**Pulmonary Nodule Analysis for Lung Cancer Detection in Low Dose CT Images**

**FINAL YEAR PROJECT REPORT**

**ABDUL QADIR AHMED ABBASI**

**SESSION 2013-2017**

****

**SUPERVISED BY**

**DR. HAFEEZ UR REHMAN**

**DEPARTMENT OF COMPUTER SCIENCE**

**NATIONAL UNIVERSITY OF COMPUTER & EMERGING SCIENCES**

**(JUNE 2017)**

###### STUDENT’S DECLARATION

I declare that this project entitled “Pulmonary Nodule Analysis for Lung Cancer Detection in Low Dose CT Images”, submitted as requirement for the award of BS (CS) degree, does not contains any material previously submitted for a degree in any university; and that to the best of my knowledge it does not contain any materials previously published or written by another person except where due reference is made in text.

ABDUL QADIR AHMED ABBASI p13-6108 SIGNATURE:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Pulmonary Nodule Analysis for Lunge Cancer Detection in Low Dose CT Images**

**APPROVAL**

THE DEPARTMENT OF COMPUTER SCIENCE, NATIONAL UNIVERSITY OF COMPUTER & EMERGING SCIENCES, ACCEPTS THE THESIS SUBMITTED BY ABDUL QADIR AHMED ABBASI P13-6108 IN ITS PRESENT FORM AND IT IS SATISFYING THE DISSERTATION REQUIREMENST FOR THE AWARD OF BACHELOR DEGREE IN COMPUTER SCIENCE.

**SUPERVISOR**

DR. HAFEEZ UR REHMAN

ASSISTANT PROFESSOR \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**FYP COORDINATOR**

SHAKIR ULLAH SHAH

ASSISTANT PROFESSOR \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**HEAD OF DEPARTMENT**

DR. OMAR USMAN KHAN

ASSISTANT PROFESSOR \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

DATED: JUNE 4, 2017

**DEPARTMENT OF COMPUTER SCIENCE**

**NATIONAL UNIVERSITY OF COMPUTER & EMERGING SCIENCES**

**ACKNOWLEDGEMENT**

I would like to extend gratitude to my supervisor **Dr. Hafeez Ur Rehman** for having taken time out from his busy schedule for supervision of my project. I am grateful to the final year project (FYP) committee for extending their cooperation to me and helping me out in completion of project.

Special thanks to **Dr. Omar Usman (NUCES, Pakistan), Dr. Nouman Azam (NUCES, Pakistan) and Dr. Usman Habib** **(NUCES, Pakistan)** for helping me out throughout the project. I would also like to thank my family and friends for their support and help throughout my academic career.

***ABSTRACT***

*Lung Cancer is a leading cause of cancer related deaths worldwide. Screening high risk individuals for lung cancer with Low Dose CT scans is now being implemented in United States and other countries are expected to follow soon. In CT Lung Cancer screening, many millions of CT scans need to be analyzed, which is enormous burden for radiologists. Therefore there is a lot of interest to develop computer aided system to optimize screening. The goal of this work is to detect Lung Cancer in an early stage, for which Pulmonary Nodules have to be located which are the early manifestation of Lung Cancers. I purpose to work for a system that will automatically detect Pulmonary Nodules in Low Dose CT images and will identify if a subject of interest has Lung Cancer or not.*

**TABLE OF CONTENTS**

**Chapter 1: INTRODUCTION 8**

Motivation 8

Objective 9

SWOT Analysis 10

**Chapter 2: RELATED CONCEPTS 11**

What is Nodule? 11

Classification of Nodules? 12

**Chapter 3: SYSTEM ANALYSIS AND DESIGN……………………….……………13**

Use case diagram 13

Sequence diagram 14

**Chapter 4: DATASET…………………………………………………….……………15**

LIDC/IDRI Database 15

Annotations 15

**Chapter 5: PROJECT IMPLEMENTATION...……………...………….……………16**

Implementation Flow Diagram 16

Patch Extraction 17

Patch Normalization 18

3D CNN Architecture 18

3D CNN Training & Testing 18

Base Results 19

**Chapter 6: FUTURE RECOMENDATIONS…………………………….……………20**

Lung Segmentation 20

Parallel Computing Power Usage 20

On Disk Memory Usage 20

Saving Machine Stats with Third Party Tools 21

Saving CNN Training Results After Every Epoch 21

Utilizing Power of Cloud 22

LIST OF FIGURES

Fig: 1 - Example of Lung Nodule 11

Fig: 2 - Classification of Lung Nodules 12

Fig: 3 - Use case Diagram 13

Fig: 4 - Sequence Diagram 14

Fig: 5 - Annotations of Lung Nodules 15

Fig: 6 – Flow Diagram 16

Fig: 7 - Example of Extracted Non Nodule Patches 17

Fig: 8 - Example of Extracted Nodule Patches 17

Fig: 9 - 3D CNN Architecture 18

Fig: 10 – 1st Epoch Baseline Results 19

Fig: 11 - Un-Segmented VS Segmented Lungs 20

**CHAPTER 1**

INTRODUCTION

**Motivation:**

Lung Cancer is a leading cause of deaths worldwide. The National Lung Screening Trial (NLST) is a randomized controlled trial in the U.S. including more than 50,000 high risk subjects, showed that lung cancer screening using annual low dose computed tomography (CT) reduces lung cancer morality by 20% in comparison to annual screening with chest radiography [1]. In 2013 the U.S. Preventive Services Task Force (USPSTF) has given low dose CT screening a grade B recommendation for high risk individuals[2] and early 2015, the U.S. Center of Medicare and Medicaid Services (CMS) has approved CT Lung cancer Screening for Medicare recipients. As a result of those developments lung screening programs using low dose CT are being implemented in the United States and other countries. Computer Aided Detection (CAD) of pulmonary nodules can play an important role when screening is implemented on large scale.

To this end, different approaches have been used for classification of cancerous nodules in CT scans. Typically they start with candidate nodules creation and then these candidates are classified on the basis of some predetermined features which are designed to differentiate between nodules & non-nodules. Feature extraction in these approaches is majorly focused on intensity distribution and nodules geometry. They require segmentation of nodule region which may not be accurately segmented due to limitations of image processing techniques and dynamic nature of nodule sizes and intensities. Though these approaches have contributed some reasonable nodule detection results but there is a huge room of improvement because of the sensitivity & effect of disease.

Machine Learning has proved itself in many visual recognitions systems, for finding discriminative features without any segmentation and increasing object detection accuracy. To use the Machine Learning effectively huge dataset is required and it’s availability in my problem motivated to go for it. Recently Deep Convolutional Neural Networks have broken the previous performance records for many medical imaging challenges. A conventional 2D CNN approach for detecting a 3D object first convert 3D data into 2D frames and then feed them as input. No doubt 2D CNN approaches have improved the problem solving and have given better results but it completely lost the 3D context of actual image which limits the performance.

This has open a ground for 3D Convolutional Neural Networks to play their part in medical imaging & analysis problems. It requires two steps to be done which are mentioned below:

1. 3D Nodule & Non Nodule candidates generation
2. Classification of candidates using 3D CNN

**Objective:**

The objective of this project is to develop a system that detects and classifies Lung Nodules in Low Dose CT Images using 3D Convolutional Neural Network.

**SWOT Analysis:**

**Strengths**

* Annotated dataset available
* Quick, painless and non-invasive approach
* Automatic detection of Lung Cancer

**Weaknesses**

* Noisy input images
* Detection with 100% accuracy not possible due to noisy nature of biological images

**Opportunities:**

* Helpful for radiologists and hospitals
* Ongoing research area

**Threats:**

* Approval from patient privacy law authorities
* No collaboration of academia with hospitals in Pakistan

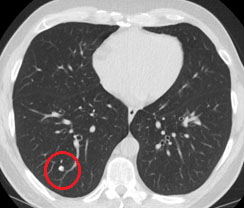
**CHAPTER 2**

RELATED CONCEPTS

**What is Nodule?**

A nodule is a small round oval shaped growth in the Lung. It may also be called a “spot on the lungs” or a “coin lesion”. Nodules are smaller than 3 centimeters in diameter. If the growth is larger than that, it is called a pulmonary mass and is more likely to represent cancer [3].

It has two types namely malignant & benign. Malignant are cancerous. Unfortunately no apparent symptoms are associated with its presences and they can only be detected with computed tomography or traditional X-rays. A nodule can be seen in figure shown below:

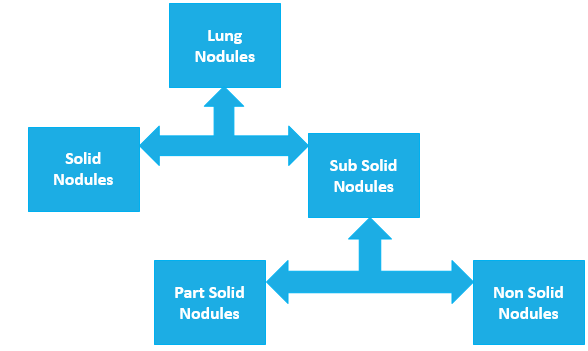


(Fig: 1- Example of Lung Nodule)

Nodules have different classes and each have a different impact on their growth rates, intensities, sizes, detection and cure.

**Classification of Nodules:**

Lung nodules can be distinguished in solid nodules and sub solid nodules. Sub solid nodules can be further classified as nonsolid nodules and part solid nodules. This classification is significant because different nodules require different approaches for their detection, measurement & management [5].

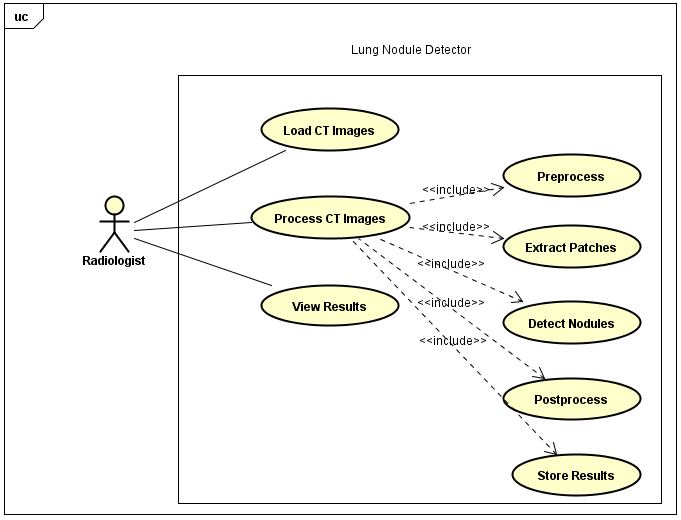


(Fig: 2 - Classification of Lung Nodules)

**CHAPTER 3**

SYSTEM ANALYSIS and DESIGN

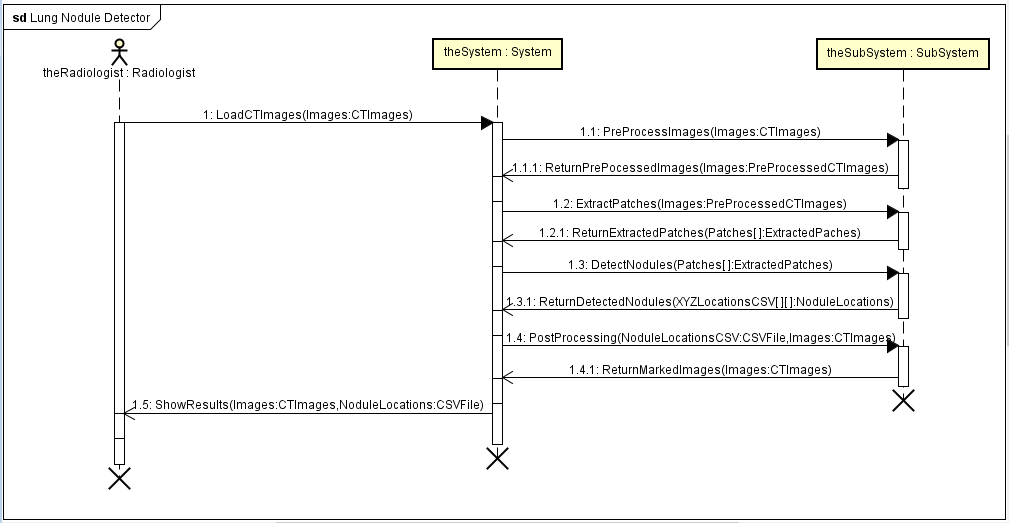
**Use case Diagram:**



(Fig: 3 – Use Case Diagram)

Picture shown above is a use case of how the proposed system shall be used by Radiologist. User comes and Loads some CT image in which nodule detection is required. System do some processing and shows results back to radiologist. In processing, a chain of sub processes starts with preprocessing of CT image, followed by 3D patch extraction which are needed for detection. Then results are computed, stored & returned to the request initiator.

**Sequence Diagram:**



(Fig: 4 – Sequence Diagram)

This is a sequence diagram which is demonstrating the methods/functions along with the parameters with their calling sequence. First method named as “LoadCTImages” is the request initiator method which takes only a CT scan as a parameter and forwards it for further processing. The method named as “ShowResults” gives back the results of detected nodules to initiator for the requested CT scan.

**CHAPTER 4**

DATASET

**LIDC/IDRI Database:**

For this problem 888 3D Ct scans are taken from the publically available LIDC/IDRI database [6]. The format of CT scans is “MHD”. Scans with slice thickness of more than 2.5mm are excluded. The dimensions are as follow:

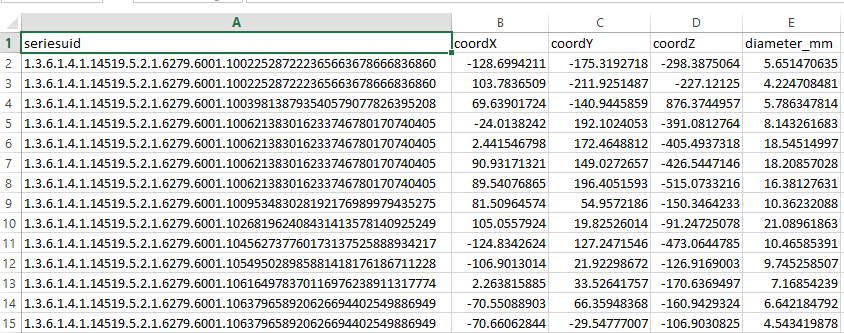
* X dimension: 512 pixels
* Y dimension: 512 pixels
* Z dimension: Varies from (100 – 250) slices

**Annotations:**

The annotations of nodules regions are also provided by LUNA Challenge forum. They are collected during a two phase process by 4 experienced radiologists. The annotations are as follow:

* Non Nodules
* Nodule < 3 mm
* Nodule >= 3 mm

The reading are stored in CSV files and a screenshot is shown below mentioning CT Scan ID, X Location, Y Location, Z Location and finally followed by Nodule Diameter in millimeters:

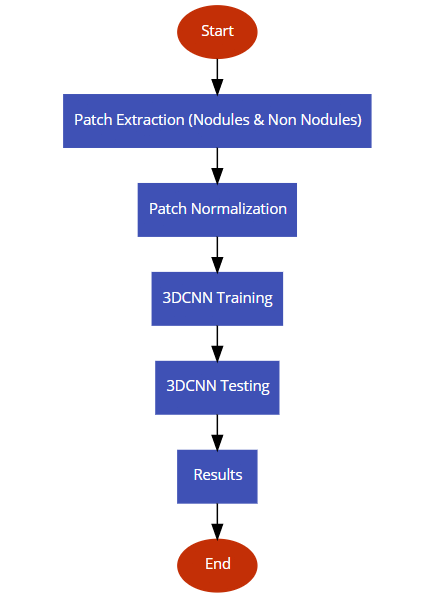


(Fig: 5 – Annotations of Lung Nodules)

**CHAPTER 5**

PROJECT IMPLEMENTATION

**Implementation Flow Diagram:**

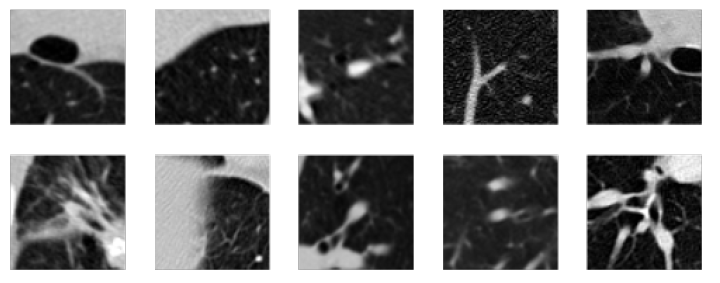


(Fig: 6 – Flow Diagram)

**Patch Extraction:**

* **Non Nodules**

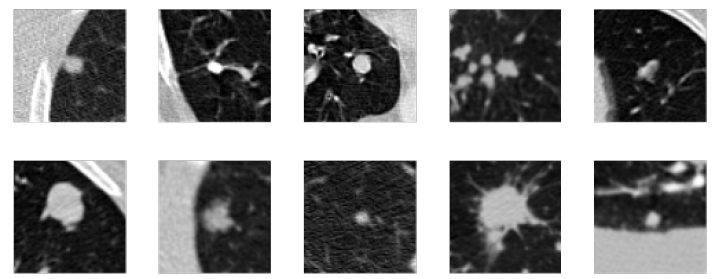
3D patches of Non Nodules are cut from CT scans using provided annotations. Total annotations for this class are more than 5 lacs. The patch size is kept as 32\*32\*32 in all three dimensions. Few middle slices in z dimension of 3D patches of non-nodules are shown below:



(Fig: 7 – Examples of Extracted Non Nodules)

* **Nodules**

3D patches of Nodules are also cut from CT scans using provided annotations. Total annotations for this class are only 1186 which are very less than 5 lacs (Non Nodule Patches). To deal with this problem and eliminating class biasness, several 3D patches are extracted around nodule locations by moving up, down, left & right one pixel in x & y dimensions. The patch size is kept as 32\*32\*32 in all three dimensions. Few middle slices in z dimension of 3D patches of nodules are shown below:



(Fig: 8 – Examples of Extracted Non Nodules)

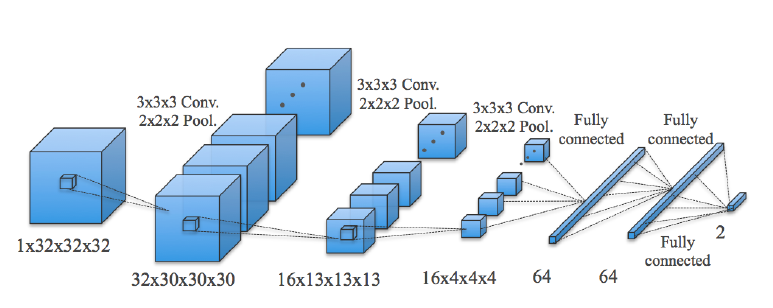
**Patch Normalization:**

3D patches of both nodules and non-nodules cut from different CT scans taken from different CT Machines are having variable CT intensity values. To deal with this problem, all 3D patches are normalized as:

* Min Value = -400
* Max Value = 1000

**3D CNN Architecture:**

The architecture proposed in paper [7] is simple but deep. A 32\*32\*32 input layer is used. After that three convolutional layers with 32, 16, and 16 small 3\*3\*3 kernels are used respectively. Each convolutional layer is followed by a max-pooling layer with overlapping 2\*2\*2 windows. Three fully connected layers with 64, 64, and 2 neurons are used respectively. Rectified linear units (ReLU) are used in each convolutional and fully connected layers. Dropout rate of 0.5 is applied after first two fully connect layers. The diagram of architecture is shown below:



(Fig: 9 – 3D CNN Architecture)

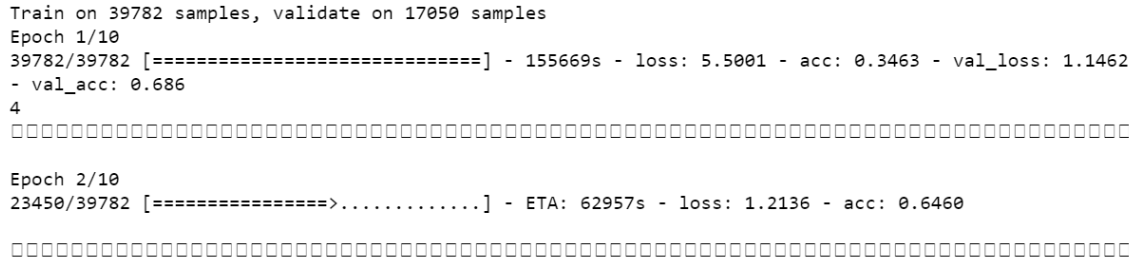
**3D CNN Training & Testing:**

Keeping in view the input size, computational and memory capacity, the 3D CNN model is trained on 39 thousand equally distributed samples of nodule and non-nodule patches. The model was required to be trained for 10 epochs and tested on 17 thousand equally distributed samples of both classes.

**Baseline Results:**

Due to the complex nature and huge size of problem, large memory and high computational power system was required with the continuity of power for approximately 40 days (4 days per epoch). To fulfill this, an Octa Core CPU was used which could handle 16 threads at a time. Memory need was fulfilled by using 64 GB ram installed in system. For power backup, a shared UPS was used with approximately 1 hour power backup time. The system was kept in university and remote access rights were given to me for using system from anywhere.

Due to huge shortfall of electricity in Pakistan during summers, the processes initiated were getting disrupted. On average it required almost 4 days for 1 epoch to train. In the mid of second epochs, system used to turnoff due to unavailability of continuous power and the process was required to start again from scratch. Due to this reason it was unfortunately impossible for me to give complete results. Anyhow the baseline results captured after 1 epoch were reasonable and there was a high probability of increase in accuracy and decrease in loss if the processing could have completed. System gave an accuracy of 64% on training and 68% on testing after 1st epoch. Screen shot of result is shown below:



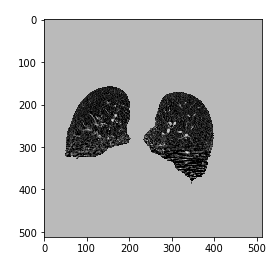
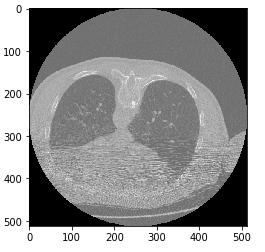
(Fig: 10 – 1st Epoch Baseline Results)

**CHAPTER 7**

FUTURE RECOMMENDATIONS

**Lung Segmentation:**

Training the model after segmenting the lung region can increase the accuracy of used 3D CNN Model rather than cutting patches from original CT scans. In most cases, segmentation helps the classifier to classify. Original CT scan is on left whereas segmented Lung is on right. All nodules will always fall inside lung region.



(Fig: 11 – Un-Segmented VS Segmented Lungs)

**Parallel Computing Power Usage:**

Parallel computing would be helpful and time efficient for this complex problem. One can exploit the power of thousands of cores in GPU and can get the work much earlier. Huge support is available in form of python packages which can run the CPU code on GPU easily & efficiently.

**On Disk Memory Usage:**

Instead of using RAM, one could use the hard drive as a memory with the help of some python packages. It will increase Input Output operations will increase but it will give more space in RAM for better computation and large data load.

**Saving Machine Stats with Third Party Tools:**

To capture the state of machine, third party tools could be used which can start the system and run high priority task by their own when the power comes again.

**Saving CNN Training Results After Every Epoch:**

It would be very handy if the results after every epoch could be stored to a file and then if the electricity goes off, then instead of initializing everything from scratch, the weights should be loaded from file and the further processing could be done. This would save a lot of time & processing.

**Utilizing Power of Cloud:**

This system can be run on Cloud where there will be no electricity shortfall and computing and memory scalability issues.

REFERENCES

[1] Aberle D. R., Adams A. M., Berg C. D., Black W. C., Clapp J. D., Fagerstrom R. M., Gareen I. F., Gatsonis C., Marcus P. M., and Sicks J. D. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med, 365:395–409, 2011.

 [2] Moyer VA, U.S. Preventive Services Task Force. Screening for lung cancer: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med, 160:330-338 2014.

[3] Cleveland Clinic. [Online]. Available: <http://my.clevelandclinic.org/health/diseases_conditions/hic_Pulmonary_Nodules>

[4] NCBI. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3569671/>

[5] LIDC/IDRI. [Online]. Available:

<http://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI>

[6] Xiaojie H., Junjie S. and Vivek V. Lung Nodule Detection in CT Using 3D Convolutional Neural Networks. GE Global Research, 2017.