Addis Ababa University

Addis Ababa Institute of Technology

Electrical and Computer Engineering Department

Classification of Candidate Pulmonary Nodules Segmented from CT Image by Using CNN

By: Nebyue Awoke

October, 2019

Addis Ababa University

Addis Ababa Institute of Technology

Electrical and Computer Engineering Department

Classification of Candidate Pulmonary Nodules Segmented from CT Image by Using CNN

By: Nebyue Awoke
Advisor: Menore Tekeba

A Thesis Submitted to the School of Electrical and Computer Engineering in Partial Fulfillment of the Requirements for the Degree of Masters of Science in Computer Engineering

October, 2019
Addis Ababa, Ethiopia

Addis Ababa University

Addis Ababa Institute of Technology

Electrical and Computer Engineering Department

Classification of Candidate Pulmonary Nodules Segmented from CT Image by Using CNN

By: Nebyue Awoke

Approval by Board of Examiners

Dr. Yalemzewed Negash _____

Dean, School of Electrical and Computer Engineering — Signature

Mr. Menore Tekeba _____

Advisor — Signature

Dr. Sosina_Mengistu_ _____

Internal Examiner — Signature

Mr. Getachew Teshome _____

External Examiner — Signature

Abstract

Lung Cancer is a leading cause of human loss globally i.e. compared with other cancer related deaths. The five-year relative survival rate of lung cancer is only 16%; however, early recognition of nodules and proper treatment of this disease reduce the death rate due to lung cancer up to 20%. To detect such nodules CT lung image analysis has been used by radiologists all over the world. However, analysis of these images is a very challenging task for radiologists, because the number of slices in one scan can be up to 600. Therefore, computer aided-detection (CAD) systems are very important for a quicker and more precise assessment of the data. False positive reduction is one of a vital element of computer aided diagnosis (CAD system), which plays an important role in lung cancer diagnosis and early treatment. In this thesis we proposed to design a framework for classification of candidate CT image slices by employing 3D Convolutional Neural Network (CNNs) to reduce a significant number of false positive candidates. 3D CNNs are favorable than 2D CNNs because 3D CNNs can encode richer spatial information and extract more representative features via their structural architecture trained with 3D samples. The proposed design has been extensively validated by using the dataset obtained from LUNA16 challenge providers.

The proposed approach mainly consists of three steps Pre-processing, Feature extraction & Classification and Fusion. In the preprocessing phase we carefully examined our data set and perform resampling to avoid image slice thickness variations because of different CT machines. In addition to that we employed data augmentation to reduce class imbalance between ground truth nodules and false positive candidates. Sizes of the nodules varied from 3mm up to 30mm, so we extract four receptive fields to encompass all nodule types (i.e. small, medium and large nodules), aiming to understand the effect of input patch sizes in the performance of the system. We designed four 3D CNNs for the corresponding four patch sizes, each CNNs model contains from 3 convolutional layers. We employed model fusion technique to acquire an improved result by using the aggregate strengths of each model. The proposed framework has been tested by the dataset provided by the LUNA16 Challenge and we achieved the competition performance metric (CPM) score of 0.8541 with a highest sensitivity of 0.8706 with 1 false positive per scan and 0.9275 at 8 false positives per scan.

Generally, from the test results we noticed that increasing the input patch size increases the overall CPM score because larger patch sizes can be able to encompass a large number of nodules within the dataset and any reasonable fusion method can be able to boost the overall classification performance.

Keywords: Pulmonary Nodule Detection, Computed Tomography, Convolutional Neural Networks, Deep Learning, false positive reduction, computer-aided diagnosis.

Declaration

I, the undersigned, certify that research work titled Candidate Pulmonary Nodule Classification Using CNN. The work has not been presented elsewhere for assessment. Where materials that have been used from other sources used for this thesis have been fully acknowledged.

Declared by:

Name: Nebyue Awoke

Signature: _____

Date: _____

Date of submission: October, 2019

Addis Ababa, Ethiopia

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

Confirmed by advisor:

Name: Menore Tekeba

Signature: _____

Date: _____

Acknowledgement

First of all, thanks to the Almighty God for His showers of blessings throughout my life and giving me the opportunity to undertake and accomplish my research successfully.

It is a great pleasure to give my sincere appreciation to my advisor Mr. Menore Tekeba for his continuous guidance and encouragements from the initial proposal to the end of this thesis research. I really admire on how you put an endless effort to make this thesis a success.

I am grateful to give my thanks for all my friends and classmates for the time we had together. My special thanks to my Family: parents and lovely brothers for your constant love, support and encouragement in my life. Dear brothers Bayeh Embiale and Yared Awoke I thank you for being there when I need your support and I will never forget your encouragement.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

LIST OF ACRONYMS

2D	Two Dimensional
3D	Three Dimensional
AI	Artificial Intelligence
ANN	Artificial Neural Networks
ANODE09	Automatic Nodule Detection 2009
CAD	Computer-aided Detection
CE-CT	Contrast Enhanced Computed Tomography
CPM	Competition Performance Metric
CPU	Central Processing Unit
CT	Computed Tomography
CR	Computed Radiography
DA	Deep Auto-encoder
DBN	Deep Belief Network
DBN	Deep Boltzmann Machine
DC-ELM	Deep Conventional Extreme Machine Learning
DICOM	Digital Imaging and Communications in Medicine
DLCST	Danish Lung Cancer Screening Trial
DNN	Deep Neural Networks
DR	Digital Radiography
FC	Fully Connected
FLD	Fisher Linear Discriminant
FNs	False Negatives
FROC	Free-Response ROC Curve
FPs	False Positives
GB	Gentle Boost Classifier
GPU	Graphical Processing Unit
HU	Hounsfield Units
ISBI	International Symposium on Biomedical Imaging
KNN	K-Nearest Neighbor
LBP	Local Binary Patterns

LDA	Linear Discriminant Analysis
LDC	Linear Discriminant Classifier
LDCT	Low-dose Computed Tomography
LIDC-IDRI	Lung Image Database Consortium - Image Database Resource Initiative
LUNA-16	Lung Nodule Analysis 2016
PNs	Pulmonary Nodules
ML	Machine Learning
MLP	Multilayer Perceptron
MRI	Magnetic Resonance Imaging
NN	Neural Network
RAM	Random Access Memory
RBF	Radial Basis Function kernel
RF	Random Forest Classifier
ReLU	Rectified Linear Unit
RGB	Red-Green-Blue
ROIs	Region of Interests
RT	Radiographic Testing
RNN	Recurrent Neural Network (RNN)
RTR	Real Time Radiography
PET	Positron Emission Tomography
SPN	Solitary Pulmonary Nodule
STP	Standard Temperature and Pressure
SVM	Support Vector Machine
TNs	True Negatives
TPs	True Positives
UTMB	University of Texas Medical Branch

Chapter One: Introduction

1.1 Background

Lung cancer is one of the most commonly known deadly cancers, caused by an abnormal (uncontrolled) cell growth within the lung tissues commonly known as “nodules”. It is the leading cause of human loss globally (i.e. compared with other cancer related diseases like prostate, colon, polyp, breast, etc. [1]). Early detection of nodules and proper treatment of this disease may reduce the death rate due to lung cancer. Up to 20% of deaths from lung cancer are estimated to be preventable with early detection and treatment [2][3]. To facilitate such detection and treatment, human radiologists use low-dose CT (computed tomography) scans [4] to look for pulmonary nodules (PNs) because PNs have high probabilities to become malignant nodules that may develop into cancer. A solitary pulmonary nodule is defined as a single nodule seen on an x-ray or CT scan, that is less than or equal to 30 mm in diameter. If a "spot" on the lung is larger than 3 cm there is a greater chance to be a cancer. It may be solid or sub-solid in attenuation. Semisolid nodules may have purely ground-glass attenuation or be partly solid (mixed solid and ground-glass attenuation) [5][6].

Accurate diagnoses of disease lung cancer depend upon image acquisition and image interpretation. Image acquisition devices have improved substantially over the recent few years i.e. currently we are getting radiological images (X-Ray, CT and MRI scans etc.) with much higher resolution [7]. The various commonly known imaging modalities are computed tomography (CT), contrast enhanced computed tomography (CE-CT), low-dose computed tomography (LDCT) and positron emission tomography (PET) for detection and diagnosis of lung cancerous cell [8]. CT examination used to predict lung nodule malignancy in patients to make noninvasive early diagnosis and treatment of lung cancer. However, a CT scan has a large number of images that must be interpreted by a radiologist, which is a challenging process. So that, the use of a computer-aided detection (CAD) system can provide an effective solution by assisting radiologists in increasing the scanning efficiency and potentially improving nodule detection and classification [9][10].

Automated pulmonary nodule detection by using a CAD system mainly consists of three steps: *Image preprocessing:* used to standardize the data, restrict the search space for nodules ROIs, and reduce noise and image artifacts (i.e. 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 [11][12][13]. This step may contain DICOM specific operation, format & window size identification and thorax part extraction (i.e. lung segmentation is essential and critical to find the accurate ROIs from the portion of the lungs)[8].

Candidate nodule detection: detect and segment the candidate pulmonary nodules from segmented lungs and classified into malignant and benign on the basis of shape, growth, texture and appearance analysis. A large number of coarse candidates are rapidly screened throughout the whole volume using a variety of criteria, e.g., intensity thresholding, shape curvedness and mathematical morphology [14][15][16]. The candidate detection stage aims to detect nodule candidates at a very high sensitivity, which typically comes with many false positive candidates.

False positive reduction: reduces the number of false positive results by classifying each candidate as nodule or non-nodule. In order to maintain a high sensitivity in candidate screening a great number of coarse candidates are selected out and forwarded to this step. In this regard, the false positive reduction track stands as the most crucial component of an automated pulmonary nodule detection system and a lot of efforts have been dedicated to improve the performance of this step [17][19]. In false positive reduction track the input is set of candidates resulted from nodule detection, but in candidate nodule detection raw CT scan images are used as an input. Effective classifiers together with discriminative features are required to reduce a large number of false positive candidates [17].

In recent years Machine Learning and Artificial Intelligence are playing an important role in medical arena like medical image processing, computer-aided diagnosis, medical image interpretation, image fusion, image registration, image segmentation, image-guided therapy, etc. ML techniques extract information from the images and represents information effectively and efficiently. Previously, machine learning techniques composed of conventional algorithms (i.e. SVM, NN, KNN etc.) these techniques enhance the abilities of doctors and researchers to understand that how to analyze the generic variations which will lead to disease, but they are not capable to handle complex problems efficiently[7]. Deep learning techniques were introduced with many layered architectures (i.e., input layer, output layer and one or more hidden layers). Deep learning is based on deep neural networks comprised of a large number of hidden

layers applied very effectively to computer vision problems in both classification and object detection – especially where large set of benchmark training data are available.

These days, convolutional neural networks (CNN) have been known as the most effective deep learning algorithm for visual recognition tasks. The remarkable successes of deep convolutional neural networks (CNNs) in image and video processing have been shown to outperform the state-of-the-art in several computer vision applications [17][18][19], it is providing an exciting solution in medical image analysis with an excellent accuracy. The representation capability of the high-level features which are learned from large amounts of training data has been broadly recognized and inspired some researchers to employ CNNs in automated pulmonary nodule detection and classification [17]. In computer vision, deep convolutional neural networks (CNNs) have been announced because they can be able to achieve a good performance and simulate like the behavior of the human vision system and learn hierarchical features, allowing object local invariance and robustness to translation and distortion in the model [20]. CNN is capable and successful in the representation of the high-level features which are learned from large amounts of training dataset. In candidate classification, effective classifiers together with categorized features are compulsory to reduce a large number of false positive candidates and to maintain a high accuracy. In this regard, in this thesis we propose to use a CNNs framework for candidate nodule classification or false positive reduction task in automated pulmonary nodule detection from CT images.

1.2 Statement of the problem

Lung cancer is responsible for a large number of deaths and significant health care costs. The five-year relative survival rate of lung cancer is only 16%; however, if nodules are detected at an early stage, the survival rate can be increased [9]. Lung nodule detection is primarily done manually by trained pulmonary radiologists with the help of CAD (computer-aided diagnosis) systems. But still accurate detection of pulmonary nodules remains a technical challenge due to subjectivity of nodules, complexity of images, extensive variations exist across different interpreters, and fatigue.

Existing CAD systems are designed to be highly sensitive during nodule generation phase, so they generate a large number of potential nodule candidates and then a radiologist looks and asses through the nodules to classify them [2]. It increase radiologists burden due to large

amount of candidates to be analyzed and different types of patient CT scans, even for highly trained radiologists, detecting nodules and predicting their relationship to cancer are challenging task for several reasons, first, nodules are typically minor, mainly in the pre-cancer stage; second, their appearance is not always different from that of other benign tissue formations in the lungs some nodules may attached to the blood vessels or the lung wall and difficult to detect, which have low contrast compared to the surrounding tissues. and third, the resolution of CT imagery can vary in ways that make precise identification challenging [5] leading to both false positive and false negative results that can negatively affect patients health.

Problems that make lung nodule detection system from CT scans most challenging task are summarized as follows:

- Pulmonary nodules have large dissimilarities in sizes, shapes and locations.
- The contextual environments around them are often diversified for different categories of lung nodules, such as solitary nodules, ground glass opacity nodules, cavity nodules and pleural nodules [21].
- Pathologically, there are many types of nodules (e.g. solids, non-solids, part-solids, calcified, etc. [22])
- Large amount of CT scans has to be analyzed for a patient to be cancer victim or not, which is an enormous burden for radiologists. The work load in analyzing number of DICOM slices leads to error prone in radiography reporting.
- Some false positive candidates carry quite similar morphological appearance to the true pulmonary nodules.

Generally, recent CAD systems for candidate nodule detection are achieving a very high sensitivity levels and can be able to improve radiologists' accuracy in the recognition of nodules, but current systems for nodule detection are reporting large number of false positives because of the above challenges. Thus there is a need for better techniques to assess CT scans for cancerous lung nodules, with a broad goal of improving predictive precision. So, we tried to design a CNN model to classify candidate pulmonary nodules to achieve high sensitivity of prediction with a small number of false positives and also we tried to address the following research questions.

- What is the effect of input patch sizes to the performance of the CNNs architecture?
- Compared with individual CNN models, may ensemble of classifiers achieve a substantial improvement in accuracy?

1.3 Objectives

1.3.1 General Objectives

The main objective of this study is to design, implement and test a CNN based model that classifies the true pulmonary nodules from a large number of candidate nodules to achieve high sensitivity with a small false positive rate and also comparing the effect of the input patch size on the classification performance.

1.3.2 Specific Objectives

Under the above general objective the study has the following specific objectives:

- To examine pre-processing methods and perform an effective data pre-processing.
- To investigate training approaches and training parameters to find an "optimal" trained model
- To compare and analyse the effect of the input patch size on the classification performance.
- Comparing the ensemble fusion result with respect to each individual CNN models
- Evaluating the competition performance metric (CPM) result with related works in the false positive reduction track.

1.4 Significance of the Study

Lung Cancer is a potentially life threatening medical case requiring adequate diagnosis and treatment from a doctor, it needs to perform a proper examination and analysis of CT scans to detect pulmonary nodules leading to cancer; many researchers have been actively studying approaches and techniques that can automatically detect pulmonary nodules in computed tomography images. False positive reduction is one of the most important steps of a computer aided detection system, which plays a significant role in lung cancer recognition and early stage treatment. This study can provide a significant research contribution in the area of CT imaging to improve accuracy of detecting malignant pulmonary nodules. The paper anticipated to provide various benefits for both local and global researchers and professionals in the area.

- The research plays an important role to understand the challenges in chest radiography and puts an approach that leads to simplicity of CT image analysis.
- It provides a research output for researchers in CT image processing and can be a comprehensive literature review to boost CT based pulmonary screening.
- Helps to reduce error prone generated from radiographers and it also initiates researchers to do lung cancer detections with different approaches.

1.5 Contributions of the Thesis

Our contribution regarding to this paper can be illustrated as:

- First of all we applied 3D CNNs successfully for candidate nodule classification task, i.e. compared with their 2D complement the 3D CNNs can encode a large amount of spatial information and can be able to extract and encode complex patterns.
- Second, we carefully examined our dataset while applying some preprocessing techniques for the improvement of classification performance and also we used ensemble of classifiers at the end, hence the 3D CNNs ensemble network outperforms each individual models by achieving a highest sensitivity of 0.8706 with 1 false positive per scan and 0.9275 at 8 false positives per scan.
- In addition to that sizes of the nodule vary from 3 mm to 30 mm and there is no a global consensus and objective study that compares the effect of input patch size to the performance of the network. Therefore, we examined and compare the effect of input patch size to the performance of the system and we noticed that increasing patch sizes improve accuracy with a cost of training time.

1.6 Scope and Limitations

1.6.1 Scope of the Study

The ultimate goal of this research is designing and implementing our CNN model to reduce false positive results. We attempt to develop a technique that uses a 3D CNN architecture for accurate classification of candidate chest radiograph images generated from a CT scan. The scope of this study is discriminating the true nodules from a large number of candidates to reduce significant false positive results and comparing the effect of input patch sizes to the performance of the network. We evaluated the proposed approach on a large-scale labeled dataset, which was provided by the LUNA16 challenge held in conjunction with ISBI 2016.

1.6.2 Limitations of the Study

The computational resources we used in this study (i.e. Intel® core™ i7-7500U CPU@2.70, ~2.9 GHz processor, GPU:Intel® HD Graphics 620, RAM: 8GB.) restrict the performance of the proposed system, increase patch extraction time during the preprocessing phase and it takes time for the training process. Since LUNA16-ISBI 2016 folds a huge data set around 124GB (i.e. total

number of candidates are 551,065), extracting, managing, analyzing the dataset and conducting demonstration was a very difficult task.

1.5 Research Methodology

To conduct our research and to implement the proposed system we follow the following steps as a methodology of our study.

1.7.1 Literature Review

We attempt to survey scholarly papers, which include the previous and the current knowledge and findings as well as theoretical and methodological contributions. A deep study was made in the literature written on this area to have a clear background about the work. Different previous papers written about CAD system, lung cancer, deep learning, image processing, CNN, etc. were reviewed to understand the various theoretical knowledge and technical approaches about automated pulmonary nodule detection and classification. The researches were selected based on their publisher rank, citation index, time of publication, and their achievement. So, as much as possible the documents are supposed to be up to date, high ranked and have good achievement

1.7.2 Data Collection

The dataset obtained from the official website of LUNA16 [23] Challenge, which is a medical image analysis challenge held in conjunction with ISBI 2016, it contains a large-scale benchmark dataset. It also provides an evaluation outline for automated nodule detection and classification. Extraction was performed from 888 patient's scans selected from the publicly available dataset LIDC-IDRI [24], containing a total of 1018 CT scans. The dataset for false-positive reduction consists of 551, 065 nodule candidates and 1120 out of 1186 ground truth nodules (with sensitivity of 94.4%). These candidates are generated by merging the candidates that were detected by Murphy et al. [16], Jacobs et al. [3], Setio et al. [15], etc

1.7.3 Data Preprocessing

We perform different data preprocessing techniques to make the dataset more suitable for training process. We perform resampling to avoid image slice thickness variations because of different CT machines. We also employed data augmentation to reduce class imbalance between ground truth nodules and FPs candidates. Sizes of the nodules varied from 3mm up to 30mm, so we extract four receptive fields to encompass all nodule types (i.e. small, medium and large nodules), aiming to understand the effect of input patch sizes in the performance of the system.

1.7.4 System Design and Implementation

We designed four 3D CNNs for the corresponding four patch sizes, each CNNs model contains from 3 convolutional layers. We employed model fusion technique to acquire an improved result by using the aggregate strengths of each model. The design process includes figuring out the architecture of the system, determining the training approach of each component of the system, determining system parameters and their value, and specifying the performance metrics of the system. The implementation part has been performed by using python programming language on Anaconda3 5.2 64 bit using Jupyter Notebook as an editor.

1.7.4.1 Implementation Tools

In this paper the following hardware and software resources have been used: ***Hardware Specification:*** the algorithms are run on Intel® core™ i7-7500U CPU@2.70, ~2.9 GHz processor, GPU:Intel® HD Graphics 620, RAM: 8GB. ***Software:*** Cuda_9.0.176_windows**,** Python 3.6.5 (Anaconda3 5.2 64 bit): Tensor flow-GPU, SimpleITK, Numpy, Pandas Matplotlib with an editor Jupyter Notebook.

1.7.5 Result Evaluation and Conclusion

The evaluation is performed by measuring the detection sensitivity of the system and the corresponding false positive rate per scan obtained by LUNA16 [23] challenge providers. The final result of our study is determined by using the average sensitivity of 7 predefined FPs points defined by LUNA16 challenge providers i.e. 1/8, 1/4, 1/2, 1, 2, 4, and 8 FPs per scan. Finally, based on our results we set our conclusion.

1.8 Organization of the Thesis

The remaining portion of the thesis is prepared as follows: Chapter two part I describes about the basic theories, concepts and different issues related with lung cancer, pulmonary nodules, CAD system, deep learning, CNN, image processing, etc. for better understanding of our research domain. Chapter two part II describes about related research works that has been done on pulmonary nodule detection and classification based on different approaches and techniques. Chapter three articulates a detailed description of method, architecture and design issues of our system. In Chapter four, the implementation of the proposed system architecture and experimental results are discussed. Finally, Chapter five discuss about the conclusion and future work of this study.

Chapter Two: Literature Review

Backgrounds and Theories

2.1 Introduction

In this topic we will discuss about theoretical backgrounds and current state of knowledge on topics related with our study. We try to survey, analyze and present papers published by accredited scholars and researchers focusing on lung cancer, pulmonary nodules, radiography, computed tomography, image processing, CAD system, and later in the chapter, fundamentals of machine learning methods, the theory behind the neural networks, and convolutional neural networks have been explained.

2.2 Lung Cancer

Lung cancer is the most common type of cancer [25], contributing with 13% to the total number of cancer cases (skin-cancer excluded). Compared with the three most common types of cancer, lung-, breast- and prostate cancers, the death rate and probability of dying is the highest with lung cancer [1]. Surgery, radiation therapy, and chemotherapy are used in the treatment of lung carcinoma [26]. In spite of that, the five-year survival rate for all stages combined is only 14% (i.e. Staging is a measure of how far the lung cancer has spread with in the lung tissue, there are five stages of lung cancer, stage I, II, IIIA, IIIB and IV [26]).

Early detection and resection of lung cancer can improve the diagnosis significantly, it is reported [27] that the survival rate for early-stage localized cancer (stage I) is 49%. The problem is that early curable lung cancers usually produce no symptoms and are often missed in mass screening chest radiography, which has been used for detecting lung cancer for a long time. CT is considered to be the most accurate imaging modality available for early detection and diagnosis of lung cancer. It allows detecting pathological deposits as small as 1mm in diameter. However, the large amount of data per examination makes the interpretation tedious and difficult, leading to both false positive and false negative results.

Lung cancer is caused by an abnormal (uncontrolled) cell growth within the lung tissues commonly known as “nodules”. The pulmonary nodules are radiologically visible as small structures that are roughly spherical opacities within the pulmonary interstitium images [28]. They have been regarded as indicators of primary lung cancer. Based on reliable detection of

lung nodules, radiologists and surgeons can perform size measurements and appearance characterizations for cancer malignancy diagnosis [29] and, if necessary, timely surgical intervention in order to increase the survival chances of patients [26]. *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 [30][31][32].

2.3 Pulmonary Nodules

Pulmonary nodule is a medical term to describe a picture on a chest x-ray or a CT images with a small spot in the lung and it is the most distinguishing feature of an early stage of lung cancer [33]. A solitary pulmonary nodule (parenchymal, non-pleural nodule) is a small, round or egg-shaped lesion in the lungs [34] or a solitary pulmonary nodule (SPN) is defined as a round opacity that is smaller than 3 cm. It may be solid or subsolid in attenuation. Semisolid nodules may have purely ground-glass attenuation or be partly solid (mixed solid and ground-glass attenuation)[35]. Juxtapleural pulmonary nodule is a small, worm-shaped lesion connected to pleura [34].

Human radiologists typically use low-dose CT (computed tomography) scans of patients' lungs to assess an individual's risk of lung cancer, usually by inspecting the images for the presence of nodules that are a common precursor to cancer [5]. 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 [35]. If the lesion is suspected to be benign, serial chest X-rays or CT scans may be taken on a regular basis for observation of the lesion. If the affected person is at high risk for lung cancer or if the CT scan appearance of the lesion suggests it is malignant, radiation therapy or surgical removal of the lesion is recommended [26].

Size: nodules are typically asymptomatic, and they are usually noticed by chance on a chest X-ray that has been done for another reason. They are usually smaller than 3–4 cm in diameter (no larger than 6 cm) and are always surrounded by normal, functioning lung tissue. Their intensity in CT scans is from -300 to 0 HU [36]. Malignancy of the nodule positively correlates with nodule diameter. As the diameter of a nodule increases, so does the likelihood of malignancy; however, a small nodule diameter does not exclude malignancy [29]. Small nodules (<4 mm) have a less than 1% chance of being a primary lung cancer, whereas the risk for malignancy

increases to 10%–20% for nodules in the range of 8 mm. Bigger, round and solid solitary nodules are easy to detect, while small, irregular nodules often fall into the false negative group. It was reported, that CAD system still has problems with detecting ground-glass nodules [37].

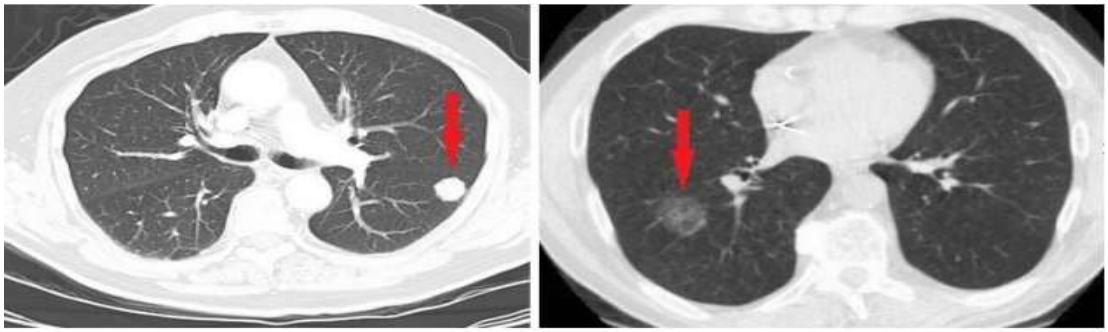


Figure 2.1: examples of solid (left) and ground-glass (right) parenchyma nodules [32]

Growth: 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 [38][32]. 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 [30][32].

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 [35]. But as we observed from different literatures, we are aware of 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.

Location: lung nodules in the right lung and in the upper lobes have a higher probability for malignancy [33]. 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 [31][39][32] 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 [38][39][32]. Calcified nodules detected at CT screening are considered by convention to be benign.

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 [30][39][32]. In case of family history when a cancer candidate's person family has nodules and is lung cancer, then the candidate has to be more likely to have cancerous nodules than those candidates without a family history.

2.4 Radiographic Testing (RT)

Radiography is used in a very wide range of applications including medicine, engineering, forensics, security, etc. Radiographic testing is method of inspecting materials for hidden flaw by using the ability of short wavelength electromagnetic radiation (high energy photons) to penetrate materials. [40] In radiography testing the test-part is placed between the radiation source and film (or detector). The material density and thickness differences of the test-part will attenuate (i.e. reduce) the penetrating radiation through interaction processes involving scattering and/or absorption. The differences in absorption are then recorded on film(s) or through an electronic means. In industrial radiography there are several imaging methods available, techniques to display the final image, i.e. Film Radiography, Real Time Radiography (RTR), Computed Tomography (CT), Digital Radiography (DR), and Computed Radiography (CR). CT is a radiographic based technique that provides both cross-sectional and 3D volume images of the object under inspection. These images allow the internal structure of the test object to be inspected without the inherent superimposition associated with 2D radiography. This feature allows detailed analysis of the internal structure of a wide range of components [40].

2.5 Computed Tomography (CT)

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 [41]. 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 [42][32]. CT machines collect X-ray projections to perform Radon transform [36]. From acquired data reconstruction is computed. The result is an output image showing the distribution of attenuation coefficient $\mu(x, y, z)$ in scanned object represented in HU [36].

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. Due to its ability to form 3-D images of the chest in greater resolution of nodules and tumor pathology, it is the preferred imaging modality for diagnosis of lung cancer. In addition, it can reveal the whole information (i.e. reveal small lesions in lungs that might not be detected on an X-ray) within the scans of the patient [43][32]. 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.

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[32][21]. 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 [32].

2.5.1 Hounsfield Units (HU)

The intensity of the image in medical X-ray imaging is measured in Hounsfield Units. The scale was established by Sir Godfrey Newbold Hounsfield, the principal engineer and developer of

computed tomography. The radio density of distilled water at standard temperature and pressure (STP) is defined as zero Hounsfield Units (HU). The radio density of air at STP is defined as -1000HU. Radio density in HU of air, common tissues and nodules is shown in Table 2.1

Substance	**HU**
Air	-1000
Nodules	-150
Fat	-180
Water	0
Muscle	40
Bone	1000

Table 2.1: materials and their radio density in HU [36]

2.6 CAD Systems

Computer aided detection (CAD) systems are applications that assist radiologists in the interpretations of medical images [44][45][32]. CAD is a technique that can help radiologists accurately interpret images and identify potential findings to avoid incorrect interpretation or overlooking of lesions due to subjective judgment. Images derived from computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), X-ray radiography, ultrasound imaging, fundus photography, etc., which may be in the form of 2-D or 3-D (sometimes 4D), can be analyzed using the CAD system [46]. It should be noted that the CAD system can only provide a second opinion and cannot replace radiologists; hence, the final diagnosis must be made by human beings [46]. However, it has the potential to assist radiologists for early diagnosis and treatment of diseases like lung cancer, by analyzing medical images generated from different imaging modalities.

CAD has now become one of the important research topics, especially in radiology, medical physics, and medical engineering. A desired CAD scheme may allow the detection of the location of abnormalities and allow the differentiation of disease categories [46]. A typical CAD setup consists of image preprocessing, candidate detection and false positive reduction as shown in Figure 2.2. The purpose of the preprocessing step is to unify the original images into a standard condition so that a computer algorithm can be developed such that it can be applicable to a wide range of input images with different qualities.

The typical preprocessing method includes extracting ROI region, smoothing or reducing noise, diminishing image size, etc.

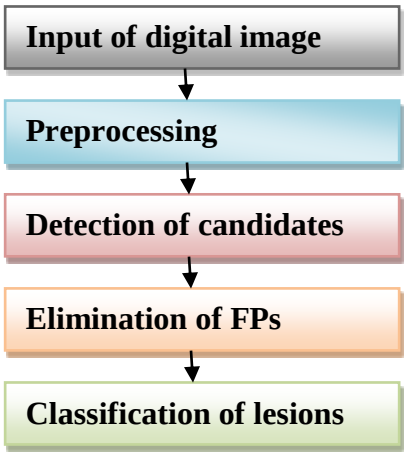


Figure 2.2: a typical flowchart of a CAD scheme [46]

Detecting abnormalities as true positives (TPs) may allow physicians to identify the locations of lesions, which can be determined on the basis of thresholding techniques, edge detection, or by using many other intelligent algorithms such as ANN, fuzzy, etc. However, the detection step always yields many false positives (FPs) that may mislead the diagnosis result. Eliminating FPs while maintaining the TP rate constant becomes an important component during programming; this is generally realized by calculating the parameters of candidates detected, for example, size, shape, intensity, circularity, etc. Finally, for a complete CAD scheme, classification of lesions into benign or malignant is expected [46]. In lung cancer, a CAD system helps to detect each individual nodule of each lung and determines which section is more to be malignant. 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 [47][42].

2.7 Image processing

Image processing is the way of manipulating images in various techniques in order to get easily visualized and detected images. Medical image processing, deals with the development of problem-specific approaches to the enhancement of raw medical image data for the purposes of selective visualization as well as further analysis [47]. Recently, image processing techniques are widely used in several medical areas for earlier detection and treatment of disease, especially in various cancer tumors such as lung cancer, breast cancer, etc. Figure 2.3 shows a general description of lung cancer detection system.

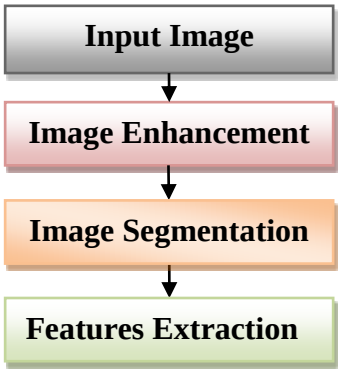


Figure 2.3: lung cancer image processing stages [50]

The first phase is we get CT scan image data. The second phase, we implement image enhancement to improve quality of image. The third phase is image segmentation which is an important step in the detection of cancer. The fourth stage is feature extraction that gives us a conclusion whether there is a lung cancer or not.

2.7.1 Image Enhancement (Preprocessing)

Image preprocessing is the technique of enhancing the image data prior to computational processing, enhancements include sharpening and color balancing. The main goal of preprocessing is to enrich the visual look of the images. It is used mainly to reduce the noise and unwanted artifacts in the image and misrepresentations of the image. The image Preprocessing stage starts with image enhancement; the aim of image enhancement is to improve the interpretability or perception of information included in the image for human viewers, or to provide better input for other automated image processing techniques [48][49][47][32]. Various techniques are used for image enhancement i.e. example Normalization.

2.7.1.1 Normalizing Image Inputs

Data normalization is an important step which ensures that each input parameter (pixels) has a similar distribution [50][51]. 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 (2.1), where $\boldsymbol{xmin}$ and $\boldsymbol{xmax}$ are the maximal and minimal values for the variable x data respectively. [32]

$$\text{Normalized} = \frac{(x - xmin)}{xmax - xmin} \qquad (2.1)$$

2.7.2 Image Segmentation

Image segmentation is an essential process for most image analysis and subsequent tasks. Many of the existing techniques for image description and recognition depend highly on the segmentation results. Segmentation divides the image into its constituent regions or objects. Segmentation of medical images in 2D, slice by slice has many useful applications for the medical professional such as: visualization and volume estimation of objects of interest, detection of abnormalities (e.g. tumors, polyps, etc.), tissue quantification and classification, and more. The goal of segmentation is to simplify and/or change the representation of the image into something that is more meaningful and easier to analyze.

Image segmentation is typically used to locate objects and boundaries (lines, curves, etc.) in images. More precisely, image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain visual characteristics. The result of image segmentation is a set of segments that collectively cover the entire image, or a set of contours extracted from the image (edge detection). All pixels in a given region are similar with respect to some characteristic or computed property, such as color, intensity, or texture. Adjacent regions are significantly different with respect to the same characteristic(s). Segmentation algorithms are based on one of two basic properties of intensity values: discontinuity and similarity. The first category is to partition the image based on abrupt changes in intensity, such as edges in an image. The second category is based on partitioning the image into regions that are similar according to a predefined criterion [48].

In lung cancer detection there are two types of segmentations, lung segmentation and nodule segmentation. First, the system isolates the lung tissue from the extraneous CT information through the segmentation process. This is carried out to reduce the computational complexity of the detection process by narrowing the region of interest to only the lung cavities. Second, The nodule detection phase passes potential nodule candidates to a second segmentation process in order to extract only those pixels belonging to the candidate nodule in question, by removing all other anatomies surrounding (and possibly attached) to the candidate nodule.

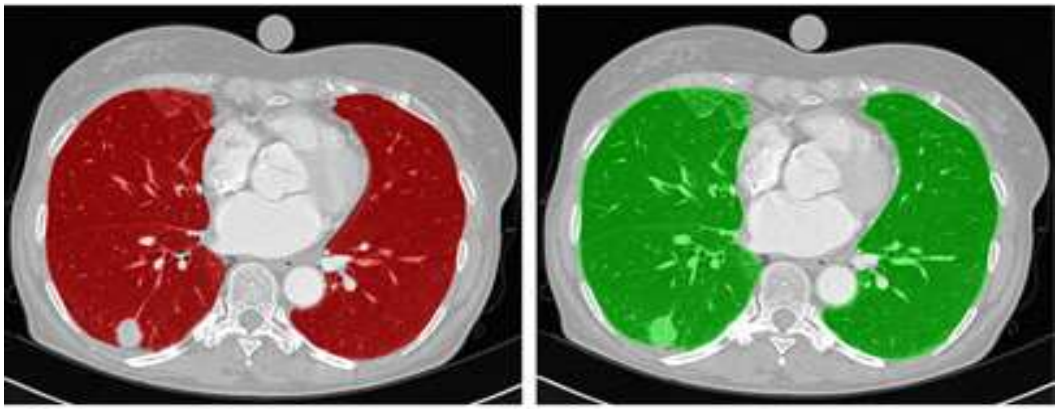


Figure 2.4: example of lung segmentation results [52]

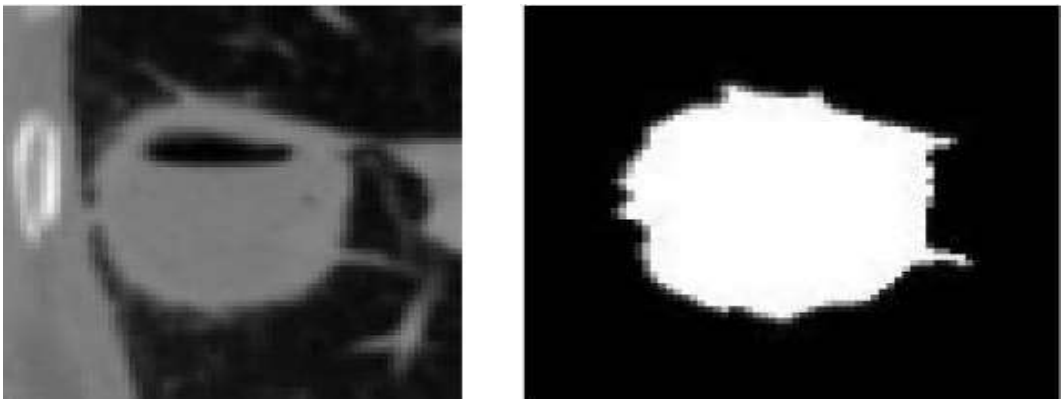


Figure 2.5: example of nodule segmentation result [53]

2.7.3 Features Extraction

Features Extraction is a process of representing raw image in a reduced form to facilitate decision making such as pattern detection, classification or recognition. It is the process of collecting discriminative information from a set of samples. Image features Extraction stage is an important stage that uses algorithms and techniques to detect and isolate various desired portions or shapes (features) of a given image [48]. Feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations. Feature extraction is a dimensionality reduction process, where an initial set of raw variables is reduced to more manageable groups (features) for processing, while still accurately and completely describing the original data set [54].

In convolutional neural network (CNNs) Feature extraction includes several convolution layers followed by max-pooling and an activation function. Convolution is the first layer to extract

features from an input image. Convolution preserves the relationship between pixels by learning image features using small squares of input data. It is a mathematical operation that takes two inputs such as image matrix and a filter or kernel [55]. CNNs are regularized versions of multilayer perceptrons. Multilayer perceptrons usually refer to fully connected networks, that is, each neuron in one layer is connected to all neurons in the next layer [54]. We will see the detail about CNNs in section 2.6.

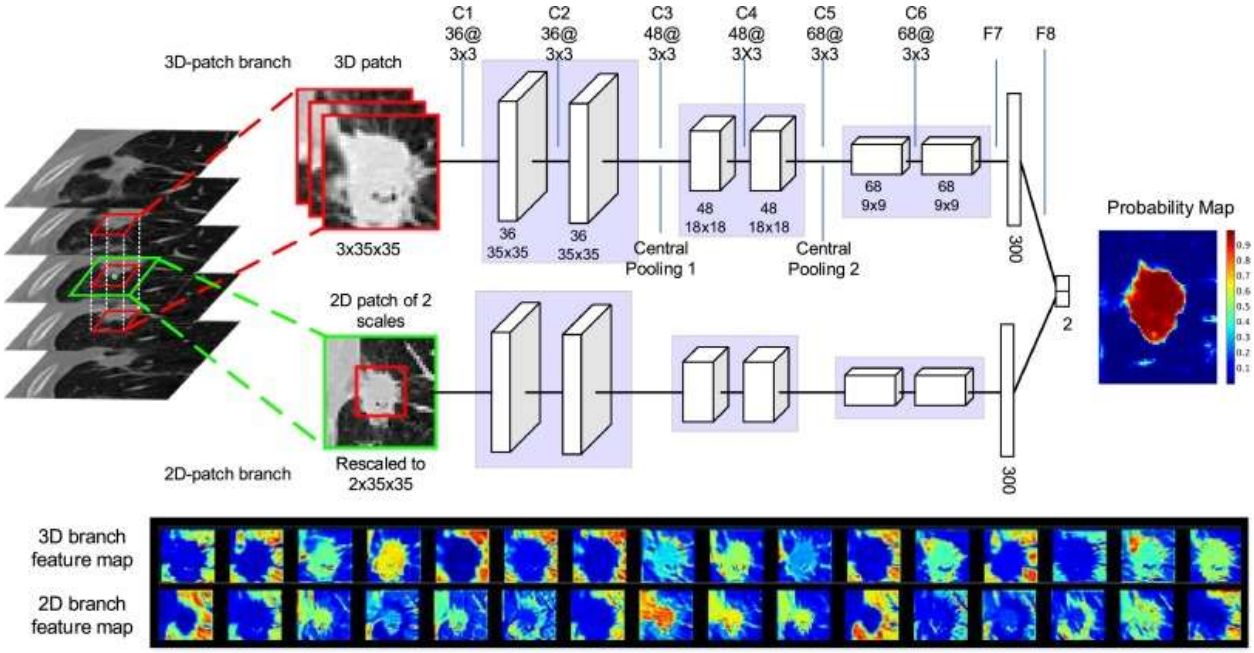


Figure 2.6: 2D and 3D Feature Extraction using CNN [56]

2.8 Machine Learning

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model of sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task [54]. Machine Learning (ML) and Artificial Intelligence (AI) have progressed rapidly in recent years. Techniques of ML have played important role in medical field like medical image processing, computer-aided diagnosis, image interpretation, image fusion, image registration, image segmentation, image-guided therapy, image retrieval and analysis [7]. There are three types of machine learning algorithms [54][32]: supervised learning, unsupervised learning and semi-supervised learning.

Supervised learning, the model is learned from the input and the expected output data. These are the most common way of learning. It uses labeled training data to learn the mapping function from the input variables to the output variables. Examples: Neural Networks (MLP), Logistic Regression, SVM, K-Nearest neighbors, Linear Regression, Decision Tree etc. 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. Examples: K-Means clustering, hierarchical clustering, mixture models, etc. 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.

Deep learning (also known as deep structured learning or hierarchical learning) is part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms. Learning can be supervised, semi-supervised or unsupervised [54]. In deep learning, a computer model learns to perform classification tasks directly from images, text, or sound. Deep learning models can achieve state-of-the-art accuracy, sometimes exceeding human-level performance. Deep learning Models are trained by using a large set of labeled data and neural network architectures that contain many layers. Thus, it plays a major role in computer vision and medical imaging. In fact, similar impact is happening in domains like text, voice, etc. Various types of deep learning algorithms are in use in researches like Convolutional Neural Networks (CNN), Deep Neural Networks (DNN), Deep Belief Network (DBN), Deep Auto-encoder (DA), Deep Boltzmann Machine (DBM), Deep Conventional Extreme Machine Learning (DC-ELM), Recurrent Neural Network (RNN) etc.[7].

2.9 The Theory Behind Artificial Neural Networks (ANNs)

Artificial neural networks structurally and conceptually inspired by human biological nervous system . ANNs consists of interconnected neurons that takes input and perform some

processing on the input data, and finally forward an output from the current layer to the coming layer. Each neuron in the network sums up the input data and apply the activation function to the summed data and finally provides the output that might be propagated to the next layer. ANNs are powerful in machine learning field. However, numbers of neurons in deep neural network systems are still not comparable to the number of neurons in human. One of the most complex neural network architectures (i.e. GoogLeNet) has nearly 6.7 million parameters [57].

In 1958, Rosenblatt [58] generated one of the earliest types of artificial neural networks known as perceptron. It is one of the earliest neural networks based on human brain system. It consists of input layer that is directly connect to output layer and was good to classify linearly separable patterns. It is a simple mathematical model of a biological neuron, whose output can be given as:

$$F(x) = \begin{cases} 1 & \text{if, } w.x + b > 0 \\ 0 & \text{otherwise} \end{cases} \qquad (2.2)$$

Where, $F(x)$ is the output of the neuron, w is a vector of real-valued weights, w. x is the dot product of the weight vector w and the vector x, and b is the bias (i.e. a neuron added to each pre-output layer that stores the value of 1), bias units aren't connected to any previous layer and in this sense don't represent a true "activity". Perceptron is accepted as one of the first artificial neural networks to be produced.

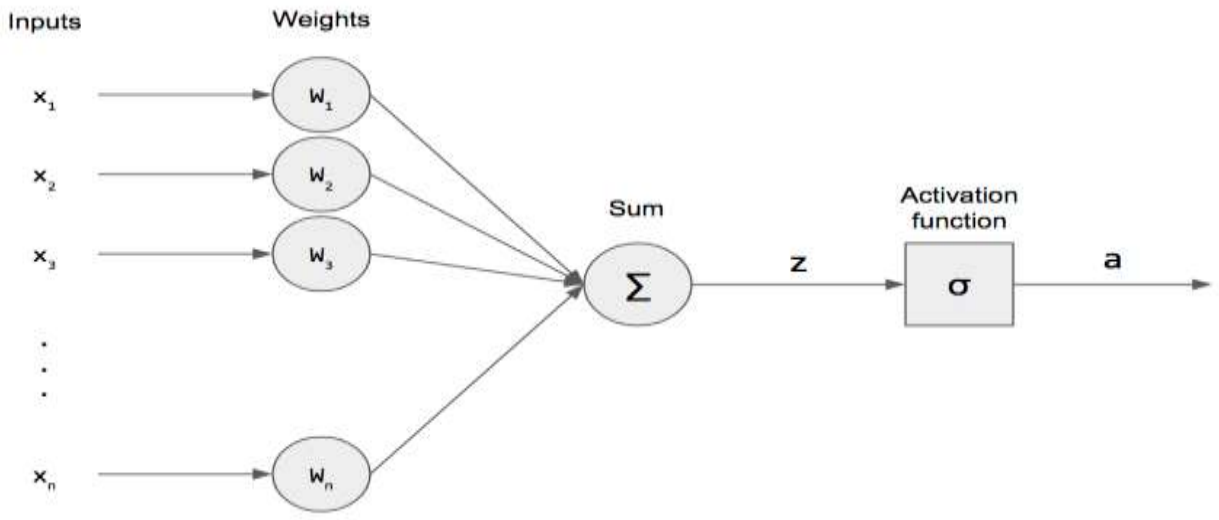


Figure 2.7: perceptron algorithm

To solve complex patterns, neural networks with a layered architecture were introduced known as Deep Neural Networks (i.e. Input layer, output layer and one or more hidden layers) [7].

Multilayer perceptron (MLP) is a kind of feed forward ANN consists of at least three layers of nodes (Figure 2.8). The first layer is called the input layer and the last layer is called the output layer. Middle layers are called the hidden layers. Due to its hidden layers, MLP can distinguish data that is not linearly separable. In a MLP system, the number of input and output nodes is determined according to the data. For example, in order to design a network architecture for handwritten digit recognition where numbers are stored in 28x28 size images, there will be 784 nodes (one input node for one pixel, 28x28 = 784) in the input layer and 10 nodes (one node for the each number) in the output nodes. Figure 2.8 shows a simple 3-layer neural network architecture and a deep neural network (1- input, 3-hidden, 1-output).

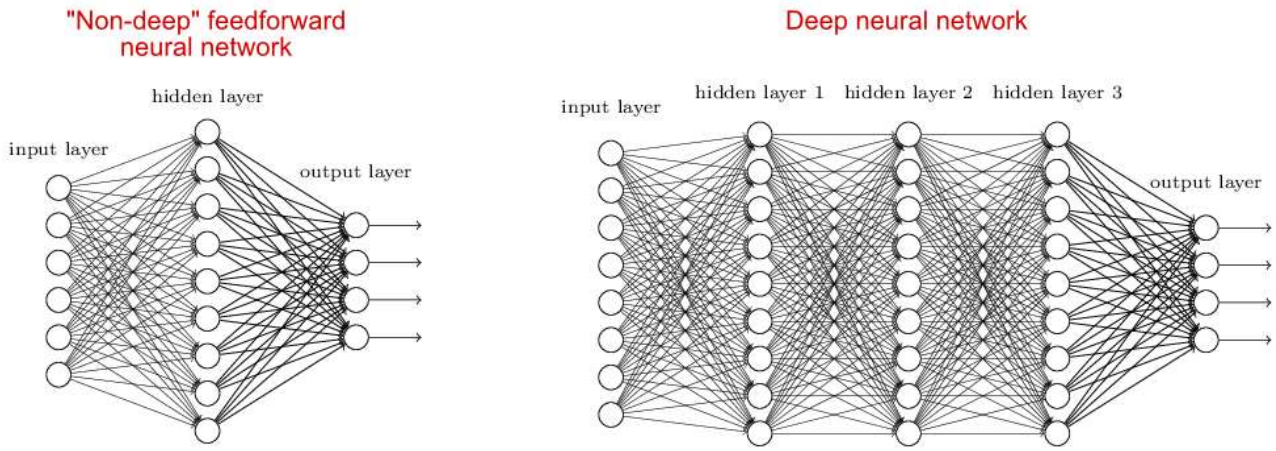


Figure 2.8: a simple and deep neural networks

Determining the number of hidden layers and the number of nodes in the hidden layer is a design issue during the training process. Too many nodes will make training longer and the network may lose its generalization ability. On the other hand, with too few nodes, the network uses too little information and may not solve the complex models.

Activation functions are non-linear complex functional mappings between the inputs and response variable. They introduce non-linear properties to our Network. Their main purpose is to convert an input signal of a node in ANN to an output signal. That output signal now is used as an input in the next layer in the stack [59]. In artificial neural networks, the activation function of a node defines the output of that node given an input or set of inputs [54]. The three commonly used activation functions in neural network are sigmoid (Eqn. 2.3 or Figure 2.9), tanh (Eqn. 2.4 or Figure 2.10) and rectified linear unit (Eqn. 2.5 or Figure 2.11).

Sigmoid (Logestic) is between 0 and 1. It is easy to understand and apply but it has major reasons which have made it fall out of popularity: vanishing gradient problem, 0 < output < 1 makes optimization harder, sigmoids saturate and kill gradients, have slow convergence. Tanh (hyperbolic tangent function), its range is between -1 and 1 i.e. -1 < output < 1. Hence optimization is easier in this method. It is always preferred over sigmoid function. But still it suffers from vanishing gradient problem. ReLU, it has become very popular in the past couple of years. It was recently proved that it had 6 times improvement in convergence from tanh function. Nowadays, almost all deep learning models use ReLU because it avoids and rectifies vanishing gradient problem [59].

$$f(x) = \frac{1}{1 + e^{-x}} \qquad (2.3)$$

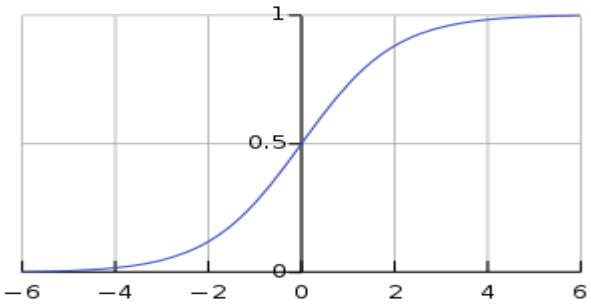


Figure 2.9: graph of sigmoid activation function

$$f(x) = \frac{2}{1 + e^{-2x}} - 1 \qquad (2.4)$$

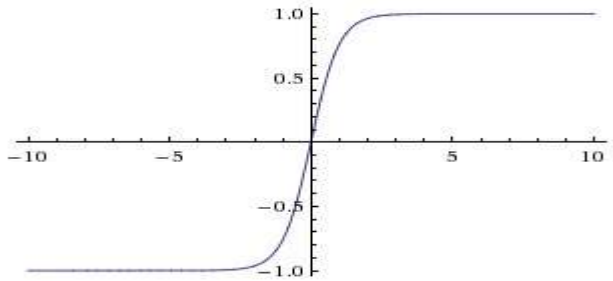


Figure 2.10:graph of tanh function

$$f(x) = \begin{cases} x & if, x \geq 0 \\ 0 & x < 0 \end{cases} \qquad (2.5)$$

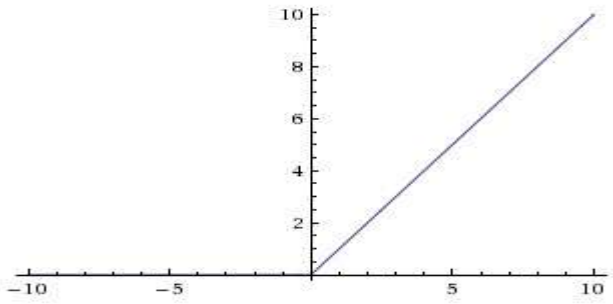


Figure 2.11:graph of ReLU function

Backpropagation is a supervised learning algorithm, for training Multi-layer Perceptrons (Artificial Neural Networks). The Backpropagation algorithm looks for the minimum value of the error function in weight space using a technique called the delta rule or gradient descent. The weights that minimize the error function are then considered to be a solution to the learning problem [58]. Backpropagation algorithms are a family of methods used to efficiently train artificial neural networks (ANNs) following a gradient descent approach that exploits the chain rule. The main feature of backpropagation is its iterative, recursive and efficient method for calculating the weights updates to improve in the network until it is able to perform the task for which it is being trained [54]. Steps involved in backpropagation are: Step – 1: Forward Propagation, Step – 2: Backward Propagation, Step – 3: Putting all the values together and calculating the updated weight value. The algorithm works as follows [58]:

- First we initialize some random value (weight) and propagate forward.
- There will be some error i.e. for each sample we calculate the output value of the nodes (feed-forward procedure)
- We propagate backwards to update weights using backpropagation
- Repeat step two and three until convergence

In the feed-forward step, nodes are multiplied with their corresponding weights and results are summed. Resulting summation is processed in a nonlinear activation function. The output of the activation function becomes the value of that node (Figure 2.12). Starting from the input nodes, this process is applied to all nodes until node values of the output layer are calculated.

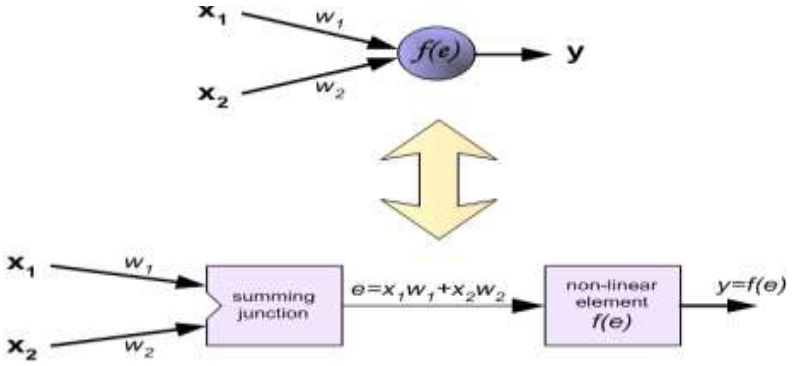


Figure 2.12: feed-forward process

Backward Step: after feedforward we calculate the error (i.e. the difference between the real output values and the calculated output values) to update the weights. Error is sometimes referred as loss or cost. This algorithm mainly looks for how much a specific weight is

responsible for the overall error. In order to get this information, derivative of total error with respect to the weights of the network must be calculated, and weights must be updated according to this information. For the Gradient Descent algorithm, updating the weights is simply achieved by the following formula:

$$w_{ij}^{+} = w_{ij} - \eta \frac{\partial E}{\partial w_{ij}} \qquad (2.6)$$

Where, E is the error, w_{ij} the current weight, η learning rate coefficient and w_{ij}^{+} is the new weight.

2.10 Convolutional Neural Network (CNN)

Convolutional neural network (ConvNets or CNNs) is one of the main categories of deep learning algorithm to do images recognition, images classification, objects detections, face recognition etc. CNN image classification takes an input image, process it and classify it under certain categories (E.g. Nodule or Non-nodule). CNNs take an input image as array of pixels and it depends on the image resolution. In CNN, raw data is represented as tensor. Tensor concept can be generalized as higher order matrices. Based on the image resolution, it will see H × W × D (H = Height, W = Width, D = Dimension). E.g., an image of 6 × 6 × 3 array of matrix of RGB (3 refers to RGB values).

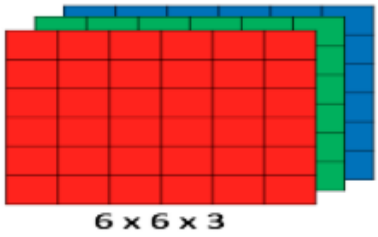


Figure 2.13: array of RGB Matrix [55]

Technically, deep learning CNN models to train and test, each input image will pass it through a series of convolution layers with filters (Kernels), Pooling layer, fully connected layers (FC) and apply softmax function to classify an object with probabilistic values between 0 and 1. CNNs extract the specific patterns by using the filters, and then pooling layers help the model ignore redundant data. Some mostly used CNN patterns can be given as follows:

- INPUT → [C → ReLU → P]^{N}→ FC
- INPUT → [C → C → P → C → C → P]^{N}→ FC

- INPUT → [[C → ReLU]^{N} → P]^{M} → FC

Figure 2.14 shows a complete flow of CNN to process an input image and classifies the objects based on values.

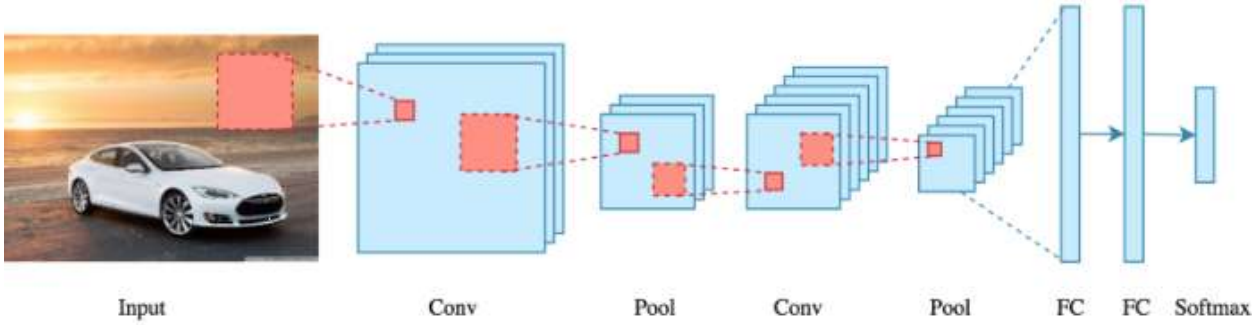


Figure 2.14: patterns involved in convolutional neural network

2.10.1 Convolution Layer

Convolution is the first layer of CNNs to extract features from an input image. Convolution preserves the relationship between pixels by learning image features using small squares of input data. It is a mathematical operation that takes two inputs such as image matrix and a filter or kernel. If the input is a 3 channel RGB image, the kernel is selected as a 3-dimensional structure. If the kernel size is very small (e.g. 2 × 2), it will not be able to extract enough features. For example, small kernels cannot detect big complex patterns. On the other hand, bigger kernel size increases computation complexity. Generally, most of the time small kernel sizes such as 3 × 3 or 5 × 5 are used in the CNN training. Strides are the number of pixels shifts over the input matrix. When the stride is 1 then we move the filters to 1 pixel at a time, when the stride is 2 then we move the filters to 2 pixels at a time through the input matrix and so on. Consider a 5 × 5 input matrix whose image pixel values are 0, 1 and filter matrix 3 × 3 as shown in figure 2.15.

Figure 2.15: image matrix multiplies kernel or filter matrix [55]

Then the convolution operation of 5 × 5 image matrix multiplies with 3 × 3 filter matrix with a stride 1 × 1 generates a feature map output as shown in figure 2.16.

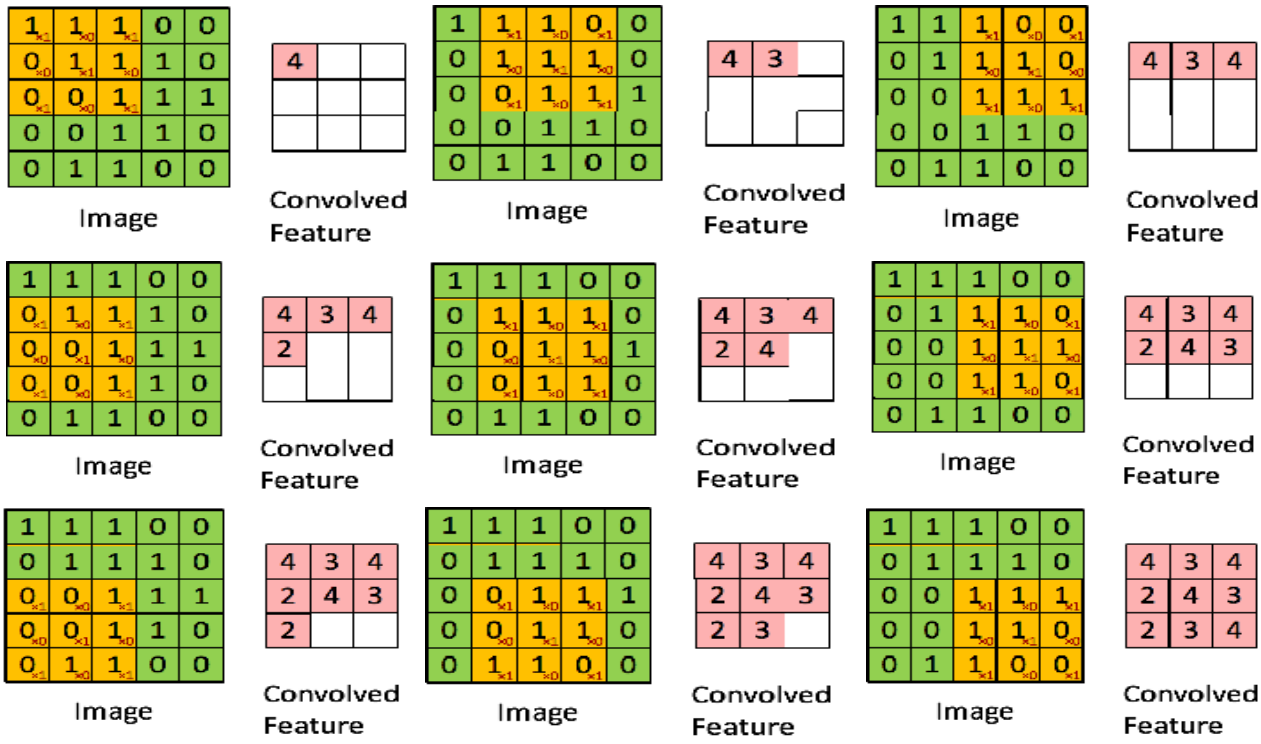


Figure 2.16: an output feature maps for 5 × 5 input convolved with a 3 × 3 kernel [55]

Number of kernels is a very important design issue because it determines the number of different features to be focused on. If the number of kernels is small, the network may miss some of the patterns in the data. On the other hand, the number of kernels should not be big enough to cause duplicate filters. For example, if the kernel size is 3×3, using 64 kernels will create duplicate filters because 3×3 sizes cannot create 64 qualified different filters. In addition to duplicate filters, too many filters will bring memory issues because each convolved image takes up space in the memory of the computer.

When each kernel is slid over the input image, they generate a feature map. These output images are concatenated after all kernels have produced their outputs. If the input image is 2-dimensional, its outputs will be 3-dimensional tensor. If the input image is 3-dimensional volumetric data, its outputs will be 4-dimensional tensor. This extra dimension comes from using many filters. Depth of the tensor and depth of the kernel must be same. For example, consider an input tensor of a size 30×30, and a kernel of a size 3×3. The resulting size of the convolution is 28 × 28. We may have more than one kernel, which are applied on the input tensor. As a result, size of the output tensor becomes K×28×28. For the next convolution operation, kernel dimension becomes K × N × M. In CNN model representations; this size is generally represented as K@N×M.

Padding: sometimes filter does not fit perfectly with the input image. We have two options:

1) Pad the picture with zeros (zero-padding) so that it fits
2) Drop the part of the image where the filter did not fit. This is called valid padding which keeps only valid part of the image.

The equation for calculating the output size for any given convolutional layer is:

$$O = \frac{W - K + 2P}{S} + 1 \qquad (2.7)$$

Where, O is the output height/length, W is the input height/length, K is the filter size, P is the padding, and S is the stride. Figure 2.17 shows the filters after the training of a CNN created by Krizhevsky et al. [60]. Shapes and colors in the filters were formed iteratively during the training process.

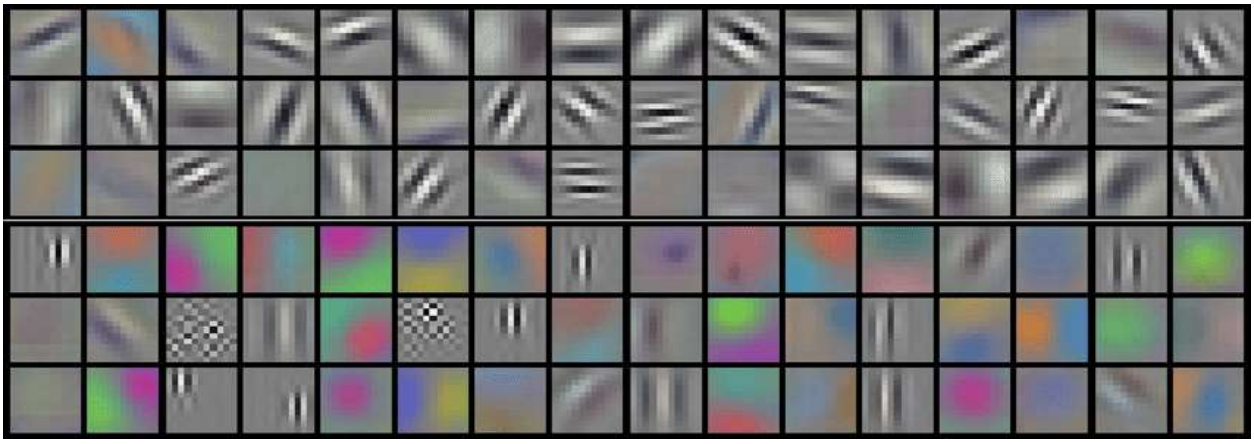


Figure 2.17: example filters learned by Krizhevsky et al. [60], filter size is 11x11x3 pixel.

A 3D ConvNet is a variant of 2D CNN to find patterns across 3 spatial dimensions; i.e. depth, height and width. It is effective in object segmentation and classification of 3D medical imaging.

2.10.2 Pooling Layer

Pooling layer reduce the number of parameters when the images are too large. Spatial pooling (down-sampling or sub-sampling) reduces the dimensionality of each map but retains the important information. Spatial pooling can be of different types: Max-pooling, Average-pooling, Sum-pooling. Max pooling take the largest element from the rectified feature map. Average pooling takes average of rectified feature map elements. Sum pooling takes sum of all elements in the feature map. For pooling operation, there are two important hyper-parameters which are window size and stride value.

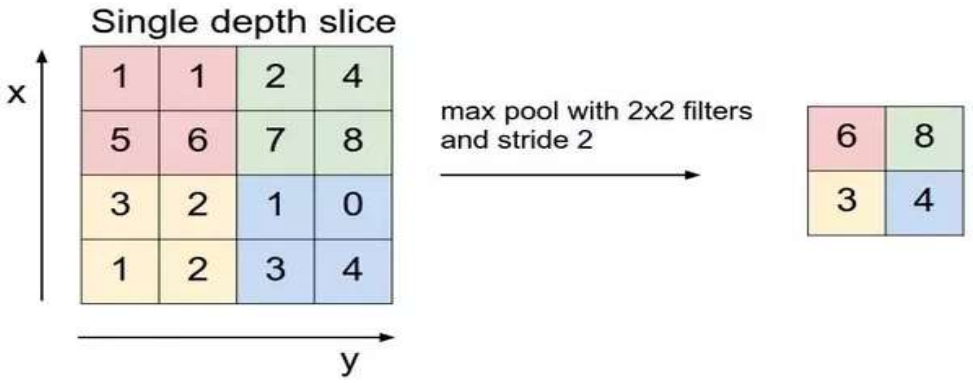


Figure 2.18: example of Max-pooling [55]

2.10.3 Fully Connected Layer

The layer we call as FC layer, we flattened our matrix into vector and feed it into a fully connected layer like neural network. Convolution and pooling layer generate rectangular shaped outputs. These outputs are converted to vector format so that they can be multiplied by the weight matrix. For example, if there are 64 feature map layers each of which has 5×5×3 voxels, in the fully connected layer these volumes are converted to a 4800×1 vector (5×5×3×64 = 4800). The layer before the fully connected layer represents high-level features. With the help of a fully connected layer, these high-level features can be multiplied by the weights of the hidden layers the remaining part of the system works like MLP do.

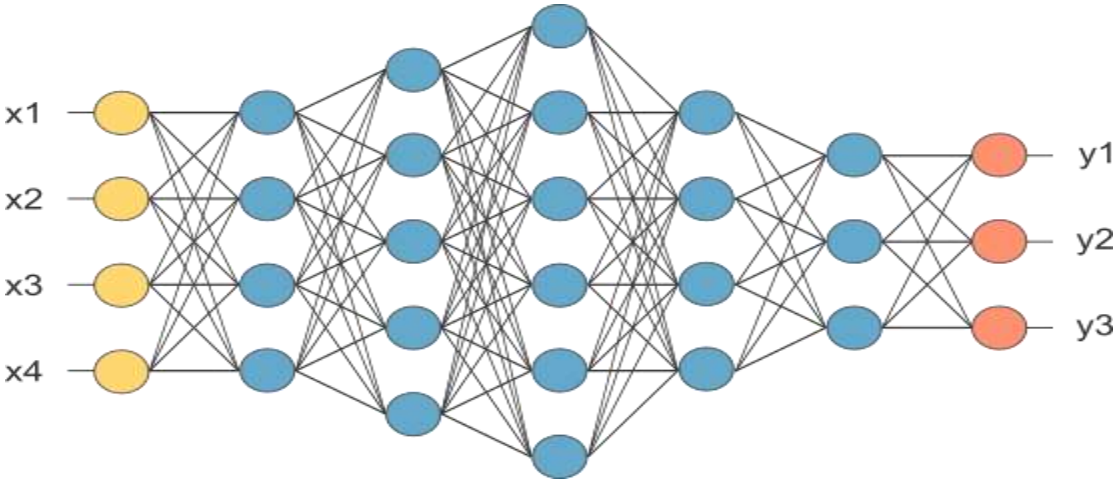


Figure 2.19: after pooling layer, flattened as FC layer [55]

In figure 2.19, feature map matrix will be converted as vector (x1, x2, x3,…). With the fully connected layers, we combined these features together to create a model. Finally, we have an activation function such as softmax or sigmoid to classify the outputs as nodule or non-nodule.

2.10.4 Softmax Classifier

The Softmax regression is a form of logistic regression that normalizes an input value into a vector of values that follows a probability distribution whose total sums up to 1. The output values are between the range [0, 1] which is nice because we are able to accommodate many classes or dimensions in our neural network model. The function is usually used to compute losses that can be expected when training a data set. It should be noted that softmax is not ideally used as an activation function like Sigmoid or ReLU (Rectified Linear Units). The classifier performs its function in convolutional neural network by using the following formula:

$$p_c(h^l) = \frac{\exp\ (h_c^l)}{\sum_{c=0}^{C-1} \exp(h_c^l)} \qquad (2.8)$$

Where, h^l is the neuron vector in the last layer C is the number of target classes in the last layer p_c prediction probability for each class **c** h_c^l the **c** th element of the neuron vector.

2.11 Cross Validation

Cross validation is a technique to evaluate predictive models by partitioning the original sample into training set to train the model, and testing set to evaluate it. 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 examine our models by using ten-fold cross validation technique. CNN is trained the same way like ANN, backpropagation with gradient descent. Due to the convolution operation mathematically it's more complex.

2.12 Summary

Lung cancer is the leading cause of human mortality worldwide compared with other cancer related diseases. It is caused by malignant nodules in the lung tissue. 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. A CT scan is a medical imaging method used to predict lung nodule malignancy in patients to make noninvasive early diagnosis and treatment of lung cancer. A CT scan generate a large amount of CT images that has to be analyzed by radiologists with the help of CAD (computer-aided diagnosis) systems.

CAD is a technique that can help radiologists accurately interpret images and identify potential findings to avoid incorrect interpretation or overlooking of lesions due to subjective judgment. A typical CAD setup consists of image preprocessing, candidate detection and false positive reduction. Image processing techniques are widely used in several medical areas for earlier detection and treatment of disease, especially in various cancer tumors such as lung cancer. Lung cancer detection can be classified into four stages. The first phase is we get CT scan image data. The second phase, we implement image enhancement to improve quality of image. The third phase is image segmentation which is an important step in the detection of cancer. The fourth stage is feature extraction that gives us a conclusion whether there is a lung cancer or not.

Machine Learning (ML) have progressed rapidly and have played an important role in medical field. Artificial neural networks (ANNs) is a supervised learning algorithm structurally and conceptually inspired by human biological nervous system. Multilayer perceptron (MLP) is a kind of feed forward ANN consists of at least one hidden layer. Deep learning Models are trained by using a large set of labelled data and neural network architectures that contain many layers. CNN is a deep learning algorithm that is capable and successful in the representation of the high-level features which are learned from large amounts of training data. A 3D ConvNet is a variant of 2D CNN to find patterns across 3 spatial dimensions; i.e. depth, height and width. It is successful in medical image classification.

Generally, in this chapter we tried to show a general overview regarding to our study specially focusing on topics related with lung cancer, pulmonary nodule, CT scan, CAD system, image processing, deep learning, etc. A fundamental working principle of CNN is also described in detail to give an initial background about our model.

Related Works

2.13 Introduction of Related Literatures

Many researchers have devoted a lot of effort intending to develop fast and efficient CAD systems. Several systems have been developed for accurate detection and classification of pulmonary nodules. In candidate classification, researchers concern up on improving the system by increasing sensitivity with a minimum of false positive rate. Sensitivity and false positive rate are the most important parameters that are commonly used in the literatures. Sensitivity, which is also called true positive rate is the ratio of correctly identified true positives with respect to all positive candidates (true positive and false positive). False positives are number of candidates incorrectly discriminate as nodules but in fact they are not true nodules. In this section we will try to survey, discuss and analyse different researches in automated lung nodule detection focusing on their methods, datasets, algorithms, and results.

2.14 Lung Nodule Detection and Classification

Messay et al. [14] designed a complete CAD system for the detection of pulmonary nodules in thoracic computed tomography (CT) imagery. Training and tuning of all modules in this system is done using dataset provided by University of Texas Medical Branch (UTMB), but testing process performed on a publicly available dataset (LIDC) comprised of 84 CT scans containing 143 nodules ranging from 3 to 30 mm. The paper describes the complete design of the CAD system and presents a detailed performance analysis on the publicly available LIDC database. The CAD system uses intensity thresholding with morphological processing to detect and segment nodule candidates simultaneously and able to detect 92.8% of all the nodules in the LIDC testing dataset (merged ground truth i.e.an average of 517.5 candidates per case/scan). For classification 245 features are computed for each segmented nodule candidates based on shape, position, intensity and gradient features from segmented nodule candidates. They have used Fisher Linear Discriminant (FLD) classifier for the CAD system. By using a 7-fold cross-validation performance analysis in the LIDC database the system achieved sensitivity of 82.66% with an average of 3 FPs/scans

Jacobs et al. [3] designed a fully automatic CAD system for detection of sub-solid nodules. The dataset was collected from the NELSON trial, a large multi center lung cancer screening trial, organized in the Netherlands and Belgium (van Klaveren et al., 2009). In total, the data set

consisted of around 20,000 scans from around 4500 subjects. All sub-solid nodule annotations with a diameter smaller than 5 mm were discarded. 128 features created for subsolid nodule candidates by using intensity (used normalized histogram), shape (calculate sphericity, compactness1, compactness2 and guess Radius.), texture (by using local binary patterns (LBP) and 2D Haar wavelets) and context features (airways, vessels and other nodule candidates). The candidate detection achieved sensitivity of 84% and 88% for training and testing sets respectively for all subsolid nodules. During classification to select the best classification scheme, several classification experiments are conducted. a k-nearest neighbor classifier (kNN), random forest classifier (RF), Gentle Boost classifier (GB), nearest mean classifier (NM), support vector machine using radial basis function kernel (SVM-RBF), and Linear Discriminant Classifier (LDC) are tested in order to find the optimal classifier for this classification task. Experiments are performed in 10-fold cross-validation and achieved a sensitivity of 80% at an average of 1.0 false positive per scan.

Arnold. A. A. Setio et al. [15] proposed detection pipeline initiated by a three-dimensional lung segmentation algorithm optimized to include large nodules attached to the pleural wall via morphological processing. Performance was evaluated using ten-fold cross-validation on the full publicly available lung image database consortium database LIDC–IDRI database. They perform preprocessing to mask structures outside the pleural space to ensure that pleural and parenchymal nodules have a similar appearance. Nodule candidates are obtained via a multistage process of thresholding and morphological operations, to detect both larger and smaller candidates. The proposed system achieves 99.2% (236/238) sensitivity for large solid nodules (>10mm and annotated by at least 3 doctors). After segmenting each candidate, a set of 24 features based on intensity, shape, blobness, and spatial context are computed. A radial basis support vector machine (SVM) classifier was used to classify nodule candidates. The CAD system reaches a sensitivity of 98.3% (234/238) and 94.1% (224/238) large nodules at an average of 4.0 and 1.0 false positives/scan, respectively (>10mm and annotated by at least 3 doctors).

Armato et al. [9] developed one of the first fully automated computerized methods for the detection of lung nodules from CT scans. Their method was based on 2-dimensional and 3-dimensional analysis of the image data. They applied multiple gray-level thresholds in order to segment the volume. An 18-point connectivity scheme was used to identify contiguous three

dimensional structures. Morphological and gray-level features were calculated for each candidate. A rule-based approach that checks the shape of candidates was applied in order to reduce the false-positives. In the final step, they applied linear discriminant analysis (LDA). The authors reported that this automated method yielded an overall nodule detection sensitivity of 70% with an average of 1.5 false-positives per scan when applied to the 43 cases. When this method was applied to 20 cases, which contained only one or two nodules per case, a corresponding detection sensitivity of 89% with 1.3 false-positives per scan was achieved.

K. Murphy et al. [16] proposed an algorithm that uses the local image features of shape index and curvedness to detect candidate structures in the lung volume and applies two successive k-nearest-neighbour classifiers in the reduction of false-positives. The nodule detection system is trained and tested on three databases extracted from a large-scale experimental screening study. In a random selection of 813 scans from the screening study a sensitivity of 80% with an average 4.2 false-positives per scan is achieved. The detection results presented are a realistic measure of a CAD system performance in a low-dose screening study which includes a diverse array of nodules of many varying sizes, types and textures.

Ross G. et al. [33], attempt to design and develop a deep learning based algorithm for the classification of potentially malignant pulmonary nodules. The dataset used for training was obtained from the public Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI). an open source deep neural network solver from Berkley Vision and Learning Center, was used in the study to model DNNs. The best performing architecture was comprised of five convolution layers (ten layers total). This model achieved an accuracy of 82.1% with a sensitivity of 78.2% and a specificity of 86.1%.

Setio et al. [19] proposed to employ 2D multi-view convolutional networks to learn representative features for pulmonary nodule detection. This method can incorporate relatively wide volumetric spatial information for detection by extracting many 2D patches from differently oriented planes. This method achieved a state of-the-art detection sensitivity of 85.4% at 1.0 false positive per scan on the benchmark of LIDC-IDRI [24] dataset. The network is fed with nodule candidates obtained by combining three candidate detectors specifically designed for solid, subsolid, and large nodules. For each candidate, a set of 2-D patches from differently oriented planes is extracted. The proposed architecture comprises multiple streams of 2-D

ConvNets, for which the outputs are combined using a dedicated fusion method to get the final classification. Data augmentation and dropout are applied to avoid overfitting. On 888 scans of the publicly available LIDC-IDRI dataset, they achieved high detection sensitivities of 85.4% and 90.1% at 1 and 4 false positives per scan, respectively. An additional evaluation on independent datasets from the ANODE09 challenge and DLCST is performed.

Qi Dou et al. [17] propose a method employing 3D convolutional neural networks (CNNs) for false positive reduction track of LUNA-16 challenge. The proposed framework has been validated by the dataset of LUNA16 challenge held in conjunction with ISBI 2016. They develop three 3D convolutional networks, each encoding a specific level of contextual information. The final classification results are obtained by fusing the probability prediction outputs of these three networks. The challenge evaluation scheme includes several false positive rates to increase the difficulty of the challenge (0.125, 0.25, 0.5, 1, 2, 4, 8 false positives per scan), which is of significance as it determines if a system can identify an acceptable percentage of modules with very few false positives, and hence increase the automation level of current computer-assisted diagnosis systems. On 888 scans of the publicly available LUNA-ISBI 2016 dataset, they achieved high detection sensitivities of 67.7 % and 92.2% at 0.125 and 8 false positives per scan, respectively as shown in the table below.

Anton D et al. [61] proposed a method aimed to recognize real pulmonary nodule among a large group of candidates. This method consists of three steps: appropriate receptive field selection, feature extraction and a strategy for high level feature fusion and classification. The dataset consists of 888 patient's chest volume low dose computer tomography (LDCT) scans, selected from publicly available LIDC-IDRI dataset. This dataset was marked by LUNA16 challenge organizers resulting in 1186 nodules. Trivial data augmentation and dropout were applied in order to avoid overfitting. The proposed system have achieved high competition performance metric (CPM) of 0.735 and sensitivities of 78.8% and 83.9% at 1 and 4 false positives per scan, respectively.

Teramoto and Atsushi [43], developed a scheme to detect pulmonary nodules using both CT and PET images. In the CT images, a massive region is first detected using an active contour filter, which is a type of contrast enhancement filter that has a deformable kernel shape. Subsequently, high-uptake regions detected by the PET images are merged with the regions detected by the CT

images. FP candidates are eliminated using an ensemble method; it consists of two feature extractions, one by shape/metabolic feature analysis and the other by a CNN, followed by a two-step classifier, one step being rule based and the other being based on support vector machines. The authors evaluated the detection performance using 104 PET/CT images collected by a cancer-screening program. The sensitivity in detecting candidates at an initial stage was 97.2%, with 72.8 FPs/case. After performing the proposed FP-reduction method, the sensitivity of detection was 90.1%, with 4.9 FPs/case.

2.15 Summary of Related Works

Many researchers attempt to increase the detection accuracy of pulmonary nodules by performing different image processing techniques and deep learning approaches. Some researchers designed their system on the basis of image processing techniques based on shape, position, intensity, texture characteristics and gradient features etc. [3][14]. Others try to design their model by using machine learning approaches for classification, such as support vector machine (SVM) classifier, successive k-nearest-neighbour classifiers, etc [15][16]. Recently, some researchers tried to propose their system on the basis of deep learning algorithms, such as CNNs using different configuration and fusion methods both in 2D and 3D.

Generally, number of researches has been done in the detection and classification of candidate pulmonary nodules, however accurate detection of nodules is still a challenging task. Many researchers have achieved high sensitivity levels by using different image processing techniques but still with many false positive results. Recently, with the remarkable successes of deep convolutional neural networks (CNNs) in image and video processing, the representation capability of the high-level features which are learned from large amounts of training data has been broadly recognized. This also inspired some researchers to employ CNNs in automated pulmonary nodule detection [17][18][19]. According to recent papers[19][17][61][33], detection of lung nodules using deep learning algorithms can be able to achieve a promising result to obtain a high sensitivity rate by reducing a large number of false positive candidates. So, we inspired to design our model on top of CNNs, which is a deep learning algorithm that achieved successful results in medical image recognition and classification.

Chapter Three: The Proposed Methods and Approaches

3.1 Introduction

As we discuss in chapter three many researchers have strived to boost the detection and classification capability of the existing CAD system, they have used different approaches and techniques. In our work we proposed to design our model on top of CNN by employing the dataset obtained from LUNA 16 challenge providers. Since we are using CNN architecture (i.e. most popular deep learning algorithm) using LUNA16 Challenge dataset is very important and suitable, because it consists of many CT images with labelled candidate points that make the dataset more suitable for supervised learning algorithms. In this chapter we try to present our architecture for candidate pulmonary nodule classification by describing the detail techniques used for data processing, parameter definition and component configuration of the proposed system. In section 3.2 we try to show the general overview of the proposed system architecture; section 3.3 discuss about a variety of data pre-processing techniques, section 3.4 presents the details of the proposed CNN model; section 3.5 deals about the training process; section 3.6, talks about fusion method and finally in section 3.7 we tried to discuss about the evaluation metrics used to assess system result.

3.2 The Proposed System Architecture

The general system architecture for a CAD system from a global point of view can be illustrated as a simple pipeline of three major stages namely image pre-processing (format conversion, ROI segmentation and noise removal), candidate nodule generation (nodule detection and segmentation), and nodules classification (reducing false positive candidates generated from stage two) as shown in figure 3.1. First, the system isolates the lung tissue from the extraneous CT information through the segmentation process. This is carried out to reduce the computational complexity of the detection process by narrowing the region of interest to only the lung cavities. In the detection phase, the candidate pulmonary nodules from segmented lungs are detected and classified into malignant and benign on the basis of shape, growth, texture, and appearance analysis. The nodule models used in this stage are generated by using appearance and shape models that are constructed with a database of previously detected and annotated lung nodules. The nodule detection phase passes potential nodule candidates to a second segmentation process in order to extract only those pixels belonging to the candidate nodule in question, by

removing all other anatomies surrounding (and possibly attached) to the candidate nodule. The coarse candidate segmented from stage two used as an input for the last stage of the pipeline called *nodule classification*, which attempts to classify the candidate in question into the categories of nodule or non-nodule. In this stage, we need to classify a number of locations in each scan as being true nodules or false-positives. A list of nodule candidates, which were computed using existing nodule detection algorithms, were supplied to this(i.e. the third) step.

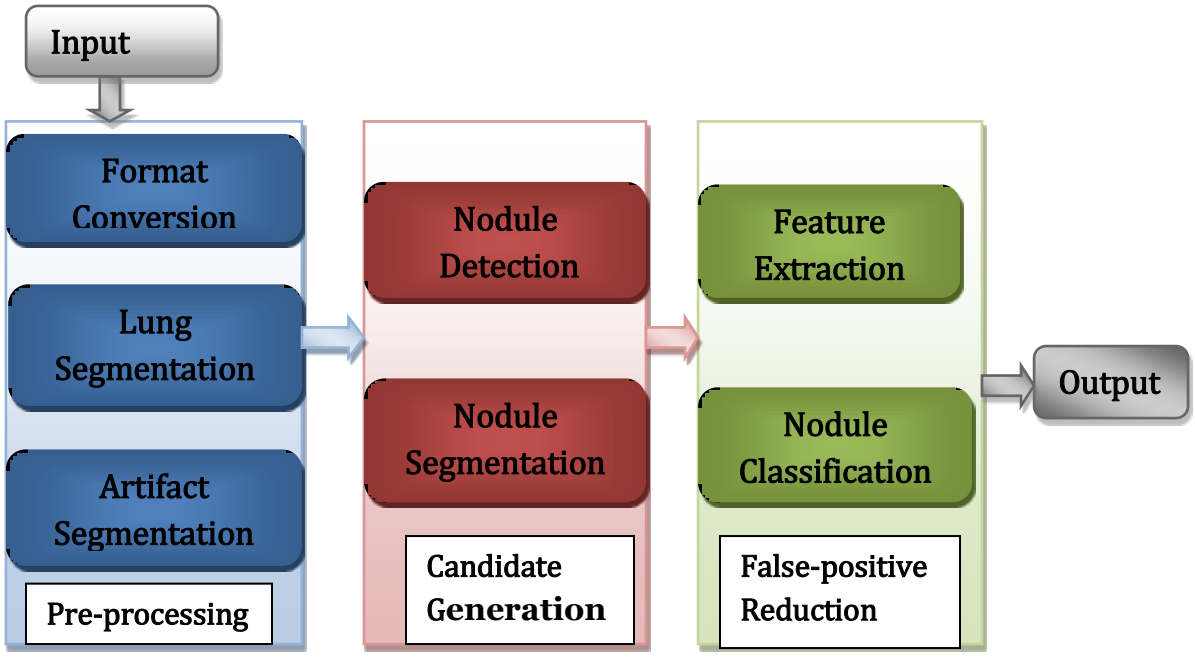


Figure 3.1: a general CAD System pipeline shown to outline the overall process

In this thesis our aim is designing and implementing a CNN model for the purpose of feature extraction and classification of nodules (i.e. stage 3) by using potential nodule candidates already generated by other algorithms. The input dataset for our study consists of 551, 065 nodule candidates and 1120 out of 1186 ground truth nodules (with sensitivity of 94.4%). These candidates are generated by merging the candidates that were detected by Murphy et al. [16], Jacobs et al. [3], Setio et al. [15], etc. But sizes of the candidate nodules vary from 3 mm to 30 mm and this makes the input patch size decision subjective.

CNNs architecture incorporates many parameters that have to be tuned. One of the important parameters is input tensor size; it can change all other parameters in the model because these parameters must be compatible with each other with in the network. Still there is no a general consensus about the effect of the input patch size to the classification performance of the system.

Therefore, we tried to observe and compare the effect of input patch size to the output of the system by using small, medium and large patch sizes taken from the dataset. In addition, we wanted to show whether different results can be used in decision fusion together in order to increase the overall classification performance or not. Figure 3.2 shows the general scheme of our model for classification of candidate pulmonary nodules.

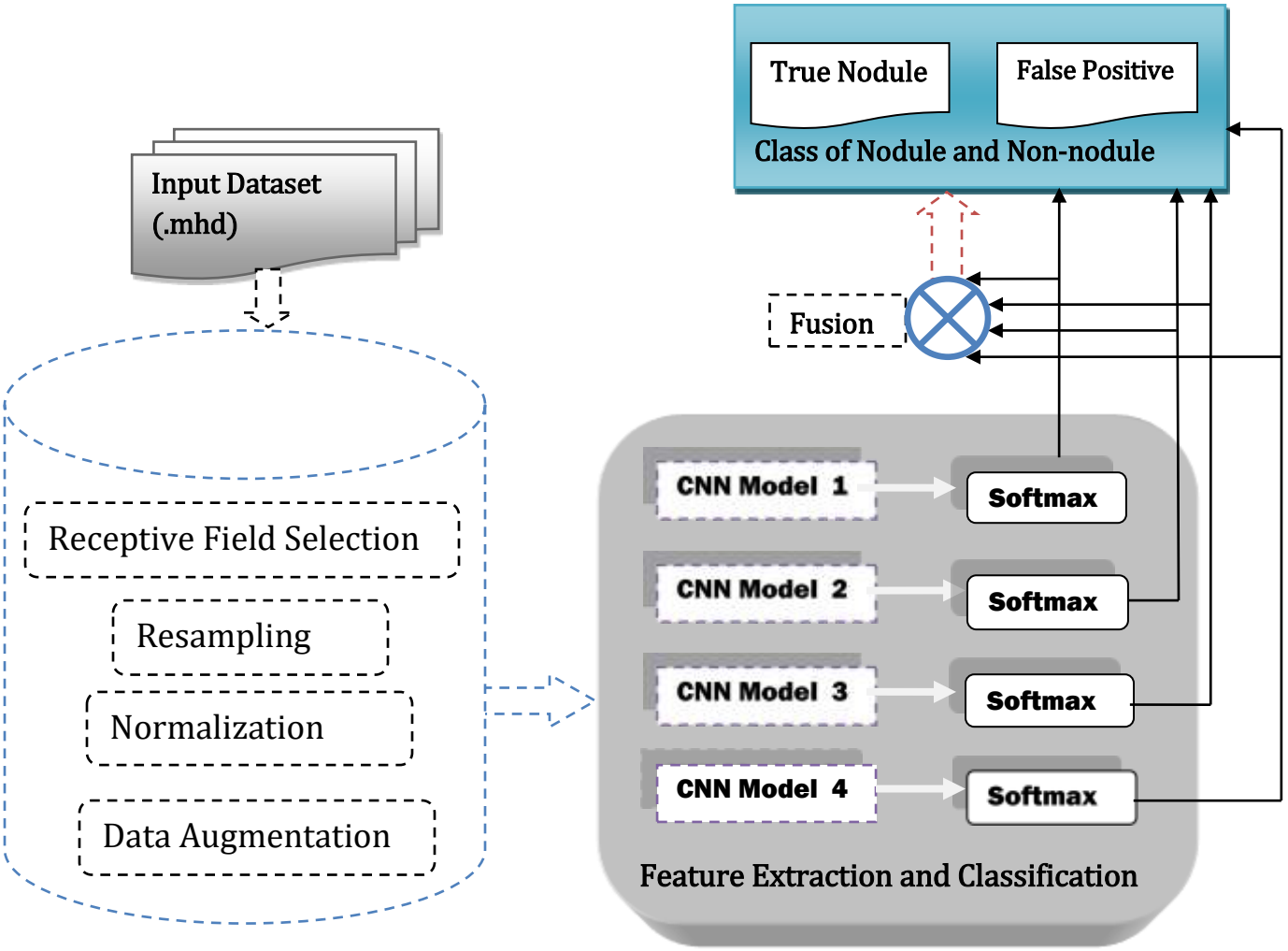


Figure 3.2: the general scheme of our model

We design four CNNs models for large, medium and small patch sizes to incorporate all annotated nodules based on the distribution of the dataset. We perform pre-processing (receptive field selection, resampling, normalization and augmentation) to make the dataset more suitable for training process. In the case of network architecture, we tried to examine two scenarios to make the evaluation more robust and acceptable. In the first case, purposefully we proposed to keep all network architectures similar, because if they were different from each other, it would be difficult to compare the effect of voxel size with respect to classification performance.

In the second case, we tried to design these four CNN models having different input patch sizes, different configuration and also different kernel sizes. Because, model-1 was modified to focus on smaller nodules, so that smaller convolutional filters have to be used in this case; model-2 and model-3 were used to focus on medium nodules so that medium convolutional filters used; model-4 intended to focus on large nodules with relatively large filter sizes. Finally, in both cases the final classification results are obtained by fusing the probability prediction outputs of these four different network architectures.

3.3 Data Pre-processing

Constructing CNN architecture requires a careful pre-processing operation on the dataset samples. To train the network, determining distance between the voxels, skewness (i.e. ratio of ground truth with respect to false-positives) of the dataset and size of the input patch is very important, because similarity of the input images, ratio of input samples and sizes of the image pixel highly affect the training process.

3.3.1 Receptive Field Selection

The size of receptive field (the surrounding range of a target position) plays an important role for the recognition performance of a network. The pulmonary nodules have large variations regarding the volume sizes (i.e. with diameter ranging from 3mm up to 30mm), shapes and many other characteristics such as subtlety, solidity, internal structure, speculation, sphericity, etc. [24]. In addition, the nodules often come with complicated contextual environments. If the patch size is small, only limited contextual information will be used to train the networks and its recognition capability will not be efficient to handle large variations of nodule candidates. On the other hand, if the receptive field is too large, more redundant information or even noises would be involved in the training process, which would degrade the performance of the networks. Because of these reasons it is difficult to figure out a single optimal receptive field for target nodules with large variations. So that, to determine patch sizes, we need to analyse the distribution of nodule diameters within the dataset. Dou et al [17] analysed the size distribution of the pulmonary nodules for all the samples in the dataset as shown in figure 3.3. Half of the nodules are less than approximately 7 mm. 80% of all nodules are less than 15 mm.

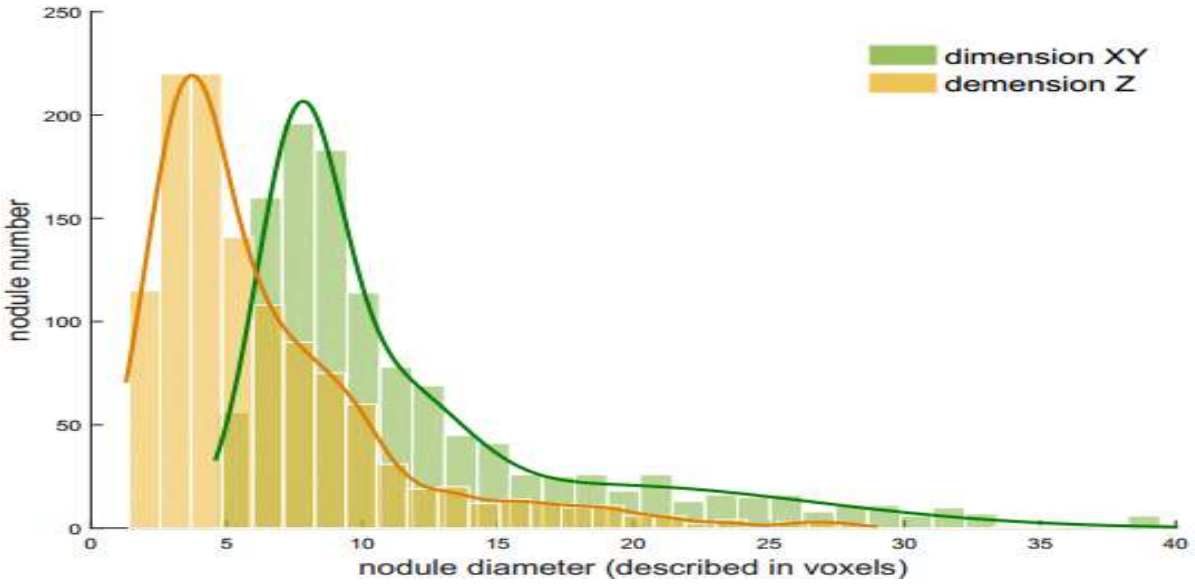


Figure 3.3: distribution of sizes of the pulmonary nodules for determining patch sizes [17].

Table 3.1 shows sample patch size with respect to their real world dimensions.

Patch Size (Voxel)	Patch Size (Real world size, mm)
12 × 24 × 24	12 × 16.8 × 16.8
18 × 30 × 30	18 × 21 × 21
20 × 20 × 6	20 × 14 × 4.2
24 × 36 × 36	24 × 25.2 × 25.2
30 × 30 × 10	30 × 21 × 7
30 × 42 ×42	30 × 29.4 × 29.4
36 × 48 × 48	36 × 33.6 × 33.6
40× 40 × 26	40 × 24 × 18.2

Table 3.1: patch sizes in voxels, and their corresponding real world dimensions in mm [17] [65].

A patch size with a receptive field of 20 × 20 × 6 can be able to encompass small sized pulmonary nodules with proper amount of context information, and it can covers up to 58% of all the nodules. Voxel 30 × 30 × 10 can represent majority of nodules by covering 86% of all nodules. On the other hand, 36 × 48 × 48 and 40 × 40 × 26 bounds over 99% of the nodules. Therefore, to compare the effect of the input patch sizes we proposed to use 20 × 20 × 6 for small nodules, 36 × 42 × 42 & 30 × 30 × 10 for medium sized nodules (i.e. the majority of the annotated nodules) and 36 × 48× 48 to cover the entire nodule including largest nodules with the surrounding environment.

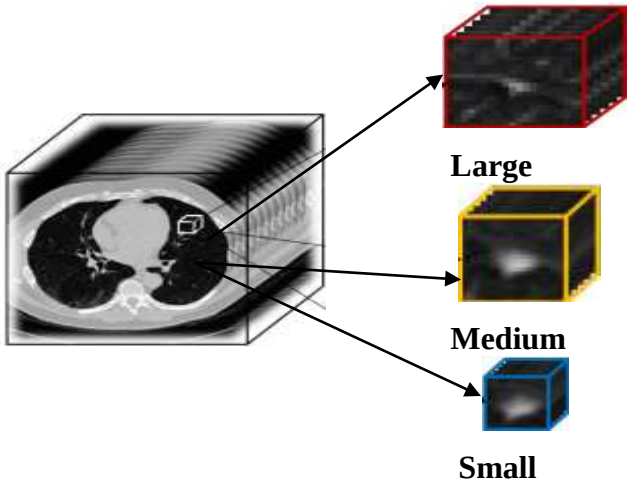


Figure 3.4: 3D Patch Extraction

3.3.2 Resampling and Normalization

Resampling is used to avoid image slice thickness differences because of different scan machines. Sizes of input tensors are measured in pixels, while real world dimensions are in millimetres. If the same input dimensions correspond to different real-world dimensions, the architecture will not give meaningful results, because it will be trained on different volumetric sizes. For example, while a 30 × 30 × 10 pixel corresponds to 30 × 21 × 7 mm for one scan, it may correspond to 28 × 28 × 10 mm for another scan. To avoid such situation, we need to examine each voxel spaces in the whole scans within the dataset. The main point here is getting the same real-world dimensions for a similar voxel sizes. Figure 3.4 shows a histogram of distances between voxels in X-Y and Z dimensions, while horizontal axis of the figure corresponds to the distance in mm, vertical axis shows how many samples exist for a certain distance.

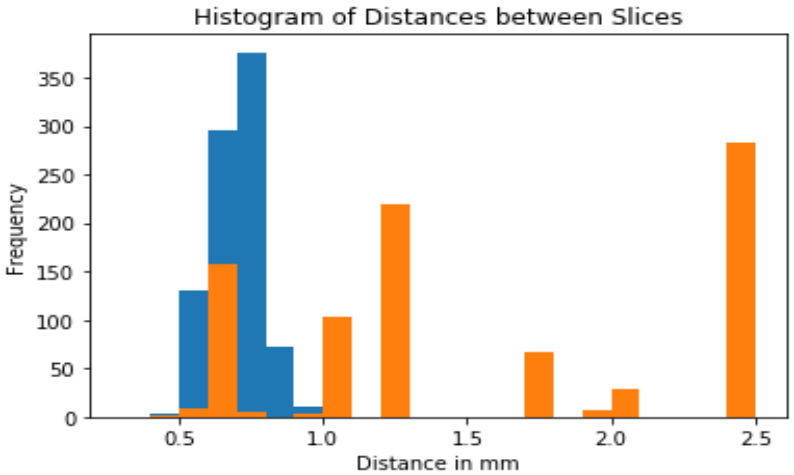


Figure 3.5: histogram of distances between voxels in X, Y and Z axes [65].

As we can see from the figure 3.5 voxel distance in the Z axis (between two transverse planes) is between 0.5 mm and 2.5 mm and voxel distances along the X-Y axis (between two sagittal plane and coronal plane) change from 0.5 mm to 1 mm. The unit space was selected as a tradeoff between minimization of data corruption due to resampling operation and maximization of nodule size. As a result, we reduce the distance in the Z-dimension to 1 mm in order to increase resolution, all volumes were resampled to new volumes so that voxel distance was 0.7×0.7×1 mm (X, Y, and Z axis, respectively). To normalize the pixel values, all CT scans were first transformed to Hounsfield Unit(HU) (i.e. section 2.5.1) scales using the information in DICOM header, to reduce impact of ribs scan intensity was clipped in range from -1000 up to 400 Housfield Unit and then normalized to the range of [0, 1].

```
#Clipping HU from -1000 to 400
def truncate_hu(image_array):
    image_array[image_array <-1000] = 0
    image_array[image_array > 400] = 0
```

```
#Normalization
def normalazation(image_array):
    max = image_array.max()
    min = image_array.min()
    image_array=(image_array-min)/(max-min)
    avg = image_array.mean()
    image_array = image_array-avg
    return image_array
```

These processes were applied by using Python programming language and Numpy library. Resizing all scans took approximately four days on Intel® core™ i7-7500U CPU@2.70, ~2.9 GHz processor, GPU:Intel® HD Graphics 620, RAM: 8GB.

3.3.3 Data Augmentation

The dataset has a skewness of around 1 nodule against 490 false positives. In a normal training case for this dataset, network parameters would learn the false positive structure because during the training process false-positive candidates would be more frequently encountered by the system. To avoid such imbalance between the false positive candidates and the true nodules (around 490:1), translation and rotation augmentations are performed for the ground truth nodule positions. Specifically, we shift the centroid coordinates by one voxel along each axis and rotated 90, 180 and 270 degrees within the transverse plane.

```
def angle_transpose(file,degree,flag_string):
    array = np.load(file)
```

```
array = array.transpose(2, 1, 0) # from x,y,z to z,y,x
newarr = np.zeros(array.shape,dtype=np.float32)
for depth in range(array.shape[0]):
    jpg = array[depth]
    jpg.reshape((jpg.shape[0],jpg.shape[1],1))
    img = Image.fromarray(jpg)
    out = img.rotate(degree)
    newarr[depth,:,:] = np.array(out).reshape(array.shape[1],-1)[:,:]
 newarr = newarr.transpose(2,1,0)
 np.save(file.replace(".npy",flag_string+".npy"),newarr)
```

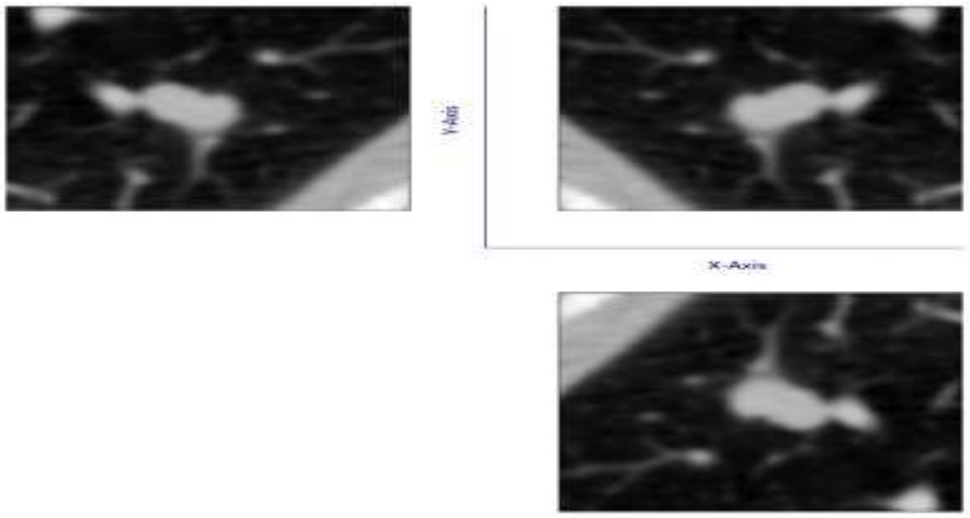


Figure 3.6: mirror images with respect to different axis [65].

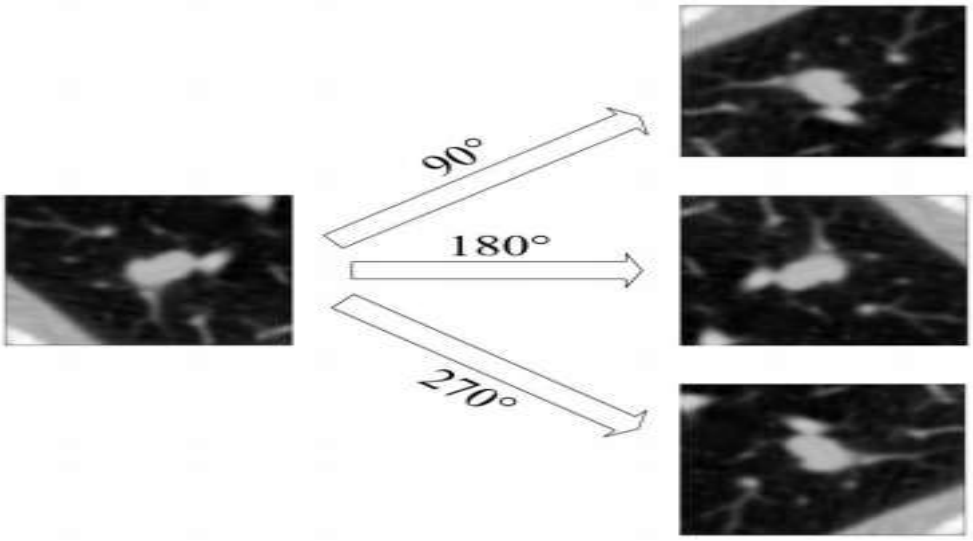


Figure 3.7: rotating images in 90, 180, and 270 degrees [65].

As we can see from figure 3.6 by transforming the original image we make those 3 mirror images, which mean in addition to the original sample slice for each ground truth nodules scans we generate, 3 different volumes (i.e. we have a total of 4 ground truth slices). When we rotate (i.e. figure 3.7) them by 90, 180 and 270 degrees: we make that 4 x 4=16 patches for each true-positive candidates. Totally we obtain 1120 x 16 = 17920 patches as a ground truth in the training algorithm. Hence, by using data augmentation we can reduce the ratio of false-positive candidates with respect to the ground truth nodules from around (1: 490) to (1: 30).

3.4 The Proposed CNN Models

Before sketching our architecture we need to know about the most important parameters to exploit CNN model. Determining number of convolution layers, pooling type, filter sizes and number of nodes in the hidden layer are the most important parameters when tuning the CNN model. In this paper we planned to address two research questions (i.e. comparing the effect of input patch sizes to the performance of the system; and evaluating results of the individual architecture with the ensemble of classifiers). Hence, in the first case we proposed to use the same CNN architecture for each patch sizes; because if they were different from each other, it would be difficult to compare the effect of voxel size with respect to classification performance. In the second case to examine each individual network with the fused result we designed two additional models having different input patch sizes, different configuration, different pooling size and also different kernel sizes.

We have made several experiments by using different number of convolutional layers, filter sizes and fully connected nodes in order to select the optimal architecture. Using only 1 convolutional layer does not extract enough features. On the other hand, using more than 3 convolution layers increases computational complexity and training time. Therefore, we compared the models, which have 2 and 3 convolutional layers by using patch size of 24 × 36 × 36 because this can be considered as unbiased patch size that can reach majority of nodules with in the dataset. As a result a CNN architecture contains from 3 convolutional layers with 3 × 5 × 5 filter size gives us the best result. Hence, to evaluate the effect of the input patch size to the performance of the CNNs architecture, we have used a similar CNN pattern and filter sizes for all models as shown in figure 3.8.

CNN Pattern: INPUT → [C → ReLU] → P → [C→ ReLU] ^{2}→ FC

Filter Size: 64@ 5 × 5 × 3

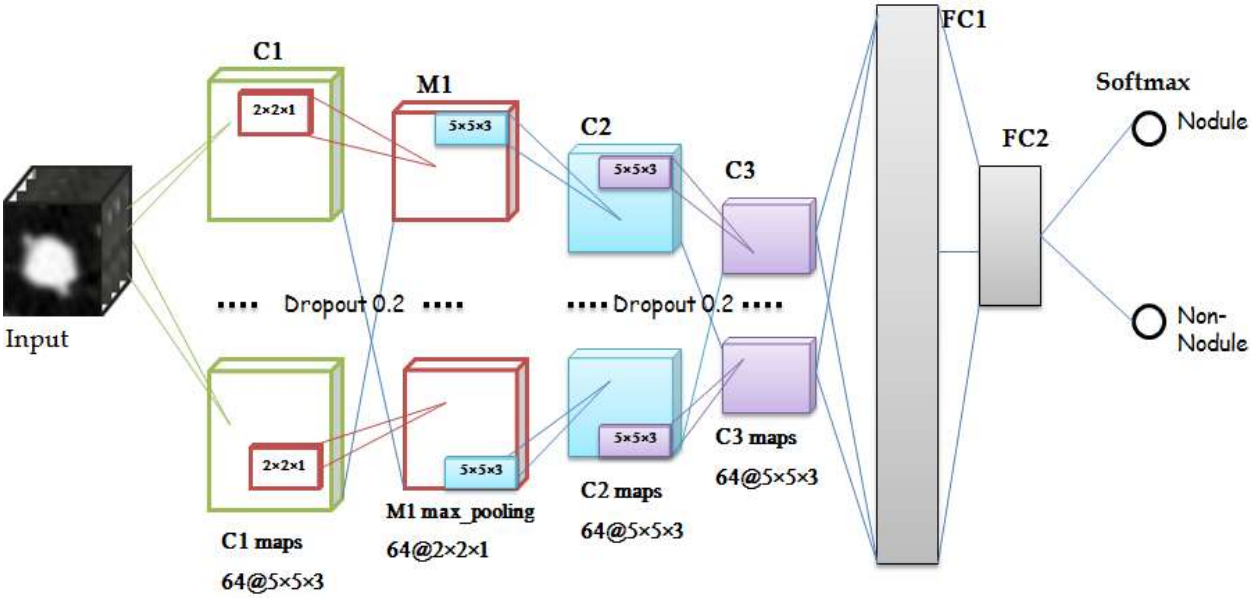


Figure 3.8: proposed CNN architecture to evaluate the effect of input patch with respect to classification performance

Convolution Layer: our CNN model contains from 3D convolutional, 3D max-pooling, fully-connected and a softmax layers. It consists of a number of channels (i.e. generates a 3D feature volume) and every channel encodes a different pattern. The 3D feature volume includes a group of neurons structured in a cubic manner. In our architecture to convolve a new feature map we use different 3D kernels and we propose to use Rectified Linear Unit (ReLU section 2.9) i.e. σ (a) = max (0, a) as an activation function by adding a bias term to derive the formula:

$$V_i^l(x,y,z) = \sigma\left(b_i^l + \sum_k \sum_{u,v,w} V_k^{l-1}(x-u, y-v, z-w) W_{ki}^l(u,v,w)\right) \quad (3.1)$$

Where, V_i^l and V_k^{l-1} represent the i^{th} 3D feature volume in the l^{th} layer and the k^{th} 3D feature volume in the previous layer, respectively; $W_{ki}^l \in R^3$ is the 3D convolutional kernel connecting V_i^l and V_k^{l-1}. $V_i^l(x,y,z)$, $V_k^{l-1}(x,y,z)$ and $W_{ki}^l(u,v,w)$ represent their element-wise values b_i^l is bias term and $\sigma(.)$ is the nonlinear activation function ReLU. The summation over k means the summation of activations from different 3D kernels.

Pooling (Down Sampling): max-pooling is more efficient and commonly used sub-sampling technique (i.e. section 2.10.2). Therefore we have used 3D max-pooling to sub-sample the convolved 3D feature volumes. 3D Max-pooling operation selects the maximum activation within neighbourhood convolved 3D feature maps and generates an output. The number of

feature volumes which remains unchanged during the pooling operation. Given the pooling kernel size of M and stride of S, the size of feature volumes is reduced as $X' = (X - M)/S + 1$ (same for Y' and Z'). In figure 3.8 we have used M = 2 × 2 × 1 for all models.

Fully Connected Layer: using too many neurons in the hidden layer will make the network lose its generalization ability. On the other hand, with too few nodes leads to use little information and may not solve complex patterns. Since patch sizes are different for each model, which makes the number of nodes in the first fully connected layer is different in each network. To implement the fully-connected layer, we first flatten the feature volumes into a neuron vector, next perform a vector-matrix multiplication, then add a bias term to it, and finally apply a non-linear function to generate the activations:

$$f^i = \sigma\left(b^i + W^i f^{i-1}\right) \qquad (3.2)$$

Where f^i the output feature vector of the $i\ th$ layer, f^{i-1} the input feature vector obtained by flattening the 3D feature volumes of the *(i-1) th* layer, W^i the weight matrix, b^i a bias term and $\sigma(.)$ represents ReLU. We perform a method known as dropout after convolutional and fully-connected layers to improve the generalization performance of our networks. In figure 3.8 we used 200 neurons in the first fully connected layer for all type of nodules.

Softmax Layer: the output layer for our architecture used to calculate the probabilities of each target class over all possible target classes. The calculated probabilities will be between 0 and 1 and the sum of all the probabilities is equals to 1. The softmax regression can be calculated as section 2.10.4.

Loss Function: to estimate the error in the last layer optimization have been performed by using cross-entropy loss function

$$E = -\sum_i y' \log(y_i) \qquad (3.3)$$

Where y is the predicted probability, and y' is the true value of the class of the sample. If the predicted probability and the true output are very close to each other, error will be close to zero. On the other hand, if the difference between the predicted value and the true output gets larger, the error will be increased exponentially.

In the second scenario i.e. to compare each individual CNN network with the ensemble one, we try to design CNN models having different input patch sizes, different pattern and also different kernel sizes. Because, model-1 was modified to focus on smaller nodules, so that smaller convolutional filters have to be used in this case; model-2 and model-3 were used to focus on medium nodules so that medium convolutional filters used; model-4 intended to focus on large nodules with relatively large filter sizes. In the same way in the fully connected layer, for different patch sizes it is better to use different number of nodes in the first fully connected layer (i.e. the larger the patch size the more number of nodes needed in the first fully connected layer). So that, for small patch sizes (i.e. $20 \times 20 \times 6$) we design CNN architecture having relatively small kernel size (i.e. $64@3\times3\times3$), small pooling size (i.e. $1 \times 1 \times 1$) and relatively small number of neurons (i.e.150) in the first fully connected layer. As we have mentioned in section 2.10 one of the most widely used patterns in CNN modeling is INPUT $\rightarrow$ [C $\rightarrow$ ReLU $\rightarrow$ P]$^{N}\rightarrow$ FC therefore we designed our CNN architecture for small sized patches (i.e. $20 \times 20 \times 6$) based on this configuration (pattern) as shown in figure 3.9.

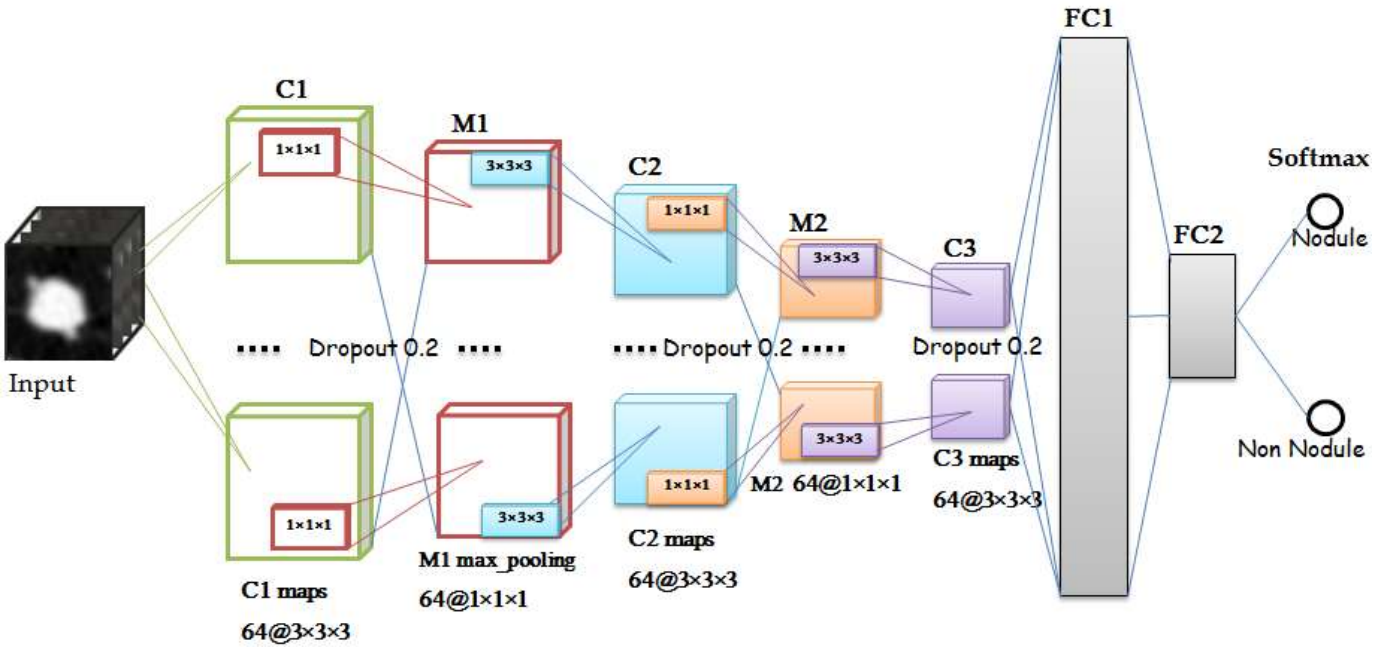


Figure 3.9: CNN architecture for small sized nodules to evaluate ensemble of classifiers

For the large nodules (i.e. $36 \times 48 \times 48$) we design CNN architecture having relatively large kernel size (i.e. $64@7\times7\times5$), large pooling size (i.e. $2 \times 2 \times 2$) and relatively large number of neurons (i.e. 250 neurons) in the first fully connected layer but a similar pattern with a model designed for small patch sizes as shown in figure 3.10.

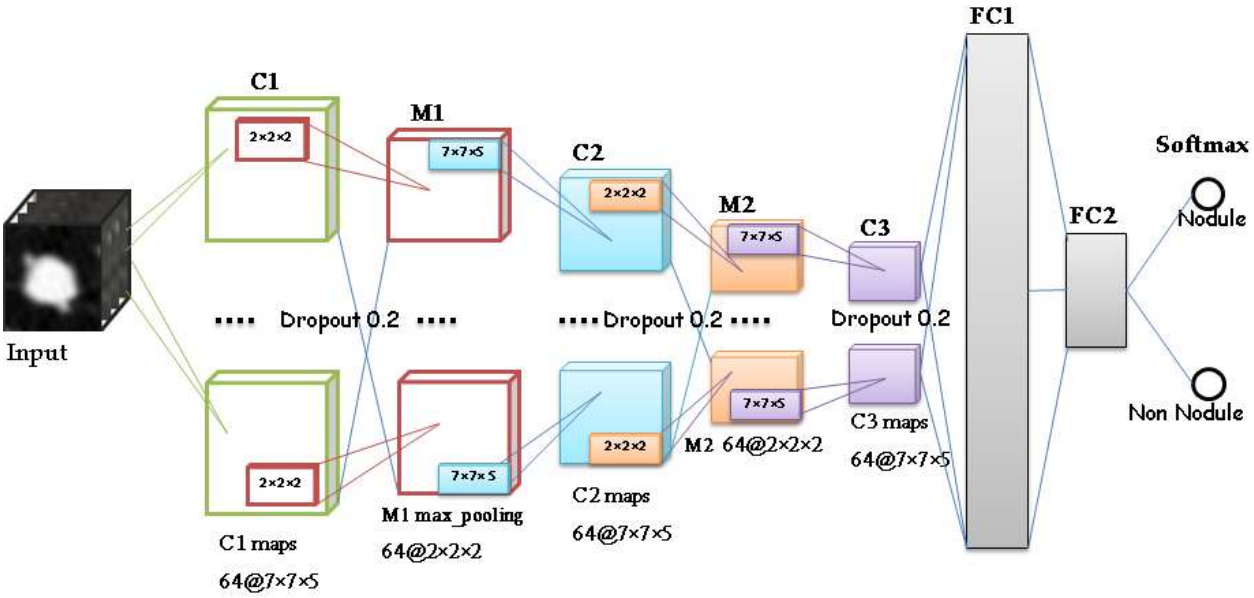


Figure 3.10: CNN architecture for large patch sizes to evaluate ensemble of classifiers

To evaluate medium sized nodules we proposed to use the same architecture and parameter configurations with figure 4.7, because during experiment (section 5.3.1) it achieves the best result for the majority of nodules having medium sizes (i.e. 24×36×36) compared with other possible CNN configurations.

Model-1	Model-2	Model-3	Model-4
Size = 20 × 20 × 6	Size = 30 × 42 × 42	Size = 30 × 30 × 10	Size = 36 × 48 × 48
C1 64@3 × 3 × 3	C1 64@5 × 5 × 3	C1 64@5 × 5 × 3	C1 64@7× 7 × 5
M1 64@(1×1×1)	M1 64@(2×2×1)	M1 64@(2×2×1)	M1 64@(2×2×2)
C2 64@3 × 3 × 3	C2 64@5 × 5 × 3	C2 64@5 × 5 × 3	C2 64@7 × 7 × 5
M2 64@(1×1×1)	C3 64@5 × 5 × 3	C2 64@5 × 5 × 3	M2 64@(2×2×2)
C3 64@3 × 3 × 3	FC1@ 200 Neurons	FC1@ 200 Neurons	C2 64@7 × 7 × 5
FC2@ 150 Neurons	FC2@ 2 Neurons	FC2@ 2 Neurons	FC2@ 250 Neurons
FC1@ 2 Neurons	-	-	FC2@ 2 Neurons
Softmax	Softmax	Softmax	Softmax

Table 3.2: the general configurations of CNN model-1, model-2, model-3 and model-4 to evaluate ensemble of classifier

3.5 Training Process

We perform 10-fold cross-validation to train and test the network. The data divided into 10 subsets of approximately equal size. From these subsets, 9 are used for training the network and

the remaining one subset is used for testing the results. As we have mentioned in section 3.3.3 we perform data augmentation to avoid class imbalance between ground truth nodules and false positive candidates, after augmentation totally we get 1120 x 16 = 17920 patches as a ground truth which increases the dataset from 551,065 to 567,865 (i.e. 551,065 + (17920-1120) divided into 10 subsets. Even though we perform data augmentation the ratio of the false positives to true positives is still skewed around (1:30). To handle such class imbalance issue we proposed to use Mini-Batch training sample. Mini-batch sizes, commonly called "batch sizes", are often tuned to an aspect of the computational architecture on which the implementation is being executed. Such as a power of two that fits the memory requirements of the GPU or CPU hardware like 32, 64, 128, 256, and so on. Hence, due to memory constraint the mini-batch size was set to be 32.

The weights θ were learned with stochastic gradient descent and the momentum was set to 0.9, and the dropout = 0.2 strategy was used in the convolutional and fully connected layers to improve the generalization ability of our networks. The weights were randomly initialized from the Gaussian distribution N (0, 0.01^2) and updated with standard backpropagation algorithm. We set a relatively high learning rate =0.3 at the beginning of the training process and decayed by 5 % every 3000 iterations because we considered that the 3D network is trained from scratch rather than fine-tuned from a pre-trained model.

The networks were implemented by using Python programming language using deep learning library Keras (i.e. plaidML for Intel HD Graphics) as a backend. The algorithm was run on Intel® core™ i7-7500U CPU@2.70, ~2.9 GHz processor, GPU:Intel® HD Graphics 620, RAM: 8GB. During the training process each model took a variable time for training, the training time depends up on the patch size, the larger the patch size the more time it takes for training. On average, each model took approximately 48 hours.

3.6 Ensemble of Classifiers (Fusion Method)

An ensemble of models (fusion) [62][63][19][17] refers to the practice of combining predictions from multiple statistical models to form one final prediction. It is a useful technique in machine learning to increase classification performance of different models. Each model can focus on different characteristics in the input data and their strengths can be used together. In this thesis, we attempt to use fusion technique in order to observe whether there is an increase in the overall

classification performance or not. For a testing nodule candidate $\mathcal{I}_j$, each model will assign a prediction probability for it; and for the final classification, we fuse the softmax regression outputs from all networks. The fused subsequent probability is estimated by weighted linear combination:

$$\mathbb{P}_{\text{fusion}}(\hat{\mathcal{Y}}_j = c \mid \mathcal{I}_j) = \sum_{\varphi \epsilon \{1,2,3,4\}} \gamma_\varphi \mathbb{P}_\varphi(\hat{\mathcal{Y}}_j = c \mid \mathcal{I}_j; \theta_\varphi) \qquad (3.4)$$

Where, $\mathbb{P}_{\text{fusion}}(\hat{\mathcal{Y}}_j = c \mid \mathcal{I}_j)$ is the fused prediction probability of $\mathcal{I}_j$, belonging to class *c* output by the whole models. $\gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 = 1$. The ensemble prediction is calculated as the average of the member predictions. The constant weights were determined using grid search on a small subset of the training data in our experiments (i.e. 0.2, 0.25, 0.25 and 3). We preferred weighted average ensemble than model average, because it allows multiple models to contribute to a prediction in proportion to their trust or estimated performance. However, model averaging ensembles are limited because they require that each ensemble member contribute equally to predictions [62].

3.7 Evaluation Metrics

The evaluation is performed by measuring the detection sensitivity of the algorithm and the corresponding false positive rate per scan. In order to assign a class (nodule or non-nodule) to the probabilities, a threshold value must be determined. To determine this threshold value, Area under the Free-Response ROC Curve (FROC) algorithm is generally used [23].

$$\text{Sensetivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad (3.6)$$

A predicted candidate location was counted as a true positive if it was located within the radius of a true nodule center. To obtain a point on the FROC curve, only those findings of a CAD system whose degree of suspicion is above a threshold (T) value are selected, and the sensitivity and average number of false positives per scan is determined. The final score is defined as the average sensitivity at 7 predefined false positive rates: 1/8, 1/4, 1/2, 1, 2, 4, and 8 FPs per scan [23]. To use FROC algorithm we use a performance metrics known as computation performance metrics (CPM). CPM measures the average sensitivity at the predefined seven operating points of the FROC curve i.e. 1/8, 1/4, 1/2, 1, 2, 4, and 8.

$$\text{CPM} = \frac{\sum_{i=1}^{7} \text{sensitivity of } i}{7} \qquad (3.7)$$

Where, $\text{sensitivity of } i$ represents the seven predefined false positive rates (i.e. sensitivity of 1/8, 1/4, 1/2, 1, 2, 4, and 8 FPs per scan). The desired goal is to achieve a very high sensitivity with a highly reduced false positive, but there is a tradeoff between sensitivity and false positives. When we set a small threshold value, we can get a very high sensitivity but false positives will also become high. This is not a desired situation because many false positives need to be cleaned out by the radiologists. On the other hand, if a high threshold value is set, false positives will be reduced but sensitivity will also be reduced. This is also not a desired situation because there will be many missing true positives, which is not applicable to the clinical use. False positives (FPs) are number of nodule predictions made by our system, but do not contain any annotated lung nodule. False negatives (FNs) are number of annotated nodules that don't predicted by our system. True positives (TPs) are the number of lung nodules that have been successfully detected by our system.

Confusion Matrix	Condition		
		P	N
Test Result	P	TP	FP
	N	FN	TN

Table 3.3: confusion matrix

FROC algorithm checks each threshold value and records its false positive rate and sensitivity. Since there can be different threshold values for a certain false positive rate per scan, there can be different sensitivity levels for the same false positive per scan. That is why there are dashed lines in the FROC graph (i.e. as shown in figure 5.); those dashed lines show maximum and minimum sensitivity values.

Chapter Four: Experimental Results and Discussions

4.1 Introduction

In this chapter we try to describe about the implementation details and experimental results of the proposed models. We perform a comprehensive set of experiments to evaluate the individual performance of each network and their ensemble result. We planned to compare the effect of different patch sizes with respect to classification performance and also to evaluate each individual model with the ensemble of classifiers. In the first case, we hypothesize that sensitivity of the network increased when we use a large patch sizes, because large patch sizes can covers a large number of nodules with many context information. In the second case, we hypothesize that fusion method can boost sensitivity, because each individual model could outperform a result based on some characteristics of the dataset. Section5.2 presents the datasets used to training and testing our models. Section 5.3 discuss about implementation and experimental results. Section 5.4 deals about the discussion.

4.2 Dataset

The dataset is obtained from LUNA16 official website; which provides a zip file (i.e. ten folders from subset0.zip to subset9.zip and .csv files) enclosing all CT images, annotations and candidate locations. It contains from 888 CT scans that have been filtered out from publicly available LIDC-IDRI [24] database. The LIDC-IDRI database contains a total of 1018 CT scans but LUNA challenge organizers discarded the scans that have a slice thickness more than 2.5 mm. In addition, they removed the scans with inconsistent slice spacing or missing slices. Remaining scans were divided into 10-folds with the objective to perform cross validation over them. The volumes have 512×512 pixels resolution in the transverse plane, 0.74×0.74 mm^2 element spacing, and variable slice thickness not larger than 2.5 mm. These scans were provided in MetaImage format (.mhd) and each .mhd file was stored with a separate .raw file that stores pixel data. Figure 4.1 shows that CT image of a single person from three different point of reference. The upper left orientation is called as the transverse plane, the upper right orientation is called as sagittal plane, and the lower right orientation is known as the coronal plane.

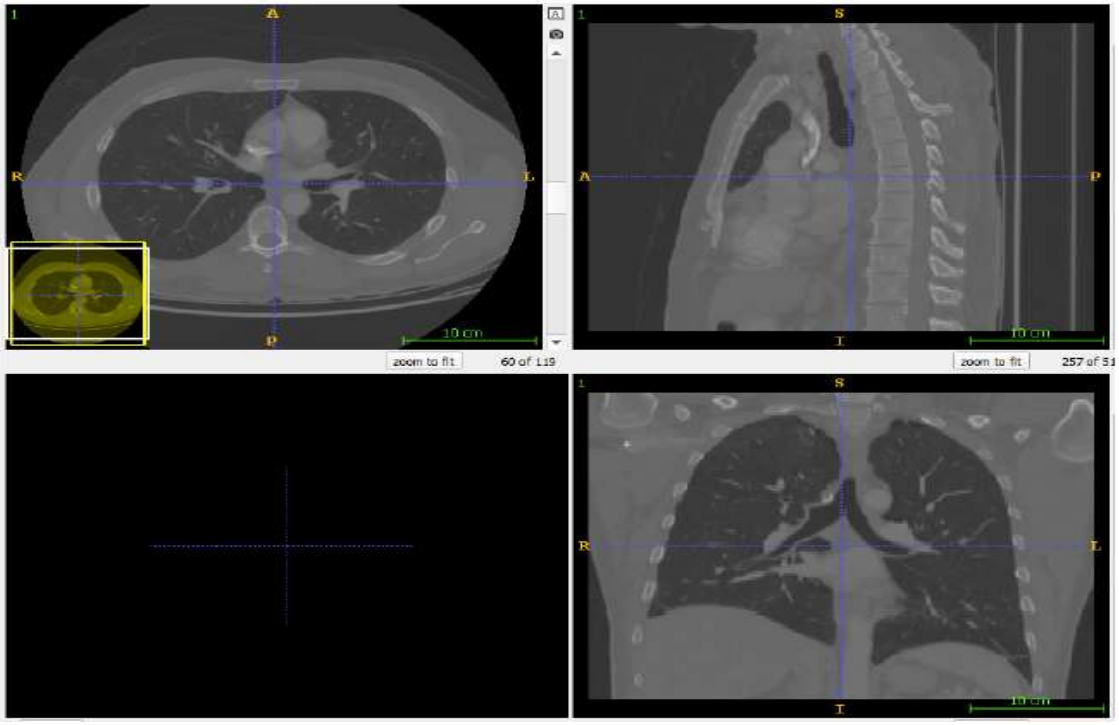
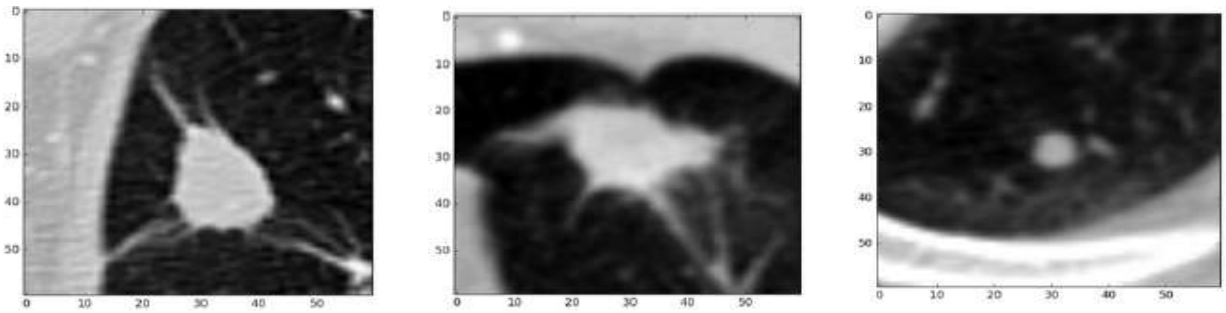


Figure 4.1: lung CT image from 3 different orientations opened in DICOM viewer software [23]

The dataset was annotated by experienced thoracic radiologists in a two phase reading process. First the annotation process was held by each radiologist independently and marked the annotated lesions as non-nodule, nodule < 3 mm, and nodules >= 3 mm. In the second phase, the whole nodules independently annotated are evaluated by each radiologist again. Then, the challenge selected a total of 1186 nodules >= 3mm approved by three or four radiologists as accepted true nodules (i.e., ground truth). Which means non-nodule structures, nodules < 3 mm, and nodules annotated by one or two radiologists are discarded from the reference standard (i.e. known as false positive candidates). The dataset consists of 551, 065 nodule candidates generated by Murphy et al. [16], Jacobs et al. [3], Setio et al. [15], etc. from which 1120 out of 1186 are ground truth nodules (with sensitivity of 94.4%). Figure 4.2 shows some ground truth nodules accepted by LUNA 16 challenge organizers.



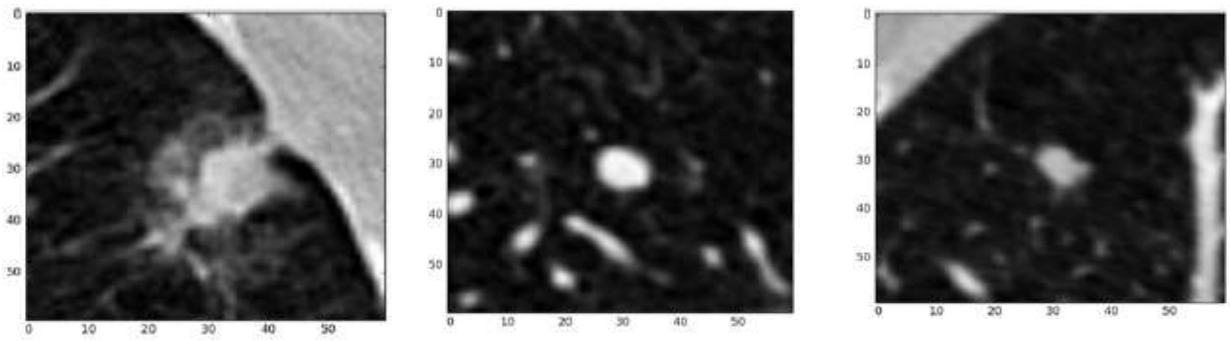


Figure 4.2: ground truth nodules accepted by the LUNA challenge organizers [23]

The .csv file contains from:

annotations.csv: consists of candidate location (i.e. X, Y and Z coordinates) and diameter (i.e. in mm), used as reference for the nodule detection track, and

candidate.csv: consists of candidate locations (i.e. X, Y and Z coordinates) and their class (i.e. being nodule or not) that are used as a reference standard for the 'false positive reduction' track.

seriesuid	coordX	coordY	coordZ	diameter_mm
1.3.6.1.4.1.14519.5.2.1.6279.6001.100225287222365663678666836860	-128.699	-175.319	-298.388	5.651470635
1.3.6.1.4.1.14519.5.2.1.6279.6001.100225287222365663678666836860	103.7837	-211.925	-227.121	4.224708481
1.3.6.1.4.1.14519.5.2.1.6279.6001.100398138793540579077826395208	69.63902	-140.945	876.3745	5.786347814
1.3.6.1.4.1.14519.5.2.1.6279.6001.100621383016233746780170740405	-24.0138	192.1024	-391.081	8.143261683
1.3.6.1.4.1.14519.5.2.1.6279.6001.100621383016233746780170740405	2.441547	172.4649	-405.494	18.54514997
1.3.6.1.4.1.14519.5.2.1.6279.6001.100621383016233746780170740405	90.93171	149.0273	-426.545	18.20857028
1.3.6.1.4.1.14519.5.2.1.6279.6001.100621383016233746780170740405	89.54077	196.4052	-515.073	16.38127631
1.3.6.1.4.1.14519.5.2.1.6279.6001.100953483028192176989979435275	81.50965	54.95722	-150.346	10.36232088
1.3.6.1.4.1.14519.5.2.1.6279.6001.102681962408431413578140925249	105.0558	19.82526	-91.2473	21.08961863
1.3.6.1.4.1.14519.5.2.1.6279.6001.104562737760173137525888934217	-124.834	127.2472	-473.064	10.46585391
1.3.6.1.4.1.14519.5.2.1.6279.6001.105495028985881418176186711228	-106.901	21.92299	-126.917	9.745258507

Table 4.1: annotation .csv coord X, coord Y, coord Z and diameter_mm are in mm

seriesuid	coordX	coordY	coordZ	class
1.3.6.1.4.1.14519.5.2.1.6279.6001.100225287222365663678666836860	68.42	-74.48	-288.7	0
1.3.6.1.4.1.14519.5.2.1.6279.6001.100225287222365663678666836860	-95.2094	-91.8094	-377.426	0
1.3.6.1.4.1.14519.5.2.1.6279.6001.100225287222365663678666836860	-24.7668	-120.379	-273.362	0
1.3.6.1.4.1.14519.5.2.1.6279.6001.100225287222365663678666836860	-63.08	-65.74	-344.24	0
1.3.6.1.4.1.14519.5.2.1.6279.6001.100225287222365663678666836860	52.94669	-92.6889	-241.068	0
1.3.6.1.4.1.14519.5.2.1.6279.6001.100225287222365663678666836860	54.65773	-104.812	-249.595	0
1.3.6.1.4.1.14519.5.2.1.6279.6001.100225287222365663678666836860	134.73	-138.79	-333.76	0
1.3.6.1.4.1.14519.5.2.1.6279.6001.100225287222365663678666836860	84.80406	-84.7527	-379.787	0
1.3.6.1.4.1.14519.5.2.1.6279.6001.100225287222365663678666836860	-76.47	-171.11	-258.23	0
1.3.6.1.4.1.14519.5.2.1.6279.6001.100225287222365663678666836860	-38.56	-218.48	-257.82	0
1.3.6.1.4.1.14519.5.2.1.6279.6001.100225287222365663678666836860	90.10228	-68.4308	-218.243	0

Table 4.2: candidates.csv column class=1 for true nodules class=0 for false positives

4.3 Implementation and Test Results

Before conducting the testing process we need to set some parameter values in order to achieve a better classification performance. Specifying some of the system parameters and their values scientifically based on experimental procedure is a critical issue in order to get an excellent result. A value of some of the parameters can be able to determine by reviewing researches, but for some other attributes it is important to undertake some preliminary experiments.

Learning rate: by analysing loss graphs in the training process, an optimum value can be decided. A high learning rate generally decreases rapidly and remains constant after a while. On the other hand, a low learning rate may not reach the global minimum due to local minima, too. Reducing the learning rate over time can help to settle our network into a nice local minimum instead of continuously stepping over it due to too large update steps. Therefore we set a relatively high learning rate = 0.3 at the beginning of the training process and decayed by 5 % every 3000 iterations i.e. 0.3 * 0.05 as [17][64].

Number of epochs: to decide on the number of training epoch an experiment has been conducted. If small number of epochs are used, the model might under-fit (cannot learn enough, both test and training set accuracies are low); if too many epochs are used, the model might over-fit (model memorizes the training samples). Over-fit means that the test set has low accuracy although the training set accuracy is very good. So, the experiment is running on the first model for 200 epochs and the loss function and accuracy have been determined. Based on our experiment we noticed that error is dropping smoothly up to around 120 training epochs, then after that it continuous almost constant. So, we decided to use 120 epochs during the training process. According to literatures this analysis can show us when to stop the training process. Finally, *Size of the mini batch* data is set to 32 due to memory constraint.

4.3.1 CNN Design Evaluation and Discussion

To determine that the chosen CNN network architecture was good enough (i.e. for evaluation of different input patch sizes), several model configurations have been tested. As we have mentioned in chapter four using only 1 convolutional layer does not extract enough features. On the other hand, using more than 3 convolution layers increases computational complexity and

training time. Therefore, we compared the models, which have 2 and 3 convolutional layers by using patch sizes of 24 × 36 × 36 i.e. because this was a neutral patch size that can reach majority of nodules with in the dataset. Table 4.3 shows three possible configurations of a CNN architecture that have been evaluated in this experiment to choose the best model for evaluation of different input patch sizes. The below three models (Table 4.3) contains from two and three convolutional layers; having relatively small, medium and large filter sizes in convolutional layer and in max-pooling layers.

Model-1			Model -2			Model -3		
Layer	Kernel size	Channel	Layer	Kernel S	Channel	Layer	Kernel Si	Channel
Input	-	1	Input	-	1	Input	-	1
C1	2 × 2 × 1	32	C1	3 × 3 × 3	48	C1	5 × 5 × 3	64
M1	1 × 1 × 1	32	M1	2 × 2 × 1	48	M1	2 × 2 × 1	64
C2	2 × 2 × 1	32	C2	3 × 3 × 3	48	C2	5 × 5 × 3	64
M2	1 × 1 × 1	32	M2	2 × 2 × 1	48	C3	5 × 5 × 3	64
FC1	-	100	C3	3 × 3 × 3	48	FC1	-	200
FC2	-	2	M3	2 × 2 × 1	48	FC2	-	2
Softmax	-	2	FC1	-	150	Softmax	-	2
			FC2	-	2			
			Softmax	-	2			

Table 4.3: three different CNN network configurations to determine the best model for input patch size evaluation

First of all, we designed Model-1 (Table 4.3) having 2 convolutional layers, filter size of 2 x 2 x 1 and a 1 x 1 x 1 max-pooling. Model-1 gave as a score of 0.656. This score seemed promising initially; therefore, we modified the other models upon this structure. In order to compare this model with another model that has 3 convolutional layers, we created Model-2 (Table 4.3). In this case we increased the number of convolutional filters and filter sizes to increase the complexities of the network. In this model we have used 48 convolutional filters which were of size 3 x 3 x 3 and size of max pooling was changed to 2 x 2 x 1. We could use 32 filters in the convolution operations but that would make Model-2 maximized version of Model-1.

Model-2 gave us a score of 0.7403 so we noticed that more number of convolutional filters and bigger filter sizes improved the performance. This shows that there are complex patterns in the nodule that can be captured with bigger filters. When we saw an increase in the performance, we wanted to observe whether increasing number of convolutional filters and kernel sizes by

reducing complexity of the network structure would generate better results; therefore, we designed Model-3 (Table 4.3) consists of 3 convolutional layers and only one max-pooling layer by increasing nodes in the fully connected layer from 150 to 200 and we got a score of 0.7505 from this network.

After having these results, we tried to observe if there is an increase in performance when we increase number of filter size from 64 to 80 and by increasing the number of nodes in the hidden layer of Model-3. But Performance dramatically decreased when it was compared with the above three Models i.e. Model-1, Model-2 and Model-3 (Table 4.3). Table 4.4 shows the mean sensitivity of the above three networks Model-1, Model-2 and Model-3 measured at seven FPs/Scan (i.e. 0.125, 0.25, 0.5, 1, 2, 4, 8). Hence, in accordance of these experiments three convolutional layers with 3 x 5 x 5 filter size, which is model -3, gave us the best result. Therefore we have used model-3 in table 5.3 (i.e. figure 3.8) for the evaluation of different input patch sizes.

Model No.	0.125	0.25	0.5	1	2	4	8	CPM
Model-1	0.4305	0.5038	0.5956	0.6855	0.7403	0.7947	0.8424	0.6561
Model-2	0.5062	0.6163	0.7	0.7747	0.8218	0.8667	0.8967	0.7403
Model-3	0.5668	0.6389	0.7118	0.7762	0.8167	0.8505	0.8926	0.7505

Table 4.4: results of Model-1, Model-2 and Model-3 shown in table 4.3

4.3.2 **Evaluation of input patch sizes**

We evaluated the performance of each model designed in chapter four by feeding the pre-processed and augmented 20 × 20 × 6, 30 × 42 ×42, 30 × 30 × 10 and 36 × 48 × 48 patches into the network structure described in figure 3.8. As we have mentioned in section 3.5 the data set was divided into 10 subsets of just about equal size. From these subsets, 9 are used for training process. The remaining one subset is then used for testing the results. This process is repeated for a total of 10 times, each time with a different subset as testing set. All training process ran for 120 epochs, with a random sample of 1200 negative patches and a mini-batch size of 32. By calculating the average of the seven sensitivities measured at several false positives per scan FPPS $\in \{0.125, 0.25, 0.5, 1, 2, 4, 8\}$ (i.e. section 3.7) CPM score can be generated. Table 4.5 shows that the mean sensitivity values at seven false positive rates and CPM scores of different patch sizes running on similar network architecture (i.e. figure 3.8).

Patch Size	0.125	0.25	0.5	1	2	4	8	CPM
20 × 20 × 6	0.4489	0.5368	0.6301	0.7057	0.7729	0.8061	0.8572	0.6797
30 × 42 ×42	0.5229	0.6320	0.720	0.8	0.8501	0.8658	0.9050	0.7565
30 × 30× 10	0.4833	0.6	0.682	0.763	0.8312	0.8598	0.9020	0.7316
36 × 48 ×48	0.5909	0.6698	0.7361	0.825	0.8511	0.8812	0.9112	0.7808

Table 4.5: average sensitivities (CPM) of different patch sizes

As we can see from table 4.5 when the patch size increase the average sensitivity of the model also increases, because large patch sizes can be able to encompass a great amount of candidate nodules as well as context information. When the average number of false positives per scan is 0.125, we get a maximum sensitivity of 0.5909. Although false positive rate is very low, relatively the maximum sensitivity of our model is also very low, which is not applicable in clinical areas. On the other hand, when the average number of false positive per scan is 8, the maximum sensitivity of our model reaches 0.9112, which means that over 90% of all the true-positives can be detected successfully. Yet, there are also false positives in the detected nodules and on average there are 8 false positives per scan. Since our dataset consists of 888 CT scans so that 7104 (i.e. 888x8=7104) false positives are generated in the classification when the sensitivity is 0.9112.

To compare the effect of the patch size with respect to the classification performance of the system we evaluated each patch size by using similar CNN architecture (i.e. figure 3.8). When we compare small patch sizes (i.e. 20 × 20 × 6) and large patch sizes (i.e. 36 × 48 ×48), there is a 0.1010 CPM score difference, which can be considered as a large gap for a similar CNN architecture. This situation can show us that the training voxel size is very important for these kinds of nodule classification problems. Even a single model that produces good results must be compared with the same model that uses different patch sizes.

Although smaller patch sizes score lower sensitivity relative to larger patch sizes, they have scored a better candidate detection probabilities for relatively smaller nodules. Table 4.6 shows probabilities of randomly selected small nodules that have sizes around 5 mm. It can be seen that smaller patch sizes can be able to achieve a better result compared to large patch sizes when small nodules are considered. It is because of that small nodules with many noises and residues around them will be hard to distinguish by using large patch sizes.

Nodule Diameter	Probability of 20 × 20 × 6	Probability of 36 × 48 ×48
5.2696	0.9892	0.7593
4.5434	0.9951	0.8022
4.4159	0.8674	0.2370
5.1188	0.9979	0.5532

Table 4.6: probabilities of the smallest nodules in small and large patch sizes

Even though small patch sizes can be able to generate a better probability of prediction in small nodules, they are limited to cover large nodules; in order to catch these large nodules it is important to use larger patch sizes.

4.3.3 **Evaluation of Ensemble of Classifiers**

Our second hypothesis was to increase the overall performance by using fusion method. In this experiment we have used two evaluation scenarios;

1. Evaluation of ensemble of classifiers by using a similar network architecture but different input patch sizes (i.e. figure 3.8)
2. Evaluation of ensemble of classifiers by using different network architecture as well as different input patch sizes (i.e. figure 3.8, figure 3.9 and figure 3.10)

Table 4.7 shows that the sensitivity values of the seven predefined false positive rates and averages of these sensitivities for the different input patch sizes (i.e. 20 × 20 × 6, 30 × 42 ×42, 30 × 30 × 10, 36 × 48 ×48) and their corresponding ensemble result in similar network architecture. As we can see from table 4.7 the fusion result increased the overall sensitivities (i.e. CPM score) when compared to any single model. In the case of 36 × 48 × 48 patch size (i.e. represent large patch sizes and scored better than the remaining three patch sizes), the fusion result was very close to it in the first 0.125 and 0.25 FP rates. Meanwhile, at 8 FP rates the Fusion model increased the score of 36 × 48 × 48 by 0.004 (i.e. 0.9152 – 0.9112) points.

Patch Size	0.125	0.25	0.5	1	2	4	8	CPM
20 × 20 × 6	0.4489	0.5368	0.6301	0.7057	0.7729	0.8061	0.8572	0.6797
30 × 42 ×42	0.5229	0.6320	0.720	0.8	0.8501	0.8658	0.9050	0.7565
30 × 30× 10	0.4833	0.6	0.682	0.763	0.8312	0.8598	0.9020	0.7316
36 × 48 ×48	0.5909	0.6698	0.7361	0.825	0.8511	0.8812	0.9112	0.7808
Fusion	0.5905	0.6685	0.7501	0.8308	0.8676	0.8967	0.9152	0.7885

Table 4.7: sensitivities of different patch sizes in a similar CNN architecture and the fusion result

In the second evaluation scenario, we try to design CNN models having different input patch sizes, different pattern and also different kernel sizes (i.e. figure 3.8, 3.9, 3.10 and Table 3.2). Because, model-1 was modified to focus on smaller nodules, so that smaller convolutional filters have to be used in this case; model-2 and model-3 were used to focus on medium nodules so that medium convolutional filters used; model-4 intended to focus on large nodules with relatively large filter sizes. In the same way in the fully connected layer, for different patch sizes it is better to use different number of nodes in the first fully connected layer (i.e. the larger the patch size the more number of nodes needed in the first fully connected layer). Table 4.8 shows the FROC scores corresponding to 4 different models introduced in section 3.4 and their aggregate fusion result. The score was calculated as mentioned in section 3.7.

Model	0.125	0.25	0.5	1	2	4	8	CPM
Model-1	0.542	0.6094	0.6607	0.73	0.7827	0.8347	0.8738	0.719
Model-2	0.5229	0.6320	0.720	0.8	0.8501	0.8658	0.9050	0.7565
Model-3	0.4833	0.6	0.682	0.763	0.8312	0.8598	0.9020	0.7316
Model-4	0.6571	0.739	0.8056	0.852	0.8747	0.9	0.9221	0.8215
Fusion	0.74	0.7801	0.8358	0.87	0.9048	0.9207	0.9275	0.8541

Table 4.8: sensitivities of different patch sizes in different CNN architecture and the fusion result

Evaluation of the four individual models (table 4.8) can be summarized as follow:

1. Each individual networks Model-1 (figure 3.9), Model-2 (i.e. figure 3.8 for (30 × 42 ×42), Model-3 (i.e. figure 3.8 for (30 × 30 ×10)) and Model-4 (figure 3.10) can be able to achieve a CPM score of 0.7190, 0.7565, 0.7316 and 0.8215 respectively.
2. For all of the three individual networks, the detection sensitivities can reach beyond 85% under the false positive rate of 8 FPs per scan, the three architectures (Model-2, Model-3 and Model-4) can achieve a sensitivity of over 90% with 8 false positives per scan.
3. Compared with Table 4.7 (i.e. sensitivities of different patch sizes in a similar CNN architecture) Model-1 and Model-4 increased by +0.0393 (i.e. 0.7190 - 0.67967) and +0.0643 (i.e. 0.8215 - 0.757214). As we can see from Table 4.8 model-4 reached detection sensitivity of 85% at 1 false positive rate per scan i.e. a highly reduced number of FPs at a promising sensitivity. This can prove that reasonably, using different network architecture for different patch sizes can make the 3D CNNs to address feature extraction and classification effectively (i.e. CT images of pulmonary nodule detection).

As we can see from table 4.8 the fusion model confirmed a strong capability in candidate classification while maintaining a better sensitivity compared with each individual CNN models (scored a CPM of 0.8541). For example, when we examine 0.125 FPs per scan, Model-1, Model-2, Model-3 and Model-4 obtained sensitivity of merely 54.2%, 52.3%, 48.4% and 65.7%, respectively. Meanwhile, our fusion model achieved a sensitivity of 74%, which exceeded that of the Model-1, Model-2, Model-3 and Model-4 by 19.8%, 21.7%, 25.6%, and 8.3%, respectively. This shows that the fusion of all models having different architecture can be able to generate the best performance.

4.4 Discussions

In this paper the development and evaluation of a CAD system for classification of lung nodules in 3D CNN have been presented. The experimental output of our design yields a promising result in the false positive reduction track. The performance of our model highly depends on the dataset size, model parameters, and the architecture of the algorithm employed. As we have seen from the test result for all models including the fusion, average sensitivities can reaches beyond 85 % in the 8 false positive rates per scan. It indicates that the 3D CNNs are a promising method in pulmonary nodule detection and classification. The achievement of the proposed architecture mainly lies in three aspects. First, compared with 2D counterpart 3D CNNs can represent a large set of spatial information, this makes 3D CNN more suitable for volumetric medical image processing. Second, we perform pre-processing on our dataset, we carefully examined the distance between two consecutive scans, distance between two voxels in the same plane and also image slice thicknesses. Third, we employed model fusion to utilize the aggregate strength of our models.

In this thesis, instead of developing the entire pulmonary nodule detection architecture which usually integrates a candidate detector and a false positive reducer, we carefully emphasize on the candidate classification task i.e. the false positive reduction track. This means that the proposed method is independent of the candidate screening operation, and therefore can be combined with any candidate detector. It is obvious that the final classification accuracy will also highly depend on the performance of the candidate screening methods. If the provided candidates come with a higher sensitivity, it is promising to achieve better results with our architecture.

In terms of computational cost each experiment was conducted on Intel® HD Graphics 620 with graphics memory of 4 GB having a RAM of 8GB by using Python programming language using deep learning library Keras (i.e. plaidML for Intel HD Graphics) as a backend. We evaluated the computational cost of each individual network in terms of number of parameters and training time. In this thesis we noticed that the training time for each model more related to the architecture of the model, but the most important parameter is the error rate rather than training time. Table 4.9 shows that the computational cost of each models for the candidate pulmonary nodule classification task (i.e. the learning rate reduced by 5% every 3000 iterations and Mom is stands for Momentum in stochastic gradient descent (SGD))

Evaluation of input patch sizes					
Patch Size	Learning Rate	Regularizer	Optimizer	Epochs	Training Time
20 × 20 × 6	0.3*0.005	Dropout (0.2)	SGD (Mom=0.9)	120	31 hrs
30 × 42 ×42	0.3*0.005	Dropout (0.2)	SGD (Mom=0.9)	120	52 hrs
30 × 30× 10	0.3*0.005	Dropout (0.2)	SGD (Mom=0.9)	120	45 hrs
36 × 48 ×48	0.3*0.005	Dropout (0.2)	SGD (Mom=0.9)	120	61 hrs
Evaluation of ensemble network					
Model 1	0.3*0.005	Dropout (0.2)	SGD (Mom=0.9)	120	26 hrs
Model 2	0.3*0.005	Dropout (0.2)	SGD (Mom=0.9)	120	52 hrs
Model 3	0.3*0.005	Dropout (0.2)	SGD (Mom=0.9)	120	45 hrs
Model 4	0.3*0.005	Dropout (0.2)	SGD (Mom=0.9)	120	69 hrs

Table 4.9: experimental computational cost for different CNN models in terms of training time

During the training process each model took a variable time for training, the training time depends up on the patch size, the larger the patch size the more time it takes for training. On average, each model took approximately 48 hours.

Table 4.10 shows the average sensitivities of different network architectures that have been designed for the false positive reduction purpose by using ConvNet and the computational resource employed for experiment. All architectures are evaluated by using CPM evaluation mechanism based on 0.125, 0.25, 0.5, 1, 2, 4 and 8 FPs/scan. Here are the selected recent researches that have been performed for the false positive reduction track.

Author	Dataset	Method	Resource	CPM
Arnaud A. Setio et al.[19]	LUNA 16	CNN	Theano (GPU GeForce GTX TITAN X)	0.814
Qi Dou et al. [17]	LUNA 16	CNN	Theano(GPU: NVIDIA TITAN Z)	0.827
Qi Dou et al. [17]	LUNA 16	CNN	Theano(GPU: NVIDIA TITAN Z)	0.908
A. Dobrenkii et al. [61]	LUNA 16	CNN	GPU: NVIDIA Tesla K40	0.735
ZNET [64]	LUNA 16	CNN	-	0.758
Ours	LUNA 16	CNN	GPU: Intel® HD Graphics 620	0.8541

Table 4.10: results of different CNN architectures in the false positive reduction track.

All the above researches employed deep CNNs for the false positive reduction task, which shows that the massive influence of deep CNNs on medical image analysis research community. Arnaud A. Setio et al. [19] trained multiple streams of 2D CNNs with a set of patches extracted from differently oriented planes. Although this method was still not able to encode a large set of features like the 3D ConvNet, it was an effective framework. Qi Dou et al. [17], uses three convolutional layers, they trained three different architectures and used a fusion method to use the aggregate characteristics of individual models. During the first entry of the challenge they scored a CPM score of 0.827 but in the second entry they achieved a better CPM performance of 0.908 because the fusion strategy they used boosts the overall performance. Dobrenkii et al. [61], re-sampled the CT scans into a homogenous voxel size and uses a ResNet architecture (i.e. variant of CNN algorithm) and can be able to achieve 0.735 CPM result. Although our architecture still needs some improvement to boost the overall accuracy, it is still comparable to many other CNN architectures in pulmonary nodule classification task.

Research questions that have been answered by the thesis:

- What is the effect of input patch sizes to the performance of the CNNs architecture?
- Compared with individual CNN models, may ensemble of classifiers achieve a substantial improvement in accuracy?

We designed four CNNs models for large, medium and small patch sizes; each model consists of three convolutional layers. We employed model fusion technique to utilize aggregate strength of all models. According to the experiments our contributions with regard to these questions can be summarized as:

- Increasing the patch size increases the overall CPM score because larger patch sizes can be able to encompass a large number of nodules within the dataset.
- Ensemble of classifiers combines the strength of different models scores a better result compared with each individual models. We can say that any reasonable fusion method can be able to boost the overall performance compared with each individual model.

Chapter Five: Conclusion and Future Work

5.1 Conclusion

Automated pulmonary nodule detection and classification helps to analyse a large number of CT images. In CT image analysis, for a single scan there are a huge amount of image slices that have to be reviewed, assessing nodules from these slices is exposed for error prone and makes radiologist task tiresome and time consuming. Sometimes, to get a good result more than one radiologist will participate on the detection process. Hence, to reduce radiologist burden there is a need for a CAD system that can be able to detect and classify nodule candidates robustly. In this paper, our aim was to design CNN architecture for candidate pulmonary nodule classification to reduce a significant amount of false positive results, to compare the effects of input patch sizes with respect to classification performance and to increase the overall performance by using fusion method. .

The proposed approach mainly consists of three steps pre-processing, feature extraction & classification and fusion technique. In the pre-processing step, we carefully examined the distance between two consecutive scans, distance between two voxels in the same plane and also image slice thicknesses. We noticed that due to different scan machines, slice numbers vary from 120 to 600 and image slice thickness changes from 0.4mm to 1mm in the X, Y plane and from 0.5mm to 2.5 mm in the Z dimension. Therefore to get a meaningful result during the classification process, we rescaled the scans to get the same real world dimension. After resampling all slices, four different patch sizes have been extracted and trained in a similar network setup in order to compare the effect of dissimilar patch sizes with respect to training, testing and classification performance of the convolutional neural networks. LUNA 16 dataset, contains a total number of 551,065 candidate points from which only 1120 candidates are true positives, this makes the skewness of the data around (1: 490) to avoid such class imbalance we perform translation and rotation data augmentations.

Our CNNs model contains from 3D convolutional, 3D max-pooling, fully-connected and a softmax layers. We employed a CNN network architecture consists of three convolutional layers to learn internal structures of the nodules better, additional convolutional layers can be used. But it makes our network more complex and the training time would be very high. In addition to that training 3D networks requires a high performance GPU power. During the evaluation of

ensemble of classifiers we have used 3×3×3 kernel size for the network corresponding to small patch size, 5×5×3 kernel size for medium sized inputs and 7×7×5 for large patches with 1×1×1 stride. It is because of that increasing kernel size like 8×8×8 captures large features, on the other hand using smaller size kernel like 1×1×1 represent specific features on the nodules. Using both type of kernel sizes might increase the accuracy of the system, but in the first case the network losses specific representation of nodules in the second case the network losses its generalization ability.

During the training process network parameters would learn the false positive structures (i.e. Due to class imbalance of the dataset) because in the training process those cases would be more frequently encountered by the system, to overcome such situation we employed mini-batch training rather than training in a full batch. After training all patches, we noticed that, each patch generated a different result. When the smallest and biggest patches were compared in similar network architecture there was a 0.101 point difference in their CPM result. In order to use strengths of different patches, fusion technique has been used. We observe that fusion method have achieved the best result compared with other individual networks. The fusion method increases the best of the four networks from 0.8215 to 0.8541. We believe that the performance of the proposed framework in this thesis is promising for future development.

Generally, from this thesis we noticed the following important points:

1. Data augmentation is an important technique to generate new data from the original ground truth nodules. In this thesis we perform that translation and rotation data augmentation methods. According to reviewed literatures [19], there are many types of data augmentation methods, zoom-in and zoom-out, shifting the nodule in different axes for a number of pixels have also been used as data augmentation techniques in literature.
2. Another point is that the receptive field of the input patch plays a significant role in the training process. Small and large patches focus on different characteristics on the training process, there for using the aggregate strength of both patches by using fusion technique gives a better result compared to each individual model.
3. CNNs require a huge number of training dataset, without large number of data network parameters only focus on certain patterns and cannot get successful results on the test dataset. Using CNN architecture requires powerful computational resources, hardware like GPU.

Thus required experimental setups must to be smart and powerful in order to get a successful result from the training process.

5.2 Future Work

We exploit different techniques to design our CNN architecture for pulmonary nodule classification and achieved an encouraging result. However, to improve the performance of nodule classification system we planned to incorporate the following points in the future work:

1. Using improved network architectures such as Inception modules or Capsule networks in the CNN architecture.
2. Using a large number of convolutional layers with a large number of filters to detect more complex features.
3. Employing different data augmentation techniques such as shifting and random zooming.

References

[1] "American cancer society". "Cancer statistics US". [online available] https://cancerstatisticscenter.cancer.org/#/. [accessed in January 2019]

[2] Kaggle Data Science Bowl 2017. https://www.kaggle.com/c/data-science-bowl-2017

[3] C. Jacobs et al., "Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images," Medical Image Analysis,vol. 18, no. 2, pp. 374–384, 2014.

[4] Darren Baker, Jen Kilpatrick, Ali Chaudhry," Predicting Lung Cancer Incidence from CT Imagery," CS 231N Final Project Report | Spring 2017

[5] Naidich DP, Bankier AA, MacMahon H, et al. "Recommendations for the management of subsolid pulmonary nodules detected at CT" a statement from the Fleischner Society. Radiology 2013;266(1): 304–317.

[6] Biegelman-Aubry, C., Hill, C. a Boulanger, X. " Evaluation of a computer aided detection system for lung nodules with ground glass opacity component on multidetector-row CT." Journals of Radiology. 90, 2009, stránky 1843-49.

[7] Muhammad Imran Razzak, Saeeda Naz and Ahmad Zaib "Deep Learning for Medical Image Processing: Overview, Challenges and Future "

[8] Satya Prakash Sahu, Narendra D. Londhe , and Shrish Verma "An Automated System for the Detection of Lung Cancer in CT data at Early Stages: Review" IJCTA, vol. 10, Nov 10,2017

[9] Armato, S.G.; Li, F.; Giger, M.L.; MacMahon, H.; Sone, S.; Doi, K. "Lung cancer: Performance of automated lung nodule detection applied to cancers missed in a CT screening program." Radiology 2002, 225, 685–692.

[10] Sluimer, I.; Schilham, A.; Prokop, M.; van Ginneken, B. "Computer analysis of computed tomography scans of the lung: A survey." IEEE Trans. Med. Imaging 2006, 25, 385–405.

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

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

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

[14] T. Messay et al., “A new computationally efficient cad system for pulmonary nodule detection in ct imagery,” Medical Image Analysis, vol. 14, no. 3, pp. 390–406, 2010.

[15] Arnold A. A. Setio, C. Jacobs, J. Gelderblom and B. van Ginneken, "Automatic detection of large pulmonary solid nodules in thoracic CT images," Medical Physics, vol. 42, pp. 5642-5653, 2015.

[16] K. Murphy et al., “A large-scale evaluation of automatic pulmonary nodule detection in chest CT using local image features and k-nearest neighbor classification,” Medical Image Analysis, vol. 13, no. 5, pp.757–770, 2009.

[17] Qi Dou, Hao Chen, Lequan Yu, Jing Qin, Pheng-Ann Heng, “Multi-level Contextual 3D CNNs for False Positive Reduction in Pulmonary Nodule Detection,” IEEE Trans. Medical Imaging, 2016

[18] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv: 14091556, 2014.

[19] Arnaud A. A. Setio, Francesco Ciompi, Geert Litjens, Paul Gerke, Colin Jacobs, Sarah J. van Riel, Mathilde Marie Winkler Wille, Matiullah Naqibullah, Clara I. S´ anchez, Bram van Ginneken “ Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks,” 0278-0062 (c) 2015 IEEE.

[20] L. Yann and B. Yoshua, “Convolutional networks for images, speech, and time series,” in The Handbook of Brain Theory andNeural Networks,vol.3361,no.10,1995.

[21] Firmino et al., “Computer-aided detection system for lung cancer in computed tomography scans: Review and future prospects,” Biomed Eng Online, vol. 13, pp. 1–16, 2014.

[22] F. Ciompi et al. “Towards Automatic Pulmonary Nodule Management in Lung Cancer Screening with Deep Learning. Scientific Reports,“ 7(46479), 2017.

[23] Thomas Bel et al. “Validation, comparison, and combination of algorithms for automatic Detection of pulmonary nodules in computed tomography images: “the LUNA 16n challenge.” 2016. [Online]. Available: https://luna16.grand-challenge.org/. accessed 2019

[24] S. G. Armato, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. R. Henschke and E. A. Hoffman, "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans.," Medcal Physics, vol. 38, no. 2, pp. 915-931, 2011.

[25] "World Cancer Research Fund International". URL: http://www.wcrf.org/int/cancer-facts-figures/worldwide-data. [accessed in January 2019]

[26] Am J Respir Crit Care www.thoracic.org "ATS Patient Education Series", American Thoracic Society Med Vol. 189, P1-P3, 2014.

[27] L. Ries et al. "SEER Cancer Statistics Review 1973–1996." National Cancer Institution, Bethesda, MD, 1999.

[28] M. Tan et al., "A novel computer-aided lung nodule detection system for ct images," Medical physics, vol. 38, no. 10, pp. 5630–5645, 2011.

[29] H. MacMahon et al., "Guidelines for management of small pulmonary nodules detected on ct scans: a statement from the fleischner society 1," Radiology, vol. 237, no. 2, pp. 395–400, 2005.

[30] 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.

[31] 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.

[32] Wogayehu Atilaw Mengesha , Menore Tekeba ,"Lung Nodules Detection from Computed Tomography Scans Using Deep Belief Networks " Addis Ababa University,October, 2018

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

[34] The Pulmonology Cannel. Solitary Pulmonary Nodule: Overview. [online] [cit. 20.12.2006], available at http://www.pulmonologychannel.com/spn/

[35] Mylene T. Truong, MD, Jane P. Ko, MD, Santiago E. Rossi, MD, Ignacio Rossi, MD, Chitra Viswanathan, MD, John F. Bruzzi, MBBCh, Edith M. Marom, MD, Jeremy J. Erasmus, MD " Update in the Evaluation of the Solitary Pulmonary Nodule " Volume 34 Number 6 Published online10.1148/rg.346130092 www.rsna.org/education/search/RG. 2014

[36] Martin Dolej, Dr. Ing. Jan Kybic, " Detection of Pulmonary Nodules from CT Scans " Available at: http://cmp.felk.cvut.cz/dolejm1/noduledetection/ January 19, 2007

[37] Biegelman-Aubry, C., Hill, C. a Boulanger, X. " Evaluation of a computer aided detection system for lung nodules with ground glass opacity component on multidetector-row CT."

Journals of Radiology. 90, 2009, stránky 1843-49.

[38] 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.

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

[40] "Radiographic Testing- NDT Inspection " https://www.twi-global.com/ [accessed 2019]

[41] 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.

[42] 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.

[43] Teramoto and Atsushi, "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.

[44] 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.

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

[46] Hiroshi Fujita et al.," An Introduction and Survey of Computer-aided Detection/Diagnosis (CAD)", 2010 International Conference on Future Computer, Control and Communication

[47] Dr. J. Thirumaran, S. Shylaja, "Medical Image Processing – An Introduction " International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

[48] Mokhled S. AL-TARAWNEH, "Lung Cancer Detection Using Image Processing Techniques" Leonardo Electronic Journal of Practices and Technologies ISSN 1583-1078 Issue 20, January-June 2012 p. 147-158

[49] Krishan A., "Evaluation of Gabor filter parameters for image enhancement and

segmentation", in Electronic Instrumentation and Control Engineering, Master. Punjab: Thapar University, 2009, p. 126.

[50] Deserno et al. "Fundamentals of biomedical image processing," Biomedical Image Processing. Springer Berlin Heidelberg, pp. 1-51, 2010.

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

[52] Awais M, Ulas B., Ziyue Xu, Brent F.,Kenneth N., Jason M., Anthony F., Jayaram K., Daniel J. "A Generic Approach to Pathological Lung Segmentation" IEEE Trans Med Imaging. 2014 Dec; 33(12): 2293–2310.

[53] Pablo G. et al., "Lung nodule segmentation in chest computed tomography using a novel background estimation method", 2016, Quantitative Imaging in Medicine and Surgery.

[54] https://en.wikipedia.org [online available] accessed 2019

[55] RaghavPrabhu, "Understanding of Convolutional Neural Network (CNN)—Deep Learning", https://medium.com [online available] accessed 2019

[56] Shuo Wang, Mu Zhou, Zaiyi Liu, Zhenyu Liu, Dongsheng Gu, Yali Zanga, Di Dong, Olivier Gevaert, Jie Tian "Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation", This is an open access article under the CC BY license. (http://creativecommons.org/licenses/by/4.0/) 2017

[57] C. Szegedy et al."Going deeper with convolutions," 2014. [Online Available] https://arxiv.org/pdf/1409.4842.pdf.

[58] F. Rosenblatt, "The Perceptron: A Probabilistic graphical model for information storage and organization in the brain," Psychological Review, vol. 65, no. 6, pp. 65-386, 1958.

[59] Anish Sing Walia, "Activation functions and its types- Which is better?", " Types of Optimization Algorithms used in Neural Networks and Ways to Optimize Gradient Descent " 2017. [Online Available] https://towardsdatascience.com/

[60] A. Krizhevsky et al., "Imagenet classification with deep convolutionalneural networks," in NIPS, 2012, pp. 1097–1105.

[61] Anton D., Ramil K, Adil K, Adin R, Asad M, " Large Residual Multiple View 3D CNN for False Positive Reduction in Pulmonary Nodule Detection ", IEEE 2017

[62] Jason Brownlee, "How to Develop a Weighted Average Ensemble for Deep Learning Neural Networks", [online] https://machinelearningmastery.com/weighted-average-

ensemble-for-deep-learning-neural-networks/ [accessed Jully 2019]

[63] Coronel S,"Ensembles of deep networks" [online] https://towardsdatascience.com/ensembles-of-convolutional-networks-3f81f59978a3 [accessed May 2019]

[64] Moira berens, et al. "ZNET - lung nodule detection", Radboud University Nijmegen, 2017

[65] Görkem Polat "Classification of lung nodules in ct images using convolutional neural networks", Middle East Technical University, January 2018