Addis Ababa University
Addis Ababa Institute of Technology
School of Electrical and Computer Engineering

Lung Nodules Detection from Computed Tomography Scans Using Deep Belief Networks

By: Wogayehu Atilaw Mengesha

Addis Ababa, Ethiopia
October, 2018

Addis Ababa University
Addis Ababa Institute of Technology
School of Electrical and Computer Engineering

Lung Nodules Detection from Computed Tomography Scans Using Deep Belief Networks

By: Wogayehu Atilaw Mengesha

Advisor: Menore Tekeba

Addis Ababa, Ethiopia
October, 2018

Addis Ababa University
Addis Ababa Institute of Technology
School of Electrical and Computer Engineering

Lung Nodules Detection from Computed Tomography Scans Using Deep Belief Networks

By: Wogayehu Atilaw Mengesha

A Thesis Submitted to the Department of Electrical and Computer Engineering in Partial Fulfillment for the Degree of Master of Science in Computer Engineering

Addis Ababa, Ethiopia
October, 2018

Addis Ababa University
Addis Ababa Institute of Technology
School of Electrical and Computer Engineering

By: Wogayehu Atilaw Mengesha

This is to certify that the thesis prepared by *Wogayehu Atilaw*, titled *Lung Nodules Detection from Computed Tomography Scans Using Deep Belief Networks* and Submitted in partial fulfilment of the requirements for the Degree of Master of Science in Computer Engineering compiles with the regulations of the University and meets the accepted standards with respect to originality and quality.

Approved by board of Examining Committee:

	Name	Signature
Dean, School of Electrical and Computer Engineering:	Dr. *Yalemzewud Negash*	_____
Advisor:	*Menore Tekeba*	_____
Internal Examiner:	_____	_____
External Examiner:	_____	_____

Addis Ababa, Ethiopia
October, 2018

Declaration

I, the undersigned, declare that this thesis is my original work and has not been presented for a degree in any other university, and that all source of materials used for the thesis have been duly acknowledged.

Declared by:

Name: **Wogayehu Atilaw Mengesha**

Signature: _____

Date of submission: October, 2018

Addis Ababa, Ethiopia

This thesis has been submitted for examination with my approval as a university advisor.

Confirmed by advisor:

Name: **Menore Tekeba**

Signature: _____

ACKNOWLEDGMENT

I first and foremost, praise and thanks to God, the almighty, for His blessings throughout my research work.

I'm absolutely delighted to say thanks to my enthusiastic adviser, Mr. Menore Tekeba for his constructive comments and advices. He is always kind to correct my work, consistence until the end of this thesis.

My acknowledgement shall also pass to Haramaya University which provide me the opportunity to join in this program and support me until I finish my study.

I want also acknowledge all staff of Addis Ababa Institute of Technology school of Electrical and Computer Engineering for all of their cooperation.

I acknowledge the National Cancer Institute and the Foundation for the National Institutes of Health for their critical role in the creation of the free publically available LIDC-IDRI databases used in this thesis work.

Last but not the least, I would like to thank my family: my parents and to my brothers and sisters for supporting me spiritually throughout writing this thesis and all my friends.

Dedicated to

My brothers passed away on 12 August 2018

Abrham and Geremew

Abstract

Lung cancer is the leading cause of cancer related deaths globally. Analyzing thousands of computed tomography (CT) scans are an enormous burden for radiologists which results for inefficient diagnosis. Hence, a need to read, detect, and provide an evaluations of CT scans efficiently exist to assist radiologists by improving accuracy, time delay to diagnose, human errors, and making bias for specific reasons. Many researches have been conducted to detect lung nodules from CT scans. However, lung nodule detection focusing on the lung than the CT image as a whole has not been conducted so far. Thus, in this research work, lung nodules detection system, which segments lung and lesions from CT images to reduce false positives and employing Deep Belief Network (DBN), is proposed to improve nodule detection accuracy.

The study comprises three main phases namely: Image Processing, DBN training and nodules classification. The process starts with DICOM to JPEG conversion. Median filter and histogram equalization are applied for noise removal and contrast adjustment. We designed lung segmentation algorithm from the concept of inverse and intersection operation to separate lung object from the whole CT image. We applied adaptive thresholding for segmenting detail elements of CT images. Lesions on the lung are segmented. This thesis is conducted using datasets, publicly available on Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI). We implemented our methods based on 201 DICOM files that consist of 36,520 samples from which a total of 221,807 lesions are used to prepare feature vectors. This feature vector sets are used as input to the DBN algorithm for training and model construction. A DBN with 5776-500-500-2000-2 architecture, learning rate of 0.1, 100 number of epochs, and backpropagation as fining tuning algorithm is used in constructing the classifier.

The DBN classifier was validated with ten-fold cross validation technique. The proposed classifier achieves a sensitivity of 81.143%, specificity of 92.47% and an accuracy of 91.247%. The classifier has also an Area under the ROC curve (AUC) value of 0.882 and 0.924 for malignant and benign cases respectively. Therefore, based on the result we found that DBN model has the potential for lung nodule detection.

Keywords: *Computed Tomography, Ten-fold Cross Validation, DBN, DICOM, AUC and ROC curve.*

Table of Contents

List of Tables

List of Figures

List of Algorithms

List of Acronyms

AE: Auto-Encoder

AUC: Area Under the ROC Curve

ANFIS: Artificial Neuro Fuzzy Inference System

ANN: Artificial Neural Network

ANODE09: Automatic Nodule Detection 2009

BPNN: Back Propagation Neural Network

DBM: Deep Boltzmann Machine

DBN: Deep Belief Network

DIP: Digital Image Processing

CAD: Computer Aided Diagnosis

CT: Computed Tomography

CNN: Convolutional Neural Network

CXR: Chest X-Ray

DICOM: Digital Imaging and Communication in Medicine

ELCAP: Early Lung cancer Action Program

FLDA: Fisher Linear Discriminant Analysis

FN: False Negative

FP: False Positive

FPRED: False Positive Reduction

GPU: Graphical Processing Unit

GLCM: Gray Level Co-occurrence Matrix

GT: Ground Truth

JPEG: Joint Photographic Experts Group

KNN: K-Nearest Neural Network

LBP: Local Binary Pattern

LDCT: Low-Dose Computed Tomography

LIDC/IDRI: Lung Image Data Base Consortium / Image Database Resource Initiative

LUNA16: Lung Nodule Analysis 2016

MRF: Markova Random Field

MTANN: Massive Training Artificial Neural network

NDET: Lung Nodule Detection

PACS: Picture Archiving and Communication Systems

PND-DBN: Pulmonary Nodule Detection using DBN

RELU: Rectifier Linear Unit

RBM: Restricted Boltzmann Machine

ROC: Receiver Operating Characteristic Curve

ROI: Region of Interest

RQ: Research Question

SDAE: Stacked Denoising Auto-Encoder

TIFF: Tagged Image File Format

TN: True Negative

TP: True Positive

VDE: Virtual Dual-Energy

VOXEL: Volumetric Pixel

XML: Extensible Markup Language

2D: Two Dimensional

3D: Three Dimensional

Chapter One: Introduction

1.1 Background

Lung nodules are the most common indicators of an early stage lung cancer. Early detection of lung nodules is extremely important for lung cancer screening radiology, because it improves the chances of successful treatment. Lung cancer is the most frequent cause of cancer related deaths globally. It is caused by malignant lung tumor. It is characterized by uncontrolled cell growth inside the lung that is also known as lung carcinoma. This disease arises from a sequence of genetic changes that move a cell from a normal state to abnormal state. It is typically diagnosed at an advanced stage to cure when survival rate is very low due to delay in early detection of nodules [1, 2]. Early detection of nodules and proper treatment of lung cancer may pull down the death rates [3].

The problem of identifying nodules as benign or malignant tumors from CT scans can be grouped into two distinct challenges. The first is the Nodule Detection (NDET) challenge, where researchers are required to develop automatic lung nodule detection systems [4]. The input for this challenge is raw CT scans. In this case candidate nodules are detected and provide accurate classification of nodules as benign or malignant. Here, benign nodules are nodules which may indicate no lung cancer, while malignant nodules are early stage lung cancer indicators. The second challenge is the False Positive Reduction (FPRED) problem, where researchers are required to find solutions for false positive reduction stage [5]. This problem would be done after the detection of nodules, by taking those candidate nodules as input obtained from the first challenge. In this problem, researchers are given a set of candidates including both true nodules and false positives. They had to mainly assign a probability for being a nodule to each set of input candidates and provide the classification. Here the input is set of candidates, but the first challenge track is using raw CT scans.

Computer-Aided Detection and Diagnosis (CAD) helps radiologists to avoid missing lung cancer during screening, since it can identify affected regions from CT images. CAD system would be better used, if there was unsupervised way to represent high-level features of lesions (or nodules) from CT images automatically. Lung nodule detection systems with deep learning applications are used to automatically detect and identify nodules in CT images [1, 6].

In order to overcome the limitations of the earlier neural networks, Geoffrey Hinton introduces a deep learning application which was DBNs in 2006, mainly to simulate the learning process of the human brain [7]. It is specifically simulating the human brain's multilayer abstraction mechanism to achieve an abstract expression of an object. Thus, deep learning can improve malignant nodule detection rate [8] and using CAD system results serving as a useful second reader for radiologists. At this point, CAD is used as a second reader means, the radiologist first reads the images without CAD and then re-reads with the knowledge of the CAD findings for accurate diagnosis. Manual detection and diagnosis of an early stage lung nodules in CT images is challenging and time-consuming task, even if radiologists will experience pressure and heavy workload considering the large number of CT scans to analyze. Therefore, we used unsupervised deep learning method to automatically extract the deep high-level features of lesions in lung nodules detection system to improve the detection performance and productivity.

Researchers [1, 9], recommend that segmenting the boundaries of the lungs from the whole chest CT images decrease false positives (FPs) and improve the performance of the system by ignoring nodules predations outside the lungs. Image enhancement techniques are used in this work as a FP reducing step before using DBN training. CT images could suffer with intensity variability, uneven illumination, and high frequency signals. To minimize the effect of those artifacts with CT images, preprocessing is used via median filter and histogram equalization to obtain enhanced images from the input data [10, 11, 12]. Segmentation techniques such as thresholding, adaptive thresholding and intersection operations are used in order to reduce false positives inside the CT images. We need to prepare and represent all lesions inside the lung section which is suitable, standardized and reduced feature vector sets used to train the DBN algorithm in model construction process.

DBN is the current state of the art applications in deep learning techniques [7, 13]. It is used to automatically extract high-level features of the input data by training stack of RBMs and build a model together with the backpropagation algorithm for testing CT images. DBN consists of several hidden layers for feature extraction and high-level feature representations of the input data. Supervised fine-tuning is done with the backpropagation neural networks (BPNN) algorithm.

1.2 Motivations for the Study

Lung cancer is a lethal disease for humans, and its treatment is started during an advanced stage where at its very low survival rate. This is because manual based analysis of lung cancer in early stage is very difficult. The disease lacks symptoms for patients themselves until it distributes throughout the whole part of the lung, which is very difficult to cure it and it decreases the number of days they will live. Thus, lung cancer detection is a hot research area and motivates me also to do on it.

Deep learning technique creates a new uprising technique for medical image analysis. It is used to develop CAD tools which can reduce the mortality rate of lung cancer. Deep learning can be used for the detection of abnormalities in different modalities of medical images to improve early diagnosis. Feature extraction using deep learning methods in lung nodule detection systems gives better accuracy in medical image analysis. Studies recommend that deep learning remains a central focus of the research [1, 14], that would help in improving the results of conventional learning methods to achieve an efficient medical image analysis for CT images. We believe that using DBN for lung nodule detection will improve detection rate of malignant nodules.

1.3 Statement of the Problem

Medical imaging problems becomes urgent research area, especially in the detection and identification of abnormalities such as lung cancer, brain tumor, breast cancer and skin lesion through different image processing, artificial neural network and deep learning techniques. As Section 1.1 presents, lung nodules are the major early signs of lung cancer victim. Lung cancer is also the major lethal disease. There have been different significant researches conducted on lung nodules detection [1, 3, 14]. Those works acknowledge that lung cancer screening puts big pressure and burden for radiologists due to large amount of data to be analyzed and different types of patient CT scans, then this increase radiologists missing of a cancer. But previous works [1, 3, 15, 16] considers nodule like objects that exist outside the lung during training their machine learning, which highly increases false positives.

Currently, detecting malignant tumor that causes lung cancer is important, and timely treatment is the most effective way to improve the patient's survival rate [17, 18]. However, it is not easy for experienced radiologists to correctly detect lung nodules in its early stage due to difficult characteristics of nodules. For large amount of CT images, it takes huge interpretation time. Therefore, it is extremely

important to study the characteristics of lung nodules in lung cancer diagnosis. Most of the nodules are detected and diagnosed at an advanced stage of the disease, where it is difficult to cure. But it does not reach to an accurate level for detection of lung nodules. So, it needs advanced systems to detect nodules.

Existing systems for lung nodules detection have been unsatisfactory result [1], and it recommends that segmenting the boundaries of the lungs can decrease the number of false positives by allowing the system to ignore the nodules predictions outside the lungs. In this study, different image processing techniques are applied initially in order to obtain the lung region from the CT chest images to ignore the boundaries outside the lungs. Segment lesions from inside the lung. Then DBN is used to get deep representations of the lesions inside the lungs with the help of BPNN as a fine-tuning stage.

The traditional systems for lung nodule detection uses nodular segmentation, morphological processing, and artificial extraction of lung nodule features [19]. Therefore, detection and classification done based on low-level features is not reliable, even-if it can reduce radiologist's workload. Since the choice regarding which features best represent lung nodules mostly depends on experience and chance. It is difficult to select best features on CT images. Additionally, the use of morphological descriptions of the lung nodules cannot be accurate [20]. For example, the definitions of the nodular edges are ambiguous and subjective. This increases misdetection of nodules.

So typically observed problems by this researcher are:

- Previous works that applied deep learning techniques consider nodule like objects that exist outside the lung, which highly increases false positives.
- Traditional studies primarily rely on handcrafted feature extractions of the nodule morphology, which cannot be able to provide an accurate description of the nodule.
- Analyzing large amount of CT images and number of DICOM slices is a huge burden for radiologists and leads them to make an error.
- In manual diagnosis radiologists lack to consider knowledge about nodules representing lung cancer victim and they made biasness for some specific reasons.

In this thesis work, we are using image processing techniques for dimensionality and image complexity reduction, as well as region of interest (ROI) identification. We employ "Hinton" deep learning [7] approach with the RBM model to learn deep representations of the input data. It was already known that these techniques can be used to improve a classification model by providing a good initialization of its weights. Under the assumptions above, we have the following research questions (RQs).

1. Can image processing techniques (preprocessing and segmentation) reduces false positives while employing DBN in detecting lung nodules?
2. Is deep belief network algorithm a promise in detecting lung nodules, for lung cancer detection problem?

1.4 Objectives of the Study

General Objectives

The main objective of this thesis is to automatically detect lung nodules on lung and classify the occurrence of either malignant or benign nodules from CT images using Deep Belief Networks.

Specific Objectives

In light of this general theme, the specific objectives of this thesis work are the following:

- To remove noise and artifacts from the input of CT images.
- To segment the image for specific region of interest identification.
- To implement lung nodules detection using Deep Belief Networks from CT images.
- To test the model and evaluate the performance of the designed system.
- To draw conclusions based on the experimental results and recommend some future works.

1.5 Significance of the Study

- Since there are many lung cancer victims in Ethiopia, and the treatment given by radiologists is not yet supported by CAD methods for their diagnosis, this work motivates researchers to develop automatic systems, based on datasets of local hospitals.
- It provides a research output for lung nodule detection researchers in the development of lung nodule detection systems from CT images.

- This thesis output helps in lowering the number of slices to be analyzed manually, by providing information about in which slice malignant nodules exist from slices selected and identify the location of the nodule in a slice.
- The research plays a great role in understanding the steps and challenges of nodule detection and identification from DICOM images through DBN.
- The study helps in recognizing benign or malignant nodules from CT images and it also initiates researchers to do lung cancer detections with different approaches such as RCNN, RNN, DBM.

1.6 Contributions of the Thesis

- In researches done before for lung nodule detection systems, they consider the whole chest CT image simply by down sampling to the same size and feed to their respective algorithms. Using the whole chest CT images leads to a high computational time and false positives. In this thesis we identify the area of ROI which are all lesions (possible nodules). So, this study contributes a lot in reducing false positives by implementing segmentation algorithms in order to segment the lung region only and from the lung extracting lesions. This improves the overall performance of the detection algorithms, since different literatures [1, 6], recommend that deep feature extraction of the input data from the lung region only improves lung nodules classification.
- Appling DBN algorithm for lung nodule detection system together with the relevance of image processing techniques is implemented, which improve earlier stage lung cancer detection.
- Making the ground truth data usable to the DBN algorithm is done as discussed in Section 4.8. Since simply an XML file is provided in the dataset of LIDC annotations, we map the annotations to the training input dataset. We should convert annotations into slices of images and get the annotated nodules. However, the information about which DICOM slice number corresponds to which ground truth (annotated nodules) information has been only described with a text file. Thus, to identify the name of DICOM slice and the corresponding slice folder containing ground truth image, mapping of annotated data to the corresponding input vector must be done.

1.7 Scope and Limitation

Scope of the study

This work attempts to detect and identify benign or malignant lung nodules via DBN which helps to find lung cancer from CT images. In general, this thesis might be wide from DICOM to JPEG

conversion, reduce false positives using image processing techniques, training DBN for high level feature representations of lesions and BPNN for fine tuning and proper classification of the nodules as benign or malignant.

Limitations for the study

- Comparing the result of DBN with different classifiers such as SVM, Softmax, and Random forest is not the scope of this work, since our ultimate goal is reducing false positives and implementing DBN with feedforward neural network classifier.
- Working with all the datasets found in LIDC-IDRI would be the difficult task in this work due to machine constraints. Since the dataset is huge (around 124GB of 1,018 DICOM files with 244,527 samples (or slices)), managing and analyzing this dataset and conducting training phase is computationally heavy. As a result, this work will not use all dataset.

1.8 Research Methodology

This thesis presents an automatic nodule detection system from the combination of image processing techniques and DBN algorithm to detect and classify lung nodules from CT images. Because the effectiveness of nodule detection systems in detecting lung cancer has not been fully investigated, we propose lung nodule detection using DBN. This proposed system will be evaluated using confusion matrix and ROC curve metrics. Therefore, in order to conduct this research work, different methodologies will be used to select and implement appropriate methods and techniques.

1.8.1 Review of Related Literature

Literature Review: reading books, articles, research papers, journal papers, materials related to the subject matter which helps to understand the problem domains and for selecting efficient algorithms.

In the past years many CAD systems for detecting and recognizing lung nodules have been developed and tested in this active field of research. Before starting the actual work, a deep study will be made in the literature written on this area to have a clear picture about the work. Researches written on lung nodule detection and classification will be reviewed to get an understanding of the various techniques and methods of automatic lung nodule detection systems. Since deep learning applications boosted and related techniques have been improved for this problem in recent times, the interest to get an accurate lung nodule detection system become high.

1.8.2 Data collection

To conduct this research, which detects and classifies lung nodules, examples of CT images are needed. The dataset that we used here is obtained from LIDC-IDRI databases [21]. LIDC-IDRI database is mainly used as a reference standard for lung nodule detection and false positive reduction algorithms from CT images. The database consists of CT images of the lung with two-phase annotated lesions. A final annotation of a lesion is made when 3 of the 4 radiologists independently agree on the lesion. Based on the annotations lesions are classified into three categories: nodules>=3mm, nodules<3mm, and non-nodules<3mm. In this thesis only, the nodules >=3mm will be considered because of (nodules, non-nodules < 3mm) are said to be irrelevant findings [1, 21] and many related works have done on the same range. In the LIDC-IDRI base, all the images are in both MHD and DICOM formats. We considered the DICOM file format here. The database supplies an XML file with contour information for the slices. So, this LIDC-IDRI database contains 1018 thoracic CT images taken from 1010 patients with nodules of different shape and size.

1.8.3 Design and Implementation Tools

Design procedures of the research comprised a series of preprocessing, deep feature extraction and classification stages. In order to find fine lesions for deep feature extraction, preprocessing of CT images such as median filtering, histogram equalization, lung segmentation, lung artifact segmentation and feature vector preparation are accomplished. All suspicious features of lesions are extracted via DBN which is implemented by training stack of RBMs. Feedforward neural network is used for fine tuning the classifier for benign or malignant nodules classification. Testing is accomplished after testing datasets are ready through the same procedure of preprocessing stages that have been used for the training dataset.

With respect to the tools, the experimentation process was implemented using Matlab software 2014_b. DeeBNet [22], is another tool used in this work, which is an object oriented Matlab toolbox that provides tools for conducting researches using DBN algorithm. We used also Lambert's tool which is a LIDC-IDRI Matlab toolbox, that contains functions for converting the LIDC-IDRI database XML files into images [18]. The toolbox will only extract the slices for which annotations are found. Finally, experimentation, discussion on the results, conclusions and recommendations will be considered.

1.9 Organization of the Thesis

The remaining of this thesis report is organized as follows. Chapter 2 presents literature review on the theoretical backgrounds of lung cancer and the business domains. Chapter 3 introduces review of related works and it discusses researches works that have been conducted on lung nodule detection and classifications based on different approaches. The design of lung nodule detection system is presented in Chapter 4. The experimental results, evaluations, test results and discussions are described in Chapter 5. Lastly, in Chapter 6 conclusions and future works are pointed out.

Chapter Two: Literature Review

This chapter, gives some theoretical background information and business domains about lung cancer and different techniques to do the design of this work. A short introduction of lung cancer, CAD, and image processing are described in sections of 2.1, 2.3, and 2.5 respectively. Then, lung nodules, DICOM, CT and DBN are treated specifically. A short overview of lung cancer screening and clinical treatments are discussed.

2.1 Introduction

Lung cancer is a lethal disease caused by uncontrolled growth of malignant cells in the tissue of lungs [23, 24]. As pointed in Section 1.1, these malignant cells do not carry out the functions of normal lung cells and do not develop into healthy lung tissue. As they grow, the abnormal cells can form tumors and interfere the functioning of the lung, which provides oxygen to the body through the blood, and leads to cancer.

Majority of patients with lung cancer are diagnosed at an advanced stage, because of lack of early detection systems, as well as limited awareness of the public and health care providers about early signs and symptoms of cancer [2, 25]. In the world for the people with a lung cancer, during diagnosis the chances of surviving are very low, so that incidences of mortality with it, increases dramatically [26]. In our view this is due to accurate automatic systems unavailability, scarcity of hospitals for cancer, cancer drugs unavailability, radiotherapy centers shortage and scarcity of trained manpower. In general, it is difficult to detect lung cancer in early stages, where preventive actions are not taken and treatment engagement is deficient. So, that accurate systems for detection of lung nodules play an important role in reducing those problems when screening with it and implemented on a large scale in everywhere.

2.2 Lung Nodules

Lung nodules are very small deviations in the lung tissue. It is a medical term that describes a picture on a chest x-ray or a CT scan with a small spot in the lung and it is the most distinguishing feature of an early stage lung cancer [3, 15]. It is a white spot on the lung of the CT scan or a shadow, round area or that is more solid than normal lung tissue as shown in Figure 1. It measures usually from 3mm-30mm in diameter [23, 21]. A lesion >30 mm in diameter on the lung is referred to as pulmonary mass

and should be considered malignant until proven otherwise. Therefore, in this thesis we consider nodules that measures from 3mm-30mm in diameter for the detection and identification process.

Figure 1: How *Lung Nodules exist*

So, lung nodules can be either benign (non-cancerous) or malignant (cancer tissue) depending on the result of screening, which is an important part in nodule detection [27]. Nodule characteristics such as nodule size, growth rate, calcification, number of nodules, family history, and smoking history can indicate or help that a nodule is more likely malignant or benign. To determine the likelihood of nodules to be malignant or benign, radiologists take those characteristics as a prerequisite for their screening and diagnosis. Nodule size and growth rate have to be taken into more considerations to assess, because they are the most appropriate characteristics for each nodule management purpose [19, 16, 28].

Nodule Size: Nodule size is one of the most important characteristics of pulmonary nodule and it is a discriminant factor for nodule management [1, 28]. It is the key factor to decide further diagnostic follow-up and larger nodules are more likely malignant than smaller one.

Growth rate: Cancerous lung nodules tend to grow fairly rapidly with an average growing time of about 20 days to 400 days, while benign nodules tend to remain the same size the whole time. An increase in the volume of a nodule over time is used as a method to differentiate benign from malignant nodules [2]. If a nodule has grown, the size and the speed of growth should be considered to define its management. A very rapid growth rate (doubling time less than one month) of the volume is more suggestive of a malignant lesion. If the nodule growth doubling time is less than 400 days, three months' follow-up and biopsy can also be performed according to the nodule size and if the doubling time is more than 400 days repeat the CT scan at one-year follow-up can be suggested [29].

Smoking sigarate: Smoking is among the predominate cause of lung nodules in lung cancer diagnosis. It is the principal risk factor for the development of lung cancer. Current and former smokers are more likely to have cancerous lung nodules than never smokers [27, 29].

Medical and Family History: If the candidate person having a history of cancer, it increases the chance that a nodule could be malignant. If the person does not have any previous cancer related cases the chance of to be benign is high [28, 29]. In case of family history when a cancer candidate's person family have nodules and is lung cancer, then the candidate has to be more likely to have cancerous nodules than those candidates without a family history.

Number of nodules: Candidates those who have multiple nodules are more likely to have cancer than those who have a single or a few lung nodules [27]. But as we observed from different literatures, we aware that large number of nodules do not means that you are directly at risk of cancer, whereas a single nodule is also don't means that it is easy for treatment than many nodules. In medication systems of lung cancer in radiological evaluation process it is very common to notice that you can have a single lung nodule or multiple lung nodules in your CT scan, whereas multiple lung nodules do not mean that you are at more risk of lung cancer, what matters are the growth rate, size and location of those nodules.

Location: Lung nodules in the right lung and in the upper lobes have a higher probability for malignancy [2]. The probability of lung nodules existence on the lung is more likely to be on the right-side lung than from the left side. Studies in [27, 28] indicate that 70% of all lung cancers are located in the upper lobes.

Calcification: Lung nodules that are calcified are more likely to be benign. If a calcium deposit is found in a nodule it may mean that it has been there for a while [2, 28]. Calcified nodules detected at CT screening are considered by convention to be benign.

2.3 Computer Aided-Detection and Diagnosis Systems

Computer aided detection (CAD) systems are applications that assist radiologists in the interpretations of medical images [30, 31]. CAD systems using machine learning techniques can facilitate the

automated detection of lung nodules. The performance of CAD systems varies significantly depending on the size and nature of the samples used to compute performance metrics as well as the typical applied algorithms employed. It has the potential to assist radiologists during their earlier stage of patient's diagnosis [16]. Considering lung cancer, CAD systems help to detect each individual nodule of each lung and determines which sections are more to be malignant (tumorous). Then from this CAD output, radiologists can use the information to quickly find the affected regions of the lung and remove the time needed previously wasted by examining healthy regions. In addition, for earlier diagnosis; the use of CAD also reduces the number of misdiagnosis made by manual examination of radiologists. Therefore, in response to the rise of radiologists' burden for lung cancer detection & diagnosis researchers developed CAD systems for the detection of lung nodules from CT scans [30, 32].

Computed Tomography

A CT scan is a medical imaging method that combines multiple X-ray projections taken from different angles to produce detailed cross-sectional images of areas inside the body [33]. It uses computer processing to create cross-sectional images, or slices of the bones, blood vessels and soft tissues inside the chest that letting the user to see inside the scanned object without cutting [34]. CT images allow doctors (or radiologists) to get very precise, three-dimensional (3-D) views of certain parts of the body, such as soft-tissues, blood vessels, lungs, brain, heart, abdomen and bones. It is the preferred imaging modality for diagnosing lung cancer; due to its ability to form 3-D images of the chest, resulting in greater resolution of nodules and tumor pathology. In addition, it can reveal the whole information from the images of a patient. It can reveal small lesions in your lungs that might not be detected on an X-ray [23]. Instead of taking one picture, like a regular x-ray, a CT scanner takes many pictures as it rotates around you while you lie on a table. A computer then combines these pictures into images of slices of the part of your body being studied. This results in a 3-dimentional image of the chest, where each volumetric pixel (voxel) has an attenuation value that is indicative to the type of material present in its location. Currently, CT is the imaging modality that is most suitable for examinations of early detection of lung cancer.

Principles of CT image formation

The formation of CT image is a distinct three phase process: which is the scanning phase, the reconstruction phase and the shades of gray conversion phase [1, 35]. The scanning phase produces

data, but not an image. During this phase a fan-shaped x-ray beam is scanned around the body. As x-ray beam is scanned around the body, forming many views, the data recorded by the detectors are stored in computer memory for later image reconstruction. The projection of the fan-shaped x-ray beam from one specific x-ray tube focal spot position produces one view. Many views projected from around the patient's body are required in order to acquire the necessary data to reconstruct an image. Then a complete scan is formed by rotating the x-ray tube completely around the body and projecting many views. This produces a complete dataset that contains sufficient information for the reconstruction of an image. In the image reconstruction phase the scan dataset is processed to produce an image. The image is digital and consist of a matrix of pixels. Filtered back projection is the reconstruction method used in CT image formation. In the third phase the digital image is converted into visible shades of gray image. In this phase the digital image, consisting of a matrix of pixels with each pixel having a CT number is converted into a visible image represented by different shades of gray or brightness levels.

What is DICOM? What is DICOM image file?

DICOM stands for Digital Imagining and Communications in Medicine standard. This standard is established by the American College of Radiology (ACR) and the National Electric Manufacturers Association (NEMA) [1, 36]. DICOM standard specifies a nonproprietary data interchange protocol, digital image format, and file structure for biomedical images and image-related information. It specifies a communication protocol for message exchange, which is an application protocol that uses TCP/IP to communicate between systems. It is a comprehensive specification of information content, structure, encoding, and communications protocols for electronic interchange of diagnostic and therapeutic images. DICOM files are represented with ". dcm" extension [19]. DICOM differs from other image formats with that of it groups information into datasets. A DICOM file consists of a header and image datasets, all packed into a single file as shown in Figure 2. All modern medical imaging systems such as X-rays, US, CT, MRI and so on support DICOM standard and use it extensively [30, 31]. Those of medical imaging equipment's create DICOM files and radiologists use DICOM viewers, such as computer software applications that can display DICOM files.

A DICOM image file is an outcome of the Digital Imagining and Communication standard. All medical images are saved in DICOM format. DICOM files contain more than just images. Every DICOM file holds patient information (name, ID, sex, birth date), important acquisition data (type of equipment

used and its settings), and context of the imaging study that is used to link the image to the medical treatment it was part of.

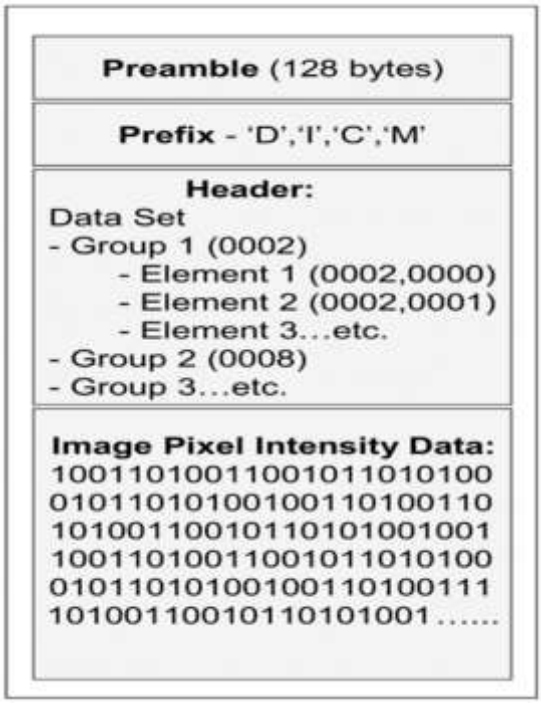

Figure 2: *Structure of a DICOM Image File*

As shown in Figure 2, the first few packets of information in a DICOM image file constitute the header (or metadata). The header stores demographic information about the patient, acquisition parameters for the imaging study, image dimensions, matrix size, color space, and a host of additional non-intensity information required by the computer to correctly display the image. This metadata is followed by a single attribute (7FE0), that contains all the pixels intensity data for the image [37]. These data are stored as a long series of 0s and 1s, which can be reconstructed as the image by using the information from the header. This attribute may contain information regarding a single image, multiple frames of study depending on the modality that has generated the image. The header data information is encoded with in the DICOM file, so that it cannot be accidentally separated from the image data. If the header is separated from the image data, the computer will not know which imaging study has been done or to whom it belongs and it will not be able to display the image correctly.

2.4 DICOM Image Acquisition

Acquisitions of data from the various imaging modalities (or databases) for input to any system is the first step of their task. Image acquisition from the inherently digital modalities such as CT, MRI and US should be a direct digital DICOM capture [38]. In this automatic lung nodule detection system, DICOM image acquisition is an activity of acquiring CT images from LIDC-IDRI databases. Therefore, the first step in this problem is DICOM image acquisition. This input data is obtained from the publically available resources [21]. Beyond to this database many researches have been conducted by collecting the images from different private hospitals, research institutions, and taking CT scans by themselves. A research paper done in [9], used private hospitals for the data to test their developed machine learning models.

2.5 Image Processing

The purpose of digital image processing (DIP) step is to prepare the dataset in the way that usable for the learning machine classifier algorithm. Image processing is the way of manipulating images in various techniques in order to get easily visualized and detected images. A large part of the work of ML nowadays is in relation with images [10, 11, 12]. There are large collections of DICOM images available on the LIDC-IDRI databases. Those DICOM images have a tendency of being complex. Moreover, the images may present a lot of difficulty to a computer model and can be of high dimensionality making them harder to process in a reasonable amount of time. Therefore, because of their complexity, dimensionality and its file format instead of a single image, we used image processing techniques in order to enhance the complexness, quality, easier manipulation and to reduce false positive objects outside the lung. It also helps to find spatial dependencies inside the image to be able to learn features that can shared across the complete image rather than be specific to pixels. The aim of image processing in our proposed system architecture is to reduce false positives of DICOM images and improve the quality of images taken from DICOM files that helps us in finding suspicious objects. It also reduces the size of the file and helps to get accurate ROI for the suspicious lesions. In addition, it makes the input data suitable for machine learning algorithms to train over the data. Because the proposed DBN algorithm only works on numeric valued and having in the same range between zero and one. So, before the dataset as input to the DBN algorithm the dataset should be prepared in such a way that usable for the system first. Thus, we use some image processing techniques such as image preprocessing, image segmentation and input vector representations.

2.5.1 DICOM to JPEG Conversion

DICOM file has two disadvantages which are large file sizes and special software is needed in viewing them on personal computers. Converting images from DICOM image format to other image format is accompanied by data compression [35]. JPEG format is the most popular format and can be read by all computer platforms [11]. It is commonly used method for compression of digital images. Because JPEG files are small in size and extremely portable, they are the preferred format. The advantage of the JPEG format is that it facilitates the use of compression to reduce file size. DICOM supports the use of JPEG image compression through the encapsulated format. JPEG is the most popular, compatible and high-quality image format, for easily manipulation [39]. As researches conducted on lung nodule detection [35, 9], suggests that it can be easily processed when they convert DICOM to JPEG before training machine learning algorithms. In terms of image quality, image compression during conversion of DICOM image format to other image format is acceptable in most areas of radiology on different devices. When converting DICOM image format to JPEG image format with low compression ratio there is no significant difference in the quality of converted images for the interpretation on CT scanner [11]. The advantage of converting into JPEG image format is to reduce the size of an image into blocks and its best lossless compression [40, 41].

2.5.2 Image Preprocessing

Preprocessing step is used mainly to reduce the noise and unwanted artifacts in the image and misrepresentations of the image [10, 11]. It is applied on images at the lowest level of abstraction. The key function of preprocessing is to improve the images observation that increases the chances of successful visual of lesions for the next components. So, our proposed approach uses histogram equalization, data normalization and median filter to enhance the image quality [11, 12]. Noise and high frequency components are some of the unwanted regions present in the CT scan images.

2.5.2.1 Normalizing image inputs

Data normalization is an important step which ensures that each input parameter (pixels in this thesis) has a similar distribution [10, 11]. This makes convergence faster while training the network. Data normalization is done by subtracting the mean from each pixel, and then dividing the result by the standard deviation. For image inputs we need the pixel numbers to be positive, we might choose to scale the normalized data in the range [0, 1] or [0, 255]. For our dataset we used the range of [0, 1].

Data normalization can be calculated using equation (1), where Xmax and Xmin are the maximal and minimal values for the variable x data respectively.

$$\boldsymbol{Normalized} = \frac{\boldsymbol{x-xmin}}{\boldsymbol{Xmax-Xmin}} \quad \text{.....(1)}$$

2.5.2.2 Histogram equalization

Histogram equalization is a method in image processing for contrast adjustment using the image's histogram [11]. This method increases the global contrast of the images, especially when the usable data of the image is represented by close contrast values. Through this adjustment, the intensities can be better distributed on the histogram. Equation (2), can be used to calculate the equalized histogram. Since, our input data is highly affected by intensity variability and uneven illumination, we apply histogram equalization for better views and better detail in CT images. We used this technique to produce high quality JPEG image, which is easily usable for the next component. As it is applied in different research works [19], it adjusts the contrast and enhance the image views of the internal structures in images from x-ray, CT, MRI and US. The key advantage of using this technique is that it is a fairly straightforward technique and an invertible operator. The histogram of a digital image with gray levels in the image $h(r_k) = n_k$, where r_k is the kth gray level and n_k the number of pixels in the image which have gray level r_k. The equalized histogram can be obtained using cumulative distribution function (CDF) formula S_k as follows:

$$\boldsymbol{S_k = T(r_k) = \sum_{j=0}^{k} \frac{n_j}{n} = \sum_{j=0}^{k} P_r(r_j)} \quad \text{..... (2)}$$

Where, $0<r_k<1$ is the normalized gray level and K=0,1,2…...L-1 L is the gray level number.

2.5.2.3 Noise removal using median filter

Image denoising is an important step specially in medical image processing, where the original images are poor due to the noises and artifacts introduced by the acquisition systems. Median filter is a non-linear operation used to reduce artifacts and salt and pepper noises [12]. As sharpening and shadows perform using high frequency signal in the image, the noise of the image will get higher. Noise removal using median filter is more effective in terms of eliminating noise and preserving edges and fine details of digital images. As it is applied in many works such as [9, 11] to reduce noise mainly they used median filter. There are a number of different filters, such as low pas, high pass, mean, median, Weiner etc. available [11, 12]. We used median filter technique for our method to remove some of the noises

because under certain conditions it preserves sharp edges of the images and fine details of digital images in the CT scan while removing noise. The median filter value is calculated using equation (3).

$$\boldsymbol{y(m,n) = median[x[i,j],(i,j)\varepsilon\omega]} \text{ ---------- (3)}$$

Where ω represents a neighborhood defined by the user center around location (m,n).

2.5.3 Image segmentation

Segmenting an image into meaningful parts is a vital operation in image processing [12]. It refers to another step-in image processing method where the inputs are images and outputs are attributes extracted from images. It subdivides an image into multiple connected regions/segments [10, 11]. It is primarily essential to processes such as CAD, quantitative analysis, visualization, registration and many more. In medical imaging aspect the selection of segmentation methods are widely depends on the specific applications and imaging modality [10, 31]. Therefore, in this current work segmentation techniques help us to segment significant regions and suspicious objects to be probable nodules. This operation extracts various structures from CT images such as blood vessels, lesions, nodules, soft bones, arteries and so on for further examinations. There are different image segmentation techniques in the state of the art such as edge-based segmentations (differential coefficient technique, Laplacian of Gaussian), special theory-based segmentations (wavelet, morphology, fuzzy, neural network) and feature based segmentations (clustering techniques, thresholding, intersection operations). We have used techniques like thresholding, adaptive thresholding, morphological operations (dilation, erosion and filling), Mathematical inverse and intersection operations for segmenting our image into meaningful parts. We have used thresholding techniques for enhancing the performance. Thresholding methods are image segmentation techniques based on image space regions [12]. Equation (4), shows the formula to compute thresholding. It provides an easy and a convenient way to separate background pixels (usually set to black) from those corresponding to the target objects (usually set to white). This method converts a gray value image into a binary image. By selecting an adequate threshold value T, the gray level image can be converted to binary image. It chooses proper thresholds T to divide image pixels into several classes and separate the objects from background. When there is only a single threshold T, any point (x, y) for which f(x, y)>T and a point on (x, y) is called an object point and f(x, y) <T is called a background. This can be done by the following equation (4)

$$F(x, y) = \begin{cases} \mathbf{1} & \boldsymbol{if\ f(x,y) \geq T} \\ \mathbf{0} & \boldsymbol{otherwise} \end{cases} \quad \ldots\ldots (4)$$

Where T is the threshold value.

We have employed also an adaptive threshold method to extract important features by segmenting the whole chest CT image. Adaptive thresholding changes the threshold value dynamically over the image, to handle changing lighting conditions in the image, those occurring as a result of strong illumination gradient or shadows [11, 12]. There are two typical adaptive thresholding methods which are adaptive mean thresholding and adaptive Gaussian thresholding. We used adaptive mean threshold technique, where the threshold value is the mean of the neighborhood area. We put a window size for this method and it calculates the mean threshold dynamically over the region to extract the features. We have tried different window size for this operation and chose 32*32 works fine for our case. The advantage of using adaptive thresholding technique is that since it changes the threshold dynamically over the image, it deals with intensity variability, images containing with strong illumination gradient and changing lighting conditions in the image.

2.5.4 Morphological operations

Morphology is a vast extent of image processing operations that modifies the images based on shapes. Binary images contain countless defects. The goal of morphological image processing is to eliminate those defects and maintain the structure of the image. There are different morphological operations such as dilation, erosion, filling, opening and closing expressed in logical AND, OR notation and set analysis. Dilation is applied to binary image but can also be applied to gray scale image. Dilation causes the objects to grow in size by adding pixels at boundaries of the objects. Whereas, erosion causes the objects to shrink in size by eroding away the boundaries of the objects which results in areas of pixels shrink in size and holes of those areas become larger [11, 12]. Fill operation is applied to fill the holes in the given input image. For binary images, it changes the background pixels to foreground pixels until it reaches the object boundaries and for gray scale images it changes the background intensity level to foreground intensity level.

2.6 *Machine Learning*

Machine learning (ML) is a set of methods that automatically detect patterns in data, and then utilize the uncovered patterns to predict future data or enable decision by making under uncertain conditions [4, 42]. It is a sub-part of AI which is designed as an approach to achieve artificial intelligence. Machine

learning algorithm is a science aiming at getting machines to learn solutions to specific problems without being explicitly programmed and that enables to learn more from data [43]. The most representative characteristics of machine learning is that it is driven by data, and the decision process is accomplished with minimum interventions by humans. The program can learn by analyzing the training data and then make a prediction when new data is put in. Machine learning models are used to solve two main tasks: classification and regression. This thesis focuses on extracting features from images automatically with ML; rather than relying on handcrafted feature extractors.

In all, ML is the practice of using algorithms to parse data, learn from it and make prediction of the world. To solve problems (classification or regression problems) three machine learning algorithms are particularly considered [4, 43].

1. Supervised learning: The model is learned from the input and the expected output data. These is the most common way of learning. It uses labeled training data to learn the mapping function from the input variables to the output variables
2. Unsupervised learning: The model is learned only from the input data. This approach is particularly useful in practice since unlabeled data is abundant while labeled data is scarce and requires a lot of effort to collect. This approach gives input data to the algorithm and it learns and predict from experience, which is mostly through association, clustering and dimensionality reduction. It mostly learns the correlations among the input data to reconstruct it again.
3. Semi-Supervised learning: In this approach both kinds of data are used to train the model. The model is first pre-trained using unsupervised data and then improved with supervised data. When a neural network is to be used for classification; it first pre-trained layer by layer using unsupervised training algorithm. Then finally the network can be trained with a standard training algorithm, for classification or prediction.

2.6.1 Deep Learning

Deep learning is the growing trend to develop automated applications and has been termed as one of the 10 breakthrough technologies in 2013 [42, 43, 13]. Today, several deep learning-based computer vision applications such as CNN, RNN, DBN, DBM, SDAE are performing tasks even better than humans. It is an improvement of artificial neural networks that consist of more hidden layers that permits higher level of representation and improved image analysis. They have the capability to learn

from DICOM images for identifying lung cancer tumors using CT and MRI scans [3]. It becomes extensively applied method due to its recent unparalleled result for several applications such as object detection, speech recognition, face recognition and medical imaging [23]. A deep neural network hierarchically stacks multiple layers of neurons, forming a hierarchical feature representation. The number of layers now extends to over 1000 with such a gigantic modeling capacity. A deep neural network can essentially memorize all possible mappings after successful training with a sufficiently large knowledge database and making intelligent predictions such as doubtful objects and non-doubtful objects from images of unseen data. Thus, deep learning is generating a major impact in computer vision and medical imagining. In fact, similar impact is happening in the domains like text, voice, etc.

2.6.2 Deep Belief Networks

DBN is one of the most important frameworks of deep learning technique and has been widely used in feature extraction, classification, recognition and other tasks [7, 13]. It was discovered by Geoffrey Hinton in 2006, that uses the principles of greedy-layer wise training to initialize parameters before performing any discriminative or generative fine-tuning [42, 24, 22]. Thus, the introduction of DBN in 2006 began the current deep learning renascence [42, 24]. It is a generative graphical model with several layers of latent variables, trained with a greedy layer wise learning algorithm. The latent variables are typically binary, while the visible units may be binary or real. DBN is the sub-part of deep learning model which is used for reduction of dimensionality of large amount of data by doing a repetitive training and testing on the data. The building blocks of DBN is a probabilistic unsupervised model called RBMs and used as building blocks for training deeper models. DBN is also a probabilistic model composed of multiple layers of stochastic, hidden variables and it uses those latent variables to learn features from the data. The learning procedure of DBN can be divided into two stages: generative learning to abstract information layer by layer with unlabeled samples firstly and then discriminative learning to fine tune the whole deep network with labeled samples to the ultimate learning target [42]. During training DBN through RBMs the joint probability distribution of the DBN can be calculated using equation (5).

The main applications of DBN in this research is deep and complex feature extraction from CT images. It is argued [1, 35, 23, 24, 31], that in order to overcome the current problems of lung nodule detection algorithms, researchers should focus on accurate feature extractions, rather than on the classification

process. So, this ensures that, the choice of our model DBN is pretty good start of our work. Because DBN through its unsupervised pre-training fashion achieves better feature extraction on the CT images, before performing any classification task.

The architectures of DBN, comprise several layers of RBMs or variety of auto encoders depending on the problems to be solved. This DBN is stacked with three RBMs as shown in Figure 3. Each RBMs are stacked on top of each other to be trained in a greedy-wise learning method. The DBN model learn the joint probability distribution over the inputs. Figure 3 presents a DBN with three-layers of an RBM.

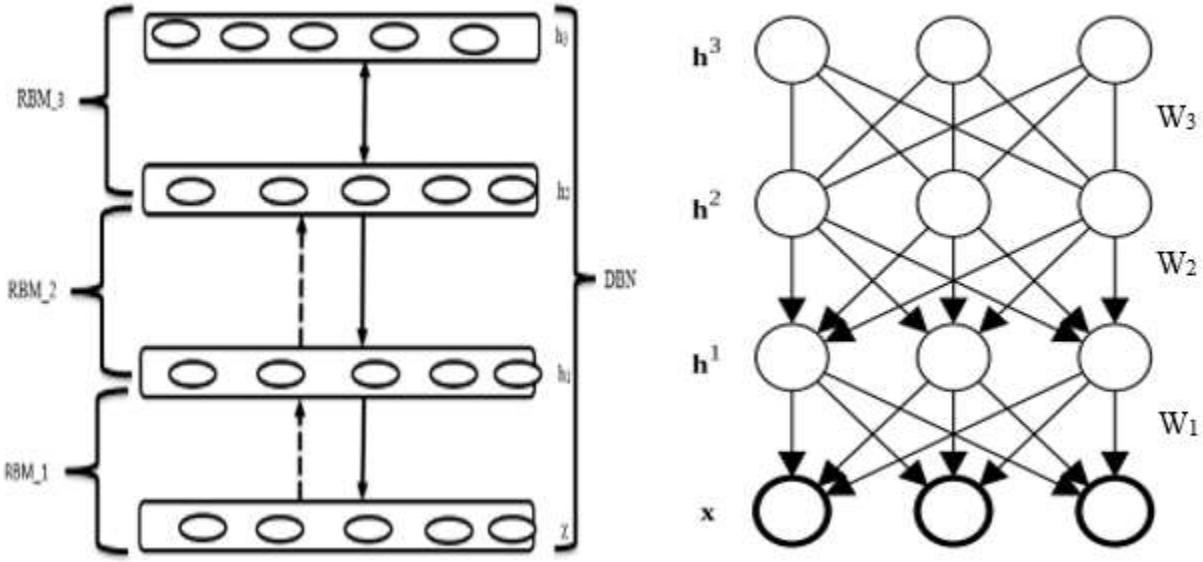

Figure 3 : *Deep Belief Network Architecture with 3-Stacked RBMs*

$P(v, h^1, \dots, h^l) = \left(\prod_{k=0}^{l-2} P(h^k|h^{k+1})\right) P(h^{l-1}, h^l)$---------- (5)

Where:

- x= h^0, $P(h^{k-1}/h^k)$ is a conditional distribution for the visible units on the hidden units of the RBM at level k
- $P(h^{l-1}, h^l)$ is the visible hidden joint distributions in the top level of RBM.

A DBN is first pre-trained, layer by layer, using RBM layer. Each layer is an RBM and they stacked each other to construct the DBN. The first step of training DBN is to learn a layer of features from the visible units, using Contrastive Divergence (CD) algorithm. The next step is to treat the activations of

previously trained features as visible units and learn features of features in a second hidden layer. Finally, the whole DBN is trained when the learning for the final hidden layer is achieved. This is a fast procedure since each layer is trained one after another. This process is done via a greedy layer-wise pre-training algorithm, in which each model in the sequence of layers is learning from a different representation of the input. This because that training RBM using CD algorithm for each layer looks for the local optimum and the next stacked RBM layer takes those optimally trained values and again look for the local optimum. At the end of this procedure, it is likely to get the global optimum as each layer consistently trained to get the optimum value.

2.6.2.1 Restricted Boltzmann Machines

RBM is a generative stochastic ANN. It is a model specially to learn a probability distribution over its inputs. It is made of two layers a visible layer and a hidden layer [7, 13, 44]. Both layers contain a certain number of units (neurons). An RBM is a variant of the normal class of Boltzmann Machine, proposed by Hinton. In an RBM restricted means the neurons form a bipartite graph i.e. there are no connections between units of the same group. This special restriction makes for more efficient algorithms to train the model with this RBM instead of Boltzmann Machine. We used the RBM as feature extractor of the inputs, while the extracted features are the activation probabilities of the output (or hidden) layer. Training an RBM means maximizing the probabilities of the input samples P(v) and the most often used algorithm to train an RBM is the contrastive divergence (CD) algorithm. Samples from an RBM can be obtained using Gibbs sampling method. So typically, RBM is used as feature extractors for higher level classifiers or to initialize the weights of a feed forward neural network. Figure 4 presents an RBM with five visible units v and four hidden units h.

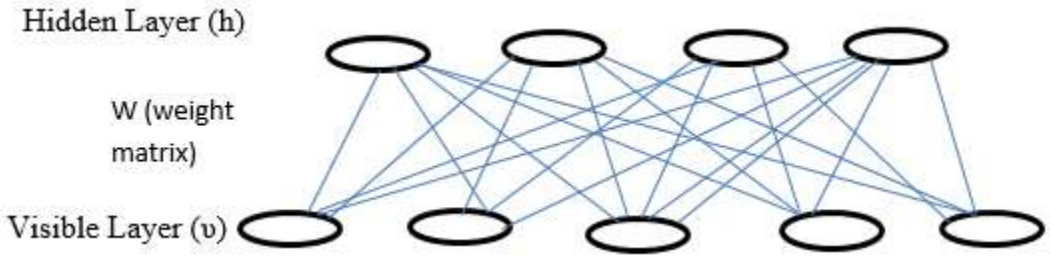

Figure 4: *RBMs with Five-Hidden Layers and Four Visible Layers*

To achieve classification, the RBM can be used to model the joint distributions of the inputs of 'm' visible units of V= (v_1, v_2……v_m) to represent observable data and 'n' hidden units H= (h_1, h_2….h_n) to

capture dependencies between observable data. It consists of a matrix of weights W= ($w_{m,n}$) associated with the connection between hidden unit 'h_n' and visible unit 'v_m' and the bias unit as a_m for the visible units and 'b_n' for the hidden units. Then it is trained with greedy algorithm with training objective that it optimizes P (V, H). Here the energy function for each RBM is calculated using equation (6).

$$\boldsymbol{E(V,M) = \sum_{i=1}^{m} \sum_{j=1}^{n} W_{m,n} h_n v_m - \sum_{i=1}^{m} a_i v_n - \sum_{j=1}^{n} b_j h_n} \text{ ---------- (6)}$$

Where:

- E (V, H) = total energy of the RBM configurations.
- $w_{m,n}$= strength of the connections between two nodes m and n in the network
- a_i, b_j = bias units of visible and hidden layers
- v_m = visible units
- h_n = hidden units of the network. The network assigns a probability to every possible pair of a visible and hidden vector via this energy function.

The probability distributions over hidden and visible vectors is defined by equation (7) and the normalizing partition function is defined by equation (8).

$$\boldsymbol{P(V,H) = \frac{1}{z} e^{-E(v,h)}} \text{ ---------- (7)}$$

$$\boldsymbol{Z = \sum_{v,h} exp\left(-E(v,h)\right)} \text{ ----------(8)}$$

Where

- P (V, H) = probability distributions of hidden and visible vectors.
- Z is the normalizing partition function.

RBM is two-layer generative models. It tries to model the training dataset over its hidden layer units. The probability that the network assigns to a visible vector 'v' is given by summing over all possible hidden vectors which can be calculated by equation (9).

$$\boldsymbol{p(v) = \frac{1}{z} \sum_{h} e^{-E(v,h)}} \text{----------(9)}$$

Gibs sampling of the visible and hidden unit pair of variables is used to estimate the gradient on the likelihood of RBM [7]. It is the process of consecutive sampling of hidden unit 'h' given visible unit 'v' then visible unit 'v' given hidden unit 'h' until the end of the chain. The chain starts from t=0 meaning from the input vector then proceed to (v_t, h_t) where 't' is the number of sampling iteration as

shown in Figure 5 [13, 44]. To fasten the learning for an RBM, contrastive divergence algorithm is used and the general idea is to update all the hidden units in parallel starting with visible units, reconstruct visible units from the hidden units, and finally update the hidden units again. This learning rule to update the hidden units is calculated based on equation (10).

$$\Delta w_{ij} = (v_i h_j)_{data} - (v_i h_j)_{recon} \quad \text{.....(10)}$$

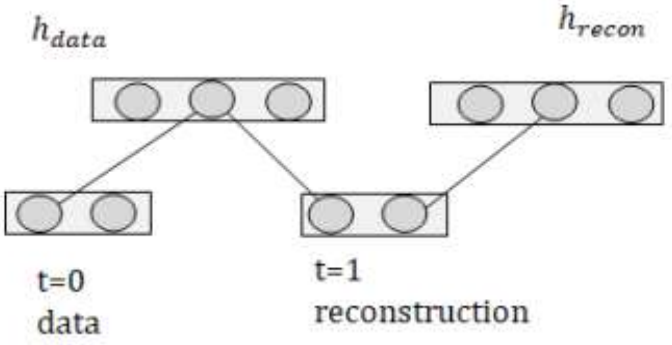

Figure 5: Contrastive divergence (CDn) with n=1

In RBM training parameters play a significant role on its performance. So, some considerations will be taken when training RBM algorithm. The main parameters are number hidden layers, number of computing units, training epoch, learning rate, batch size and others [7, 43]. To implement this work preliminary experiments are done to select the appropriate parameter values and some of them are taken as their default value on the RBM.

Number of hidden layers: The number of hidden layers affects the fitting degree of data directly. In theory the more layers of network there are the more complicated the network structure is, making the network express data precisely and ultimately obtaining a higher accuracy [7, 43, 44]. However only increasing the number of hidden layers may lead to difficulty in neural network training, greatly extend the learning time and decrease the accuracy. The number of hidden layers is studied in this thesis. Setting the number of hidden layers as 2, 3, and 4 respectively (excluding the input and output layers), the accuracy is calculated and we used 3 as the number of hidden layers in this thesis by making preliminary experiment in Chapter Five.

Number of computing units of hidden layer: Hidden layer is feature extraction part of RBM. For such purpose it uses a number of computing units. Because the number of units of hidden layers is

difficult to ascertained and the selection method is very subjective, there is no convincing study on it [44]. Its range is extended in positive integer. So, to decide the specific number of units in each hidden layer preliminary experiments should be taken.

Learning rate: Learning rate directly affects the stability and convergence of the network. It is the controlling parameter for weight and bias update values. If the learning rate is too high, the reconstruction error may grow dramatically, and weight may change too much and skip optimal solution. This mean the weight change in each iteration is large then finally the weight will explode and the system may overfit earlier than it was expected. If the learning rate is too low, the reconstruction error may be significantly reduced. The network will stay near local extreme for long time, greatly extending the convergence rate [7, 44]. The value range is between zero and one and the recommended values are 0.05, 0.1 and 0.3 in different literatures [43]. For this thesis it will be determined through simple experiments.

Training epoch: It is the total training steps of the machine learning algorithm using the specified dataset. For each single epoch the RBM will compute contrastive divergence (CD_n) through consecutive Gibbs samplings, then it updates its weight and bias values and finally it calculates the reconstruction error statistics. Its value extends in the range of positive integer and the specific value for this work will be determine using preliminary experiment in Chapter Five.

Batch size: Batch size is the number of training instance per batch. The typical value depends on the training data. Literatures recommend value ranges [10, 100] and 100 will be used as batch size in all experiments in this thesis.

Momentum: Momentum simply adds a fraction of the previous weight update to the current one [44]. The momentum parameter is used to prevent the system from converging to a local minima or saddle point. A high momentum parameter can also help to increase the speed of convergence of the system. However, setting this parameter too high create a risk of overshooting and the minimum which can cause the system to become unstable. A momentum coefficient that is too low cannot avoid local minima, and can also slowdown the training of the system. The value of momentum range in [0, 1]. In this thesis we used momentum value of 0.9 as recommended by [7].

Weight and bias initialization: Weight initialization has been widely recognized as one of the most effective approaches in speeding up the training of machine learning. In fact, it influences not only the speed of convergence, but also the probability of convergence and generalization [13]. Using too small or too large values could speed up the learning, but at same time, it may end up performing worse. In addition, the number of iterations of the training algorithm and the convergence time would vary depending on the initialized values. In neural network weight is used to tune the connection between computing units and between two consecutive layers. Then the algorithm learns the pattern from the training dataset by adjusting its weight parameter through its learning epoch. The bias is used as initialization signal for computing units of their respective layers. Since DBN is stack of RBMs, the first RBM weight will initialized using small random values between 0 and 1, that have zero mean Gaussian distribution with standard deviation of about 0.01 using equation (11) as recommended by [7, 13, 44]. The rest upper RBMs weight will initialized to the transpose of their respective lower RBM weight. The visible and hidden biases will initialize to zero as recommended by [7, 44].

$$\boldsymbol{w = 0.01 * rand\left(NV_{un_i ts}, NH_{un_i ts}\right)} \quad \text{(11)}$$

Where:

- NV_{units} are number visible units
- NH_{units} are number hidden units

Weight decay: Through training epoch weight decay shrinks the weights towards smaller values and this tends to control overfitting of the model [13]. There are two version of weight decay, the first one is absolute value decay (L_1) that push a lot of the weights to be exactly zero while allowing some to grow large and the other one is square value (L_2) which tends to derive all the weights to smaller values. L_2 is used for this thesis work. The recommended value of L_2 by [43, 44] ranges from 0.01 to 0.00001 and 0.00001 is used in this thesis.

2.6.2.2 Greedy Layer-wise Training of Deep Belief Networks

According to [7, 43, 13], DBN is a stack of RBM and can be trained in unsupervised greedy-wise way; which is one layer at a time for pre-training. It is the way of training deep architectures by training each part separately and stacking them to form the full DBN structure. For DBN training greedy-layer-wise training is the way fast training in unsupervised manner for pre-training. By training each layer of DBN in sequential way and feeding lower layer output to upper layer as input and initializing upper

layer weight from transpose of lower layer weight, it is possible to pre-train DBN as shown in Figure 6. Since each layer of DBN is composed of RBM, training each layer of DBN is equivalent to training of respective RBMs. It results better optimization of a network than traditional stochastic descent training. The algorithm of greedy-layer-wise unsupervised training for DBNs with RBMs as the building blocks for each layer can be generalized as follows.

1. Train the first layer as an RBM that models the raw input x = h(0) as its visible layer.
2. Use the first layer to obtain a representation of the input as data for the second layer. This representation can be chosen as being the, mean activations p(h(1) =1/h(0)) or samples of p(h(1)/h(0)).
3. Train the second layer as an RBM taking the transformed data (samples of mean activations) as training examples for the visible layer of that RBM.
4. Iterate steps 2 and 3 for the desired number of layers, each time propagating upward either samples or mean values.
5. Fine-tune all the parameters of the unsupervised DBN network and training by gradient descent on a supervised training criterion.

DBN is unsupervised training but can be applied to labeled data by learning a model that generates both the label and the data [8]. It can be confirmed that to bring a better generalization by initializing a local minimum (or local criterion) that helps to formulate a representation of high-level abstractions of the input to the network. Figure 6 illustrates how greedy-wise-training a DBN algorithm works to extract high level features from low level features.

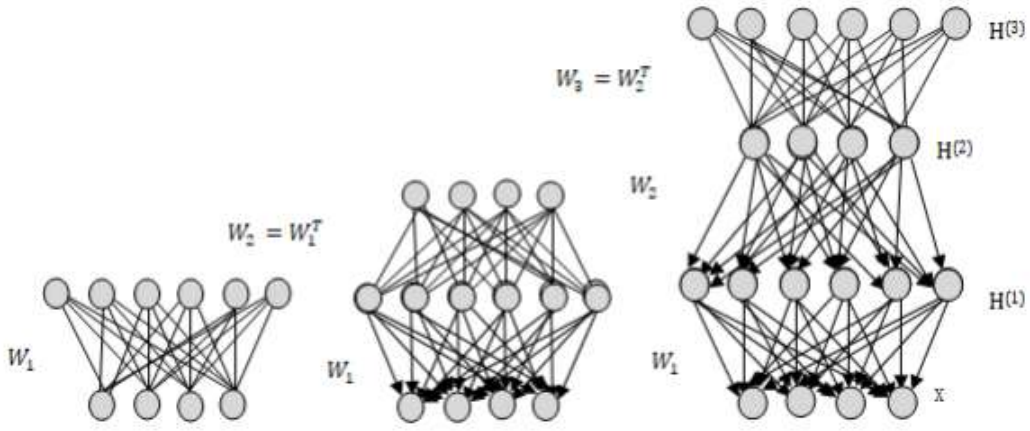

Figure 6: *Greedy Layer-Wise Training Procedure*

2.6.2.3 Supervised Fine-tuning

Supervised fine-tuning is considered as the last training stage in DBN training process. It is possible to fine-tune the parameters of all the layers together for better performance of the system. It uses the same training dataset and network structure with supervised machine learning algorithms. The most usual supervised algorithm which is used for DBN fine-tuning is BPNN.

2.6.3 Backpropagation Training Algorithm for Feedforward Neural Networks

Neural networks are effective tools in the field of pattern classification using training and testing data to build a model. It derives the power for classification due to their massively parallel structure and also ability to learn from experience. It was proposed based on the working principle of natural neural cell and their connection between them to form nerves system [10]. Neural cell is composed of nucleus where electrochemical reaction takes please, dendrites that used to connect to other neural cell axon for synapses input, and axon that expend to connect other neuron cell dendrites as output of synapses. The strength of the axon is a matter of the level of signal received by the next neurons for electro-chemical reaction. Then in this way number of them connects to form nerves system.

Backpropagation algorithm is the method of training multilayer feedforward neural networks using the gradient optimizations method [19]. The basic BPNN consists of three steps. The input pattern is given to the input layer of the network. Then these inputs are propagated to through the network until they reach the output units. This forward pass produces the actual or predicted output pattern. BPNN is a supervised learning algorithm where the desired outputs are given as part of the training vector.

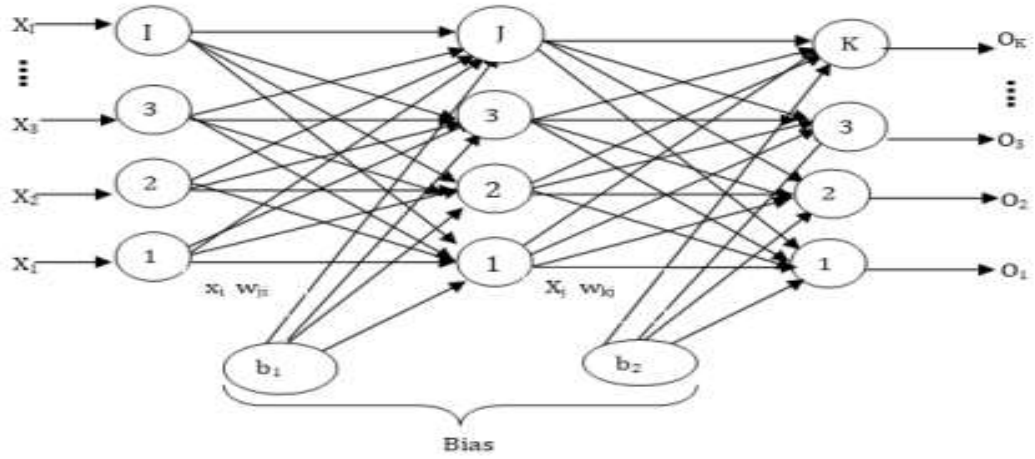

Figure 7: *The architecture of Feedforward Neural Network*

Where:

- x_i is the input signal
- w_{ji} and w_{kj} is the weight matrix between the input layer and the hidden layer and hidden layer and output layer.
- b_1 and b_2 are bias signal for input and output layers respectively.

Feed forward backpropagation has two phases to train the given neural network. The first phase is feed forwarding of the input signal through each layer until output layer. The second phase is back propagating the error between the desired and the resulting output back to each layer until before reaching the input layer. The output of any neuron in respective layer computed using equation (12) in combination with the activation function.

$$\boldsymbol{x_j = \delta\left(\sum_{i:\perp}^{I} x_i w_{j_i} + b_i\right)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (12)$$

Where:

- δ is the activation function.

The activation function is the nonlinear differential mathematical formula to compute the output of the specified neurons output using the weighted input signals. Its final value is bounded between two values. There are number of mathematical formulas that can be used as activation function of neuron which are sigmoid, hyperbolic tangent, etc. and let se sigmoid function. First let's consider a training dataset X input having N features and Y output having M class. Then the sigmoid activation formula and its derivative are described in equation (13) and (14) respectively.

$$\boldsymbol{f(y) = \frac{\perp}{\perp + e^{-y}}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 13$$

$\frac{\partial f(y)}{\partial y} = f(y)\big(1 - f(y)\big)$ 14

$y = \sum_{i=1}^{N} x_i w_i$ 15

Where:

- Xi is the i^{th} feature input
- Wi is the i^{th} feature input connection weight

The result of one layer will feed to the next layer until the output layer and when it reaches the output layer it will be end of feed forward computation and it became the beginning of the next backpropagation process. The basic element of this algorithm is the energy function that defined as a quadratic sum of the difference between the actual output signals and the desired values as defined by equation (16).

$$E = \frac{1}{2} \sum_{i=1}^{p} \sum_{k=1}^{m} \left(y_k^{(i)} - d_k^{(i)} \right)^2 \text{---------- (16)}$$

Where:

- E is the energy function which is the square of the error between the output and the desired signal.
- P is number of training vector.
- M is number of classes or output layer neurons.

So, to reduce the above energy function value through training, BPNN use the following steps.

1. Apply the actual input signal vector X.
   (i) Calculate the output signal under each hidden and output layer using equation (12)
   (ii) Calculate the gradient of activation function in each neuron of each layer using the derivative of the activation function using equation (14).
2. Create the backpropagation network by reversing the direction of signal transmission.
   (i) Replace the activation function by its derivative
   (ii) The input vector at former output layer and the current input layer is the error between the actual and the desired value.
   (iii) The weight modifications proceed on the bases of the result in one feedforward and backward propagation using equation (19).

3. Repeat one and two for all training samples as much time until the stopping criteria of the algorithm is reached.

The weight modification in each training steps can be computed using the equation (17).

$w_{ij}(t+1) = w_{ij}(t) - \varepsilon \nabla E(w)$---------- (17)

Where:

- w_{ij} is the weight of connection from neuron i to j.
- $E(w)$ is the gradient of the energy function.
- ϵ is the training coefficient
- $t+1$ is the next training time and t is the current training time.

The above formula states that the current weight update is the difference between the previous weight and training coefficient multiplied by the gradient of the energy function. The previous weight is obviously known but the gradient of the energy function obtained by differentiating the energy function with respect to respective weight of neurons as in equation (18).

$$\boldsymbol{\nabla E(w) = (o - t) * x_{ji} \frac{df(y)}{dy}}$$---------- (18)

Where:

- $o - t$ is the error of the current training step.
- x_{ji} is the signal of the current connection which is under consideration to modify its weight.
- $\frac{df(y)}{dy}$ is the derivative of the activation function.

The final formula for weight update of backpropagation algorithm is between two neurons of consecutive layer is computed using equation (19).

$$\boldsymbol{w_{ji}(t+1) = w_{ji}(t) - \varepsilon(o - y)^{*} x_{ji} \frac{df(y)}{dy} + \partial \Delta w_{ji}(t)}$$---------- (19)

$$\boldsymbol{\Delta w_{j_i}(t) = w_{ji}(t) - w_{ji}(t-1)}$$---------- (20)

Where

- t is current execution time
- $t+1$ is next execution time and $t-1$ previous execution time
- o is desired output, y is current system output, $\frac{df(y)}{dy}$ is the gradient of the activation function and ∂ is momentum co-efficient.

In this thesis work the input data to the DBN are used as a visible layer (feature vectors extracted from lesions). After low-level RBM learning, the results of the hidden layer are the input of the visible layer of high-level RBM, followed layer by layer. Here, in Figure 8, DBN is used to extract the deep features of lung nodules and provide better features for the BPNN classification. The LND-DBN feature extraction network is made up of three RBMs; the structure of the DBN is as shown in Figure 8.

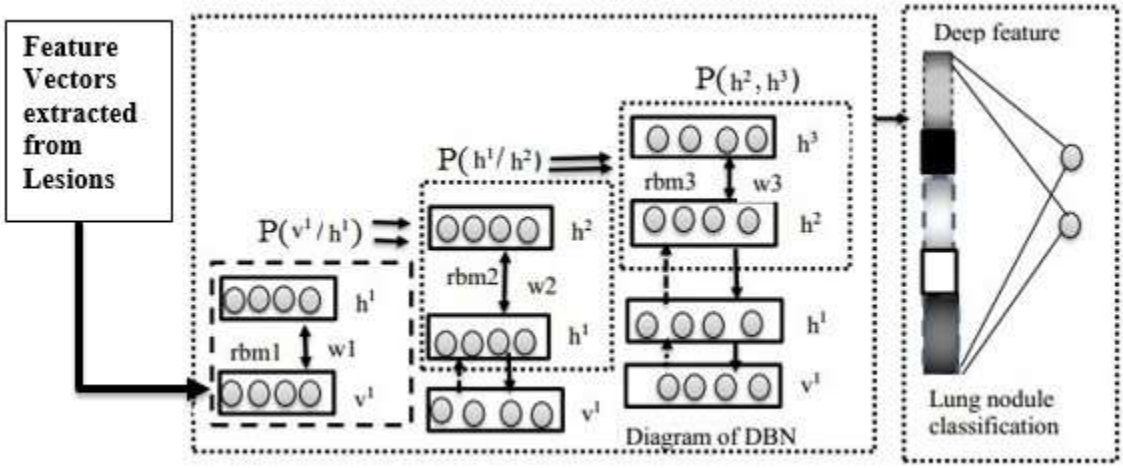

Figure 8: *DBN Feature Extraction and Nodule Classification (Schematic Diagram)*

2.7 Performance metrics and Classification

Cross validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it [45]. But maximizing both the training and testing dataset is the main tradeoff issue, because maximizing training dataset means best modeling and maximizing the testing dataset result the best system validation. In K-fold cross validation, the original sample is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k-1 subsamples are used as training data. The cross-validation process is then repeated k times (folds), with each of the k subsamples used exactly once as the validation data. The k results from the folds can then be averaged (or otherwise combined) to produce a single estimation. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once. We used our model ten-fold cross validation technique.

Finally, after designing and implementing the system there should be some performance evaluation mechanisms. They are the means of measuring the performance of the system based on the output of the system. There are a number of classification task evaluation mechanisms as stated by literatures [1, 46], Based on the number of class in the system output classification system can be grouped in to two groups. Binary classifier with two classes and multi-class classifier with more than two classes. This proposed lung nodule detection system is binary classifier because the dataset contains two classes (benign or malignant).

2.8 *Lung Cancer Screening and Clinical Treatments*

Lung cancer screening is looking for cancer before you have any symptoms that can help to find cancer at its early stage [25]. Screening may provide new hope for early detection and treatment of lung cancer. A severe truth about lung cancer is that it doesn't usually cause symptoms until the cancer is already advanced and difficult to be cured. That is why the idea of screening is taken in any of the people who should want to know him/herself about his/her cancer victim status. Looking for lung cancer in people, even who do not have any symptoms is recommended by expertise. It has the potential of finding the cancer earlier, when it's easier to treat.

If lung cancer is suspected as a result of a screening procedure, then clinical treatments follow-up. There are different clinical treatments mindful by doctors after the screening results for lung cancer victim. Surgery, radiation therapy, and chemotherapy are used in the treatment of lung cancer [25].

After screening procedure, if the nodule in your lung is benign, it may be the result of an infection or irritation. It might also be scar tissue from a previous infection. As a result of examination if the nodule is very small, your doctor may have you to take antibiotics for a few weeks to see if the nodule grows away. Weather you take antibiotics or not, from small nodules your doctor wants to repeat the CT scan in about three months [25]. But if the nodule is malignant, then the patients should discuss with their doctor about their follow-up appointments for treatment and diagnosis. Since nodules can reaper after they have been removed, it is important that patients should have made routine follow-up appointments with their doctors.

2.9 Summary

Lung cancer is a malignant lung tumor characterized by uncontrolled cell growth in tissues of the lung. Lung nodules are a white spot on CT scan of lung tissue and very small deviations on the lung. They commonly indicate an early stage lung cancer. Lung nodules exist only inside the lung region. All lesions inside the lung have the chance to be a nodule or not, thus all of them should be given to the DBN algorithm. As we have seen the dataset, the CT scan includes the whole chest image of the patient and it contains lesions from outside and inside of the lung tissue. Lesions outside the lung increase false positives, since they are not lung nodules. Objects that are outside the lung cannot be considered in our proposed method. Nodules are usually less than or equal to three centimeters in diameter. Any lesions bigger than a nodule is called a mass, which is more straightforward to detect and classify due to its larger size. Image processing techniques such as segmentations of lung and lung artifacts is considered to reduce those false positives. These techniques extract all lesions inside the lung which is very important for DBN algorithm training features.

Detecting nodules early is critical in diagnosing lung cancer to treat it effectively. Nodule detection is the task to find malignant tumors. The nodule detection system reduces the burden for radiologists during lung cancer screening. Several lung nodule detection systems play a huge role for this task. CAD systems assist radiologists in the interpretations of CT images. Developing effective CAD systems for lung nodules have a great impact for the diagnosis of patients and can increase the patient's chance of survival.

DBN and BPNN algorithms are used to train the data for model construction. DBN is a graphical generative model which is trained with a stack of RBMs. An RBM is a model made to learn a probability distribution over the input data. While an RBM may able to learn features from simple input data during training, it is limited in what it can represent. For this reason, RBMs are stacked together in order to form a higher level of representations of the input samples. As for feature extraction, a DBN can learn features that are relevant to the input in that they are able to reconstruct the input using these learned features. Better initializations of the weights with unsupervised pre-training using RBMs allows for fast convergence and generally requires less refinement of the fine-tuning stage. So, in a DBN the pre-training network tries to minimize the reconstruction error on the network, while the fine-tuning network tries to minimize its classification error. Finally, the model is used to classify the nodules.

Chapter Three: Related Works

3.1 Introduction

Automated lung nodule detection system is an active field of research in medical image analysis today [36]. Several systems have been proposed for the detection of lung nodules to improve the performance. Its importance for radiologists in the interpretation of CT images as a second opinion to make decisions makes researchers to be motivated. Researchers select different tools and techniques depending on the problem they attempted. Lung nodule detection procedure varies from dataset to dataset, algorithm to algorithm and so on. Thus, we have tried to analyze and discuss different studies from different aspects of their work, such as methods, algorithms, datasets, limitations, their experimental setups and results.

As discussed in Chapter Two, the main step in the analysis of lung cancer screening is the exact feature extraction of lung nodules, which represent early stage lung cancer victim. Many systems have already been proposed for this task, in order to focus on the detection and classification of lung nodules based on the LIDC-IDRI and private hospital dataset. By discussing those different related studies from different angles, we were trying to observe limitations and gabs. Then we have tried to investigate how to fill those gabs using our proposed methods.

3.2 Detection of Lung Nodules Using Different Approaches

3.2.1 Deep learning-based approaches

The Authors proposed in [1], discuses about automatic detection of pulmonary nodules from CT images. The two primary stages that they have done were feature extraction and classification of nodules. So, they mainly focus to extract valuable features from the input data and detect lung nodules in sub-volumes of CT images. They used a deep CNN algorithm which is trained using backpropagation algorithm for fine tuning. They have done the task based on the publicly available LIDC-IDRI dataset. Their method follows as, first it takes input data from raw CT images, then they perform pixel by pixel convolution to search for a particular pattern from the images, next to that they applied max pooling for dimensionality reduction purpose in their data and finally fully connected layers used for classification of nodules together with the backpropagation algorithm. They used activation functions RELU for the convolutional layers, a threshold activation function for the fully connected layers and a softmax function for the output layer. In the result section they have scored 78.9% sensitivity as 20 FPs per scan or 71.2% sensitivity as 10 FPs per scan. But their result was still minimum according to

the deep CNN algorithm high capability of solving such complex problems. And, that is why, they did it without any false positive reducing techniques such as lung segmentation, segmenting lesions from lung before feeding the input data to the CNN algorithm, which have a potential value to improve detection rate. Even-if it has an advantage, with decreasing the computational time, it has negative impact on its performance also. We believed that, the detection rate of nodules using such a deep learning algorithm gives a promising result together with those different false positive removing and segmentation techniques. So, here we assumed that the performance could be achieved to a better one by adding those techniques in our proposed method.

The researchers in [14], discuses on the detections of lung tumors with the use of 3D CNN based on size and texture methods for lung nodule classification. The general approach of this study was to classify lung nodules from CT scans. They proposed to directly learn the 3D features needed for classification through the use of a convolutional neural networks. Their approach consists of probabilistic, size-based classifier, and a convolutional neural network which operates on a size normalized representation of the nodule. They applied data augmentations on the LIDC dataset to increase their total dataset size, which gives a promising result compared to other approaches including size-based classifiers. They consider both binary and multi-class formulations for the prediction problem and they observed that similar network architectures can be effective for both classifications. Their work was done based on databases of LIDC-IDRI datasets which contains 2, 434 identified nodules with its useful labels. Finally, they score 95 % true positive rate with 0.096 false positives per scan. But their work was limited to only size normalized input data and texture to classify nodules and this misses nodules due to the diffuse nature of the size of nodules and only this two information's alone are insufficient to predict the malignancy of nodules. Beyond to that 3D-CNNs for nodule detection is a natural direction to explore and didn't perform well on the detection task. So, the problem of lung nodule classification is still not well solved and needs further study on this area.

The researchers in [46], presented a methodology using Stacked Autoencoder (SAE). SAE is a deep learning technique for lung nodule classification. Their proposed system uses deep features extracted from an autoencoder to classify lung nodules as benign or malignant. The objective of their CAD system was to assist radiologists by offering a second opinion and making the whole process faster. They have done their research based on the LIDC-IDRI databases, of 4303 instances which consists

4323 nodules. Using the LIDC-IDRI dataset they showed that their proposed method convincingly better results to some state-of-the-art methods on overall accuracy metric as compared from belief decision tries. Finally, they scored an overall accuracy of 75.01% with a sensitivity of 83.35% and they got 0.39 FPs per patient. They scored those results over 10-fold cross-validation.

The Authors in, developed [34], their method for lung nodules detection and classification with deep learning algorithms. They considered three deep learning algorithms: CNN, DBN, and SDAE to classify nodules. In their study these three different deep learning frameworks were implemented and compared the performance on the same dataset. The architecture of their CNN model contains 8 layers including convolutional layer, pooling layer and subsampling layer. They trained this CNN with a batch size of 100 and its learning rate is 1 for 100 epochs. The second-deep learning algorithm they tried was DBNs, which was obtained by training and stacking four layers of RBMs in a greedy fashion. Each layer contains these RBMs stacked together and trained to initialize a feed forward neural network for classification. The third deep learning model they implemented was three-layer SDAE and each autoencoder was stacked on the top of each other. The structure was similar to the DBN model they implemented here. The structure of the SDAE model was 2000-1000-400 hidden neurons in each autoencoder. For both DBN and SDAE the size of the batches was set to 100, and the learning rate was 0.01 for all 100 epochs. In this study they tested the feasibility of using deep structured algorithms in lung cancer image diagnosis. They used 1,018 CT images of the LIDC-IDRI database for the evaluation. Finally, the accuracies of CNN, DBNs, and SDAE were 79.76%, 81.19%, and 79.29% respectively. The highest accuracy they got was 81.19% using DBNs. However, in lung cancer image detection and classification defining the area of ROI is a very important step because the better representation of ROI for lung nodules is one of the most important indicators for malignancy likelihood. This is considered in our proposed work.

3.2.2 Neural Network and other Image processing-based approaches

A research proposed by [47], developed an advanced computerized system for detection of lung nodules by incorporating the virtual dual-energy (VDE) imaging technology for their scheme. The VDE technology suppressed rib and clavicle opacities in chest x-rays (CXRs), while maintaining soft tissue opacity by the use of massive training artificial neural network (MTANN) technique that had been trained with real dual-energy imaging. Their scheme detects nodule candidates on VDE images by the use of a morphologic filtering technique. Sixty morphologic and gray level-based features were

extracted from each candidate, from both original and VDE chest x-rays. To train their CAD scheme they collected 300 cases with nodules and 100 normal cases from six medical institutions by use of screen-film systems, CT systems, and digital radiography systems. They used a publicly available database containing 140 nodules in 140 CXRs and 93 normal CXRs for testing their CAD scheme. Their work focused on the size of nodules from 5 mm to 40 mm. A nonlinear support vector classifier was employed for classification of the nodule candidates. Their original CAD scheme without VDE technology achieves a sensitivity of 78.6% with 5 FP per image. Finally, by the use of the VDE technology more nodules overlapping with ribs or clavicles were detected and they improved the sensitivity to 85% with 5 FPs per image. Therefore, the performance of their CAD scheme, by the use of VDE technology for detection of nodules especially subtle nodules in CXRs improved significantly.

The Authors on [48], employed Fuzzy KNN to classify potential nodules as non-nodule or nodule. They developed their system for early automatic detection and classification of pulmonary nodules from CT images. They were used different stages as follows. First, they segment the lung parenchyma (ROI) from the CT image using a thresholding method. In this step they modify their segmentation by not losing the nodules attached to the lung walls. They obtained the lung section by using adaptive thresholding, low-pass filtering, lung border elimination, and lung reconstruction techniques. Then they applied Gaussian filters for noise reduction and nodule enhancement. Afterwards, they used intensity and volumetric shape index for detecting suspicious nodule candidates that include both nodules and vessels. In addition to that by using the shape index they recognized nodules that are attached to vessels, pleural wall and mediastinal surface. Then features such as sphericity, mean and variance of the gray level, elongation and border variation of potential nodules are extracted to classify detected nodules as malignant or benign groups. Finally, they used Fuzzy KNN classifier with Euclidian distance between normalized feature vectors, for classification of nodules as benign or malignant. The dataset they have been exploited for their proposed method was 63 CT images acquired from the LIDC database. Their proposed method finally achieved a sensitivity of 88% for nodule detection with 10.3 FPs per image.

The researchers proposed in [37], deals about improving the performance of computer-aided detection systems by developing a new CAD scheme for lung nodule detection based on dynamic self-adaptive template matching and Fisher linear discriminant analysis (FLDA) classifier. This proposed method consists of three basic image processing and feature analysis steps. The steps are preprocessing on the

CT scan image, rough detection and false positive reduction. The preprocessing stage consists of lung segmentation using OTSUs algorithm to segment the lung section, removal of critical section and isotropic interpolation. In rough detection stage suspicious regions of interest (ROI) are extracted and filtered by applying 3D dot filtering and thresholding method. Then, pulmonary nodule candidates are roughly detected with 3D dynamic self-adaptive template machining. Finally, they optimally select 11 image features and apply FLDA classifier to reduce false positive detections. They computed their experiments based on two groups of different datasets which were the lung image database consortium (LIDC) and automatic nodule detection 2009 (ANODE09) datasets. By using a 10-fold cross validation experiment their system finally achieved a sensitivity of 90.24% with 4.56 FPs per scan on the LIDC dataset and 84.1% with 5.59 FPs per scan on the ANODE09 dataset.

The researchers in [24], developed a methodology for classifying lung nodules using image processing and pattern recognition techniques. The steps in their method includes image acquisition from LIDC-IDRI dataset, which consists of 833 CT scans selected from 1018 CT scans based on slice thickness, pixel spacing and size of nodules. Then they applied feature extraction based on only shape analysis for classifying lung nodules and non-nodules. After that they computed the feature vectors and pattern recognition was followed based on SVM classifier. The proposed method has been used additional classifiers such as shape diagrams, proposed proportion measurements and cylindrical-based approach including to SVM. So, they basically considered enhancing the parameters (accuracy, Sensitivity, specificity and the ROC curve) for classifying g nodule or non-nodule from the CT images. They were finally achieved a sensitivity of 91.99 %.

The Authors in [20], proposed an automated system for lung nodule classification based on wavelet feature descriptor and support vector machine to design their own CAD system. The stages in their proposed method were the following: Extraction of the region of interest, wavelet transform, feature extraction, and finally classification of nodules as nodule or non-nodule. They pointed out that some CT scans have different shapes and nodule information, because the CT scans were acquired from different scanners. Thus, in order to eliminate the differences between the CT images, they extracted the ROI for each CT image using a supervised extraction method. After this preprocessing stage, they obtained the images from the ROI extraction, and transformed it from the spatial domain to the transformed domain using wavelet transform, used to separate the study region from other organs and

tissues in the CT scan. Then they have been used gray-level co-occurrence matrix (GLCM) to extract the texture information of the lung nodules. Finally, their pattern classification stage was made using a support vector machine (SVM). They used a clinical test dataset which consists of 45 CT images and 61 CT images for the training collected from LIDC and ELCAP databases. In total they implemented on 106 datasets. They classify nodules from 2 to 30 mm in diameter. Finally, they got a result of 90.90 % sensitivity and 73.91 % specificity.

The Authors in [45], discussed automated lung nodule classification based on artificial neuro fuzzy inference system (ANFIS). This study has been conducted based on LIDC-IDRI dataset. They have been used 617 nodules from 151 CT scans with 166 nodules as true positives from LIDC database. Additionally, they used 10 scans of 10 patients with diagnosis information from SPIE-AAPM database for their system. They used the SPIE-AAPM dataset to get better reliability of the system when applied together with the LIDC-IDRI dataset. They were used median and Gaussian filters to remove variety of noises from the images. They were applied iterative or adaptive thresholding to segment the images and also global thresholding was used for some scans failed in the iterative or adaptive thresholding. To enhance the structure of the segmented images, reduce the false detection and avoid missing nodules attached to the lung walls, they employed mathematical morphology functions in their proposed system. After that they extracted geographical and texture-based features of each nodule using GLCM. In their system false positives were detected and removed using ANFIS classifier. Finally, the method identifies nodules 3-30mm in diameter as nodule or non-nodule by using ANFIS algorithm. They scored a final result of 94.44% sensitivity and 85% accuracy with 0.22 false positives per case.

Researchers in [29], perform classification approach for pulmonary nodule detection from CT images. They used morphological features of nodule patterns and ensemble learning. Ensemble learning approaches were used for classification tasks in the detection processes. Their proposed detection system consists of the following steps. They used morphological image processing for feature extraction. The ensemble learning classifiers such as Bagging, AdaBoost and Random Subspace were used in the training and testing case to classify nodule and non-nodule. Feature extraction were done based on geometrical features on the base of morphological shape information and patient information. The dataset used here was obtained from CT images of 103 patients collected from Radiology

Cerrahpassa School of Medicine. All CT images were in size of 512x512 pixels and stored as DICOM format files directly from the CT modality. They achieved a sensitivity of 80.7% on their score line.

The Authors in [17], developed a methodology for improving prior detection and treatment of lung cancer using EK-Mean clustering. First in their preprocessing stage they used median filter and Weiner filter to enhance the nature of the image and to make the component extraction stage more reliable. They remove edges using median filtering while evacuating noise and Weiner filter to figure out a factual estimation of an obscure sign utilizing a related sign as an information and sifting that referred to motion to create the evaluation as a yield. Then they used the K-means clustering unsupervised learning algorithm to segment the region of interest. To the clustered result, EK-mean clustering is applied. Then the features are extracted from the suspicious regions of the lung using GLCM. Features like entropy, correlation, homogeneity, PSNR and SSIM can be used for classification of the tumor images. Finally, they have been used backpropagation neural network for classification. They classified the image as normal image or a tumor image and achieve an accuracy of 90.87% as a result.

The researchers in [49], proposed and implemented lung nodule detection in CT images using ANNs. In their scheme image preprocessing was applied to enhance the image quality. They remove unwanted artifacts and noises by using 3x3 median filter and balance the distribution of pixel value on the image using histogram equalization. Then lung lobes were extracted from the preprocessed images using morphological operations. They used double thresholding method to remove sides and edges of the remaining images and acquired the lung region successfully. They have been used also GLCM for lung nodules feature extraction and then the most appropriate features were selected using the PCA feature reduction method to reduce the dimensionality of their dataset. Finally, they perform the classification of nodules as benign or malignant using ANN trained with BP algorithm. They have done their work based the dataset of 128 CT images from 47 different patients collected by themselves. As a result, they achieved an accuracy of 90.63%, a sensitivity of 92.30% and specificity of 89.47% with their method. They score high result, but they used a very small amount of data and created by themselves, which cannot be compared to the methods that are done with datasets of LIDC databases. They used also artificial feature extraction method which leads missing nodules due to lack of selecting best features.

The Authors in [50] , proposed lung nodule detection using Fuzzy Min-max neural network from CT images. In the first step they used Otsu threshold method and adaptive border marching algorithm for

lung volume segmentation from the chest CT images. Since there are different regions in CT images like thorax, lung and other regions and those regions have different gray levels, Otsu's method was implemented to segment the lung volume in each CT slice. Then they used adaptive border marching technique to correct the segmentation defects. The second step in their work was detecting candidate nodules on the segmented lung volume. Region growing and rule-based method were used to detect candidate nodules. After candidate nodules have been detected 11 features are extracted to reduce FPs. These extracted 11 features are based on shape, size and intensity of candidate nodules. Afterwards, they used the fuzzy min-max neural network classifier together with the compensatory neurons for classification of lung nodules as benign or malignant. Their proposed method was applied based on 19 CT images consisting of 5766 slices collected from private hospitals. Finally, the performance of their proposed system was achieved a sensitivity of 84% for the nodules with 2.6 FPs per scan.

3.3 Summary

Current research works of lung nodule detection system from CT images suggests that there were limited research outputs in the medical image processing fields using deep learning techniques. Based on reviews of this Chapter, different literatures recommend that there is great scope of deep learning-based research in the area of lung cancer detection and medical image analysis. Research contributions reported on CAD systems for detection of abnormalities in CT images still have the problem of accurate lung nodule detection. So far, there is no satisfactory attempt using DBN as a learning machine, which is recent deep learning technique and it is mostly used for solving problems of complex patterns.

Researchers conducted for lung cancer detection based on the conventional approaches depends on the artificial extraction of lung nodules, which are low-level features. So, selecting best features that represent nodules depends on the knowledge and experience of medical experts that affects the overall performance of the detection system. Additionally, some of the works were conducted based on datasets collected by themselves instead of using the international dataset. For example, if a system works well for a very small and selected images by themselves, does not mean that the same system works well for all images in any condition plus to that it cannot handle the problem of overfitting and under-fitting.

Many research works reported in the literature were mainly focused on the training part of deep neural networks with large amount of training data. But current literatures in the detection of lung nodules using different approaches suggests that a promising result will be obtained by reducing false positives via image processing techniques together with deep learning methods [1, 17, 23]. On top of that none of them considered segmentations of the lung tissue only from the chest volume, segmentation of lesions inside the lung and prepare feature vector sets. Not including these techniques, increases false positives that decreases the diagnostic performance of nodule detection system. It lacks also robustness that can deal with all CT images due to large amount of non-nodule lesions present outside the lung. Therefore, this work, tries to fill the gabs mentioned above and to solve the problems pointed out in Section 1.3, by adopting lung nodules detection from CT images using DBN algorithm as discussed in the coming chapter.

Chapter Four: Design of Lung Nodule Detection System

4.1 Introduction

As discussed in Chapter Three, automatic lung nodule detection system had been considered by different researchers with different approaches. In this thesis, therefore, an automatic lung nodule detection system with the design goal that enables to attain accurate measurement in detecting lung nodules is proposed. In the following sections, the details about the techniques and the model developed for the proposed solution are described. In Section 4.2, the proposed system architecture for automatic lung nodule detection is presented. Section 4.3 presents DICOM file acquisition process. Section 4.4 provides how DICOM to JPEG conversion processed. Section 4.5 reveals how input image is preprocessed for further activities. Section 4.6 describes how the lung and lung artifacts (lesions) are segmented. Here, "artifacts" are exchangeable to "all possible lesions" on the lung. In section 4.7 annotated data illustrated. Section 4.8 presents how input vector preparation is done. Section 4.9 describes training and model construction. Section 4.10 shows the way how to classify nodules, and finally, Section 4.11 summarizes this chapter.

4.2 The Proposed Lung Nodule Detection System Architecture

In this section we discussed, the proposed system architecture for automatic detection of lung nodules which is illustrated in Figure 9. The architecture of the proposed method contains different image processing techniques with the integration of state of the art of deep learning techniques, DBN learning machine. Hence, those activities made the proposed method different from the previous attempts.

The system architecture consists of three major stages namely image preparation (ROI segmentation), DBN training, and nodules classification (fine tuning). In the image preparation stage: preprocessing component does the job of removing noise and balance poor illumination. The output of this preprocessing component is fed into segmentation component. The segmentation component is responsible for isolating lung object from the CT scan and isolate objects (or lesions) on the lung from the background. We used three sub components for segmentation namely: Lung Segmentation, Foreground Object Segmentation and Segmentation of Lesions from the Lung.

In this first stage of image processing the last component, is input vector preparation, with a suitable size for all the lesions on the lung and performs the input data usable to the next stage DBN training which provides standardized, reduced and compact format as input vectors used by DBN algorithm.

The second stage is DBN training for the extraction of features of lung nodules as illustrated in Figure 9. In this stage DBN is trained to extract the deep features of lung nodules from the input data using three stacked RBMs and provide better features for BPNN fine tuning stage. DBN takes reduced and standardized input data from the output of the input vector preparation component, that is used as its visual layer at its first RBM. Then the second RBM takes the input from the output of the first RBM and learn features of features automatically. The third RBM takes its input from the second RBM output layer and learn complex features of features, then provide trained features of lung nodules to BPNN algorithm. So, in this stage training DBN, does the job of pre-training layer by layer with a stack of three RBM machines in a greedy fashion. Finally, in the third stage, lung nodules classification is obtained from the BPNN classifier. The result obtained from pre-training stage is fed to BPNN for fine-tuning. This is used to build a model that could be used as a source for classifying malignant vs benign.

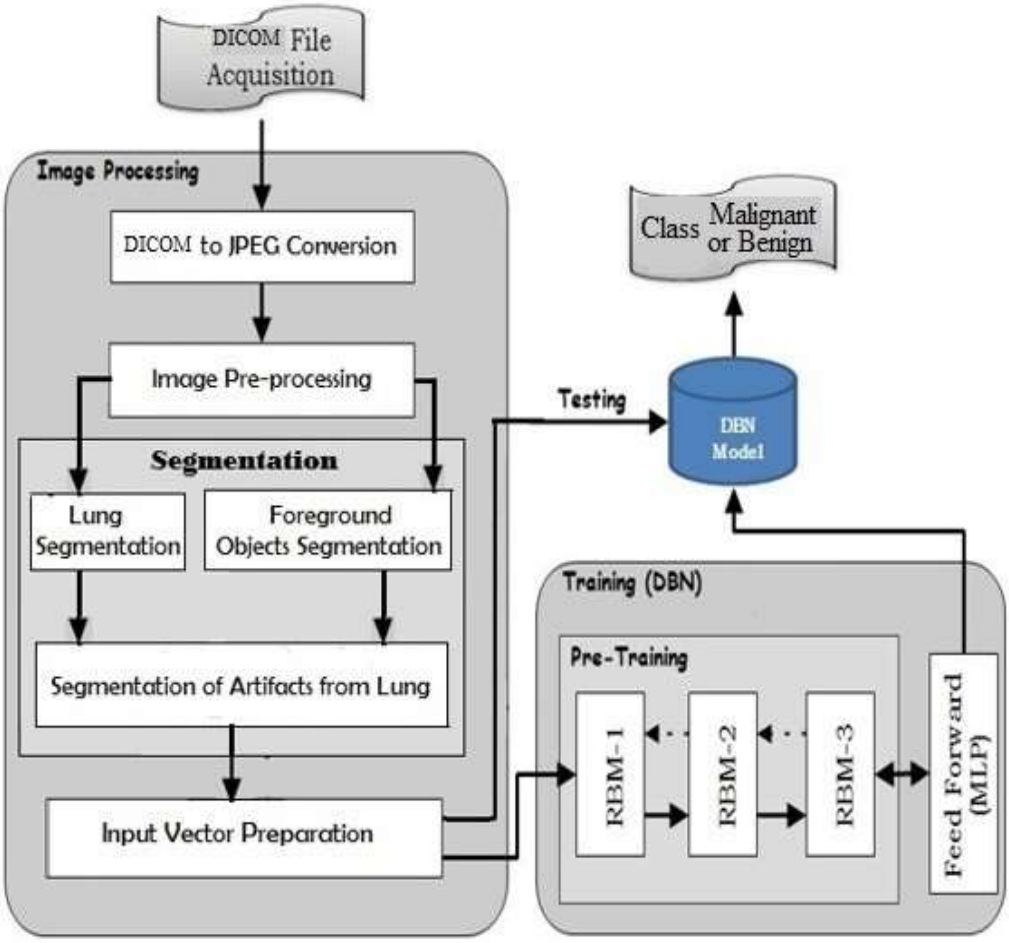

Figure 9: *Architecture of the Proposed Lung Nodule Detection System*

4.3 DICOM File Acquisition

The first step in our proposed method is DICOM file acquisition from the resources of LIDC-IDRI databases. It is the process of acquiring lung CT images. As discussed in Section 2.3, DICOM file is a standard file format or a protocol to interchange medical imaging derived from CT scan images. We got this DICOM file from the publically available databases [21]. The dataset is annotated data and annotations are made by four radiologists which is stored for researchers. It is aimed for the development, training, and evaluation of CAD methods for lung cancer detection and diagnosis. The original dataset is in a DICOM standard as shown in Figure 10.

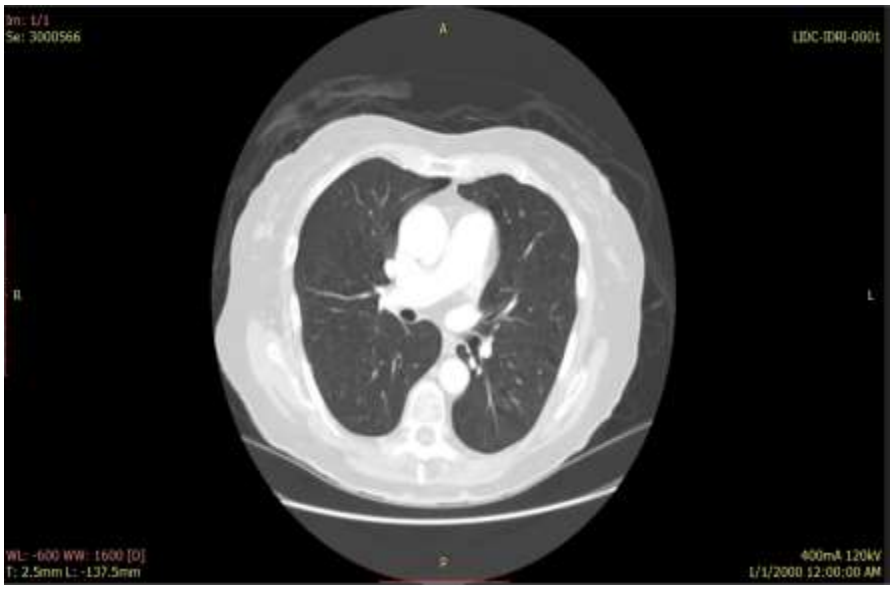

Figure 10: *Original DICOM Image*

4.4 DICOM to JPEG Conversion

As discussed in Chapter Two, LIDC-IDRI datasets are in DICOM format. After we acquired DICOM image data from LIDC-IDRI database, this component is responsible to convert DICOM file into JPEG file format. Algorithm 1 describes the conversion process. We convert into JPEG file format since JPEG is common for easily manipulation in images and the most popular, compatible and high-quality image format. In terms of image quality, image compression during the conversion of DICOM image format into other image format is acceptable in most areas of radiology on different devices [10]. When we convert DICOM image to JPEG image with low compression ratio there is no significant difference in the quality of converted images used for interpretation on CT scanner [39]. Figure 11 shows the result when DICOM image is loaded and converted into JPEG using Matlab function.

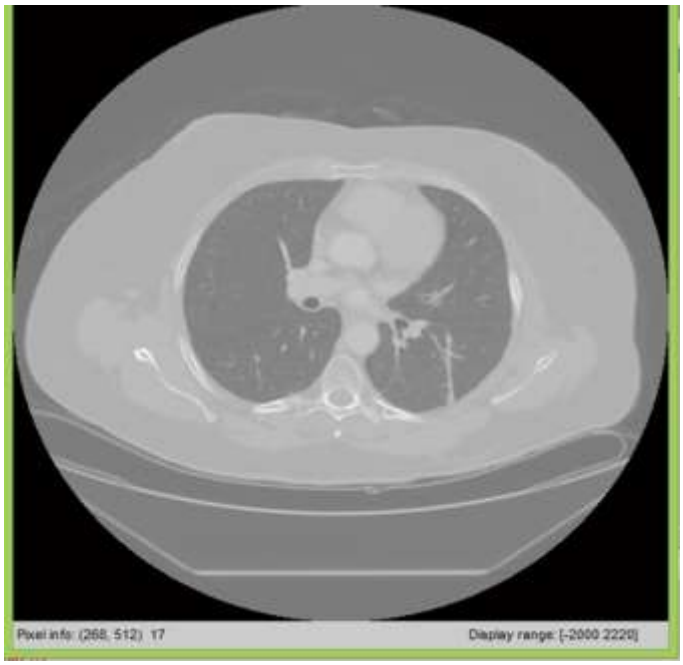

Figure 11: *Converted DICOM to JPEG Format*

```
Input:  A sample DICOM slice: S
Output: A DICOM2JPEG image file
        Load original DICOM file S
        Apply DICOM to jpeg conversion on S
        Return S
```

Algorithm 1: DICOM to JPEG conversion

4.5 Image Preprocessing

The output obtained from DICOM to JPEG conversion component is fed into Image Preprocessing component which is responsible for improving the quality of images taken from DICOM files. Algorithm 2 performs the preprocessing task. We remove noises that can affect the detection of nodules. As we discussed in Section 2.5, different techniques can be applicable to enhance and improve the quality of image. Median filter and Histogram equalizations are the major ones [11, 12]. Median filter is a non-linear filter used to remove high frequency signal noises and salt and pepper noises from images. Histogram equalization is a method in image processing for contrast adjustment using the image's histogram. Since our input data is highly affected by intensity variability and uneven illumination, we apply histogram equalization. It is used for intensity transformations, that changes the

given image distribution to a uniform distribution. We used these techniques to produce high quality DICOM image, which is easily usable for the next component.

```
Input:  A DICOM2JPEG image file: S
Output: A DICOM2JPEG image free of noise
        Get JPEG converted DICOM slice: S
        Normalize S in to double [0 1]
        Apply histogram equalization on S
        Apply median filter on S
        Return S
```

Algorithm 2: Image Preprocessing Algorithm

So, this step, first convert into double conversion and normalize the data between zero and one [0, 1] using equation (1). Then median filter and histogram equalization are applied, result illustrated in Figure 12.

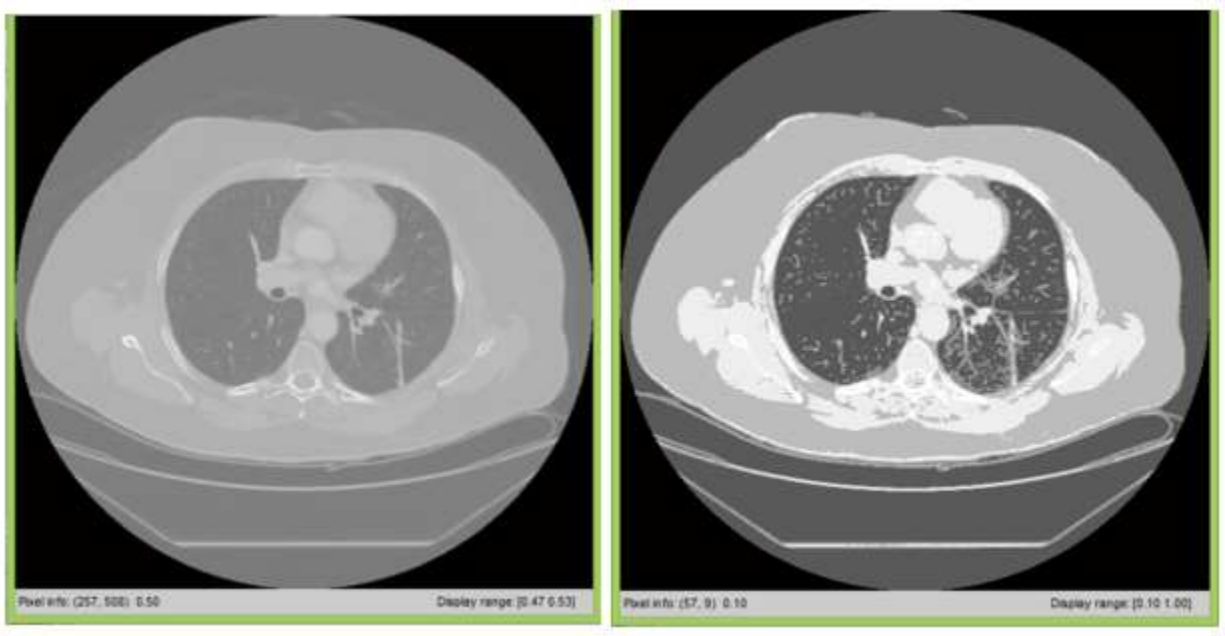

Figure 12: *Preprocessed Gray Images (Normalized & Histogram Equalized)*

4.6 Segmentation

Image segmentation is the process of dividing an image into multiple parts [11, 12]. This is typically used to identify objects or other relevant information from images as noted in Section 2.5. In the

proposed system architecture, the output taken from the preprocessing is fed into the segmentation components. Adaptive thresholding, inverse and intersection operation techniques are employed. Intersection operation is applied between the output of adaptive thresholding and segmentation of the lung section to extract lung artifacts (all lesions inside the lung), which provides us an effective result to reduce false positives for DBN algorithm.

4.6.1 Foreground Objects Segmentation

The result obtained from Image Preprocessing component is fed in to this component which can realize the implementation of adaptive segmentation. In this case the effect of illumination and intensity variability could be handled. This component is responsible for isolating foreground objects from background of the whole image. It helps us to segment detailed artifacts (lesions) which presents on the whole image and can accommodate changing lighting conditions in the image such as shadows, shading, reflections and strong illumination gradient. Our local window size for this thresholding is chosen to be a 32x32 pixel to compute the segmentation and it has been chosen as it renders best result than another window size that we have tried such as 28x28, 36x36. It changes the threshold value dynamically over the image of the window size. One of the previous challenges while segmenting DICOM images taken from CT scan is the occurrence of uneven illumination and intensity variability of slices with in the DICOM file. As a result, false detection increases due to existence of such artifacts. Algorithm 3 performs the process of segmenting the foreground object from the background. Figure 13 shows the result of Foreground Objects Segmentation Component.

Figure 13: *Segmented Image Using Adaptive Thresholding*

```
Input:  A DICOM2JPEG image free of noise: I
Output: Foreground pixels
        For each pixel location J in I
            Get adaptive threshold T1 in [32 X32] window
            If   I(J) < T1
                 I(J)=0;
            Else
                 I(J)=1;
            End
        End
        Return I
```

Algorithm 3: Adaptative Thresholding Algorithm for segmentation

4.6.2 Lung Segmentation

What we need in this work is not the whole CT scans, but only the lung section is our target of interest. So, we have used thresholding followed by intersection, compliment and morphological operations to segment the lung section. Hence, the output that we got from Image Preprocessing component is fed in to Lung Segmentation component. This is responsible to isolate only the lung section from the remaining bodies. As shown in Figure 12, the background color is similar to the lung section, thus it needs large effort to isolate the lung section only. This is very challenging task in our image processing stage since segmentation of similar objects is difficult.

Therefore, we used morphological operations (dilation, erosion and filling) for the purpose of segmenting the lung section from the remaining part of the CT. After we have tried different threshold values that cannot represent the lung successfully we excluded with threshold value of 6000pixels. This threshold removes all connected components (objects) that have fewer than 6000 pixels. We take only area pixels of the lung 6000 and above which represent lung well. This threshold is obtained with preliminary experiments. We segment lung section from each slice of the DICOM file. Together with the above thresholding value, we applied dilation, erosion and filling operations. So, for this work we apply 5 pixels morphological dilation to the boundaries of lung section of the given image, while erosion is applied in order to remove 8 pixels from the boundaries of lung parts. This operation is done based on the rules for dilation and erosion [11, 12]. In case of dilation the value of the output pixel is the maximum value of all the pixels in the input pixel's neighborhood, while in erosion the value of the

output pixel is the minimum value of all the pixels in the input pixel's neighborhood [34]. The process and result of segmenting the lung section is illustrated in Figure 14 and Figure 15.

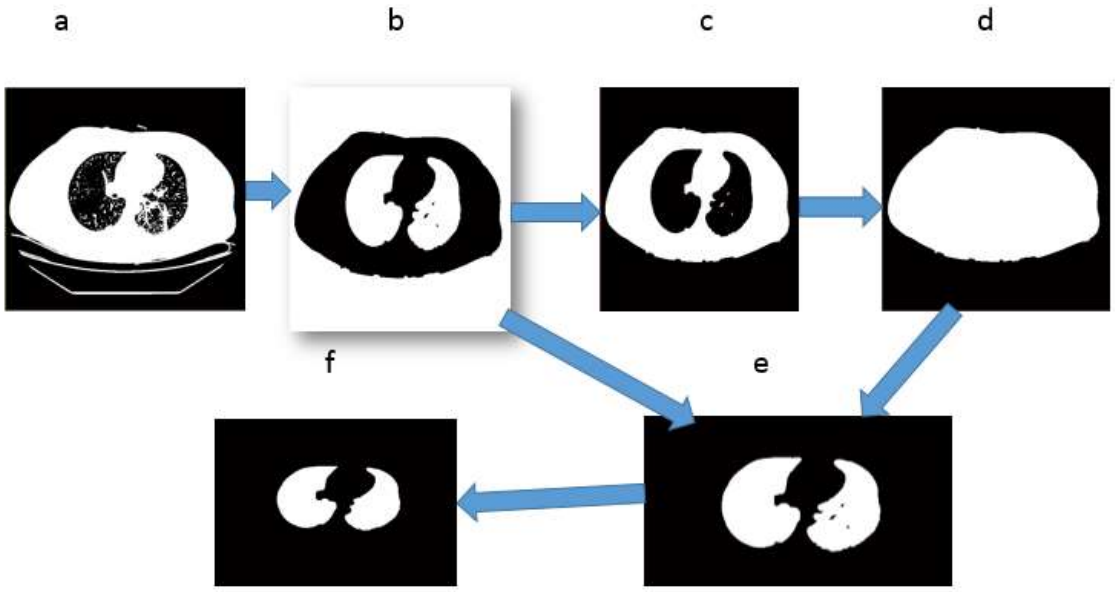

Figure 14: *Lung Tissue Segmentation Process*

The lung segmentation process in Figure 14(a) starts by converting the preprocessed image into binary. Here the converted image consists only 0 and 1, where the background and lung section become black and the remaining is white. Still lung section is black together with the background and considered as unwanted object. Then Figure 14(b), performs first compliment the image obtained in Figure 14(a) and apply a 5-pixel morphological dilation for the purpose of expanding objects by 5 pixels on the complimented white (objects) parts. Figure 14(c) performs first get inverse of Figure 14(b), and apply morphological filling that fills holes in the input binary image. Here a hole is a set of background pixels (black regions) (all 0 values) surrounded by foreground objects. Figure 14(d) is the result of filling operation. The process in Figure 14(e) is segmentation of lung section only by applying intersection operation from the results obtained in Figure 14(b) and Figure 14(d). Then we applied filling operation to fill holes in the lung, as a result we obtained only the lung section as shown in Figure 14(f). The final result of lung segmentation by doing steps from Figure 14(a) to Figure 14(f) in the above, yields segmented lung section as shown in Figure 15.

Figure 15: *Segmented lung Image*

```
Input:  A DICOM-JPEG image free of noise,
Output: Lung Section
        Get preprocessed DICOM2JPEG image: I
        Convert I in to binary: B
        Get inverse Of B: B'
        Apply morphological filling operation on B: M
        Apply 'AND' operation between B' and M: L
        Apply morphological filling Operation on L
        Return L
```

Algorithm 4: Lung segmentation algorithm

4.6.3 Segmentations of Lung Artifacts (Lesions)

Our target area (ROI) is only lesions on the lung section. Here we applied intersection operation over the segmented lung tissue of Figure 15 and the foreground objects shown in Figure 13. This intersection operation is applied on similar objects available on both Figure 13 and Figure 15 with an AND operation. So, by this the most possible lesions to be nodules inside the lung are extracted. Additionally, we have used a threshold value of greater than or equal to 3mm in diameter to segment lung lesions, which are the most probable lung nodules. These threshold value >=3mm (3-30mm) is taken because of non-nodules, nodules <3mm are referred as irrelevant findings [1, 21], and assumed those does not have an impact on the patient. Those nodules (non-nodules, nodules <3mm) are not indicators of lung

cancer as discussed in Chapter Two, since they are too small in size and have high probability to be other infections.

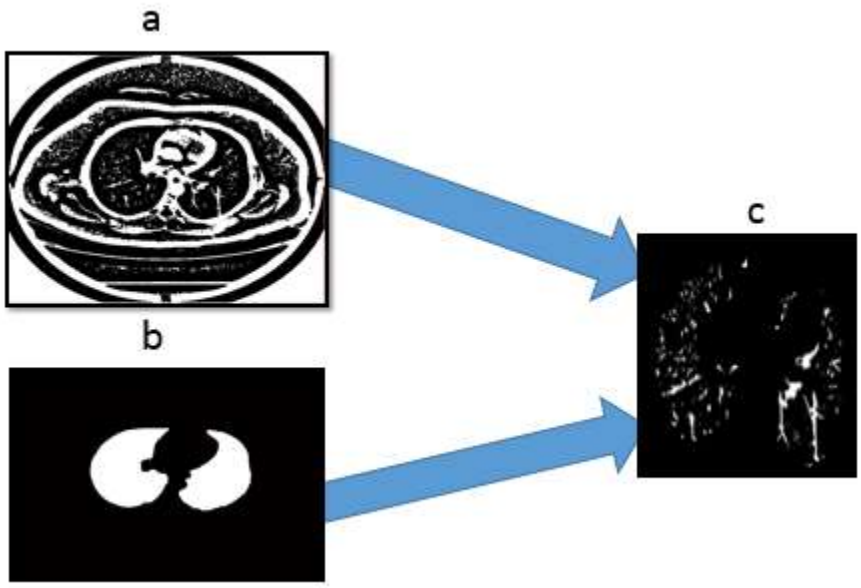

Figure 16: *Segmented Lesions from Lung Using AND Operation*

Here Figure 16 illustrates how lung artifacts or lesions inside the lung are segmented using AND operation. The AND operation is made using Figure 16 (a), segmented foreground objects taken from Figure 13 and Figure 16(b), segmented lung tissue taken from Figure 15. The intersection operation results Figure 16 (c) which shows all segmented lesions inside the lung. In getting lesions, we design and implement Algorithm 5.

```
Input:  Lung Section: S, Foreground pixels: P
Output: Lung artifacts or lesions.
        Prepare empty image: I, size(I)=size(S)
        For each pixel J in S and P
            If   S(J) > 0 and P(J) >0
                 I(J)=1;
            End
        End
        Return I
```

Algorithm 5: Lesion Segmentation Algorithm

4.7 Input Vector Preparation

For the purpose of getting usable dataset for the DBN input, we perform dataset preparation. Input feature vector is prepared by taking the original gray value from the target section (or Lung). This could be done by mapping on the encapsulated binary nodules on the segmented lung artifacts to the corresponding gray nodules on the original DICOM2JPEG converted image. Figure 17 shows the representations of the lesions inside the lung, probable lung nodules.

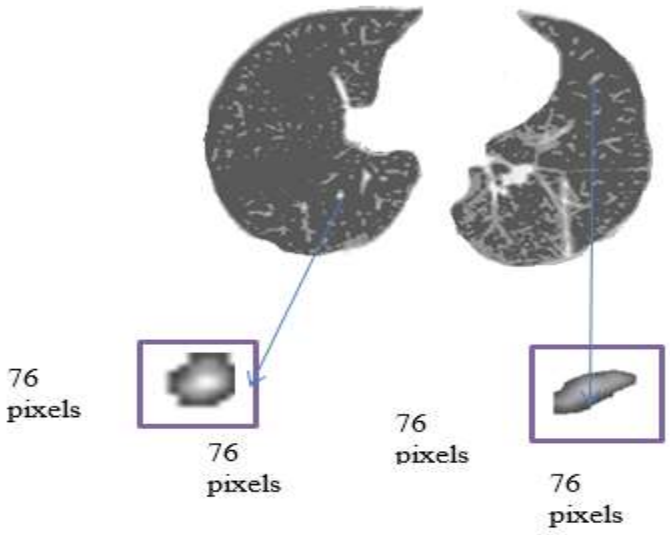

Figure 17: *Input Vector Representation for Lesions*

We represent the input feature vector to the visible layer of the DBN as 76x76 pixel image, which can capable to hold objects or lesions inside the lung. We take this pixel size representation because of the maximum in plane diameter of a lung nodule in the dataset is 76 pixels [1, 22] and the original pixel values of the input data (low-level features) should be fed to the DBN algorithm. We have tried less than this representation but it may lose more information from the lesions. This range holds all lesions or objects such as nodules, blood vessels and other artifacts inside the lung. Then we apply normalization on those pixels of gray values and take the normalized gray values from the JPEG image and put the record of normalized (0, 1) gray values as corpus data.

4.8 *Ground Truth Data*

Data annotations of the LIDC-IDRI database were kept in the form of an XML file. The XML file should be converted into images with a resolution of 512x512pixels. This could be done based on Lampert's tool box [18], which can convert the XML file into its labels of binary image form. The tool is designed for LIDC-IDRI annotated datasets. First the XML file is converted in to slices of 512x512 image and in each DICOM file there is a text file with notepad that automatically generated which says slice-correspondence. The slice-correspondence consists information about original data path, DICOM number, DICOM slices found (slices in order in that specific DICOM file which nodules are found), and Study-Instance ID. Then we prepare an algorithm to map and put this annotation as ground truth data for the validation of our model. Since this notepad information do not give the same information to the original DICOM number, we convert the notepad file from plain text to string file format to get clear knowledge. Then we apply an algorithm to read all the text inside the slice-correspondence and put into an array, by preparing an empty array. So, it reads each string until it gets a space in the notepad and put it in one line of the array using Matlab tools. Because Matlab has built-in functions for manipulating images like to merge, to remove, to separate and so on. Then from the array, we sort all strings with .dcm files in order and map to the original input data. Finally using centroid of every object, we take the value as 0 or 1. The details of this process clearly illustrated below.

4.8.1 Data Annotation Preprocess

In this phase, the given annotated datasets have been prepared to suit our proposed learning algorithm. In this research we used a toolbox proposed in [18]. The toolbox contains functions for converting the LIDC database XML annotation files into images by extracting the readings for each individual marker in the database, and then creates a TIFF image related to each slice of the scan. The process of mapping annotated data to the corresponding vector has been illustrated as shown in Figure 18. The LIDC dataset has been searched recursively for all XML files and the processing will be performed on each. Since the annotated information is formed from this XML file, the LIDC dataset without the presence of XML files are discarded and excluded from the input vector (dataset). If the images and XML files are found in the LIDC dataset, three folders will be created: GTS, images, and masks. Each of these folders will contain folders as shown in Figure 16(A) that are named after the Study Instance ID of the relevant scan (i.e. the first 1.3.6.1.4.1.14519.5.2.1.6279.6001.', which seems to be constant throughout the dataset), for this work, the GTS contents are enough to construct the whole annotation process as the

rest folders contains irrelevant information. Hence, within each content of the GTS folder several folders named slice $_1$, slice $_2$ . . . slice $_n$. where n is the number of slices for which reader annotations were found. And a text file named 'slice_corspondance' are automatically generated as shown in Figure 18(B). Each of the Slice folders contains binary images named as GT_id $_1$, GT_id $_2$, …GT_id $_n$ as shown in Figure18(D). The text file contains detail information each binary files (i.e. slice number, Instance UID (unique for each slice), and the DICOM filename) as shown in Figure 18(C). However, the information about which DICOM slice number corresponds to which ground truth information (in binary, image format) has been only described by the text file. This needs further algorithm to read a row text characters and form a string data type to identify the name of DICOM slice and the corresponding slice folder containing ground truth image as shown in Figure 18(E). Therefore, Algorithm 6 does the process of mapping which DICOM slice file name corresponds to the exact ground truth image.

```
Input:   text file:T
Output: *.dcm, Slice k, DICOM ID

 Get the text file 'Slice_corspondance.
 LineCount=1;INI=0;
 While (EndOf LineCount)
   Read all characters at Line LineCount: C
   Concatenate all characters in C to form string: S
   While (EndOf S)
     Extract string ending with .dcm and push to .DCM[INI+]
     Extract string ending with 'Slice' and push to SLICE[INI+]
     Increment LineCount by 1
   END
 END
 Extract string ending with 'LIDC-IDRI':DICOM ID
```

Algorithm 6: Data Annotation Algorithm

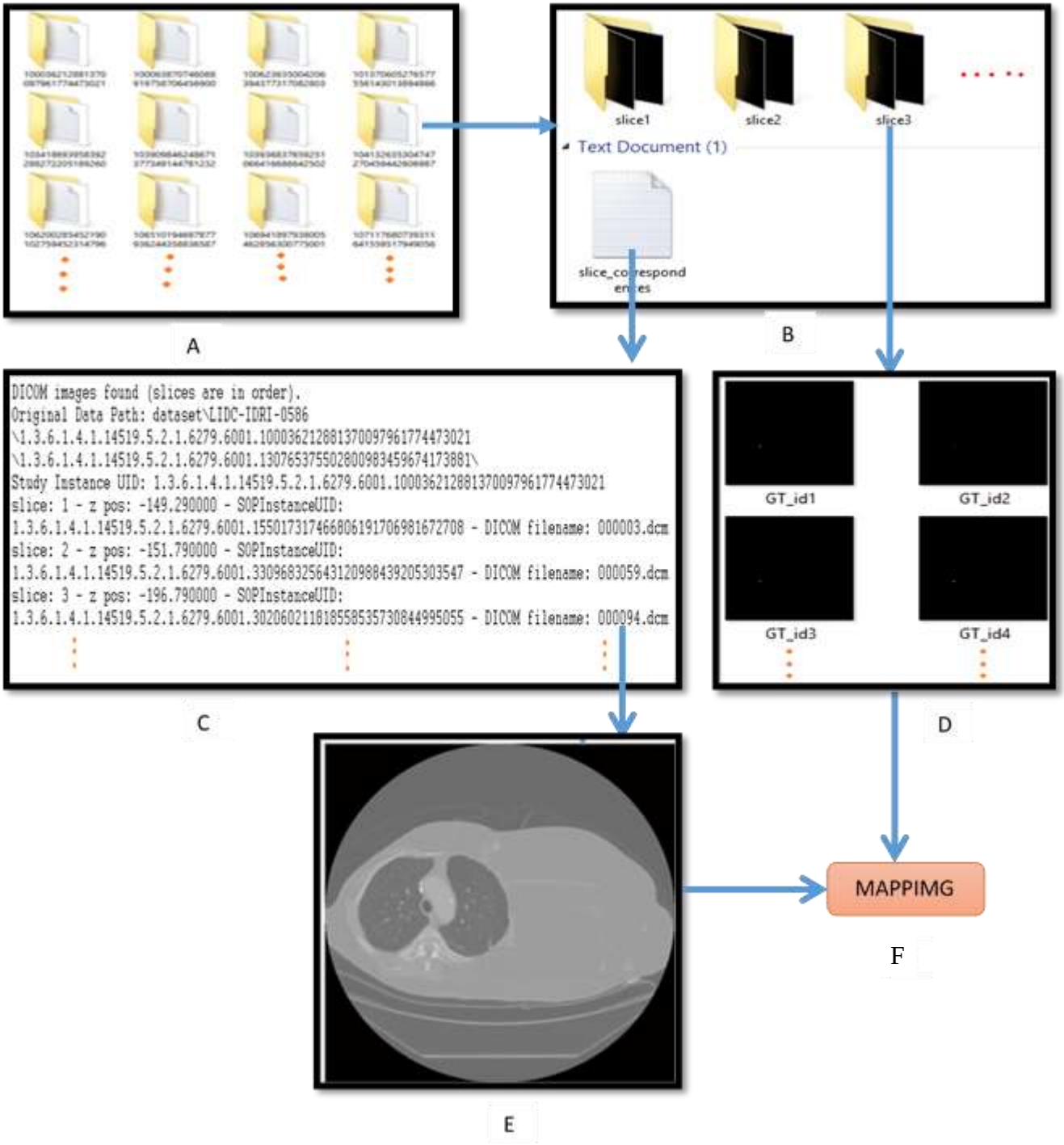

Figure 18: *Data Annotation Pre-Process*

4.8.2 Data annotation Post-process

This is the processes of mapping the ground truth data with the predicted one. We designed and implement Algorithm 7.

```
Input: Lung artifacts (Binary), DICOM_ID, Slice_number
Output: Ground truth Dataset
 To hold input data prepare an array: double InputData[];
 To hold output data prepare an array: int outputData[];
 n=0; //counter
 For each object J in the Lung
   Represent object J by 76 X76 pixel box: P
   Convert P to a row vector and put to dataset: inputData[n]
   Get annotated DICOM file list: Fi from DICOM_ID
   If isFound(Slice_number in Fi)
      Extract centroid of J: C
      For each GT_id image
          If (GT_id (C) ==1) {
            outputData[n]=1
            BREAK
          }ELSE
            outputData [n]=0
          END

      END
   END
 End
Return InputData, outputData
```

Algorithm 7: Data Annotation postprocessing algorithm

4.9 Training and Model Construction

4.9.1 DBN Training

As discussed in Chapter Two, DBN is a multi-layer generative model which try to model the input data through deep hierarchical architecture. The training algorithm that follows to train DBN is Greedy layer-wise unsupervised for pre-training and BPNN is used for fine tuning. We used DBN to extract features of lung nodules from CT images and construct a model for classification. In this research, RBM with binary units are used to construct the DBN network structure. This DBN network consists of three RBMs, in which we got better result that can improve the performance of the algorithm. The

layers are labeled as RBM_1, RBM_2, and RBM_3. For training these RBMs, we first randomly initialize the units and parameters. Then, there are two phases in contrastive divergence algorithm the positive phase and the negative phase. During the positive phase, the binary states of the hidden units are determined by calculating the probabilities of weights and visible units. Since it is increasing the probability of training data, it is called positive phase. On the other hand, the negative phase decreases the probability of samples generated by the model. A complete positive-negative phase is considered as one epoch and the error between generated samples by the model and actual data vector is calculated at the end of the iteration. Finally, weights are updated by taking the derivative of the probability of visible units with respect to weights, which is the expectation of the difference between positive phase contribution and negative phase contribution.

If we train the whole DBN at a time without a greedy-layer-wise strategy, its many layers will lead to the low efficiency of learning [7, 43]. Through layer-wise unsupervised learning of the DBN we extract the features of lesions in unsupervised way from the input data. The extracted features of one layer become the input to the next layer. Then the algorithm learns to extract the deep features of features of the input data by this greedy-layer-wise learning approach. Each RBM layer is pre-trained for 100 epochs. First, we train the first RBM by inputting the original data. Then we use this output as input of the second RBM and the rest can be done in the same way. Unsupervised pre-training method is used to initialize the hidden layer weights while building deeper model. The trained RBM will be used to construct the pre-trained layer of DBN. Starting from the first up to the last layer they will be modeled as generative RBM but the last layer of DBN with the classifier layer will be modeled as discriminative RBM. The training process will continue from the input layer up to the last layer in greedy layer-wise manner for unsupervised pre-training of DBN.

Our method uses the structure of 5776-500-500-2000-2 (meaning 500 units for the first and second hidden layer and 2000 units for the third hidden layer) for lung nodule detection. The number of computing units in the input layer is equal to the number of input pixel sizes in the dataset and also in the case of output layer it is equal to the number of classes in the dataset. The structure of this DBN algorithm was selected based on experiments as discussed in Chapter Five. The number of computing units varies from layer to layer.

4.9.2 Model Construction

DBN model is constructed from RBM training and BPNN fine tuning. As we can see in Figure 19, the constructed model takes 5776 input data as an input neuron. These input neurons are classified in three consecutive RBMs. The first and the second RBM outputs 500 neurons from 5776 input neurons. The third RBM gives an output neuron of 2000 for the last layer, which is the classifier. Finally, the classifier classifies the input layer in to two neurons (malignant or benign lesions). The model is tested using tenfold cross validation technique.

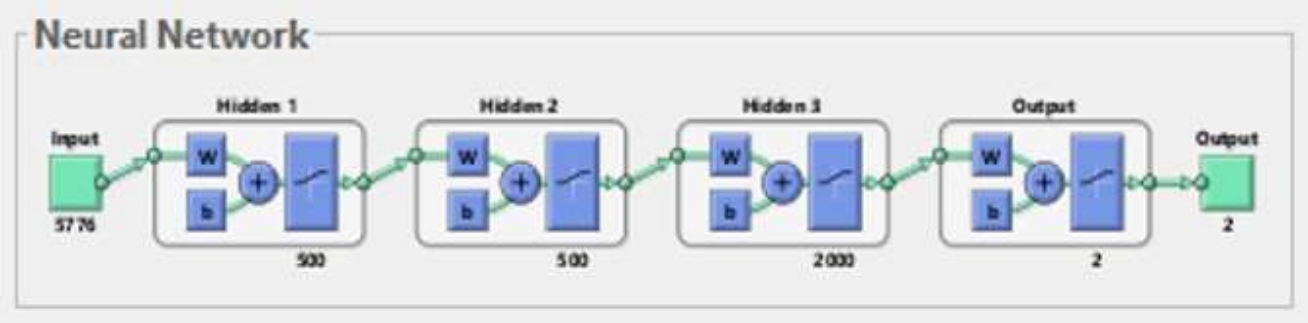

Figure 19: Constructed DBN model

4.10 Classification

Finally, after classifier is constructed we used unseen test data to evaluate the model performance. To evaluate the performance of this proposed DBN algorithm, state of the art neural network algorithm, feedforward neural network [36], was applied to the same classification problem using the same input data and target output. It is trained with backpropagation algorithm.

We have used 10-fold cross-validation technique to test the performance of this designed model. Cross-validation method is a technique that divides data randomly into k equal size subsamples [25]. In 10-fold cross-validation, the original sample is randomly partitioned into 10 equal size subsamples. Of the 10 subsamples, a single subsample is retained as the validation data for testing the model, and the remaining 9 subsamples are used as training data. The cross-validation process is then repeated 10 times, with each of the 10 subsamples used exactly once as the validation data. Then we averaged all 10 results from this fold to produce a single estimation of our model. The advantage of using this method is that all observations are used for both training and validation, and each observation is used

for validation exactly once. The overall overview of this proposed work actions and tasks are summarized below in Figure 20.

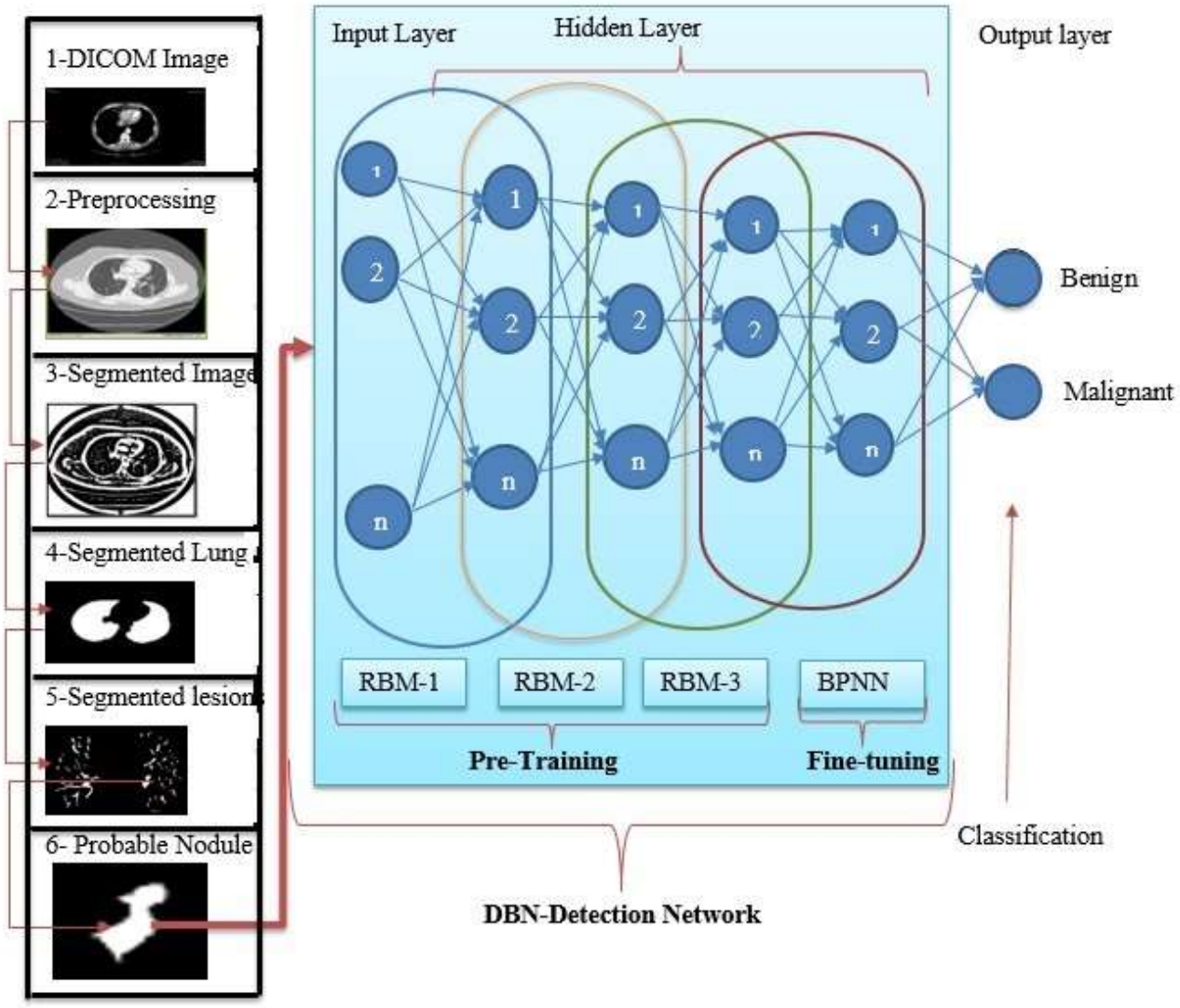

Figure 20: *The General Overview of Our LND-DBN Algorithm*

4.11 Summary

The chapter systematically went through the design of Lung Nodule Detection System via DBN. The designed system has components working together. In the Image processing components, DICOM files are acquired and converted into JPEG image. The image is preprocessed to reduce noise and false positives. This increases the performance of DBN in classifying nodules into malignant or benign. In the segmentation component, lung region is segmented. Lung lesions are also segmented from the lung using adaptive thresholding and intersection operation. The segmented lesions are prepared as input feature vector sets. The input vector or data to DBN is represented with 76x76 pixel images and inputted to RBM_1.

In the training components, the input vectors or datasets are trained with three consecutive RBMs (pre-training components) and the last training component which is a feedforward with BPNN algorithm. The pre-training component takes 5776 input neurons from which 2000 output neurons are feed into the feed forward. The feedforward taking 2000 feedforward as an input classifies the lesions into two classes malignant or benign. From the training component we construct a DBN model. The model is tested using tenfold cross validation.

Chapter Five: Experimental Results and Discussions

5.1 Introduction

This chapter describes the implementation details and experimental result of the proposed design for Lung Nodule Detection system. A comprehensive set of experiments was performed to verify the performance of the proposed approach. Section 5.2 presents the datasets used in training and testing the system. Section 5.3 describes the implementations of the proposed system. Section 5.4 describes evaluation methods we used to evaluate our proposed approach. Section 5.5 presents the test results found. Finally, discussions are made in the last section of this chapter.

5.2 Datasets

In order to test the proposed design for detection of Lung Nodules, we used an international dataset from LIDC_IDRI [21]. The whole dataset of this international source is collected from 1,010 patients which consists 1,018 DICOM files. To demonstrate our work, we used 201 DICOM files of 36,520 slices due to machine constraints from the total of 1,018 DICOM files of 244,527 images (or slices) from the LIDC-IDRI dataset. The dataset we used in this work is illustrated in Table 1.

Table 1: *Summary of the Dataset*

Data type, Modality	Resolution	Used DICOM file characteristics			Input vector size To DBN
DICOM, CT	512x512pxls	Total No. of DICOM	Avg. No of Slices per DICOM	Total No. of images/slices	76x76x221,807
		201	240	36,520	

Experiments have been performed using LIDC-IDRI dataset created by the collaborations of seven academic centers and eight medical imaging companies [21, 25]. This LIDC-IDRI database is a publicly available data for the medical imaging research community. It is web-accessible international resource database for development, training, and evaluation of lung nodule detection systems. The dataset is generated to support the development of computer-aided diagnostic methods for lung nodule detection and identification. The dataset contains 1018 CT images that originated from a total of 1010

patients. The data is available in both MHD and DICOM file formats. We have used the DICOM file format. Annotations are provided in XML format. Each folder includes DICOM images from a clinical CT scan and associated XML file. The XML file records the results of a two-phase annotation process. In LIDC-IDRI dataset, the images are formatted as [65, 764] x512x512 voxels per scan, where [65, 764] is the range of values for the number of slices in the 3-D images, and 512x512 is the in-plane pixel resolution of each of the 2-D slices. The average number of slices per scan in the dataset is 240, whereas the minimum is 65 and maximum is 764 [1, 37, 29]. The LIDC-IDRI data contains annotations which were collected during a two-phase annotation process. Each scan has been examined by four experienced radiologists in a two-phase image annotation process. In the first phase each radiologist independently analyzed each CT scan and marked lesions to one of three categories (nodule>=3mm, nodule<3mm, and non-nodule>=3mm). In the second phase each radiologist independently analyzed their own marks along with the anonymized marks of the three other radiologists to render a final opinion. This reference standard consists of all nodules >= 3mm accepted by at least 3 out of 4 radiologists [21, 25]. Annotations that are not included in this reference standard (non-nodules, nodules <3mm, and nodules annotated by only 1or 2 radiologists) are referred as irrelevant findings. As shown in Figure 21, the images have the whole chest CT image which needs further process to eliminate the area outside the lung section. We have used annotations of nodules >= 3mm in diameter for this current work. We have used ten-fold cross validation for testing our model.

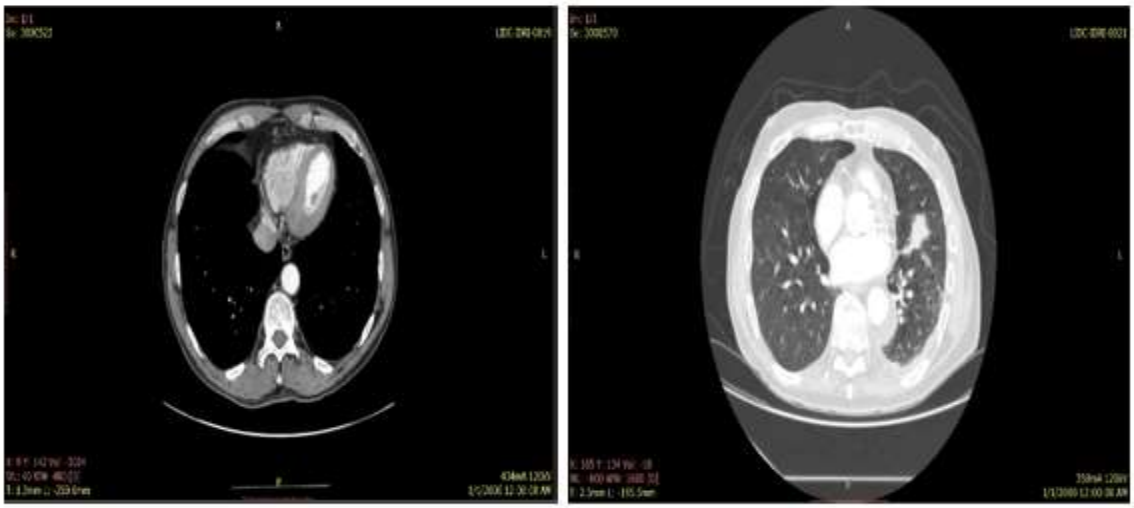

Figure 21: *Sample Data from LIDC-IDRI Datasets*

5.3 Implementations

The prototype is implemented using MATLAB. The prototype takes a DICOM as an input using the graphical user interface as shown in Figure 22, which describes detection of lung nodules from lung section, when applying DBN algorithm. The designed system is implemented using HP DELL OPtiplex GX core-i7 3.6 GHZ, 16GB RAM.

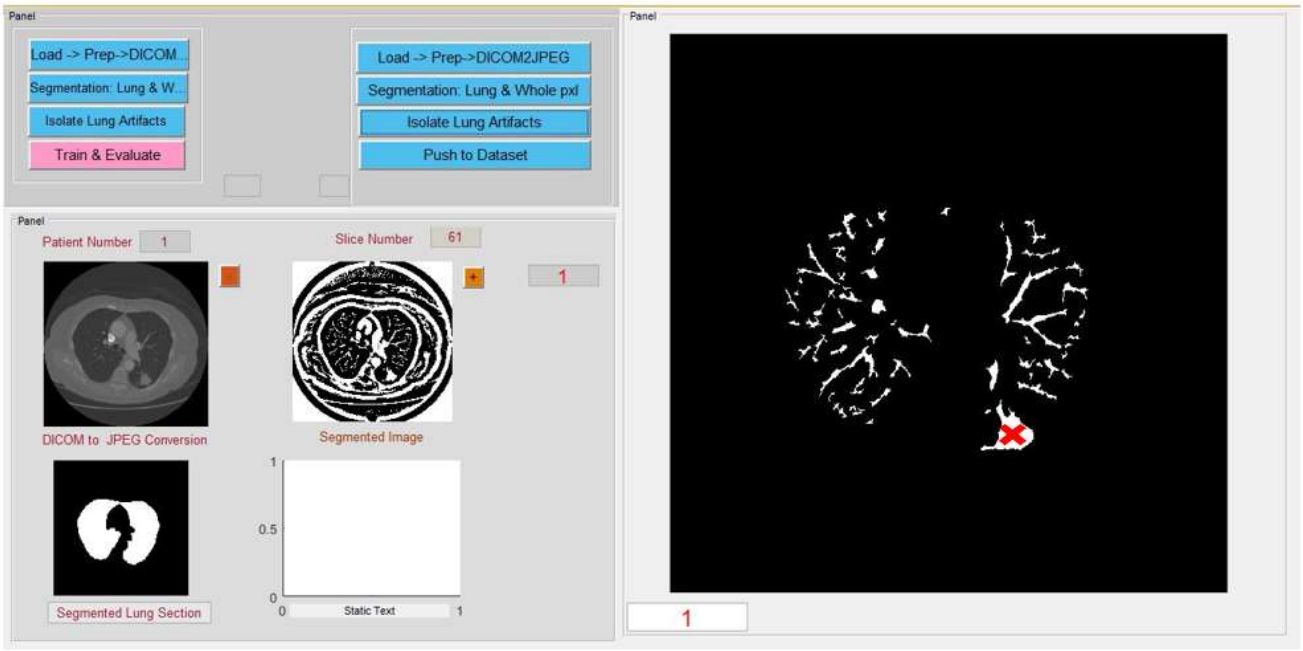

Figure 22: *Running Prototype*

5.4 Evaluation Method

We have used a confusion matrix for the detailed breakdown of correct and incorrect classifications for each class in our results. It is a table that is often used to describe the performance of the classification model on a set of test data for which the true values are known [37]. At first, we set the parameters of the properties in which malignant lung nodules are positive (P) and benign lung nodules are negative (N). We define the number of false positives (FP) to be the number of nodule predictions made by our system that do not contain any annotated lung nodule. Also, we define the true positive (TP) value to be the number of lung nodules that have been successfully detected by our system, and the false negative (FN) values to be the number of lung nodules that have not been detected by our system. The

true negative (TN) values to be the number of benign lung nodules that are predicted as benign nodules. Table 2 shows representations of confusion matrix in this experimental result.

Table 2 : *Confusion Matrix*

Confusion Matrix		Conditions	
		P	N
Test Results	P	TP	FP
	N	FN	TN

We then define four types of evaluation metrics: accuracy, sensitivity, specificity and precision. Accuracy is defined as the ratio of correctly detected samples and full samples. This metrics describes how the model (classifier) is correct and can be calculated based equation (21).

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}} \quad \text{.....} \quad (21)$$

Given the TP and FN values the sensitivity (true positive rate) of the system can be computed based equation (22).

$$\boldsymbol{Sensitivity} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \quad \text{.....} \quad (22)$$

Specificity is the ratio of the correctly detected benign nodules and all benign nodules, as equation (23) reflecting the rate of misdiagnosis.

$$\boldsymbol{Specifity} = \frac{\boldsymbol{TN}}{\boldsymbol{TN+FP}} \quad \text{.....} \quad (23)$$

Precision is defined as the number of true positives (TP) over the number of true positives plus the number of false positives (FP) which is defined by equation (24).

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \quad \text{.....} \quad (24)$$

Monitoring training progress: There should be some quantitate mechanism to monitor the progress of the system training to see how it is going through training process. For RBM training the reconstruction error is used to monitor the training progress as recommended by [7, 44]. It is difference between the reconstructions and the original data. For backpropagation mean square error between the expected and the predicted value will be used [40]. Training time: It is the total time to train the system using the training dataset.

To evaluate the detection performance more objectively, it is more popular to quantitively analyze the classification using receiver operating characteristics curve (ROC) and area under the ROC curve (AUC) analysis [3]. It is used to visualize the tradeoffs between sensitivity and specificity in a binary classifier. ROC curve describes the performance of a model across the entire range of classification threshold [36]. It is an excellent tool for assessing class separation such as nodule classification. It shows the relationship between sensitivity and specificity values (or the true positive rate versus false positive rate). Computating the AUC is one way to summarize the ROC in a single value. The AUC measures how well the ability of the classifier can distinguish between benign and malignant groups. The value of AUC is usually between 0.5 and 1.0, where 0.5 denotes a bad classifier and 1 denotes an excellent classifier [26]. The closer the curve is to the upper left, the larger the AUC value is, indicating that the system performance is better.

5.5 Test Results

As we have depicted results in Table 3, in testing the model we used tenfold cross validation technique. In 10-fold cross validation techniques, each fold is used as training and testing in turn and make an average of each folds as discussed in Section 2.9. Average values of accuracy, specificity, sensitivity, precision and error rate is computed for the performance of the proposed method. However, before conducting the testing activity we set the following parameters settings in order to achieve better classification result.

5.5.1 Experimental Setups and Discussion

Specifying some of the system parameters and their values scientifically based on experimental procedure is critical to the designed system model performance and has a great relationship with network parameter selection. In this proposed method different parameters were tested, and the model with smallest error rate was selected. Where the number of hidden layers is 3, the network structure is set to 5776-500-500-2000-2, the learning rate is 0.1, the size of the mini batch data is set to 100, and the number of epochs is 100. Here three parameters, namely RBM training epoch, RBM learning rate, number of computing units in RBM hidden layers, are taken in to consideration with experiments. In those experiments while testing the specified parameters, the other parameter values are set to either previously explored value or random value is selected from acceptable value range for temporary use until experimenting on it.

5.5.1.1 Number hidden layers used in this DBN training

In our experiment we tried to observe reconstruction errors in deciding the number of hidden layers to be used to achieve better performance. We observed that, the reconstruction errors decrease significantly as the number of hidden layers is increased from two to three as shown in Figure 23. When we increase the number of hidden layers from three to four, the reconstruction error slightly decreases. However, it is affected by large number of overfittings which has a negative impact on the overall classification performance of the algorithm (DBN). As a result, in this study we decided the number of hidden layers to be three.

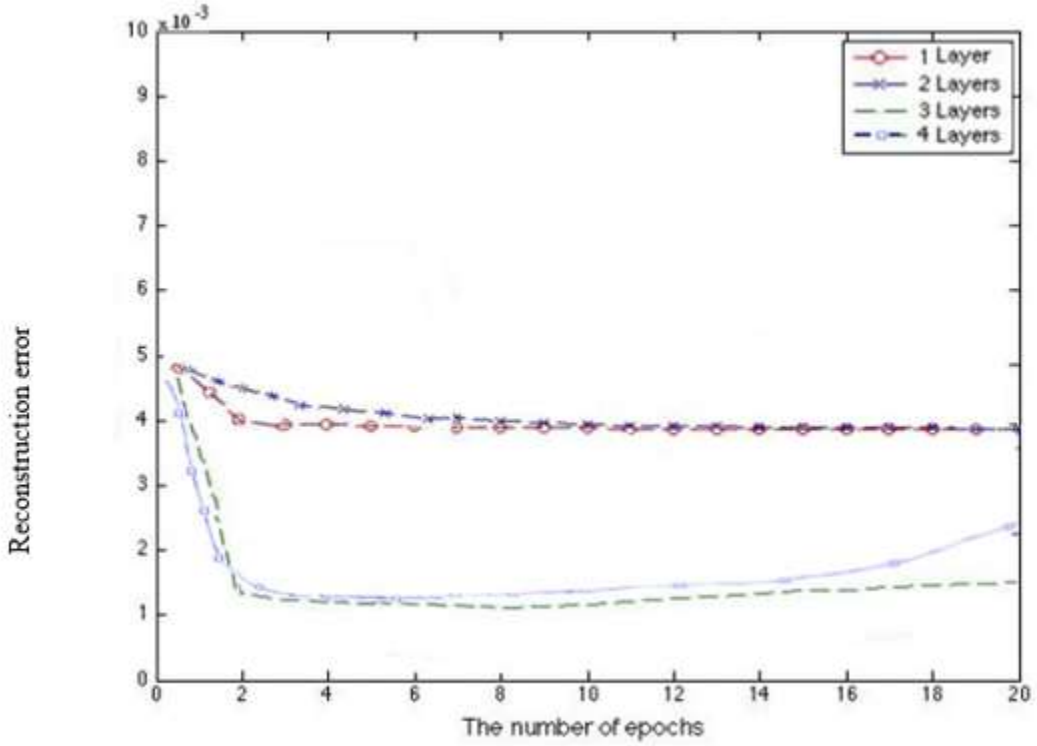

Figure 23: The effect of different number hidden layers

5.5.1.2 Number of Computing Units in DBN Hidden Layers

The number of hidden units in each layer corresponds to the features of input data stored in the system. To decide the number of computing units in the hidden layer of DBN the following experiment is undertaken. Since it is not possible to experiment on all possible positive integer computing units first it is wise to guess the most probable range of values in initiative way. The different number of hidden

units of 3-layer DBNs, 500-500-1000, 500-500-2000, 500-1000-2000, and 1000-1000-2000 are trained and the best number of hidden units in each hidden layer is acquired in Figure 24.

As shown in Figure 24, 1000-1000-2000 DBN has the highest error and its performance does not improve much until 20 iterations of epochs. The errors for DBN with 500-1000-2000 hidden units are almost identical with 1000-1000-2000 structure. DBN with 500-500-1000 hidden units, has faster training time since there are less hidden units than to train 500-500-2000 DBN. But in our system, 500-500-2000 DBN is selected because it is the best setting for the classification as it has faster training time and the least error.

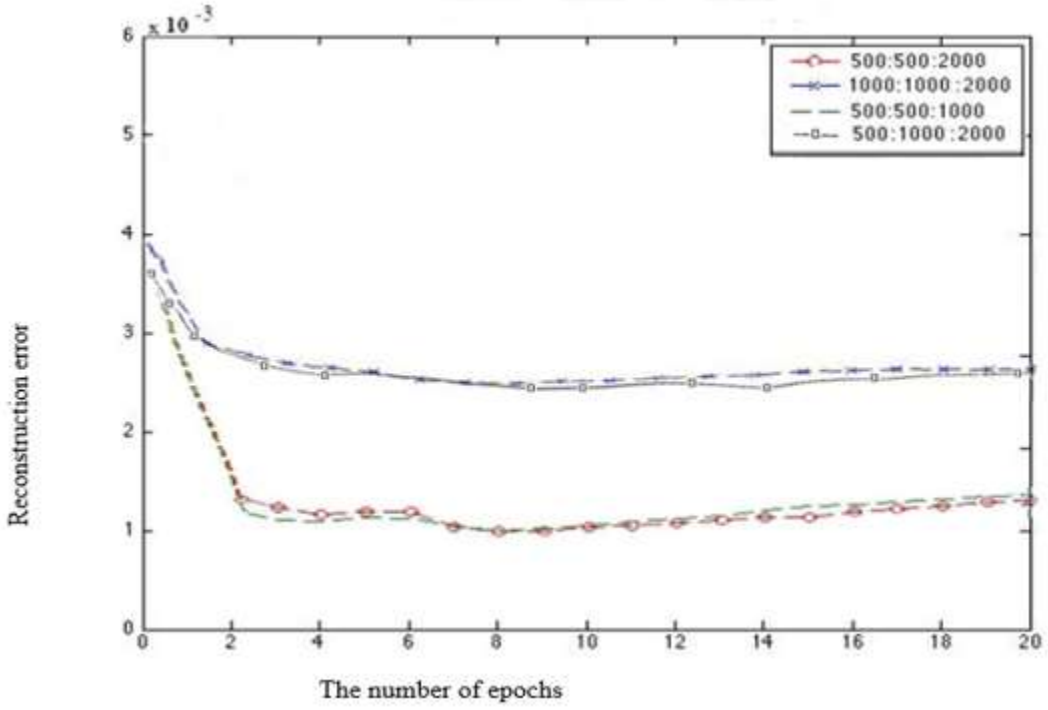

Figure 24: The effect of different number computing units

5.5.1.3 Number of Epoch for RBM Training

To decide on the number of training iteration of RBM the following experiment is under taken. The range of epoch for this experiment is between 1 up to 200. So, the experiment is running on a single RBM for 200 epochs and the reconstruction error is determined as shown in Figure 25. It shows that the reconstruction error is dropping smoothly up to around 100 training iterations, then after that it continuous constant and even it increases a little bit which shows that the occurrences of overfitting.

So, based on the above result 100 is selected as the number of epochs for model construction in this study.

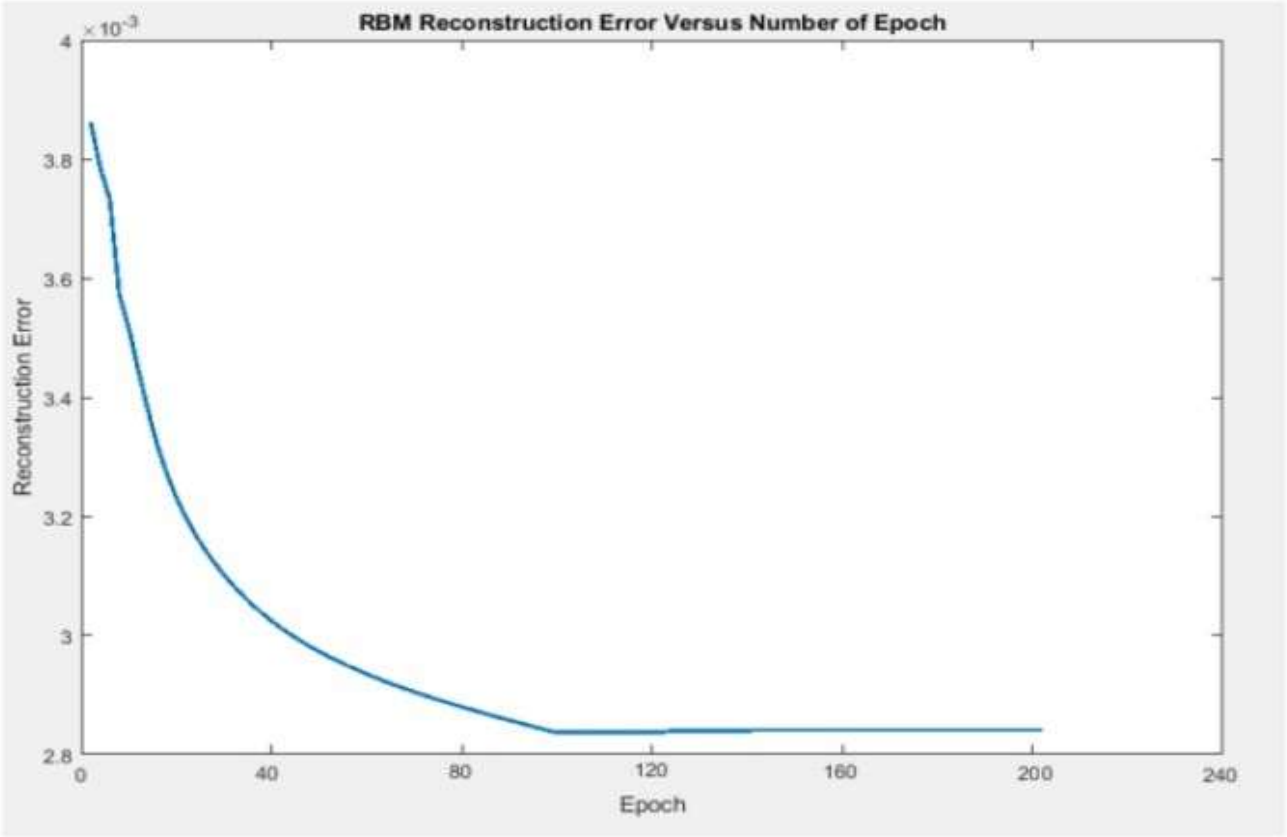

Figure 25: *The Effect of Number of Epochs on the Training Performance of the System*

5.5.1.4 Learning Rate for RBM Training

To decide the learning rate of RBM the following experiment was undertaken. The learning rate range for this experiment is between 0.05, 0.1, and 0.3 as shown in Figure 26. For each value in the specified range the RBM is trained and its reconstruction errors are recorded. It shows that RBM trained with 0.1 learning rate results smaller reconstruction error than learning rate with 0.05 and 0.3. Since small reconstruction error shows that the fast convergence and better model building process of the system 0.1 is selected as learning rate.

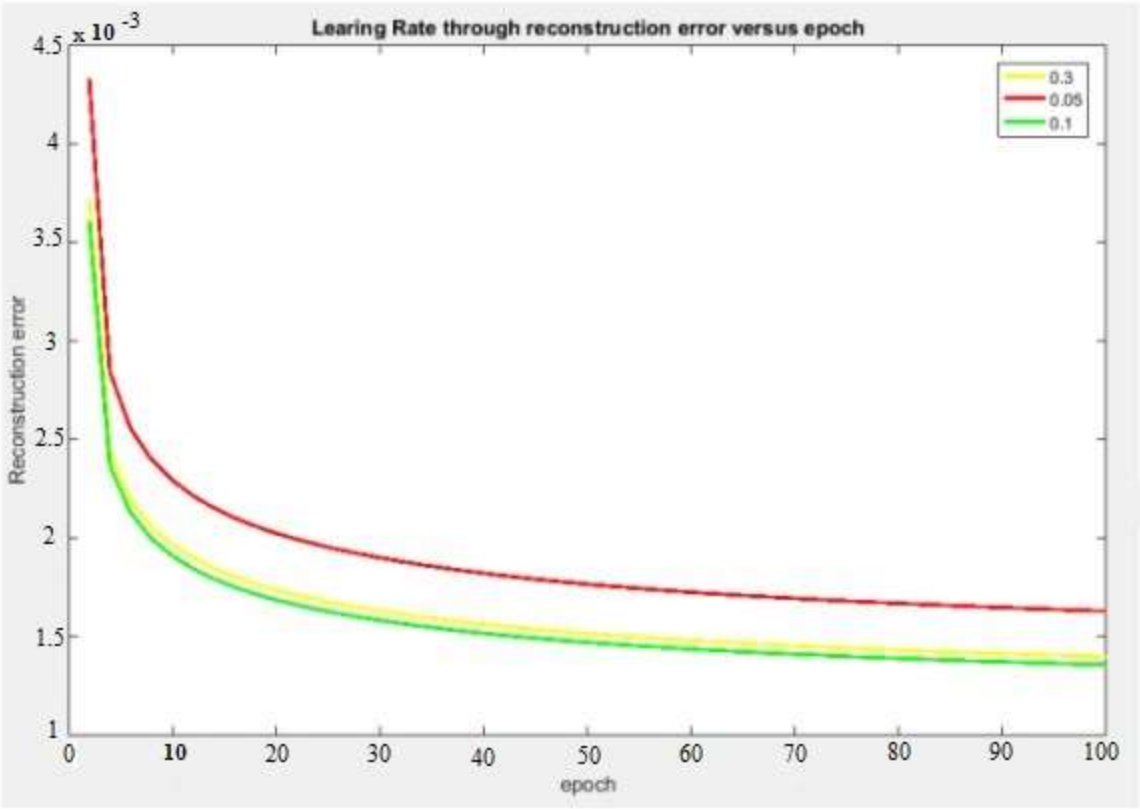

Figure 26: *The effect of Learning Rate on the Training Performance of the System*

5.5.2 The Test Result with Ten-Fold Cross-Validation

Considering the above settings, we trained the DBN algorithm to construct a Model. The model is tested using cross validation technique. Table 3 shows the accuracy, Specificity, Sensitivity, Precision and Misclassification Rate of the constructed Model. From the table we can see that the average accuracy, sensitivity, specify, precision and misclassification rate are 91.25, 81.14, 92.45, 59.06, and 8.68 percent respectively.

Table 3 : *Performance of the Constructed Model*

Fold #	Accuracy (%)	Specificity (%)	Sensitivity (%)	Precision (%)	Misclassification Rate (%)
1	93.8	95.07	80.12	60.3	6.2
2	94.25	95.55	81.11	64	5.75
3	93.6	95.61	77.77	59.47	5.75
4	93.35	94.69	74.8	50.45	6.65
5	91.54	92.86	81.6	60.23	8.45
6	92.44	93.75	84.08	67.75	7.55
7	88.79	89.52	84.08	55.74	11.21
8	90.19	90.96	85.95	63.21	9.8
9	87.93	89.06	80.74	53.54	12.06
10	86.58	87.63	81.18	55.94	13.41
Avg.	**91.247**	**92.47**	**81.143**	**59.063**	**8.683**

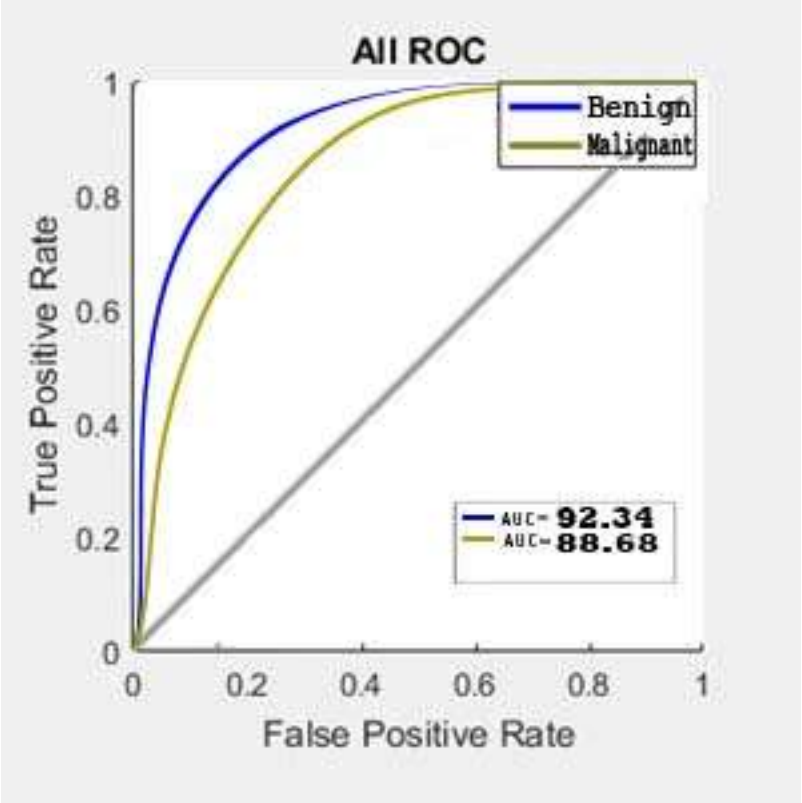

Figure 27: ROC Curves

The ROC curve above is a plot of the true positive rate against the false positive rate. Probably the most straightforward and intuitive metric for classifier performance is accuracy. Unfortunately, there are circumstances where simple accuracy does not work well. Unlike accuracy, ROC curves are insensitive to class imbalance dataset. AUC is the most important metrics for any classification models of skewed datasets to success in their performance. Thus, our proposed design would have an AUC value of 0.9234 for benign cases and 0.887 for malignant cases as shown in Figure 27.

5.6 Discussions

We have proposed lung nodule detection method using Deep belief network algorithm. Our yielded experiment entails that a model constructed from DBN is found to be comparable classifier to classify nodules as compared to CNN model which have sensitivity of 78.9% and ours is 81.143%. As shown in Table 4, SAE model produced a result of 75.01% accuracy and ours is 91.247%. Based on the review on [34], SDAE deep learning model produced a result of 79.29% sensitivity as discussed in Chapter Three. Thus, based on the result of this proposed method we got an encouraging result in accuracy, sensitivity and specificity. The performance of our model highly depends on the dataset size, model parameters, and the architecture of the algorithm employed. Nodules among different lesions can be effectively detected, due to the ability of this method to successfully preprocess and segment the lesions from the given input data and ability of DBN to automatically extract deep features of the input data starting from low level features that represent relevant information's of the input. Our model has the ability to classify nodules as benign or malignant as per the test results. Furthermore, each lesion is localized by its respective DICOM file id, slice number and centroid information.

The model has been quantifiably evaluated on 201 DICOM files yielding nearly 36,520 slices from which lung section could be extracted. Tenfold cross validation shows that an average accuracy of 91.247 % with sensitivity of 81.143%, Specificity of 92.47 %. Besides a ROC analysis showed that DBN gives satisfactory result.

As mentioned by previous attempts [1], which uses LIDC DICOM files, score a sensitivity of 78.9%. Their implementation uses deep CNN based on LIDC-IDRI databases. On this attempt they directly employ their model with CT images without any FPs reducing techniques before using the model of CNN. Attempts in [46], which used DICOM files of LIDC database and has got an accuracy of

75.01%and sensitivity of 83.35%. Here the Authors used SAE deep learning algorithm to detect and classify nodules. Attempts in [20, 26, 48, 45] used less number DICOM files for lung nodule detection. They achieved different percentages in sensitivity using different learning machines as we tried to summarize in Table 4. False positives are considered in their attempt because those attempts directly fed DICOMs into their learning algorithms. Segmentation of lung from its surrounding together with lesions is required like what we did using image processing techniques. It has significant effect in reducing false positives, computational time and improve the accuracy of our model.

Table 4: *Comparison with Recent Studies*

Studies	Methods (Models)	Database	**#samples**	**Sensitivity (%)**	**Specificity (%)**	**Accuracy (%)**
Our system	DBN	LIDC-IDRI	201	81.143	92.47	91.247
Rotem Golan [1]	CNN	LIDC-IDRI	888	78.9	71.2	NaN
Kumar,D; Wong, Clausi, D. [46]	SAE	LIDC-IDRI	157	83.35	NaN	75.01
Namin; Sarah, Taghavi, [48]	Fuzzy KNN	LIDC-IDRI	63	88	NaN	NaN
Orozco; Madero; Hiram [20]	SVM	LIDC-IDRI	106	90.90	73.91	NaN
Henry and Krewer [26]	KNNA	LIDC-IDRI	33	85.71	94.74	90.91
Htwe and Khaing, Zin, [45]	ANFIS	LIDC-IDRI	151	94.44	NaN	85

In Chapter One, two research questions were presented. According to the experiments our contributions with regard to these questions are summarized below.

RQ-1: Can image processing techniques (preprocessing and segmentation) reduces false positives while employing DBN in detecting lung nodules?

In this section, in order to answer this question, a separate preliminary experiment is done to show weather the relevance of image processing, before conducting DBN training, enhance to reduce false

positives or not. Based on this, 10 randomly selected DICOM files (2,500 slices) are used to test if nodules (lesions) are detected outside the lung section. This could be done by conducting experiment first without applying image processing technique that is, feeding the DICOM files as it is and secondly, feeding lesions inside lung section only by employing image processing techniques such as preprocessing and segmentation.

Hence, two scenarios are conducted by rendering DBN models using the same parameter setting. The results for nodule detected as malignant cases for both, with and without image processing, are counted and recorded for making comparisons with respect to ground truth nodule count. Besides, the centroid of each malignant nodules is also traced if it is inside or outside the lung object and recorded for making comparison.

According to nodules count shown in Figure 28, the number of truly detected nodules in each DICOM slices without employing image processing is higher than the case with image processing techniques (preprocessing and segmentation). As shown in Figure 28, the number of truly detected nodules in the first DICOM is 6, 7 and 9 as ground truth, with image processing, and without image processing respectively. This shows that out of the seven detected nodules with preprocessing and segmentation there exist one detected nodule which is found to be false positives inside the lung when compared to the ground truth nodules count. On the contrary, out of nine detected nodules without preprocessing and segmentation, there exist two false positives outside the lung object and one false positive inside the lung when compared to the ground truth. This shows that there is existence of false positives due to occurrence of nodule like objects outside the lung object.

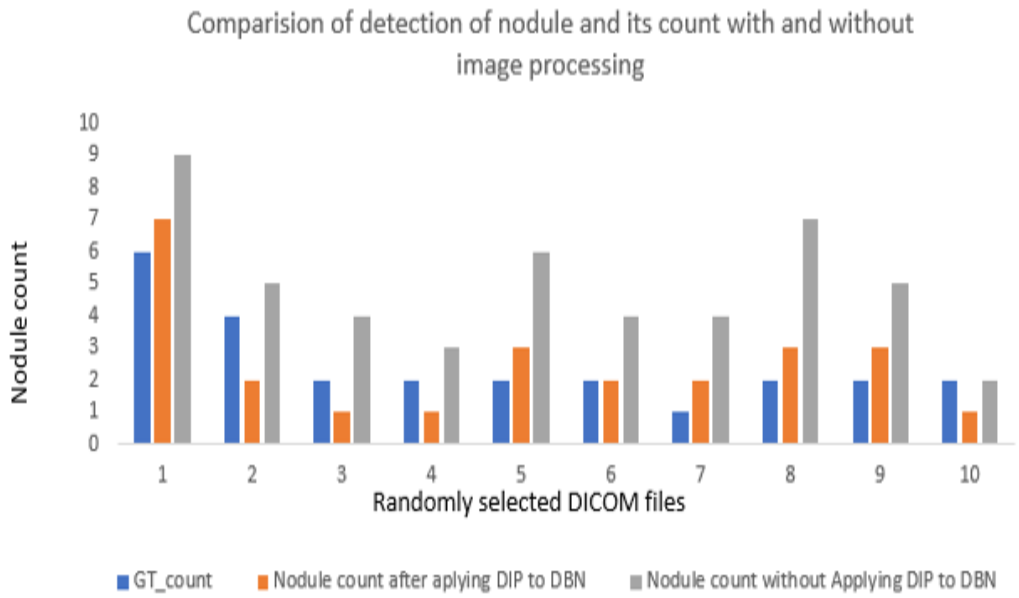

Figure 28:*The number of nodules detected in the ten randomly selected DICOM files*

The line chart specified in Figure 29, shows the existence of FPs from inside or outside the lung and entails that DBN with image processing produce a better result in reducing false positives as the number of false positives outside the lung is maximum. Therefore, based on this scenario employing DBN with image processing reduces false positives.

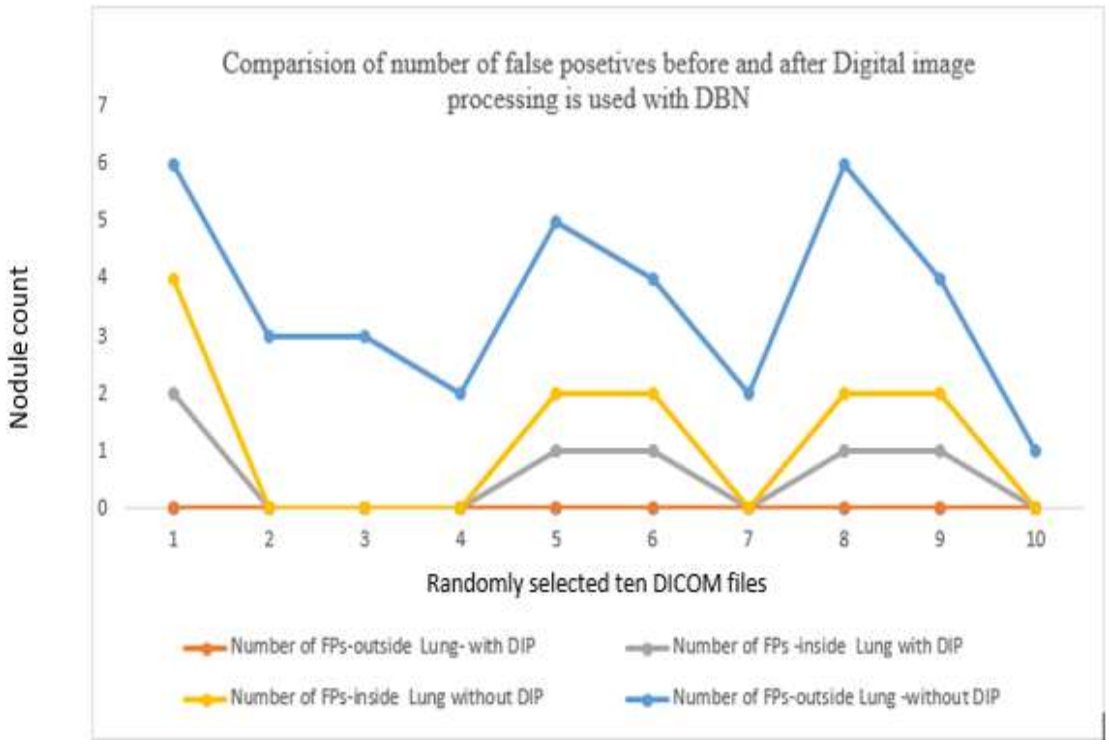

Figure 29: *The variation of false positive count with and without employing image preprocessing and segmentation*

RQ-2**:** Is deep belief network algorithm comparable in detecting lung nodules, for lung cancer detection problem?

Based on the results presented here, DBN algorithm provides comparable result for lung nodule detection system based on LIDC_IDRI CT images, as discussed in Chapter Three. Its detection accuracy and sensitivity are comparable to the works of deep CNN, SAEs, deep neural networks and other comparable approaches. Due to DBN constructs the model from high-level features (trained data) instead of low-level features its capability to produce correct performance is better than conventional approaches. Since it is an emerging technology in deep learning it incorporates various features where other models lack to consider. Therefore, DBN ensures that it can be a hopeful algorithm in the direction of lung nodule detection as well as other complex computer vision problems.

Chapter Six: Conclusion and Future Work

6.1 Conclusion

In this study, we have designed and implemented Lung Nodule Detection system. We adopted and designed different techniques for different steps of the recognition process. In the image processing components, we applied different techniques such as preprocessing and segmentation techniques. Median filter and histogram equalization are used to remove noises and improve the quality of the JPEG image. Adaptive thresholding and intersection operations are applied in segmenting the lung region and lesions from the lung only. Segmentation of artifacts (lesions) from the lung has its own advantage in reducing false positives. From our literatures, we have found that nodules exist only inside the lung. Lesions outside the lung which looks like nodules must be ignored. Considering lesions only on the lung reduce FPs.

From the segmented lesions, we have prepared input vectors for each lesion (probable nodules). The input vectors are features of the probable nodules or lesions. The input vectors are collected from image size of 76 x 76 pixels of set of data. This size selected, in taking all the features of the probable nodule lesions. Because of the maximum in plane diameter of a lung nodule in the dataset is 76 pixels. These input vectors are low level features feed in to DBN algorithm for training and model construction.

DBN network consists of three Restricted Boltzmann Machines (RBM), in which we got those layers improve the performance of the algorithm. The layers are labeled as RBM_1, RBM_2, and RBM_3. This DBN is composed of two major parts RBMs and BPNN. The first part is designed to extract valuable features from the input data and it is composed of different stack of RBMs. RBMs are trained in greedy layer manner of unsupervised fashion. This stage is called the pre-training stage, which advantages to get an initialized weight for the classifier. The second part of this DBN is the fine-tuning stage with BPNN.

As we have discussed in Chapter Five, we have implemented a prototype to construct DBN model for the classification of a nodules as malignant or benign. The model is tested using tenfold cross validation techniques. We have trained and tested the constructed modules using dataset collected from an international dataset, which is found in LIDC-IDRI database.

In this thesis, deep belief network was used for automatic lung nodule detection which helps to get the nodules in earlier stage and leads to better diagnosis. Three key important stages were employed in this thesis work: Image processing, DBN training and nodule classification. The prediction in the classification of benign or malignant pulmonary nodules was evaluated by LIDC_IDRI dataset. The major contributions of this thesis were accurate lung tissue segmentation to reduce FPs, segmentation of all lesions inside the lung which reduce FPs and computational time, detection of candidate nodules through DBN, and classification of nodules as benign or malignant. The experimental results suggested that DBN achieve better performance of nodule detection rate.

We have achieved an average accuracy of 91.247% in classifying a nodule as malignant or benign. Our proposed method achieves a sensitivity of 81.143% with 8.683% misclassification rates for lung nodules detection method. These results are comparable to previous works in terms of detection sensitivity and better in terms its FPs value per data. Our system is validated on 201 number CT images. Using the LIDC dataset we showed that the proposed method convincingly performs an encouraging result on overall accuracy and sensitivity metric.

6.2 Future works

We have adopted and designed different techniques for detection of nodules from the lung and achieved an encouraging result. However, we believe that incorporation of the following ideas as a future works, would achieve better result:

- Implementing the whole dataset of LIDC_IDRI DICOM files improve the robustness and sensitivity of the model.
- Using unsupervised DBN for pre-training and fine-tune it with SVM, Random Forest or Softmax classifiers and try to compare each other will widen to choose the best result.
- Using improved model architecture selection mechanisms and batch normalization will improve the model performance.
- Developing a system which able to provide levels to the malignant cases.

REFERENCES

[1] R. Golan, C. Jacob and J. Denzinger, "Lung nodule detection in CT images using deep convolutional neural networks," *In Neural Networks (IJCNN), International Joint Conference on, IEEE.*, pp. 243-250, 2016.

[2] B. Al Mohammad, P. C. Brennan and C. Mello-Thoms, "A review of lung cancer screening and the role of computer-aided detection," *Clinical Radiology,* vol. 72.6, pp. 433-442, 2017.

[3] Gruetzemacher, Richard and A. Gupta, "Using deep learning for pulmonary nodule detection and diagnosis," 2016.

[4] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma and Y. Wang, "Artificial intelligence in healthcare: past, present and future. Stroke and Vascular Neurology," 2017.

[5] B. Van Ginneken, A. A. Setio, C. Jacobs and F. Ciompi, "Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans," *In Biomedical Imaging (ISBI), IEEE 12th International Symposium IEEE,* pp. 286-289, 2015.

[6] F. Rossi and A. Rahni, "Combination of low level processing and active contour techniques for semi-automated volumetric lung lesion segmentation from thoracic CT images.," *In Biomedical Engineering & Sciences (ISSBES), IEEE Student Symposium, IEEE.*, pp. 26-30, 2015.

[7] G. E. Hinton, S. Osindero and Y.-W. Teh, "A fast learning algorithm for deep belief nets," in *Neural computation 18.7 (2006):*, 2006.

[8] G. Ian and e. al., Deep learning, vol. Vol. 1, Cambridge: MIT press, 2016.

[9] Tartar, Ahmet and A. Akan, "Ensemble learning approaches to classification of pulmonary nodules," in *Control, Decision and Information Technologies (CoDIT), International Conference on. IEEE,*, 2016.

[10] Deserno and T. M., "Fundamentals of biomedical image processing," *Biomedical Image Processing. Springer Berlin Heidelberg,* pp. 1-51, 2010.

[11] Dougherty and Geoff, "Digital image processing for medical applications," Cambridge University Press, Cambridg, 2009.

[12] SHAKTI and SHIV, "Comparative study of various image segmentation methods," *International Journal of Multidisciplinary Acadamy,* pp. 1-12, 2013.

[13] H. G. E., Deep belief networks, Scholarpedia, 2009.

[14] Hua and Kai-Lung, "Computer-aided classification of lung nodules on computed tomography images via deep learning technique," OncoTargets and therapy, 2015.

[15] Z. Luo, M. A. Brubaker and M. Brudno, "Size and Texture-Based Classification of Lung Tumors with 3D CNNs.," *In Applications of Computer Vision (WACV), IEEE,* pp. pp. 806-814, 2017.

[16] X. Li, "Pulmonary nodules detection algorithm based on robust cascade classifier for CT images.," in *Control and Decision Conference (CCDC), IEE*, Chinese, 2017.

[17] Shin and Hoo-Chang, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging,* vol. 35(5), pp. 1285-1298, 2016.

[18] P. B. Sangamithraa and G. S, "Lung tumour detection and classification using EK-Mean clustering," in *Wireless Communications, Signal Processing and Networking (WiSPNET), International Conference on. IEEE.*, 2016.

[19] T. Lampert, A. Stumpf and P. Gancarski, "An Empirical Study of Expert Agreement and Ground Truth Estimation," *IEEE Transactions on Image Processing* , vol. 25 (6), p. 2557–2572, 2016.

[20] Akram and Sheeraz, "Artificial Neural Network Based Classification of Lungs Nodule Using Hybrid Features from Computerized Tomographic Images," *Applied Mathematics and Information Sciences,* vol. 9(1), no. 183, 2015.

[21] Orozco, Madero and Hiram, "Automated system for lung nodules classification based on wavelet feature descriptor and support vector machine," 2015.

[22] LUNA-16, "https://luna16.grand-challenge.org/Data/," LUNA-16, 2016. [Online]. Available: https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI. [Accessed 15 October 2016].

[23] M. A. Keyvanrad and M. M. Homayounpour, "A brief survey on deep belief networks and introducing a new object oriented toolbox (DeeBNet)," arXiv Aug. , 2014.

[24] A. Teramoto, "Automated detection of pulmonary nodules in PET/CT images: Ensemble false-positive reduction using a convolutional neural network technique," *Medical Physics 43.6 (2016): 2821-2827.,* vol. 43(6), pp. 2821-2827, 2016.

[25] Clark and Kenneth, "The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository," *Journal of Digital Imaging,* vol. 26(6), pp. 1045-1057, 2013.

[26] K. Henry, "Effect of texture features in computer aided diagnosis of pulmonary nodules in low-dose Computed Tomography," in *IEEE international conference,* 2013.

[27] Niknam and Farshid, "Approach to Multiple Pulmonary Nodules: A Case Report and Review of Literature," *The Scientific World Journal,* vol. 11, pp. 760-765, 2011.

[28] Wallace and M. B., "Minimally invasive endoscopic staging of suspected lung cancer," *Jama,* vol. 299(5), pp. 540-546, 2008.

[29] Amir, G. J., H. P. and Lehmann, "After Detection:: The Improved Accuracy of Lung Cancer Assessment Using Radiologic Computer-aided Diagnosis.," *Academic Radiology,* vol. 23(2), pp. 186-191, 2016.

[30] Traverso and Alberto, "Computer-aided detection systems to improve lung cancer early diagnosis: state-of-the-art and challenges," *Journal of Physics: Conference Series.,* vol. 841 (1) , pp. Journal of Physics: Conference Series. Vol. 841. No. 1. IOP Publishing, , 2017.

[31] A. El-Baz, "Computer-aided diagnosis systems for lung cancer: challenges and methodologies," *International journal of biomedical imaging,* 2013.

[32] Jalalian and Afsaneh, "Foundation and methodologies in computer-aided diagnosis systems for breast cancer detection," *EXCLI Journal ,* vol. 16, p. 113, 2017.

[33] Parveen, S. Shaik and C. Kavitha, "A Review on Computer Aided Detection and Diagnosis of lung cancer nodules," *International Journal of Computers and Technology,* vol. 3(3), pp. 393-400, 2012.

[34] Sun, Wenqing, B. Zheng and W. Qian, "Computer aided lung cancer diagnosis with deep learning algorithms," *SPIE Medical Imaging, International Society for Optics and Photonics,* 2016.

[35] Firmino and Macedo, "Computer-aided detection system for lung cancer in computed tomography scans: Review and future prospects," Biomedical engineering, 2014.

[36] Ejaz, Naveed, S. Javed and Z. Sajid, Implementation of Computer Aided Diagnosis System for Lung Cancer Detection, Lecture Notes on Software Engineering, 2013.

[37] Gong and Jing, "Computer-aided detection of pulmonary nodules using dynamic self-adaptive template matching and a FLDA classifier," Physica Medica, 2016.

[38] Jacobs and Colin, "Computer-aided detection of pulmonary nodules: a comparative study using the public LIDC/IDRI database.," *European radiology,* pp. 1-9, 2015.

[39] Desai, M. Bharatbhai, S. V. Patel and B. Prajapati, "ANOVA and Fisher Criterion based Feature Selection for Lower Dimensional Universal Image Steganalysis," *International Journal of Image Processing (IJIP)*, vol. 10(3), p. 145, 2016.

[40] P. Golik, D. Patrick and N. Hermann, "Cross-entropy vs. squared error training: a theoretical and experimental comparison," *Interspeech*, vol. 13, 2013.

[41] Sun and Zhaojie, "Cross-entropy-based antenna selection for spatial modulation," *IEEE Communications Letters*, vol. 20(3), pp. 622-625, 2016.

[42] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," *Nature*, pp. 436-444, 2015.

[43] Hua, Yuming, J. Guo and H. Zhao, "Deep Belief Networks and Deep Learning," in *Intelligent Computing and Internet of Things (ICIT), 2014 International Conference on. IEEE*, 2015.

[44] G. Hinton, "A practical guide to training restricted Boltzmann machines," in *Momentum 9.1*, 2010.

[45] Htwe and Z. Khaing, "Automated lung nodule classification by artificial neural network and fuzzy inference system," in *Consumer Electronics, IEEE 5th Global Conference on. IEEE*, 2016.

[46] D. Kumar, A. Wong and D. A. Clausi, "Lung nodule classification using deep features in CT images," *In Computer and Robot Vision (CRV) IEEE*, pp. 133-138, 2015.

[47] K. S. Sheng Chen, "Computerized Detection of Lung Nodules by Means of "Virtual Dual-Energy" Radiography," in *IEEE*, 2013.

[48] Namin and T. Sarah, "Automated detection and classification of pulmonary nodules in 3D thoracic CT images," *Systems Man and Cybernetics (SMC), IEEE International Conference on. IEEE*, 2010.

[49] E. Dandil and e. al, "Artificial Neural Network Based Classification System for Lung Nodules on Computed Tomographyic Scans," in *International Conference of Soft Computing and Pattern Recognition, IEEE*, 2014.

[50] Z. Zhiwei, S. Daifeng, C. Yuanzhi and G. Haoyan, "Computer-aided Detection of Lung Nodules with Fuzzy Min-max Neural Network," in *Sixth International Conference on Intelligent Human-Machine Systems and Cybernetics*, 2014 .

Annex A: Fragment of Matlab Code Implemented

Image Acquisition, DICOM to Image conversion

```
% Ask the user what data he wants to load
% [filename, pathname] = uigetfile({'*.*'},'File Selector');
% fullPathname = strcat(pathname, filename);

 dataset = load ('dataset.mat');
 dataset = cell(1,1);
 save('dataset.mat','dataset');
 Dir = 'Input_Dataset\';

 % Read images from Images folder
 Ctr = 0;
 Dicom2Image = cell(0,0);
 dicom_id = cell(0,0);
 Each_Slice_path = cell(0,0);
 inputDicom=cell(0,0);

for j=1:36520
     files = dir(fullfile(strcat(Dir, 'dicom_', num2str(j),'\', '*.dcm')));
  for k=1:numel(files)
      file_name=files(k).name;
      image_name=strcat(strcat(Dir,'dicom_', num2str(j) , '\'),file_name);

      %DICOM to Image conversion
      I=dicomread(image_name);

      %Preprocessing
      I=histeq(Input);
      I=medfilt2 (I);

      Dicom2Image{j,k}=I;
      inputDicom{j,k}=I;
      dicom_id{j,k}=j;
      Each_Slice_path{j,k}=image_name;

  end

end
```

Lung Segmentation

```
% hObject    handle to pushbutton2 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
% Initalization
 data_seg_whole=cell(0,0);
 data_seg_Lung=cell(0,0);
 segctr=0;

 %#########################################################
   inputDicom=handles.inputDicom;
   for i=1:size(inputDicom,1)
     for j=1:size(inputDicom,2)
           if ~isempty(inputDicom{i,j})

                %Normalization
                Input=im2double(inputDicom{i, j});
                L=im2bw(L);
                L=~L;

                %Applay morphological operation
                se=strel('disk',5);
                L=imdilate(L,se);
                V=~L;
                V=imfill(V,'holes');

                %Extract only the Lung Part
                 L=L&V;
                 L=bwareaopen(L,6000);
                 L=imdilate(L,se);
                 L=imfill(L,'holes');
                 se=strel('disk',8);
                 L=imerode(L,se);
                 data_seg_Lung{i,j}=L;
 %#########################################################
%Segment the whole images extracted from the dicome using adaptive
segmentation
                 R2 = adaptivethreshold((Input),[32 32],.001,0);
                 se=strel('disk',1);
                 bin=imerode(R2,se);
                 data_seg_whole{i,j}=bin;
           end

      end
  end
```

Extract the most probable Lesions inside the Lung

```
% hObject    handle to pushbutton3 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)

data_seg_whole=handles.data_seg_whole;
data_seg_Lung=handles.data_seg_Lung;
data_seg_for_postprocessing=cell(0,0);
artifactctr=0;

%#########################################################################
%Refine only the most probable lesions inside the lung only

for i=1:size(data_seg_Lung,1)
      for j=1:size(data_seg_Lung,2)
           if ~isempty(data_seg_Lung{i,j})
               I1=data_seg_whole{i,j};
                UB=1000;
                LB=10;
                I3=I1;
                I3(data_seg_Lung{i,j} ~= 1)=0;
                I3=xor(bwareaopen(I3,LB),bwareaopen(I3,UB));
               %Lasons lessthan approximatly 3mm are unnecessary
               Data_seg_for_postprocessing{i,j}=I3;
               %Segmented elements inside Lung only
               artifactctr=artifactctr + 1;

           end
      end
end
```

Input Vector Preparation

```
% hObject    handle to pushbutton6 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
%#####################################################################

%Initalize parameters
   data_seg_for_postprocessing=handles.data_seg_for_postprocessing;
   Each_Slice_path=handles.Each_Slice_path;
   data_seg_for_postprocessingcpy=data_seg_for_postprocessing;
   Post_processesd_data_bin=cell(0,0);
   Post_processesd_data_col=cell(0,0);
   Post_processesd_data_Blob=cell(0,0);
   Nodule_Dictionary=cell(0,0);
   Index_data_blob={};
   datasetctr=0;
%hold color input
   data=handles.data;
   dataset = load ('dataset.mat');
   dataset=dataset.dataset;
   record=cell(1,1);output=cell(1,1);
   pi=3.14;
   Index=0;
   ctrblob=0;
   ctrnodule=0;
% End of initalization
%#########################################################################
%Annotate each dicom file based on Lampert resulut found inside groundtruth
%directory

for k=1:size(data_seg_for_postprocessingcpy,1)
     id=strcat('dicom_',num2str(k));
      %Read slice_correspondences.txt inside 'groundtruth/dicom_k' and
      %extract slice name and its corsponding slice folder
       dicomStr =  extractDicom (id);

    %For each dicom slice name(.dcm extension), check if it is Slice+ or
    %Slice- where Slice+ tells us the presence of nodule otherwise, non
    %nodule
    for x=1:size(data_seg_for_postprocessingcpy,2)
```

…Continued

```
                    if num == x
                        found=1;
                        break;
                    end
                end
  % annotate  if dicomfile is detected
              if found
               %data_seg_for_postprocessingcpy{k,x}
          Dir = strcat('groundTruth\dicom_',num2str(k));
files = dir(fullfile(strcat(Dir, '\slice', num2str(sliceno),'\', '*.tif')));
                      for m=1:numel(files)
                          file_name=files(m).name;
   image_name=strcat(strcat(Dir, '\slice',num2str(sliceno),'\'),file_name);
                      I=imread(image_name);

data_seg_for_postprocessingcpy{k,x}=annotate(data_seg_for_postprocessingcpy{k,x},I);
                           end
              %End of if found
               end
            %End of if not empty
         end
        %End of for loop (x)
     end
    %End of for loop(k)
end
%####################################End of anotating mapped dicom files
%For each nodule inside the lung section  should be represent in to a row
%vector interms of input data vs  output data
for i=1:size(data_seg_for_postprocessing,1)

      for J=1:size(data_seg_for_postprocessing,2)

         if ~isempty(data_seg_for_postprocessing{i,J})
       %   Z= zeros(size(data{i,J}, 1), size(data{i,J}, 2));
      %    emptyZ=1;
          cc=(data_seg_for_postprocessing{i,J});
                % cc=imfill(cc,'holes');
              [L,no]=bwlabel(cc,8);
               Attrib=regionprops(L,'Area'
,'Centroid','BoundingBox','MajorAxisLength');
            for k=1:no
                val=logical(displabel(L, k, data_seg_for_postprocessing{i,J}));
%                Area=Attrib(k).Area;
                  Centroid=Attrib(k).Centroid;
                  BB=Attrib(k).BoundingBox;
%                 p=Attrib(k).Perimeter;
%                 roundness = 4*pi*Area/p^2;
                MA=Attrib(k).MajorAxisLength;
                  %##############################################
                  % save('dataset.mat', 'RowDset');
                  %##############################################
```