Asian Journal of Convergence in Technology Volume IV Issue I
ISSN NO: 2350-1146 I.F-5.11

Lung Nodule Detection and Classification using Machine Learning Techniques

Ruchita Tekade[1], Rajeswari Kannan[2]
[1,2]*Computer Department,*
[1,2]*Pimpri Chinchwad College of Engineering*
[1,2]Savitribai Phule Pune University
[1,2]Pune, Maharashtra
[1]ruchitatekade@gmail.com
[2]kannan.rajeswari@pccoepune.org

***Abstract*—As lung cancer is second most leading cause of death, early detection of lung cancer is became necessary in many computer aided dignosis (CAD) systems. Recently many CAD systems have been implemented to detect the lung nodules which uses Computer Tomography (CT) scan images [2]. In this paper, some image pre-processing methods such as thresholding, clearing borders, morphological operations (viz., erosion, closing, opening) are discussed to detect lung nodule regions ie, Region of Interest (ROI) in patient lung CT scan images. Also, machine learning techniques such as Support Vector Machine (SVM) and Convolutional Neural Network (CNN) has been discussed for classifying the lung nodules and non-nodules objects in patient lung ct scan images using the sets of lung nodule regions. In this study, Lung Image Database Consortium image collection (LIDC-IDRI) dataset having patient CT scan images has been used to detect and classify lung nodules. Lung nodule classification accuacy of SVM is 90% and that of CNN is 91.66%.**

Keywords*—Computer Tomography (CT), thresholding, clearing borders, morphological operations, Region of Interest (ROI), Support Vector Machine (SVM), Convolutional Neural Network (CNN), classification, Lung Image Database Consortium image collection (LIDC-IDRI).

I. INTRODUCTION

According to the World Health Organization (WHO), cancer is the second leading cause of death globally, and it causes 8.8 million deaths in 2015. Among those, 1.69 million deaths were caused due to Lung Cancer [1]. It is very difficult to analyze the cancer at its early stage manually. Various Computer Aided Diagnosis (CAD) systems have been designed for the early diagnose of lung tumor. Lung nodule detection plays a very important role in diagnosis of lung cancer. Medical Imaging is helpful to detect lung nodules also useful for treatment of lung cancer. It is more accurate and efficient method for the diagnosis than manual diagnosis methods as they are time consuming. In medical Imaging different types of images are being used, but for the detection of lung nodules, Computed Tomography (CT) images are being preferred for several researches as they give better clarity of lung objects. In this study, the dataset of Lung Image Database Consortium image collection (LIDC-IDRI) [3] has been used. This dataset is initiated by the National Cancer Institute (NCI), further advanced by the Foundation for the National Institutes of Health (FNIH). The database contains a total of 1018 helical thoracic CT scans acquired from 1010 different patients [3]. Using these lung CT scans the objects can be visualized easily. Relative study of lung cancer diagnosis is discussed in section II.

This study is divided into two parts: One is Lung Nodule Detection and other is Lung Nodule Classification. In lung nodule detection, Region of Interest (ROI) are detected and in lung nodule classification the classification of nodules from non-nodules is done. ROI can be detected by image pre-processing methods which are discussed in section III. ROI is nothing but an area where objects like lung nodules are present. These ROI are further provided to SVM and CNN to classify the lung nodules from non-nodule objects. Working of SVM and CNN is discussed in section IV.

The main aim of this study is to compare the lung nodule classification efficiency of SVM and CNN. The comparison parameter of this study is lung nodules classification accuracy. The SVM classifier gives the accuracy of 90% and CNN gives the accuracy of 91.66% for the dataset. CNN is the most efficient algorithm for classification of lung nodules and non-nodules objects in lung ct scan images.

II. LITERATURE REVIEW

In [4], a novel CAD system to detect lung nodules from lung CT scan images is proposed. Lung Image Database Consortium image collection (LIDC-IDRI) dataset [3] has been used for the experiment of this CAD system. This system

is designed for analyzing the 3D structure [14] of lung nodules for better visualization of connections between the objects in the CT scan. Thresholding [4] and morphological closing operations [4] has been applied to lung CT scan to extract lung region. Contrast stretching is also applied for visualizing the CT scan image. Nodule detection and segmentation is done using k-means clustering [4] and morphological opening [4]. Grouping of lung nodules has been done according to the thickness and wall connections like 2D and 3D features [4]. This grouping is used for classification of small and large lung nodules using SVM classifier. This classification is done at accuracy of 96.22%.

In [5], novel SVM classifier has been proposed for classification of lung nodules and non-nodules. This SVM works with the combination of random undersampling (RU) [5] and synthetic minority oversampling technique (SMOTE) [5]. This combination of sampling methods helps to balance the training samples used for classification of lung nodules. It also helps to reduce noise and redundancy of samples. Before RU the noise samples on the boundary of minority class are get removed. The samples are drawn in RU then the like dual drawn out method duplicated samples are removed from sample space. SMOTE algorithm uses similarity measures between features to remove maximum similar samples. And for calculating similarity KNN algorithm [5] is used. After sampling, SVM algorithm is used for classification of lung nodules and non nodules using the sets of lung nodule regions. The classification accuracy of this method is 93.76%.

In [6], Multigroup patch based CNN [6] is proposed for better efficiency of lung nodule detection. In this study, two groups of image patches are used for classification. Lung Image Database Consortium image collection (LIDC-IDRI) dataset [3] is used for lung nodule detection. These lung ct scan images are preprocessed using highlighting lung regions with the help of local maximum variance [6], morphological operations including opening and closing [6], frangi filter [6], etc. After this preprocessing the lung nodule regions are cropped and one group is formed. And another group of patches is of original lung CT scan images. These original image patches group and preprocessed image patches group are provided to CNN for classification of lung nodules and non-nodules objects. This method gives the 94% sensitivity.

In [7], the modified architecture of CNN is given for better efficiency rather than traditional CNN architecture. Typical CNN architecture [17] has convolutional layer, pooling layer and fully connected layer with convolutional and pooling layer applied alternatively. Convolutional layer is responsible for feature extraction of image and pooling layer extracts important features from convolved features and fully connected layer is connected for classification of image after feature extraction. The pooling layer is modified in this study using central pooling [7]. In central pooling the way of applying max pooling is different than traditional pooling. In traditional max pooling operation the fixed size window is put all over the input image and the maximum value in the window is extracted. But in central pooling, the size of window varies for input space. The central pooling is applied row wise and column wise then max pooling is applied. This modification of pooling layer affected on the accuracy of classification and it gives the accuracy of 81.66%.

III. Lung Nodule Detection

Lung CT scan images contains various objects including lungs, vessels, etc. So the ROI has to be detected before classification. The region of our interest is only the areas where lung nodules are present. For this purpose, some image processing methods are applied to lung CT scan imges as shown in Fig 1. The image preprocessing methods are discussed further.

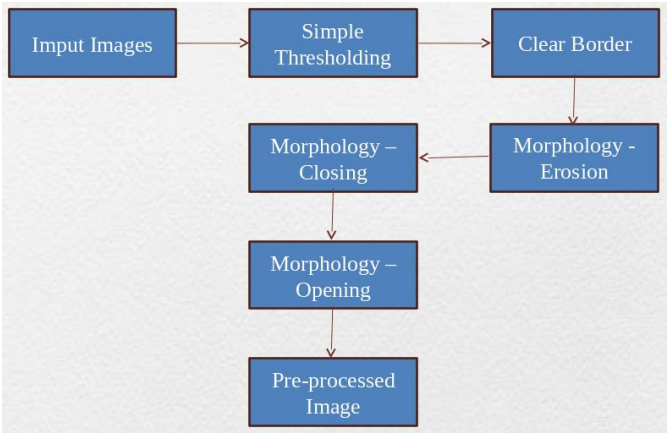

Fig 1 System Model

A. Thresholding

Thresholding is the simplest method in image processing which is mainly used to convert gray scale images to binary images.

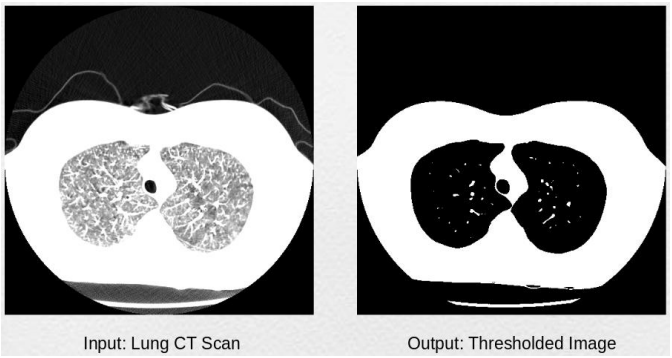

Fig 2 Thresholding

The simplest thresholding methods check for intensity of pixel $I_{i,j}$. If the image intensity $I_{i,j}$ is less than some fixed constant T (that is, $I_{i,j}<T$), then each pixel in an image is replaced by a black pixel, and if the image intensity $I_{i,j}$ is greater than that constant T (that is, $I_{i,j}>T$), then each pixel in an image is replaced by a white pixel. This T is nothing but a threshold set by user according to the requirement of object visualization. The result of thresholding is shown in Fig. 2. Left hand side image shows input lung ct scan image and right hand side image shows the output image after thresholding.

B. Clear Border

After thresholding, the objects connected to the image borders need to be removed as they are non-ROI image objects. Clear Border method has been used to remove such objects. It also suppresses structures that are lighter than their surroundings. The result after clearing border is shown in Fig 3. Left hand side image shows input thresholded image and right hand side image shows the output image after clearing border.

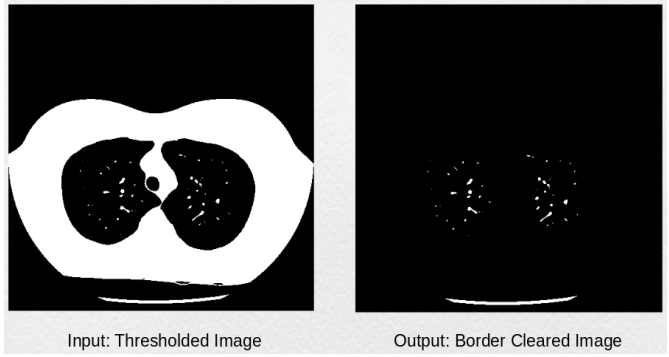

Fig 3 Clear Border

C. Morphology Erosion

The erosion of a binary image f by a structuring element s (denoted as $f \ominus s$) produces a new binary image $g = f \ominus s$. while placing s on f, if s fits f then, g(x,y) = 1 or else, g(x,y) = 0, repeating for all pixel coordinates (x,y).

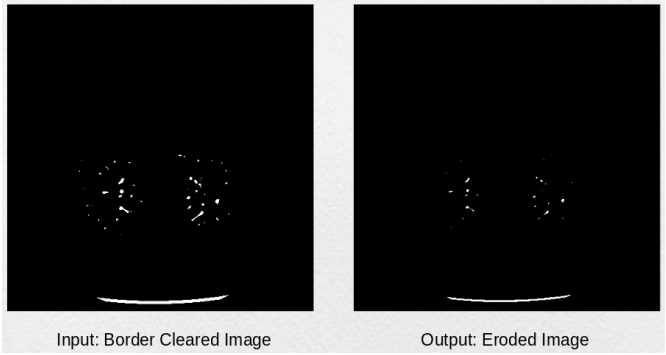

Fig 4 Morphological Erosion

The holes and gaps between different regions become larger, and small details are eliminated. The result after erosion is shown in Fig 4. Left hand side image shows input border cleared image and right hand side image shows the output image after erosion.

D. Morphology Closing

The closing of an image f by a structuring element s (denoted by $f \bullet s$) is a dilation followed by an erosion. The dilation of an image f by a structuring element s produces a new binary image $g = f \oplus s$. If structuring element s hits f, g(x,y) = 1 or else g(x,y) = 0, repeating for all pixel coordinates (x,y). Dilation has the opposite effect to erosion, it adds a layer of pixels to both the inner and outer boundaries of regions. In simple words, dialation is used for increasing size of an object. After applying closing operation, the holes enclosed by a single region and gaps between different regions become smaller, and small intrusions into boundaries of a region are filled in.

The result after closing is shown in Fig 5. Left hand side image shows input eroded image and right hand side image shows the output image after closing.

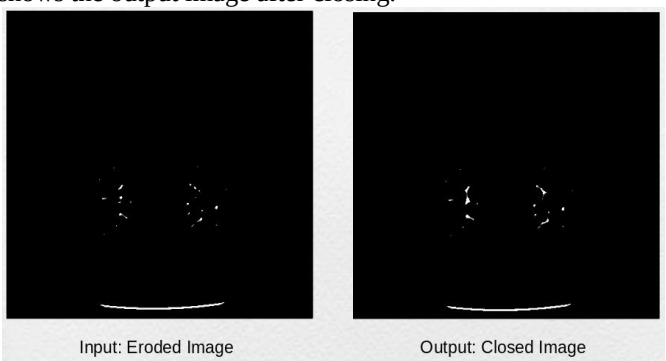

Fig 5 Morphological Closing

E. Morphology Opening

The opening of an image f by a structuring element s (denoted by) is an erosion followed by a dilation. This method is also known as filling because it can open up a gap between objects connected by a thin bridge of pixels. Any regions that are affected by erosion operation are restored to their original size by the dilation operation.

The result after opening is shown in Fig 6. Left hand side image shows input closed image and right hand side image shows the output image after opening.

IV. Lung Nodule Classification

In lung nodule detection [13] or ROI detection, the objects like lung nodules are detected in lung CT scan images. These objects can be seen by human eyes very easily.

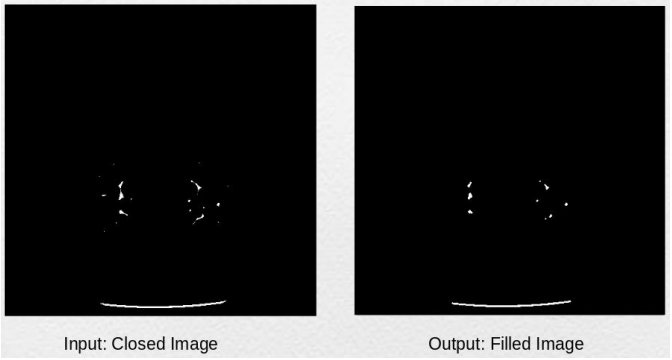

Fig 6 Morphological Opening

The objects are cropped out of image to have two region sets as nodules and non-nodules. Fig 7 shows the 6 nodules and 6 non-nodules of training dataset. These two sets are provided to SVM and CNN for training. The features are extracted from both the sets and then the input image is classified into noduleand non-nodule classes.

Providing image patches to the classifier instead of whole image reduces the complexity of classification. CNN extract the features of non-imaging region very correctly which can reduce accuracy that's why image patching is helpful for increasing efficiency. The working of SVM and CNN is discussed further.

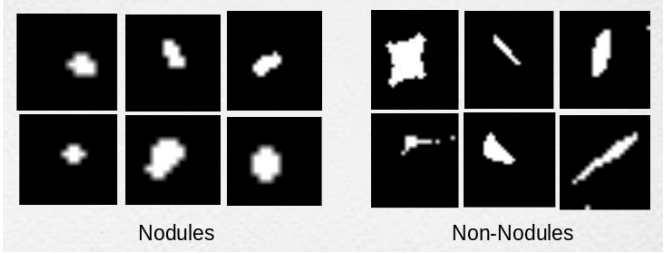

Fig 7 Nodules and Non-nodules

A. Support Vector Machine (SVM)

SVM [9] is a supervised model where data used for classification are first analysed by the features.

SVM is originally useful for a binary classification i.e., when two classes are present then each data element is get analysed and get categorized into one of the classes based on the features and similarity. After evolution, SVMs for classification of data into multiple classes are developed. But, when data to be classified in only two classes then linear SVM [10] is useful as it is binary classifier. In addition to perform linear classification, SVMs can efficiently perform a non-linear classification by implicitly mapping their inputs into high-dimensional feature spaces. SVM model [10] is a representation of mapping of the data elements to the respective classes as points in space, so that the data elements of the separate categories are divided by a clear gap that is as wide as possible.

The main aim of SVM is to find the optimal separating hyperplane which maximizes the margin of the training data. Hyperplane is nothing but a separation between two classes. In one dimention this separation is nothing but a point, in two dimention it is a line, in three dimention it is a plane and in more dimention it is a hyperplane.

Fig 8 shows example of SVM, where optimal hyperplane is shown between circle and square classes. The optimal hyperplane is the hyperplane which as far as possible from the data element of both the classes. The margin is the maximum distance between two classes. Calculating the margin is useful for finding optimal hyperplane.

As the equation of line, $y = ax + b$ is same as $y - ax - b = 0$
Given two vectors $w(-b - a\ 1)$ and $x(1\ x\ y)$:
$w^T x = -b \times (1) + (-a) \times x + 1 \times y$
ie., $w^T x = y - ax - b$

So the equation of an hyperplane is defined by :

$$w^T x = 0 \quad (1)$$

Given the two classes +1 and -1,
from equation (1),

$$w \cdot x^{+1} + b = +1 \quad (2)$$
$$w \cdot x^{-1} + b = -1 \quad (3)$$

Subtracting equation (2) from equation (1),

$$w \cdot (x^{+1} - x^{-1}) = 2 \quad (4)$$

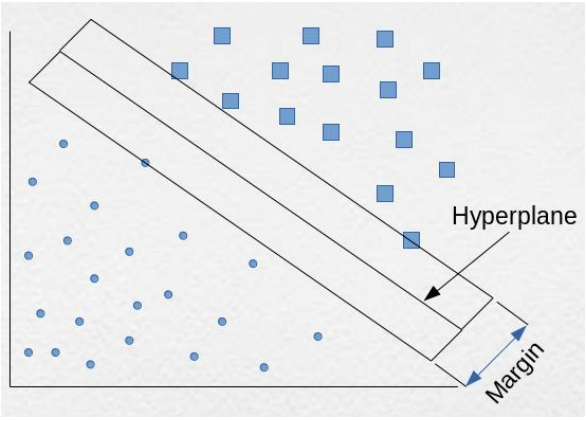

Fig 8 Optimal Hyperplane separating two classes

Margin width will have the equation as,
From equation (4),

$$M = \frac{(x^{+1} - x^{-1}).w}{\|w\|} = \frac{2}{\|w\|} \quad (5)$$

When ‖w‖=1, then m=2
When ‖w‖=2, then m=1
When ‖w‖=4, then m=1/2

B. Convolutional Neural Network (CNN)

In machine learning, a convolutional neural network (CNN, or ConvNet) [8] is a class of deep artificial neural networks that has successfully been applied to analyzing visual imagery. The main advantage of CNN is, it uses very less pre-processing of images as compared to other image classification algorithms, so the human efforts needed is less.

A CNN [11] consists of an input and an output layer, as well as multiple hidden layers as:

1. Convolutional layer
2. Pooling layer
3. ReLU layer
4. Fully connected layer
5. Loss layer

Typically, convolutional layers and pooling layers are get applied alternatively [15]. The convolution layer is the main building block of a convolutional neural network. Convolutional layer is responsible for feature extraction of images. Image patches of lung nodules are provided to convolutional layer.

Given a two-dimensional image, I, and a small matrix, K (convolutional kernel or filter) and slide it over the complete image and along the way take the sum of dot product between the kernel and pixels of the input image to get convolved image I * K. The example of convolution is shown in Fig 9.

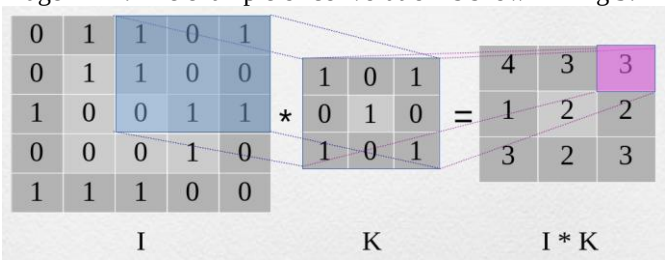

Fig 9 Convolutional Layer

A pooling layer is another important building block of CNN. As convolutional layer extract the features from images, pooling layer is useful for extracting important features from convolved features. There are two types of pooling as max pooling and average pooling as shown in Fig 10. In pooling, a window of 2*2 size is used to have a group of features. Pooling works on each group of features independently.

In max pooling, the maximum value among the features in a group is extracted. In average pooling, the average of each value in a group is calculated. Pooling is also used for reducing the computation space and it is also known as sub-sampling.

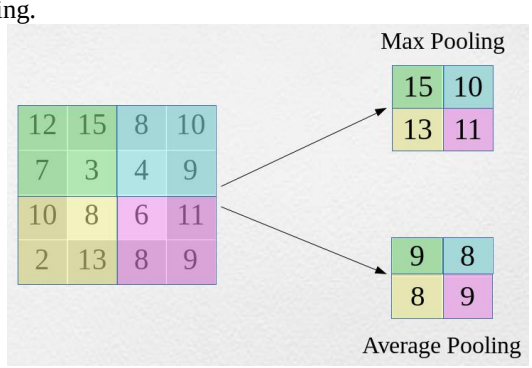

Fig 10 Pooling Layer

After extracting the features of image the input space is in the range of $(-\infty, +\infty)$ and activation layers squash the values into a range, typically [0,1] or [–1,1].Activation function layer is also known as ReLU layer as ReLU is the most common activation function used for CNN. ReLU is the abbreviation of REctified Linear Unit. The function is defined as, f(x) = max(x, 0). The output of this function is x when x>0 and it is 0 for x<=0. The major benefit of using ReLU is that it has a constant derivative value for all inputs greater than 0. The constant derivative value helps the network to train faster.

The last layer of CNN is fully connected layer and it is also known as dense layer. In fully connected layer each and every node is connected to each and every node from previous layer.

The features extracted in convolutional and pooling layer are used in fully connected layer for classification applying classification algorithms. Softmax function is commonly used function over svm for classification. This classifier gives the output of probabilities of image can be assigned for each classes. The highest probability class will be assigned to the respective input image.

After classification of an image the loss function is used for calculating the classification loss. It is also known as cost or error function. In this study the categorical cross entropy function has been used for calculating the loss of classification. The function is defined as given in equation (6),

$$H(p,q) = -\sum_{x} p(x)\log(q(x)) \quad (6)$$

This function gives the cross entropy between calculated approximate probability distribution and true probability distribution. Here p is the true distribution and q is the coding distribution. The loss calculated by this function is provided for back propagation of neural network. The weights assigned for convolution are then get changed according to the loss. This gets repeated till the weights do not change and we get the desired classification output.

V. Results

As discussed in section III and section IV, the image pre-processing techniques are applied on the patient lung ct scan images to identify the lung nodule regions, also the machine learning techniques such as SVM and CNN [12] are applied to the pre-processed image patches to classify the lung nodules and non-nodules in the ct scan. The ultimate goal of this experiment is to idenfy which technique is more efficient to classify the lung nodules to diagnose the lung cancer. The comparison parameter used in this study is classification accuracy of each classifier. It is defined as, the ratio of samples which are truly classified to the total no. of samples and multiplied by 100. It can be shown in equation as equation (7),

$$Accuracy = \frac{Truely\ classified\ samples}{Total\ no.\ of\ samples} * 100 \quad (7)$$

Table 1 shows the comparison of classification accuracy of Support Vector Machine (SVM) and Convolutional Neural Network (CNN).

TABLE I
COMPARISON OF CLASSIFICATION ACCURACY

Sr. No.	Technique	Accuracy
1	Support Vector Machine (SVM)	90%
2	Convolutional Neural Network (CNN)	91.66%

Classification of lung nodules and non nodules is more accurate using Convolutional Neural Network (CNN) and it is relatively less accurate using Support Vector Machine (SVM).

VI. Conclusion

The purpose of conducting this study is to analyze the effectiveness of machine learning techniques such as Support Vector Machine (SVM) and Convolutional Neural Network (CNN) for lung nodule detection and classification. By going through various research papers it is discovered that SVM and CNN [14] has played an very important role for lung cancer diagnosis. As lung nodule detection is very important step for lung cancer diagnosis and treatment the classifiers from machine learning are used for classifying lung nodules from other objects in the patient lung CT scan images. Preprocessing of images intend to better accuracy instead of providing whole image, as it reduce false positive rate. Providing preprocessed lung ct scan with annotations given by radiologists, feature extraction and classification of lung nodules with better efficiency can be performed. SVM has given the classification accuracy of 90% and CNN has given as 91.66%. By this study, it can be said that, CNN is proven to be the best method for feature extraction and classification of lung nodules in CAD systems.

References

[1] [1] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lorter Tieulent, and A. Jemal, “Global cancer statistics, 2012”, CA Cancer J Clin., vol. 65, no. 2, pp. 87–108, 2015.

[2] [2] I. Sluimer, A. Schilham, M. Prokop, and B. Ginneken, “Computer analysis of computed tomography scans of the lung: a survey”, IEEE Trans. Med. Imaging, vol. 25, no. 4, pp. 385–405, 2006.

[3] [3] Armato III, Samuel G., McLennan, Geoffrey, Bidaut, Luc, McNitt-Gray, Michael F., Meyer, Charles R., Reeves, Anthony P., … Clarke, Laurence P. (2015). Data From LIDC-IDRI. The Cancer Imaging Archive. http://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX.

[4] [4] Muzzamil Javaid, Moazzam Javid, Muhammad Zia Ur Rehman, Syed Irtiza Ali Shah, “A novel approach to CAD system for the detection of lung nodules in CT images”, Elsevier B.V., Computer methods and programs in biomedicine, 2016.

[5] [5] Yuan Sui, Ying Wei, Dazhe Zhao, “Computer-Aided Lung Nodule Recognition by SVM Classifier Based on Combination of Random Undersampling and SMOTE”, Pubmed, Computational and Mathematical Methods in Medicine, 2015.

[6] [6] Hongyang Jiang, He Ma, Wei Qian, Mengdi Gao and Yan Li, “An Automatic Detection System of Lung Nodule Based on Multi-Group Patch-Based Deep Learning Network ”, IEEE Journal of Biomedical and Health Informatics, 2017.

[7] [7] Shuo Wang, Mu Zhou, Zaiyi Liu, Zhenyu Liu, Dongsheng Gu, Yali Zang, Di Dong, Olivier Gevaert, Jie Tian, “Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation”, Elsevier, Medical Image Analysis, 2017.

[8] [8] Marios Anthimopoulos, Stergios Christodoulidis, Lukas Ebner, Andreas Christe, Stavroula Mougiakakou, "Lung Pattern Classification for Interstitial Lung Diseases Using a Deep Convolutional Neural Network", IEEE Transactions on Medical Imaging, 2016.

[9] [9] Jingjing Yuan, Xinglong Liu, Fei Hou, Hong Qin, Aimin Hao, "Hybrid-feature-guided lung nodule type classification on CT images", Elsevier, Computers & Graphics 2017.

[10] [10] Mohsen Keshani, Zohreh Azimifar, Farshad Tajeripour, Reza Boostani, "Lung nodule segmentation and recognition using SVM classifier and active contour modeling: A complete intelligent system", Elsevier, Computers in Biology and Medicine, Volume 43, Issue 4, 1 May 2013, Pages 287-300.

[11] [11] Kingsley Kuan, Mathieu Ravaut, Gaurav Manek, Huiling Chen, Jie Lin, Babar Nazir, Cen Chen, Tse Chiang Howe, Zeng Zeng, Vijay Chandrasekhar, "Deep Learning for Lung Cancer Detection: Tackling the Kaggle Data Science Bowl 2017 Challenge", arXiv:1705.09435v1 [cs.CV] 26 May 2017.

[12] [12] Diaz JM, Pinon RC, Solano G., "Lung cancer classification using genetic algorithm to optimize prediction models", IISA 2014, 5th Int. Conf. Information, Intell. Syst. Appl. Chania: IEEE; 2014: 1–6.

[13] [13] Choi WJ, Choi TS., "Automated pulmonary nodule detection based on three-dimensional shape-based feature descriptor", Computer Methods Programs Biomed. 2014;113:37–54.

[14] [14] Valente IRS, Cortez PC, Neto EC, Soares JM, de Albuquerque VHC, Tavares JMRS, "Automatic 3D pulmonary nodule detection in CT images: a survey", Computer Methods Programs Biomed. 2015;124:91–107.

[15] [15] Tianyi Liu, Shuangsang Fang, Yuehui Zhao, Peng Wang, Jun Zhang, "Implementation of Training Convolutional Neural Networks", arXiv:1506.01195.au, Ministry of Home Affairs, Government of India, 2014.