Pulmonary Lesion Detection and Staging from CT Images Using Watershed Algorithm

Mehak Khatri
Department of Computer Science and Engineering, GZS Campus College of Engineering and Technology, Bathinda, Punjab, India
mehak.khatri20@gmail.com

Munish Kumar
Department of Computational Sciences,
Maharaja Ranjit Singh Punjab Technical University, Bathinda, Punjab, India
munishcse@gmail.com

Abhilasha Jain
Department of Computer Science and Engineering, GZS Campus College of Engineering and Technology, Bathinda, Punjab, India
abd_jain@rediffmail.com

***Abstract*— Nowadays, various image processing methods are broadly being used as a part of the biomedical zones. It is crucial to diagnose the disease and to classify the specific stage for the radiologists to give reasonable remedial to the patients. Lung cancer is the most widely recognized known cancer among individuals, which can be delegated little cell and non-little cell. In this paper, we have proposed a model for the detection of pulmonary lesions at the initial and advanced stages of lung disease on CT (Computed Tomography) images. The proposed framework consists of four stages; change of RGB to grey scale image, smoothing will be performed using median filter to lessen the effect of noise from images, segmentation will be performed using thresholding and watershed techniques and after that the features are extracted for processed image. A framework has been tested with 12,645 images, a dataset of 50 patients. We have noticed that the proposed model perform better than already existing techniques and performance of this model is zero false positive acceptances.**

Keywords- Cell; Computed Tomography; Image Enhancement; Lung; Segmentation.

I. Introduction

Lung cancer is a deadly and most normal leading reason for the cause of deaths around the world. It might create in light of hereditary inclination, abnormal gene mutation which builds the patient's vulnerabilities to cancer-causing stimuli, for example, cigarette smoking, radon gas or different cancer-causing agents. Out of which smoking is the key guideline to contaminate the lungs. There were around 1.69 billion deaths brought about because of lung disease, as indicated by the World Health Organization (WHO). The American Cancer Society (ACS) has evaluated the events of lung disease for 2017 in the United States as; near about 222,500 new cases, out of which 116,990 will be for men and 105,510 will be for women. What's more, near about 155,870 deaths from lung growth, out of which 84,590 will be in men and 71,280 will be for women [1]. Lung carcinoma another name for lung cancer is characterized among little cell and non-little cell. Non-Small Cell Lung Cancer (NSCLC) can be recognized with the proposed model by scanning CT images. There are different frameworks accessible to recognize the infected nodule inside the lung zone, yet likewise we have attempted to accomplish the suitable phase of cancer by using different feature extraction techniques. The proposed model consists of pre-preparing, segmentation, feature extraction and classification of risk stage. The proposed framework is presented by using different image processing techniques. Image processing in the field of medical science is formally known as medical imaging. Various techniques of image processing are being used with medical sciences as to detect and examine the foundation of the infection or issue of the patients. Cancer is a terrible disease which is crucial to detect in early stages, so that specialists or experts to manage the patients and analyze the disease timely. Image processing is also called as imaging science is the handling of images by using a few strategies, operations and methods. There are different operations can be executed as sharpening, image smoothing, image enhancement, image segmentation and so forth. In the planned framework the CT scanned image is the input for the model, the picture is then pre-processed, segmentation is performed after pre-preparing, Feature are extracted by using distinct properties and by diagnosing these properties, classification of stages is assessed. The public dataset has been used for training the model which is obtained from public lung imaging library. Database contains a number of CT scan images that highlight a hefty portion of the key issues in measuring infected nodules or clusters in the lung. The model has the functionality to upload image for detection. The proposed framework has been tested with public dataset provided by VIA and I-ELCAP. It comprises of images of 50 patients caught in single relax.

II. Related work

Penedo et al. [2] have presented an automatic CAD system for radiographic images of the thorax using artificial neural network based approach. The proposed system is intended for detecting nodules in the early or initial stages. The curvature peak space has been exhibited for order of different anatomical structures. Werghi et al. [3] have proposed techniques for sputum cell detection in their initial stages of lung disease. The detection has been proposed by using Bayesian classification, by using thresholding approach and histograms. By observing colour quantization in bigger histograms, the cell detection is performed. The mean shift technique and k-mean clustering is used for the segmentation of sputum cell. Taher et al. [4] have presented a CAD system for early lung cancer detection based on analysis of sputum colour images. The artificial neural network and support vector machine classification techniques have been used for training and testing the system. Different parameters are used for performance analysis such as sensitivity, precision, specificity and accuracy. The ROC (receiver operating characteristic) curve has been used for the assessment purpose. A set of different features like nucleus to cytoplasm ratio, curvature, eigenvectors ratio and density of nucleus region were extracted from nucleus region. 97% high accuracy was evaluated by SVM over ANN technique.

978-1-5386-6678-4/18/$31.00 ©2018 IEEE

And 3% error rate was identified over 10% of ANN. Image enhancement methods for tumor detection in the lung area have been presented by Dimililer et al. [5]. They have applied thresholding over images to clear the highly contrasting pixels out of the images. After that, they applied erosion and median filtering techniques, for noise removal, and then image subtraction is to be performed for removal of small objects from the median filtered images, to locate the tumor. Ng et al. [6] have proposed a watershed algorithm with texture based post-segmentation merging, which is fit for creating the noteworthy decreases in over-division maps acquired utilizing watershed without the proposed texture based post-division consolidating process. Sivaramakrishna et al. [7] have proposed an algorithm for the automatic detection and segmentation of lung nodules in pulmonary CT images. The algorithm proposed is the collaboration of image processing techniques with pattern-recognition and ROC curve analysis. They have observed the performance of the algorithm with 89% of sensitivity. Armato et al. [8] have proposed a computerized method for automated detection of pulmonary nodule on helical thoracic CT scans. Gray-level thresholding segmentation approach has been used to segment the lung area in their work. Multiple gray-level thresholds have been applied in conjunction with volume criterion to identify three dimensional nodule candidates. Fahoum et al. [9] have also presented an automated intelligent system for nodule detection. The system is capable to detect the size of the tumour; margin is one of the features which have been used for the diagnosis purposes that represent the tumour size. The system calculates the width and length distances of the cancer area. Tarawneh [10] has presented image improvement techniques for the detection of cancer in the paper. Image quality assessment as well as enhancement has been adopted on low pre-processing techniques based on Gabor filter within Gaussian rules. Efficient segmentation principles have been proposed for detection of ROI (region of interest) area, and feature extraction techniques are used for detecting particular stage. Fadel et al. [11] have proposed a system for lung nodule detection in CT scan image by using Matlab. In first stage, auto thresholding is used for image extraction from CT scan image. After that connected component labelling and lung masking is to be performed. In second stage, lung nodule extraction using thresholding is to be performed. After thresholding, various geometric properties (shape features) are calculated for nodule selection. Abdullah et al. [12], have conducted research aims to create pulmonary object nodules in three dimensions (3D). The nodules are obtained from CT scan of the extracted lung images reconstruction. The techniques employed in this study are bit-plane slicing, thresholding, elimination and outlining techniques. Reconstruction methods to form a 3D modelling are done by combining the extracted image slices. The result of the extracted images is packed in a simple animation to assist the doctors in diagnosing the nodules whether the nodules are benign or malign.

III. MATERIAL AND METHODS

The proposed framework for cancer detection comprises of different stages like, image acquisition, pre-processing, image enhancement, image segmentation, feature extraction, staging and classification.

A. Dataset

The dataset used for training the model is gotten from the public library provided by VIA (Cornell University Vision and Image Analysis Group) and I-ELCAP (International Early Lung Cancer Action Program). It contains various CT scan images that highlight a considerable lot of the key issues in measuring pulmonary lesions or infected nodules inside the lung area. This database was a possibility of collaboration between the ELCAP and VIA research groups. It was made to make accessible a typical dataset that might be utilized for the execution assessment of various computer aided detection systems. The database right now comprises of a picture set of 50 low-measurements archived entire lung CT examines for recognition, the aggregate of 12,645 images. The CT scans were obtained in a single breath hold with a 1.25 mm slice thickness [13].

B. Methodology

The methodology for the proposed framework consists of various activities as discussed on following sub-sections and scheme of the proposed framework is depicted in Fig. 1.

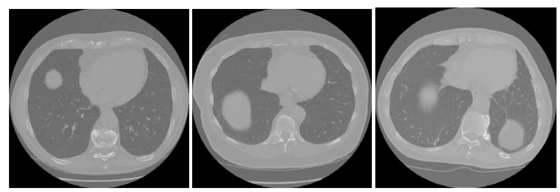

Fig.1. Few samples of dataset

1) *Image Acquistion*

Image acquisition is the process of creating computerized images of an object. In the proposed model, we will consider CT scanned images for the extraction of the lesion from the lung area. CT images have low noise effect as compared to X-Ray and MRI images; therefore we have considered CT images for developing the technique. The main advantage of computed tomography images is that, it gives better quality with less distortion. Originally CT images obtained from the library were of DICOM file type, which is of .dcm extension. Following are the original images obtained from the library.

2) *Pre-processing*

Pre-processing of image is important for the proposed model as the system can work on jpg files and greyscale images. Pre-processing is required as to convert the images according to the system pre-requisite. Public dataset contains DICOM images are of .dcm format and these files are converted into jpg file format to process the images. CT scanned images are of RGB type but the system processes over greyscale images. So, RGB images are converted into greyscale images.

3) *Image Enhancement*

The main purpose of image enhancement is to improve the quality of an image by adjusting brightness, contrast, sharpness etc for the better perception. It is the critical stages in medical imaging as for detection, analyzing and to diagnose the root of the disease. Following are the enhancements to be performed:

- **Image smoothing:**

First, Image smoothing is to be performed in which median filters are employed to reduce the effect of noise in the images. The noise is of different types such as Gaussian noise, salt and pepper noise, short noise, quantization noise or uniform noise, film grain and anisotropic noise. The system uses a median filter to reduce the noise out of the image. In this filtering, the value of the output pixel is to be calculated by considering the median of neighborhood pixels. Median filters are much more sensitive than other filters when images of outliers (extreme values) are considered. Therefore, Use of median filters rather than averaging or linear filters is better to remove the outliers without affecting the sharpness of the image.

- **Auto enhancement:**

Auto enhancement technique has been used in the attempted system, which automatically adjusts the image brightness, sharpness and contrasts up to some optimal levels.

4) *Image Segmentation*

Image segmentation is the process, in which the image is divided into consecutive segments based on their characteristics as similarities and dissimilarities. Image segmentation is the important step in medical imaging as by segmenting it is simpler to analyze the image. Image segmentation is usually being used to detect and locate the boundaries of an object. Segmentation in medical imaging can be used for various purposes as to locate the tumors, measure the tissue volumes, and so forth.

- **Thresholding**

Using Thresholding approach in the proposed model, the preprocessed and enhanced grayscale image is converted into a binary image. In this method, a threshold value (T) is to be a fixed constant then intensity ($I_{i,j}$) of black and white pixel values is to be compared with the threshold value. If the intensity I, of any pixel is less than that of the threshold value then the pixel is replaced with black pixel and if the pixel value is prominent than that of threshold value then the pixel value is replaced with white.

If ($I_{i,j} < T$) then replace a pixel with black pixel.

If ($I_{i,j} > T$) then replace a pixel with white pixel.

- **Watershed Segmentation:**

Watershed segmentation is a technique to identify the presence of an object by extracting the seed and unique boundaries such as curves, lines, and so forth and can distinguish and isolate the touching boundaries of the objects. The watershed transformation is based on local topography or elevation. Each pixel value in an image is acted as an elevation, in which white pixels have high elevations and black pixels have low elevations. The underlying seed is defined by the user and the algorithm floods basins from the initial point or marker. Watershed transformations can be performed by different methods as by flooding, topographical distance, drop of water principle, topological and inter-pixel watershed. We have connected flooding watershed strategy, in which initial an arrangement of introductory seeds are browsed where flooding ought to begin, then the closest neighbor pixel to each seed is to be set apart by some mark as of the seed. Every one of the pixels that are not yet named will begin again by choosing the seeds from those pixels and rehashes until every one of the pixels are named or stamped.

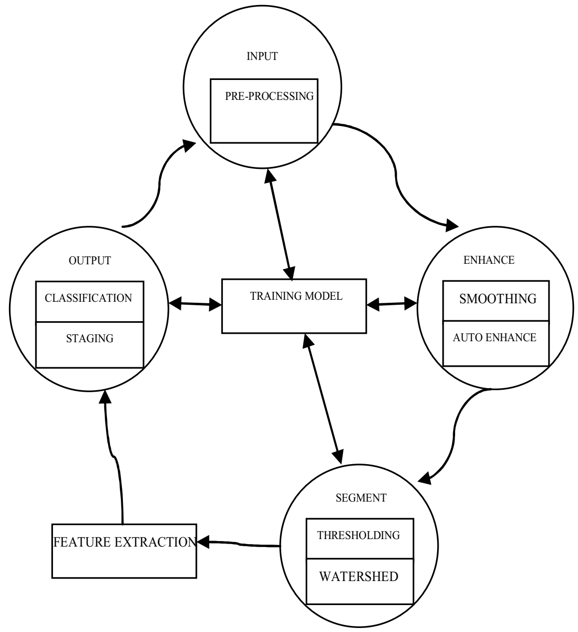

Fig. 2. Block diagram of proposed model

- **Watershed Segmentation:**

Watershed segmentation is a technique to identify the presence of an object by extracting the seed and unique boundaries such as curves, lines, and so forth and can distinguish and isolate the touching boundaries of the objects. The watershed transformation is based on local topography or elevation. Each pixel value in an image is acted as an elevation, in which white pixels have high elevations and black pixels have low elevations. The underlying seed is defined by the user and the algorithm floods basins from the initial point or marker. Watershed transformations can be performed by different methods as by flooding, topographical distance, drop of water principle, topological and inter-pixel watershed. We have connected flooding watershed strategy, in which initial an arrangement of introductory seeds are browsed where flooding ought to begin, then the closest neighbor pixel to each seed is to be set apart by some mark as of the seed. Every one of the pixels that are not yet named will begin again by choosing the seeds from those pixels and rehashes until every one of the pixels are named or stamped.

5) *Feature Extraction*

Feature extraction is as vital as segmentation in medical imaging, which is utilized to recognize the features out of the image. Using regional properties of the binary image we can calculate various parameters according to the prerequisite. Various algorithms and methods are utilized to identify and ascertain the diverse elements. It is the last stage to determine the abnormality and normality of an image. Feature like area, perimeter, centroid and diameter are to be ascertained from the recognized area of abnormal nodule of clot to decide the diverse phases of lung cancer disease. To process the classification and staging accurately, it is crucial to calculate the appropriate values of the defined features. Following are the parameters to be calculated to extract the pulmonary nodule.

- Area: Area is the total number of pixels in the extracted region. Area is identified by number of foreground pixel in a particular region.
- Perimeter: It is the total number of pixels in the boundary region. It is used to find the two adjoining pixels across the boundary of the region.
- Centroid: Centroid or center of mass is used to calculate the centre of the detected area. Centroid is to be calculated by finding the row and column having pixel value one.

Diameter: Diameter can be defined as the longest distance between two connected points. The particular stage of cancer is to be detected by measuring diameter of the extracted region by using region properties. Each of the above mentioned parameter can be calculated by using region properties method.

IV. Experimental Results

Staging is the last phase of the proposed framework, in which the proposed framework will distinguish the phase of the cancer in patient in view of the extracted features as appeared in Figure 3, comprises of CT scanned image, then the input image is pre-processed in which enhancement is performed and after image noise removal, segmentation is performed in which watershed segmentation is performed and after segmentation cancer spot is detected in lung area and other area is ignored. The identified nodule is the cancer spot and the framework has distinguished the number of nodules and the size of the nodule. Organizing of lung cancer is presently arranged inside four phases. The framework will ascertain the parameters and out of these parameters, distance across is the element which is utilized to recognize the specific stage. Lung cancer stages begin from stage 0 to stage 4[3]. Organize 0 is when growth is set up and has not developed into the adjacent tissues and spread outside the lung. It is named as *Carcinoma in Situ*. In Table 1, we have delineated classifications of cancer stages distinguished in the proposed framework. Table 2, depicts the comparison of methodologies and results between different studies.

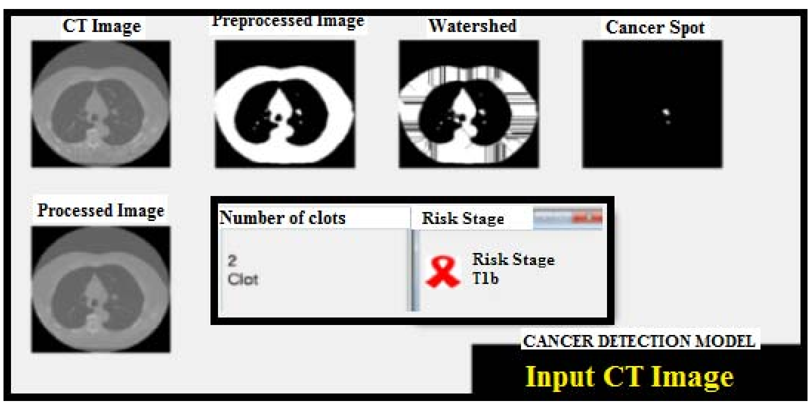

Fig. 3. Working of proposed framework.

Table 1: Classifications of cancer stages

Tumour Stage		Criterion (in Diameters)
T1	T1-A	Less than 30mm
	T1-B	More than 30mm and Less than 50mm
T2	T2-A	Less than 70mm and Greater than 50mm (far from lymph nodes)
	T2-B	Less than 70mm and Greater than 50mm (has spread to lymph nodes)
T3	T3-A	Size greater than 70mm(Inside lung)
	T3-B	Size greater than 70mm (spread into chest area i.e. outside lung)
T4	T4-A	Tumour of any size greater than 80mm and spread within the chest
	T4-B	Tumour of any size greater than 80mm and spread outside the chest

Table 2: Comparison of our proposed work with State-of-the-art work

Authors	Dataset	Methodology	Results
Messay *et al.* [14]	Online dataset contains 456 CT scans from LIDC- IDRI Dataset	Regression Neural Network (RNN).	The performance (average overlap) = 77.58 ± 8.63%.
Brown *et al.* [15]	Offline Dataset of 31 Subject from GE Medical Systems, Milwaukee, WI	Fuzzy Logic.	Efficiency = 86% Average false positives (FP) per case = 11.
Ballangan *et al.* [16]	Offline dataset contains 30 PET-CT provided by GE medical systems.	Fuzzy logic and image processing techniques.	Average DSC (Dice's Similarity Coefficient) = 0.73 (compared to other methods in the study).

Li *et al.* [17]	62,492 online samples obtained from LIDC-IDRI dataset.	Deep Convolution Neural Network (CNN).	Performance based on cross validation (CF) and training data (DD) tests: Accuracy = 0.864, Sensitivity is 0.890.
Magdy *et al.* [18]	Online dataset of 83 CT images of 70 patients obtained from The Cancer Imaging Archive (TCIA).	Amplitude-Modulation Frequency-Modulation (AM-FM) and Linear classifiers.	Linear classifier is concluded with: Accuracy = 95%, Sensitivity = 94% Specificity = 97%.
Lee *et al.* [19]	20 offline clinical cases involving 557 sectional images.	Genetic algorithm (GA) and Convolution template matching.	Detection rate = 72% No. of False positives (FP) = 1.1 per sectional image (approx).
Kim *et al.* [20]	GE offline 827 images from 24 cases obtained from Chungnam National University Hospital.	CAD (computer-Aided Diagnosis), gray-level thresholding and 3D segmentation technique.	Sensitivity = 96% No false-positives
Khatri, Jindal, Jain	Online public dataset from VIA and I-ELCAP, contained 12,645 CT scans of 50 patients.	Watershed Segmentation	No false-positives.

V. Experimental Results

In this paper, a framework for lesion detection using the watershed algorithm has been presented. The proposed framework follows the six phased model, out of which enhancement, segmentation and feature extraction are the core stages to detect the pulmonary nodule in the lung area. Furthermore, by analyzing various parameters in feature extraction, the classification and staging of the cancer are determined. We have utilized open dataset acquired from VIA (Cornell University Vision and Image Analysis Group) and I-ELCAP (International Early Lung Cancer Action Program). It comprises of 12,645 DICOM images of the CT scanner. In this model, first the dicom images are pre-processed and converted in jpg file format and the RGB images are converted into greyscale images. The proposed model also has the possibility to upload the CT scanned images for detection of infected clots. The specific stage will be related to the location and size of nodule present in the lungs. The primary purpose of the model is to identify the pulmonary lesion inside the lung area and to detect the number of nodules as well as size of the cancer clot and to make it simple for specialists to recognize the specific phase of the growth. The system is reliable and tested over different type of CT scanned lung images for detection.

References

[1] "Key Statistics for Lung Cancer", 'American Cancer Society', January 2017. https://www.cancer.org/cancer/non-small-cell-lung-cancer/about/key-statistics.html.

[2] M. G. Penedo, M. J. Carreira, A. Mosquera, and D. Cabello, "Computer-Aided Diagnosis: A Neural-Network- Based Approach to Lung Nodule Detection" *IEEE Transaction on Medical Imaging*, Vol. 17(6), pp. 872-880, 1998.

[3] N. Werghi, C. Donner, F. Taher, and H. Al-Ahmad, "Detection and Segmentation of Sputum Cell for early Lung Cancer Detection", *in the proceedings of the 19th International Conference on Image Processing*, pp.2813-2816, 2013.

[4] F. Taher, N. Werghi and H. Al-Ahmad, "Computer Aided Diagnosis System for Early Lung Cancer Detection", *in the proceedings of the International Workshop on Systems, Signals and Image Processing* (IWSSIP), pp. 5-8, 2015.

[5] K. Dimililer, B. Ugur, Y. K. Ever, "Tumor Detection on CT lung images using image enhancement", *The Online Journal of Science and Technology*, Vol. 7(1), pp. 133-138, 2017.

[6] H. P. Ng, S. Huang, S. H. Ong, K. W. C. Foong, P. S. Goh, and W. L. Nowinski, "Medical Image Segmentation Using Watershed Segmentation with Texture-Based Region Merging", *30th Annual International IEEE EMBS Conference*, pp. 4039-4042, 2008.

[7] R. Sivaramakrishna, J. S. Birbeck, B. E. Higgs, R. M. Cothren,Y. R. Parisky, and W. D. Boswell, "Computer-aided Detection of Lung Cancer from Computer Tomography Images", *Technology Review Journal*, pp.79-92, 2002.

[8] S. G. Armato, M. L. Giger, C. J. Moran, J. T. Blackburn, K. Doi, and H. Macmahon, "Computerized Detection of pulmonary nodules on CT scans", *Imaging & Therapeutic Technology*, Vol. 19, no. 5, pp. 1303-1311, 1999.

[9] A. S. Al-Fahoum, E. B. Jaber, and M. A. Al-Jarrah, "Automated detection of Lung Cancer using Statistical and Morphological Image Processing Techniques", *Journal of Biomedical Graphics and Computing*, Vol. 4, No. 2, pp. 33-42, 2014.

[10] M. S. AL-Tarawneh, "Lung Cancer Detection Using Image Processing Techniques", *Leonardo Electronic Journal of Practices and Technologies*, 20, pp. 147-158, 2012.

[11] M. E. Fadel, R. A. Ali and A. A. Ukasha, "Computerized System For Lung Nodule Detection in CT Scan Images by using MATLAB", *in the proceedings of 3rd International Conference on Automation, Control, Engineering and Computer Science, Engineering & Technology (PET)*, pp. 11-17, 2016.

[12] A. A. Abdullah, Rodiah, E. Haryatmi and Fitrianingsih, " Lung Cancer Nodule Extraction and 3D modelling using Bit-Plane slice and outlining", *International Journal of Engineering & Technology*, Vol. 5 (9), pp. 384- 389, 2016.

[13] "Lung Image Database", 'ELCAP public lung image database', December 2003. https://veet.via.cornell.edu/lungdb.html.

[14] T. Messay, R. C. Hardie, and T. R. Tuinstra, " Segmentation of pulmonary nodules in computed tomography using a regression neural network approach and its application to the Lung Image Database Consortium and Image Database Resource Initiative dataset", *Medical Image Analysis*, Vol. 22(1), pp.48-62, 2015.

[15] M. S. Brown, M. F. McNitt-Gray, J. G. Goldin, R. D. Suh, J. W. Sayre, and D. R. Aberle, "Patient-Specific Models for Lung Nodule Detection and Surveillance in CT Images", *IEEE Transactions on Medical Imaging*, Vol. 20(12), pp. 1242-1250, 2001.

[16] C. Ballangan, X. Wang, M. Fulham, S. Eberl, Y. Yin, and D. Feng, "Automated Delineation of Lung Tumors in PET Images Based on Monotonicity and a Tumor-Customized Criterion", *IEEE Transactions on Information Technology in Biomedicine*, Vol. 15(5), pp. 691- 702, 2011.

[17] W. Li, P. Cao, D. Zhao, and J. Wang, "Pulmonary Nodule Classification with Deep Convolutional Neural Networks on Computed Tomography Images", *Computational and Mathematical Methods in Medicine*, 2016.

[18] E. Magdy, N. Zayed, and M. Fakhr, "Automatic classification of Normal and Cancer Lung CT Images Using Multiscale AM-FM Features", *International Journal of Biomedical Imaging*, 2015.

[19] Y. Lee, T. Hara, H. Fujita, S. Itoh, and T. Ishigaki, "Automated Detection of Pulmonary Nodules in Helical CT Images Based on an Improved Template-Matching Technique", *IEEE Transactions on Medical Imaging*, Vol. 20(7), pp. 595-604, 2001.

[20] D. Y. Kim, J. H. Kim, S. M. Noh, J. W. Parkm, "Pulmonary nodule detection using chest CT images", *Acta Radiologica*, Vol 44(3), pp. 252–257, 2016.