

# Malignant Nodule Detection: A new approach towards early prediction of Lung Cancer using X-ray data

**Authors: Sia Sehgal<sup>(a)</sup>; Reetu Jain<sup>(b)</sup>**

<sup>(a)</sup> Hill Spring International School, Balkrishna Nakashe Marg, Janata Nagar, Tardeo, Mumbai, Maharashtra 400034  
[sia13sehgal@gmail.com](mailto:sia13sehgal@gmail.com)

<sup>(b)</sup>Mentor, On My Own Technology Pvt Ltd, 1018, Samartha Aishwarya, Lokhandwala, Oshiwara, Andheri west, Mumbai – 400053, India.  
[reetu.jain@onmyowntechnology.com](mailto:reetu.jain@onmyowntechnology.com)

---

Early detection of lung cancer is essential for the survival of individuals, but at the same time, it poses a challenge for the medical community. Generally, chest radiography and computed tomography scans are used during initial screening for diagnosis of malignant nodules. This is because during early stages, benign and malignant nodules show close resemblance to each other. This paper focuses on deep learning-based convolutional neural network model proposed for early identification of malignant nodules. Only 16% of lung cancer cases are diagnosed at an early stage. More than half the people with lung cancer die within one year of being diagnosed, since most are diagnosed in stage 3 or 4. CNNs are primarily used to solve difficult image-driven pattern recognition in images. This study considers the use of x-ray images, one of the difficult data studies for Neural Networks. Thus, a dynamic study using drop out layer and ReLU (Rectified Linear) activation function is used in the model. Our research showed effective results in predicting the type of nodule with the accuracy score of approximately 68 %. This is a noninvasive solution detecting lung cancer based on the size, pattern and pixel values of the nodule.

**Keywords:** Malignant nodules, Benign nodules, Convolutional Neural Network (CNN)

---

## 1. Introduction

Lung nodules are commonly found clumps of cells located in the lungs. Most lung nodules are scar tissues from past lung infections. They do not cause negative effects under usual circumstances, and a majority of them are not signs of lung cancer. If the CT scan shows small nodules (less than a centimeter wide, or about the size of a green pea) also called as Benign nodules, the probability of them being cancerous is low. Larger nodules are more worrisome. Rounded nodules are less likely to be cancerous than spiculated (having jagged edges) ones also called Malignant nodules. Lung nodules show up on imaging scans like X-rays or CT scans. Below image shows x-ray images of chest having Lung Nodule.

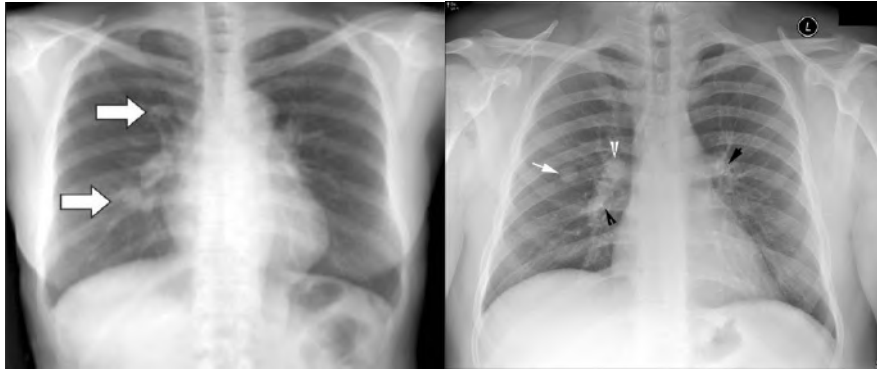


Fig 1.: X-ray images of lung nodules

Benign and malignant nodules have considerable feature overlaps, but still are differentiated on the basis of morphology and location at early stages [1]. Benign nodules are usually located at the peripheral, with smooth surface and triangular shapes filled with fat and calcium, while malignant nodules often show speculation with edges, lobulated, vascular convergence, cystic air spaces, pleural indentations and sub-solid morphology [2]. Early detection of lung cancer is challenging, due to the various shapes, sizes, densities and locations of nodules. Studies have adopted machine learning Approaches such as segmentation, clustering, Artificial Neural Network and Support Vector Machine (SVM) to solve this problem [3,4,5,6,7,8]. However, relatively few approaches require to be incorporated while using Convolution Network (CNN) to classify nodules in images. CNN is inspired by the biological functions of neurons.

A typical CNN framework consists of several convolutional layers, regularization layers, subsampling and fully connected layers. CNN has demonstrated to achieve better results in any image processing tasks, including segmentation, object detection, etc. Compared to previous approaches that involve complicated image pre-processing and designing optimal vision features, CNN is able to learn to identify visual features automatically and could potentially provide better classification accuracy in nodule (such as between nodule and non-nodule region of interests with high sensitivity and low false positive rate).

This study explores how a CAD system can be created for developing lung cancer diagnosis, which has the potential of reducing the time for the radiologists to create large-scale, labeled datasets of lung x-ray images. Section 1 of this paper introduces the challenge at hand, and explains the application of CNN, while Section 2 summarizes the recent literature in this research domain. Section 3 describes the methodology adopted for solving the said problem and presents the results, and Section 4 concludes the study.

## Literature Review

We made use of a convolutional neural network where they extracted features from a CNN pre-trained on the ImageNet dataset and applied a simple linear SVM classifier to classify a perifissural nodule (a benign nodule in the lung with a high false positive rate in lung cancer detection) [9]. They used the ensemble of classifiers for 3D slices of an object to obtain the final classifier and compared its performance with a bag of frequency descriptors with similar accuracy of 86.8%. Hua explored a 3-layer CNN and deep belief network to approach the problem of nodule classification in computed tomography images [10]. et al.. considered three deep learning algorithms: CNN, Deep Belief Network (DBNs) and Stacked Denoising Auto coder (SDAE) to classify nodules They showed that the performance on down sampled data of the three methods are similar and are around 80% [11]. Shen et al. proposed a learning framework using Multi Scale CNN (MSCNN). The method captures nodule heterogeneity extracting features [12].

Many combinations of features [13, 14, 15, 16.] proposed an artificial neural network approach using statistical parameters like mean, standard deviation skewness, kurtosis, fifth central moment and sixth central moment to

achieve the classification accuracy of 91.1% [17]. Abdulla et al. used area, perimeter, and shape as features to train an artificial neural network for classification with an accuracy of 90% [18]. Ada and Kaur carried out a computational procedure that was also effective. Using similarity, groups of images were created. They first pre-processed the data with morphological and histogram normalization, then they selected features, used Principle Component Analysis (PCA), and trained a Neural Network Classifier with a single hidden layer [19].

The categorization of nodules in lung CT scans has been thoroughly investigated using image processing techniques. For better nodule recognition, many researchers used segmentation, morphological operations, and contour filter techniques [20, 22, 21, 23, 22]. In order to segment CT scans and classify focal areas in the lung region, Nathaney and Kalyani used the ROI processing and adaptive threshold method to identify lung cancer with the lowest possible rate of false negatives [24] Another image processing technique, called LCDS, was created by Chaudhary and Singh. In this method, pre-processing steps include smoothing, Gabor filter enhancement, and Watershed segmentation. Then, morphology and colorimetric techniques were used to determine features like area, perimeter, and roundness. To establish the stage of lung cancer for each patient, they evaluated the statistics of various parameters [25]. In order to identify cancer nodules, Sharma and Jindal used segmentation algorithms after applying fundamental image processing techniques, such as erosion, median filtering, dilation, outlining, and lung border extraction [26].

## Materials and Methods

The aim of the present study is to develop an intelligent system to detect and differentiate between benign nodules and malignant nodules. In this regard, an intelligent system can be developed that integrates image processing with DL techniques to differentiate the two types of nodules. The DL technique that is used in this paper is the Convolutional Neural Network (CNN). The CNN model classifies the data into two classes: Malignant and benign. The model considers the pixel value of the x-ray images and maps it to the two classes. The image dataset needed to train and test the CNN model is x-ray images from the lungs.

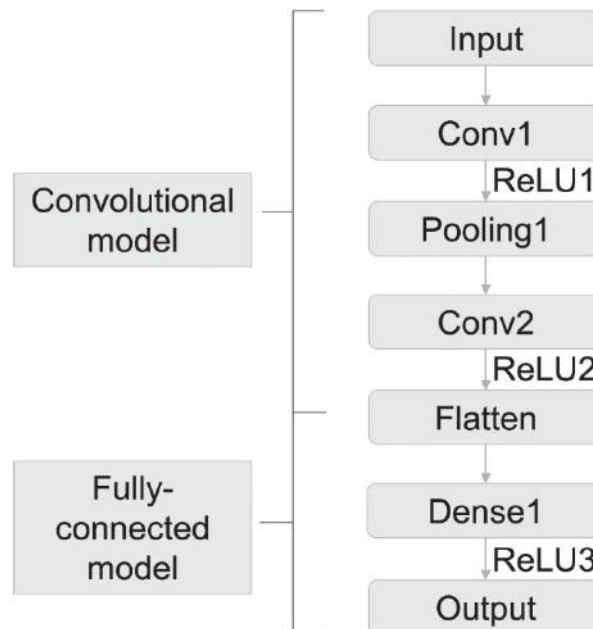


Fig 2.: CNN Pipeline

### 3.1 Dataset

This dataset consists of several thousand examples formatted in multipage TIFF (for use with tools like ImageJ and KNIME) and HDF5 (for Python and R). The data was preprocessed and extracted partially from the LUNA16 competition [27] and should be used with the same policy that data has. The dataset is more for practice with medical images and CNNs but it would be interesting to know how the best manually created features (Hog, SIFT, ...) perform against various Deep Learning approaches. It would also be valuable to visualize exactly which parts of an image made the algorithm guess malignant or benign. Figure 3 shows the images of the two classes in the dataset.

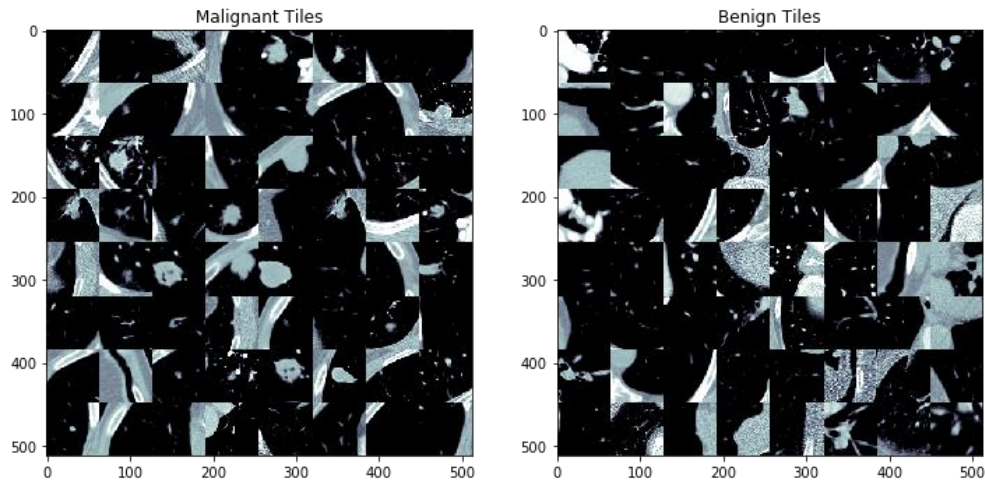


Fig 3.: Dataset

It is difficult to spot the difference between the two classes with the naked eye, and therefore feature engineering was conducted to determine obvious differences using EDA as shown in figure 4.

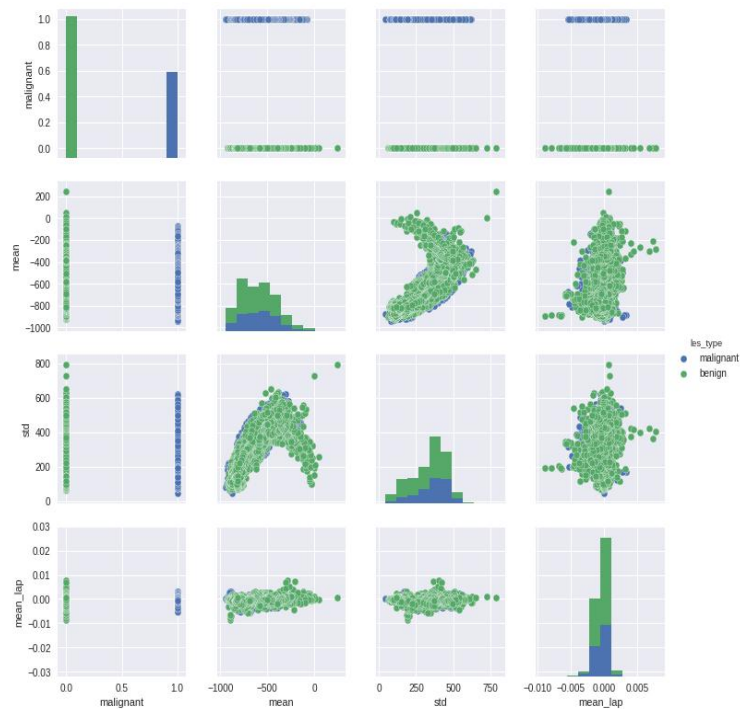


Fig 4.: EDA on dataset

### 3.2 Preliminaries and methodology

This section of the paper briefly describes the preliminary concept required to analyze the images and subsequent development of the DL model. The concept used to create the DL model and to analyze the images and classify them into two classes (Malignant and Benign) is Convolutional Neural Networks

### 3.3 Deep Learning (DL)

Multi-layer processing computer models may recognise patterns in data using many levels of abstraction thanks to deep learning (DL) [28]. The DL techniques have significantly increased the efficiency of AI and are widely used in the fields of facial recognition, object recognition, speech recognition, and many other complex domains. The backpropagation technique is used by the DL algorithm to identify patterns among the various dataset parameters. Image, video, speech, and audio processing have advanced thanks to CNN, a DL model [29].

The CNN model is used in the study to identify and classify the Lung nodule into two classes namely Malignant and Benign. The Characteristics are classified as pixel values and based on the pixel value it will give the prediction to which class the images of lung nodule belong to.

Training and testing are the first and second segments of CNN. The model is established for all significant weight values and the bias from labelled examples throughout the training phase. The model created during training is tested for a sample dataset in the second section. The accuracy and precision of the developed model are evaluated by comparing the predicted label with the actual label. The following are the many operations needed for CNN prediction:

#### 3.3.1 Convolution operation

The CNN convolution operation is used to separate the relevant noise from the features that are extracted from the training images. To make it easier to extract an image's features and reduce unnecessary noise, the convolution procedure splits an image into smaller pieces. The phrase "image matrix" refers to this shattered image ( $Im$ ). The layer holding the matrix of the  $N$  filter ( $Fi$ ) is slid over the width and height of the image matrix. Image and filter matrix multiplication yields the final matrix ( $Re$ )

Mathematically,

$$Im \times Fi = Re \quad (1)$$

#### 3.3.2 Activation operation

The activation operation of a node in a neural network determines that node's output for a certain input or collection of inputs. In this study, the activation operator is a rectified linear unit (ReLU). The ReLU is written mathematically as:

$$y = 0, \text{ if } x < 0 \quad (2a)$$

$$y = \quad (2a)$$

$$x, \text{ if } x \geq 0 \quad (2a)$$

$$\geq \quad (2a)$$

$$y = x, \text{ if } x \geq 0 \quad (2b)$$

ReLU enables more rapid and efficient training of deep neural architectures on complicated datasets as compared to sigmoid function or similar activation functions. [30].

### 3.3.3 Pooling operation

The pooling method decreases the number of learning parameters, which in turn reduces the required computational work. This method is beneficial for condensing the features that are present in a specific area of the feature map that the convolution layer produced. The majority of CNN's pooling operations use average and maximum pooling. The average value for patches on a feature map is used to construct a down sample in the average pooling technique. The main drawback of average pooling is that it does not provide an accurate result if there are multiple outliers. Maximum pooling, which uses a feature map's maximum value for patches, can be utilized to get around this. [31].

### 3.3.4 Layer stacking

Convolution, activation, and pooling operations are repeatedly used in layer stacking until the output is a reduced matrix of the input picture.

### 3.3.5 Fully connected layer

A CNN model ends with this layer. Neurons from the preceding layers are completely linked to those in this layer. Because of this, this layer is referred to as a fully connected (FC) layer. It is in charge of categorizing the given category and forecasting the output and label for that class.

### 3.3.6 Classification and Prediction

Classification is categorization, and each FC layer neuron is labelled on the map. The input class whose label has the greatest amount of characteristics comparable to those in the test images is predicted by the FC layer. The SOFTMAX activation function is utilized in this investigation to categorize the label. The multinomial probability distribution is predicted using the SOFTMAX activation function. [32]. The SOFTMAX activation function's mathematical expression is:

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} \quad (3)$$

### 3.3.7 Results from the CNN model

The performance and validation of the CNN model are briefly reviewed in this section of the text.

### 3.3.8 Performance of the CNN model

The Kera's and TensorFlow package, which is written in Python 3.10 and operates on a computer, is used to create the CNN model. Using Adam optimizer, the CNN model's parameters are improved. The advantageous characteristics of the 'Gradient Descent with Momentum' are carried over to the Adam optimizer. and 'Root Mean Square Propagation' algorithms. The proposed CNN model is propagated for 5000 epochs with 9 steps each and 1 step for validation. The best model is chosen for the epoch with the lowest loss value. The CNN model's training accuracy and training loss graph for identifying and classifying the nodule type is shown in fig 4.

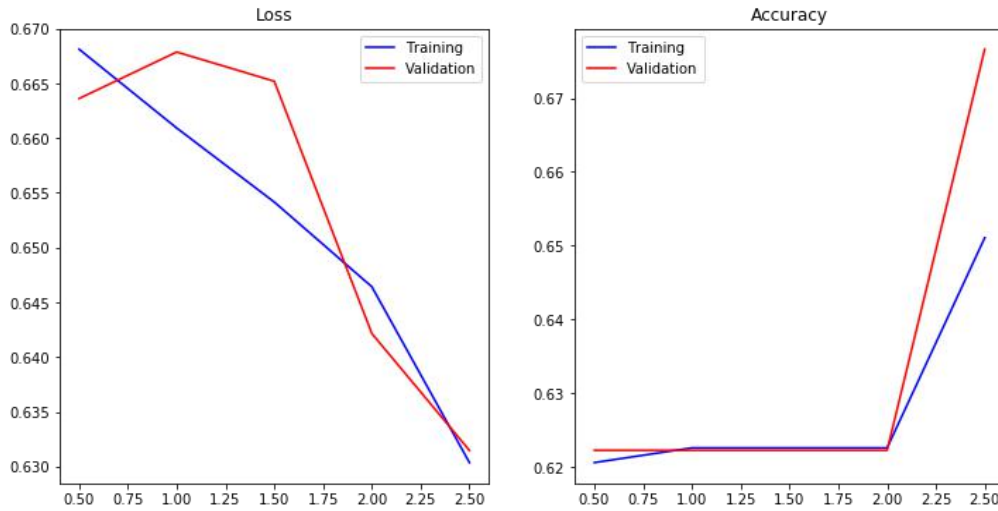


Fig 4.: accuracy of loss of the model

The CNN model's first epoch's training accuracy and loss are 0.3036 and 0.9106 respectively. However, the 50th epoch's training accuracy and loss values are 0.6766 and 0.6389 respectively.

### 3.3.9 Validation of the CNN mode

The CNN model is validated with a picture of a lung nodule which was not used to train and evaluate the system. The ROC graph is shown in fig. 5.

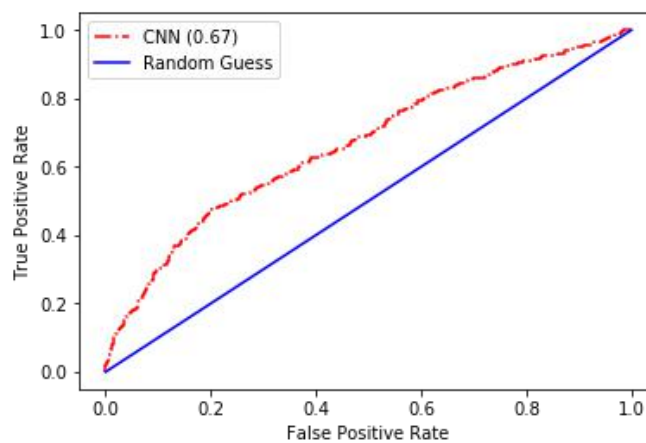


Fig 5.: ROC Curve

#### 4. Conclusion

In this study evaluated Deep Learning based on CNN and carried out a reader survey. Evidently, CNN has the potential to assist radiologists in their daily diagnosis. In contrast to previous methods, CNN's judgments may be followed with individual lesion ratings and box predictions. There is still room for improvement and accuracy; it must be determined in subsequent work whether training with a larger dataset is sufficient or whether a second CNN is required to recognize aberrant cases and respond appropriately. Due to developments in healthcare digitization, data might soon be accessible in machine-readable format. Such data might be used to alter CNN's conclusion (e.g. to invalidate lesion scores in the region of a known pacemaker). The intension is to gather and annotate more data in order to improve categorization performance.

#### References:

1. American Thoracic Society, PATIENT EDUCATION | INFORMATION SERIES, what is a Lung Nodule, <https://www.thoracic.org/>
2. 12. Spinhoven M.J., Desbuquoit D., Snoeckx A., van Meerbeeck J.P., Reyntiens P., Parizel P.M., Van Schil P.E. Evaluation of solitary pulmonary nodi: Size matters, but do not ignore the power of morphology. *Insights Imaging*. 2017; 9:73–86. [PMC free article] [PubMed] [Google Scholar]
3. Francesco Ciompi et al. “Automatic classification of pulmonary peri-Fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-box” In: *Medical image analysis* 26.1 (2015), pp.195-202
4. Sheeraz Akram et al. “Artificial neural network- based classification of lung nodules using hybrid features from computerized tomographic images  
In: *Applied Mathematics & Information Sciences* 9.1 (2015), p. 183.  
Anindya Gupta et al “A tool for lung nodule analysis based on segmentation and morphological operation”. In: *th International Symposium on . IEEE*. 2015, pp.
5. Lin Lu et al. “Hybrid detection of lung nodules on CT scan images”. In: *Medical physics* 42.9 (2015), pp. 5042–5054.
6. Temesguen Messay, Russell C hardie, and Timothy R Tuinstra “Segmentation of pulmonary nodule in computed tomography using a regression neural network approach and its application of the lung image database consortium and image database resource initiative dataset”. In: *Medicial Image analysis* 22 (2015), pp. 48– 62.
7. Atsushi Teramoto rt al. “Automated detection of lung tumors in PET/CT images using active contour filters”. In: *Medical Imaging 2015: Computer-Aided Diagnosis*. Vol. 9414. International Society for Optics and Photonics 2015, p. 94142V.
8. Francesco Ciompi et al. “Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box”. In: *Medical image analysis* 26.1 (2015), pp. 195–202.
9. Kai-Lung Hua etal. “Computer-aided classification of lung nodules on computed tomography images via deep learning technique”. In: *OncoTargets and therapy* 8 (2015)
10. Wenqing Sun, Bin Zheng and Wei Qian. “Computer aided lung Cancer daignosis with deep learning Algorithms”. In: *Medical Imaging 2016: Computer-Aided Diagnosis*. Vol. 9785. International Society for Optics and Photonics. 2016, 97850Z.
11. Wei Shen et al. “Multi-scale convolutional neural networks for lung nodule classification”. In: *International Conference on Information Processing in medical Imaging*. Springer. 2015, pp. 588–599.
12. Sheeraz Akram et al. “Artificial neural network based classification of lung nodules using hybrid features from computerized tomographic images”. In: *Applied Mathematics & Information Sciences* 9.1 (2015), p. 183.
13. Lin Lu et al. “Hybrid detection of lung nodules on CT scan images”. In: *Medical physics* 42.9 (2015), pp. 5042–5054.
14. Weisheng Wang et al. “Data analysis of lung imaging database consortium

15. Constrium and image database resource initiative". In: Academic radiology 22.4 (2015), pp. 488–495.
16. Koichiro Yasaka et al. "High-resolution CT with new model-based iterative reconstruction with resolution preference algorithm in evaluations of lung nodules  
Comparison with conventional model-based iterative reconstruction and adaptive statistical iterative reconstruction". In: European journal of radiology 85.3 (2016), pp. 599– 606.
17. Jinsa kuruvilla and K Gunavathi. "Lung cancer classification using neural network For CT images
18. ". In: Computer method and programming in biomedicine. (2014), pp. 202–209.
19. Azian Azamimi Abdullah and Syamimi Mardiah Shaharum, "Lung cancer cell classification method using artificial neural network". In: information engineering letters 2.1 (2012).
17. Rajneet Kaur Adal Early detection and prediction of lung cancer survival using neural network classifier 2013
20. Anindya Gupta et al. "A tool for lung nodule analysis based on segmentation and morphological operation". In: Intelligent Signal Processing (WISP), 2015 IEEE 9th International Symposium on. IEEE. 2015, pp. 1–5.
21. Temesguen Messay, Russell C Hardie, and Timothy 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". In: Medical image analysis 22.1 (2015), pp. 48– 62.
22. Atsushi Teramoto et al. "Automated detection of lung tumors in PET/CT images using active contour filters". In: Medical Imaging 2015: Computer-Aided Diagnosis. Vol. 9414. International Society for Optics and Photonics. 2015, p. 94142V.
23. Sudipta Mukhopadhyay. "lung CT images". In: journal of digital imaging 29.1 (2016), pp. 86–103.
23. Ms Gangotri Nathaney and Kanak Kalyani. "Lung Cancer Detection Systemon Thoracic CT Images Based on ROI Processing  
". In: Lung Cancer 4.4 (2015), pp. 173– 176.
25. Anita Chaudhary and Sonit Sukhraj Singh. "Lung cancer detection on CT image using image processing". IEEE. 2012, pp. 142–146.
26. Disha Sharma Gagandeep Jindal. "Identifying lung cancer using image processing techniques' International Conference on Computational Technique and Artificial Intelligence (ICCTAI). 2011, pp. 115–120
27. LUNA16 competition, lung nodule analysis 2016(<https://luna16.grand-challenge.org/description/>)
28. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning.
29. Wlodarczak, P., Soar, J., & Ally, M. (2015, October). Multimedia data mining using deep learning. In 2015 Fifth International Conference on digital information Processing and Communications (ICD IPC) (pp. 190-196). IEEE.
30. Behnke, S. (2003). Hierarchical neural network for image interpretation (vol. 2766) Springer.  
D.Song, Z., Liu, Y., Song, R., Chen, Z., Yang, J., Zhang, C., & Jiang, Q. (2018) A sparsity-based stochastic pooling mechanism for deep convolutional neural network, neural network 105, 340-345
31. E.Gao, B., & Pavel, L. (2017) On the properties of the SoftMax function with the application in game theory and reinforcement learning. arXiv preprint arXiv:1704.00805.