JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

# LUNG CANCER NODULE DETECTION USING IMAGE PROCESSING

**[1]Anushka Kushwaha, [2]Ashna Khan, [3] Dr. Nidhi Saxena**

[1]Undergraduate, [2] Undergraduate, [3]Assistant Professor
[1,2,3]CSE Department,
[1,2,3]SRMCEM, Lucknow, India
[1]anushkamsd7@gmail.com, [2] ashna.khan77@gmail.com, [3]nidhi.shivansh@gmail.com

*Abstract:* Lung cancer is one amongst the foremost deadly diseases within the world. The recent estimates provided by the World Health Organization (WHO) says that around 7.6 million deaths worldwide each year are due to carcinoma. Moreover, humanity because of cancer is imagined to continue rising, to become around 17 million worldwide in 2030. Discovering carcinoma within the early stage is the only method for its cure. Different methods are available for diagnosis of carcinoma, namely, MRI, isotope, X-ray and CT. CT scan image don't seem to be easy to know, but using CNN with Image Segmentation is a straightforward approach to detect carcinoma. Convolutional neural network (CNN) is one of of the deep structured algorithms widely applied to investigate the flexibility to visualize and extract the hidden texture features of image datasets.

This study aims to automatically extract the self-learned features using an end-to-end learning CNN and compares the results with the traditional state-of-art and traditional computer-aided diagnosis system's performance.

***Keywords— Image processing; Neural Networks; Feature Extraction; Segmentation; Preprocessing.***

## I. INTRODUCTION

Lung cancer is one among the deadliest cancers worldwide. However, the first detection of carcinoma significantly improves survival rate. Cancerous (malignant) and noncancerous (benign) pulmonary nodules are the little growths of cells inside the lung. Detection of malignant lung nodules at an early stage is important for the crucial prognosis [1]. Early-stage cancerous lung nodules are considerably just like noncancerous nodules and want a medical diagnosis on the premise of slight morphological changes, locations, and clinical biomarkers [2] . The challenging task is to measure the probability of malignancy for the first cancerous lung nodules [3]. Various diagnostic procedures are employed by physicians, in connection, for the first diagnosis of malignant lung nodules, like clinical settings, computed tomography (CT) scan analysis (morphological assessment), PET (metabolic assessments), and needle prick biopsy analysis [4]. However, mostly invasive methods like biopsies or surgeries are employed by healthcare practitioners to differentiate between benign and malignant lung nodules. For such a fragile and sensitive organ, invasive methods involve many risks and increase patients' anxieties. The most accurate method used for the investigation of lung diseases is computed tomography (CT) imaging [5]. However, CT scan investigation features a high rate of false positive findings, with carcinogenic effects of radiations. Low-dose CT technique uses significantly lower power radiation contact than standard-dose CT. The results show that there's no significant difference in detection sensitivities between low-dose and standard-dose CT images. However, cancer-related deaths were significantly reduced within the selected population that were exposed to low-dose CT scans as compared to chest radiographs, which is depicted within the National Lung Screening Trial (NLST) database [6]. The detection sensitivity of lung nodules grows with complexed anatomical details (thinner slices) and better image registration techniques. However, this increases the datasets to a really large extent. Depending on the thickness of slice, up to 500 sections/slices are produced in one scan [7]. An experienced radiologist takes approximately 2–3.5 min to observe one slice [8]. The workload of a radiologist increases greatly to screen a CT scan for the possible existence of a nodule. additionally, to section thickness of the CT slices, detection sensitivity also depends on nodule features like size, location, shape, adjacent structures, edges, and density. Results show that only 68% of the time carcinoma nodules are correctly diagnosed when just one radiologist examines the scan and are accurately detected up to 82% of the time with two radiologists. The detection of cancerous lung nodules at an early stage may be a very difficult, tedious, and time-consuming task for radiologists. Screening plenty of scans with care requires lots of time by the radiologist, meanwhile it's significantly error-prone within the detection of small nodules [9]. After successful making masks next step is to create a model on VGG16 NET transfer learning model for better accuracy. eventually both the trained and optimized model are used and combined to create a whole final model for the classification of carcinoma. Referring to the paper "Deep learning for lung Cancer" by A. Asuntha & Andy Srinivasan. For the input layer, lung nodule CT scans are used and collected for various steps of the project. For the dataset we have the LUNA16 dataset. The dataset is a subset of LIDC-IDRI dataset, within which the heterogeneous scans are filtered by different criteria. Since pulmonary nodules are often very small, a skinny slice should be chosen. Therefore, scans having a slice thickness greater than 2.5 mm were discarded. Furthermore, scans with inconsistent slice spacing or missing slices were also discarded. This led to 888 CT scans, with a complete of 36,378 annotations by radiologists. during this dataset, only the annotations categorized as nodules $\geq 3$ mm are considered relevant, because the other annotations (nodules $\leq 3$ mm and non-nodules) aren't considered relevant for carcinoma screening protocols. Nodules detected by different readers that were closer than the sum of their radii were merged. during this case, positions and diameters of those merged annotations were averaged. This ends up in a group of 2290, 1602, 1186 and 777 nodules annotated by a minimum of 1, 2, 3 or 4 radiologists, respectively.

## II. DATASET COLLECTION AND VISUALIZATION

Visualization of dataset is a crucial a part of training; it gives better understanding of dataset. But CT scan images are hard to visualize for a standard pc or any window browser. Therefore, this module uses the pydicom library to unravel this problem. The Pydicom library gives a picture array and metadata information stored in CT images like patient's name, patient's id, patient's birth date, image position, image number , doctor's name , doctor's birth date etc. Luna16 dataset may be a directory which contains many subdirectories named on patient's ids. a whole subdirectory is 3d image of lungs which is stored in around 180- 2D image slices in keeping with their image number.

## III. IMAGE PROCESSING

For preprocessing of the image with image processing methods for uniformity and noise reduction, our first approach will be to make segmentation of CT scan images after preprocessing. We are using a watershed algorithm to make segmentation. Watershed algorithms make a mask for cancer cells in lung images. refer to article ["Serge Beucher and Fernand Meyer. The morphological approach to segmentation: the watershed transformation. In Mathematical Morphology in Image Processing (Ed. E. R. Dougherty), pages 433-481 (1993)."]

Given the sets of CT scans, this layer will construct the 3D lung scan, and extract only the lung region. This is especially important that edge detection is done with high accuracy to minimize the error rate of our model. We also utilize segmentation and machine learning techniques to pre-process the image, including auto-detecting boundaries that surround the volume of interest. The images represent the 2D slices of the patient's thoracic region in DICOM format.

### 3.1 Watershed Algorithm

The watershed could be a classical algorithm used for segmentation, that is, for separating different objects in a picture. A watershed could be a transformation defined on a grayscale image. It refers metaphorically to a geographically watershed, or drainage divide, which separates adjacent drainage basins. The watershed transformation treats the image it operates upon sort of a topographic map, with the brightness of every point representing its height, and finds the lines that lie the tops of ridges." The topological watershed" was introduced by M. Couprie and G. Bertrand in 1997 ranging from user-defined markers, the watershed algorithm treats pixels values as a local topography (elevation). The algorithm flood basins from the markers until basins designated to different markers connect on watershed lines. In many cases, markers are selected as local minima of the image, from where basins are flooded. First , we extract internal and external markers from CT scan images with the assistance of binary dilations and add them with a whole dark image using watershed methods. And it removes external noise from the image and provides a watershed marker of lungs and cancer cells. As we are able to see within the below figure watershed marker removes external noise and applies a binary mask on the image , black pixels in lungs represent cancer cells.

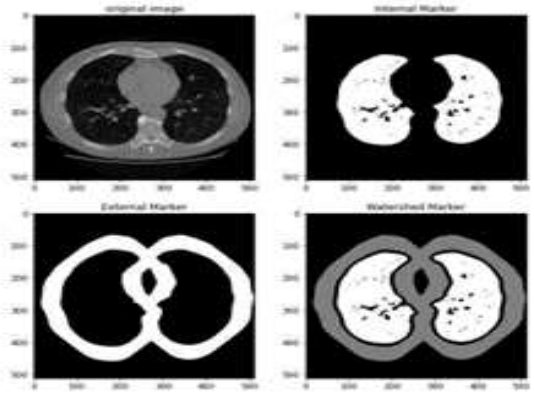


Figure 1 : Watershed marker

## IV. PROPOSED MODEL

The model is a CNN based upon lung segmentation on CT scan images. After preprocessing, the second step is making lung segmentation using a watershed algorithm. Watershed algorithm emphasizes the lung part and makes binary masks for lungs with semantic segmentation approach. Having 16 layers, VGG16 is an extremely effective image processing algorithm. Utilising a pre-trained model to use the most up to date technologies in machine learning and data mining. By training the CNN on the dataset, we are able to auto detect features in each DICOM image.

### 4.1 TRANSFER LEARNING: VGG16-NET

VGG Net is the name of a pre-trained convolutional neural network (CNN) invented by Simonyan and Zisserman from Visual Geometry Group (VGG) at University of Oxford in 201and it was the 1st runner-up of the ILSVRC (ImageNet Large Scale Visual Recognition Competition) 2014 in the classification task. The VGG Net has been trained using the ImageNet ILSVRC dataset which contains images of 1000 classes divided into three sets each of 1.3 million training images, 100,000 testing images and 50,000 validation images. The model obtained 92.7% test accuracy in ImageNet. VGG Net has attained success in many real-world applications such as estimating the heart rate based on the body motion, and pavement distress detection.

VGG Net has learned to extract the features (feature extractor) that can diffrentiate the objects and is applied to classify unseen objects. VGG was invented for the purpose of enhancing classification accuracy by increasing the depth of the CNNs. VGG 16 and VGG 19, having 16 and 19 weight layers, used for object recognition. VGG Net takes input 224×224 RGB images and passes them across a stack of convolutional layers with the fixed filter size of 3×3 and the stride of 1. There are five max pooling filters embedded among convolutional layers to down-sample the input representation (image, hidden-layer output matrix, etc.). The stack of convolutional layers is followed by 3 fully connected layers, having 4096, 4096 and 1000 channels, respectively. The last layer is a soft-max layer.

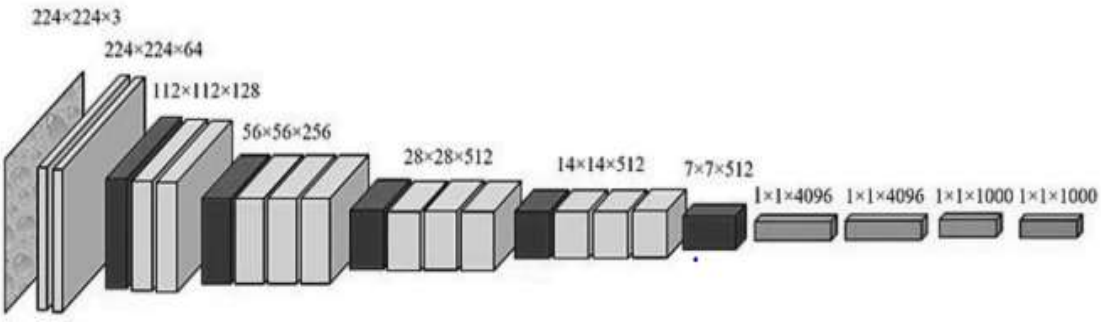



Figure 2 : VGG16-net Working Architecture.

## V. CLASSIFICATION AND VALIDATION

For the end layer we are using a single node for binary classification as we want to classify between cancer and non- cancer lungs. The final column in the data frame will be a binary value: 1 if the patient has cancer, and 0 if the patient does not have cancer. 75% of training data is given to train the model and 25% is used for testing purposes.

## VI. RESULT AND ANALYSIS

By using a hybrid of approaches in image processing and classification, we are able to develop an end-to-end process that detects lung cancer nodules with high accuracy. Further, by placing a heavy emphasis on automation of image processing as well as a reduction of false positives, we were able to develop a full model that runs with 70% accuracy on test data. The current research is extremely helpful to many students who would like to implement the findings into their projects and as well as to do further research. This project's findings also help in improving the detection of cancer at an early stage and to reduce the detection error.

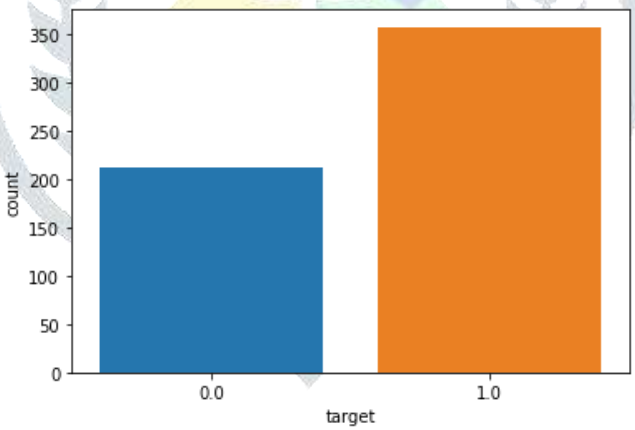



Figure 3: Count plot of target class

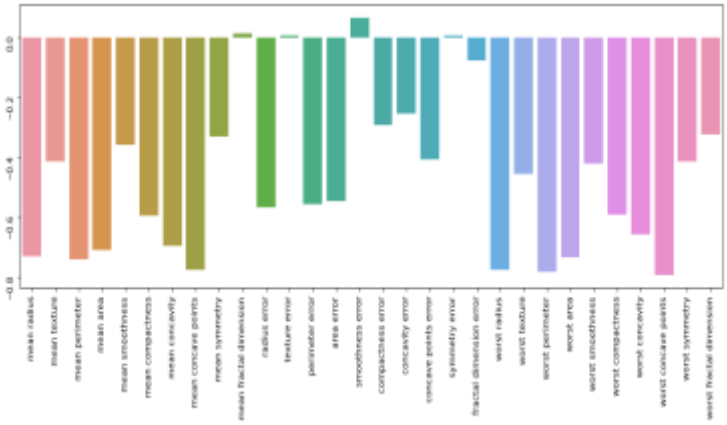


Figure 4: Correlation bar plot

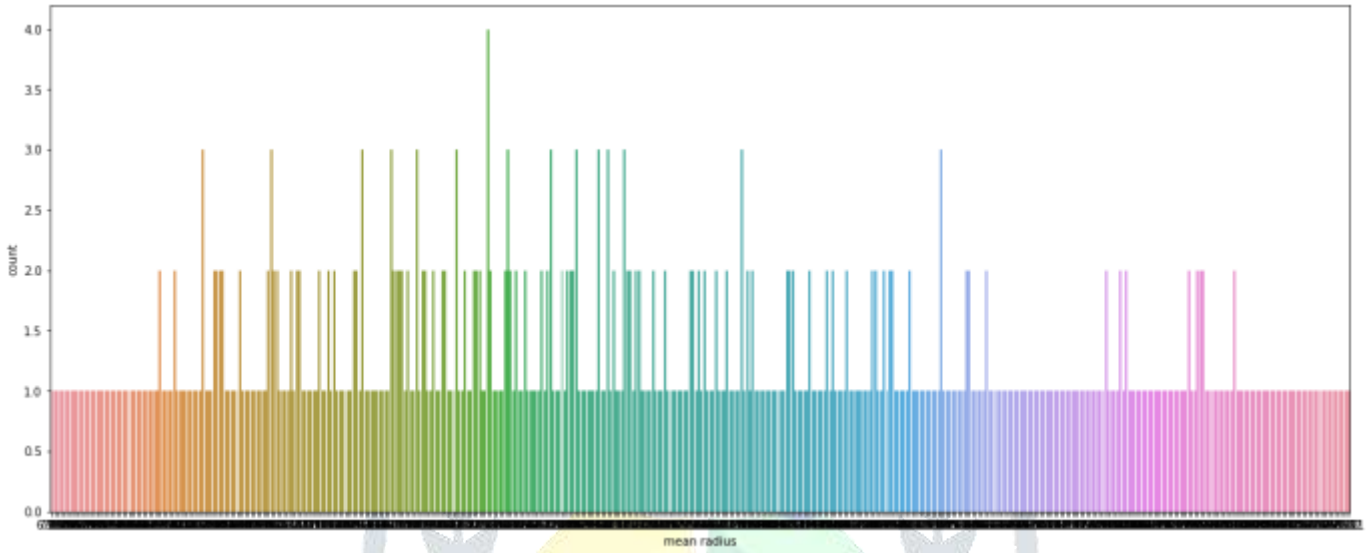
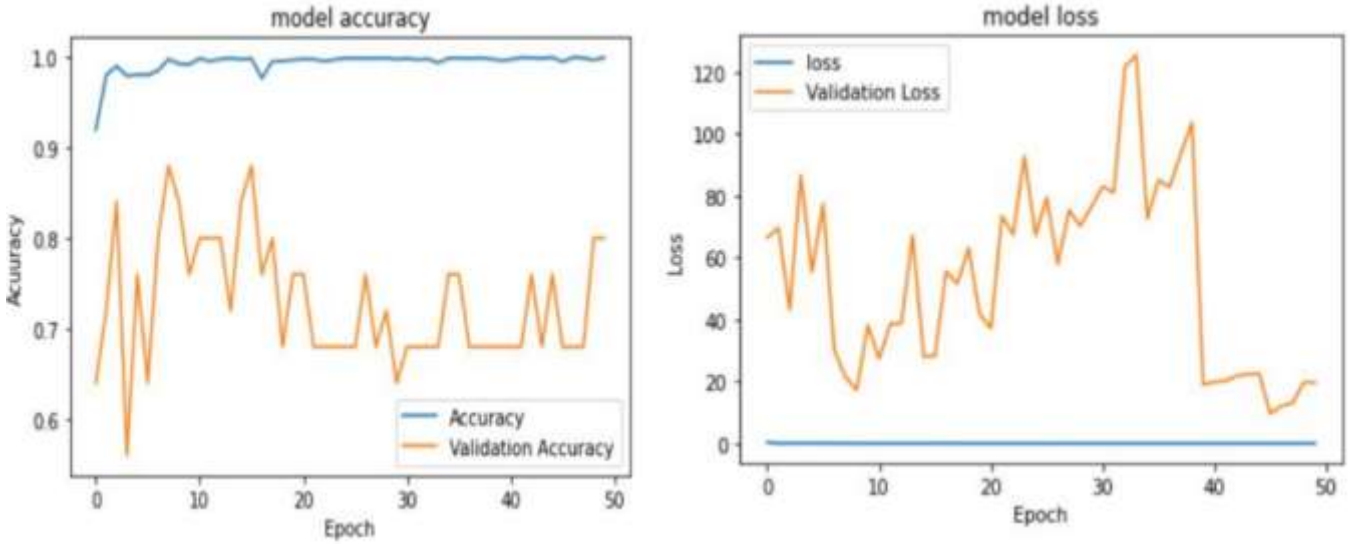



Figure 4: counter plot of feature mean radius




## VII. CONCUSION AND FUTURE SCOPE

- The current research is extremely helpful to many students who would like to implement the findings into their projects and as well as to do further research.
- This project's findings also help in improving the detection of cancer at an early stage and to reduce the detection error.

- For feature extraction, we have use VGG16 the simplicity of the model makes it easy to implement. However, there are other pre-trained CNN models available too for feature extraction. Other pre-trained models could be used like Resnet, Googlenet, etc.
- At each stage, one could use a novel approach to retain the maximum features, which would overall lead to a better model.

## VIII. ACKNOWLEDGMENT

We would like to thank all the working staff of Shri Ramswaroop Memorial Group of Professional College of Engineering and Management and a great thank to our project guide Dr. Nidhi Saxena who gave her best to make us do this great project. Moreover, thanks to all the authors whose papers and books we referred during this project. Without help of all these resources, we would have never completed this project.

## REFERENCES

**[1]** Bjerager M., Palshof T., Dahl R., Vedsted P., Olesen F. Delay in diagnosis of lung cancer in general practice. Br. J. Gen. Pract.;56:863–868. 2006.

**[2]** Nair M., Sandhu S.S., Sharma A.K. Cancer molecular markers: A guide to cancer detection and management. Semin. Cancer Biol. 2018;52:39–55. doi: 10.1016/j.semcancer.2018.02.002.

**[3]** Silvestri G.A., Tanner N.T., Kearney P., Vachani A., Massion P.P., Porter A., Springmeyer S.C., Fang K.C., Midthun D., Mazzone P.J. Assessment of plasma proteomics biomarker's ability to distinguish benign from malignant lung nodules: Results of the PANOPTIC(Pulmonary Nodule Plasma Proteomic Classifier) trial. Chest.;154:491–500. doi: 10.1016/j.chest.2018.02.012.2018

**[4]** Shi Z., Zhao J., Han X., Pei B., Ji G., Qiang Y. A new method of detecting pulmonary nodules with PET/CT based on an improved watershed algorithm. PLoS ONE.;10:e0123694. 2015

**[5]** Lee K.S., Mayo J.R., Mehta A.C., Powell C.A., Rubin G.D., Prokop C.M.S., Travis W.D. Incidental Pulmonary Nodules Detected on CT Images: Fleischer. Radiology. 2017

**[6]** Diederich S., Heindel W., Beyer F., Ludwig K., Wormanns D. Detection of pulmonary nodules at multirow detector CT: Effectiveness of double reading to improve sensitivity at standard-dose and low-dose chest CT. Eur. Radiol.;15:14–22. 2004.

**[7]** Demir Ö., Çamurcu A.Y. Computer-aided detection of lung nodules using outer surface features. Bio-Med. Mater. Eng.;26:S1213–S1222. doi: 10.3233/BME-151418. 2015

**[8]** Bogoni L., Ko J.P., Alpert J., Anand V., Fantauzzi J., Florin C.H., Koo C.W., Mason D., Rom W., Shiau M., et al. Impact of a computer-aided detection (CAD) system integrated into a picture archiving and communication system (PACS) on reader sensitivity and efficiency for the detection of lung nodules in thoracic CT exams. J. Digit. Imaging.;25:771–781. doi: 10.1007/s10278-012-9496-0. 2012

**[9]** Al Mohammad B., Brennan P.C., Mello-Thoms C. A review of lung cancer screening and the role of computer-aided detection. Clin. Radiol.;72:433–442. doi: 10.1016/j.crad.2017.01.002. 2017