International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 3, March - 2014

Feature Extraction and Classification of Lung Cancer Nodule using Image Processing Techniques

Khin Mya Mya Tun
M.E Thesis Student, Department of Electronic Engineering
Mandalay Technological University
Mandalay, Myanmar

Aung Soe Khaing
Associate Professor, Department of Electronic Engineering
Mandalay Technological University
Mandalay, Myanmar

***Abstract*— The common cause of death among people throughout the human race is lung cancer. In this paper, median filter is used for image pre-processing. For segmentation, Otsu's thresholding method is used. In feature extraction, physical dimensional measures and gray-level co-occurrence matrix (GLCM) method are used. Artificial neural network (ANN) is applied for classification of disease stages. CT (computed tomography) scan image is suitable for lung cancer diagnosis. This paper is to implement feature extraction and classification of lung cancer nodule using image processing techniques. To implement the algorithm, MATLAB software is developed. This technique can help radiologists and doctors to know the condition of diseases at early stages and to avoid serious disease stages for lung cancer patients.**

Keywords*— Median filter,Otsu's thresholding, GLCM, ANN, MATLAB

I. INTRODUCTION

Mortality from lung cancer are expected to continue rising, to become around 17 million worldwide in 2030. Early detection of lung cancer can increase the chance of survival among people. There are many techniques to diagnose the lung cancer, such as Chest Radiograph (X-ray), Computed Tomography (CT), Magnetic Resonance Imaging (MRI scan) and Sputum Cytology. However, most of these techniques are expensive and time consuming. Therefore, there is a great need for a new technology to diagnose the lung cancer in its early stages. Image processing techniques provide a good quality tool for improving the manual analysis.

The lungs are a pair of sponge-like, cone-shaped organs [1]. The right lung has three lobes, and is larger than the left lung, which has two lobes. Anatomy of lung is shown in Fig.1. Lung cancer is a disease of abnormal cells multiplying and growing into a nodule. Fig.2 describes the beginning of the cancer. The types of lung cancer are divided into four stages. In stage I, the cancer is confined to the lung. In stages II and III, the cancer is confined to the chest (with larger and more invasive tumors classified as stage III).Stage IV cancer has spread from the chest to other parts of the body.

In previous technical literature done by A.Amutha and R.S.D Wahidabanu [3], Level Set-Active Contour Modeling was used as a method in diagnosing lung tumor. First step was removing noise from image using kernel based non-local neighborhood denoising function and done feature extraction based on histogram to classify between normal and abnormal classes. At the final step or in tumor detection, level set-active contour modeling with minimized gradient to the image was used. In another study [4], Autoenhancement, Gabor filter and Fast Fourier transform (FFT) were used to enhance the image and used Thresholding and Watershed segmentation to segment the image. While for feature extraction, Binarization and Masking approach were applied. N.A. Memon et. al [5] proposed thresholding method which select the threshold based on the object and background pixel means. Region growing is used then to extract the exact cavity region with accuracy. In this paper presents image collection, image preprocessing, image segmentation, feature extraction, and classification of disease stages.

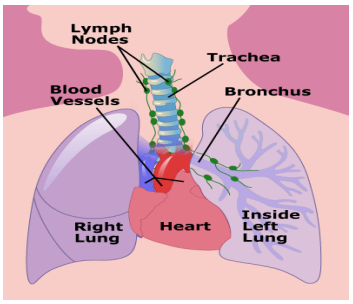

Fig. 1. Anatomy of lung [2]

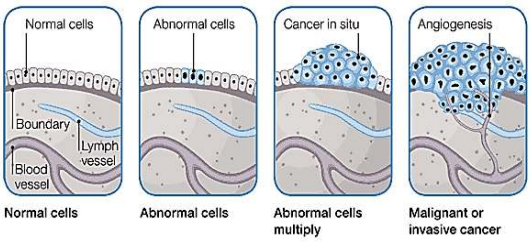

Fig.2. The beginning of cancer [2]

II. METHODS

In this paper, the system presents five basic steps. The first step starts with taking a gathering of CT images (stage1, stage2, stage3 and stage4) from the available database. The second step applies median filter for image pre-processing to get best level of quality and clearness. The third step is image segmentation which uses Otsu's thresholding method and the fourth step contains the calculation of feature extraction .The final step is the classification of disease stages using neural network. Fig.3 illustrates a block diagram of lung cancer nodules feature extraction and classification of this system.

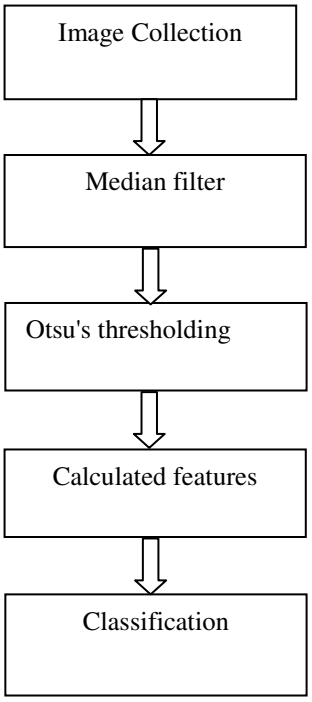

Fig. 3.Block diagram of Lung Cancer Nodule Feature Extraction and Classification System

A. *Image Collection*

The foremost step in medical image processing is collection of images. The lung CT images are collected from Mandalay General Hospital in Myanmar. The medical data is usually in DICOM format, which is the standard for storage and transfer of medical images [6]. Computed tomography (CT) images have better clarity, low noise and distortion for lung diagnosis. So, CT scan of lung images are given as input for this system. Dimensions of images are 512x512 pixels in size. Fig.4. shows the original CT lung image with nodule. The input of CT image contains noises such as white noise, salt and pepper noises etc. Therefore, image preprocessing stage is needed to eliminate noises.

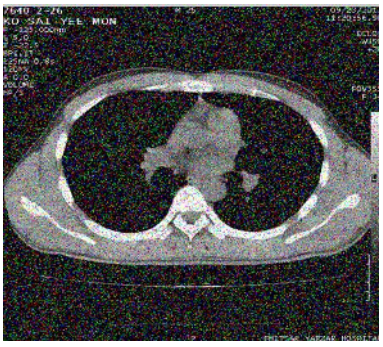

Fig.4.Original CT lung image with nodule

B. *Median filter*

Median filter is one of the filter methods of image pre-processing. Image preprocessing is a way to improve the quality of image, so that the filtered image is better than the original one. The median filter is a non-linear tool. Mean filtering is a simple, intuitive and easy to implement of smoothing images, *i.e.* reducing the amount of intensity variation between one pixel and the next, than other filters. The median filter is normally used to reduce salt-and-pepper noise in an image. It often does a better job than the other filters of preserving useful detail in the image. The median is calculated by first sorting all the pixel values from the surrounding neighborhood into numerical order and then replacing the pixel being considered with the middle pixel value [7]. If the neighborhood under consideration contains an even number of pixels, the average of the two middle pixel values is used. In general, 3x3 mask size of filter is mostly used. In this paper, the mask size of filter is 15x15 because the larger the mask size, the more eliminate the noise. The output of median filtered image is shown in Fig.5. Median filter is used to remove the noise of images. This filtered image is used as the input for image segmentation.

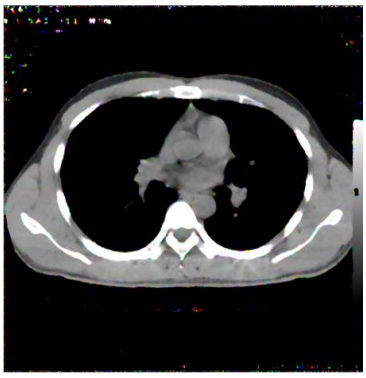

Fig.5. Output of median filtered image

C. *Otsu's thresholding*

Thresholding is one of the most powerful tools for image segmentation. Thresholding method has the advantages of smaller storage space, fast processing speed and ease in manipulation, compared with gray level image which usually contains 256 levels [8]. It is divided into two approaches: global thresholding and local thresholding . Otsu's thresholding method is one of the global thresholding Otsu's thresholding is a non-linear operation and converts a gray-

scale image into a binary image where the two levels are assigned to pixels that are below or above the specified threshold value[9].

$$g(x,y) = \begin{cases} 1 & f(x,y)>T \\ 0 & f(x,y) \le T \end{cases} \quad (1)$$

where, g(x,y)=output image
f(x,y) =input image
T = threshold value

It is based on threshold range by statistical. Otsu suggested minimizing the weighted sum of within-class variances of the object and background pixels to establish an optimum threshold. Recall that minimization of within-class variances is equivalent to maximization of between-class variance [10]. Threshold value based on this method is between 0 and 1.

Otsu's thresholding method is based on selecting the lowest point between two classes (peaks). Frequency and Mean value are the following equations to be calculated.

Frequency:

$$\omega = \sum_{i=0}^{T} P(i) \quad , P(i) = \frac{ni}{N} \quad (2)$$

Where, N=total pixel number
n_i= number of pixels in level i

Mean:

$$\mu = \sum_{i=0}^{T} iP(i)/\omega \quad (3)$$

The variation of the mean values for each class from the overall intensity mean of all pixels:

between-classes variance σ_b^2 ,

$$\sigma_b^2 = \omega_0 (\mu_0 - \mu_t)^2 + \omega_1(\mu_1 - \mu_t)^2, \quad (4)$$

Substituting $\mu_t = \omega_0 \mu_0 + \omega_1\mu_1$, (5)

$$\sigma_b^2 = \omega_0\omega_1(\mu_1 - \mu_0)^2 \quad (6)$$

$\omega_0, \omega_1, \mu_0, \mu_1$ stands for the frequencies and mean values of two classes, respectively. Derived from this method, threshold value represents between 0 and 1 and the segment of image will be achieved. In this paper, the Otsu's threshold value is 0.498. Fig.6 shows the output of Otsu's thresholding image.

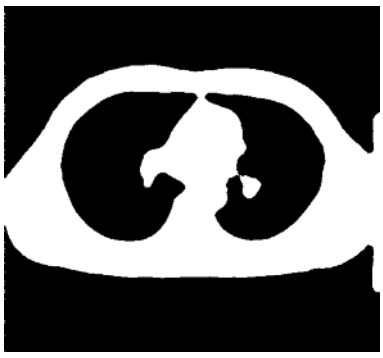

Fig.6. Output of Otsu's thresholding image

After segmentation of images, morphological operations are needed to obtain individual lung and to remove unnecessary parts. Morphology is a technique of image processing based on shapes. A structuring element is a shape mask used in the basic morphological operations [11]. The basic morphological operations are dilation and erosion.

Dilation is an operation that 'grows' or 'thickens' objects in a binary image. The formal definition of dilation of a set *A* by another set *B* is denoted A⊕B, and defined by:

$$A \oplus B = \left\{ z \,|\, (\hat{B})_z \cap A \neq \phi \right\} \quad (7)$$

Where, B^ is the reflection of B. This definition means that dilation of A by B is done by reflecting B and then shifting B over A by z. Then all the displacements of B are set such that B and A overlap by at least one element, which gives the dilation. Set B is also referred to as the dilation mask or structuring element (STREL).

Erosion is an operation that 'shrinks' or 'thins' objects in a binary image. Erosion produces an opposite effect of dilation. In other words, erosion of A by B is set of all points traversed by center of B such that B is totally contained within A at all times.

$$A \Theta B = \left\{ z \,|\, (B)_z \cap A^C \neq \phi \right\} \quad (8)$$

Applying morphological operations on the output of Otsu's thresholding image, it removes the unwanted parts of background shown in Fig.7.This paper focuses on region of interest (ROI) is only lung nodule. So, erosion is further used to get the output of segmented lung nodule shown in Fig.8.

Fig.7. Output of Morphological operation for abnormal lung

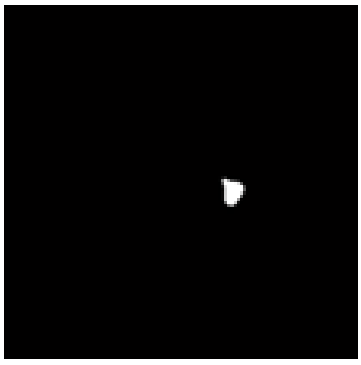

Fig.8. Output of segmented lung nodule

D. *Feature extraction*

After the segmentation is performed, the segmented lung nodule is 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 geometric and intensity-based statistical features are extracted. Shape measurements are physical dimensional measures that characterize the appearance of an object. Only these features were considered to be extracted; area, perimeter and

eccentricity. The physical dimensional measures are defined as follows:

1) Area: The area is obtained by the summation of areas of pixel in the image that is registered as 1 in the binary image obtained [12].

$$A=n\{1\} \quad (9)$$

where, n{ } represents the count of number of the pattern within the curly brackets.

2) Perimeter: The perimeter [length] is the number of pixels in the boundary of the object. Perimeter P is measured as the sum of the distances between every consecutive boundary points [13]. Mathematically,

$$P = |s_n s_1| + \sum_{i=1}^{n-1} |s_i s_{i+1}| \quad (10)$$

where, $s=\{s_1,\ldots,s_n\}$ is a set of the boundary points

3) Eccentricity: The eccentricity is the ratio of the distance between the foci of the ellipse and its major axis length. The value is between 0 and 1.

After calculating of physical dimensional measure, then texture feature extraction is also calculated on the quantized image by using Gray level co-occurrence matrix (GLCM) method, one of the most known texture analysis method. A gray level co-occurrence matrix is a second order statistical measure introduced by Haralick [14]. GLCM is the gray-level co-occurrence matrix (GLCM), also known as the gray level spatial dependence matrix. The Gray-Level Co-occurrence Matrix (GLCM) is based on the extraction of a gray-scale image. The GLCM functions characterize the texture of an image by calculating how often pairs of pixel with specific values and in a specified spatial relationship occur in an image, creating a GLCM, and then extracting statistical measures from this matrix[15]. Statistical parameters calculated from GLCM values are as follows:

4) Entropy: the statistical measure of randomness that can be used to characterize the texture of the input image.

$$\text{Entropy} = -\sum\sum p(i,j)\log p(i,j) \quad (11)$$

Where, p is the number of gray-level co-occurrence matrices in GLCM .

5) Contrast: Measures the local variations in the GLCM. It calculates intensity contrast between a pixel and its neighbor pixel for the whole image. Contrast is 0 for a constant image.

$$\text{Contrast} = \sum\sum (i-j)^2 p(i,j) \quad (12)$$

Where, P(i,j) = pixel at location (i,j)

6) Correlation: Measures the joint probability occurrence of the specified pixel pairs.

$$Correlation = \frac{\sum_{i=0}^{G-1}\sum_{j=0}^{G-1}(i-\mu_i)(j-\mu j)p(i,j)}{\sigma_i\sigma_j} \quad (13)$$

7) Energy: Provides the sum of squared elements in the GLCM. It is also known as uniformity or the angular second moment.

$$\text{Energy} = \sum\sum (p(i,j))^2 \quad (14)$$

8) Homogeneity: Measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal.

$$Homogeneity = \sum_{i,j} \frac{P(i,j)}{1+|i-j|} \quad (15)$$

In this paper, eight features are extracted from the segmented image in Fig.9.These information of the feature extraction is used as the input to classify the lung cancer stages.

```
The  statistical information of extracted features are
            Area: 14
    Eccentricity: 0.8399
       Perimeter: 13.0711

Entropy is 0.002679
The  GLCM features are
       Contrast: 0.0056
    Correlation: 0.7058
         Energy: 0.9995
    Homogeneity: 0.9999
```

Fig.9.Extracted Feature Values from the Lung Nodule Image

E. Classification

Classification is the final step of determination of disease stages to have lung cancer nodule or not of the patient lung. Artificial neural network (ANN) is one of the classification methods commonly used in image processing techniques.ANN is collections of mathematical models that emulate the real neural structure of the brain[16]. ANN has three layers. They are input layer, hidden layer and output layer. Architecture of a general ANN is shown in Fig.10.

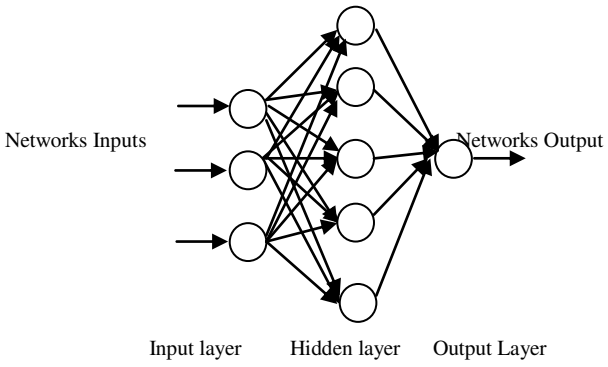

Fig.10. Architecture of a general ANN

The common terminologies used in ANN include weight, bias and activation functions. In this paper, feed-forward neural network is used. The feed-forward neural networks are the simplest type of artificial neural networks devised. In this network, the features information moves in only one direction, forward from the input nodes, through the hidden nodes (if any) and to the output nodes. The input layers consider eight features from the feature extraction step. The output layers contain four stages. The hidden layers present 20 layers. Transfer function is used log-sigmoid which is more suitable than other transfer function for this research. The train images

contain four images for stage 1, stage2, stage3 and stage 4. The extracted features for train images describe in Table.1.

TABLE I. INPUTS OF EXTRACTED FEATURES FOR TRAINING

	Stage1	Stage2	Stage3	Stage4
Area	206	341	491	608
Perimeter	54.2843	77.5980	94.5269	109.0122
Eccentricity	0.7270	0.6897	0.7909	0.9225
Entropy	0.0092379	0.014346	0.01967	0.023641
Contrast	0.0056	0.0101	0.0131	0.0165
Correlation	0.9271	0.9207	0.9286	0.9275
Energy	0.9983	0.9972	0.9960	0.9950
Homogeneity	0.9999	0.9998	0.9998	0.9997

After creating and training the neural from the train file including the known lung CT images and the test file including the unknown lung CT images. Fig.11. is the training of neural network. And then the training process starts, mean square error and epochs graph can be seen. When the mean square error reaches to zero or the training time reaches to the defined epochs, the training program stops automatically. The graph of mean squared error (mse) and epochs is shown in Fig.12. The elapsed time for training this program is 0.01 seconds. The test features data sets consist of ten images. The tested image and a result box appear as shown in Fig.13.The identification result obtains using the neural network approach the success of its efficient use lung cancer detection system.

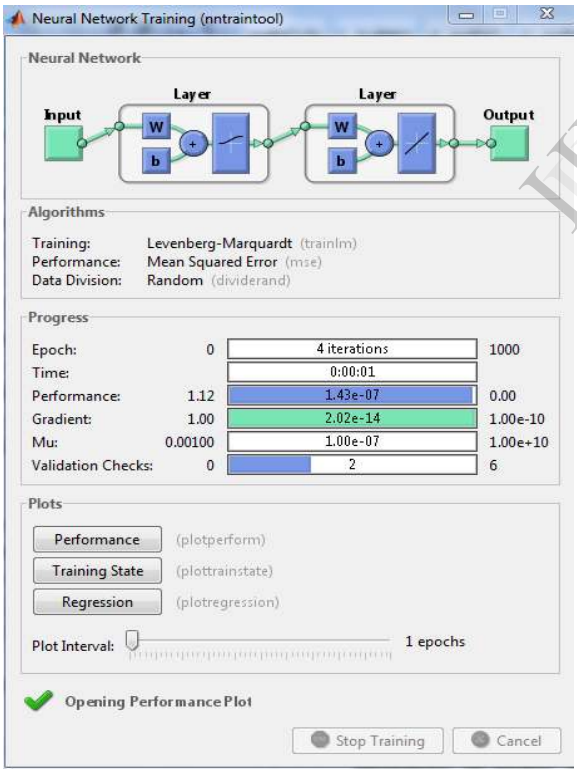

Fig.11.Neural Network Training

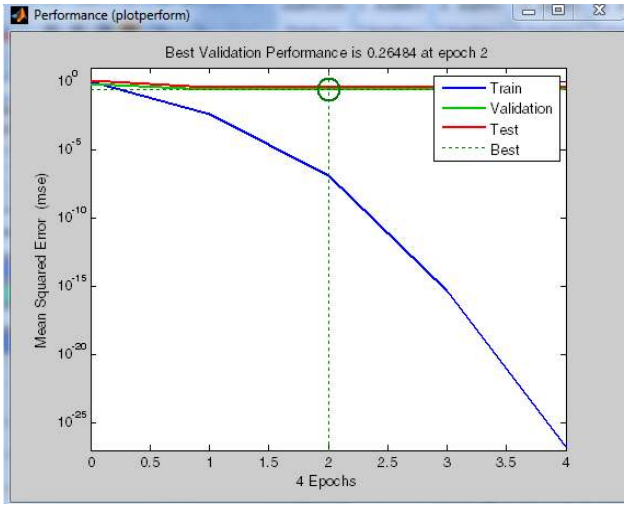

Fig.12.Mean Squared Error Vs Epoch Graph

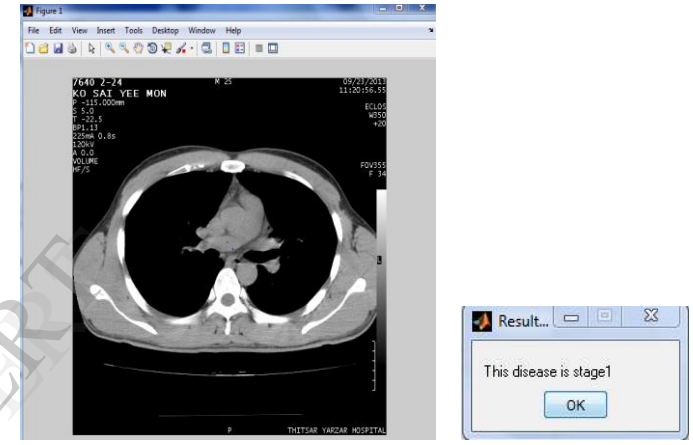

Fig.13. (a) Tested Image.(b)Result Box

III. DICUSSIONS

In the preprocessing step, the median filter is used to remove noise from the lung CT original image. Median filter is more suitable than other filters for this research because the main advantage of median filtering is that even after pixel intensity values are changed the edges of the images are preserved. The increasing mask size is more effective in minimizing the impact of noise. So, the median filter is chosen for this research.

In segmentation step, Otsu's thresholding converts a gray scale image to binary image which threshold value is 0.498 for this paper. This value is autothreshold value using 'graythresh' command in MATLAB. Therefore, Otsu's thresholding method is straightforward than other segmentation methods. After segmentation of image, morphological operation is applied to get individual lung and to eliminate unnecessary parts using erosion and dilation. By doing morphological operations, it gets not only the individual lung but also apparent the lung nodule. For feature extraction step, features are calculated from their formulas and they are used for classifying the disease stages of the lung nodule. For classification step, feed-forward neural network is used. After

training, the test set of unknown categories of lung CT images is passed through the ANN classification system. The 20 hidden layers and log-sigmoid transfer function are appropriate to improve the correct classify of the disease stages.

Finally, the performance of the system is evaluated. The following equations in (16) and (17) are to evaluate the correct classification and the incorrect classification of this system.

$$Correctclassification = \frac{TP+TN}{TP+TN+FP+FN} \text{x}100\% \qquad (16)$$

$$Incorrectclassification = \frac{FP+FN}{TP+TN+FP+FN} \text{x}100\% \qquad (17)$$

Where, TP=True Positive; the image has positive result and the diagnosis system has the disease,
TN=True Negative; the image has negative result and the diagnosis system does not have the disease,
FP=False Positive; the image has negative result and the diagnosis system has the disease, and
FN=True Negative; the image has positive result and the diagnosis system does not have the disease,

```
Percentage Correct classification   : 90.000000%
Percentage Incorrect classification : 10.000000%
>>
```

Fig.14.Performance Evaluation

In this research, the train images are four images (stage 1, stage 2, stage3 and stage4).The test images are 10 images; five images are true positive (TP) images and the rest images are true negative (TN).When the testing of images, TP images can be exactly correct classify. FP image is found 1 time and FN is not found for incorrect classification. Therefore, this system offers correct classification of lung disease stage is 90% and incorrect classification is 10% shown in command window of MATLAB in Fig.14. The result of this system and the analysis of doctor together increase the accuracy of detecting lung cancer nodule. If the lung cancer is detected in its earlier stage, then the chance of surviving the patient increases.

IV. CONCLUSIONS

The mortality rate of lung cancer is the maximum among all other types of cancer. In this paper, image preprocessing and image segmentation are implemented to obtain the diagnosis result. By using these steps, the nodules are detected and some features are extracted. The extracted features are calculated for classification of disease stages. In classification, feed-forward neural network is used to classify the lung cancer stages. This system can know the condition of lung cancer at early stages, so it can play a very important and essential role to avoid serious stages and to reduce the percent of lung cancer distribution in the humanity. This technique helps the radiologists and the doctors by providing more information and taking correct decision for lung cancer patient in short time with accuracy. Therefore, this method is not expensive and few time consuming.

ACKNOWLEDGEMENT

First and foremost, the author would like to thank her supervisor, Associate Professor Dr. Aung Soe Khaing, for providing his valuable advice, helpful guidance and knowledge. The author sincerely wishes to thank to the Head of Department of Electronic Engineering, Associate Professor Dr.Hla Myo Tun at Mandalay Technological University. The author is very thankful to all her teachers from Department of Electronic Engineering at Mandalay Technological University. Last, the author appreciates the help from Mandalay General Hospital for supporting a large collection of lung CT images which have been valuable for this research.

REFERENCES

[1] Non-Small Cell Lung Cancer, [online available], http://www.katemacintyrefoundation.org, Adapted from National Cancer Institute (NCI) and Patients Living with Cancer (PLWC), 2007, accessed on 13 July 2013.
[2] Anatomy of lung picture and beginning of cancer, [online available], www.allreferhealth.com, accessed on 4 June 2013.
[3] A.Amutha and R.S.D Wahidabanu, "A Novel Method for Lung Tumor Diagnosis and Segmentation using Level Set- Active Contour Modelling", European Journal of Scientific Research, Vol.90, No.2, November 2012, pp.175-187
[4] Mokhled S. Al-Tarawneh, "Lung Cancer Detection Using Image Processing Techniques", Leonardo Electronic Journal of Practices and Technologies, Issue 20, January- June 2012, p.147-158. ISSN 1583-1078
[5] N.A. Memon et. al, "Segmentation of Lungs from CT Scan Images for Early Diagnosis of Lung Cancer," World Academy of Science, Engineering and Technology. 2006.
[6] Digital Imaging and Communications in Medicine (DICOM) Part 1: Introduction and Overview, National Electrical Manufacturers Association, 2006, pp. 11.
[7] S Jayaraman, S Esakkirajan and T Veerakumar, "Digital Image Processing",3rd edition, Tata McGraw Hill,2010,ISBN(13):978-0-07-014479-8, ISBN (10): 0- 07014479-6.
[8] Gonzalez R.C., Woods R.E., "Digital Image Processing using MATLAB", Upper Saddle River, NJ Prentice Hall, 2008.
[9] Nunes É.D.O., Pérez M.G., "Medical Image Segmentation by Multilevel Thresholding Based on Histogram Difference", presented at 17th International Conference on Systems, Signals and Image Processing, 2010
[10] Huang Q., Gao W., Cai W., "Thresholding technique with adaptive window selection for uneven lighting image", Pattern Recognition Letters, Elsevier, p. 801-808.
[11] Sudha.V, Jayashree.P.," Lung Nodule Detection in CT Images Using Thresholding and Morphological Operations", International Journal of Emerging Science and Engineering (IJESE) ISSN: 2319–6378, Volume-1, Issue-2, December 2012
[12] Pratt,William K., "Digital Image Processing", New York, John Wiley & Sons,Inc.,1991,p.634
[13] Cigdem Demir and B¨ulent Yener., "Automated cancer diagnosis based on histopathological images: a systematic survey", Technical report, Rensselaer Polytechnic Institute, Department of Computer Science, TR-05-09.
[14] Haralick(1979),"R.M. Statistical and structural approaches to texture", Proceedings of the IEEE,vol.67, pp.786-804.
[15] Haralick,R.M., and L.G.Shapiro., "Computer and Robot Vision",Addison-Wesley,vol.1,p-459,1992
[16] Mark Hudson Beale., Martin T. Hagan., Howard B. Demuth, "Neural Network Toolbox™ User's Guide" ,R2014a Matlab, MathWorks, Inc.

BIOGRAPHIES

Khin Mya Mya Tun was born in Mandalay, Myanmar on 15.3. 1990.She received her Bachelor of Technology (B.Tech) degree in 2011 and Bachelor of Engineering (B.E) degree in 2012 in Electronics Engineering from Mandalay Technological University, Myanmar. She is now Master of Engineering (M.E) thesis student in Mandalay Technological University, Myanmar. Her research interests include bio-medical image processing, texture analysis, and computer vision.

Aung Soe Khaing was born in Pyawbwe Township, Mandalay Division, Myanmar on 27.3.1982.He received Bachelor of Engineering in Electronics from Mandalay Technological University, Mandalay, Myanmar, in 2004 and Master of Engineering in Electronics from Yangon Technological University, Yangon, Myanmar, in 2006. He has continued his PhD dissertation in 2006. From October 2008 to September 2010, he was doing research on Spatial Frequency Analysis of the Human Brain at the Institute of Biomedical Engineering and Informatics, Technical University Ilmenau, Germany. He received his PhD in Electronic Engineering from Mandalay Technological University, Mandalay, Myanmar, in 2011.

He is now Associate Professor at Department of Electronic Engineering, Mandalay Technological University, Myanmar. His research interests include computer based Electrocardiogram (ECG) system, biomedical signal and image processing, bioinstrumentation and telemedicine.

Dr. Aung Soe Khaing was responsible for the ECG laboratory for the biomedical engineering students at the Institute of Biomedical Engineering and Informatics, Technical University Ilmenau, Germany from October 2008 to September 2010.