Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 125 (2018) 107–114

Procedia Computer Science

www.elsevier.com/locate/procedia

6th International Conference on Smart Computing and Communications, ICSCC 2017, 7-8 December 2017, Kurukshetra, India

Lung Cancer Detection using CT Scan Images

Suren Makaju[a], P.W.C. Prasad[*a], Abeer Alsadoon[a], A. K. Singh[b], A. Elchouemi[c]

[a]*School of Computing and Mathematics, Charles Sturt University, Sydney, Australia*
[b]*Department of Computer Applications, National Institute of Technology, Haryana, India*
[c]*Walden University, USA*

Abstract

Lung cancer is one of the dangerous and life taking disease in the world. However, early diagnosis and treatment can save life. Although, CT scan imaging is best imaging technique in medical field, it is difficult for doctors to interpret and identify the cancer from CT scan images. Therefore computer aided diagnosis can be helpful for doctors to identify the cancerous cells accurately. Many computer aided techniques using image processing and machine learning has been researched and implemented. The main aim of this research is to evaluate the various computer-aided techniques, analyzing the current best technique and finding out their limitation and drawbacks and finally proposing the new model with improvements in the current best model. The method used was that lung cancer detection techniques were sorted and listed on the basis of their detection accuracy. The techniques were analyzed on each step and overall limitation, drawbacks were pointed out. It is found that some has low accuracy and some has higher accuracy but not nearer to 100%. Therefore, our research targets to increase the accuracy towards 100%.

Keyword: Lung Cancer Detection, CT Scan Image, Cancer, Image Processing

*Corresponding author
E-mail address: CWithana@studygroup.com

10.1016/j.procs.2017.12.016

1. Introduction

Lung cancer is one of the causes of cancer deaths. It is difficult to detect because it arises and shows symptoms in final stage. However, mortality rate and probability can be reduced by early detection and treatment of the disease.
Best imaging technique CT imaging are reliable for lung cancer diagnosis because it can disclose every suspected and unsuspected lung cancer nodules [1]. However, variance of intensity in CT scan images and anatomical structure misjudgment by doctors and radiologists might cause difficulty in marking the cancerous cell [2]. Recently, to assist radiologists and doctors detect the cancer accurately computer Aided Diagnosis has become supplement and promising tool [3]. There has been many system developed and research going on detection of lung cancer. However, some systems do not have satisfactory accuracy of detection and some systems still has to be improved to achieve highest accuracy tending to 100%. Image processing techniques and machine learning techniques has been implemented to detect and classify the lung cancer. We studied recent systems developed for cancer detection based on CT scan images of lungs to choose the recent best systems and analysis was conducted on them and new model was proposed.

2. Literature Review

Several researchers has proposed and implemented detection of lung cancer using different approaches of image processing and machine learning. Aggarwal, Furquan and Kalra [4] proposed a model that provides classification between nodules and normal lung anatomy structure. The method extracts geometrical, statistical and gray level characteristics. LDA is used as classifier and optimal thresholding for segmentation. The system has 84% accuracy, 97.14% sensitivity and 53.33% specificity. Although the system detects the cancer nodule, its accuracy is still unacceptable. No any machine learning techniques has been used to classify and simple segmentation techniques is used. Therefore, combination of any of its steps in our new model does not provide probability of improvement.

Jin, Zhang and Jin [5] used convolution neural network as classifier in his CAD system to detect the lung cancer. The system has 84.6% of accuracy, 82.5% of sensitivity and 86.7% of specificity. The advantage of this model is that it uses circular filter in Region of interest (ROI) extraction phase which reduces the cost of training and recognition steps. Although, implementation cost is reduced, it has still unsatisfactory accuracy.

Sangamithraa and Govindaraju [6] uses K mean unsupervised learning algorithm for clustering or segmentation. It groups the pixel dataset according to certain characteristics. For classification this model implements back propagation network. Features like entropy, correlation, homogeneity, PSNR, SSIM are extracted using gray-level co-occurrence matrix (GLCM) method. The system has accuracy of about 90.7%. Image pre processing median filter is used for noise removal which can be useful for our new model to remove the noise and improve the accuracy.

Roy, Sirohi, and Patle [7] developed a system to detect lung cancer nodule using fuzzy interference system and active contour model. This system uses gray transformation for image contrast enhancement. Image binarization is performed before segmentation and resulted image is segmented using active contour model. Cancer classification is performed using fuzzy inference method. Features like area, mean, entropy, correlation, major axis length, minor axis length are extracted to train the classifier. Overall, accuracy of the system is 94.12%. Counting its limitation it does not classify the cancer as benign or malignant which is future scope of this proposed model.

Ignatious and Joseph [8] developed a system using watershed segmentation. In pre processing it uses Gabor filter to enhance the image quality. It compares the accuracy with neural fuzzy model and region growing method. Accuracy of the proposed is 90.1% which is comparatively higher than the model with segmentation using neural fuzzy model and region growing method. The advantage of this model is that it uses marker controlled watershed segmentation which solves over segmentation problem. As a limitation it does not classify the cancer as benign or malignant and accuracy is high but still not satisfactory. Some changes and contribution in this model has probability of increasing the accuracy to satisfactory level.

Gonzalez and Ponomaryvo [9] proposed a system that classifies lung cancer as benign or malignant. The system uses the priori information and HousefieldUnit(HU) to calculate the Region of Interest(ROI). Shape features like area, eccentricity, circularity, fractal dimension and textural features like mean, variance, energy, entropy, skewness, contrast, and smoothness are extracted to train and classify the support vector machine to identify whether the nodule is benign or malignant. The advantage of this model is that it classifies cancer as benign or malignant, however the limitation of it is that prior information is required about region of interest. Model's classification of benign or malignant using support vector machine can be useful in our new model.

Analyzing the literature reviews, on the basis of accuracy and advantages of the steps used, the system proposed by Ignatious and Joseph [8] is current best solution. In image pre processing it uses Gabor filter to enhance the image and uses marker controlled watershed method for segmentation and detects the cancer nodule. This model also extracts the features like area, perimeter, and eccentricity only of the cancer nodules. It shows the comparison with other previously proposed models and highlights its accuracy 90.1% which is higher than of those.

Even the system is current best solution (fig. 1), it has some limitations. They are highlighted below.

- Only few features has been extracted for cancer nodules
- No preprocessing like noise removal, image smoothing which can probably assists in increasing the detection of nodules accurately has been implemented
- No classification as benign or malignant of extracted cancer has been performed

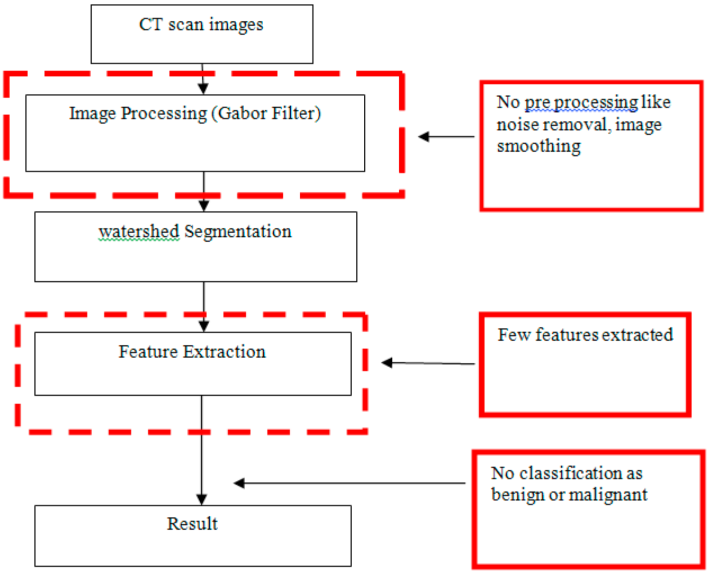

CT scan images
Image Processing (Gabor Filter)
No pre processing like noise removal, image smoothing
watershed Segmentation
Feature Extraction
Few features extracted
No classification as benign or malignant
Result

Fig. 1. Current Best model and its limitations (Ignatious and Joseph,2015)

3. Proposed Model

Changes on current best solution have been made and new model has been proposed as below in the figure 2. Instead of Gabor Filter, Median filter and Gaussian filter have been implemented in pre- processing stage. After pre processing the processed image is segmented using watershed segmentation. This gives the image with cancer nodules marked. In addition to features like area, perimeter and eccentricity, features like Centroid, Diameter and pixel Mean Intensity have been extracted in feature extraction stage for the detected cancer nodules. The best model ends after the detection of cancer nodule, it's feature extraction and calculation of accuracy. But, its classification as benign or malignant has not been implemented. Therefore, additional stage of classification of cancer nodule has been performed using Support Vector Machine. Extracted features are used as training features and trained model is generated. Then, unknown detected cancer nodule is classified using that trained prediction model.

3.2 Image Preprocessing

Firstly, in image pre-processing median filter is used on grayscale image of CT scan images. Some noises are embedded on CT Images at the time of image acquisition process which aids in false detection of nodules. Noise may be detected as cancer nodules sometimes. Therefore, these noises have to be removed for accurate detection of cancer. Median filter removes salt and pepper noise from the CT images [10]. After median filter, Gaussian filter is implemented. It smoothes the image and removes speckle noise from image.

3.3 Segmentation

This process locates objects or boundaries which help in acquiring the region of interest in the image [11]. It partitions the image into regions to identify the meaningful information. In lung cancer detection it segments the cancer nodule from the CT scan image. In the proposed model watershed segmentation is implemented. Its main feature is that it can separate and identify the touching objects in the image. This feature helps in proper segmentation of cancer nodules if it is touching to other false nodules.

3.3 Features extraction

In this stage, features like area, perimeter, centroid, diameter, eccentricity and Mean intensity. These features later on are used as training features to develop classifier.

3.4 Classification

This stage classifies the detected nodule as malignant or benign. Support vector machine (SVM) is used as classifier. It is supervised machine learning method. SVM defines the function that classifies data into two classes [9].The function is defined as $D(x)=w^Tx_i+ b$ where x_i are training inputs, $w^T$is m dimensional vector, and b is bias term. Here, i=1…. M.

$D(x)=w^Tx_i+ b \geq 1$ for $y_i=1$

$D(x)=w^Tx_i+ b \leq -1$ for $y_i=-1$

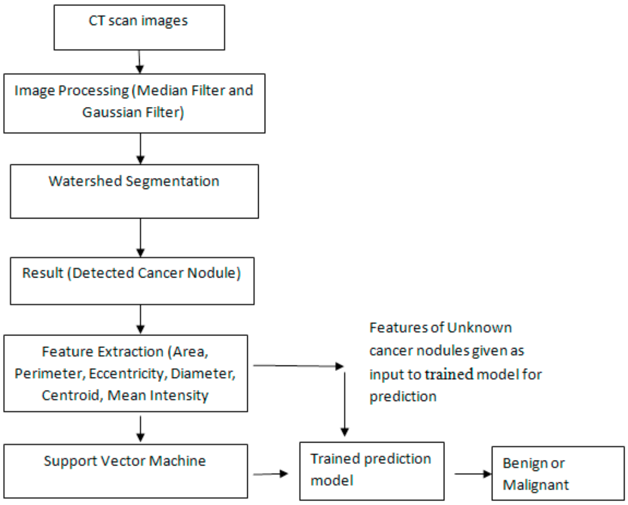

CT scan images
Image Processing (Median Filter and Gaussian Filter)
Watershed Segmentation
Result (Detected Cancer Nodule)
Feature Extraction (Area, Perimeter, Eccentricity, Diameter, Centroid, Mean Intensity
Features of Unknown cancer nodules given as input to trained model for prediction
Support Vector Machine
Trained prediction model
Benign or Malignant

Fig. 2. Proposed model

Main strengths of the proposed model are pointed as below:

- Increase in accuracy of cancer nodule detection than the best current model.
- Classifies the detected lung cancer as malignant or benign.
- Removes salt-pepper noises and speckle noise that creates false detection of cancer

Together with strength, model has some weakness too. They are pointed as below:

- There is increase in the accuracy but still it has not reached to best level i.e. nearer towards 100%
- It classifies the cancer as just malignant or benign but does not classify into different stages like stage I, II, III, IV.

4. Implementation

For implementation, real patient CT scan images are obtained from Lung Image Database Consortium(LIDC) archive [12]. It is the database of lung cancer screening CT images for development, training, and evaluation of computer assisted diagnostic methods for lung cancer detection and diagnosis. It was initiated by National Cancer

Institute. It consists of 1018 cases of dataset contributed by seven academic center and eight medical imaging companies. Images are in DICOM format with size 512*512 pixel. DICOM format is difficult to process; therefore those images are converted to JPEG Gray scale image using software MicroDicom software. MicroDicom opens the DICOM CT scan images and can also convert to appropriate JPEG format.

The proposed model is then developed in MATLAB R2016a. MATLAB is one of the tools for research development and analysis [10]. Both detection and features extraction are implemented in MATLAB and classification is implemented using machine learning toolbox. Classification learner toolbox aids in developing the trained prediction model from the features extracted easily and very fast. 5 folds cross validation was used to prevent from overfitting during the training process. Different 16 DICOM images from LIDC are used for training the classifier and result is validated using 5 images with total 15 nodules.

Some challenges were faced during our implementation of proposed model. They are pointed as below:

- LIDC database was very large in size (124 GB) which was very tedious to download.
- Cancer annotations for the images of LIDC database was defined in xml file which was very difficult to understand.

5. Result and Evaluation of Implementation

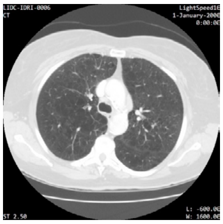


Fig. 3. Original Grayscale image

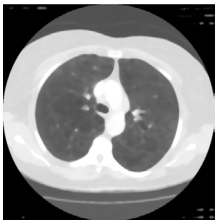


Fig. 4. Median filtered Image

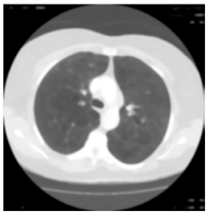


Fig. 5. Gaussian filtered Image

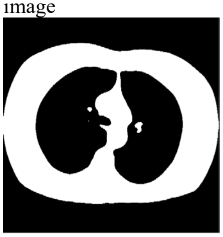


Fig. 6. Binarized Image

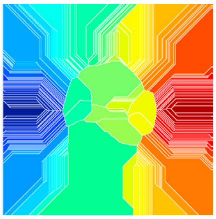


Fig. 7. Watershed segmented Image

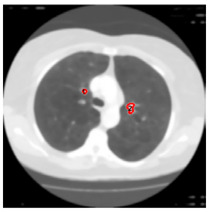


Fig. 8.Cancer marked image

In above, figure 3, 4, 5, 6, 7, 8 are original grayscale image, median filtered image, Gaussian filtered image, binarized image, watershed segmented image and cancer marked image respectively.

Result of all CT scan images are tabularized as below in table 1. Number of true positive nodes, true negative nodes, false positive nodes and false negative nodes detected by the system is recorded

Comparing the accuracy of proposed model with current model it can be seen that there is progressive increase in accuracy from 88.4% to 92%. Sensitivity remained same. Specificity increased from 40% to 50%

From the detected cancer nodes, features like Area, Perimeter, Centroid, Diameter, Eccentricity and Mean Intensity of the Pixels were extracted. Extracted features were used to Train Support vector machine and trained model was developed. Training time for classification learner app was 5.93 seconds. Classification learner app evaluates the prediction time for the developed trained model to be 310 observations per second. Scatter plot of trained model are as below.

Table 1. Comparison analysis of Proposed and Current model

For Proposed model,	**For current best model,**
Total number of nodes detected=23	Total number of nodes detected=24
Number of True Positive (TP)= 21	Number of True Positive (TP) = 21
Number of True Negative (TN)= 2	Number of True Negative (TN) = 2
Number of False Positive (FP) = 2	Number of False Positive (FP) = 3
Number of False Negative (FN) = 0	Number of False Negative (FN) = 0
Accuracy= (TP+TN)/(TP+TN+FP+FN)=23/25=0.92=92.0%	Accuracy= (TP+TN)/(TP+TN+FP+FN)=23/26=0.884=88.4%
Sensitivity=TP/(TP+FN)=21/(21+0)=1=100%	Sensitivity=TP/(TP+FN)=21/(21+0)=1=100%
Specificity=TN/(TN+FP)=2/4=0.5=50%	Specificity=TN/(TN+FP)=2/5=0.4 = 40%

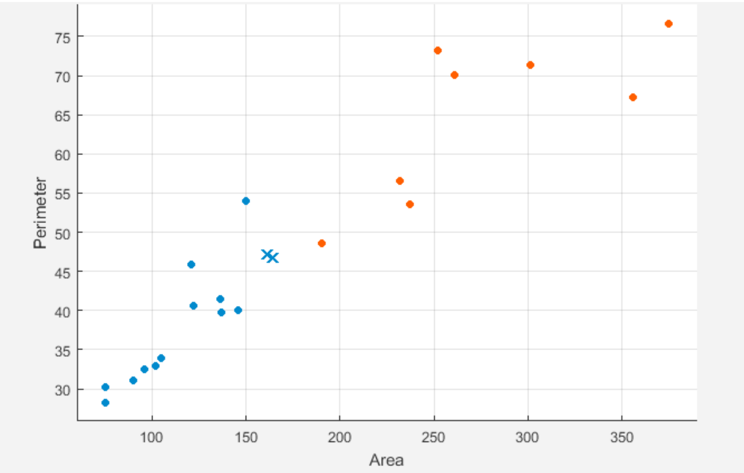


Fig. 9. Scatter plot for Area vs Perimeter of Trained Model

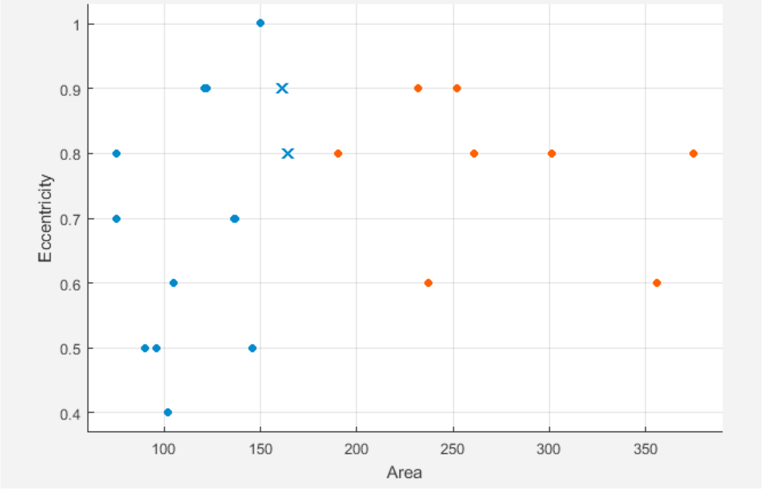


Fig. 10. Scatter plot for Area vs Eccentricity of Trained Model

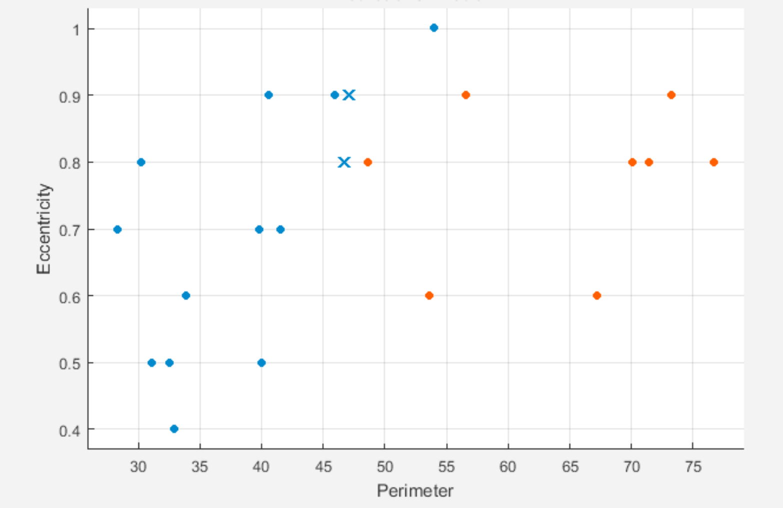
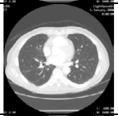


Fig. 11. Scatter plot for Perimeter vs Eccentricity of Trained Model

In above figure 9,10,11 blue dots represent correct benign classes, red dots represent correct malignant classes and cross marking represents incorrect. Result of classification of 5 CT scan images cancer nodules is given in table 2 below.

Table 2: Nodes classification by proposed system

Image		Nodules	Classification	Remark
Image 17		Nodule 1 Nodule 2 Nodule 3	Malignant Malignant Malignant	True True True
Image 18		Nodule 1 Nodule 2 Nodule 3	Benign Benign Benign	True True True
Image 19		Nodule 1 Nodule 2 Nodule 3	Benign Malignant Malignant	True True True
Image 20		Nodule 1 Nodule 2 Nodule 3 Nodule 4	Malignant Malignant Benign Benign	False False True True
Image 21		Nodule 1 Nodule 2	Malignant Malignant	True True

Total number of nodules classified=15
Total number of false classification=2
Total number of True classification=13
Accuracy=13/15=0.866=86.6%
Therefore, from above result we can say that our proposed model classifies as benign or malignant with accuracy of 86.6%. The classification of nodule as malignant or benign which was not performed in the best model has been successfully implemented.

6. Conclusion

The current best model has no satisfactory result of accuracy and does not classify degree of cancer of detected nodules. Therefore new system is proposed. The proposed system is used to detect the cancerous nodule from the lung CT scan image using watershed segmentation for detection and SVM for classification of nodule as Malignant or benign. Proposed model detects the cancer with 92% accuracy which is higher than current model and classifier has accuracy of 86.6%. Overall, we can see improvement in the proposed system in comparison to current best model However, this proposed does not classifies into different stages as stage I, II, III, IV of cancer. Therefore, as future scope improvement in this can be done by implementing classification in different stages. Also, further accuracy can be increased by proper pre-processing and eliminations of false objects.

References

[1] Gindi,A. M., Al Attiatalla, T. A., & Sami, M.M. (2014) "A Comparative Study for Comparing Two Feature Extraction Methods and Two Classifiers in Classification of Earlystage Lung Cancer Diagnosis of chest x-ray images." Journal of American Science, **10(6):** 13-22.

[2] Suzuki, K., Kusumoto, M., Watanabe, S. I., Tsuchiya, R., & Asamura, H. (2006) "Radiologic classification of small adenocarcinoma of the lung: radiologic-pathologic correlation and its prognostic impact," The Annals of Thoracic Surgery. **81(2):** 413-419.

[3] Xiuhua,G., Tao, S., & Zhigang, L.(2011) "Prediction Models for Malignant Pulmonary Nodules Based-on Texture Features of CT Image." In Theory and Applications of CT Imaging and Analysis. DOI: 10.5772/14766.

[4] Aggarwal, T., Furqan, A., & Kalra, K. (2015) "Feature extraction and LDA based classification of lung nodules in chest CT scan images." 2015 International Conference On Advances In Computing, Communications And Informatics (ICACCI), DOI: 10.1109/ICACCI.2015.7275773.

[5] Jin, X., Zhang, Y., & Jin, Q. (2016) "Pulmonary Nodule Detection Based on CT Images Using Convolution Neural Network." 2016 9Th International Symposium On Computational Intelligence And Design (ISCID). DOI: 10.1109/ISCID.2016.1053.

[6] Sangamithraa, P., & Govindaraju, S. (2016) "Lung tumour detection and classification using EK-Mean clustering." 2016 International Conference On Wireless Communications, Signal Processing And Networking (Wispnet). DOI: 10.1109/WiSPNET.2016.7566533.

[7] Roy, T., Sirohi, N., & Patle, A. (2015) "Classification of lung image and nodule detection using fuzzy inference system." International Conference On Computing, Communication & Automation. DOI: 10.1109/CCAA.2015.7148560.

[8] Ignatious, S., & Joseph, R. (2015) "Computer aided lung cancer detection system." 2015 Global Conference On Communication Technologies (GCCT), DOI: 10.1109/GCCT.2015.7342723.

[9] Rendon-Gonzalez, E., & Ponomaryov, V. (2016) "Automatic Lung nodule segmentation and classification in CT images based on SVM." 2016 9Th International Kharkiv Symposium On Physics And Engineering Of Microwaves, Millimeter And Submillimeter Waves (MSMW). DOI: 10.1109/MSMW.2016.7537995.

[10] Miah, M.B.A., & Yousuf, M.A. (2015) "Detection of lung cancer from CT image using image processing and neural network." 2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT): 1-6.

[11] Khobragade, S., Tiwari, A., Patil, C., & Narke, V. (2016) "Automatic detection of major lung diseases using Chest Radiographs and classification by feed-forward artificial neural network." IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES): 1-5.

[12] Armato, I., Samuel McLennan, G., McNitt-Gray, F. R., Michael, Charles, Reeves, Anthony P., … Clarke, Laurenc, (2015) "Data From LIDC-IDRI. The Cancer Imaging" Archive.http://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX.