifying the ratio of correctly classified pixels and total pixels in the region of interest.Support vector machine (SVM) is an algorithm that attempts to find a linear separator (hyper-plane) between the data points of two classes in multidimensional space. SVMs are well suited to dealing with interactions among features and redundant features.

Steps for SVM classifier:

1) The goal of a support vector machine is to find the optimal separating hyper plane which maximizes the margin of the training data.
2) We could trace a line and then all the data points representing men will be above the line, and all the data points representing women will be below the line.
3) Find the optimal hyper plane with the margin.
4) We can they find the closest data point.

Fig: 6.1.1 Using SVM classifier

B. Performance Analysis

The performance of classifier depends on various factors like Sensitivity, Specificity and the area under Receiver operating characteristics curve (Lu et al. 2004). These values are calculated by considering confusion matrix, which is given as below:

TP	FN
FP	TN

Where,

TP: Number of True Positives,
FP: Number of False positives,
TN: Number of True negatives,
FN: Number of False negatives.

True Positive (TP): The test result is positive in the presence of the lungs cancer.
True Negative (TN): The test result is negative in the absence of the lungs cancer.
False Positive (FP):The test result is positive in the absence of the lungs cancer.
Sensitivity-It is the statistical measure of how well a binary classifier correctly identifies thepositive cases.

$$\text{Sensitivity}=\frac{TP}{TP+FN}$$

Specificity-It is also a statistical measure of how well a binary classifier correctly identifies thenegative class.

$$\text{Specificity}=\frac{TN}{TN+FP}$$

C. Result Analysis

Proposed method has produced 96.74% of accuracy, 91.60%sensitivity and 84.48 % specificity values. For further research, improvement can be made by selecting another better filter, as well as by investigation the morphological operation in findingthe optimum values for detecting more accurate lungs cancer.

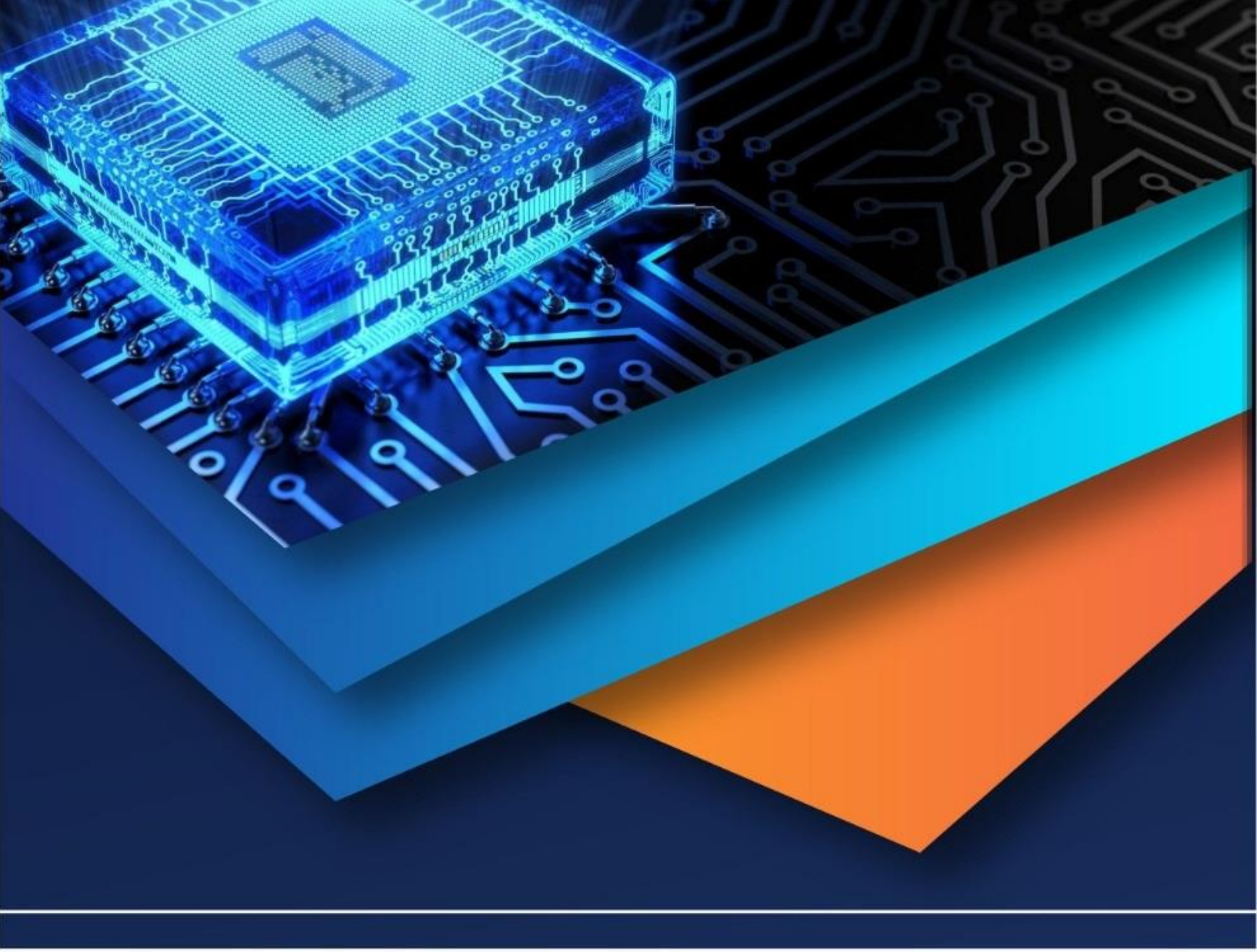
ImageID	Contrast	Correlation	Energy	Homogeneity	Mean	Sensitivity	Specificity	Accuracy%
1	0.1181	0.2007	0.9598	0.9883	0.0015	96.2861	83.3333	96.0758
2	0.1483	0.1396	0.9702	0.9906	0.0018	96.2861	83.3333	96.0758
3	0.2311	-0.0024	0.9906	0.9959	0.0024	96.2861	83.3333	96.0758
4	0.2816	-0.0029	0.9886	0.9950	0.0029	96.2861	84.3333	96.0758

VII. CONCLUSION

This proposed system addresses the image processing techniques to recognize the lung cancer in CT images. Our Proposed system develops an automatic detection of lung cancer in CT images using watershed segmentation, GLDM feature and SVM classifier. Comparatively GLDM features give more accurate results than GLCM. The accuracy of the tumor detected is checked using SVM classification techniques.

REFERENCE

[1] Alizadeh, G., Frounchi, J., BaradaranNia, M., Asgarifar, S. and Zarifi, M.H. "An FPGA Implentation of an Artificial Neural Network for Prediction of Cetane Number", in Proc. IEEE International Conference on Computer and Communication Engineering, Kuala Lumpur, Malaysia, pp. 605-608, 2008.
[2] American Cancer Society, "Cancer Statistics, 2005", CA: A Cancer Journal for Clinicians, http://caonline.amcancersoc.org/cgi/content/full/55/1/10, Vol. 55, pp. 10-30, 2005.
[3] Antonelli, M., Frosini, G., Lazzerini, B. and Marcelloni, F. "Lung Nodule Detection in CT Scans," World Academy of Science, Engineering and Technology 1, pp. 128-131, 2005.
[4] Armato, S.G., Giger, M.L. and MacMahon, H. "Automated Detection of Lung Nodules in CT Scans: Preliminary Results", Medical Physics, Vol. 28, pp. 1552–1561, 2001.
[5] Armato, S.G., Maryellen L.Giger, Catherine J.Moran, James T. Blackburn, KunioDoi, Heber MacMahon "Computerized detection of pulmonary nodules on CT scans," Radiographics, Vol.19, pp. 1303-11, 1999.
[6] Armatur, S.C., Piraino, D. and Takefuji, Y. "Optimization Neural Networks for the Segmentation of Magnetic Resonance Images", IEEE Transactions on Medical Imaging, Vol. 11, No. 2, pp. 215-220, 1992.
[7] Banakar, A. and FazleAzeem, M. "Artificial wavelet neural network and its application in neuro-fuzzy models," Applied Soft Computing, Vol. 8, pp. 1463-1485, 2008.
[8] Banik, S., Rangayyan, R.M. and Desautels, J.E.L. "Detection of architectural distortion in prior mammograms of interval-cancer cases with neural networks", Annual International Conference of the IEEE on Engineering in Medicine and Biology Society (EMBC), pp.6667-6670, 2009.
[9] BaoyuZheng, Wei Qian, and Clarke, L.P. "Multistage neural network for pattern recognition in mammogram screening", IEEE International Conference on IEEE World Congress on Computational Intelligence, Vol. 6, pp. 3437 – 3441, 1994.
[10] Barreno, M., Cardenas, A. A. and Tygar. J. D. "Optimal ROC curve for a combination of classifiers," in Proc. NIPS, pp. 57–64, 2007.
[11] Behnamghader, E., Ardekani, R.D. and Fatemizadeh, E. "Another Approach to Detection of Abnormalities in MR-Images Using Support Vector Machines", 5th International Symposium on Image and Signal Processing and Analysis (ISPA 2007), pp. 98–101, 2007.
[12] Berlin. L. "Liability of performing CT screening for coronary artery disease and lung cancer," American Journal of Roentgenology, Vol. 179, pp. 837–42, 2002.
[13] Brown, M.P.S., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, CW., Furey, T.S., Ares, M. and Haussler, D. "Knowledge-based analysis of microarray gene expression data by using support vector machines", Proc. Nat., Acad., Sci., USA, Vol. 97, pp. 262-267, 2000.
[14] Brown. M.S. "Method for segmenting chest CT image data using an anatomical model: Preliminary results," IEEE Trans. Med. Imag., Vol. 16, pp. 828–39, 1997.
[15] Cacoulos, R. "Estimation of a multivariate density", Ann. Inst. Stat. Math. (Tokyo), Vol.18, pp. 179–189, 1966.
[16] Chen, J.J. and White. C. S. "Use of CAD to evaluate lung cancer on chest radiography," J. ThoracImag., Vol. 23, No. 2, pp. 93–96, 2008.
[17] Cheng, H.D., Chen, C.H. and Freimanis, R.I. "A neural network for breast cancer detection using fuzzy entropy approach", International Conference on Image Processing, Proceedings, Vol. 3, pp. 141–144, 1995.
[18] Cheran, S.C. and Gargano, G. "Computer Aided Diagnosis for Lung CT Using Artificial Life Models", Seventh International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, 2005.
[19] Early detection of Lung Cancer, URL http://www.disabled-world.com/health/cancer/lung/detect.php University of Bonn – Published: 2011-05-16.
[20] El Hamdi, R., Njah, M. and Chtourou, M. "Breast cancer diagnosis using a hybrid evolutionary neural network classifier", 18th Mediterranean Conference on Control & Automation (MED), pp. 1308 – 1315, 201



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY