DETECTION AND DESCRIPTION OF PULMONARY NODULES THROUGH 2D AND 3D CLUSTERING

AMERA ABDULWOHID FUNJAN AL-FUNJAN

Thesis Submitted in Partial Fulfilment of the Requirements
of the Degree of Doctor of Philosophy
School of Computing, Science and Engineering
University of Salford, Salford, UK

2019

Contents

CHAPTER 1: INTRODUCTION 1

1.1 Introduction 1

1.2 Research Motivation 3

1.3 Aim and Objectives 4

1.4 Research Questions 5

1.5 Research Contributions 5

1.6 Thesis Organization 6

CHAPTER 2: RESEARCH METHODOLOGY 8

2.1 Introduction 8

2.2 The scope of the Research 8

2.3 Research Methodology 9

2.4 Summary 13

CHAPTER 3: LITERATURE REVIEW AND BACKGROUND 14

3.1 Introduction 14

3.2 Introduction to Medical Imaging 14

3.3 Lung Anatomy on CT-Scan 14

3.4 CT and MRI-Scans of Lung 15

3.5 Pulmonary Nodule Types 16

3.6 Image Pre-Processing Techniques 18

3.7 Image Conversion 19

3.8 Image Enhancement 19

3.9 Lung Segmentation Techniques 20

3.9.1 Threshold-Based Lung Segmentation 21
3.9.2 Deformable Models -Based Lung Segmentation 22
3.9.3 Edge Detection -Based Lung Segmentation 23
3.9.4 Clustering -Based Lung Segmentation 23

3.10 Nodules Segmentation Techniques 27

3.10.1 Nodule Segmentation Based on Clustering 27
3.10.2 Nodule Segmentation Based on Active Models 28
3.10.3 Nodules Segmentation Based on Feature Extraction and Classification 29
3.10.4 Features Selection and Dimension Reduction Techniques 32
3.10.5 Pattern Recognition Techniques 33

3.11 Computer-Aided Detection/Diagnosis Systems (CADs) 45

3.12 Advantages and Disadvantages of Detection Techniques 48

3.13 Summary and Lessons learned from the Literature Review 49

CHAPTER 4: DATA COLLECTION AND SEGMENTATION OF CT IMAGES 51

4.1 Introduction 51

4.2 Data Collection 53
4.2.1 Standard Data 53
4.2.2 DICOM Standard 54
4.2.3 Data Diagnosis Reports 56

4.3 Pre-Processing 57
4.3.1 CT Image Enhancement 58
4.3.2 Lung Segmentation (Fully Automated) 58

4.4 Lung Segmentation Evaluation 62

4.5 Experiments and Discussion 63
4.5.1 Denoising and Adjustment 64
4.5.2 Image Histogram 65
4.5.3 Lung Segmentation 66

4.6 Lung Segmentation Evaluation 68

4.7 Post Processing 69
4.7.1 Vessels and Nodules Segmentation Together 69

4.8 Vessels and Nodules Segmentation Results 74

4.9 Summary 76

CHAPTER 5: NODULE CLASSIFICATION 77

5.1 Introduction 77

5.2 Image Features 77
5.2.1 Features Analysis 78

5.3 Classification phase 83
5.4 Validation Phase 83
5.5 The Evaluation Phase 85
  5.5.1 Confusion Matrix 85
  5.5.2 Other Metrics 86
5.6 Results and Discussions 87
5.7 Summary 109
CHAPTER 6: 3D CLUSTERING AND MEASURING NODULE VOLUME 111
6.1 Introduction 111
6.2 3D Clustering Advantages 112
6.3 Research Materials 112
6.4 3D Clustering 112
6.5 Case Study 115
6.6 3D Clustering Evaluation 119
  6.6.1 DBSCAN Method 120
  6.6.2 Plot.ly Application 122
6.7 Measuring the Volume 125
6.8 Nodule Volume Measuring Methods 126
  6.8.1 Manually Nodule Volume Calculation 127
  6.8.2 Converting the Area Unit from Pixel to MM 128
  6.8.3 Automated Nodule Volume Measurement 130
  6.8.4 3D-DBSCAN Performance in Automated Volume Measuring and Evaluation of 3D Clustering 132
6.9 Comparison 137
6.10 Summary 139
CHAPTER 7: SYSTEM DESIGN AND IMPLEMENTATION 140
7.1 Introduction 140
7.2 Main Menu 140
7.3 Loading and Enhancing Image 141
7.4 Lung Segmentation 143
7.5 2D-Clustering 143

7.6 Feature Extraction and Classification 144

7.7 3D-Clustering 147

7.8 Measuring Volume 147

7.9 Summary 148

CHAPTER 8: CONCLUSION AND FUTURE WORKS *149*

8.1 Introduction 149

8.2 Review of Methodology 149

8.3 Conclusions 151

8.4 Research Limitations 151

8.5 Future Works 152

8.6 Summary 152

References *153*

LIST OF FIGURES

Figure 1-1: Distribution for all Cancer, Males and Females Combined According to Incidence Rate Ratio, Ninawa/Iraq 2001-2010 (AL-Hashimi and Wang 2013) 4

Figure 2-1: Research Methodology; Main Stages 11

Figure 3-1: Images of CT Lung Slices: (a) CT Planes Passing Through the Body; (b) Coronal CT Slice of Lung; (c) Sagittal CT Slice of Lung; (d) Axial CT Slice of Lung (Sluimer 2005) 15

Figure 3-2: pulmonary nodules types' classification (Sluimer 2005) 17

Figure 3-3 The Axial Image Slices of Nodule Types: (a) Solid Nodule; (b) Part-Solid Nodule; (c) Pure Ground Glass Nodule (Sluimer 2005) 18

Figure 3-4: CT Lung Slice Saved in DICOM Format with Dimensions of 512×512 Pixels 19

Figure 3-5: Image Histogram of Two Thresholds, Region 1 belongs (0–50), Region 2 belongs (50–150), and Region 3 belongs (150–255) of Grey Scale Values 21

Figure 3-6 Individual segments of lung CT image: (A) original, (B) segmented image in pseudocolour, (C–E) individual segments 26

Figure 3-7: An Overview of Shape Description Techniques (Ping et al., 2013). 30

Figure 3-8: Illustration of Linear Regression on a Data Set (Draper and Smith 2014) 37

Figure 3-9 Two Line Segments to Fit the Dots in Global Logistic Regression(Hsieh et al., 1998) 38

Figure 3-10 (Global) Logistic Regression for Classification :(a) a Curve of Global Logistic through Data (b) a Curve of Global Logistic through Noise (c) a Curve of Global Logistic does not Work Because of more Than Two Segments (Hsieh et al., 1998) 38

Figure 3-11: Hyperplane Separates Two Classes *(Al-Waeli 2017)* 40

Figure: 3-12 Multilayer Perceptron Configuration (Hu and Hwang, 2001) 43

Figure 4-1 General Block Diagram of the Detection System 51

Figure 4-2 Block Diagram of the Proposed System Processes 52

Figure 4-3 a) Human Lung; b) Continuous HLCT Images; c) Normal HLCT Slice (Huaqing & Chen, 2012) 54

Figure 4-4: DICOM Tag 55

Figure 4-5 Sample of Dataset before Updating 56

Figure 4-6 Sample of Dataset after Updating 57

Figure 4-7: Block Diagram of Lungs Segmentation by Fast Fuzzy C-means and Morphological Operations 61

Figure 4-8: Overall Block Diagram of Lung Segmentation 62

Figure 4-9: Denoising and Adjusting of CT- Scan Lung Image, A) Original CT Lung Image; B) Denoising CT Lung Image by Gaussian Filter; C) Increasing Contrast of CT Lung Image by Adjusting Algorithm and D) Denoising and Increasing contrast of CT Lung Image by Gaussian Filter and Adjusting Algorithm 64

Figure 4-10: Corresponding Histograms Representation for Each One of a) Denoising CT lung image by Gaussian Filter; B) Increasing Contrast of CT Lung Image by Adjusting Algorithm and C) Denoising and Contrast Increasing of CT lung image by Gaussian Filter and Adjustment Algorithm 65

Figure 4-11: CT Lung Image with Pseudo Colour of Algorithm 66

Figure 4-12: Fast Fuzzy C-means Algorithm Applied on CT Lung Image, Shows Pixels Belonging Degree to Each Cluster in the Image 67

Figure 4-13: Samples of CT Original Lung Images and Segmented Lungs Images by Fast Fuzzy C-Means 67

Figure 4-14: CT Lung Images Delineated the Boundaries with Green Colour by Radiologist and Segmented of Background by Fast Fuzzy C-Means Method 68

Figure 4-15 Flow Chart for Image Intensity- Based Clustering 73

Figure 4-16: a) Clustering with 45 Threshold; b) Clustering with 90 Threshold; c) Clustering with 120 Threshold 76

Figure 5-1: Accuracy Rates of Five Classifiers Through 4-Cross-Validation 87

Figure 5-2: Accuracy Rates of Five Classifiers Through 10-Cross-Validation 88

Figure 5-3: Accuracy in Classifying Vessels and Nodules of Four Classifiers 89

Figure 5-4: Accuracy in Classifying Vessels and Nodules of Naïve Bayes Classifier 89

Figure 5-5: Accuracy in Classifying Vessels and Nodules of SVM-poly 4-Cross-Validation 90

Figure 5-6: Accuracy in Classifying Vessels and Nodules of SVM-POLY 10-Cross-Validation 90

Figure 5-7: Accuracy in Classifying Vessels and Nodules of SVM-RBF in 4-Cross-Validation 91

Figure 5-8: Accuracy in Classifying Vessels and Nodules of SVM-RBF in 10-Cross-Validation 91

Figure 5-9: Accuracy in Classifying Vessels and Nodules of SVM-PUK in 10-Cross-Validation 92

Figure 5-10: Recall of Nodules for all Classifiers 93

Figure 5-11: TNR Vessels for all Classifiers 93

Figure 5-12: FNR of Vessels for all classifiers 94

Figure 5-13: FPR of Nodules for all classifiers 94

Figure 5-14: Precision Measure of Nodules and Vessels for all classifiers and Average of both 95

Figure 5-15: F- Measure Score for all classifiers 96

Figure 5-16: MCC Measure for all classifiers 97

Figure 5-17: AUC Measure for all classifiers 97

Figure 5-18: ROC Curve for TPR & FPR of all classifiers 98

Figure 5-19: Curves of Circularity Values for Vessel and Nodule 102

Figure 5-20: Curves of Ellipticity Values for Vessel and Nodule 103

Figure 5-21: Curves of Slenderness Values for Vessel and Nodule 104

Figure 5-22: Curves of Area Values for Vessel and Nodule 105

Figure 5-23: Curves of Perimeter Values for Vessel and Nodule 106

Figure 5-24: Curves of Compactness Values for Vessel and Nodule 107

Figure 5-25: Curves of Rectangle Degree Values for Vessel and Nodule 108

Figure 6-1: Original CT Lung Images of Slices 60-64 with Red Referral of Two Attached Nodules and Assessed by 4 Experts, a. Candidate Nodule in the 60 slice; b, c, d. Similar Structures of Nodules through 61, 62, 63; e. Candidate Nodule in 64 Slice, f. The Normal Image after 64 Slice 117

Figure 6-2: 3D Clustering of CT lungs with Referral to Two Diagnosed and Attached Nodules in 60 and 64 Slices 118
Figure 6-3: 3D Clustering of Attached Nodules in 60 and 64 slices and Another One in the 46 slice with Centre Pixels Corresponding Radiologists' Assessment 119
Figure 6-4: 3D-DBSCAN Method Clusters the Linked Nodules in Yellow and the Remaining Nodules with Different Colours of the Same Patient 121
Figure 6-5: 3D-DBSCAN Method Clusters the Linked Nodules with Yellow Colour, Centres Pixels, Which Were Reported in Radiologist Evaluation, and the Remaining Nodules with Different Colours 121
Figure 6-6: DBSCAN Method Just Clusters the Linked Nodules and Proves One Nodule 122
Figure 6-7: 3D Plot of the Areas of Nodules with Similar Structures Between Different Levels from Slice 60 to 64 (going upwards) 123
Figure 6-8: 3D Plot from Different Angle to Figure 6-7, Black Areas (Bubbles) Represent the Overlap Between Them to Show the Connected Regions 124
Figure 6-9: (A, B) Show the Overlap between Centre Pixels of Clustered and Actual Nodule in 60 Slice with Two Colours Red and Green Respectively 125
Figure 6-10: Pixel Spacing Values against of Two Pixels Centres in Vertical and Horizontal of 2D Pixels Matrix (Treichel et al. 2012) 126
Figure 6-11: (a) Frustum Model, (b) CT Slices Contain Nodule's Structures *(Mahmood, Abbas, and Ali 2014)* 128
Figure 6-12: Segmented and Clustered CT Lung Images for Nodule Starts from Slices 60 to 64 with Column for Convex Areas in Pixels and MM of Nodules, Manually Calculated Volume 129
Figure 6-13: Pseudo-Code for Automatic Nodule Volume 133
Figure 6-14: Nodule and False Positives Clustered by 3-DBSCAN and Measuring the Volume for each Cluster 134
Figure 6-15: The Centre Pixels for Nodules in Slices 46, 60 and 64 135
Figure 6-16: the Centre Pixel of Nodule in the Slice 12 135
Figure 6-17: A. 3D- DBSCAN Clustering of 2D Classification Results, B. Identifying the Attached Nodules Location in DBSCAN Figure, C. The Cluster of Attached Nodules are Clustered in B in a 3D View 137
Figure 7-1: Main Menu of Proposed System 141
Figure 7-2: Loading and Enhancing Image 142
Figure 7-3: Lung Segmentation Stage 143
Figure 7-4: 2D-Clustering 144
Figure 7-5: Results and Execution Time before Features Reduction for 4,10-Cross Validation 145
Figure 7-6: Results and Execution Time after Features Reduction for 4,10-Cross Validation 146
Figure 7-7: 3D-Clustering 147
Figure 7-8: Measuring Volume Stage 148

LIST OF TABLES

Table 3-1: Description and definition regarding pulmonary nodule types 17
Table 4-1: Hausdorff Values for Ten Segmented Lung Images 68
Table 5-1: Accuracy Values for Five Classifiers Through 4-Cross-Validation 88
Table 5-2 Accuracy Values for Five Classifiers through 10-Cross-Validation 88
Table 5-3: Time Values for Five Classifiers through 4-Cross-Validation before and after Features Reduction 99
Table 5-4: Time Values for Five Classifiers through 10-Cross-Validation before and after Features Reduction 99
Table 5-5: Samples of Features Vectors for Nodule and Vessels 101
Table 5-6: Circularity Values of Vessel and Nodule 102
Table 5-7: Ellipticity Values of Vessel and Nodule 103
Table 5-8: Slenderness Values of Vessel and Nodule 104
Table 5-9: Area Values of Vessel and Nodule 105
Table 5-10: Perimeter Values of Vessel and Nodule 106
Table 5-11: Compactness Values of Vessel and Nodule 107
Table 5-12: Rectangle Degree Values of Vessel and Nodule 108
Table 5-13: Accuracy, Sensitivity and Specificity Rates for Four Optimal Classifiers 109
Table 6-1: Other works compared with this study in terms of accuracy and measurements the nodule characteristics automatically. 138

Acknowledgement

First of all, my great thanks to ALLAH for his mercy and blessing. Also, this thesis would never be completed without the excellent supervision and guidance that I received from Prof. Dr Farid Meziane and Dr Rob Aspin and Prof. Dr Eman AL-Shamery. There are no words in the English dictionary can express my gratitude to them. Also, I would like to thank my sponsors, the Iraqi Ministry of Higher Education and Scientific Research and the University of Babylon for funding my PhD study. I would say heartfelt thanks to my mother for her prayers, which bless me all the time. Besides, I would like to pay my regards to all relatives and friends, especially (Prof. Dr Eman AL-Shamery and Assist. Prof. Dr Abdul Sameea Alkilabi) for the help that they showed. I would like to express my profound respect and sincere appreciation to my husband "Aqial", my sister and my children for their patience and encouragement that gave me so much hope and support.

To my Mother’s Soul

To my Husband

To my Sister and Sons

ABBREVIATIONS

CADs	Computer Aided-Detection Systems
CT	Computed Tomography Imaging
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DICOM	Digital Communications in Medicine Services
FFCM	Fast Fuzzy C-Means
FPR	False Positive Rate
GGN	Ground-Glass Nodule
IDRI	Image Database Resource Initiative
IIBC	Image Intensity-Based Clustering
LIDC	Lung Image Database Consortium
LR	Logistic Regression
MK-M	Modified K-Means
MLP	Multi-Layer Perceptron Neural Network
MRI	Magnetic Resonance Imaging
MCC	Matthews Correlation Coefficient
NB	Naive Bayes
PSN	Part-Sold Nodule
PGGN	Pure Ground-Glass Nodule
RBF	Radial Basis Function
SSN	Sub-Sold Nodule

SVM	Support Vector Machine
TNR	True Negative Rate
TPR	True Positive Rate

Abstract

Precise 3D automated detection, description and classification of pulmonary nodules offer the potential for early diagnosis of cancer and greater efficiency in the reading of computerised tomography (CT) images. CT scan centres are currently experiencing high loads and experts shortage, especially in developing countries such as Iraq where the results of the current research will be used. This motivates the researchers to address these problems and challenges by developing automated processes for the early detection and efficient description of cancer cases. This research attempts to reduce workloads, enhance the patient throughput and improve the diagnosis performance. To achieve this goal, the study selects techniques for segmentation, classification, detection and implements the best candidates alongside a novel automated approach. Techniques for each stage in the process are quantitatively evaluated to select the best performance against standard data for lung cancer. In addition, the ideal approach is identified by comparing them against other works in detecting and describing pulmonary nodules. This work detects and describes the nodules and their characteristics in several stages: automated lung segmentation from the background, automated 2D and 3D clustering of vessels and nodules, applying shape and textures features, classification and automatic measurement of nodule characteristics. This work is tested on standard CT lung image data and shows promising results, matching or close to experts' diagnosis in the nodules number and their features (size/volume, location) and in terms the accuracy and automation. It also achieved a classification accuracy of 98% and efficient results in measuring the nodules' volume automatically.

CHAPTER 1: INTRODUCTION

1.1 Introduction

Radiology and medical imaging are efficient technologies used to diagnose inner body diseases by offering an internal insight and visual representation of organ tissues and recognising abnormalities even in complex cases. The threat imposed by serious diseases such as tumours to human health increases the importance of using medical imaging in diagnosis. Consequently, this technology has become a subject of interest in other fields, which can improve diagnosis with the development of analysis and processing techniques applied to medical images. Computer science and medical image processing and analysis are combined to extract vital information that is relative to pathological patterns (Deserno 2011). Within these fields, systems and applications that are able to automatically and precisely detect tumours are developed, and they can compete with the diagnosis made by radiologists in terms of precision and speed when using medical images. In addition, they address the challenges and problems faced by CT centres resulting from the increasing number of lung cancer patients around the world (Salim et al., 2011) and (AL-Hashimi and Wang 2013). Therefore, automated detection systems are sought to be adopted by those centres in order to reduce the workload, improve throughput and accurate detection of nodules in the early stages of lung cancer without the help of specialists. These systems are based on the segmentation and classification of lung cancer (pulmonary nodules) in 2D and 3D medical images (CT lung images).

This introduction addresses the key areas of this study. Some of the relevant questions that this research is attempting to answer are discussed below.

What is the main subject of this work?

The subject of this study is concerned with the precise detection of pulmonary nodules in CT lung images to be conducted automatically, without human intervention. This area includes various approaches in the computer science and image processing fields (Eadie et al., 2012). The work is based on the segmentation, detection and classification of both 2D and 3D CT lung images to correspond to the automatic system requirements.

- Why is it timely to consider this now?

Currently, lung cancer is a common disease throughout the world, especially in developing countries because of high rates of smoking, pollution, wars and accidents. Consequently, lung cancer patients' need for healthcare has increased and generated a working pressure on CT centres. Requirements for greater numbers of radiologists and experts and large budgets have led to the lack of follow-up and early diagnosis, denying patients of higher chances of survival by not receiving effective treatments (Lederlin et al. 2013). Furthermore, the work pressure causes a rush in the reading of images by radiologists who may miss small nodules or complicated ones that are hidden in dense tissues. Moreover, the time-consuming process spent on normal cases leads to a reduction in the CT centres' throughput. An automatic detection system is considered as a worthwhile and ultimate solution for most of these issues.

-Who will potentially benefit from the research presented here?

This study may benefit both patients and radiologists. For the patient, the early discovery of cancer raises the possibility of controlling on the disease, which can possibly help their recovery with appropriate treatments. Improving throughput of patients will be achieved via the use of a system detects the nodules automatically.

Automated systems offer several benefits to radiologists. The precise and early detection of nodules by systems increase the confidence of radiologists with their diagnosis. Furthermore, because the systems usually reject normal cases through a previous reading of images before the radiologist scan, it ultimately reduces the workload of radiologists and CT centres. Another benefit of the automated systems to both radiologist and patient is detecting the complicated nodule patterns, which are typically hard to be recognised in 2D images which based by radiologists for diagnosis purposes. Moreover, an automatic form of detecting and classifying pulmonary nodules helps to avoid the need for additional radiologist evaluations.

- How will the research be conducted?

This work attempts to develop an approach that can introduce a more efficient solution as compared with other works in this field. The automatic system performance will be validated, evaluated and compared using appropriate metrics and methods.

1.2 Research Motivation

Lung cancer is a fatal disease from which many people around the world suffer, particularly in Iraq. In an evaluation for all cancer types in Ninawa, which is the second biggest province in Iraq after Baghdad (the Capital of Iraq), the lung cancer accounted for around (13.6%) of all cancer cases, as shown in Figure 1.1. This was due to smoking, pollution and accidents that have risen from the significant environmental pollution during the First and Second Gulf Wars in 1990 and 1997 respectively (Al-Rahim, Ch, and Cm 2007) and (AL-Hashimi and Wang 2013). The motivations behind constructing an automatic detection system are due to two pressing issues: the health services deterioration and diagnosis challenges faced by the radiologists despite the availability of the CT-scan technology and the impact of the increase in the number of cancer patients who visit the CT-scan centres. This pressure and time limit hinders precise diagnosis and patient follow-up and requires more experts and CT centres. Equally, losing time and effort in dealing with normal cases ultimately reduce the patient's throughput in CT centres. Consequently, these factors have led to the degradation of medical services in CT centres. As for the second issue, the challenges that face radiologist in diagnosis are present too, such as the variance of lung tissue, nodule size (smallness) and the lack of precision when considering the difficulty in measuring nodule features. The above problems and the significant development of CT-scan technology have motivated researchers in computer science and image processing to construct automated systems (Sieren et al. 2010). These systems, which exploit CT technology development, are considered by radiologists in CT centres as applicable and reliable, based on segmentation and classification in detecting pulmonary nodules automatically and precisely. It offers early nodule detection that ensures proper treatment management of tumours by radiologists reduces workload by rejecting normal cases and therefore avoids the need for additional experts. Furthermore, improving the throughput of patients will be conducted in CT centres and consequently enhance the diagnosis performance notably. Moreover, to describe the nodule progress, the systems ensure a precise measurement for nodule characteristics, which are vital in guiding treatment management (resection, follow-up or treatment) of the nodule. In addition, the systems provide accurate detection of complicated cases and small nodules that the radiologist may miss through the rushed reading of images. Therefore, an automatic detection system offers solutions for both problems in terms of the tumour detection precision and full automation.

However, the requirements of the optimal systems are still not addressed and need more investigation (Javaid et al. 2016) and (Kuruvilla and Gunavathi 2014).

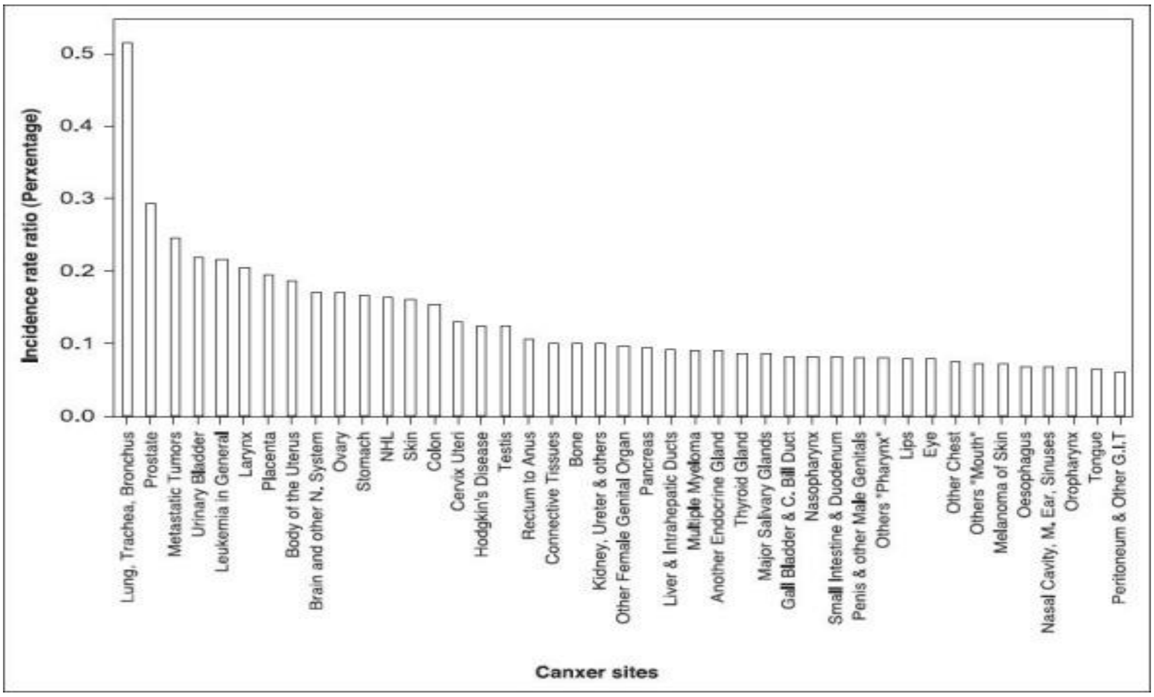


Figure 1-1: Distribution for all Cancer, Males and Females Combined According to Incidence Rate Ratio, Ninawa/Iraq 2001-2010 (AL-Hashimi and Wang 2013)

1.3 Aim and Objectives

The main aim of this study is to propose a new system to detect the tumours in CT lung images automatically. Then, identifying the nodule numbers that show the disease spreading, determining the location and shape which provide information about tumour type. In addition, measuring the size that refers to nodule progress. The system has matched the assessments made by radiologists. This aim is achieved using techniques based on segmentation, classification and detection in 2D and 3D CT lung images.

Specifically, the research focuses on the following objectives to address the previously identified problems:

1. Segmenting the lungs from background according to qualitative and quantitative evaluation.

2. Extracting computation characteristics of the nodule to reflect diagnostic information which determines the treatment management and the malignancy rate.

3. Detecting the complicated cases that radiologists have missed in the diagnosis.

4. Constructing a proposed system which corresponds to most Computed Aided Detection (CAD) requirements regarding false positive reduction, precise nodules number and types even the nodules that are difficult to be recognized by radiologists such as small and glass type.

1.4 Research Questions

In addition to the questions answered in the introduction section, this attempts to address the following academic questions:

1. What are the best techniques for determining the segmentation, classification and detection of pulmonary nodules?

2. Do the best techniques offer the best-integrated solution for the automated system?

3. Do the best techniques used in the automated system accurately identify the different tumours in CT lung images?

1.5 Research Contributions

A new approach is proposed to detect and describe the pulmonary nodules in CT images with high accuracy and automatically based classification and detection in 2D& 3D of CT lung images. The novelty and specific contributions of this study are summarised as follows:

- An Image Intensity-Based Clustering (IIBC) is proposed a new algorithm for 2D images to separate the nodules and vessels from normal tissue of lungs to concern the search space.

- A new 3D-DBSCAN algorithm is introduced to cluster the nodule depending on location (x,y) and thickness instead of intensity values that are used in the previous works. Consequently, measuring volume has automatically become possible by applying the trapezoid method while the volume has been measured manually in traditional ways of radiologists.

- The detection of complicated cases of nodules that radiologists have missed through the diagnosis, which indicates deep view and correct decision.

1.6 Thesis Organization

This thesis is organised in eight chapters, and these are summarised as follows:

Chapter 1:

This chapter introduces the subject and the motivation for the study, followed by the aim and objectives, research questions and the main contributions of the study. The chapters are organised in sequence with providing a summary of signification and content of the different chapters in the thesis.

Chapter 2:

This chapter is concerned with the methodology that is applied in the study; it describes the research framework and defines the scope and methodology objectives.

Chapter 3:

This chapter introduces a background of medical images and survey of previous works on automated analysis and nodule pattern recognition of CT-scanning images, including segmentation, feature extraction, classification techniques and 2D and 3D clustering. Also, it surveys about computer-aided detection/diagnosis systems in medical images processing field. It finally reviews the various methods and their characteristics, advantages and disadvantages concerning this work.

Chapter 4:

This chapter discusses data collection, image processing, and lung segmentation as a pre-processing task, with presentation and discussion of results. The chapter also includes post-processing with the CT lung image-clustering algorithm in 2D and discusses the results.

Chapter 5:

This chapter displays the stages of image features extraction, the classification, validation, accuracy rates and an evaluation of classification.

Chapter 6:

This chapter evaluates the performance of 3D techniques compared with 2D in describing nodule characteristics, measuring the nodule volume, identifying the complex case of the nodules and discusses the outcomes.

Chapter 7:

The chapter displays a complete software supported with graphical user interfaces review the implementation stages of work systematically to be as guide for researcher and radiologist during dealing with the system.

Chapter 8:

This chapter summarises the outcomes in conclusions for this study, reviews the methodology with the limitations and the future suggestions potentials of the work.

CHAPTER 2: RESEARCH METHODOLOGY

2.1 Introduction

The purpose of this work is to construct an automated system based on detecting and classifying CT lung tumour images with high precision levels with the aim of detecting the characteristics of the tumour. It seeks to meet the criteria reported by current works to measure system accuracy in detecting tumours (Valente et al. 2016). This system aims to deliver an automated solution to improve diagnosis performance.

2.2 The scope of the Research

This work is concerned initially with the automated detection of lung tumours (pulmonary nodules) with high levels of precision, based on classification and detection using 2D&3D CT images. It mainly focuses on the standard dataset of lung cancer patients and, in particular, axial images. The images were diagnosed and reported by four experts in two reports that accompany the dataset obtained from the public Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI) (Samuel et al., 2011) and (Clark et al. 2013). In the first report, nodules are numbered and classified by size discriminating, between those greater and smaller than 3mm in diameter for each patient. In the latter, the report was updated only to include information about the nodules that are greater than 3mm.The case report consists of the patient ID and the number, volume, diameter, location and slice number for each nodule in the updated file. One of the targets of this work is to compare and match radiologists' assessments and the automatic detection system output, at least in terms of nodule numbers and location. In addition, the nodule characteristics detection accuracy is an interesting point in the diagnosis because they play a vital role in identifying tumour progress. In recent years, computer-aided detection/diagnosis systems (CADs) have made significant improvements in medical image processing. CADs have become the focus of many researchers in this field, where the objectives are: to increase detection precision, false positive reduction, full automation and speed. However, all efforts to build such systems could not address all CAD requirements and achieving a level of precision required in the automatic detection of tumours. The work is aimed at constructing a system based on the detection and classification of 2D and 3D CT images in an automated manner. In addition, it includes an approach that delivers precision levels in detecting and exactly matching or being close to radiologist assessment in the nodules number and location. This modified approach will be evaluated against previous works' findings for a final comparison. The work is based on understanding the characteristics of the datasets and automatic detection features required to construct an optimal system. Therefore, the research scope includes a definition of tools and techniques that efficiently detect tumour patterns with the intention to form a framework based on automatic detection and classification. Moreover, the research techniques used in this study will be compared with other approaches reported in the literature.

2.3 Research Methodology

Many image-processing tools can be applied in this research to solve the problem-domain-specific challenges described in the literature review of other works. The definitions of process/method/algorithm and framework are known within the field of image processing will be reviewed, and this will include problem analysis, process implementation and automation (Deserno 2011), (Ramesh et al., 2004), (Liao 2005) and (Ayash 2014).

Some interesting points in this work are the overlap between the approaches based on image processing techniques and the development in the medical modality imaging, as the initial conceptual implementation draws from both. This research aims at creating and evaluating a novel approach towards high precision levels in tumour detection. Furthermore, it comprehensively addresses CADs requirements necessary for constructing an optimal detection system. This new approach presents an automatic system to improve diagnostic performance by offering higher precision levels in the detection and description of nodules. Among the objectives of the research are improvements in diagnosis turnaround times. The objectives provide solutions to problems reported in previously published work (Van Rikxoort and Van Ginneken 2013). The improvements produced in this work will frequently be assessed and evaluated through quantitative analysis.

The starting point for this work is to draw from a current understanding of medical image diagnosis problems, CADs requirements, datasets and the impact of detection precision on diagnostic performance. The approaches will be detailed in the literature review, which sets out state of the art (Senthil Kumar et al., 2017), (Oseas et al. 2014) and (Javaid et al. 2016) regarding nodule detection. As a part of the literature review, there is a significant need for the investigation, study and understanding of the nature of the datasets to identify the tools and techniques that need to be developed. Also, an identifying group of approaches based on the same dataset of this study in order to compare the new approach performance. Many published articles present some analysis of tumour detection systems in terms of accuracy levels (Valente et al. 2016). However, these publications did not offer ideal solutions to cancer detection problems, and neither do they fully address the detection system requirements. To summarise, a precise and full automated framework, which supports the research of an integrated system, is absent in these publications. The automated framework of nodule detection is constructed and guided by the literature review in this work. Also, it adopts existing principles and includes efficient evaluation metrics.

Continuous development of selected and potential techniques within an evaluated framework is established through objective analysis to find the best technique. The improvement of the approach is compared against current research based on the same datasets. The accuracy of the detection system is assessed at each stage to determine the progress in addressing the CADs requirements. This processing of development in approach intends to determine the matching of approach with a goal. However, the ultimate aim of this research is to develop an approach that will be comparable to current methods, while allowing the possibility of future enhancement. To attain this, evaluation of the technique must be in comparison to the existing approaches in terms of available metrics required for computer-aided detection/diagnosis systems, in applications mapped to the problem domain.

Figure 2.1 summarises the research methodology adopted in this research. The research starts with understanding the problem domain and defining candidate solutions arising from state of the art and the literature review to establish the best solutions for the detection and classification. In this framework, the research questions and objectives are intersected to be addressed in the methodology structure of this study. The best solutions are finally evaluated and validated. Research questions1, 2 and 3 are answered to achieve the aim. The methodology will be reviewed through a discussion at the end of the thesis. The figure also shows how the aim was accomplished according to the methodology, to develop an automatic detection system of lung tumours with a high level of precision and accuracy.

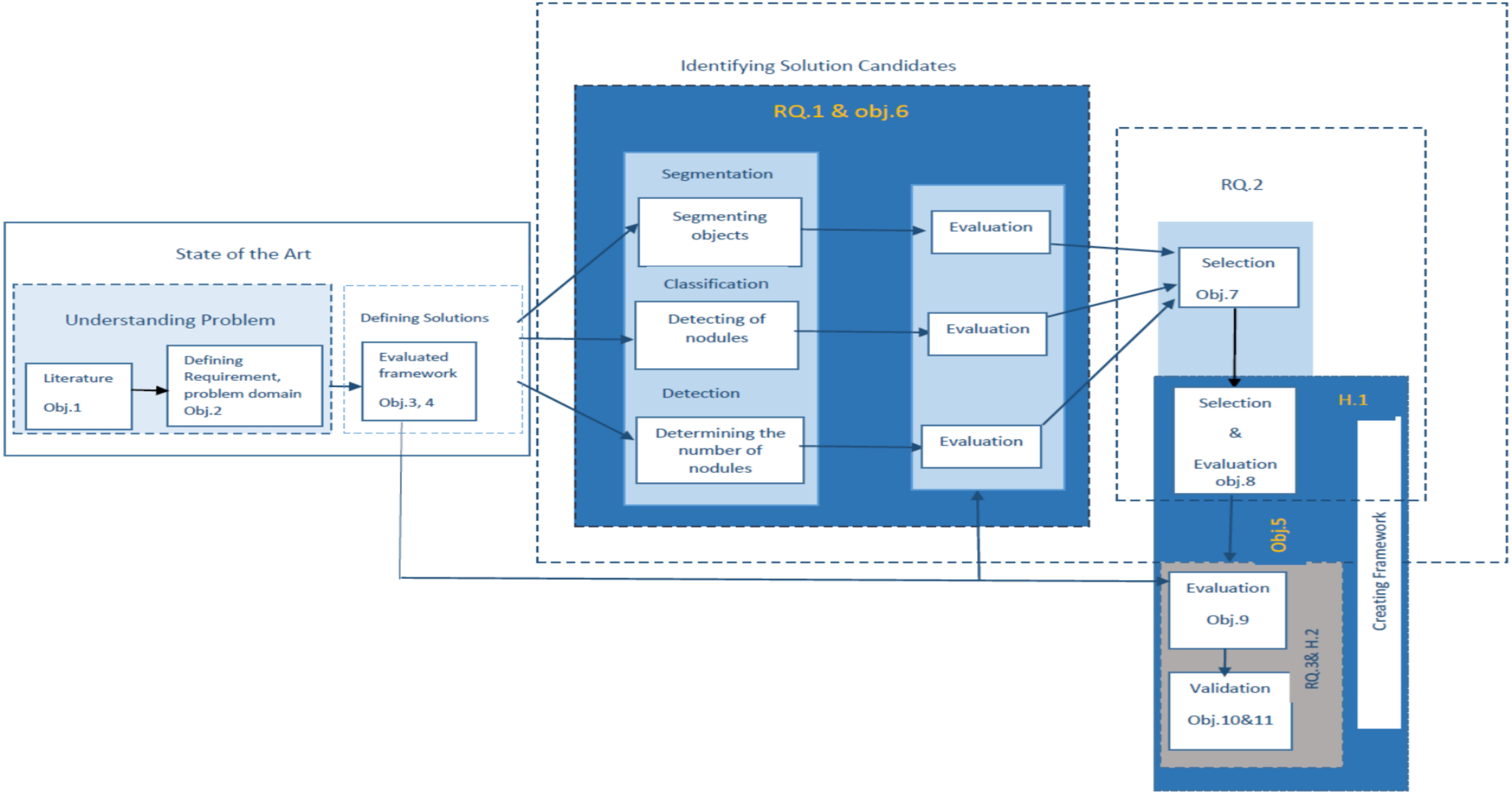


Figure 2-1: Research Methodology; Main Stages

The research is focused on the following methodology objectives to deal with the previously identified problems:

1. A comprehensive review of the literature on lung cancer detection systems, to define the current state of the art in the field and current approaches for automating detection and classification.

2. Defining the characteristics of the problem and challenges that face the diagnosis performance.

3. Identifying requirements related to the workflow of an automatic tumour detection system.

4. Identifying the required precision and accuracy levels, which will be proposed for the evaluation framework.

5. Designing and implementing an application in an automatic form, based on detection and classification techniques.

6. Evaluating the application to select the appropriate method for processing.

7. Defining an evaluation framework to achieve the precision level that is required for the tumour detection system.

8. Selecting the evaluated proposed framework that offers high accuracy in the results.

9. Defining the appropriate criteria for evaluating the novel approach.

10. Determining whether the approaches to detect/count/measure match expert assessment.

11. Assessing the novel technique for accuracy levels in counting, size and location by measuring performance using objective parameters.

2.4 Summary

This chapter has outlined the research methodology for this study. The structure of the research is defined in relation to the aims, with a map of the objectives and research questions defined within the workflow shown in Figure 2.1.

CHAPTER 3: LITERATURE REVIEW AND BACKGROUND

3.1 Introduction

This chapter provides background to the work presented in this thesis and the state of the art review of related work based on a thorough literature review. Furthermore, the limitations and major issues of previous works are identified. The survey includes the methodologies, processes and techniques used in segmenting, classifying and detecting pulmonary nodules in the CT-scan image of the lung. In addition, the requirements of automated Computer Aided Detection (CADs) systems are identified. These methodologies concentrate on the analysis and pattern-recognition of the nodule objects in 2D &3D images to give a detailed description for a nodule. These methodologies offer approaches for the evaluation and validation of the techniques and tasks that are adopted in this field such as clustering, classification and segmentation.

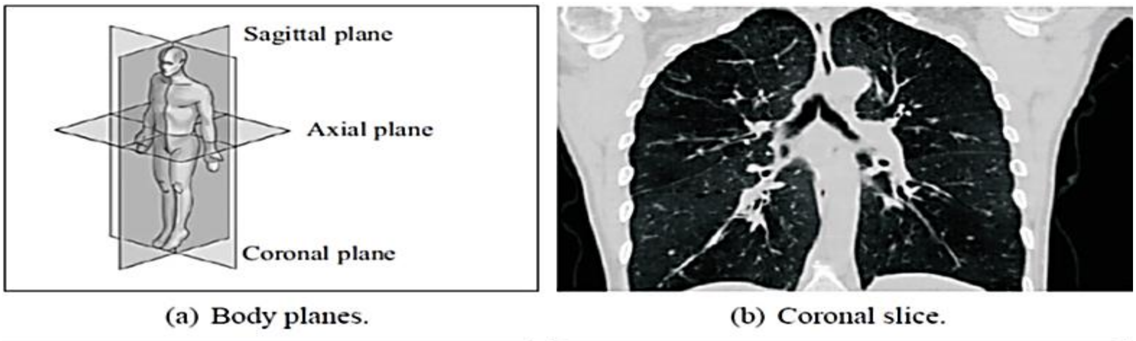
3.2 Introduction to Medical Imaging

A medical image is one of the image applications, which includes development techniques and based in the medical field. Medical imaging displays the internal structure of the human body and the various pathology patterns in an efficient non-aggressive manner to improve the diagnosis. Moreover, these tools are considered harmless techniques, can be used to identify the pathology patterns within the body and support clinicians by providing an internal vision of the organs and vessels, tumours and broken bones (Dzung et al., 1998). In light of the rapid development of medical image technologies, the competition between scientists and researchers has become a productive path to processing and analysing the medical images, producing promising results for serious diseases that cannot be recognised precisely by the radiologists. The CT-scan image is the main tool of different works relevant to the detection of various tumours such as lung, liver and brain tumours.

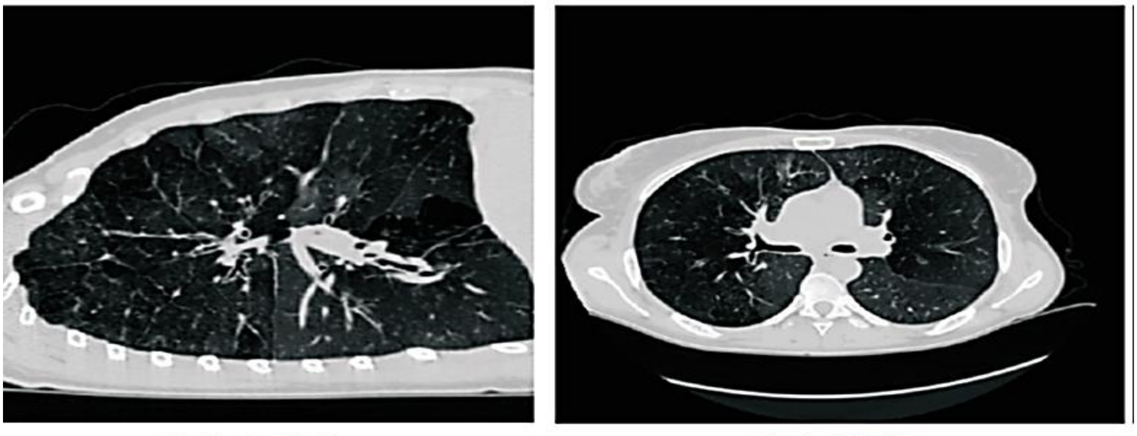
3.3 Lung Anatomy on CT-Scan

The lungs are comprised within the chest, and they are wrapped into two pleural membranes close to the thorax walls. Each lung contains a tree of venous vessels, and where arteries enter the lung, their diameter is about to 30 mm. The vessels' diameter decreases when the veins and arteries become branched. The lungs include compartments called lobes; the left lung consists of two lobes, while the right lung comprises three.

These lobes are divided by fissures, which represent thin sheets in the lung tissue. The CT-lung image is represented in three planes that are sagittal, coronal and axial. Figure 3-1 shows CT-lung slices with different planes, where a coronal slice divides the body to the frontal and posterior, while the sagittal slice cuts the body to the right and left, and the axial slice is in the horizontal plane of the body.



(a) Body planes. (b) Coronal slice.



(c) Sagittal slice. (d) Axial slice.

Figure 3-1: Images of CT Lung Slices: (a) CT Planes Passing Through the Body; (b) Coronal CT Slice of Lung; (c) Sagittal CT Slice of Lung; (d) Axial CT Slice of Lung (Sluimer 2005)

3.4 CT and MRI-Scans of Lung

In general, improving patient care is the main objective of using radiology imaging techniques using different modalities, the most popular ones being the CT Computed Tomography (CT) Image and the Magnetic Resonance Imaging (MRI). The distinguished performance of these modalities in pulmonary imaging is due to the different detection methods used for addressing specific challenges in radiology imaging and the output image characteristics. Since its introduction in the 1970s, the CT provides enormous insight into the structural and anatomical aspects of different lung diseases, as well as ventilation, perfusion and other mechanisms. One of the most encouraging factors of using a CT-scan is its availability with a high spatial resolution. In terms of time this technique can be considered as a useful tool to detect diseases in a very short time (~ 3 sec), and this is a crucial point for the patients since most of them cannot hold their breath for longer than 3 sec during the examination.

Furthermore, the high resolution of the CT-scan images is one of the successes of this technology. Moreover, the accurate detection of the nodules and characterisation of anatomic changes in the chest has been obtained using these tools. When compared to MRI, the CT-scan could be considered more accurate, because the CT density changes when the lung expands to provide more precise measurements of the lung volumes. In contrast, MRI imaging has problems due to its low proton density (Sieren et al. 2010) and (Coxson et al., 2012).

3.5 Pulmonary Nodule Types

The tumours types in the lung are usually classified according to their characteristics appear in the CT image. Pulmonary nodules can be single or multiple and circular structures which exhibit various diameters about 3 to 30 cm and are surrounded by lung airways (Z. Li et al. 2015). Most of the studies combined the features of the CT image to describe the shape, internal structures and size of the nodules. The boundary of nodules could be lobulated, speculated or smooth and usually appear as concave margins, spherical and polygonal shapes. The difficulties in the nodule segmentation are due to a) the nodule being attached to a vasculature or other entities such as the fissures, pleura or abnormalities; b) the nodule being too small; c) the nodule being non-solid or part-solid, in this case, it is too difficult to recognise the boundary, and d) the data has a high level of noise. Therefore, the features that is relative to a nodule appearance represent challenges for segmentation algorithms in the medical image processing. Figure 3-2 shows the nodule types based on the density level of nodules in their classifications. Table 3-1 explains the nodules' characteristics in different aspects such as their shape, density, position, size and their effects on lung details that are displayed as vascular structure and vessels. Figure 3-3 displays the axial slices of CT lung images with nodule types (solid, part-solid, and pure ground glass) (Sluimer 2005).

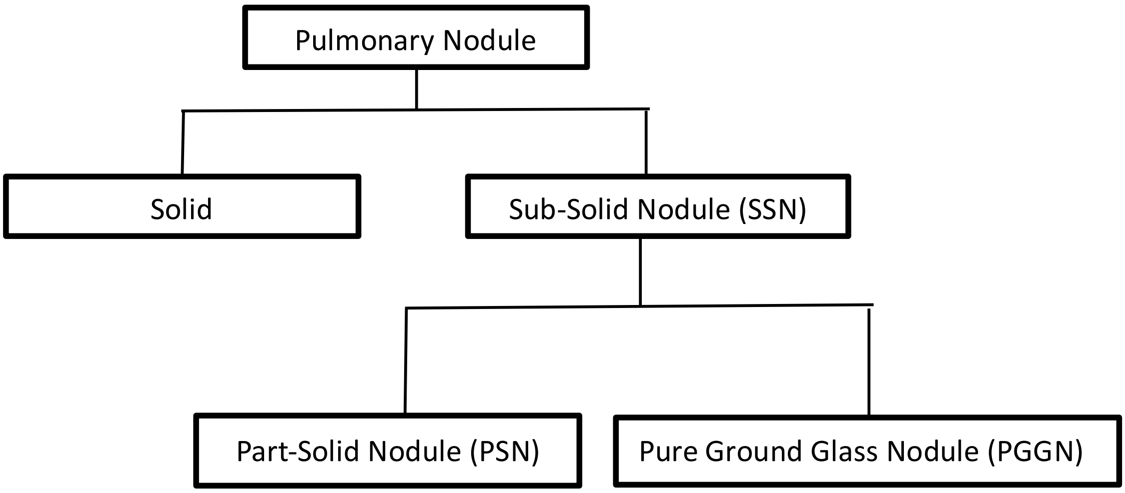
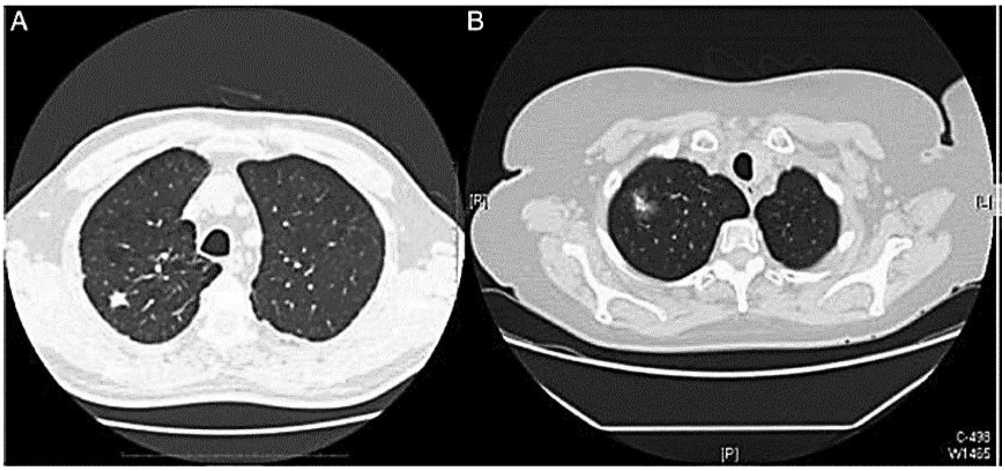


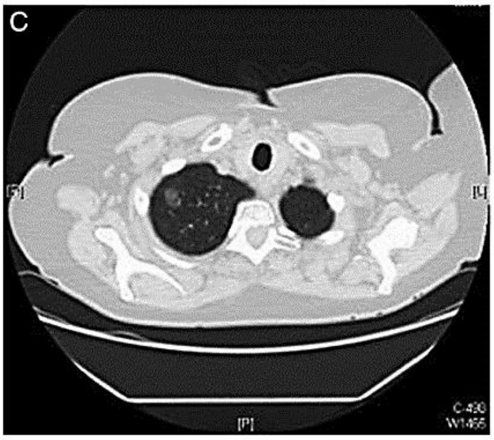
Figure 3-2: Pulmonary Nodules Types' Classification (Sluimer 2005)

Table 3-1: Description and Definition Regarding Pulmonary Nodule Types

Pulmonary Nodule	Focal, rounded opacity <=3 cm diameter, mostly surrounded by aerated lung, including contact with pleura, but without potentially-related abnormalities in the thorax
Sub-Solid Nodule (SSN)	A part-solid or PGGN
Part-Solid Nodule (PSN)	A focal opacity that has both solid and ground-glass component <=3 cm diameter
Pure Ground-Glass Nodule (PGGN) (synonymous with solid component)	A focal ground-glass opacity <=3 cm diameter that does not obscure the underlying broncho-vascular structure
Ground Glass	Opacification that is greater than that of the background, but through which the underlying vascular structure is visible



Solid nodule Part-solid nodule



Pure ground glass nodule

Figure 3-3 The Axial Image Slices of Nodule Types: (a) Solid Nodule; (b) Part-Solid Nodule; (c) Pure Ground Glass Nodule (Sluimer 2005)

3.6 Image Pre-Processing Techniques

Image pre-processing is a significant stage in a medical image processing system. Typically, it is responsible for preparing, smoothing and enhancing the image in order to eliminate the effects of deformation and corruption that may contaminate the vision of the medical image through the acquisition or transmission (Wanget al., 2012) and (Sharma and Anand 2013). This stage includes several steps such as removing noise, image adjustment and the reformation of the image for its conversion to greyscale (intensity values).

3.7 Image Conversion

This process involves saving the CT lung image in a DICOM format that is the standard for storing, printing, handling, and transmitting the medical imaging information. DICOM files can be passed through two entities, which can receive the patient data and images in DICOM format. DICOM conformance enables the equipment to work with efficient electronic health record systems (Mustra et al., 2008) and (Blume and Hemminger 1997). Subsequently, there is a series of procedures that are required during the image processing process such as converting the CT lung scanning image to greyscale and binary values according to the processing requirements. Figure 3-4 shows the axial slice of a CT lung image in the DICOM format with dimensions of 512×512 pixels.

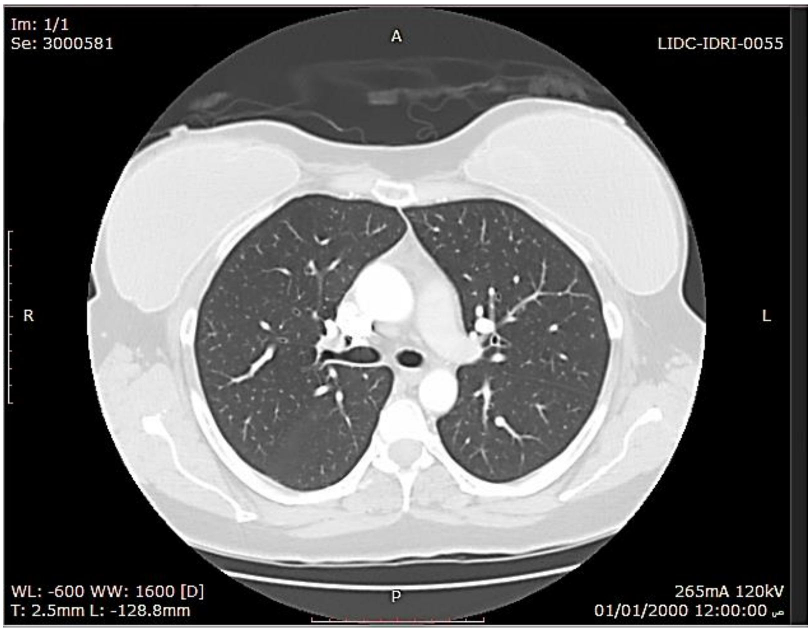


Figure 3-4: CT Lung Slice Saved in DICOM Format with Dimensions of 512×512 Pixels

3.8 Image Enhancement

Image enhancement techniques are vastly employed to refine and improve the visibility of the internal structures and the region of interest in medical images. Image enhancement supports the operators in recognising the details inside medical images, which could be invisible and disorganised after the acquisition of the original image (Pratt 2001). Image enhancement is essentially about improving the perception or interpretability of the information in images for an observer and preparing suitable inputs for automatic image processing systems. The principal target of image enhancement seeks to modify image attributes in order to be appropriate for specific processing or expert evaluation of the image. Also, image enhancement is important for the tumour detection systems to avoid the noise problems that hinder recognising the pathologic patterns. This will present a great amount of subjectivity into the selection of image enhancement schemes that enhances the image without damaging it. There exist two categories of image enhancement methods: spatial domain methods and frequency domain methods. The first group (spatial domain methods) directly deals with the pixel values of images to attain the desired enhancement. In the frequency domain methods, the Fourier Transform is computed for the initial transfer of the image to the frequency domain, where the enhancement operations are achieved, and then the inverse of the Fourier Transform is completed to obtain the resultant image. All the enhancement operations are accomplished in order to increase and modify the image's contrast, brightness or the grey levels' distribution. As a consequence, the intensities of the pixel value of the output image will be modified according to the transformation function applied through the input values (El-Shenawy, 2013; (Burger and Burge 2009)

3.9 Lung Segmentation Techniques

The main goal of the detection systems in the lung segmentation is to reduce the search area of the nodule as a pre-processing phase in the computer-aided detection systems (CADs). However, lungs-segmentation is a challenging step for the detection systems because of the similar structures within lung region such as veins, bronchi and arteries; in addition, the use of different devices and protocols in the medical imaging is equally challenging. Also, the varying contrasts and intensive, the high computational complexity for some methods, the difficulty in keeping the lung information and characteristics that are particularly relevant to diseases are factors that hinder the segmentation results improvement. Consequently, finding accurate, efficient and robust methods in the entire lung segmentation have motivated and encouraged researchers in this area. Therefore, over the last decade, automatic lung segmentation techniques have been developed in the medical imaging field to improve performance (Fu and Mui 1981). In this area, the relevant methods in the lung segmentation can be classified in four groups (Maintz 2005): techniques based on a threshold(Van Rikxoort and Van Ginneken 2013), deformable shape and models, edge-based models with some techniques, and clustering methods (Pal and Pal 1993). In this section, some of the related works of this study are briefly presented.

3.9.1 Threshold-Based Lung Segmentation

In these methods, the segmentation is based on the threshold of the histogram properties, which could be the simplest technique and suitable for brightness uniform regions or objects within images. This technique is performed to segment the foreground and background. However, this method does not work with multiple grey levels of image regions. To process this limitation, the multi-threshold technique is applied to achieve good segmentation results. Figure 3-5 shows the response of the image histogram to two thresholds dividing the image into three regions within greyscale values.

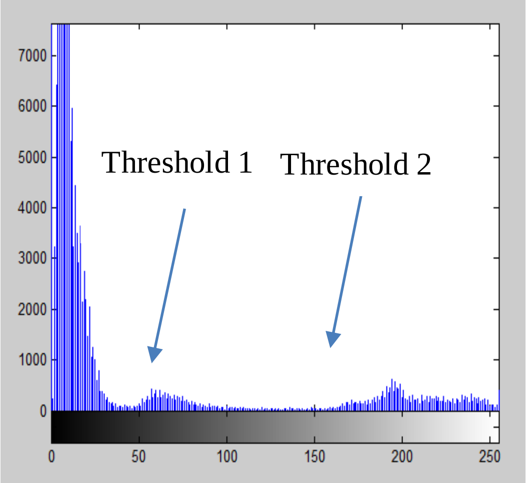


Figure 3-5: Image Histogram of Two Thresholds, Region 1 belongs (0–50), Region 2 belongs (50–150), and Region 3 belongs (150–255) of Grey Scale Values

These techniques have been applied in the lung segmentation of the background. Previous works of a group of researchers have focused on the threshold techniques to segment the lungs in CT images. The technique of Van Rikxoort et al.(2009) depends on the image pixels intensity that is evaluated by a quantitative criterion of radiodensity which is called Hounsfield unit (HU) and was developed by Newbold in 1972. The grey scale range represents intensity values of pixels. The range from 1000-400 HU has been applied in many types of research to segment the lung regions. However, this fixed range is not efficient for the separation of the lungs object from all CT scan images because of the image contrast variety. Therefore, Tseng and Huang (2009) adopted an updated threshold with specific value iteratively increasing for different images until reaching the satisfactory criteria in the lungs segmentation works. To improve image contrast, Zhou et al.(2016) applied an enhanced threshold using stretch techniques to the image histogram for intensities distribution unified between areas with high and low Hounsfield values. In addition, Wang et al.(2009) employed experimental threshold value selected from -400HU to -200HU range to display an initial prediction of normal and abnormal lung regions. However, threshold techniques have some disadvantages in lung segmentation. The density of CT lung is affected by factors such as air volume, organ tissue volume, the degree of inspiration, imaging devices protocol and physical material characteristics of the lung parenchyma.

3.9.2 Deformable Models -Based Lung Segmentation

The essential approach in this segmentation is that the organs' structure has a geometric form to model the different shapes of organs. The probabilistic models of image deform according to parameters represent texture features for an estimated model that is used for segmentation. Model-based techniques of segmentation include the active appearance model and shape, level-set based and deformable models. However, the disadvantages of these methods are that they require manual interaction to locate an initial model and select appropriate parameters and have poor detection of concave boundaries using standard deformable models (Pham et al., 2000). The deformable models are other techniques that proved effective in the medical images segmentation by discovering the salient edges of images precisely. Recently, (Rebouças Filho et al. 2013) and (Rebouc et al. 2015) used a crisp active contour to segment a whole lung from CT lung in two separate works. The first research reduces the curve energies, drawn inside the region of interest, to be pushed towards the object edges by successive iterations. The second research is based on the topology curve information for matching between the curve points and the contour expansion towards the lung boundary. To guide the active contour,(Annangi et al. 2010) employed an active contour depending on a suitable contrast of lung edge and extracted some features using canny edge applied upon the image histogram. In order to handle the local minima problems attached to active contour models in the segmentation, the region and shape terms are employed with an active contour in this method. However, the variance contrast and local minima are still a challenge for active contour models.

3.9.3 Edge Detection -Based Lung Segmentation

An edge-based segmentation is a common method to detect the edges and boundaries, which are split into distinct regions. The edge detection technique is able to detect and recognise the differences and discontinuities in grey level, colour, etc., which often represent boundaries among objects. The gradient (derivative) function is based on the edge detecting operators and is available in, for example, the Prewitt, Sobel, Roberts (1st derivative type), Laplacian (2nd derivative type), Canny, Marr-Hilclrath edge detector. However, these methods are affected by the weak edges and fake edges or the presence of noise that have an unsatisfactory impact on the segmentation results. For lung segmentation, edge detection algorithms are one of the techniques that focus on lungs boundaries information for accurate segmentation. These algorithms depend on some operators for the lung separation. In the work of (Saad and Hamid 2014), some filters such as Sobel, Prewitt, Laplacian and Canny edge filter with morphological techniques to segment lung regions in CT images were used. The results were unsatisfactory because of the effect of image noise on high pass filter performance. To accurately segment lung borders, Talakoub et al. (2007) proposed an algorithm which is based on edge detection; it uses wavelet analysis that stands a noise effect and discriminates the intensity of various types of edges. However, these algorithms are more interested in lung border detection than lung tissue and usually fail to be implemented for images with weak segment boundaries.

3.9.4 Clustering -Based Lung Segmentation

Similarly, the methods that are based on clustering included many helpful features for segmenting. The principle of homogeneity is based on the region properties where pixels that have similar features are clustered as a homogenous region. The measure for absolute homogeneity is the grey level of points (Sonka et al., 1999), and this measure can be demonstrated by the following conditions:

$$R_1 \cup R_2 \cup R_3 \ldots . \cup R_i = I$$

Where $R_1, R_2, R_3, \ldots R_i$ are the regions within the image I, and further,

$$R_1 \cap R_2 \cap R_3 \ldots . \cap R_i = 0$$

Further classification of region-based segmentation is divided into three categories that are derived from the region growing principle and these are, summarised as follows:

a. Region Merging

The process determines the initial seed points; the segmentation results' success is due to the selection of seed points. By the criterion of merging, the iterative merging of the neighbouring pixels is performed for region growth. This process is stopped when each pixel is assigned to its particular region according to the merging criterion.

b. Region Separating

The principle of this process is the inverse of the region merging and iteratively splits the whole image until no further separation of a region is possible.

c. Separating and Merging

This is the integration of two processes (separation and merging) using their merits. In this method, the representation of the data is as quad-quadrant tree splits image segments into four quadrants as long as the original segment reflects non-uniform features. The uniformity of the segment (regions) ensures the merging of the four neighbouring squares. This splitting and merging process is continued until there are no new results for both the split and merge.

As part of the mentioned techniques above, the K-means clustering is the one that is based on a region-based segmentation and the clusters' pixels having similar grey levels. It can separate the object with most of the characteristics of the original image and has efficient clustering because it is able to shift the cluster centre of data to be in the right position among two or more clusters.

The clustering conditions are adopted by various approaches such as the fuzzy clustering that applied to the lung segmentation broadly (Wang et al., 1996),(Gevers and Smeulders 1997) and (Ney 1992). The methods of fuzzy clustering are considered a soft segmentation which is an ideal option to allow the region or clusters to be overlapped, where the pixels can possess multiple memberships with different degrees of membership parameter in the various regions. In hard segmentation, the overlapping is not allowed for the separated region, and the pixels are assessed in the region according to their maximum membership value. Thus, soft segmentation holds more information of the original image through allowing the pixels to retain membership in several regions.

For hard segmentation, the pixel has a binary membership defined as in Eq. 3-1:

$$m_{k.j} = \begin{cases} 1 & if\ j\ \in R_k \\ - & - \\ 0 & otherwise \end{cases} \qquad 3\text{-}1$$

Where j is a j^{th} pixel belonging to the image (I), and $m_{k,j}$ is the membership function of j^{th} pixel in region R_k.

On the other hand, in the soft segmentation, the pixel has several memberships in various regions, and the membership function must implement the following constraints:

$0 \leq m_{k.j}$ 1 for all k, j and

$$\sum_{k=1}^{N} m_{k.j} = 1\ \forall j \qquad 3\text{-}2$$

Here, N = the total number of separation regions in the image (I). By the value of membership of pixel j in R_k (k^{th} region), one may measure how robustly the pixel belongs to the region; the more robust membership value is the member of region R_k. At the edge region, the pixel may have a diverse membership in several regions.

Fuzzy clustering is an efficient method regarding soft segmentation and is widely based on an unsupervised algorithm in the segmentation of CT images. One of these methods is the fuzzy c-means.

- **Fuzzy C-Means Segmentation**

The partition of imaging data into various cluster regions within an image space is based on the similar intensity of the image values. Medical images usually introduce intensity overlapping of grey-scale for different tissues, and this method is found to be more suitable for medical image processing. Fuzzy c-means can be defined as follows:

Let $X = \{x_1 . \ldots . x_c\}$ represent a data set. Assume c represents a positive integer that should be greater than one. A segmentation of X *into* c clusters can be defined by mutually disassembled sets $X_1 . \ldots . X_c$ such that $X_1 \cup \ldots \cup X_c = X$ or are equivalent to the indicator functions $\mu_1 . \ldots . \mu_c$ *such as* $\mu_i(x) = 1$ if x is in X_i and $\mu_i(x) = 0$ if x is not in X_i for all $i = 1. \ldots . c$. All these called cluster in X into c clusters $X_1 . \ldots . X_c$, which is known as a hard c-partition$\{\mu_1 . \ldots . \mu_c\}$. The common method is k-mean (or called hard c-mean) and is an iteration method to reduce the objective function J_{HCM} and is defined as:

$$J_{HCM}(\mu . a) = \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_i(\chi_j) \left\| \chi_j - a \right\|^2 \qquad 3\text{-}3$$

Here, $a1, \ldots, ac$ are usually the c cluster centres. The fuzzy expansion allows $\mu_i(x)$ to become a membership functions in fuzzy sets μ_i on X having values in the interval [0, 1] such that $\sum_{i-1}^{c} \mu_i(x) = 1$ for all x in X. In this state, $\{\mu_1 . \ldots . \mu_c\}$ is a so-called fuzzy c-partition of X. Consequently, the objective function $JFCM$ of the fuzzy c-mean (FCM) will become:

$$J_{FCM}(\mu . a) = \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_i^m(\chi_j) \qquad 3\text{-}4$$

Here, $\{\mu_1 . \ldots . \mu_c\}$ is a fuzzy c-partition and m is a constant number greater than one to display the degree of fuzziness. The FCM clustering method is a repetition by the necessary conditions to minimise J_{FCM} .

The separated segments of this image have been obtained using the fuzzy-c-means algorithm and are illustrated in Figure 3-6, which shows a segmented image by fast fuzzy c-means and the pseudo colour covers the individual segments representing each one the belonging degrees of region's pixels.

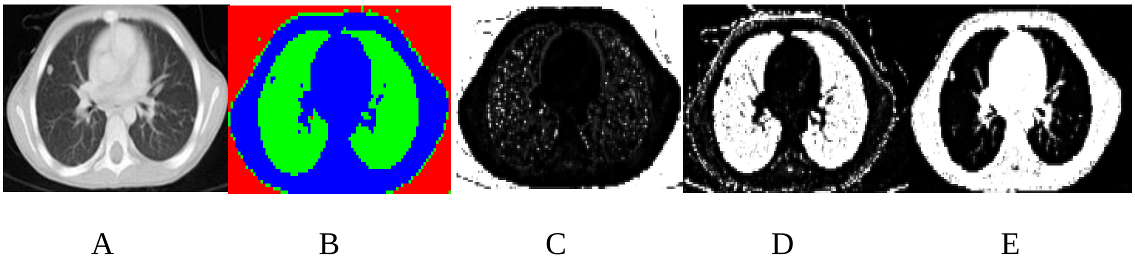


Figure 3-6 Individual Segments of Lung CT Image: (A) Original, (B) Segmented Image in Pseudocolour, (C–E) Individual Segments

However, the original fuzzy c-means is modified and enhanced to fast fuzzy c-means in later research to show more robustness against the outlier, and reduce its sensitivity to noise and the segmentation process accelerating in the fuzzy c-means algorithm. The grey information and local spatial play a vital role in the fuzzy c-means clustering acceleration which is based on the number of grey-levels instead of an image size that needs longer segmentation time; therefore, the image segmentation time is extensively consumed according to the image size (Biniaz et al., 2012).

Some of the researchers, (Sivakumar and Chandrasekar 2012) and (Javed et al. 2013), have considered the weight of fuzzy and some features as a standard of lungs image segmentation. (Sivakumar and Chandrasekar 2012) proposed a segmentation approach for CT lung images using Fuzzy C-Means (FCM) and Weighted Fuzzy C-Means (WFCM) algorithms to validate and evaluate the segmentation performed by the two approaches. Meanwhile, the lungs CT scan images are enhanced by a Median filter, which is used to remove the noise. Haralick features are then calculated from the co-occurrence matrices as input data to the FCM and WFCM algorithms. (Javed et al. 2013)suggested a method of segmentation of the CT lung image using fuzzy logic where the weights to pixels are assigned by fuzzy logic. The pixels have higher weights if its entropy and local variance are less, while lower weights are assigned to pixels having a high entropy and local variance.

3.10 Nodules Segmentation Techniques

Pulmonary nodules detection amongst lung pathologies is considered as one of the challenges that face researchers in this field. The challenge is not only due to the nodule patterns variation in the CT lung image but also a result of the difference in the intensities, contrast, lung tissue and the recognition difficulty of the connected vessels and nodules attached to the lung wall. A huge number of studies adopted diverse techniques and involved various solutions for these problems. They will be explained in detail below.

3.10.1 Nodule Segmentation Based on Clustering

The clustering techniques are broadly applied in the nodules segmentation. (Nie et al. 2012) proposed a weighted kernel of fuzzy c-means that employed the information of vessels and class as weights for semi-solid nodules segmentation in CT lung images.

Manual detection of the nodule is defined in the centre slice where the possible nodule pixels are clustered by improved weighed kernel fuzzy c-means (IWKFCM) from the centre slice and its attached slices. The method validated 36 nodules and achieved accurate results. In their investigation, (Jinke Wang and Cheng 2015) have employed an adaptive fuzzy C-Means (AFCM) technique to improve the training phase of nodules. In this technique, Mahalanobis distance is implemented for the classification of the candidate test nodules that are identified by a proper threshold. This method was validated through 35 thoracic CT images, and a 2.8 false-positive rate was obtained from the experiment per scan. Furthermore, (K. Chen et al. 2013) proposed a curvature coefficient based on the fuzzy clustering algorithm for identifying nodule surrounding tissues. The fuzzy clustering obtains edge information as an initial segmentation and then reduces the false-positive object by filtering. This algorithm demonstrated its results with juxta-vascular and GGO (ground glass opacity) nodules.

3.10.2 Nodule Segmentation Based on Active Models

The most efficient methods used in the nodule segmentation are active models. In their work of (Keshani et al. 2010) used the active contour in the lung segmentation to transfer non-isolated nodule linked to chest wall apart. Then, the active contour extracts the contours of the region of interest (ROI) by applying 2D stochastic features. In another work, (Keshani et al. 2013), these features with 3D anatomical features are used by a support vector machine to detect the nodule. The active contour with mask techniques is adopted for the nodule contour extraction that transfers non-isolated nodules into isolated ones. The performance of the first method was 3 false positive (FP) per scan and 89% detector rate for the second method. (K. Zhou and Weng 2015) suggested a snake model to segment suspicious nodule edges in the lung tissue using pathological features extracted from nodules to decrease interference and indicate suspect nodules in CT scans. Furthermore, (Farag et al. 2012) based on template matching formulation of the active appearance (AAM), measured a similarity degree between the input image and an AAM template. This template deals well with various scans and the location of nodules because of the template possession of the sufficient rotation parameters have been applied for a different radiology imaging.

3.10.3 Nodules Segmentation Based on Feature Extraction and Classification

Features extraction represents a crucial stage in multimedia processing. To extract optimal features that present the fundamental content of the images as much as possible, this task is still a challenging problem for researchers in the computer vision field. The attention focused on a feature extraction provides a comprehensive study for images feature representation methods. The extracted features are various and describe different aspects of the image such as shape features as given in figure 3-7, texture, statistical, structural and spectral features.

i. Shape Features

It is acknowledged that shape description is an important scheme for human beings in order to recognise and identify the real-world things and objects by some geometrical forms such as straight lines in various directions. The classification of shape-featured extraction methods can often be divided into two classes (D. Zhang and Lu 2004); region and contour-based methods. The first measures the shape features for an entire region of the shape, while the second method calculates features specifically from the boundary of the shape. Both techniques are performed on an image concerning the spatial relationship presence, which can report the object location and relationships among objects in an image(Yang et al., 2005) and (Ping et al., 2013).

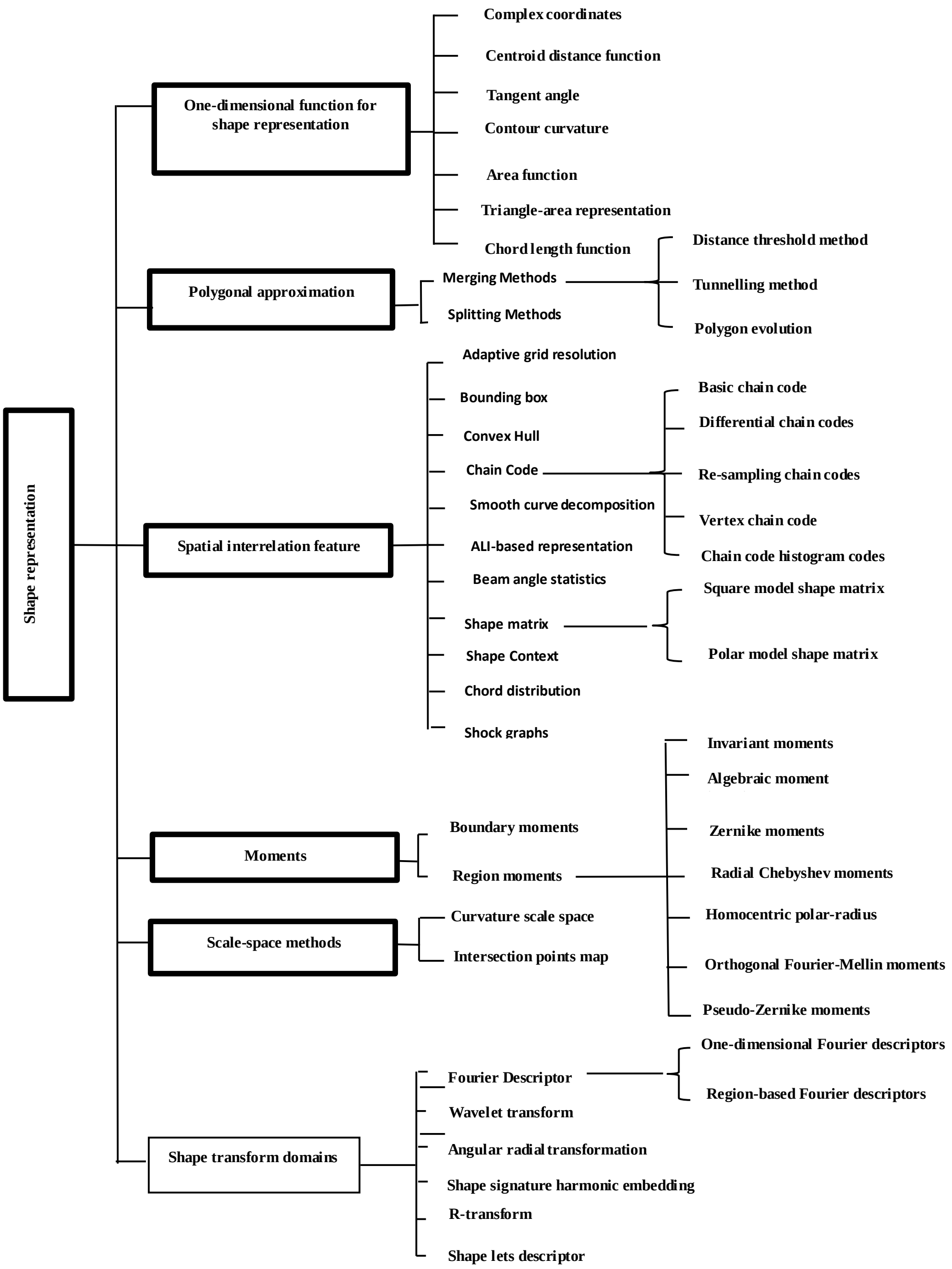


Figure 3-7: An Overview of Shape Description Techniques (Ping et al., 2013).

ii. Textural Features

The viewpoint of image segmentation and classification is that the textural features are important tools that perform image segmentation and classification or both. The texture can be described as something containing mutually attached elements reflecting smoothness or coarseness. This will be according to the spatial relationship of the pixels in the texture structure or the pixel intensity properties of the texture tone. The goal of the texture-based segmentation scheme is subdividing the image into segments having various textural features while the classification aims to sort the prior segmented regions by other segmentation methods. The texture features-based approaches are divided into the structural, the statistical approach and the spectral approach as follows:

iii. Statistical Methods

The spatial distribution of grey-level pixel values establishes statistical methods for extracting texture features that apply to different tasks in the computer vision. The texture features of an image are calculated statistically depending on the intensity of the positions’ distribution. These methods produce a wide range of textural features that include first order, second order and then higher order statistics. The techniques that are related to the statistical features’ orders are a histogram of image intensity and autocorrelation features representing the second order, while the higher features’ order is correlations as superior statistics than the second order.

iv. Structural Methods

In the case of the structural approach, the texture is described as producing a complete pattern based on placement rules that are acquired by modelling the geometric relation between the textures’ primitives or studying their statistical features. Individual pixels, lines segments or an object with similar grey-levels represent the texture primitives to organise the texture spatially. The local binary pattern is a common technique to extract structural features.

v. Spectral Methods

The textures, in this case, are defined through the spatial domain based on the filtering technique in deriving the textures’ features. In the frequency domain, gradient filters with adjusted frequencies perform the filtering for extracting lines, edges, and isolated dots, etc. Following this notion, the filter banks are applied as mathematical models to convolve the input image in order to extract the spectral features.

There are methods found for extracting and classifying the previous approaches such as textural features derived from the co-occurrence matrix of image values and defined by the statistical description of the image greyscale (Argenti et al., 1990) and (Haralick, Shanmugam, and others 1973). The fractal texture description method, the run of the length of the grey level method (Chaudhuri and Sarkar 1995), the syntactic method and the Fourier filter technique are common to generate robust descriptors. Comparing the above-mentioned three approaches, the spectral frequency-based methods are less efficient. While the statistical methods are particularly useful for random patterns/textures and for complex patterns, the syntactic or structural methods still give better results. Texture-based methods are the best for the segmentation of medical images when compared to the segmentation of medical images using simple grey level-based methods (Koss et al. 1999).

3.10.4 Features Selection and Dimension Reduction Techniques

In view of other researches, the features reduction is an essential step has been applied before the classification stage in the machine learning field(Saeys, Inza, and Larrañaga 2007). The reduction of features dimensionality may facilitate a complicated understanding of questions about the research area and interest and reduce the cost of computations through the training and testing the dataset. The features reduction can be divided into supervised and unsupervised techniques. Both techniques of reducing the features have been divided into three categories are wrapper, filter and embedded methods. First, wrapper techniques based on objective functions belong to the regression machine learning model(Guyon and Elisseeff 2003). Second, filter techniques involve t-test, Pearson correlation coefficient and ANOVA (Ying Wang et al. 2010) and (C. Chen 2015). Final, the embedded methods are applied to choose relevant features through forcing penalties onto a machine learning model, thus submitting a subset of relevant features(Tibshirani 2011).

3.10.5 Pattern Recognition Techniques

The classification presents a process to isolate the objects in images into sorts to support and automate the diagnosis in medical imaging and several other applications, such as speech recognition and robotics(G. Dougherty 2009). Also, it assigns image classes based on features' space partitioning which are extracted from an image with known labels for each class. In addition, it contains two phases of the process: the first phase is the training phase or learning phase implemented on a known number of data classes that are usually derived from the application problem specification. In the second phase, the classifier is applied for classifying the test data determined in the dataset, which is called the "test phase" (Larose 2005) and (Han et al., 2011).

The classification techniques are defined in two groups: the structural approach and statistical approach. The statistical approach consists of probability computations of the class and measurement of the standard deviation and mean to reflect the ideal class representation in data such as support vector machine (SVM), linear discriminant analysis (LDA), and K-nearest neighbour (KNN) (Dougherty, 2009; Han et al, 2011.; Li et al., 2006). The intelligent approach exhibits various learning capabilities including the use of AI techniques in the classification process such as ANN (Hagan et al. 1996). This approach is predominantly based on some features representing the classes and provides prior knowledge to support the classification process. Furthermore, features extraction is a complicated operation that requires pre-processing such as relevance analysis, which removes irrelevant features or redundant ones that do not contribute to the classification process and thus reduces the cost of the computations. Moreover, a transformation of extraction features through normalisation to a new form suitable for the classification avoids the features with large ranges in order to outperform the features with small ranges. The models that used for classification are explained their basic work principles.

1- Bayesian Classifier

The Bayesian classifier representation is quite intuitive and easy to understand the issue that is often a significant interest in machine learning. It is considered a probabilistic statistical form that performs a probabilistic relationship between a set of variables with their wide conditional dependences. Bayes' theory relies on the use of advance information about unknown landmarks(Ghosh et al., 2007). Therefore, it solves several problems and applications in particular in artificial intelligence. The formula can be defined as:

$$P(A\cap B)= P(A/B)P(B) \tag{3-5}$$

$$=P(B/A)P(A) \tag{3-6}$$

$$[=P(A)P(B)] \tag{3-7}$$

$$P(B/A)= \frac{P(A/B)P(B)}{P(A)}=\frac{P(A/B)P(B)}{P(A/B)P(B)+P(A/B)P(B)} \tag{3-8}$$

$$P(B/A)\propto P(A/B)P(B) \tag{3-9}$$

$$P(B/A)=P(A/B)*P(B)/P(A) \tag{3-10}$$

Where:

1. P(A) is the marginal or prior probability of A (since it is the probability of a prior to having any information about B).
2. P(B) is the marginal or prior probability of B.
3. P (A|B) is **the likelihood** function for A given B.
4. P(B|A) is **the posterior probability** of B given A (since it depends on having information about A).

In view of Bayesian about subjective probability emphasizes that completely unknown parameters are handle as uncertain and therefore they need to be described through a probability distribution. Consequently, there are three ingredients attached to Bayesian statistics (Domingos and Pazzani 1997). The first ingredient represents the background knowledge about the parameters belonging the tested model. It indicates to all knowledge available is held in the so-called prior probability and before understanding the data. The second ingredient represents the data information themselves. It is the observed proof extracted in terms of the likelihood function of the data given the parameters. The third ingredient is established by merging the first two ingredients, which is declared posterior probability. Both (1) and (2) are merged via Bayes' theorem and are summarized through the supposed posterior probability, which is an adjustment of the observed evidence and the prior probability. The posterior probability considers one's updated knowledge; stability observed data with prior probability (Van de Schoot et al. 2014).

Advantages

- The Bayesian framework reduces many of the contradictions connected to traditional hypothesis testing.
- The Bayesian framework presents a more direct expression of suspicion, including full ignorance. Bayesian methods are different from other frameworks as they are the only that can combine background knowledge within the analyses through means of the prior distribution.
- It endorses updating knowledge to avoid a null hypothesis test over again.
- The Bayesian paradigm often leads to reproducing others' conclusions or sometimes strengthening them.

2- Naive Bayes Classifier

This is a classifier based on Bayes' Theorem with a supposition of independence between predictors. A Naive Bayes technique assumes that the existence of a particular property in a category is unrelated to the existence of any other property (Ozekes and Osman 2010). In other words, this classifier can produce its decision even if the features are based on each other or depend on the presence of other features. To summarise the Naive Bayes model characteristics, they can be defined as follows (McCallum et al., 1998).

- The Naive model is easy in construction and performs well in multiclass prediction.
- It is useful for large data groups.
- Naive Bayes is known to outperform the highly complicated classification methods.
- It quickly predicts the class of test data set and needs little training data (Lewis 1998).

In this study, the Naive Bayes could be suitable classifier and efficient to classify the feature vectors for nodule and vessels and issues a decision that detects a potential nodule from the vessel in a CT image. Mathematically, the Naive Bayes classifier can be defined as:

$$P(y|x) = P(x|y)P(y)/P(x) \qquad 3\text{-}11$$

Where:

P(y/x) is the posterior probability of class (y, target) given predictor (x, attributes)

P(y) is the prior probability of class

P(x/y) is the likelihood, which is the probability of predictor given class

P(x) is the prior probability of predictor

Naïve Bayes Weakness

In a close look at Naïve Bayes, it has some drawbacks that can be summarised as: (i) if the attributes are not independent, it decreases the Naïve Bayes classification accuracy and (ii) it cannot deal with nonparametric continuous attributes. Therefore, the dependence presence among image attributes values leads to the unsuccessful classification of Naïve Bayes. To process these problems, there are many proposed classifiers as alternative methods to improve prediction accuracy and deal with dependent and irrelevant attributes such as linear regression that suggests different models be tested in several domains with better results (Ng and Jordan 2002).

3. Linear Regression

The linear regression is a statistical process that describes the relationship between one or more independent variables (input) and one dependent variable (output). The main task of regression analysis fits a single line at the centre of scattered spots. Figure 3-8 shows a regression line with dataset points (Draper and Smith 2014). The simplest form with one dependent and one independent variable is defined by the formula:

$$y = c + b * x \qquad 3\text{-}12$$

Where y = estimated dependent, c = constant, b = regression coefficients, and x = independent variable.

The linear regression process can be divided in two processes depending on the number of independent variables (x); if the x is one variable with one or more independent variables, the process called simple linear regression. However, multivariate linear regression process predicts multiple correlated dependent variables are for more than a single independent variable.

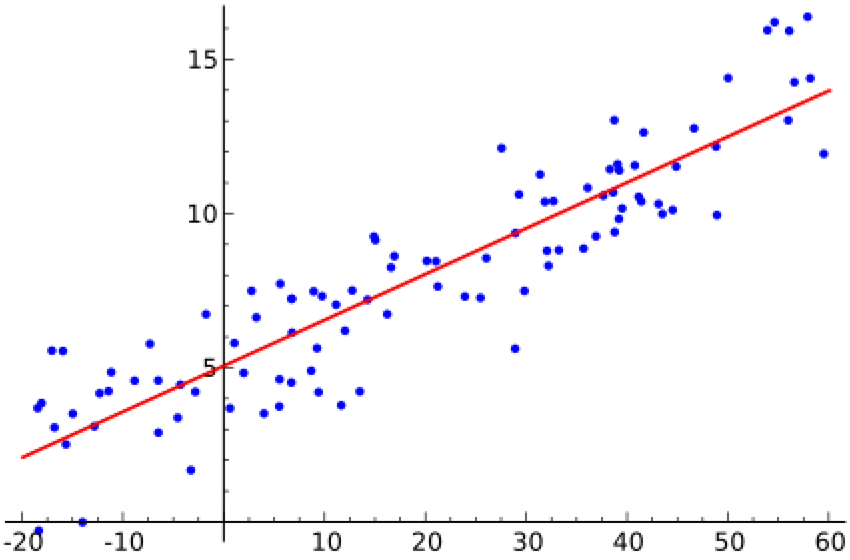


Figure 3-8: Illustration of Linear Regression on a Data Set (Draper and Smith 2014)

- Logistic Regression

Logistic Regression predictor is most common model applied for regression analysis. The model is developed to predict binary outcomes for classification, by David Cox in 1958. This model is designed to estimate binary dependent variables for one or more independent variables. The logistic regression based on characteristics of other memory-based methods in classification. Thus, it works with two analysis algorithms: global logistic regression and locally weighted version (Hsieh et al., 1998). A global logistic regression statistical algorithm, its output represents Boolean, resists the noise because the curve of the logistic function can go in midway through the data points. A logistic function can be similarly employed here as a sigmoid function, which represents a continuous function between 0 and 1, to fit the data points in memory as shown in Figure 3-9 . The global logistics fails when more than two segments of data in the memory as shown in Figure 3-10 , therefore, a locally weighted version, which able to deal with several segments of data and based assigning the weights of data in order to avoid noise, is implemented. The independent variables (input variables) of logistic regression in this study are some features extracted from the region of interested (ROIs) while logistic regression outputs are two classes (binary classification) and represent nodule and vessels. The logistic regression is defined in the Eq.3-13.

$$P(x) = \frac{1}{1+e^{-xb}} \tag{3-13}$$

$x\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$. Here β are the regression coefficients of input variables.

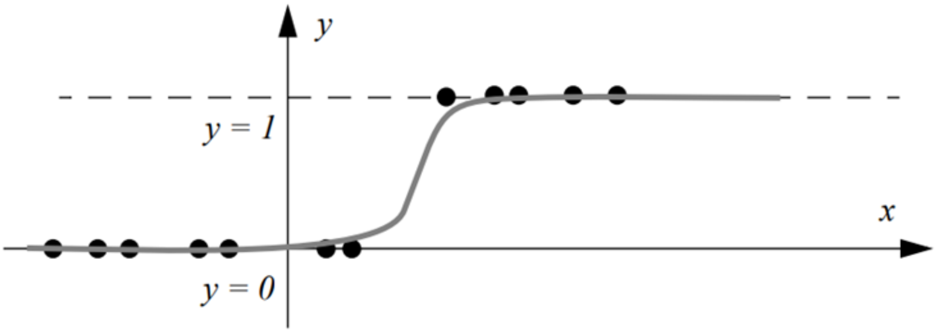
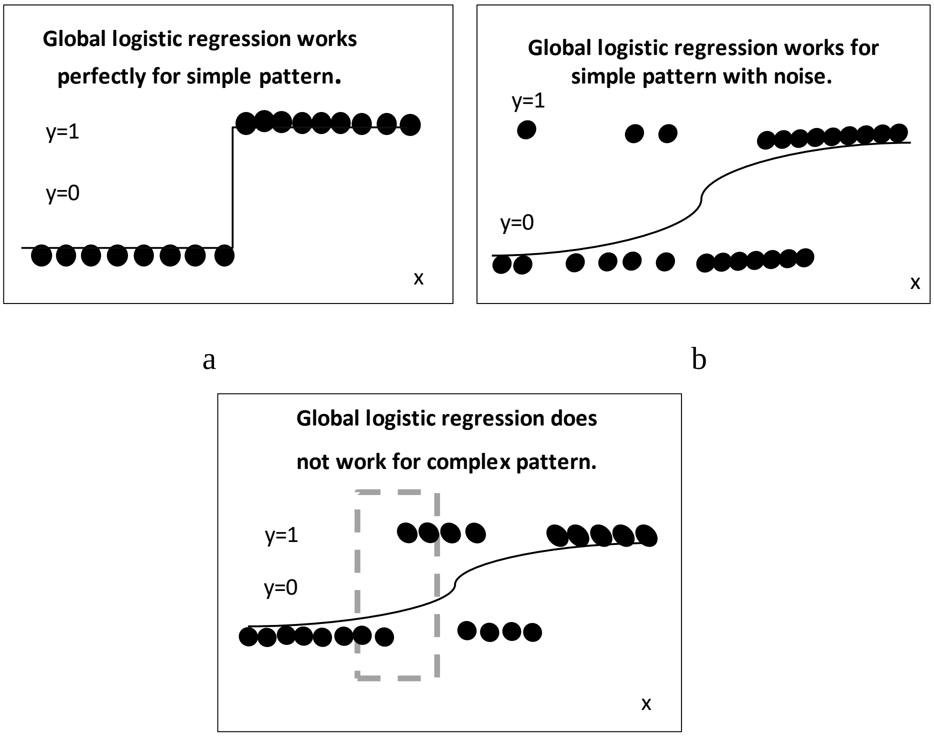


Figure 3-9 Two Line Segments to Fit the Dots in Global Logistic Regression(Hsieh et al., 1998)


Global logistic regression works perfectly for simple pattern.
y=1
y=0
x
Global logistic regression works for simple pattern with noise.
y=1
y=0
x
Global logistic regression does not work for complex pattern.
y=1
y=0
x

a b

c

Figure 3-10 (Global) Logistic Regression for Classification :(a) a Curve of Global Logistic through Data (b) a Curve of Global Logistic through Noise (c) a Curve of Global Logistic does not Work Because of more Than Two Segments (Hsieh et al., 1998)

4- Support Vector Machine

The support vector machine (SVM) is one of the supervised learning models and developed in 1992 by Vapnik, Boser and Guyon. It is used in different applications known such as object and handwriting recognition, identifying the speaker and medical diagnosis (Han et al., 2011). It is considered a statistical approach performs both nonlinear and linear classification using nonlinear mapping to transform the features into a higher dimension. To optimal separation, it searches the linear or nonlinear hyperplane for the features separation into classes. SVM presents accurate classification even though in cases the features number is large while the sample is small (Dubitzky et al., 2007). The optimal separation of the hyperplane is established when the two closest data points of two classes have the maximum distance between them shown in Figure 3-11 (Han et al., 2011; Dubitzky et al., 2007).

The decision function is represented by Eq. 3-14.

$$d(x, \omega, b) = \omega . x + b = \sum_{i=0}^{n} \omega_i x_i + b \qquad 3\text{-}14$$

Where x is the vector of attributes, $\boldsymbol{\omega}$ is the vector of weight, b is a scalar and known as a bias or as an additional weight ω_0 too, and n represents the attributes number. In case, two attributes are found (A1, A2), then X= (x1, x2), where values' vectors are x1 and x2 of these two attributes. When the sets=0, the decision function can estimate the separating hyperplane function as given in Eq. 3-15.Figure 3-11 shows how hyperplane function to separate two classes.

$$\omega . x + \omega_0 = \sum_{i=1}^{n} \omega_i x_i + \omega_0 = 0 \qquad 3\text{-}15$$

Therefore, for each point that lies over the separating hyperplane will offset Eq. 3-16:

$$\omega_0 + \omega_1 x_1 + \cdots + \omega_n x_n > 0 \qquad 3\text{-}16$$

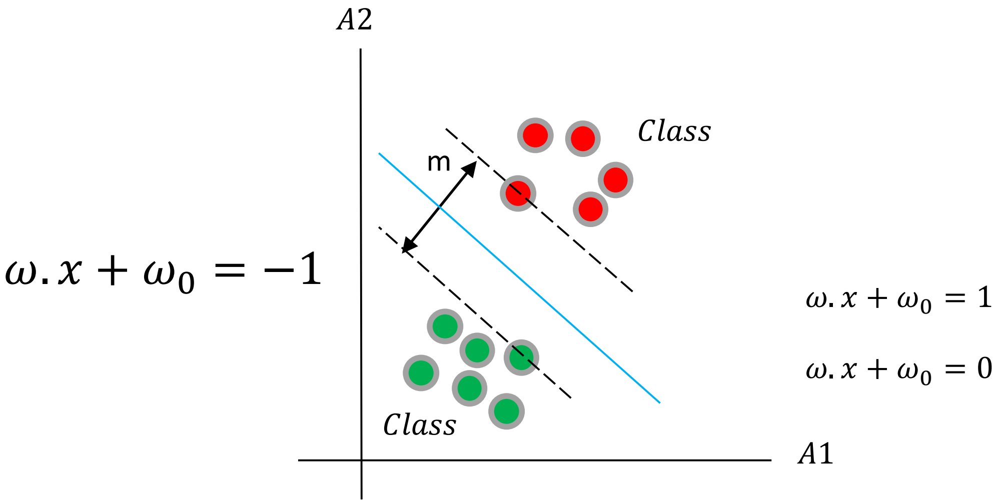

$A2$
$Class$
m
$\omega . x + \omega_0 = -1$
$\omega . x + \omega_0 = 1$
$\omega . x + \omega_0 = 0$
$Class$
$A1$

Figure 3-11: Hyperplane Separates Two Classes (Al-Waeli 2017)

Correspondingly, for per point, which lies below the separating hyperplane, will offset

$$\omega_0 + \omega_1 x_1 + \cdots + \omega_n x_n < 0 \qquad 3\text{-}17$$

Whilst the expression of the margins “sides” can be written as Eq. 3-18 and Eq. 3-19

$$H1: \omega_0 + \omega_1 x_1 + \cdots + \omega_n x_n \geq 1 \qquad 3\text{-}18$$

$$H2: \omega_0 + \omega_1 x_1 + \cdots + \omega_n x_n \leq 1 \qquad 3\text{-}19$$

Where H1 and H2 are considered the hyperplanes and also defined support vectors.

By merging Eq. 3-18 and Eq. 3.19, Eq. 3-20 is acquired.

$$\omega_0 + \omega_1 x_1 + \cdots + \omega_n x_n \geq 1, \forall_{\mathrm{i}} \qquad 3\text{-}20$$

Each attribute sample that locates on the support vectors H1 or H2 will offset Eq. 3-20 and it gets as Eq. 3-21:

$$\omega_0 + \omega_1 x_1 + \cdots + \omega_n x_n = 1 \qquad 3\text{-}21$$

Thus, the size of the maximal margin (m) is obtained easily and which considers the distance for each point placed on the support vector H1 to the hyperplane is$\frac{1}{\|\omega\|}$, where the $\|\boldsymbol{\omega}\|$ represents the Euclidean norm of $\boldsymbol{\omega}$, which it is $\sqrt{\omega_1^2+\omega_2^2+..+\omega_n^2}$. And, this equals to the distance from any point that is located on the support vector H2 to the hyperplane. Thus, the greater distance among the support vectors is $\frac{2}{\|\omega\|}$.

To get the superior separability, the maximal distance must be maximized $\frac{2}{\|\omega\|}$, or sometimes minimized $\|\boldsymbol{\omega}\|$. The latter term may be solved using the Lagrangian method through minimizing Eq. 3-22 and it is equal to minimizing $\frac{1}{2}\|\omega\ \|$ (Sonka et al., 2014) and (Hamel 2011).

$$L=\frac{1}{2}||\omega||^2-\sum_n \alpha_n[y_n(\omega.x_n+\omega_0)-1] \qquad 3\text{-}22$$

Via deriving Eq.3-22, Eq. 3-23 is acquired

$$\frac{\partial L}{\partial \omega}=\omega-\sum_n \alpha_n y_n x_n=0 \qquad 3\text{-}23$$

And ω is defined in Eq. 3-24

$$\omega=\sum_n \alpha_n y_n x_n \qquad 3\text{-}24$$

Through substituting $\boldsymbol{\omega}$ in Eq. 3-22, Eq. 3-25 is acquired

$$\frac{\partial L}{\partial \omega_0}=-\sum_n \alpha_n y_n=0 \qquad 3\text{-}25$$

Thus, SVM classifier can be acquired the discriminant function as Eq. 3-26.

$$d(x^T)=-\sum_n y_n \alpha_n x_n x^T+\omega_0 \qquad 3\text{-}26$$

Where y_n represents the class label for support vector x_n, x^T represents the set of attributes while $\propto_n$ represents numeric parameter which is automatically determined through the optimization process (Hamel 2011).

5. Artificial Neural Network

The artificial neural network (ANN) is an intelligent model that is built to simulate the human brain neurons (Ross and others 2004) in the processing. It solves several problems in the different areas, such as pattern recognition, object recognition, classification, signal processing, and robotics. These networks are specified by the pattern of connectivity, processing elements learning rules, the strength of weights and training. Initial weights are set, and these weights adjust during implementation (Birry 2013) and (Wilson and Ritter 2000).

- Multilayer Perceptron

The multilayer perceptron (MLP) neural network is a model well known in ANNs. It has provided a nonlinear mapping of its input and the output of net to solve several problems in the prediction areas and medical applications diagnosing (Birry, 2013; Jiang et al., 2010) and (Hwang and Hu 2001). Typically, the MLP network is shown in Figure: 3-12. It consists of levels of input; hidden layers are one or more and an output layer. Such that the number of input layers corresponds to the number of parameters in features vector. Usually, no neuron function performed in that layer (Hwang and Hu 2001) While the output layers based on the output of the processed problem. The neurons of the net are completely connected where every neuron a layer connected to whole neurons of next layer (Larose 2005), (Han et al., 2011) and (Günther and Fritsch 2010). Typically, for choosing the hidden layers no a theoretical limited but the maximum number does not exceed one or two layers for processing of pattern recognition problems. Also, increasing the hidden layers affect the classification accuracy because they cause the overfitting, computation cost, memorizing the training set. While increasing the neurons number in hidden layer increases the optimization for tackling complex cases. Therefore, the number selection of the hidden layers and neurons is experimental.

The most common functions that effectively contribute to decision-making neurons for classification process are the hyperbolic tangent and sigmoid functions. Such functions accepting the nonlinear functions, approximation conditions of ANNs, and more differentiable (Lekutai 1997), (Özkan and Erbek 2003) and (Negnevitsky 2005).

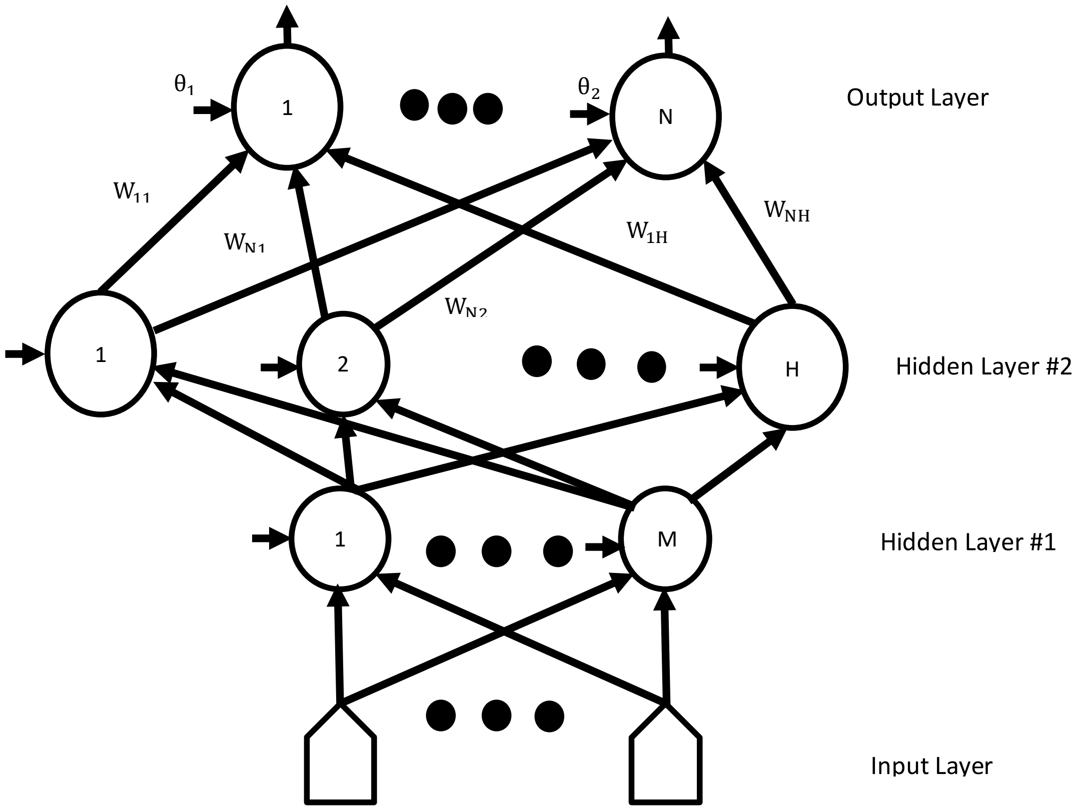


Figure: 3-12 Multilayer Perceptron Configuration (Hu and Hwang, 2001)

The sigmoid function is defined as Eq. 3-27.

$$f(x) = \frac{1}{1+e^{-x}} \qquad 3\text{-}27$$

This function maps a large input for any value between plus and minus infinity and within the small range from 0 to 1. Also, it is a nonlinear function, therefore, permits the MLP to classify the linearly inseparable data (Han et at., 2011). As well as, the error of a sigmoid function has a positive derivative and is flat or smooth (Birry, 2013; Rojas, 2013).

The second function is used in the hyperbolic tangent function; it is defined as following Eq. 3-28 (Graupe 2013):

$$f(x) = \tanh(x) = \frac{1-e^{-2x}}{1+e^{-2x}} \qquad 3\text{-}28$$

It is a bipolar version of the sigmoid function to map data into the infinity range of -1 to +1. The hyperbolic tangent function is more active classification than sigmoid function and has a faster convergence of learning algorithm (Özkan and Erbek 2003).

The backpropagation is a popular learning algorithm in the training of MLP, was developed by Rumelhart Hinton and Williams in 1986 (Lekutai 1997). To minimize the mean squared error between actual output and target of the MLP network, an iterative descent method is used (Larose 2005). Two issues affect training and designing of MLP network: the number of neurons and hidden layers that affect the final output in the network. Therefore, the number of hidden layers and neurons must be selected carefully (Panchal et al. 2011).

Recently, these classifiers are based in the works for classifying the nodules to nodule or non-nodule in automatic and semi-automatic systems for lung nodules detection. The efforts could be discussed in brief below:

The work of (Campadelli et al., 2006) reduced the false positive number obtained from an effective recognition of true nodules by sensitive SVM training. Experimental results were performed on two data sets, applying for Gaussian, features selection means and appointing various parameters to train SVMs. The compared results of the best SVM gained approximately 1.5 false positives for each image with sensitive increases equalling 0.71. The study of (Eskandarian and Bagherzadeh 2015) was based on SVM for the solitary nodule diagnosis in an attempt to gain more accuracy. SVM is combined with a threshold to describe lung areas and to reduce the false positive ratio. The obtained result demonstrated 89.9% sensitivity and false positive per scan of 3.9. The approach of (Ying, Tong, and Ming-Xiu 2011) employs different algorithms in the segmentation and detection of the solitary nodule, features selection and classification. It provides a solitary nodule segmentation based on multi-scale morphological filtering while the features are selected by separating their probability distribution, to be classified using SVM. The sensitivity was 94.6% of nodules detection for twenty cases. Interestingly, (Jacobs et al. 2014)have applied a privileged set of 128 features, which defined the shape, texture and aspects of intensity, to detect sub-solid nodule candidates. These features improve the classification performance of the CAD system in recognizing the sub-solid nodule within computed tomography images. This system used different classifiers with SVM to validate the algorithm and evaluated by a spacious dataset from sites of a multi-centre lung cancer screening trial. The proposed system reached 80% sensitivity at an average of only 1.0 false positive detections per scan.

3.11 Computer-Aided Detection/Diagnosis Systems (CADs)

Since the early years, many researchers have been attempting to develop computer-aided diagnosis/detection systems to detect, segment and diagnose lung tumours from CT scans (Ma et al., 2009). These systems still face challenges in terms of sensitivity, specificity, accuracy, false positive rates, full automation, speed in the nodule detection and dataset size. The challenges are the result of varying intensities and irregular shapes of lung tumours in CT images. Therefore, the complex characteristics of tumours pose difficulties in predicting the size, location and types (benign or malignancy) of the nodule and in making the radiologist distrustful CAD to be based in the medical practice. Although researchers attempt to construct efficient and active systems, the ideal results require more efforts to be achieved. Consequently, several methodologies in computer applications continuously handle the problems above in this area. Some works are summarized to explain computer-aided detection systems processing to 2D &3D CT lung images in detecting the nodules. In the work of (Javaid et al. 2016) proposed an approach to segment and detect the challenging pulmonary (juxta-vascular and juxta-pleural) nodules. The k-means and shape specific morphology is applied to detect the nodule after lung segmentation by a threshold. The next step is that potential nodules are divided into six groups depending on their percentage of connectivity with lung walls and thickness. These groups have eliminated false positives (FP) in each set of nodules by describing their salient features. The sensitivity is 93.8% for the proposed system as well as the overall sensitivity of 91.65% of receiver operating characteristic (ROC) curve. However, the system's sensitivity of small nodules is still lower than the total sensitivity of the algorithm. (X. Zhang et al. 2005) suggested a new automated method to segment nodules attached to vessels. A scheme developed on a parametrically deformable geometric model was based on processing the problem of segmenting juxta-pleural nodules. The false positives were 7.5 per exam.(Hara et al. 2006) developed a recognition approach to detect small nodules that have a diameter within range 3-16 mm. This algorithm presented a high sensitivity of 94% with 2.5 false positives rate per scan. However, the algorithm validation that was accomplished on 139 nodules that were acquired from a private database that may result from an unreliable diagnosis, as well as the number of nodules is low. (Ozekes et al., 2008) suggested new approach segments some regions of interest using Genetic Cellular Neural Networks technique. In this approach, the lung is separated, and the eight directional search technique then processes its interest regions. The technique had a sensitivity of 97% with 10.5 false positives for the exam. (Q. Wang et al. 2013) obtained acceptable results in the system sensitivity rate, but the false positive rate is still of a high value of 9.1 per image. In addition, the research-based database is not public, and hence avoids the replication of results. (Cascio et al. 2012) offered sensitivity of 97%, but with 6.1 false positives per exam. When the sensitivity decreased to 88%, the technique achieved a better rate of false positives, existing in this case at 2.5 per exam. In the methodology of (Santos et al. 2014), the small lung nodules that have diameters between 2 to 10 mm are detected by Tsallis entropy, Gaussian mixture models, and SVM. Some techniques, such as Shannon's entropy and Hessian matrix have been implemented in this work as well. The algorithm achieved specificity of 85%, sensitivity of 90.6%, accuracy of 88.4% and 1.17 false positives per slice. For the training stage, 112 exams with 118 nodules were used while 72 nodules of 28 exams of LIDC database have validated the algorithm. A developed approach of (Badura and Pietka 2014) has segmented types of pulmonary nodules in CT lung. The method adopted evolutionary computation and fuzzy connectivity (FC). LIDC–IDRI databases were selected for algorithm validation and included juxta-vascular, juxta-pleural, isolated and low-density nodules that have diameters between 3 and 30 mm. A total of nodules were 23 and 551 from LIDC and LIDC–IDRI respectively used and 50% was the true positive rate of the exams. In a study of (Gong et al. 2018), three Machine learning (SVM, naïve Bayes classifier and linear discriminant analysis) classified all nodules to benign and malignant in a CADs scheme which is implemented on CT lung images. And, the performance assessment of the three models has been validated by applying cross-validation method. In addition, the area under curve (AUC) has been computed to compare the discriminant power of classifiers, which were fluctuated between 0.88 and 0.99. The early detection of lung cancer has achieved in the study of (Tirz\"\ite et al. 2018) which used the logistic regression analysis (LRA) to discriminate the stage of cancer. The overall sensitivity was 95.8% for smokers and 96.2% for non-smokers.

Successful works in three-dimension have been performed to detect and describe lung nodule characteristics. The 3D-view offers an accurate description of shape, size, features of the nodule, and distinguishes the nodule from other structures in the lungs.

(Fetita et al. 2003) proposed a method to be used in the 3D space of the thorax's volume, based on a specific grey-level mathematical morphology operator in order to discriminate lung nodules from other dense (vascular) structures. The method presents the false positives rate equals 8.5 per exam. In the work of (Way et al. 2006), 3D active contour method has applied to 96 nodules which have been extracted the features and classified by a linear discriminant analysis classifier. (Ozekes and Osman 2010) developed an algorithm of computer-aided detection (CAD) system based on three-dimensional (3D) to detect lung nodules for 16 exams from LIDC with 16 nodules which have diameters between 3.5 and 7.3 mm. The algorithm is evaluated by feed forward neural networks (NN), support vector machines (SVM), Naïve Bayes (NB) and logistic regression (LR) methods. To compare the results of methods, the methods were trained and tested via K-fold cross-validation. The algorithm offered a sensitivity of 100% with 13.37 false positives per exam. (El-Baz et al. 2013) created a developed algorithm to detect lung nodules using Genetic Algorithm Template Matching within three stages: isolating nodules, arteries, veins, bronchi and bronchioles from other attached components; isolating the nodule by deformable 3D and 2D templates; finally, reducing the false positives by robust defining of three geometric features and the grey levels distribution of a nodule of the same kind. The algorithm presented rates of 82.3% and 9.2% for sensitivity and false positives respectively, wherein it was validated in a private database, i.e., in only 12 samples of the whole data set and 130 nodules for three types of nodules, all with diameters that were greater than 10 mm. (Choi and Choi 2014) developed an automated technique which detects the previously segmented nodules by applying enhancement filtering of multi-scale dot and extracting the characteristics that describe object shape in 3D. After that, a technique was used to eliminate the nodule edge for refinement. To validate the proposed technique, the LIDC dataset (acquired up to 2009) was adopted in this work. Juxta-pleural and juxta-vascular nodules were detected in 84 exams, which included 148 isolated nodules with diameters ranging between 3 and 30 mm. The sensitivity, accuracy, and specificity of the algorithm were 97.5%, 99.0% and 97.5% respectively. In addition, the false positives number was around 6.76 per examination. (Hamidian et al. 2017) trained a 3D CNN using volumes of interest (VOI) to detect pulmonary nodules in chest CT images automatically, and then converted the 3D CNN which has a fixed field of view to a 3D fully convolutional network (FCN) which can generate the score map for the entire volume efficiently in a single pass. And this screening FCN is used to generate difficult negative examples which are used to train new discriminant CNNs.

In discussing the works above in terms of the challenges that face the computer-aided diagnosis/detection to be applicable in practice, it still needs further investigation. Despite the reasonable sensitivity that is presented by different algorithms and that reached 100%, the false positive rate remains a high value, and the researchers still validate their algorithm with a low nodules number.

3.12 Advantages and Disadvantages of Detection Techniques

To summarise, this chapter presented some of the former studies, which could be outlined as follows:

The previous works introduced efficient methods that proved their potentials in the detection, segmentation, and classification of pulmonary nodules in CT-scan images using different approaches. Although these algorithms are promising concerning the experimental results in lung pathologies detection, they are still insufficient with various conditions accompanying the medical images. The mentioned techniques included some challenges which be summarised in the following points:

1- Disadvantages of lung segmentation techniques (Van Rikxoort et al. 2009), (Tseng and Huang 2009) and (Jiahui Wang, Li, and Li 2009).

- Although the optimal threshold can be found, there is still a need for many segmentations of the lung.
- Threshold adjustment and determining lung segments boundaries are time-consuming.
- Some of the issues such as unwanted background and air pixels accompany the threshold results.
- The inaccurate segmentation of images that have large airways and trachea attached to the lung.

2- Drawbacks of pulmonary nodule segmentation techniques and CAD systems

- The segmentation techniques struggle in detecting the attached nodules to dense pathologies

- The presence of the lung diseases with their different contrast could be considered as one of the challenges in nodule segmentation.
- A poor ability of segmentation methods to detect the diverse pathologies patterns in the radiology imaging.
- Absence the full automation, which considers one of the CAD requirements to efficient detection, is an existing challenge.

3.13 Summary and Lessons learned from the Literature Review

The primary idea was built using knowledge included in the scope and understanding of the problem domain. The previous cognizance of medical imaging characteristics (CT-scan) and comprehensive of dataset form a view of appropriate techniques to accomplish detection and classification tasks in CT lung images perfectly (Motoyama et al. 2007) and (Washko et al., 2012). The former studies declared a viewpoint which showed the defects of the studies that inspected in CT lung image field (Mansoor et al. 2015) and (Van Rikxoort et al. 2009). State of the art for this research summarized the failure of CADs to be addressed in other works. The defects and challenges are various and seek for finding ideal solutions in terms of accuracy and precision in detecting the pulmonary lung nodules. The challenges could be mentioned briefly as:

A detection difficulty of potential nodules which have irregular shapes in lung and thus no an ideal segmentation of nodule. Also, a low density and similar nodules tissue with the background are additional factors that affect nodules detection and increasing false positives rates and false negative. Full automation in nodules detection and description avoids the detection system to be based in CT centres.

Also, there challenges in the diagnostic dataset which is taken from the public Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI) (Samuel *et al.*, 2011) and (Clark et al. 2013) on which this research bases. Four radiologists assessed it and revealed estimated measurements (volume, location, and diameter) of nodules greater than 3mm while the nodules that are less than 3mm in size were only their numbers mentioned. Therefore, the detection chance of such nodules became an impossible task. The size and diameter length of the nodule are very vital parameter based by the radiologist to estimate the nodule malignancy rate and to identify the management strategy of the pulmonary nodule treatment. Whether the followed strategy is resection, follow-up or treatment of the nodule as declared that in the medical researchers (Lederlin et al. 2013). There is a significant interest in different approaches used to measure the nodule size and diameter to match or, at least, approximately meet of radiologists' assessment, in order to obtain accurate information about the nodule diameter length for estimating tumour progress.

In general, this work focused on the target of improving diagnosis performance by constructing an automatic detection and classification system for and pulmonary nodules detection. This increases the radiologist's confidence with such systems to be based in medical practice (Roy, Sirohi, and Patle 2015),. To evaluate the detection systems, qualitative and quantitative measurements were sought in this study (Taha and Hanbury 2015).

CHAPTER 4: DATA COLLECTION AND SEGMENTATION OF CT IMAGES

4.1 Introduction

The aim of this study is to construct a system that is able to detect and describe pulmonary nodules precisely, offering the potential for the early detection of cancer and rejection of non-cancer cases. This will save the radiologists’ effort and time, which are spent in diagnosing and scan for long images stack of the case may be normal. In addition, reducing the workloads in CT centres, fasting work by automatic detection systems is sought in the medical practice. These systems support radiologists’ interpretation of medical images and the detection of lung cancer, which includes several challenges that obstruct the successful diagnosis. The challenges are such as increasing faint contrast and extremely fuzzy margins of pulmonary nodules, irregular and non-spherical shapes of nodules that confuse the radiologist’s diagnosis (Naresh and Shettar 2014). Therefore, the automatic detection systems of lung cancer remain the researcher’s goal in this field in order to solve diagnosis problems.

The system is evaluated and compared using appropriate tools for all its stages. The work is divided into five areas: lung segmentation, vessels and nodule segmentation together, shape and texture features extraction, classification and developed 3D clustering to describe the nodule by a novel approach to detecting the nodules.

The segmentation and detection approach to present the results of the nodules greater than 3mm in this work, which seek to determine the ideal approach and completes the final validation. Figure 4-1 and Figure 4-2 show two block diagrams for the general stages and proposed system respectively. The first figure identifies the main processes of work while the second figure describes the processing stages in details.

Segmentation
Classification
Detection and description

Figure 4-1 General Block Diagram of the Detection System

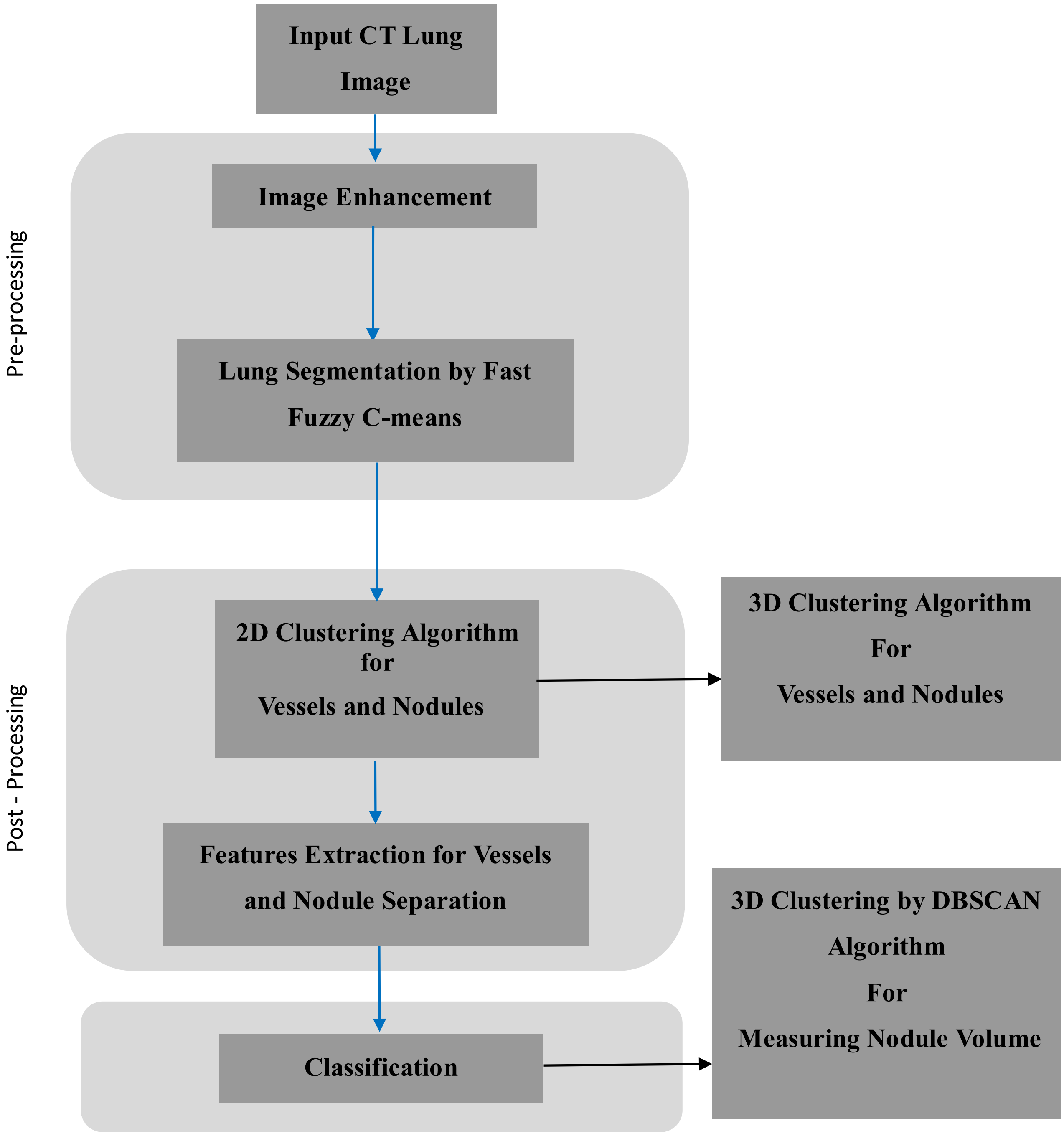

Input CT Lung Image
Image Enhancement
Lung Segmentation by Fast Fuzzy C-means
Pre-processing
2D Clustering Algorithm for Vessels and Nodules
3D Clustering Algorithm For Vessels and Nodules
Features Extraction for Vessels and Nodule Separation
Post - Processing
3D Clustering by DBSCAN Algorithm For Measuring Nodule Volume
Classification

Figure 4-2 Block Diagram of the Proposed System Processes

4.2 Data Collection

The most important part of this thesis is the data collection to test the developed prototype. 180 nodules from 2D CT-scan lung images of 512*512 pixels were chosen for testing and validating the algorithm. Initially, the data were real and had been collected from different Iraqi hospitals. Then, the work plan was modified for many reasons, which can be summarised as follows. First, the missing of early diagnosis of cancer due to the limited ability of some centres in terms of the archive of patient's history, which created difficulty in diagnosis, the cases in the early stages. Secondly, there are the patients' rights, since some of the health centres refused to provide the data of their patients. Therefore, it was too difficult to collect sufficient numbers of nodule cases required to test and validate the algorithm. The third problem is presented regarding the image quality because the data as mentioned previously were collected from various hospitals and devices, and therefore the images involved a high level of noise. And, the data is usually diagnosed by only one radiologist that may be not adequate to reliable diagnosis results. According to the above problems, the precise diagnosis of the data from these Iraqi hospitals was impossible. Therefore, the standard data is considered a good alternative has been completely used in this work. Because of this data has been diagnosed by four experts, reported diagnosis information in the attached files of the site, it has a lot of cancer cases with their numbers and characteristics of the CT lung images so that it can validate the approach (Armato *et al.*, 2011).

4.2.1 Standard Data

The pulmonary nodules data were obtained from the Cancer Imaging Archive[1]. These data were provided with a diagnosis to four radiologists at least in one file which describes abbreviated information about the nodules such as the total number of nodules, which

---

[1] https://public.cancerimagingarchive.net/ncia/login.jsf

sometimes exceed 30 nodules, and their size being greater or less than 3 mm. This data was updated to contain further information about the nodules that are just greater than 3mm in the separated file later. The file includes volume, diameter, location and slice number of nodules for each case. The patient stack images are saved as DICOM. However, these images are still raw data that include high levels of noise and the common outliers in the medical images. Additionally, this needs to pass through the pre-processing stage such as lung separation of the image background, enhancing the image contrast and removing noise, which was caused by motion, streak and metal artefacts. The CT scanning includes a series of slices. Figure 4-3 displays a) the CT scanning of the human lung; b) an example of CT lung scanning slices with continuous CT sequences, where the width and length are 512 and the thickness of slices is 2.5 mm according to the information is saved in tags DICOM of CT-scan; and c) normal axial CT lung slice.

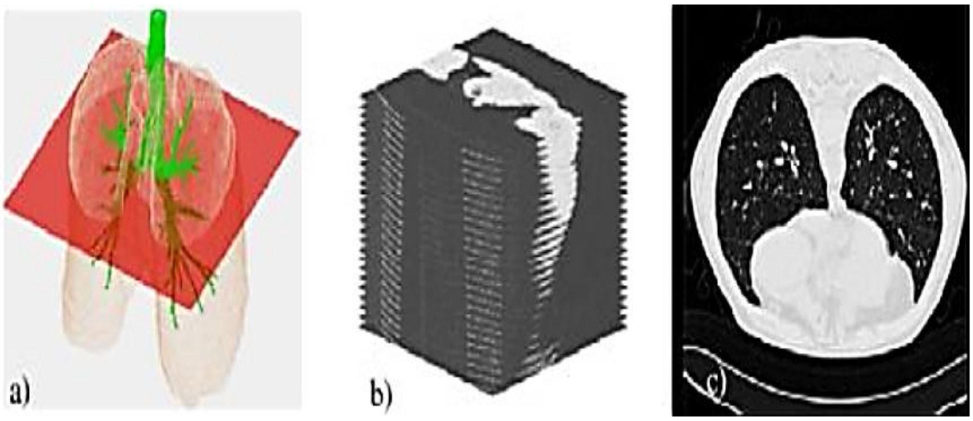
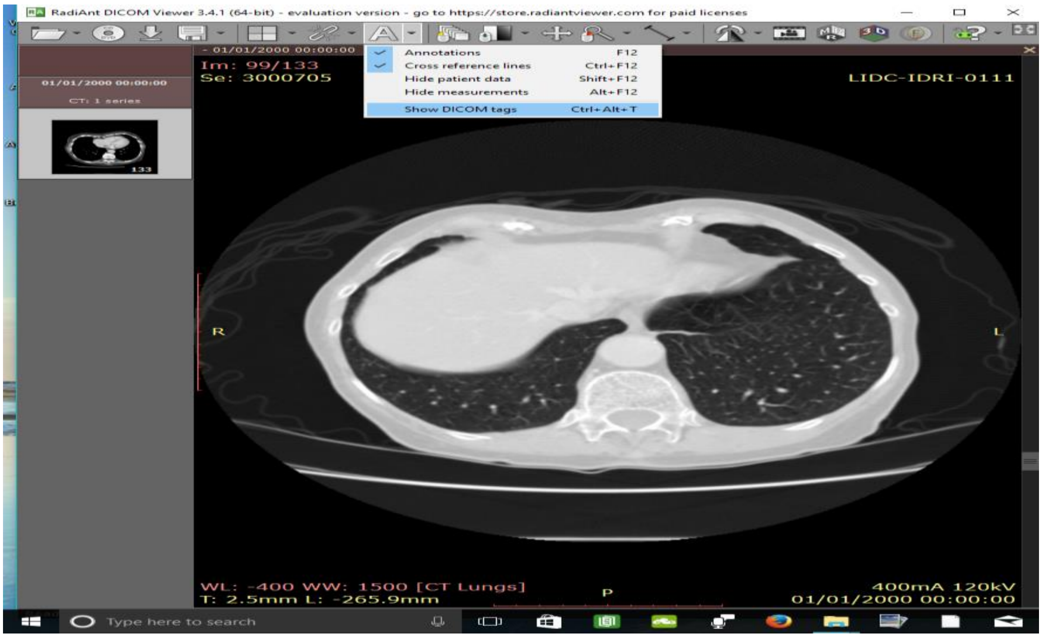
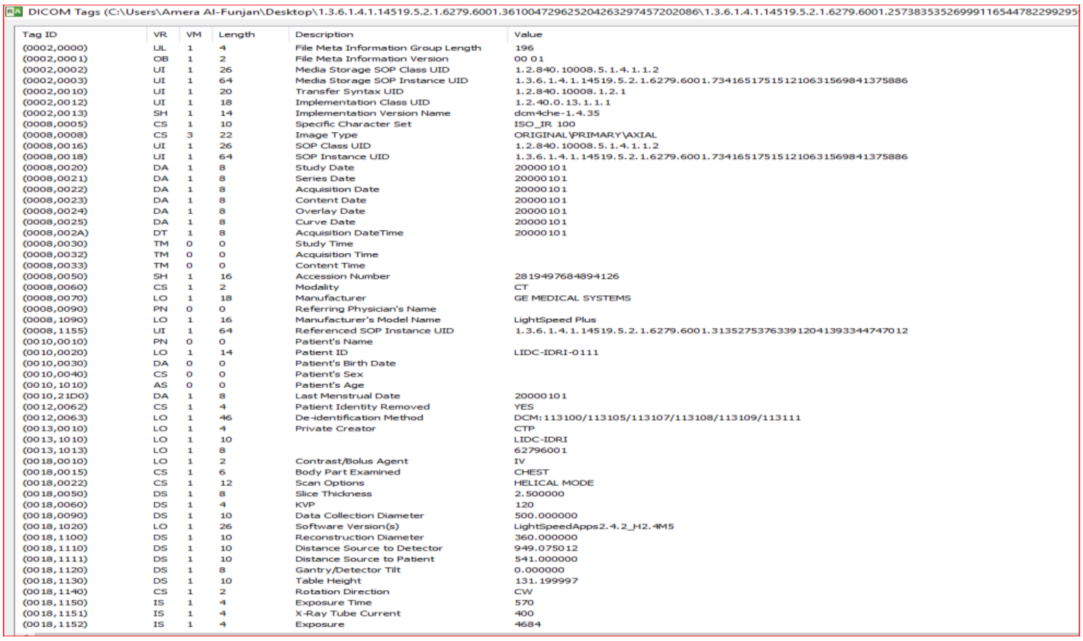


Figure 4-3 a) Human Lung; b) Continuous HLCT Images; c) Normal HLCT Slice (Huaqing & Chen, 2012)

4.2.2 DICOM Standard

Recently, the data structure of medical images has been included in a version named Digital Communications in Medicine Services (DICOM). This version is a simple data transfer and uses the model for data structure and identifiers that are responsible for network-oriented services and objects, which consist of images, patients' data and reports. DICOM is considered as an evolutionary archive of the pictures, and it presents interfacing among the medical information systems. The file format of DICOM supports other useful information for describing the image and data exchange, which are easy, safe and fast, avoiding any potential confusion caused by multiple files from the same study. Furthermore, it provides convergence for greyscale perception, ensuring consistency, management and efficient quality of presentation (Blume and Hemminger, 1997; Mustra et al., 2008). Figure 4-4 shows how to access the DICOM tags that contain all information about scan image.


RadiAnt DICOM Viewer 3.4.1 (64-bit) - evaluation version - go to https://store.radiantviewer.com for paid licenses
- 01/01/2000 00:00:00
01/01/2000 00:00:00
CT: 1 series
133
Im: 99/133
Se: 3000705
Annotations F12
Cross reference lines Ctrl+F12
Hide patient data Shift+F12
Hide measurements Alt+F12
Show DICOM tags Ctrl+Alt+T
LIDC-IDRI-0111
R
L
WL: -400 WW: 1500 [CT Lungs]
T: 2.5mm L: -265.9mm
P
400mA 120kV
01/01/2000 00:00:00
Type here to search



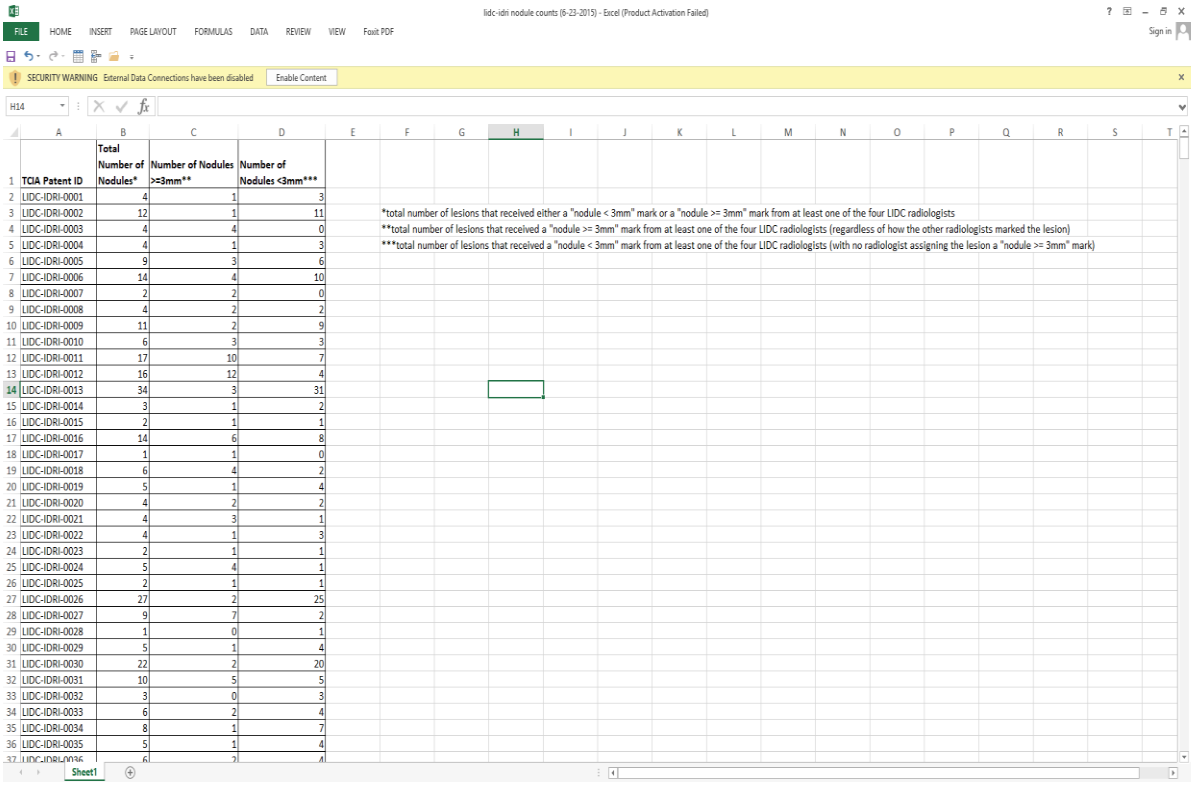
DICOM Tags (C:\Users\Amera Al-Funjan\Desktop\1.3.6.1.4.1.14519.5.2.1.6279.6001.361004729625204263297457202086\1.3.6.1.4.1.14519.5.2.1.6279.6001.25738353526999116544782299295

Tag ID	VR	VM	Length	Description	Value
(0002,0000)	UL	1	4	File Meta Information Group Length	196
(0002,0001)	OB	1	2	File Meta Information Version	00 01
(0002,0002)	UI	1	26	Media Storage SOP Class UID	1.2.840.10008.5.1.4.1.1.2
(0002,0003)	UI	1	64	Media Storage SOP Instance UID	1.3.6.1.4.1.14519.5.2.1.6279.6001.734165175151210631569841375886
(0002,0010)	UI	1	20	Transfer Syntax UID	1.2.840.10008.1.2.1
(0002,0012)	UI	1	18	Implementation Class UID	1.2.40.0.13.1.1.1
(0002,0013)	SH	1	14	Implementation Version Name	dcm4che-1.4.35
(0008,0005)	CS	1	10	Specific Character Set	ISO_IR 100
(0008,0008)	CS	3	22	Image Type	ORIGINAL\PRIMARY\AXIAL
(0008,0016)	UI	1	26	SOP Class UID	1.2.840.10008.5.1.4.1.1.2
(0008,0018)	UI	1	64	SOP Instance UID	1.3.6.1.4.1.14519.5.2.1.6279.6001.734165175151210631569841375886
(0008,0020)	DA	1	8	Study Date	20000101
(0008,0021)	DA	1	8	Series Date	20000101
(0008,0022)	DA	1	8	Acquisition Date	20000101
(0008,0023)	DA	1	8	Content Date	20000101
(0008,0024)	DA	1	8	Overlay Date	20000101
(0008,0025)	DA	1	8	Curve Date	20000101
(0008,002A)	DT	1	8	Acquisition DateTime	20000101
(0008,0030)	TM	0	0	Study Time	
(0008,0032)	TM	0	0	Acquisition Time	
(0008,0033)	TM	0	0	Content Time	
(0008,0050)	SH	1	16	Accession Number	2819497684894126
(0008,0060)	CS	1	2	Modality	CT
(0008,0070)	LO	1	18	Manufacturer	GE MEDICAL SYSTEMS
(0008,0090)	PN	0	0	Referring Physician's Name	
(0008,1090)	LO	1	16	Manufacturer's Model Name	LightSpeed Plus
(0008,1155)	UI	1	64	Referenced SOP Instance UID	1.3.6.1.4.1.14519.5.2.1.6279.6001.313527537633912041393344747012
(0010,0010)	PN	0	0	Patient's Name	
(0010,0020)	LO	1	14	Patient ID	LIDC-IDRI-0111
(0010,0030)	DA	0	0	Patient's Birth Date	
(0010,0040)	CS	0	0	Patient's Sex	
(0010,1010)	AS	0	0	Patient's Age	
(0010,21D0)	DA	1	8	Last Menstrual Date	20000101
(0012,0062)	CS	1	4	Patient Identity Removed	YES
(0012,0063)	LO	1	46	De-identification Method	DCM:113100/113105/113107/113108/113109/113111
(0013,0010)	LO	1	4	Private Creator	CTP
(0013,1010)	LO	1	10		LIDC-IDRI
(0013,1013)	LO	1	8		62796001
(0018,0010)	LO	1	2	Contrast/Bolus Agent	IV
(0018,0015)	CS	1	6	Body Part Examined	CHEST
(0018,0022)	CS	1	12	Scan Options	HELICAL MODE
(0018,0050)	DS	1	8	Slice Thickness	2.500000
(0018,0060)	DS	1	4	KVP	120
(0018,0090)	DS	1	10	Data Collection Diameter	500.000000
(0018,1020)	LO	1	26	Software Version(s)	LightSpeedApps2.4.2_H2.4M5
(0018,1100)	DS	1	10	Reconstruction Diameter	360.000000
(0018,1110)	DS	1	10	Distance Source to Detector	949.075012
(0018,1111)	DS	1	10	Distance Source to Patient	541.000000
(0018,1120)	DS	1	8	Gantry/Detector Tilt	0.000000
(0018,1130)	DS	1	10	Table Height	131.199997
(0018,1140)	CS	1	2	Rotation Direction	CW
(0018,1150)	IS	1	4	Exposure Time	570
(0018,1151)	IS	1	4	X-Ray Tube Current	400
(0018,1152)	IS	1	4	Exposure	4684

Figure 4-4: DICOM Tag

4.2.3 Data Diagnosis Reports

The diagnostic reports of standard data, on which this study is based, are attached to the site and contain the description details of cancer cases with further information about nodules. The first report is available on the site, which consists of the nodules number divided into nodules that are less and greater than 3mm for each patient. This report was modified to include specific characteristics of nodules (volume, diameter, location and slice number) that are just greater than 3mm in an individual file; four experts evaluated both reports. These files attached to sites in the reference of Armato et al. (2011), for example, samples of the content of the file are shown in figures below. Figure 4-5 shows the radiologists' report (first report) demonstrates the patient ID, total nodules number with two columns that included their classification according to the size (greater and smaller than 3mm). Figure 4-6 shows an updated report for nodules that are higher than 3mm only with some characteristics (volume, diameter, location and slice number) for each case was reported in the previous report.


lidc-idri nodule counts (6-23-2015) - Excel (Product Activation Failed)
FILE
HOME
INSERT
PAGE LAYOUT
FORMULAS
DATA
REVIEW
VIEW
Foxit PDF
Sign in
SECURITY WARNING External Data Connections have been disabled
Enable Content
H14
Sheet1

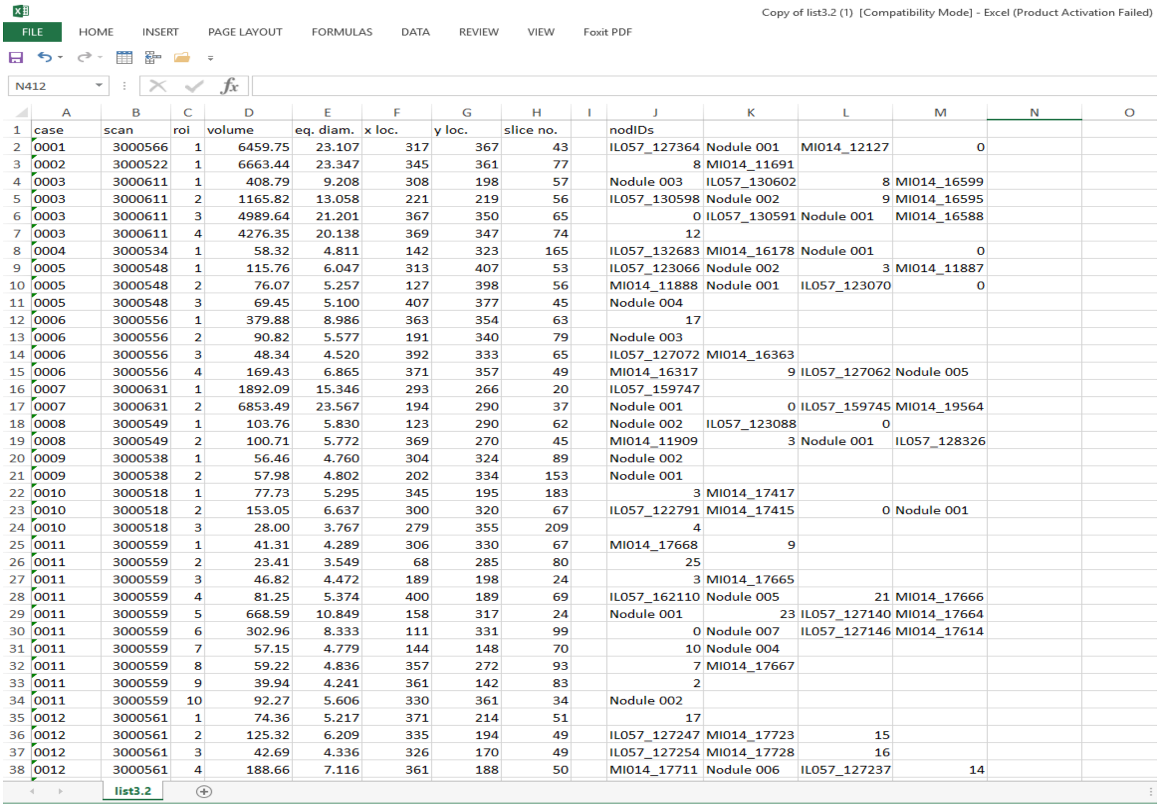
TCIA Patent ID	Total Number of Nodules*	Number of Nodules >=3mm**	Number of Nodules <3mm***
LIDC-IDRI-0001	4	1	3
LIDC-IDRI-0002	12	1	11
LIDC-IDRI-0003	4	4	0
LIDC-IDRI-0004	4	1	3
LIDC-IDRI-0005	9	3	6
LIDC-IDRI-0006	14	4	10
LIDC-IDRI-0007	2	2	0
LIDC-IDRI-0008	4	2	2
LIDC-IDRI-0009	11	2	9
LIDC-IDRI-0010	6	3	3
LIDC-IDRI-0011	17	10	7
LIDC-IDRI-0012	16	12	4
LIDC-IDRI-0013	34	3	31
LIDC-IDRI-0014	3	1	2
LIDC-IDRI-0015	2	1	1
LIDC-IDRI-0016	14	6	8
LIDC-IDRI-0017	1	1	0
LIDC-IDRI-0018	6	4	2
LIDC-IDRI-0019	5	1	4
LIDC-IDRI-0020	4	2	2
LIDC-IDRI-0021	4	3	1
LIDC-IDRI-0022	4	1	3
LIDC-IDRI-0023	2	1	1
LIDC-IDRI-0024	5	4	1
LIDC-IDRI-0025	2	1	1
LIDC-IDRI-0026	27	2	25
LIDC-IDRI-0027	9	7	2
LIDC-IDRI-0028	1	0	1
LIDC-IDRI-0029	5	1	4
LIDC-IDRI-0030	22	2	20
LIDC-IDRI-0031	10	5	5
LIDC-IDRI-0032	3	0	3
LIDC-IDRI-0033	6	2	4
LIDC-IDRI-0034	8	1	7
LIDC-IDRI-0035	5	1	4
LIDC-IDRI-0036	6	2	4

*total number of lesions that received either a "nodule < 3mm" mark or a "nodule >= 3mm" mark from at least one of the four LIDC radiologists

**total number of lesions that received a "nodule >= 3mm" mark from at least one of the four LIDC radiologists (regardless of how the other radiologists marked the lesion)

***total number of lesions that received a "nodule < 3mm" mark from at least one of the four LIDC radiologists (with no radiologist assigning the lesion a "nodule >= 3mm" mark)

Figure 4-5 Sample of Dataset before Updating


Copy of list3.2 (1) [Compatibility Mode] - Excel (Product Activation Failed)
FILE
HOME
INSERT
PAGE LAYOUT
FORMULAS
DATA
REVIEW
VIEW
Foxit PDF
N412
list3.2

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	case	scan	roi	volume	eq. diam.	x loc.	y loc.	slice no.		nodIDs			
2	0001	3000566	1	6459.75	23.107	317	367	43		IL057_127364	Nodule 001	MI014_12127	0
3	0002	3000522	1	6663.44	23.347	345	361	77		8	MI014_11691		
4	0003	3000611	1	408.79	9.208	308	198	57		Nodule 003	IL057_130602	8	MI014_16599
5	0003	3000611	2	1165.82	13.058	221	219	56		IL057_130598	Nodule 002	9	MI014_16595
6	0003	3000611	3	4989.64	21.201	367	350	65		0	IL057_130591	Nodule 001	MI014_16588
7	0003	3000611	4	4276.35	20.138	369	347	74		12			
8	0004	3000534	1	58.32	4.811	142	323	165		IL057_132683	MI014_16178	Nodule 001	0
9	0005	3000548	1	115.76	6.047	313	407	53		IL057_123066	Nodule 002	3	MI014_11887
10	0005	3000548	2	76.07	5.257	127	398	56		MI014_11888	Nodule 001	IL057_123070	0
11	0005	3000548	3	69.45	5.100	407	377	45		Nodule 004			
12	0006	3000556	1	379.88	8.986	363	354	63		17			
13	0006	3000556	2	90.82	5.577	191	340	79		Nodule 003			
14	0006	3000556	3	48.34	4.520	392	333	65		IL057_127072	MI014_16363		
15	0006	3000556	4	169.43	6.865	371	357	49		MI014_16317	9	IL057_127062	Nodule 005
16	0007	3000631	1	1892.09	15.346	293	266	20		IL057_159747			
17	0007	3000631	2	6853.49	23.567	194	290	37		Nodule 001	0	IL057_159745	MI014_19564
18	0008	3000549	1	103.76	5.830	123	290	62		Nodule 002	IL057_123088	0	
19	0008	3000549	2	100.71	5.772	369	270	45		MI014_11909	3	Nodule 001	IL057_128326
20	0009	3000538	1	56.46	4.760	304	324	89		Nodule 002			
21	0009	3000538	2	57.98	4.802	202	334	153		Nodule 001			
22	0010	3000518	1	77.73	5.295	345	195	183		3	MI014_17417		
23	0010	3000518	2	153.05	6.637	300	320	67		IL057_122791	MI014_17415	0	Nodule 001
24	0010	3000518	3	28.00	3.767	279	355	209		4			
25	0011	3000559	1	41.31	4.289	306	330	67		MI014_17668	9		
26	0011	3000559	2	23.41	3.549	68	285	80		25			
27	0011	3000559	3	46.82	4.472	189	198	24		3	MI014_17665		
28	0011	3000559	4	81.25	5.374	400	189	69		IL057_162110	Nodule 005	21	MI014_17666
29	0011	3000559	5	668.59	10.849	158	317	24		Nodule 001	23	IL057_127140	MI014_17664
30	0011	3000559	6	302.96	8.333	111	331	99		0	Nodule 007	IL057_127146	MI014_17614
31	0011	3000559	7	57.15	4.779	144	148	70		10	Nodule 004		
32	0011	3000559	8	59.22	4.836	357	272	93		7	MI014_17667		
33	0011	3000559	9	39.94	4.241	361	142	83		2			
34	0011	3000559	10	92.27	5.606	330	361	34		Nodule 002			
35	0012	3000561	1	74.36	5.217	371	214	51		17			
36	0012	3000561	2	125.32	6.209	335	194	49		IL057_127247	MI014_17723	15	
37	0012	3000561	3	42.69	4.336	326	170	49		IL057_127254	MI014_17728	16	
38	0012	3000561	4	188.66	7.116	361	188	50		MI014_17711	Nodule 006	IL057_127237	14

Figure 4-6 Sample of Dataset after Updating

4.3 Pre-Processing

Removing noise and segmenting organ from the image background is an efficient process used to reduce the search area about nodules within the image data and is considered one of the challenges that confront researchers in the medical image processing field (Van Rikxoort and Van Ginneken 2013). This process requires the separation of the objects that contain the region of interest in a way that ensures the preservation of the important information of the disease and lung characteristics. The low contrast and noise are considered the main factors that prevent the pathology features' vision in the medical image for both the radiologists and computer-aided diagnosis (CADs) systems. This work presents the segmentation of lungs from the CT image background, which is evaluated subjectively and objectively as an important part of the work stages.

4.3.1 CT Image Enhancement

The low contrast of the medical image and noise affect the visualisation and analysis of the image. Specifically, the radiologists depend on the bright and high-resolution of the image when reading the scan to see the internal anatomy of the organ and present the final diagnosis with high precision. Generally, the medical image encounters duplicated noise problems resulting from many sources, such as the movement of patients within the scan, the motion of molecules in the scanned tissue and streak and metal artefacts (Huaqing Chen, 2012). In this study, both Gaussian filter and adjustment function used to enhance the contrast of the CT lung image and smooth the image (Pratt 2001).

- Gaussian filter

Spatial domain methods are classified as enhancement algorithms for the image. The Gaussian filter is one of these methods, which is able to suppress noise before the lungs' segmentation of the image background. Furthermore, it shows spatial frequencies that have a fair range of the high and low filters frequencies (Kumar and Nachamai, 2012; Hamad et al., 2014).

- Adjusted image

Image adjustment aims to increase the contrast of the output image and make specific features easier to view by adjusting the intensities of the image. This technique maps the new range of image intensities' value in the greyscale and enhances the visual image.

4.3.2 Lung Segmentation (Fully Automated)

The importance of entire lung segmentation is to comprise the whole disease parts. Such challenge motivates the researchers to find and develop methods for this task. Also, the various contrast and homogeneity of lung organ are another obstacle hinders robust segmentation of the fuzzy lung edges. Furthermore, the juxta-pleural is a common nodule type attached to the lung wall and is often missed in the incorrect segmentation of lung boundaries (Jirapatnakul et al. 2011). Therefore, the implemented method to segment the lungs of the image background in this study is concerned with the image homogeneity. As well as, it will be evaluated qualitatively and quantitatively by a suitable metric that deals with complex boundaries.

- Fast Fuzzy C-Means

The fuzzy clustering algorithms are traditional in image applications (Suganya and Shanthi 2012). In fuzzy clustering, the data points are assigned as membership values for each of the clusters. And, fuzzy clustering algorithms allow the clusters to grow inside them. The algorithms have introduced promising outcomes in medical image especially. A fast fuzzy c-mean method is one of the improved clustering algorithms in the medical image applications and addresses speed problems from which fuzzy clustering suffers. Therefore, this method is necessary to accelerate the segmentation in this study and then save the diagnosis time. This algorithm could combine the grey information and local spatial to form a framework against noise and outlier problems that are common in the medical imaging. Furthermore, the algorithm reduces the segmentation time by depending on the number of grey-levels instead of the image size(Yon et al., 2004). Moreover, this method exploited fuzzy c-means advantages in dealing well with pixels and avoid disadvantages in terms of long computational time and sensitivity to noise and outliers (Ahmed et al. 2002) and (Biniaz et al., 2012).

- Morphological Operations

Morphological image processing is a group of non-linear operations connected to the morphology of features for an image. These operations are applied to binary and grey scale images and investigate the input image with a small template is called a structuring element. This template consist of an array of pixels is located at all possible positions in the image where each pixel corresponds to its neighbourhood. The aim of the structuring element is to distinguish expressive shape information in the image. In this study, using a disk-structuring element is an essential part of the dilation and erosion of the morphological operations and is applied to probe the input image. The structuring element is a matrix that recognizes the processed pixel in the image and identifies the neighbour that is processed with the pixel. For the input image, the structuring element is chosen to match the shape and size of the lungs (Pratt 1991) and (E. R. Dougherty and Lotufo 2003).

In this study, the method is fully automated for lung segmentation. It contains the following steps:

Pre-processing is as a step used an adjustment function and Gaussian filter to increase the contrast of the image and improve the visualisation of the image. Fast Fuzzy C-means is a popular method and is applied in the medical image segmentation which is proposed in the reference of Yon et al.(2004). This method is implemented in this work to segment the lung from the image background in the CT scan as part of the pre-processing stage. It identifies three clusters (background, lung wall and two lungs) in a 2D image (512*512).

Algorithm

Name: Fast Fuzzy C-means (FFC-means)

Input: array of image pixels (512*512)

Output: two objects (left and right lungs)

Begin

- Denoising image by the Gaussian filter.
- Adjust the contrast of greyscale image data and brightness increase by using `‘imadjust (I)'` function in Matlab.

Apply fast fuzzy c-means to cluster the image into three regions as follows:

Divide the image data partitions to 3 ranges, represented as 3 labels according to Eq.4-1:

$$\text{image}=\text{Grey scale values}/3 \tag{4-1}$$

Label1=0-85

Label2=86-170

Label3=171-255

Calculate the pixels-belonging degree to each region by membership function of the fuzzy c-mean method

Apply morphological operation (disk structuring element) to segment the lungs

Figure 4-7 shows the lung segmentation by Fast Fuzzy c-means, and the result is a segmented lungs image. Figure 4-8 shows the overall of processes of lung segmentation.

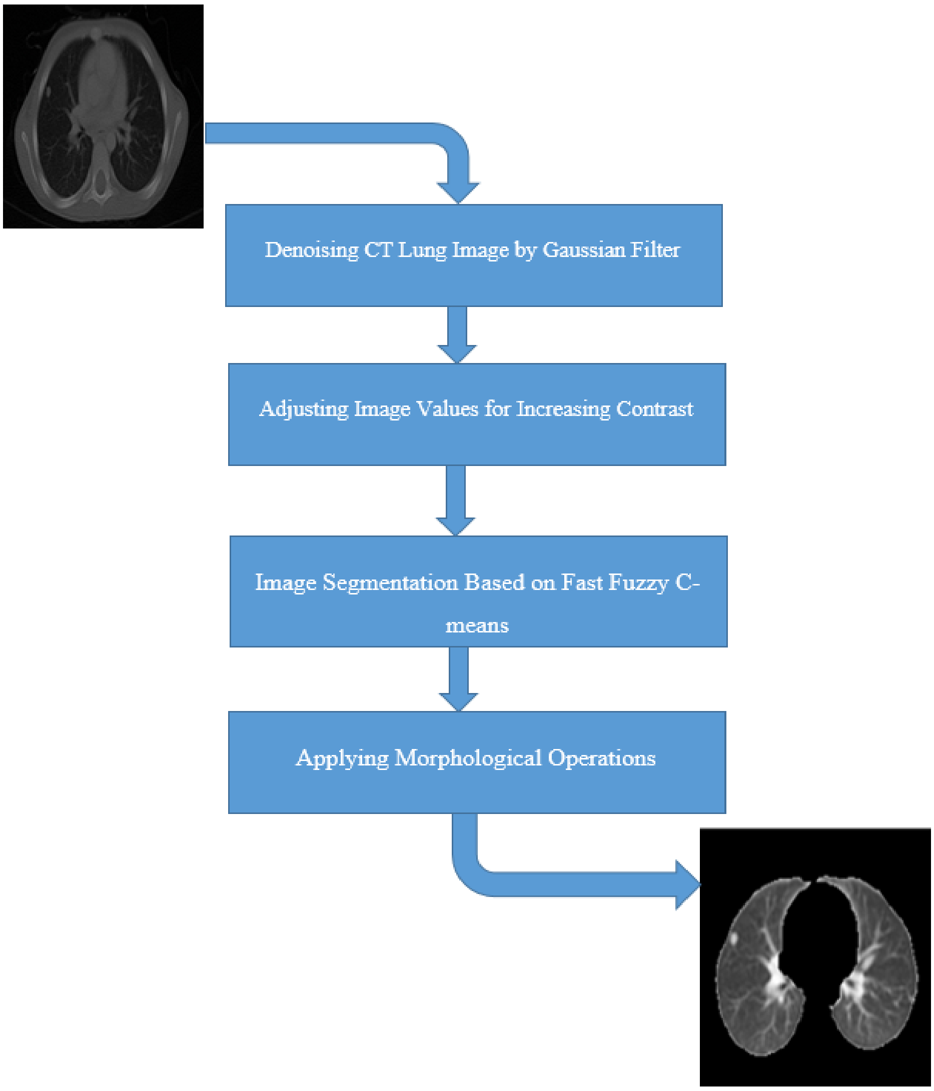

Denoising CT Lung Image by Gaussian Filter
Adjusting Image Values for Increasing Contrast
Image Segmentation Based on Fast Fuzzy C-means
Applying Morphological Operations

Figure 4-7: Block Diagram of Lungs Segmentation by Fast Fuzzy C-means and Morphological Operations

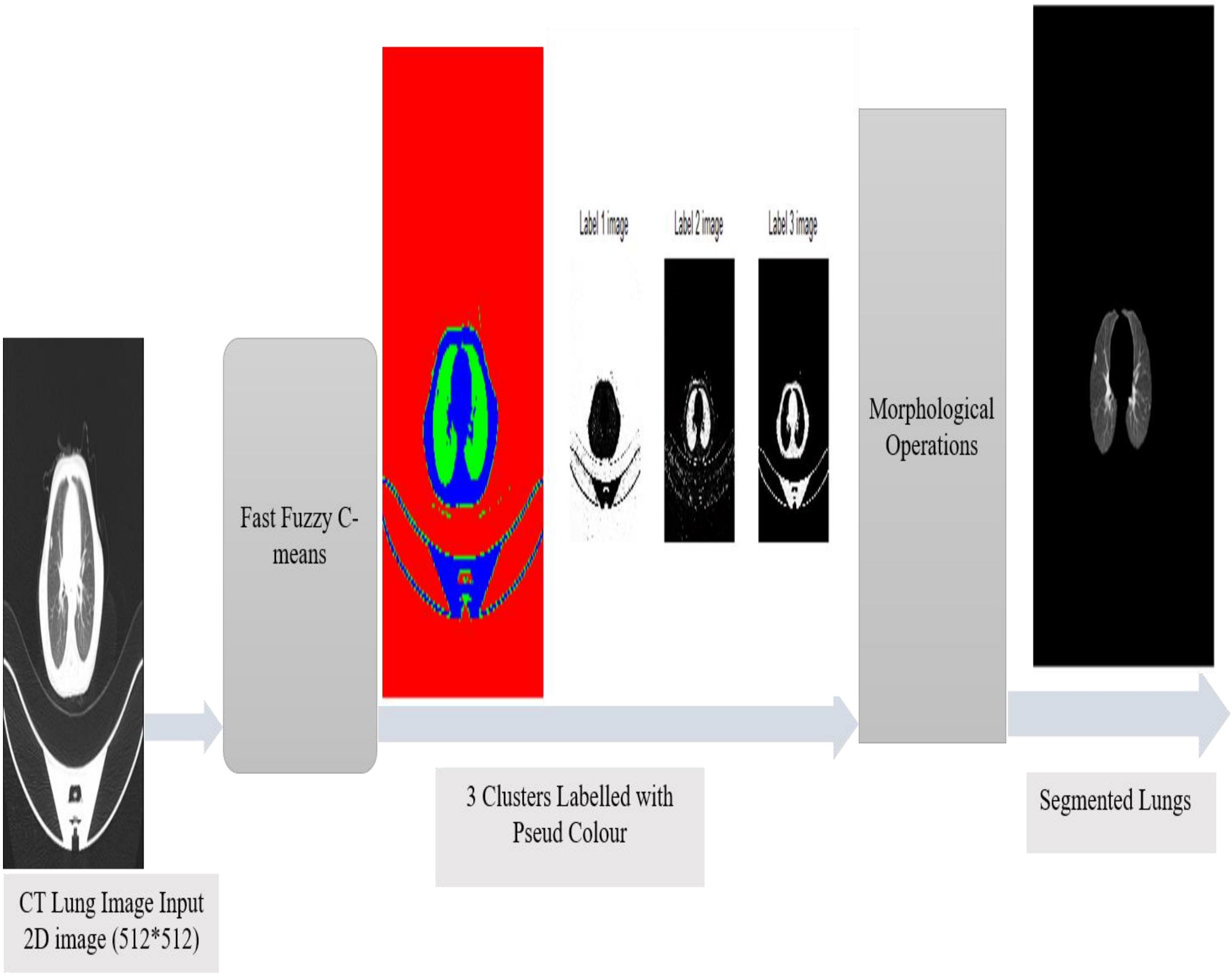


Figure 4-8: Overall Block Diagram of Lung Segmentation

4.4 Lung Segmentation Evaluation

The research methodology of the present work offers an evaluated framework that determines precision levels of the automatic detection system and evaluates each task within this study. The evaluation of lung segmentation task is necessary to determine the accuracy the separation of lung boundaries that often have attached nodules. Both quantitative and qualitative evaluations have made for this stage of the work. The CT lung images are assessed subjectively by a radiologist who manually delineates the lung edge to be compared with segmented lungs by the algorithm. For the objective evaluation, Hausdorff distance metric is a useful tool which assesses the lung boundary segmentation quantitatively (Taha and Hanbury 2015).

- Hausdorff Distance

One of the standard metrics which measure the distance between two points sets is Hausdorff Distance is considered sensitive to pixel position and deals with complex edges such as the irregular boundary of objects. Therefore, it is a suitable metric used to evaluate segmentation accuracy of lungs edge (Taha and Hanbury 2015), (Zhao, Shi, and Deng 2005). The mathematical formal of this metric can be defined in Eq. 4-2, Eq. 4-3.

$$\mathrm{HD}(\mathrm{A},\mathrm{B}) = \max(\mathrm{h}(\mathrm{A},\mathrm{B}), \mathrm{h}(\mathrm{B},\mathrm{A})) \qquad 4\text{-}2$$

Where A,B represent two sets of pixels of the evaluated images, $\mathrm{h}(\mathrm{A},\mathrm{B})$ represents Hausdorff distance that is given by

$$\mathrm{h}(\mathrm{A},\mathrm{B}) = \max_{\mathrm{a}\in\mathrm{A}}\min_{\mathrm{b}\in\mathrm{B}}\|\mathrm{a}\text{-}\mathrm{b}\| \qquad 4\text{-}3$$

$\|\mathrm{a}\text{-}\mathrm{b}\|$ is usually some norms such as Euclidean distance.

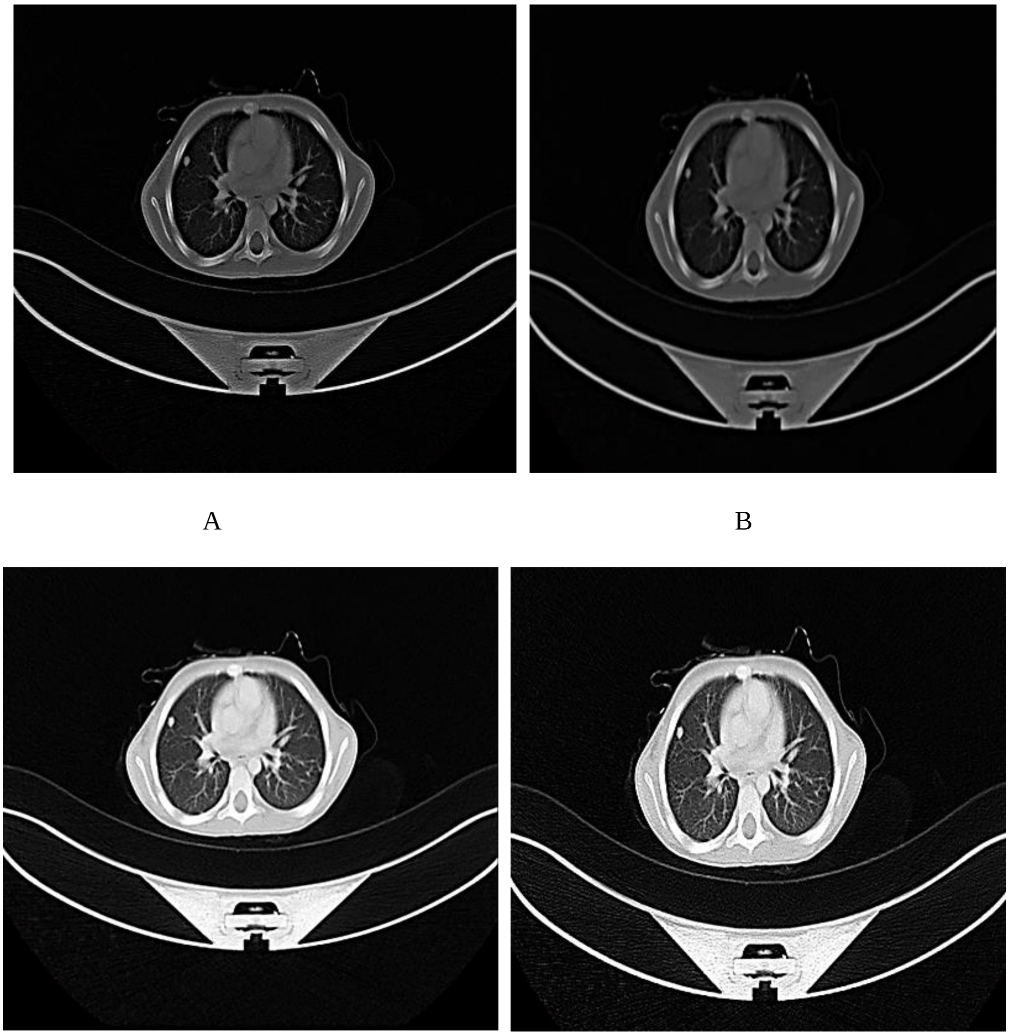
This metric calculates the distance between pixels of the segmented image and its reference. Where the reference image represents the original image that was manually segmented by a radiologist. The outcome of the metric is a value indicates the difference between the segmentation of images. The lowest difference value identifies a proper segmentation and refers to a better converge or matching between the segmented and reference images. The optimal value of the metric is zero which demonstrate that there is ultimately matching between images boundaries.

4.5 Experiments and Discussion

Dataset of CT lung images has used to validate the automatic algorithm. The images that used in this work are obtained from LIDC database, constituting a total number is 40 scans consist of 180 nodules. The image has a resolution of 512*512 and the total number of slices for scan stack from 60 to 380. The pulmonary nodules sometimes exceed 35 nodules in the 1 scan, whose size varied between less and greater than 3mm. The algorithm has validated the nodules with size greater than 3mm in this work.

4.5.1 Denoising and Adjustment

In terms of denoising and increasing contrast, two tools have been used. First, the Gaussian filter is applied to remove the image noise that leads to inaccurate segmentation. The second tool is the adjustment function that increases contrast, improves the intensity of the image, and provides brightness for image details such as vessels, nodules, airways. Figure 4-9 shows an original axial CT lung image is filtered by the Gaussian filter once, adjusted by applying adjustment function only and undergone by both the tools (Gaussian filter, adjustment function) once.



A B

C D

Figure 4-9: Denoising and Adjusting of CT- Scan Lung Image, A) Original CT Lung Image; B) Denoising CT Lung Image by Gaussian Filter; C) Increasing Contrast of CT Lung Image by Adjusting Algorithm and D) Denoising and Increasing contrast of CT Lung Image by Gaussian Filter and Adjusting Algorithm

4.5.2 Image Histogram

The histogram of the image shows the differences of an image responding for filtering and adjustment the same image. Figure 4-10 shows CT lung image with the histogram response for different enhancement using Gaussian filter and adjustment algorithm.

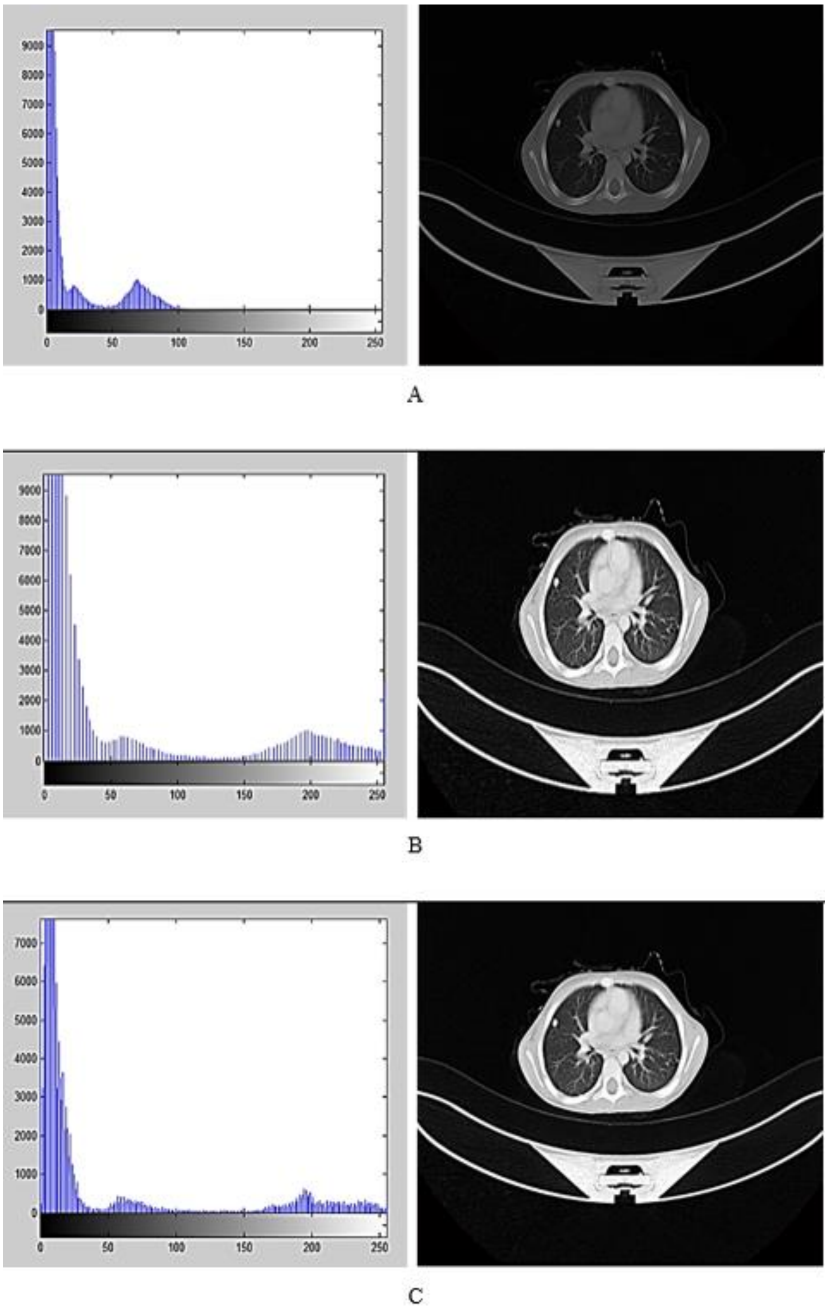


Figure 4-10: Corresponding Histograms Representation for Each One of a) Denoising CT Lung Image by Gaussian Filter; B) Increasing Contrast of CT Lung Image by Adjusting Algorithm and C) Denoising and Contrast Increasing of CT Lung Image by Gaussian Filter and Adjustment Algorithm

4.5.3 Lung Segmentation

The fast fuzzy c-means (FFCM) algorithm is implemented to separate the lung of image background using MATLAB version R2014a. In Figure 4-11, the algorithm labels the image with pseudocolour to three regions (two lungs, lung wall and background). Figure 4-12 displays three labelled images demonstrate the belonging degree the pixels of the three clusters regions that are divided into grey scale values. The segmentation of proposed method presents segmented lungs as shown in Figure 4-13 shows original CT lung images with segmented lungs by the algorithm.

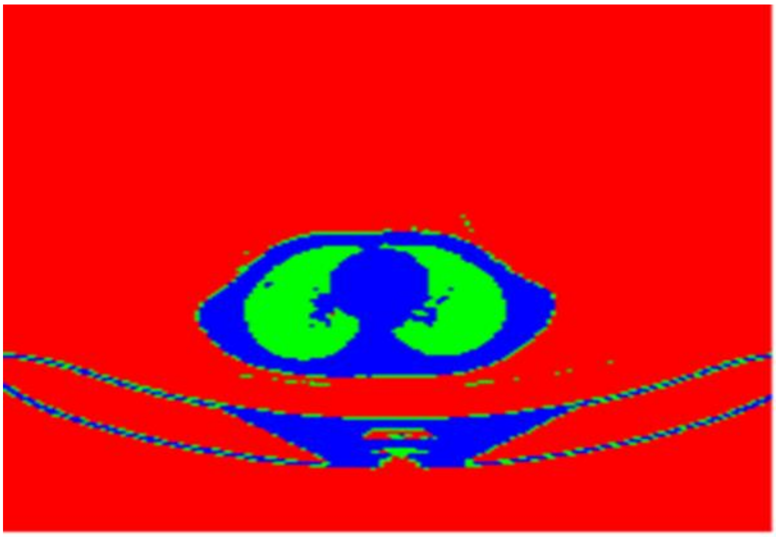


Figure 4-11: CT Lung Image with Pseudo Colour of Algorithm

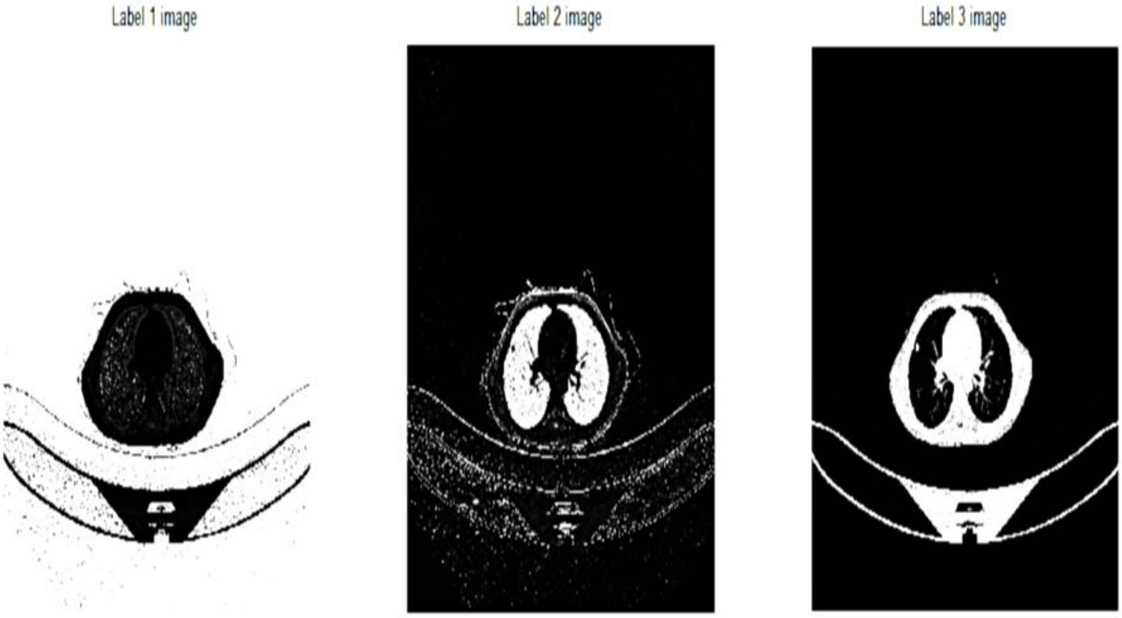


Figure 4-12: Fast Fuzzy C-means Algorithm Applied on CT Lung Image, Shows Pixels Belonging Degree to Each Cluster in the Image

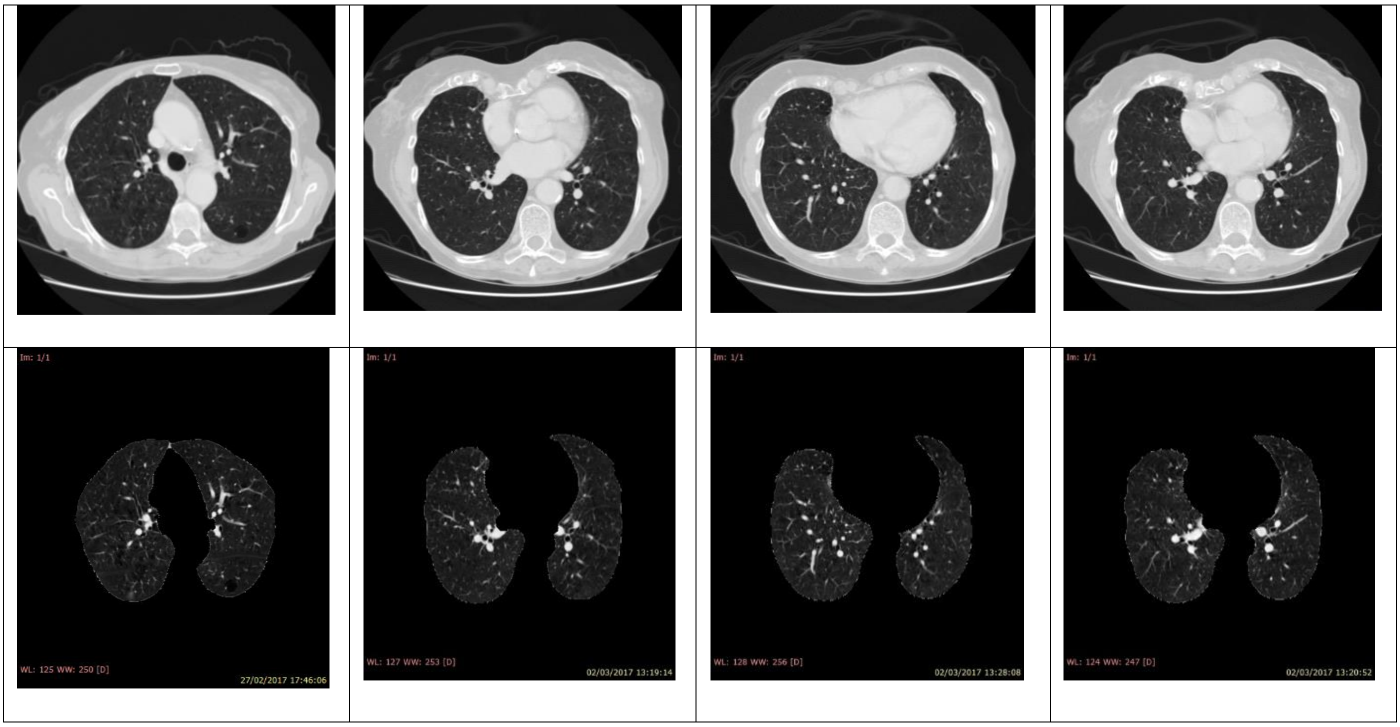


Figure 4-13: Samples of CT Original Lung Images and Segmented Lungs Images by Fast Fuzzy C-Means

4.6 Lung Segmentation Evaluation

The suitable metric to evaluate the Fast Fuzzy c-means method segmentation is Hausdorff distance. The metric has applied on 10 images are segmented by fast fuzzy c-means and their reference which was accessed by the radiologist. Table 4-1 shows the metric's values for the 10 images. Figure 4-14 shows samples of images were delineated their boundaries with green colour by a radiologist and then the same images were segmented by the fast fuzzy c-means method.

Table 4-1: Hausdorff Values for Ten Segmented Lung Images

Images	1	2	3	4	5	6	7	8	9	10
Hausdorff Values	6.659	5.012	4.590	6.048	8.201	7.011	5.25	4.390	6.432	5.352

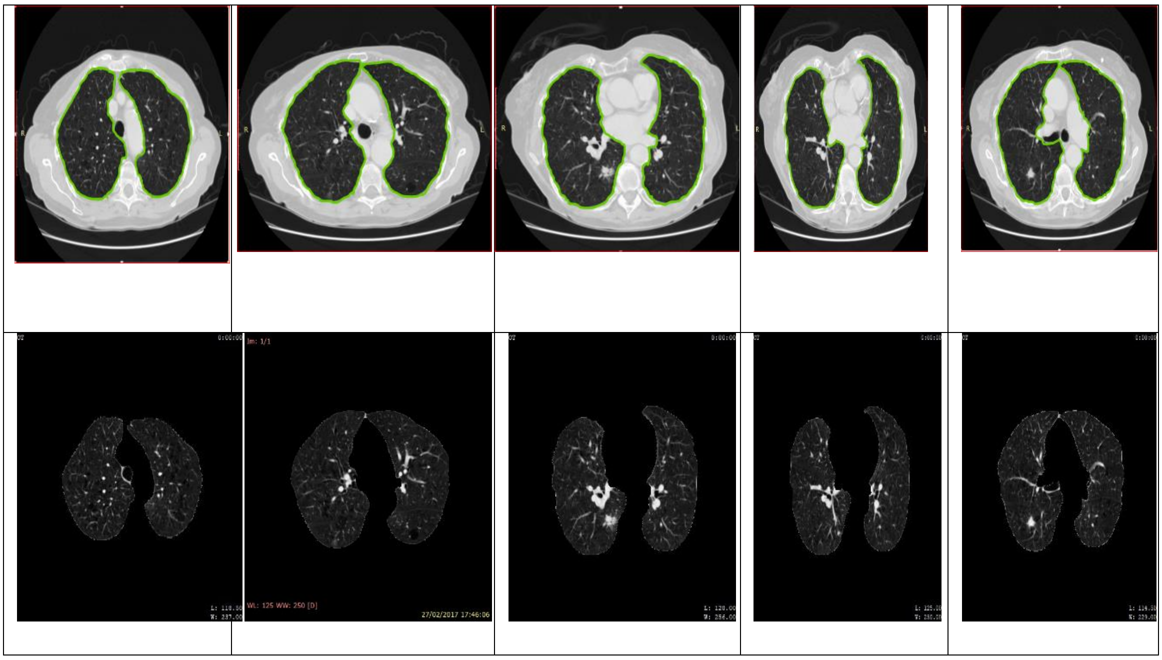


Figure 4-14: CT Lung Images Delineated the Boundaries with Green Colour by Radiologist and Segmented of Background by Fast Fuzzy C-Means Method

4.7 Post Processing

The main objective of this work is the precise detection of nodules. To achieve this aim, the nodules should be distinguished the vessels and normal tissue of the lung. This stage adopts more than one process to reach this target. The first step is a segmentation process of both vessels and nodules from the normal tissue of the lung, depending on the intensities of the image values. Also, some shape and texture features are applied to describe the resulting objects of the previous step. The features vector is the input to classifiers that classify the clusters into two classes (vessel and nodules) in a later step. For more description of the nodule, this processing develops 2D clustering algorithm for 3D clustering to measure the volume of the nodule.

4.7.1 Vessels and Nodules Segmentation Together

The pathology patterns often reflect the high intensities in the medical image. The various contrast of medical image pixels has become a search point whereby several algorithms focus intensity values used to separate the image regions. The significant challenge in the lung nodules detection is the similarity in intensity values between the nodule and the lung components. The main goal of this process is to exclude the normal tissue to reduce the research area of the nodules within the lung. Also, thus, the following processing concentrates on the objects that include the vessels and nodules. To detect the nodules easily, the selection of a segmentation method is based on its precision in describing nodule features (circularity, round, irregular shapes) to be classified correctly in the classification stage.

- Image Intensity -Based Clustering (IIBC)

The intensity characteristic is a vital factor in the various contrast regions segmentation of medical images. The proposed method offers active characteristics in clustering competing with the common methods. The algorithm presents a potential of representing the irregular shapes efficiently, dealing with the large data and the outliers in the best way. The image is divided into blocks to manage and process the large data faster. Each block has an equal size (m*m). The memory is divided to tree contains the block pixels and a heap the space to process these pixels. Also, The algorithm adopted a dynamic, empirical threshold is less than the middle value of the grey scale values to keep most intensity values which may consider a part of a tumour (Oseas et al. 2014). The purpose of the threshold is to separate image data into two classes. The class's values that are greater or equal to the threshold represent nodules and vessels while the background (normal tissue) of the lung is considered the class values that are less than the threshold value. There are two procedures for the proposed algorithm: the threshold conducts the separation procedure, and a merging is based on the connectivity and similarity principles. In the separation procedure, each pixel in a heap is checked by the threshold and classified to a class. For the merging procedure, the cluster that has similarity and overlap in the class will be merged into the single object. The merging is done if both the similarities and overlap are available together among clusters. The merging is performed into each block in the image and then among.

The algorithm potential to process the clusters without shifting position enables the irregular object patterns to be identified efficiently. The proposed method can cluster the irregular shape and manage the time by dividing the memory into two places: one for the data structure and the second for the data processing.

Algorithm

```
Name: Image Intensity-based Clustering (IIBC) algorithm
Input: $x_{i.j}$ where $x$ represents image points (pixel intensity), i,j are the pixel value index
Output: $CW_k$ white cluster of objects, where k is the number of white objects
        $CB_n$ black cluster of objects, where n is the number of black objects
Begin
    $W_0 = 0$, represents white object pixels, k=0
    $B_0 = 0$, represent black object pixels, n=0
  Divide the image to N-blocks (m*m), where N=number of blocks, m*m=block size
```

```
For each block
Divide the memory (tree, heap), where T=tree of block pixels, H=heap=pixels processing location of T
Order $x$ ascending in the heap
While size (heap) <>0 do
For each $x$ inside block
    if $x \geq R$ (R: threshold value)
    $w.cluster := x$
    if $W_0$ complete
    k=k+1
    insert $(W_0, w.cluster)$
    else
    $b.cluster := x$
    if $B_0$ complete
        $n = n + 1$
    insert$(B_0, b.cluster)$
    End
    delete$(heap.x)$
Call merging procedure
relocate$(T.W_0)$
relocate$(T.B_0)$
$CW_k = W_0$
$CB_n = B_0$
```

Pseudo-code of Merging

Procedure for merging $(u.v)$

Begin

```
$w.m$ = the connectivity of eight neighbours between $(u\&v)$
For each pixel in the corner of the cluster
Begin
If i=1
$\text{minDist} = dist(p.w,m)$
if $\text{dist (u.v)} \leq \text{minDist}$ (minDist represents the close distance between two objects that belong to the same class)
$\text{w} = \text{u} \cup \text{v}$
$\text{no. of objects} = \text{no. of objects} - 1$
End
$\text{return(w)}$
End
```

Figure 4-15 shows a flowchart for the proposed algorithm and includes two procedures for separating the data and merging to cluster the nodule and vessels into a class and normal tissue in another class.

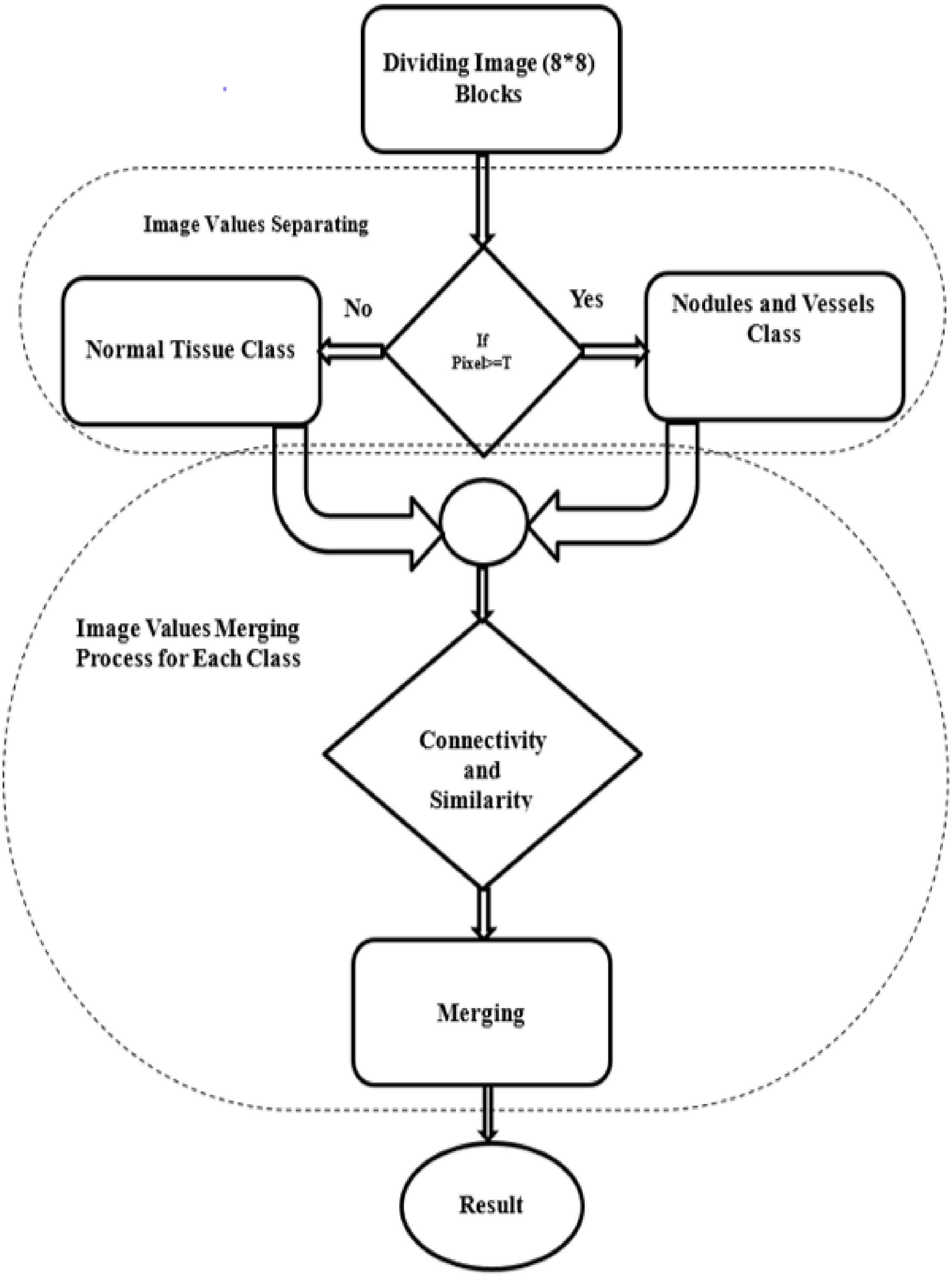
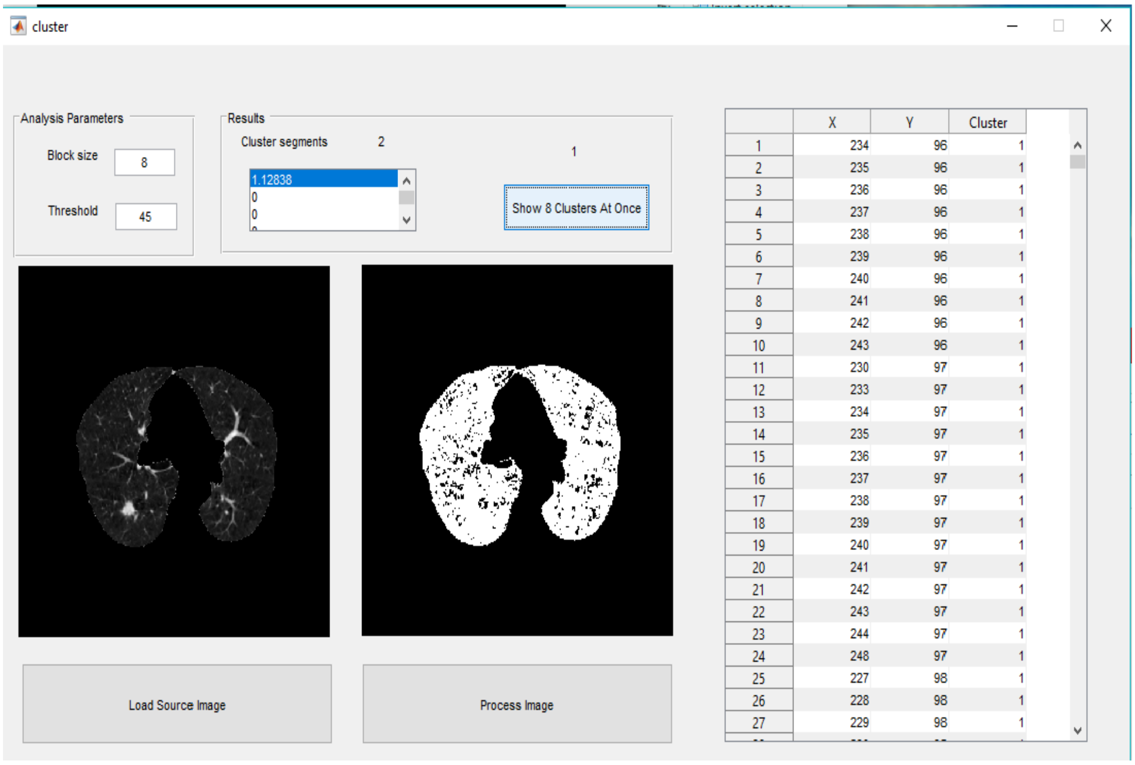

Dividing Image (8*8) Blocks
Image Values Separating
No
Yes
Normal Tissue Class
If Pixel>=T
Nodules and Vessels Class
Image Values Merging Process for Each Class
Connectivity and Similarity
Merging
Result

Figure 4-15 Flow Chart for Image Intensity- Based Clustering

4.8 Vessels and Nodules Segmentation Results

In our experiments, to exclude normal tissue and reduce the search area of the nodule, Image Intensity-based Clustering (IIBC) algorithm is applied to the segmented lungs in the pre-processing stage. This algorithm clusters the lungs image in objects according to two procedures; the first one separates the image values (grey scale) depending on empirical and dynamic threshold value into two classes. The second procedure merges the class' pixels that have a similarity and connectivity among them in objects belonging to each class. The resulting objects have high-intensity values that are upper the threshold and represent nodules and vessels together in a class. The second class includes the normal tissue that its values are less than the threshold.

The threshold values are chosen according to experiments based on the middle value of grey scale values approximately and the middle value of the optimal threshold that is identified through tests. The experiments demonstrate that the threshold 90 is the most suitable threshold clusters the majority of intensity values that reserve disease regions. And, it to be the optimal threshold of the algorithm, as shown in Figure 4-16 -b-. The 90 value clusters the image with the concern of objects that are attached to boundaries. Figure 4-16-a-,-c- show the segmented lungs image is clustered with other thresholds to identify the optimal in clustering. The 120 threshold misses parts of lung edges with the lack of objects compared with the 90 threshold clustering. The experiments display that no clustering with 45-threshold value of the image has been done where the threshold clusters two regions (lungs) have high density. For each clustered image, there is an index for the object pixels for each cluster, the selected threshold value and the block size shown in the GUI of Matlab 2014a.



Analysis Parameters

Block size 8

Threshold 45

Results

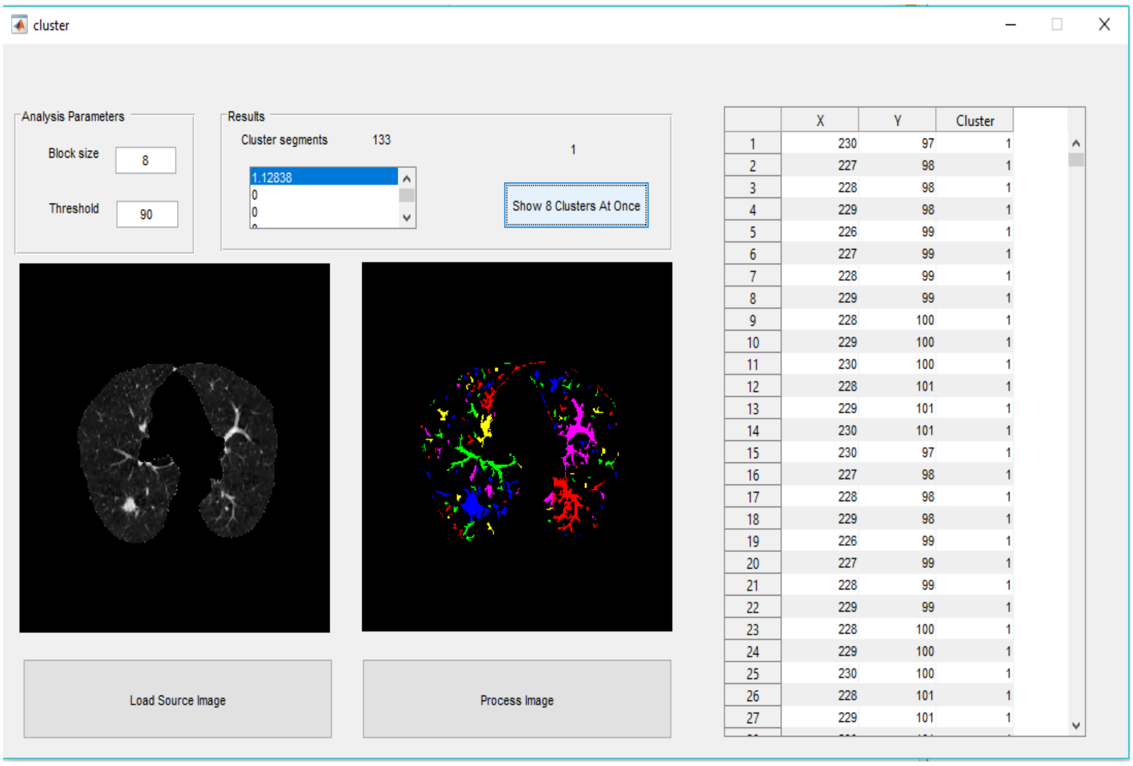
Cluster segments 2

1.12838
0
0

1

	X	Y	Cluster
1	234	96	1
2	235	96	1
3	236	96	1
4	237	96	1
5	238	96	1
6	239	96	1
7	240	96	1
8	241	96	1
9	242	96	1
10	243	96	1
11	230	97	1
12	233	97	1
13	234	97	1
14	235	97	1
15	236	97	1
16	237	97	1
17	238	97	1
18	239	97	1
19	240	97	1
20	241	97	1
21	242	97	1
22	243	97	1
23	244	97	1
24	248	97	1
25	227	98	1
26	228	98	1
27	229	98	1

-a-



Analysis Parameters

Block size 8

Threshold 90

Results

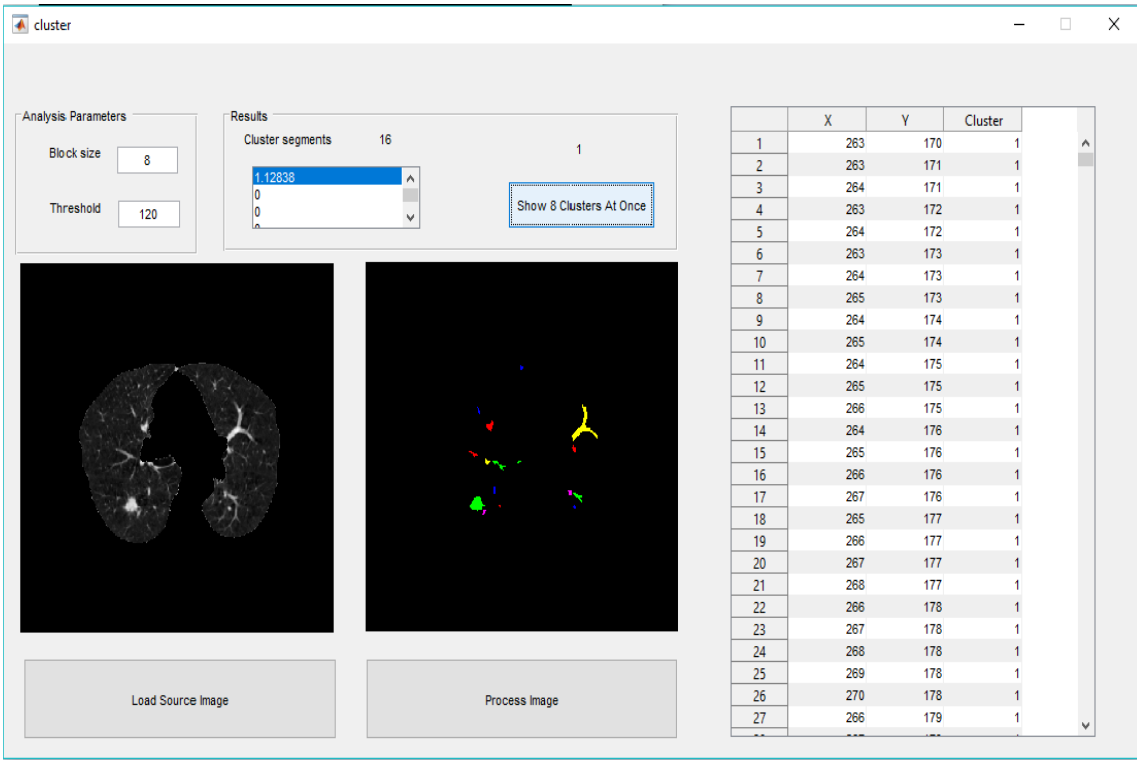
Cluster segments 133

1.12838
0
0

1

	X	Y	Cluster
1	230	97	1
2	227	98	1
3	228	98	1
4	229	98	1
5	226	99	1
6	227	99	1
7	228	99	1
8	229	99	1
9	228	100	1
10	229	100	1
11	230	100	1
12	228	101	1
13	229	101	1
14	230	101	1
15	230	97	1
16	227	98	1
17	228	98	1
18	229	98	1
19	226	99	1
20	227	99	1
21	228	99	1
22	229	99	1
23	228	100	1
24	229	100	1
25	230	100	1
26	228	101	1
27	229	101	1

-b-



Block size 8

Threshold 120

Cluster segments 16

1.12838
0
0

1

	X	Y	Cluster
1	263	170	1
2	263	171	1
3	264	171	1
4	263	172	1
5	264	172	1
6	263	173	1
7	264	173	1
8	265	173	1
9	264	174	1
10	265	174	1
11	264	175	1
12	265	175	1
13	266	175	1
14	264	176	1
15	265	176	1
16	266	176	1
17	267	176	1
18	265	177	1
19	266	177	1
20	267	177	1
21	268	177	1
22	266	178	1
23	267	178	1
24	268	178	1
25	269	178	1
26	270	178	1
27	266	179	1

-c-

Figure 4-16: a) Clustering with 45 Threshold; b) Clustering with 90 Threshold; c) Clustering with 120 Threshold

4.9 Summary

This chapter has involved various tasks relative to the research. The data collection on which the study based, comprehensive understanding of implied challenges of data and their characteristics have made. Also, the segmentation task of the lung has done as a part of pre-processing of the image. In addition, a qualitative and quantitative evaluation has made of lung segmentation corresponding the research methodology that was set previously. Furthermore, the modified clustering method is implemented to separate the vessels and nodule in the pre-processing stage. The method clusters the vessel and nodule within objects efficiently. The objective of nodule segmentation has achieved in this chapter, to be prepared for classification stage subsequently.

CHAPTER 5: NODULE CLASSIFICATION

5.1 Introduction

Classification is a process that involves sorting the image objects into separated classes based on vectors of features. It represents the final step of image analysis, consisting of sorting patterns into classes. The efficient classification relies on the variance of features, which belong to different classes, to raise the classifier's precision for discriminating the patterns. Common classifiers are used to accurately assign clusters, obtained from images in the previous stage, to nodules and vessels. The most discriminant descriptors increase classification performance through the precise description of object characteristics. An evaluation of the classification process has been carried out in this study to determine the optimal classifiers in terms of accuracy using appropriate metrics (G. Dougherty 2009),(Han et al., 2011) and (Russ 2016).

5.2 Image Features

To identify meaningful structures and find the desired objects within an image, some descriptors were obtained from the clusters to determine which image pixels belong together. These measurements are considered to be an essential step to interpret regions and recognise the groups of pixels that are called objects. In medical images, descriptors are important to the representation of objects that reflect pathological patterns in organ shape, defined by characteristics such as surface area, length, roundness, elongation, and compactness. Determining an appropriate measurement selection to reduce the amount of information attached to an object is a difficult task for medical applications. In this work, attention is given to the features that describe shape, size and texture of a nodule and which demonstrate the differences between vessels and nodules to achieve separation and detection. An active and segregating set of shape and texture parameters are extracted from objects within the image as an ideal group of features for the classification stage. The purpose of the group is to reduce the dimensionality of the problem, caused by redundant characteristics that influence classifier performance and data training time, as discussed in the next sections (Haralick et al., 1973) and (Ping et al., 2013).

5.2.1 Features Analysis

There are various approaches to extract textural and geometric descriptors based on the geometry; histogram and textures characteristics of objects form a medical image. These feature types were discussed in the literature review and background chapters in more details (Mingqiang et al., 2008), (Montero and Bribiesca 2009) and (X. Yang et al. 2012). The features extracted in this study introduced below.

5.2.1.1 The Shape and Geometric Descriptors

- Area

This parameter counts the pixels of which the object consists. The area parameter measures the object size and provides information on whether the object is a real or noise using Eq.5-1.This parameter is not capable of producing information about the object's length; therefore, the perimeter parameter can report the length information of the object.

$$A = \sum f(x.y). in\ 2D\ space \qquad 5\text{-}1$$

Where x, y represent the coordinates of pixel intensity value in the image.

- Perimeter

The perimeter measures the length of an object boundary. The boundary is the collection of all edge pixels, which are all the object pixels that do not have an identical neighbour.

$$Perimeter(L) = n_0 + \sqrt{2}n_e \qquad 5\text{-}2$$

Here $n_0 . n_e$ are the odd and even code numbers of the chain code, respectively.

Chain code represents pixels form a grid in x and y-direction to assign an orientation to the segments connecting every pair of pixels. The segments specify length and direction based on N-4 or N-8 connectivity (Mingqiang et al., 2008).

- Diameter

In centroid calculation of objects, the diameter parameter measures m_{pq} where p,-order origin moment is calculated using Eq.5-3.

$$a = 2 \times \left[[2(\mu_{20} + \mu_{02} - \sqrt{(\mu_{20} - \mu_{02})^2 + 4{\mu_{11}}^2})]/\mu_{00} \right]^{1/2} \qquad 5\text{-}3$$

$$\mu_{pq} = \sum_x \sum_y x^p y^q \, f(x, y)$$

$$\mu_{00} = Area$$

- Circularity

  This geometrical parameter reflects the roundness of a 2D object. It represents a relation between structure's area and perimeter by using Eq. 5-4

$$c = \frac{4 * \pi * A}{l^2} \qquad 5\text{-}4$$

Here l is the perimeter, and A the area of the object. The factor 4π ensures that c equals 1 (the lowest possible value) for a circle shape.

- Compactness

Compactness gives the low values for very compact objects, while high values demonstrate that the objects are less compact. It is the reciprocal of circularity, as seen in Eq. 5-5.

$$F = \frac{1}{c} \qquad 5\text{-}5$$

- Ellipticity

This parameter uses Ellipses in two orientations of the object and is calculated by the ratio of the longest chord and shortest chord of an object using Eq.5-6.

$$e = a/b \qquad 5\text{-}6$$

a, b are the long axis and short axis, respectively, in a 2D space

- Slenderness

A parameter to measure the elongation of an object. It represents the ratio of major axis to minor axis of the object using Eq.5-7.

$$\mathrm{S} = \min(\mathrm{W.H})\ /\max(\mathrm{W.H}) \qquad 5\text{-}7$$

H, W are the width and height of the potential nodule object, respectively, both are considered in the slenderness and rectangle degree parameters.

- Rectangle Degree

Also called the extent or rectangularity of an object, it is defined as the ratio of the object area to the product of width and height of the object using Eq.5-8. A perfect rectangle would have a degree of 1.

$$\mathrm{R} = \mathrm{A}/(\mathrm{W} \times \mathrm{H}).\ \text{in 2D space} \qquad 5\text{-}8$$

- Concavity Ratio

This parameter reflects the concavity of the object boundary and is the ratio of the difference between the original region and a convex hull and concave area using Eq.5-9

$$\text{Concavity Ratio (E)}=\frac{\mathrm{S_e}}{\mathrm{S}} \qquad 5\text{-}9$$

S is the area of a concave region, $\mathrm{S_e}$ represents the difference between the original region and its convex hull.

5.2.1.2 Texture Descriptors

In this work, texture parameters are obtained from the co-occurrence matrix that is a statistic derived from a second-order histogram that describes the relationship between sets of two pixels in the object region. This matrix contains some rows and columns of grey scale levels that produce some features (Haralick et al., 1973), (Ying et al., 2011) and (X. Yang et al. 2012). These are defined as follows below.

- Contrast

The contrast descriptor is a measure that describes the intensity contrast for each a pixel and its neighbour in the image. It is defined using Eq. 5.10 (Haralick et al., 1973).

$$\text{Contrast(Con)} = \sum_{n=0}^{N_{g-1}} n^2 \{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j)\}, |i-j| = n \qquad 5\text{-}10$$

Where N_g represents the number of grey scales in image, and $p(i,j)$ represents an element at the cell coordinates i and j in co-occurrence matrix.

- Correlation

A statistical coefficient that demonstrates the strength of the relationship between the real and predicted values. The coefficient increases when trends in the predicted values track real trends. It also depends on standard deviation and means values (μ_x, μ_y, σ_x and σ_y) for P_x and P_y, in ROI.

$$Correlation = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} P(i,j) \frac{(i-\mu_x)(j-\mu_y)}{\sigma_x \, \sigma_y} \qquad 5\text{-}11$$

$$P_x(j) = \sum_{j=0}^{N_g-1} P(i,j)$$

$$P_y(i) = \sum_{i=0}^{N_g-1} P(i,j)$$

$$\mu_x = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} i.\mathrm{p(i,j)}$$

$$\mu_y = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} j.p(i,j)$$

$$\sigma_x = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} (i-\mu_x)^2.p(i,j)$$

$$\sigma_y = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} (j-\mu_y)^2.p(i,j)$$

- Entropy

It is a scale of the smooth region of interest (ROIs) by using low values and a measure of randomness.

$$\mathrm{Entropy(ENT)} = -\sum_{i=0}^{N_g-1}\sum_{j=0}^{N_g-1} P(i,j)\log P(i,j) \qquad 5\text{-}12$$

- Energy

A descriptor measures the pixel value distribution together with the grey-level range. Images with higher grey levels have higher energy than those with lower grey levels.

$$\mathrm{Energy} = \sum_{i=0}^{N_g-1}\sum_{j=0}^{N_g-1} P(i,j)^2 \qquad 5\text{-}13$$

- Homogeneity

It is the quality or state of being homogeneous.

$$\mathrm{Homogeneity} = \sum_{i=0}^{N_g-1}\sum_{j=0}^{N_g-1}\{P(i,j)\}^2 \qquad 5\text{-}14$$

- Mean

This parameter measures the average of intensity values that belong to the region of interest (ROI).

$$Mean(\mu) = \frac{1}{N}\sum_{i,j} p(i,j) \qquad 5\text{-}15$$

Where $p(i,j)$ represents the pixels intensity position in the image, N determines the total number of pixels in the image.

- Skewness

It describes the degree of asymmetry in the pixel distribution around its mean. Skewness produces a number that describes only the shape of the distribution.

$$Skewness(S) = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left[\frac{p(i,j)-\mu}{\sigma}\right]^3 \qquad 5\text{-}16$$

- Kurtosis

It measures the sharpness or flatness of a distribution relative to a normal distribution with the same mean and standard deviation.

$$Kurtosis(K) = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left[\frac{p(i,j)-\mu}{\sigma}\right]^{4} \qquad 5\text{-}17$$

- Variance

  It considers the changes of the grey value in the ROI.

  $$Variance = \sum_{i,j}(i-\mu)^{2}\, p(i,j) \qquad 5\text{-}18$$

- Inverse Different Moment (IDM)

  It raises the low contrast of ROIs because of the dependence on $(i-j)^{2}$.

  $$\text{Inverse Different Moment(IDM)} = \sum_{\mathrm{i,j}}\frac{\mathrm{p(i,j)}}{1+(\mathrm{i-j})^{2}} \qquad 5\text{-}19$$

5.3 Classification phase

The study objective is to achieve the highest possible accuracy in pulmonary nodule detection from CT lung scan images. This requires identifying all tumours greater than 3mm with minimal errors and early detection. In addition, the false positive rate should be reasonably low. Recently, for the various characteristics of the medical image and the best classification performance, some classifiers have received multiplied attention due to their high performance. The most common classifier models are the Support Vector Machine(SVM), Logistic Regression(LR), Bayesian and Multilayer Perceptron (MLP) and Naive Bayes, all used to predict cancers (Ozekes and Osman 2010), (H. Chen et al. 2012). In this study, these five common classifiers are used to accurately classify the clusters into nodules and vessels. To evaluate and analyse the classifiers' performance, the K-fold cross-validation is used.

5.4 Validation Phase

The evaluation of the classification task in nodule detection accuracy is done by cross-validation (Witten et al. 2016). It is essentially used where the goal is to predict and to estimate how a classification is accurately performed in practice. Also, it describes the statistical analysis of the classification results which assess the classifiers' accuracy using a statistical approach. In this study, the cross validation is carried out using a two-dimensional matrix. The rows and columns in the matrix represent, respectively, the class instance number and the descriptor for each class (Han et al., 2011). K-fold cross-validation method divides the image data to N separate sets where the values at N-1 are used for training while the Nth set is used for testing. The merit of this model is to perform for both validation and training over iterated random sub-sampling. Each observation is selected for validation precisely once. In this work, two classes are defined; the first class denotes vessels and the second class refers to nodules. Two different fold of cross-validation models validates the five classifiers. For both two-fold, four classifiers achieved a high level of accuracy. Each data array for validation contains 19 descriptor values applied on 400 instances (400*19). The instances consist of 180 candidate nodules (class b) and 220 vessel objects (class a). Firstly, 4-fold cross validation is employed on the array for training and testing with the five classifiers. In this fold, the data is divided into 4 groups; 3 groups are used for classifier training while the left out group is used for testing. The cross-validation method is repeated with 10-fold dividing the data into 9 groups for training and the left out group for testing (Witten et al. 2016). The sensitivity, false positive rate and accuracy are then obtained for each trained classifier. The posterior probabilities are achieved for every K-fold of cross-validation, which are obtained from the assigned groups for training in the fold and per classifier. The process of K-fold cross-validation is conducted thus; the training groups cover the left out-group to leave other groups for testing purposes and so on. The process stops when the posterior probabilities are calculated for all the candidate nodules (L. Ma et al. 2015).

In this study, the five classifiers are compared in terms of performance to identify the optimal. The classifier functions and parameters significantly contribute to improving classification performance and are selected according to a previous study(Ozekes and Osman 2010). For SVM, three basic functions or kernels are used to get its approximate accuracies, such as Polynomial and Radial basis function (RBF) and PUK (Abakar and Yu 2014). The hyperparameters values of SVM are Epsilon= 1.0E-12, C=1.0, Tolerance parameter= 0.001 for POLY, RBF and PUK. For MLP, there are two nodes for the output layer corresponding to two classes (nodules, vessels), Learning Rate =0.1 to 0.3 with the same results, and input and hidden layers contain 20 and 21 nodes respectively. The logistic regression has a Ridge parameter of 1.0E-8. The dataset used in this work consists of 180 nodules with size greater than 3mm, which were initially taken from 40 patients with further processing for each slice.

5.5 The Evaluation Phase

The goal of the classification is discriminating the nodules from other components efficiently into the clustered image. There are good tools that are able to assess the performance of the classifiers and to identify the robust classifier that achieves the highest accuracy. First, the confusion matrix is popular and presents statistical parameters that are calculated for each slice for evaluation purposes. Second, other metrics for further statistical measurements, which evaluate the accuracy and precision of classifiers, have been used in this study.

5.5.1 Confusion Matrix

The confusion matrix is a good tool to evaluate the performance of classifiers (Xu et al., 1992), (L. Ma et al. 2014). It deals with four terms known for evaluation purposes: TP (true positive), the abnormal case which is correctly classified as abnormal; FP (false positive), the normal case which is incorrectly classified as abnormal; TN (true negative); and FN (false negative). TN is the normal case that is correctly classified as normal, and FN is the abnormal case that is incorrectly classified as normal.

Confusion matrix parameters are considered in the statistical computations to analyse and quantitatively evaluate a classifier's performance in terms of accuracy, sensitivity and specificity. The overall accuracy demonstrates the efficiency of a classifier; however, it could be misleading with imbalanced data. The total accuracy is calculated using Eq. 5-20. Sensitivity provides the information about the correct classification rate of abnormal cases, while the percentage of normal cases is estimated by specificity in the medical image using equations 5-21 and 5-22 respectively.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} * 100\% \qquad \text{5-20}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} * 100\% \qquad \text{5-21}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} * 100\% \qquad \text{5-22}$$

5.5.2 Other Metrics

Other metrics related to accuracy, sensitivity and data balance are used to evaluate the classification performance in this study. These are the rates of True positives (TP) also known as Recall, True Negatives, False Positives (FP), False Negatives, Precision, F-measure, Matthews Correlation Coefficient (MCC) and the Area under curve (Sun et al. 2013),(Al-fahoum, Jaber, and Al-jarrah 2014). They require parameters from the confusion matrix for their calculation. The formulas for these metrics are given below:

$$\text{True Positive Rate(TPR) or Recall} = \frac{\text{TP}}{\text{P}} = \frac{\text{TP}}{\text{TP+FN}} \tag{5-23}$$

$$\text{True Negative Rate(TNR)} = \frac{\text{TN}}{\text{N}} = \frac{\text{TN}}{\text{TN+FP}} \tag{5-24}$$

$$\text{False Positive Rate(FPR)} = \frac{\text{FP}}{\text{FP+TN}} = 1 - \text{TNR} \tag{5-25}$$

$$\text{False Negative Rate(FNR)} = \frac{\text{FN}}{\text{FN+TP}} = 1 - \text{TPR} \tag{5-26}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{5-27}$$

$$\text{F}_{-\text{measure}} = 2 * \frac{\text{Precision}*\text{Recall}}{\text{Precision+Recall}} \tag{5-28}$$

$$\text{Matthews Correlation Coefficient(MCC)} = \frac{\text{TP}*\text{TN}-\text{FP}*\text{FN}}{\sqrt{(\text{TP+FP})(\text{TP+FN})(\text{TN+FP})(\text{TN+FN})}} \tag{5-29}$$

$$\text{Area Under Curve(AUC)} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP+FN}} + \frac{\text{TN}}{\text{TN+FP}}\right) \tag{5-30}$$

5.6 Results and Discussions

Both 4, 10-fold cross validation have been used to evaluate the classifiers' potential to distinguish the nodule and vessels. The obtained accuracy results are as follows: the Bayesian, logistic regression (LR), multilayer perceptron (MLP), and support vector machine with PUK kernel (SVM-PUK) achieved the highest value of 98%. This was followed by accuracies of 76% and 76.25%, which are obtained by Naïve Bayes (NB) that has the lowest accuracy rate among all other classifiers in both 4 and 10-fold cross-validation method.

The accuracy results are shown in Figure 5-1, Figure 5-2, Table 5-1 and Table 5-2 all exhibit that the best accuracy rate reached 98% for the four classifiers. The lowest accuracy for the Naïve Bayes occurred because of reasons that have been discussed in details in chapter 3. Figure 5-4 shows the Naïve Bayes decreasing to half in the classification of the nodule patterns and has the worst classification rate. Furthermore, Figure 5-3 shows an error in the nodule classification rate sets of 0 and thus, the sensitivity rate is 100% percent as calculated by Eq. 5-21.

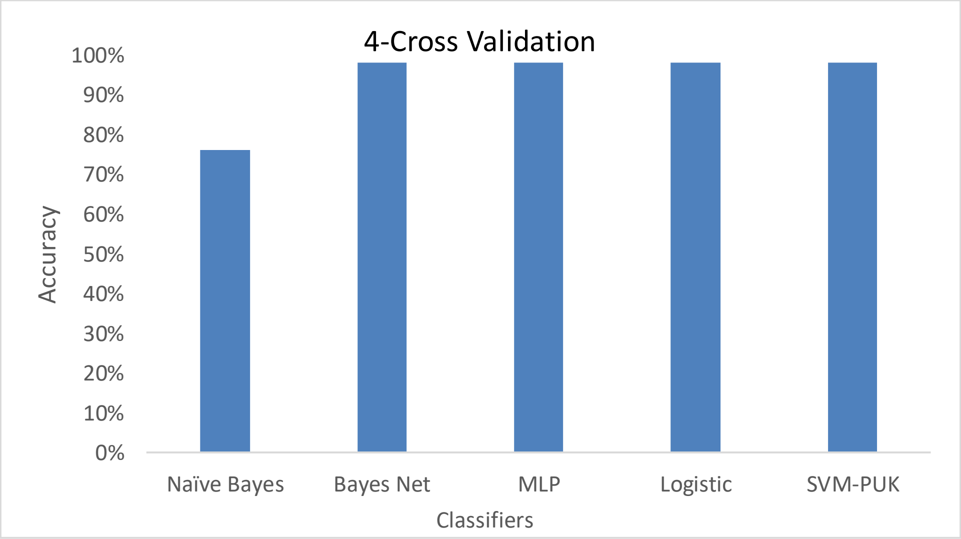


Figure 5-1: Accuracy Rates of Five Classifiers Through 4-Cross-Validation

Table 5-1: Accuracy Values for Five Classifiers Through 4-Cross-Validation

Classifiers	Naïve Bayes	Bayesian	MLP	Logistic	SVM-PUK
Accuracy	76%	98%	98%	98%	98%

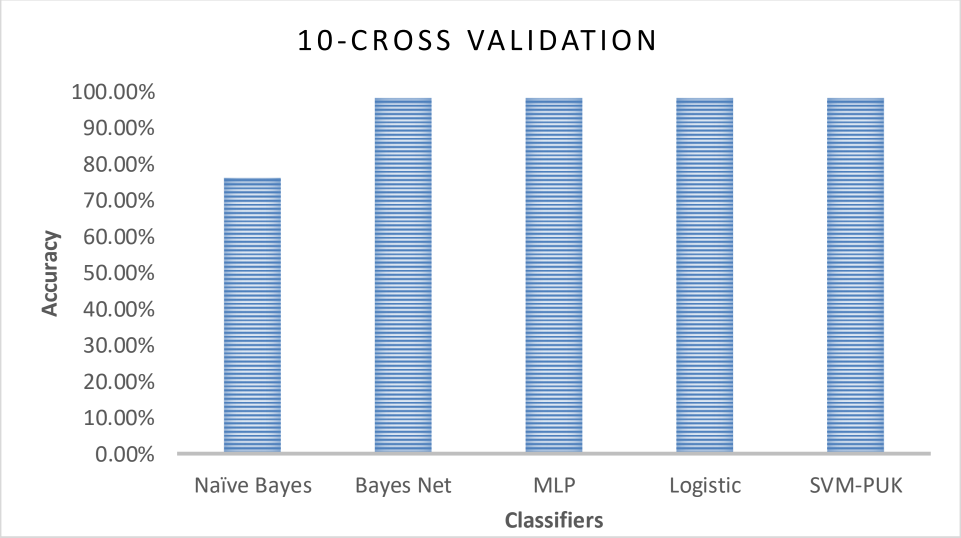


Figure 5-2: Accuracy Rates of Five Classifiers Through 10-Cross-Validation

Table 5-2 Accuracy Values for Five Classifiers through 10-Cross-Validation

Classifiers	Naïve Bayes	Bayesian	MLP	Logistic	SVM-PUK
Accuracy	76.25%	98%	98%	98%	98%

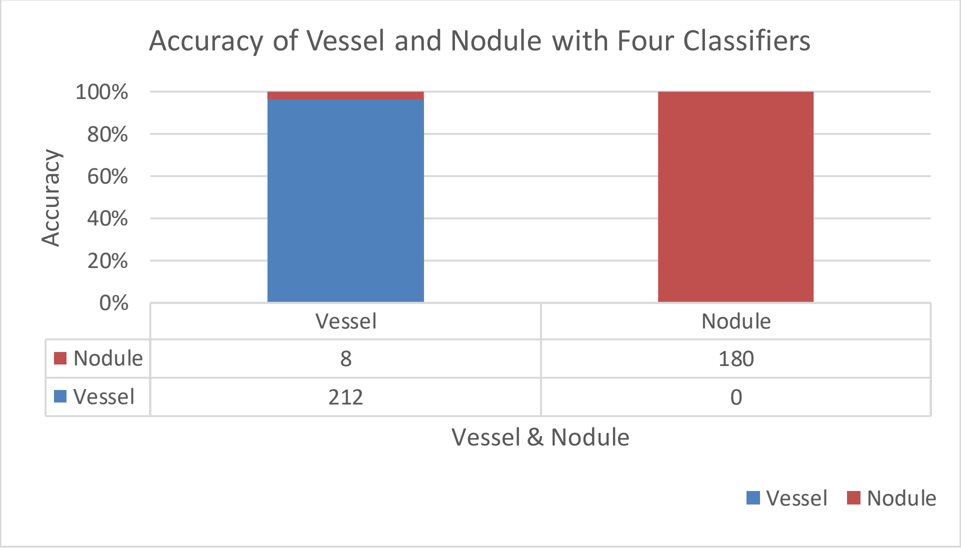


Figure 5-3: Accuracy in Classifying Vessels and Nodules of Four Classifiers

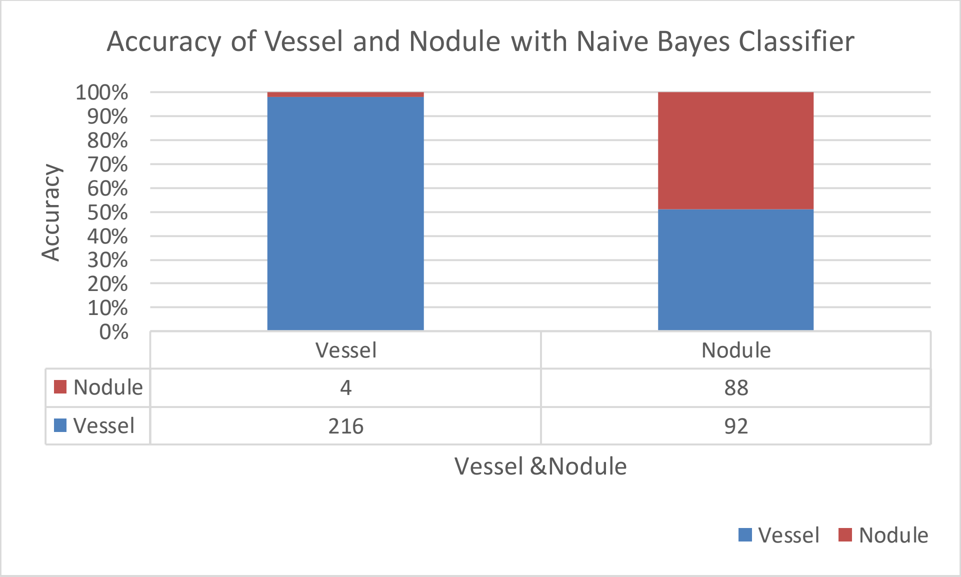


Figure 5-4: Accuracy in Classifying Vessels and Nodules of Naïve Bayes Classifier

The SVM classification accuracy has various values according to its kernels (POLY, RBF and PUK) in both 4, 10 fold. The SVM accuracy dramatically decreases with POLY, RBF kernels, as shown in Figure 5-5, Figure 5-6, Figure 5-7 and Figure 5-8, however, it increases to reach 98% accuracy with PUK function Figure 5-9 showing the difference among SVM Kernels.

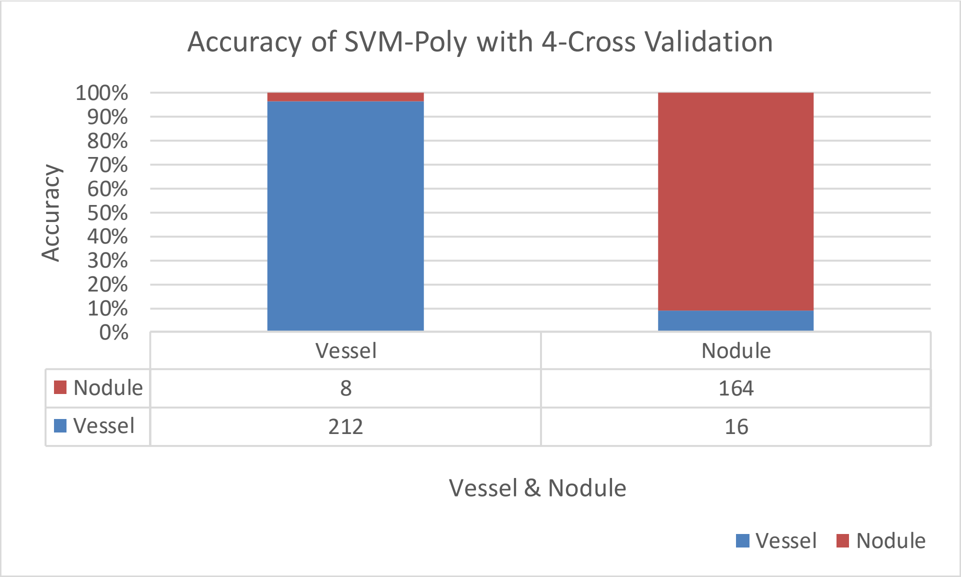


Figure 5-5: Accuracy in Classifying Vessels and Nodules of SVM-poly 4-Cross-Validation

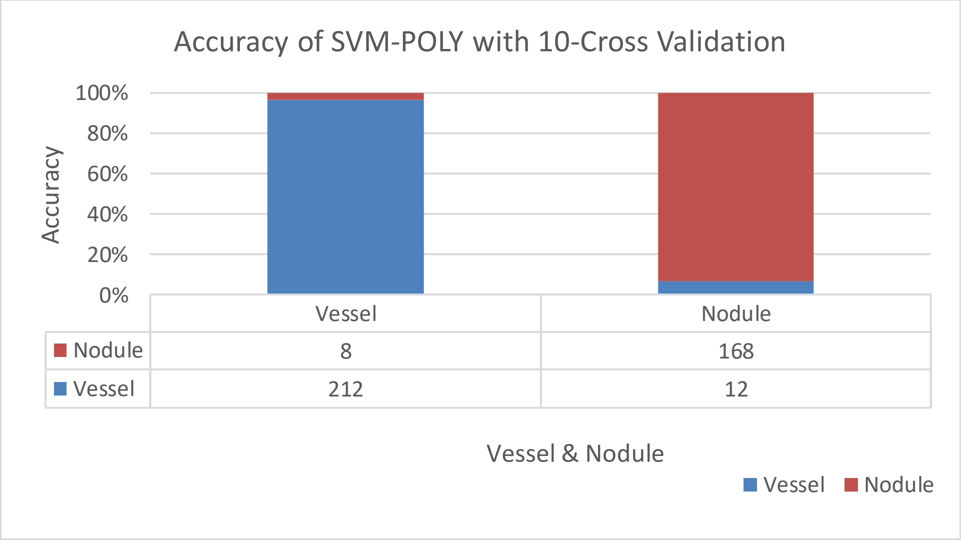


Figure 5-6: Accuracy in Classifying Vessels and Nodules of SVM-POLY 10-Cross-Validation

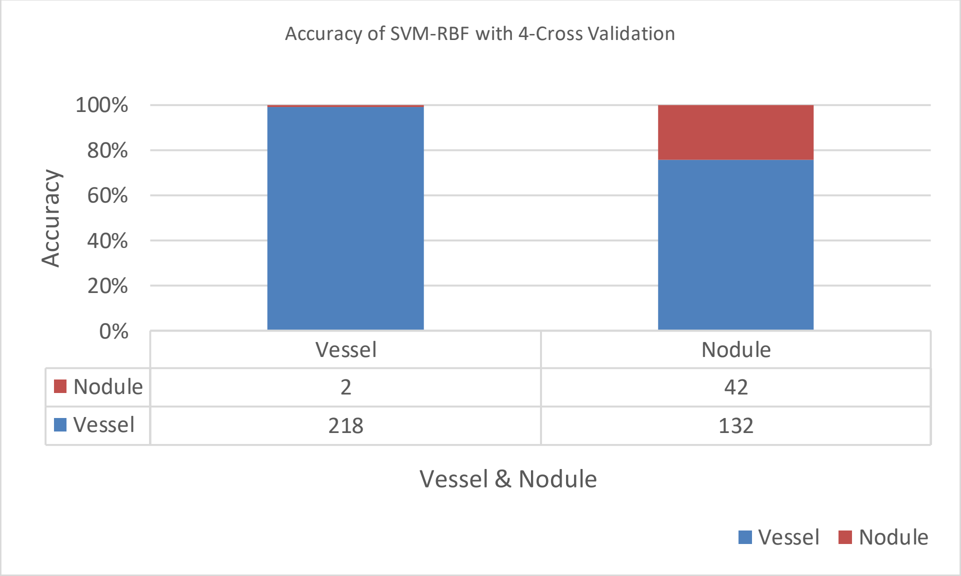


Figure 5-7: Accuracy in Classifying Vessels and Nodules of SVM-RBF in 4-Cross-Validation

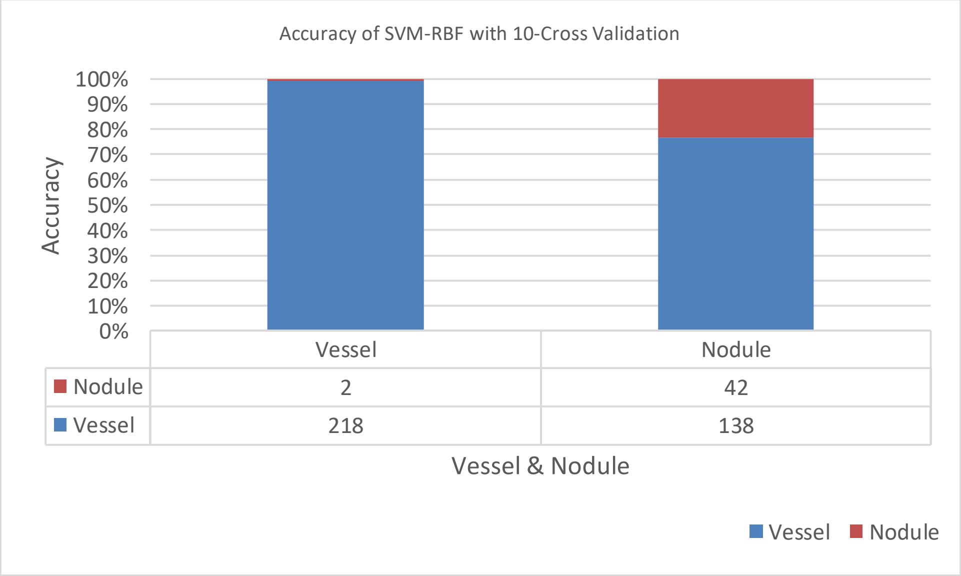


Figure 5-8: Accuracy in Classifying Vessels and Nodules of SVM-RBF in 10-Cross-Validation

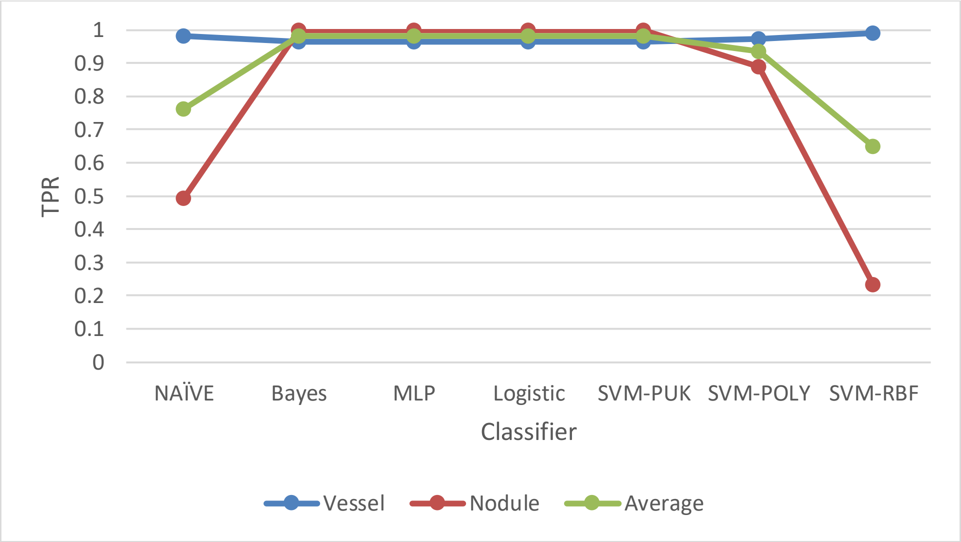
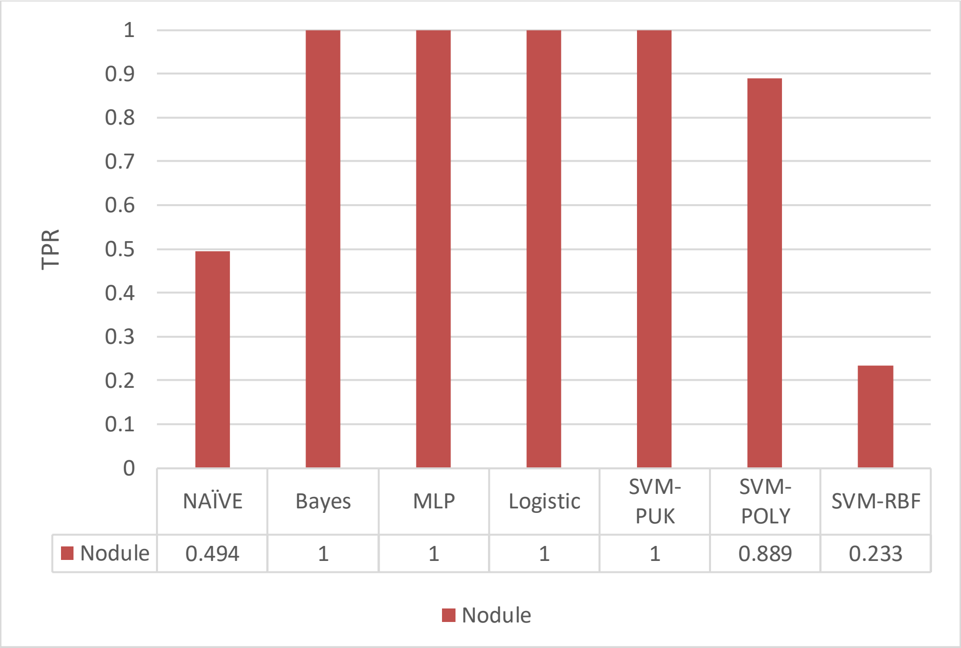


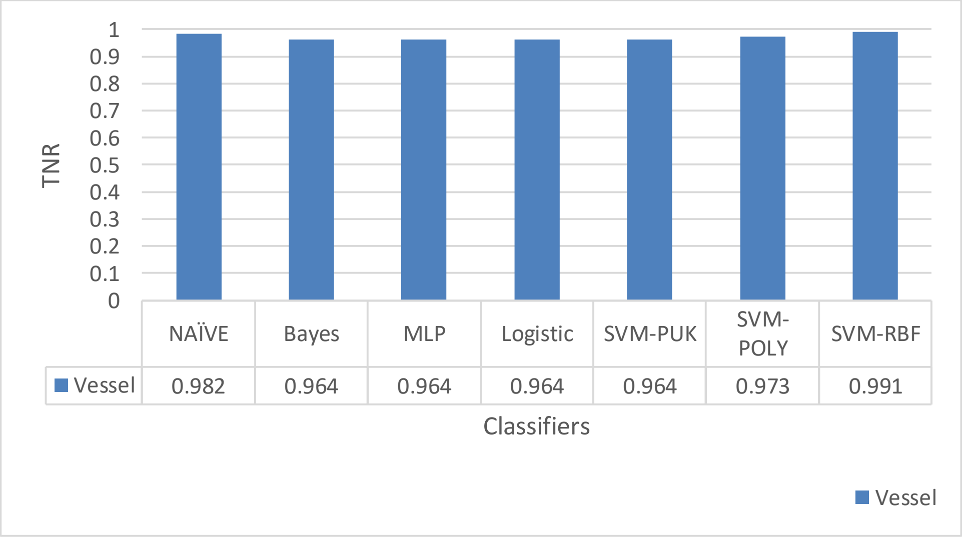
Figure 5-9: Accuracy in Classifying Vessels and Nodules of SVM-PUK in 10-Cross-Validation

To evaluate the classification performance, Figure 5-10 shows the **TPR** or **Recall** measure that represents the sensitivity of the five classifiers in detecting the TPR of the nodules. In this study, the TPR or Recall rate of nodule classification of four classifiers (Bayesian, MLP, LR, and SVM-PUK) is 1 while it is 0.494, 0.889, and 0.233 for Naïve, SVM-POLY and SVM-RBF respectively. The sensitivity rate is calculated as 100% percent using Eq.5-21 for the four classifiers that prove a very accurate classification of the nodule. That means that no nodules in the dataset were missed. The TNR decreases to 0.964 in vessels detection for the first four classifiers while the rate fluctuates for other classifiers depending on their accuracy using Eq.5-23 and Eq.5-24. Also, the TNR for Naive, SVM-POLY and SVM-RBF are 0.982, 0.973 and 0.991 respectively, as demonstrated in Figure 5-11.



	NAÏVE	Bayes	MLP	Logistic	SVM-PUK	SVM-POLY	SVM-RBF
🟥 Nodule	0.494	1	1	1	1	0.889	0.233

Figure 5-10: Recall of Nodules for all Classifiers



	NAÏVE	Bayes	MLP	Logistic	SVM-PUK	SVM-POLY	SVM-RBF
🟦 Vessel	0.982	0.964	0.964	0.964	0.964	0.973	0.991

Figure 5-11: TNR Vessels for all Classifiers

FPR indicates that vessels are mistakenly classified as nodules, while FNR represents the rate of nodules also incorrectly classified as vessels. In this work, the FNR for vessels is 0 which means no nodules are mistaken classified for vessels, as shown in Figure 5-12, and the FPR for nodules is 0.036 by four classifiers as shown in Figure 5-13. These rates are satisfactory, as there is no misclassification of the nodules for the four classifiers (Bayesian, MLP, LR, and SVM-PUK). Other classifiers have higher misclassification of nodules, which passively affect FPR and FNR that are computed using Eq. 5-25 and Eq.5-26.

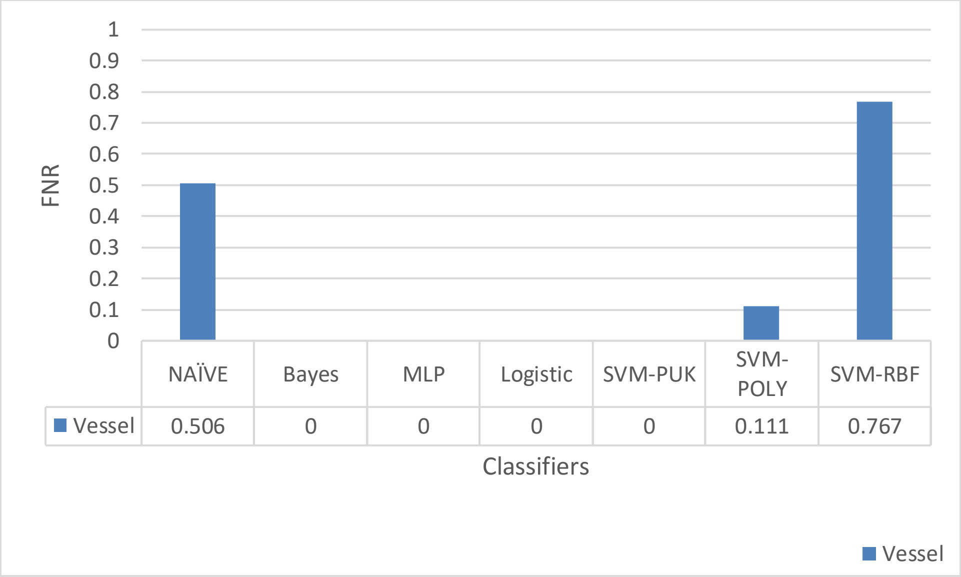


Figure 5-12: FNR of Vessels for all Classifiers

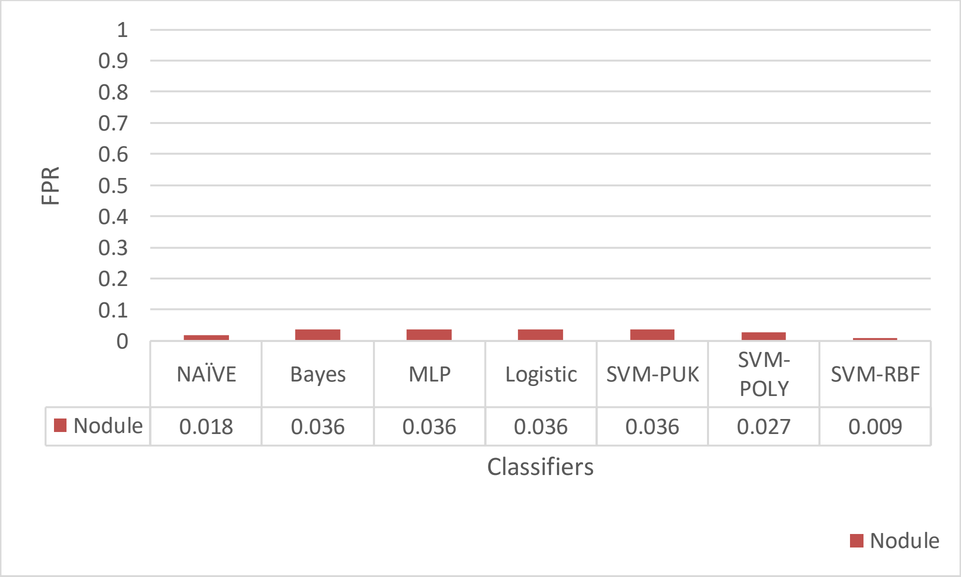
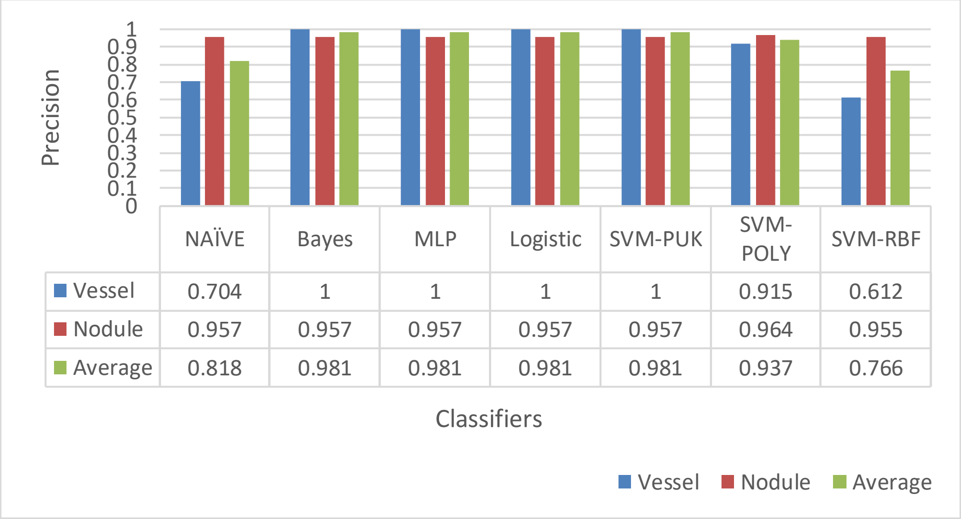


Figure 5-13: FPR of Nodules for all Classifiers

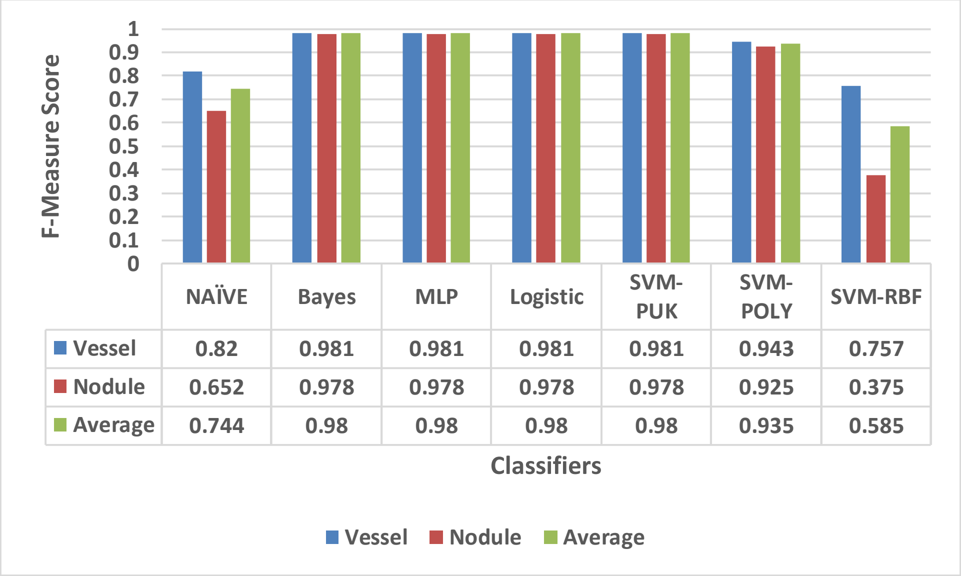
Precision metric is the percentage of instances assigned to the correct class. It does not determine the number of instances that are classified into the wrong class. The precision of each classifier is shown in Figure 5-14. A perfect score of 1 was found for four classifiers (Bayesian, MLP, LR, and SVM-PUK) for vessels meaning no nodules were mistakenly classified as vessels. The score for nodules is just below 1 for all classifiers, indicating that some vessels were mistakenly classified as nodules in all cases using Eq.5-27.



	NAÏVE	Bayes	MLP	Logistic	SVM-PUK	SVM-POLY	SVM-RBF
🟦 Vessel	0.704	1	1	1	1	0.915	0.612
🟥 Nodule	0.957	0.957	0.957	0.957	0.957	0.964	0.955
🟩 Average	0.818	0.981	0.981	0.981	0.981	0.937	0.766

Figure 5-14: Precision Measure of Nodules and Vessels for all Classifiers and Average of both

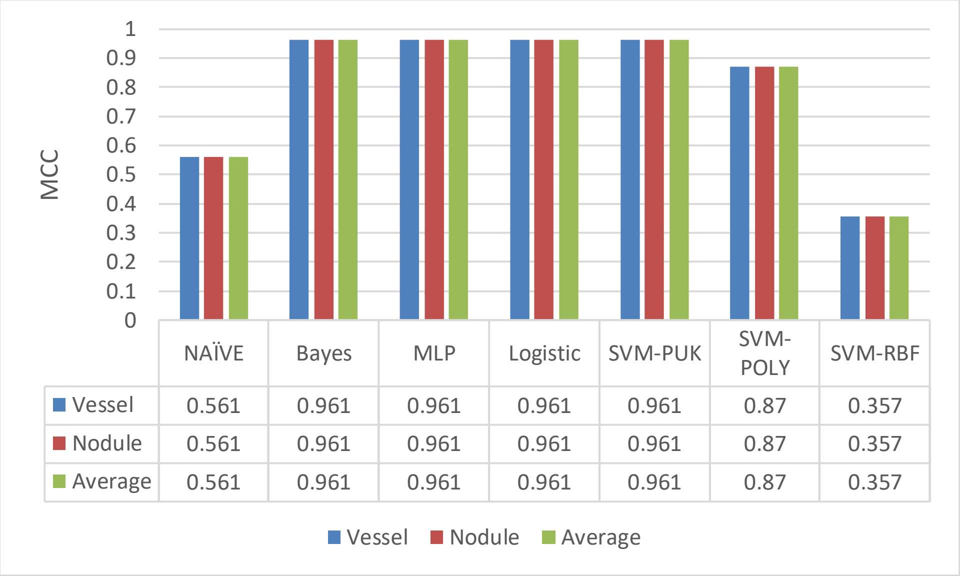
F-measure (F-score) is a metric for evaluating the classification accuracy. It is considered as a type of rates averages in math and is computed for a set of classes as statistical analysis for the binary classification. To measure the F-score, both recall and precision rates are considered in the test as in Eq.5-28. F-measure is between 0 and 1 indicating the perfect score. The F-score for both nodules and vessels tends to be 1 for four classifiers that demonstrate the best performance in classification, as shown in Figure 5-15. For Naïve, SVM-POLY and SVM-RBF, the F-score decreases to 0.652, 0.925 and 0.375 respectively in nodule classification. For vessel classification, the F-score is 0.82, 0.943, 0.757, scores for Naïve, SVM-POLY and SVM-RBF.



	NAÏVE	Bayes	MLP	Logistic	SVM-PUK	SVM-POLY	SVM-RBF
🟦 Vessel	0.82	0.981	0.981	0.981	0.981	0.943	0.757
🟥 Nodule	0.652	0.978	0.978	0.978	0.978	0.925	0.375
🟩 Average	0.744	0.98	0.98	0.98	0.98	0.935	0.585

Figure 5-15: F- Measure Score for all Classifiers

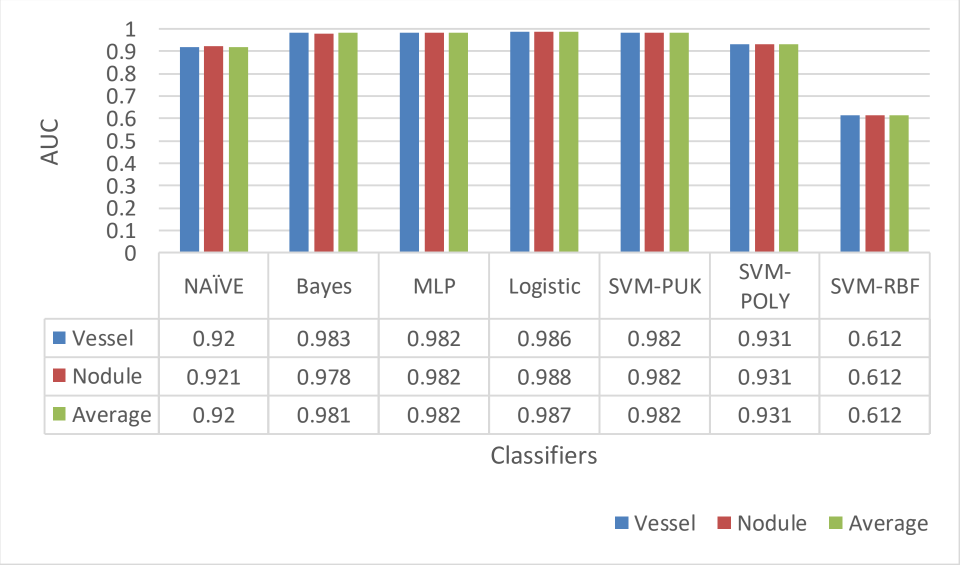
Matthew Correlation Coefficient (MCC) is a metric used in machine learning to measure the quality of binary classification. It is considered a measure that is less affected by imbalanced data has different class sizes and represents a correlation coefficient of observation and prediction in binary classification. This metric presents a range between 1and -1 and a perfect prediction gives a value of 1. The values that are greater than zero better than the values that equal to or lower than zero, and -1 indicate a conflict between observation and prediction in Eq. 5-29. For four classifiers, the coefficient value is 0.961 while 0.561, 0.87 and 0.357 are the values for Naïve, SVM-POLY and SVM-RBF respectively as shown in Figure 5-16.



	NAÏVE	Bayes	MLP	Logistic	SVM-PUK	SVM-POLY	SVM-RBF
🟦 Vessel	0.561	0.961	0.961	0.961	0.961	0.87	0.357
🟥 Nodule	0.561	0.961	0.961	0.961	0.961	0.87	0.357
🟩 Average	0.561	0.961	0.961	0.961	0.961	0.87	0.357

Figure 5-16: MCC Measure for all Classifiers

Area Under Curve (AUC) is a numerical method used to evaluate the performance of the binary classifiers and to identify the optimal classifier. Each point of the ROC curve distinguishes a probabilistic classifier where each point of this curve corresponds to a discrete classifier. Most of the four classifiers tend towards a value of 1 that is the perfect score of the AUC metric for nodules, vessels and their averages, and computed by Eq.5-30. This value shows which classifier has the optimal performance, as shown in Figure 5-17; four classifiers (Bayes, MLP, LR and SVM-PUK) displayed the best performance in this study.



	NAÏVE	Bayes	MLP	Logistic	SVM-PUK	SVM-POLY	SVM-RBF
🟦 Vessel	0.92	0.983	0.982	0.986	0.982	0.931	0.612
🟥 Nodule	0.921	0.978	0.982	0.988	0.982	0.931	0.612
🟩 Average	0.92	0.981	0.982	0.987	0.982	0.931	0.612

Figure 5-17: AUC Measure for all Classifiers

Receiver Operating Characteristic (ROC)

A ROC curve is a visualization approach for exhibiting the trade-off between the TPR and FPR of a classifier. The ROC curve is used to compare the performance of several classifiers for many experiments. In a ROC curve, the horizontal axis represents the FPR while the vertical axis represents TPR. A good classifier should be set as close as possible to the upper left corner of the ROC curve figure, while a classifier that displays random behaviours should locate along the main diagonal, connecting the points 0,0 and 1,1 for TPR and FPR respectively.

In this study, the ROC curve is employed to evaluate five classifiers by applying eight experiments for each classifier to the same data set. These experiments are three, four, five, six, seven, eight, nine and ten folds-cross validation. Figure 5-18 illustrates the best classifiers performance through the approaching of cut-points of TPR against FPR to 1 of a diagnostic examination. The ROC curve emphasizes the best performance of the five classifiers.

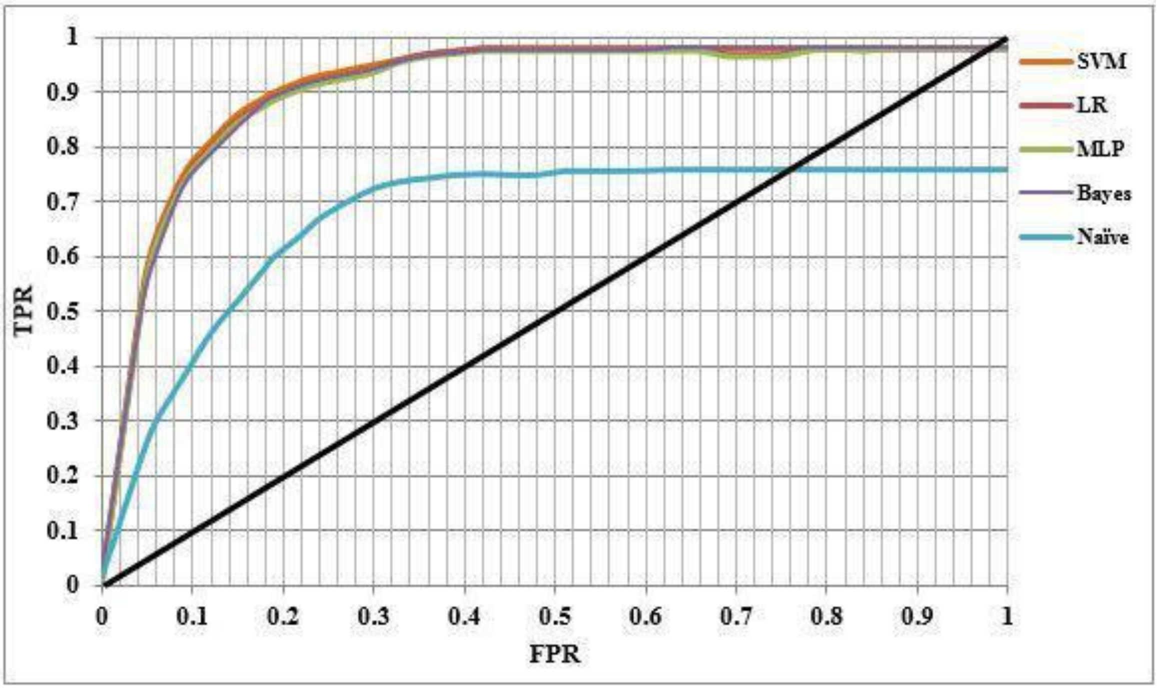


Figure 5-18: ROC Curve for TPR & FPR of all Classifiers

Accordingly, the ROC curve figure could exhibit a plot of great discriminatory power and similar performance of four classifiers (MLP, SVM, Bayes and LR) with TPR that is so close to a value of 1. Naive Bayes displays a weak classification compared with the other four classifiers and its discriminatory power with TPR is near to 0.7 value in this study.

- **Execution Time**

Besides the accuracy, the execution time is taken into account as another vital factor. To reduce the complexity in the systems, the dimension reduction is considered a crucial process that is done by the feature selection methods evaluating the feature's importance in the classification. The wrapper method is one of the feature selection methods depending on the leave -one –out (LOO) principle. This method excludes the feature that influences negatively or no effect on the accuracy from the dimension of the features, while it retains the feature that has a positive effect. The experiments of feature selection have indicated that there is a slight difference in the execution time. Certain tables (table 5-3 and table 5-4) show the execution time for all classifiers before and after applying feature selection through two folds (4 and 10 folds-cross-validation). Table 5-3 shows the time for all classifiers with 19 and 16 features before and after features reduction in 4-fold cross-validation while the table 5-4 shows the time for all classifiers with 19 and 16 features, before and after features reduction in 10-fold cross-validation.

Table 5-3: Time Values for Five Classifiers through 4-Cross-Validation before and after Features Reduction

4-Cross-Validation	**Naïve**	**Bayesian**	**MLP**	**Logistic**	**Poly**	**PUK**	**RBF**
Before	0.02	0.06	2	0.09	0.04	0.07	0.07
After	0.01	0.01	1.44	0.06	0.02	0.06	0.05

Table 5-4: Time Values for Five Classifiers through 10-Cross-Validation before and after Features Reduction

10-Cross-Validation	**Naïve**	**Bayesian**	**MLP**	**Logistic**	**Poly**	**PUK**	**RBF**
Before	0.03	0.08	2.15	0.23	0.08	0.11	0.14
After	0.01	0.02	1.65	0.14	0.04	0.07	0.08

- **Descriptors Performance**

In this stage, nineteen shape and textures parameters are applied on objects, which are clustered using the previously described clustering algorithm in chapter four, to describe the nodule patterns. Table 5-5 shows numeric values of some feature vectors for 20 instances (10 nodules and10 vessels), chosen as a sample of descriptors which have power or no effect on the precision of classification outcomes and discrimination. The experiments prove that three descriptors are ineffective compared with other features in this study. The remaining features hold significant information about the shape and size of each object. To explain the ineffective parameters in the classification, Figure 5-19, Figure 5-20 and Figure 5-21 demonstrate overlap between the parameter curves for both vessel and nodule for three parameters. Circularity, Ellipticity and Slenderness show noticeable overlap for nodule and vessels as a result of the convergence of values for both classes as in Table 5-6, Table 5-7 and Table 5-8. These parameters, therefore, have no an effect on classification accuracy through the validation by cross-validation method. The remaining features more strongly support classification. Figure 5-22 shows the area of both vessels and nodules and shows a significant difference and divergence between the two classes. Table 5-9 shows the area of both vessel and nodule objects. The perimeter of vessels is always greater than that of nodules, as shown in Table 5-10. The perimeter is thus considered a robust parameter in the description of the vessel and nodule as shown in Figure 5-23. Compactness is very important, low values of this parameter indicate that a nodule is close to a circular shape. Therefore, this parameter could be used to describe accurately the vessel and nodule as shown in Figure 5-24 and, Table 5-11 that show the compactness of both classes. Also, the rectangle degree parameter tends to diverge and helps discrimination between nodules and vessels as shown in Figure 5-25 and Table 5-12.

For more investigation of three inactive parameters, the t-test has done to have significant values around 0.070, 0.656 and 0.650 for Slenderness, Ellipticity and Circularity respectively while the remaining parameters have 0.05 as optimal criteria value in this test.

Table 5-5: Samples of Features Vectors for Nodule and Vessels

Area	Perimeter	Compactness	Slenderness	Diameter	Ellipticity	Circularity	Rectangle Degree	Class
55	91.154329	7.5822003	2.5	7.8825568	1.2051307	1.0977845	1.375	Vessel
120	154.85281	9.7380502	0.5	6.7553702	0.9506304	0.9750028	1.2244898	Vessel
144	169.3381	10.686062	0.5333334	8.0982339	0.9888245	0.9943966	1.2	Vessel
126	152.85281	10.358741	0.6153846	5.4048402	0.8472454	0.9204593	1.2115385	Vessel
66	99.39697	8.3441222	2	5.6770729	0.9837553	0.9918444	1.32	Vessel
105	148.61017	8.8787254	0.4285714	4.6032484	1.3110964	1.1450312	1.25	Vessel
154	165.3381	11.704629	0.7692308	1.9204957	0.843648	0.9185031	1.1846154	Vessel
228	212.79394	13.464352	0.6111111	7.4544388	0.8974361	0.9473311	1.1515152	Vessel
121	140.61017	10.813804	1	5.3042599	0.792134	0.8900191	1.21	Vessel
100	126.12489	9.9634342	1	4.5463619	0.7465584	0.8640361	1.2345679	Vessel
Area	Perimeter	Compactness	Slenderness	Diameter	Ellipticity	Circularity	Rectangle Degree	Class
49	82.669048	7.4484003	1	3.5330617	0.7683859	0.8765762	1.3611111	Nodule
30	61.941125	6.0862814	0.8	4.0789541	1.3328226	1.1544793	1.5	Nodule
56	90.911688	7.7406631	0.8571429	3.7803895	1.0107388	1.005355	1.3333333	Nodule
48	80.669048	7.4772891	1.4	5.0236848	0.782398	0.8845326	1.3714286	Nodule
42	74.426407	7.0914019	1.2	4.5737239	0.596839	0.7725536	1.4	Nodule
42	74.426407	7.0914019	1.2	3.7750535	0.9412414	0.970176	1.4	Nodule
20	45.455844	5.5290451	1.333333	3.2454343	1.1196016	1.0581123	1.6666667	Nodule
48	80.669048	7.4772891	1.4	5.6724492	1.1547636	1.0745993	1.3714286	Nodule
42	74.426407	7.0914019	1.2	2.7702703	1.270979	1.127377	1.4	Nodule
36	70.426407	6.4235755	2.666667	6.3370351	1.8526187	1.3611094	1.5	Nodule

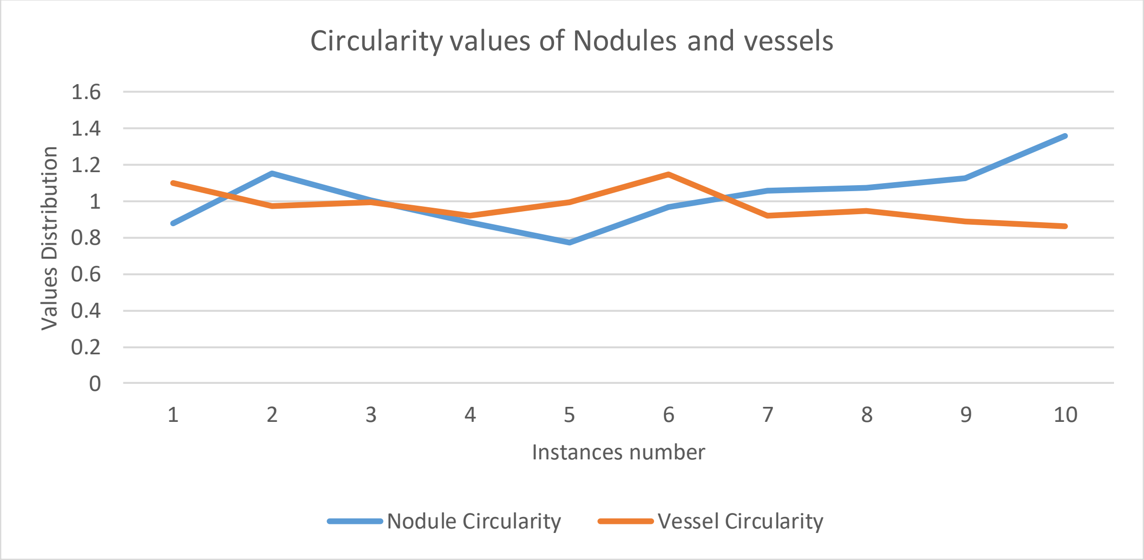


Figure 5-19: Curves of Circularity Values for Vessel and Nodule

Table 5-6: Circularity Values of Vessel and Nodule

Nodule Circularity	Vessel Circularity
0.8765762	1.0977845
1.1544793	0.9750028
1.005355	0.9943966
0.8845326	0.9204593
0.7725536	0.9918444
0.970176	1.1450312
1.0581123	0.9185031
1.0745993	0.9473311
1.127377	0.8900191
1.3611094	0.8640361

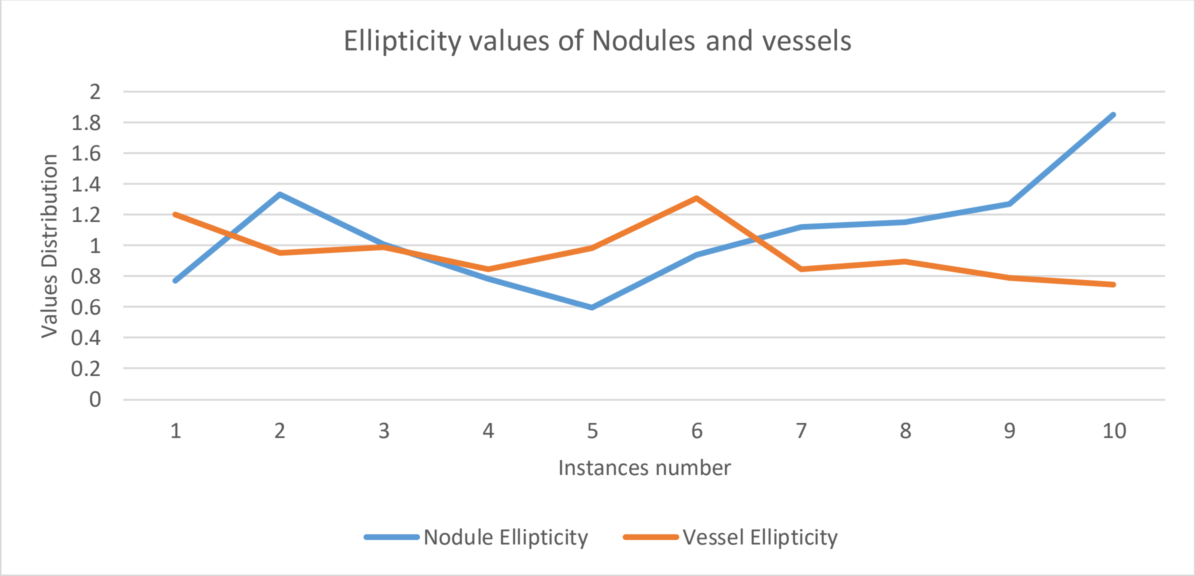


Figure 5-20: Curves of Ellipticity Values for Vessel and Nodule

Table 5-7: Ellipticity Values of Vessel and Nodule

Nodule Ellipticity	Vessel Ellipticity
0.7683859	1.2051307
1.3328226	0.9506304
1.0107388	0.9888245
0.782398	0.8472454
0.596839	0.9837553
0.9412414	1.3110964
1.1196016	0.843648
1.1547636	0.8974361
1.270979	0.792134
1.8526187	0.7465584

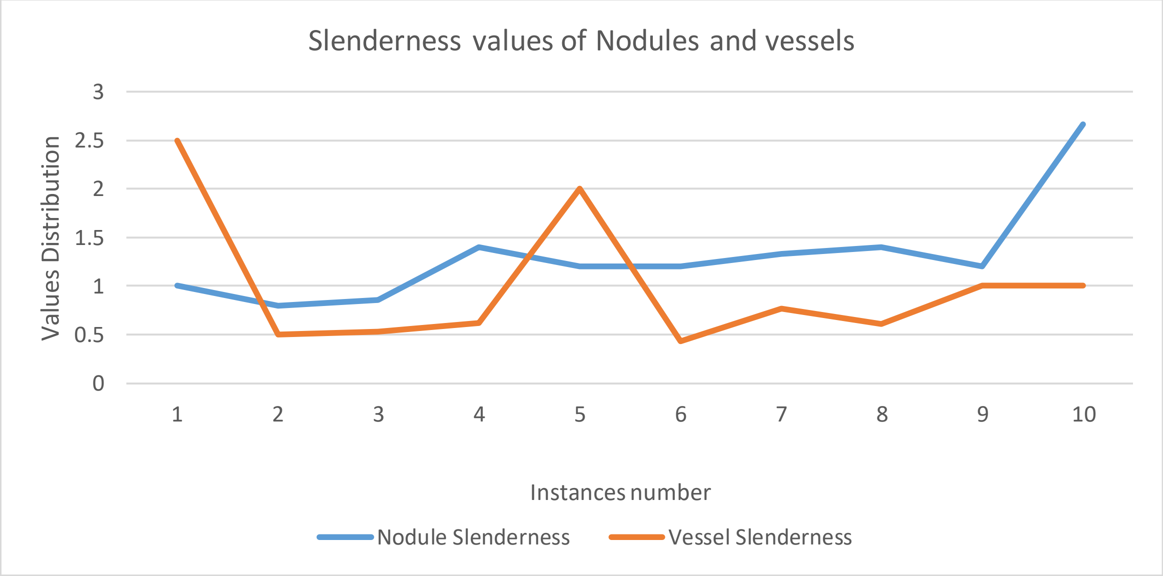


Figure 5-21: Curves of Slenderness Values for Vessel and Nodule

Table 5-8: Slenderness Values of Vessel and Nodule

Nodule Slenderness	Vessel Slenderness
1	2.5
0.8	0.5
0.8571429	0.5333334
1.4	0.6153846
1.2	2
1.2	0.4285714
1.333333	0.7692308
1.4	0.6111111
1.2	1
2.666667	1

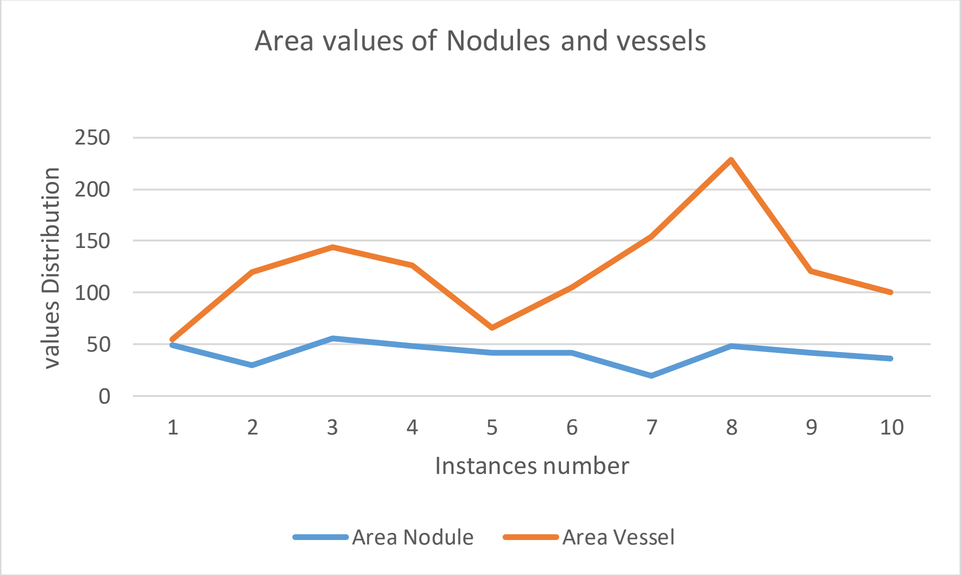


Figure 5-22: Curves of Area Values for Vessel and Nodule

Table 5-9: Area Values of Vessel and Nodule

Area Nodule	Area Vessel
49	55
30	120
56	144
48	126
42	66
42	105
20	154
48	228
42	121
36	100

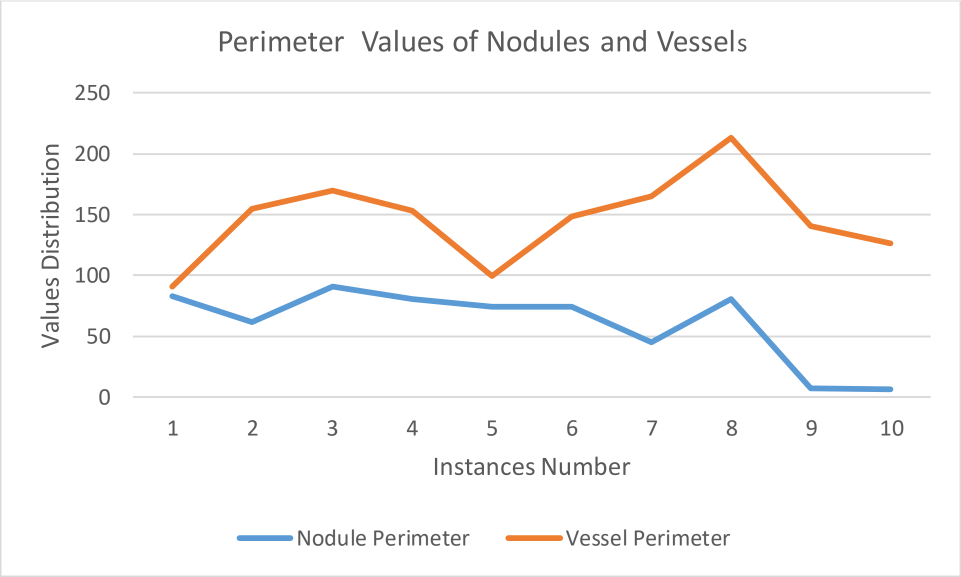


Figure 5-23: Curves of Perimeter Values for Vessel and Nodule

Table 5-10: Perimeter Values of Vessel and Nodule

Nodule Perimeter	Vessel Perimeter
82.66904756	91.15432893
61.9411255	154.8528137
90.91168825	169.3380951
80.66904756	152.8528137
74.42640687	99.39696962
74.42640687	148.6101731
45.45584412	165.3380951
80.66904756	212.7939392
7.091401936	140.6101731
6.423575505	126.1248917

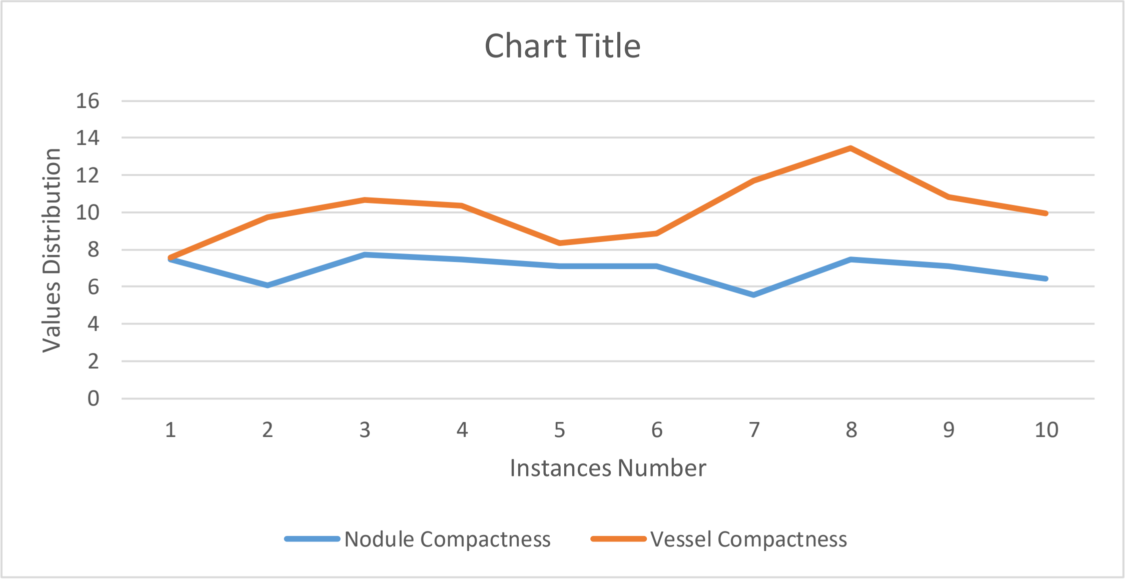


Figure 5-24: Curves of Compactness Values for Vessel and Nodule

Table 5-11: Compactness Values of Vessel and Nodule

Nodule Compactness	Vessel Compactness
7.448400318	7.58220034
6.086281375	9.738050199
7.740663142	10.68606191
7.477289093	10.35874093
7.091401936	8.344122197
7.091401936	8.878725375
5.529045102	11.70462907
7.477289093	13.46435199
7.091401936	10.81380395
6.423575505	9.963434217

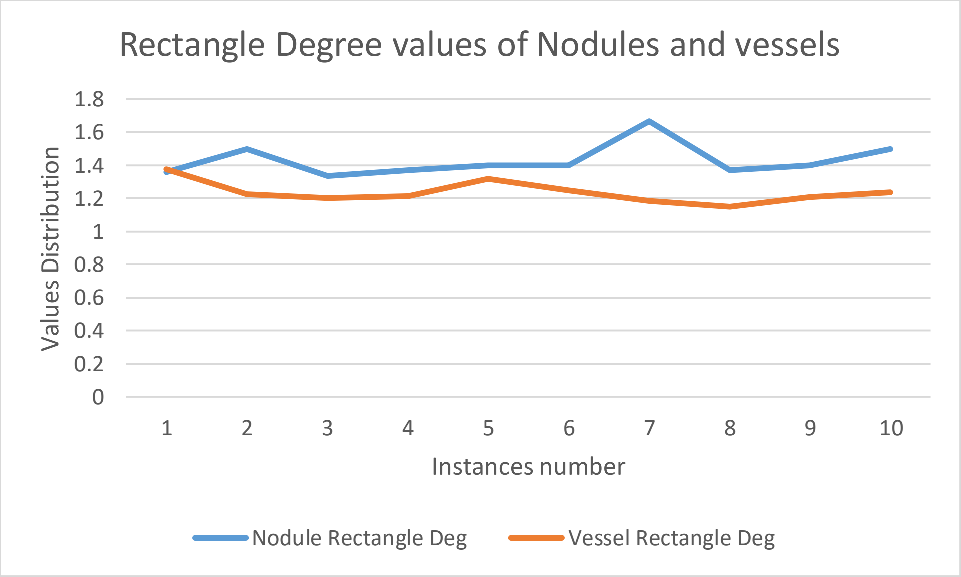


Figure 5-25: Curves of Rectangle Degree Values for Vessel and Nodule

Table 5-12: Rectangle Degree Values of Vessel and Nodule

Nodule Rectangle	Vessel Rectangle
1.361111	1.375
1.5	1.22449
1.333333	1.2
1.371429	1.211538
1.4	1.32
1.4	1.25
1.666667	1.184615
1.371429	1.151515
1.4	1.21
1.5	1.234568

Finally, this study provides the highest accuracy and stability at 98% for 4 classifiers with a sensitivity of 100% in nodule detection. The final evaluation of the completely computer-aided detection performance has been carried out using different tools. The performance of classification is evaluated by confusion matrix and other metrics that establish robust features in describing the instances of balanced data in this work. The Receiver Operating Characteristic (ROC) curves are one of these tools and present graphical summaries of a classifiers' performance in the automatic systems where it identified the optimal classifier through the access to 1 score value. Table 5-13 summarizes the accuracy, sensitivity and specificity rates for four optimal classifiers evaluate the classification performance in this study.

Table 5-13: Accuracy, Sensitivity and Specificity Rates for Four Optimal Classifiers

Classifiers	Bayesian	MLP	LR	SVM-PUK
Accuracy	98%	98%	98%	98%
Sensitivity	100%	100%	100%	100%
Specificity	96.4%	96.4%	96.4%	96.4%

5.7 Summary

One of the objectives of the research is to achieve high accuracy in detecting nodules. The study uses five common classifiers and identifies the optimal one for medical image analysis. Robust description of both vessels and nodule by textural and geometric descriptors plays a vital role in detecting the precision for four classifiers (Bayesian, MLP, LR, and SVM-PUK). Three geometric descriptors (Slenderness, Ellipticity and Circularity) have no effect on classification results; the number of descriptors reduces from 19 to only 16 features, which reduces training time and computation cost. In this study, an accuracy rate of 98% has been achieved by four classifiers, which were also evaluated by confusion matrix and other metrics that are used to measure the classification performance and balance of data. The high sensitivity that reaches 100% confirms no nodules were missed in classification, with the reasonable low false positive rate. For this purpose, the software

WEKA is used for training and testing the data through 4, 10 fold-cross-validation method. The Naïve Bayes classifier had the worst accuracy, as it fails in the presence of dependence among image attributes.

CHAPTER 6: 3D CLUSTERING AND MEASURING NODULE VOLUME

6.1 Introduction

In the previous chapters, the precise detection of pulmonary nodules has been achieved from 2D CT lung images in terms of accuracy and sensitivity. However, the full description of nodules, the accurate characteristic measurements, identification of small nodules and the existence of exceptional cases have not been defined in the studies that employed 2D images. As some nodule features require a multi-dimensional view to be detected, such as volume and type (juxta-pleural and juxta-vascular nodules), 2D techniques have not been considered adequate. Theoretical advantages of applying 3D volumetric computations include better estimation of total nodules bulk, more precise assessment of nodules development by adding a 3-view measurement, and a better evaluation of irregular and attached nodules. 3D clustering offers an accurate description of denser structures and low-density regions of lung components. Moreover, it reveals information on nodule shapes (spherical or irregular) and helps to address false positives as a requirement for computer-aided detection (CAD). The outcomes of 3D clustering have the capacity to improve the precision of diagnosis and support the radiologist in selecting appropriate treatment management (Lederlin et al. 2013) according to measuring the nodule volume. This chapter reviews the performance of 3D techniques in the enhancement of the 2D clustering results and the accurate description of the pulmonary nodule’s characteristics regarding size, location, type and the current nodule progress for early detection. A 3D clustering demonstrates the actual nodule numbers by identifying complicated cases that were suspected have missed by radiologists during the diagnosis. Moreover, measuring size has been automatically calculated for nodule (area and volume) and by the 3D-DBSCAN method for the first time. Furthermore, the descriptive form of the precession of the 3D clustering offers a potential detection of nodules that are less than 3mm. Finally, the outcomes are validated by matching the nodule centres in the system with the radiologists’ assessments.

6.2 3D Clustering Advantages

The 2D clustering has succeeded in automatically detecting the nodules and some of their geometric features (e.g. area, diameter, circularity and location); however, it is not able to describe the nodule progress on consecutive exam slices. Hence, 3D clustering offers a more precise description of nodule types and shapes of other lung components (vessels, bronchi, etc.)(Magalhães Barros Netto *et al.*, 2012). In fact, the accurate calculation of nodule volume has been achieved by 3D clustering. Also, good improvement in rating the false positives that resulted in the classification stage has been made using a 3D-DBSCAN method that rejects a lot of the classification results as a noise. Moreover, it also has the potential to identify nodules that are smaller than 3mm in diameter, which perhaps exist in the false positives rate.

6.3 Research Materials

A 3D clustering was implemented on 20 nodules in the CT lung images on the axial plane, with an image size of 512 ×512 pixels. It only describes nodules larger than 3mm of which characteristics (number, volume, diameter and location) were reported by radiologists' report. The images also have nodules smaller than 3mm, and only their total number was available in the radiologists' report.

6.4 3D Clustering

In this study, the empirical threshold that is adopted in the 2D tests of CT lung images is applied in the clustering to reduce parenchyma tissue and the search area of nodules within the image and to separate the intensity values that include the vessels and candidate nodules together. Through the Modified K-means (MK-M) method is used to cluster intensity values in two clusters (vessels and nodules) to ignore the unwanted spots, which are with threshold results, and to label the data by two clusters. The MK-M is applied for each image in the scan separately, and the clustered images are stacked to be undergone the 3D graphics functions later. The K-means method is considered an efficient clustering technique because it keeps clusters in their appropriate location by allowing the centre of each cluster to shift and fit the cluster. It also clusters data in a non-supervised mode (Anand and Jeffrey 2011), (Peter and Karnan 2013) and (Celebi et al., 2013). A 3D graphic of labelled clusters is computed by some scalars with the integration of techniques and functions, which are provided by Mathworks for reconstructing a 3D structure of data. Geometric computations of clustered data are measured by the Isosurface function that presents an improved volumetric visualisation of objects. The function inputs are an array of pixel locations (x, y, and z) and value belonging to the pixel coordinates for each cluster, to generate the 3D volume. Equally, the Isocaps and Isonormal computations, patch and view functions contributed to the production of an enhanced 3D vision of the nodules and vessels compared to 2D images.

The algorithmic steps applied to cluster each lung nodules and vessels in 3D as follows:

A. Thresholding (T)

In this process, the algorithm is based on an empirical threshold to reduce the regions and components surrounding the region of interest (ROI) to detect the nodule easily.

$$\text{Pixel intensity} > \mathrm{T}$$

Where values of intensity greater than the threshold define intensity in the image, which represent the nodules and vessels values.

B. MK-M

The next step is applying the K-means clustering on 2D CT lung images to remove the undesirable spots, which are produced from the threshold of the previous step. In addition, it identifies and labels the data in the two clusters for 3D functions processing purposes. The clusters represent the intensity values (nodules and vessels). K-means clustering carries out the following steps.

1. Selecting k to identify cluster centre, C=c_1, c_2,…, c_k

2. Calculating the distance between each data pixel and cluster centres.

$$S = \sum_{i=1}^{m} \sum_{j=1}^{n} || \, x_i - c_j \, ||^2 \qquad 6\text{-}1$$

Where x_i is an image pixel, m and n are the image’s dimension

3. Assigning the data pixel to a cluster that has the minimum distance between its centre and the data point.

4. Updating the cluster centre according to Eq. 6-2:

$$C_i = \frac{1}{N_{i,j}} \sum_{j=1}^{N_{i,j}} x_{i,j} \qquad 6\text{-}2$$

Where c_i represents the i^{th} cluster centre, $N_{i,j}$ represents whole data pixels that belong to the i^{th} cluster centre while $x_{i,j}$ represents the data pixel belonging to i^{th} cluster centre.

5. Repeating the steps 2 to 4 until the objective function S (4) becomes minimized.

6. After that, the labelled clusters pixels undergo the 3D functions processing to generate the volumetric data.

C. 3D Functions

A 3D graphic is one of the image processing applications, which plots the data volume according to computations that are figured for the image pixels. The following are techniques combined to create 3D graphics of K-means outcomes to reflect nodule's shape and type (Varios el at., 2003).

1) **Isosurface** is a built-in Matlab functions presents an active treatment of volume visualization. It is applied to gather the inputs **x, y, z** which represent the coordinates of data and the **V** parameter is a value belonging to the above coordinates positions in the 3D space. **Isosurface** output is a structure which contains the faces and vertices values (Engel et al., 1999) and (Bankman and Morcovescu 2002).

2) For the current axes of vertices and faces computed by Isosurface, the **Patch** function provides a patch with the colour given to faces in 3D space. The end-cap geometry of **Isosurface** data is computed by **Isocaps** functions to create an effect of isolating the surface and showing the values distribution on that plane with a red structure.

3) **View** is a function creates axes in a 3D view if no axes were computed for the formed faces in the previous step. It deals with two parameters **(**Azimuth and Elevation). As the Azimuth parameter represents a polar angle which is positive angles in the *x*-*y* plane

rotating counter-clockwise of the viewpoint while Elevation parameter is the angle above as positive angle once or below (negative angle) the *x*-*y* plane.

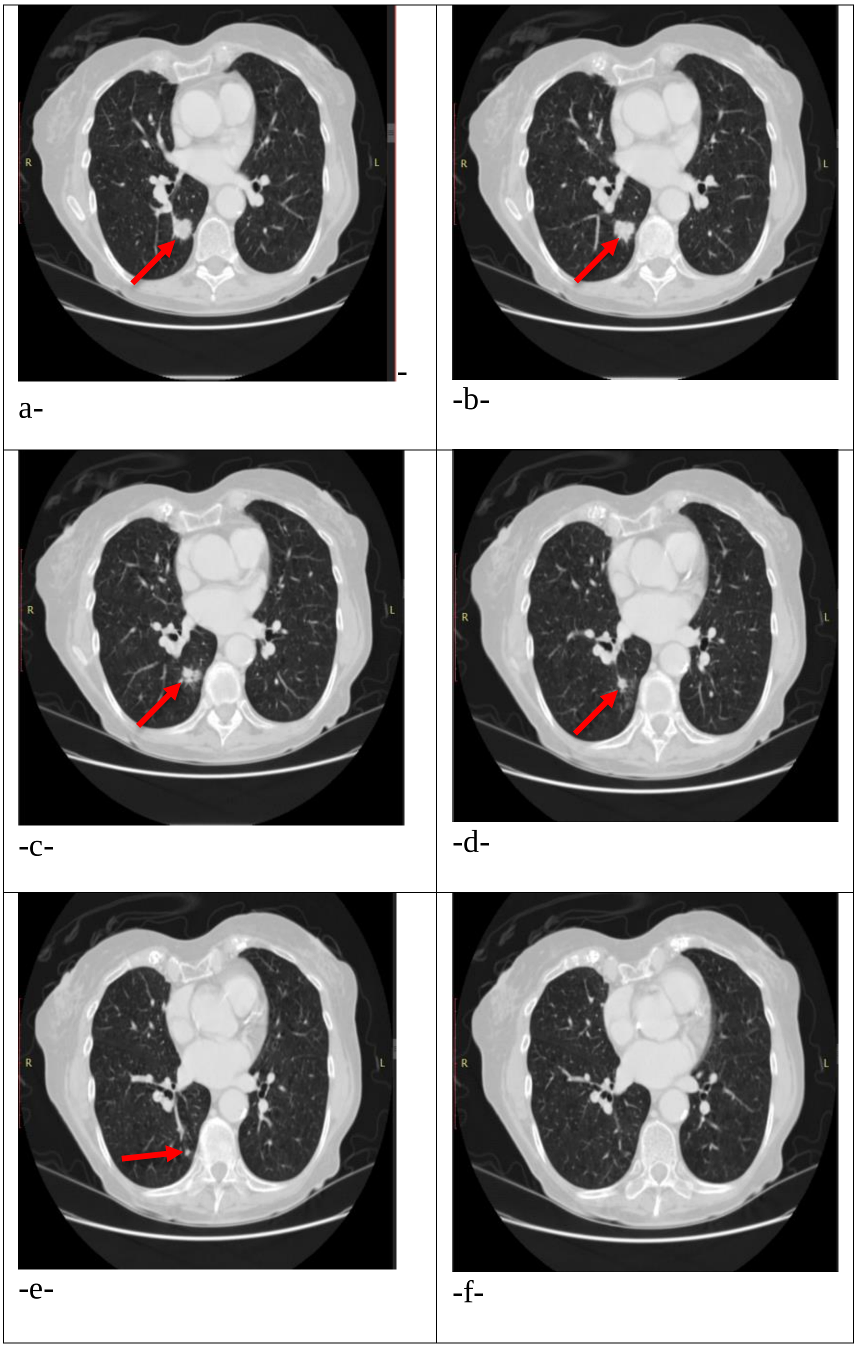
4) **Isonormals** calculates the normal of Isosurface vertices are handled for patch function to form a shadow of the surface through vertices gradient.

In 3D clustering, the cluster data identified by the modified K-means algorithm will be plotted in 3D space with the above functions to build a 3D enhanced visualisation of nodules and vessels. The 3D method improves on a 2D clustering to provide the final description of the nodule characteristics that were not mentioned in the radiologists' report such as the nodule types. In addition, it could distinguish a compound case for developing nodule patterns and its progress through the CT lung layers.

6.5 Case Study

This study identifies a complicated case that was invisible through radiologists' diagnosis. According to the scan reported in a radiologists' assessment, a patient has six nodules in different positions in the 2D images of CT lung. The 3D analysis describes six nodules in more details providing information regarding shape and location. Two nodules are found in slices 60 and 64 as shown in Figure 6-1 which also shows the original CT lung images, with a resolution of 512 ×512, of slide 60 to slide 65 which contain two nodules evaluated by four experts into 2D CT images. Also, Figure 6-1 refers, with a red referral, to the two nodules that are located in slices 60 and 64 while slices, which are among them, have similar and attached structures of the nodule progress in the 60th slice. After the experiments, the 3D clustering of exam images reported a connection between the two nodules attached by a thin tissue as shown in Figure 6- 2. The tissue line between the two nodules is invisible in 2D images. Therefore, the radiologists could not recognise it, which led to evaluate the one nodule as two isolated nodules. Generally, the method outcomes have emphasised that the nodule in the 60th slice is the largest and significantly developed through slices 61, 62 and 63 to be thinly linked to a smaller mass in slice 64. Consequently, the radiologists' evaluation of one nodule as two in different slices is imprecise.

Interestingly, although the radiologists' evaluation did not demonstrate the nodule type, a 3D method has identified the types of the nodule in this work in which the nodule classification represents the juxta-vascular nodule and a nodule attached to the lung wall (juxta-pleural). Also, the 3D clustering's ability to detect the thin line, which connects the nodules in the slices 60th and 64th, reveals a potential for early nodule detection, which is thought to increase the chance of patient survival with early treatment.



a-

-b-

-c-

-d-

-e-

-f-

Figure 6-1: Original CT Lung Images of Slices 60-64 with Red Referral of Two Attached Nodules and Assessed by 4 Experts, a. Candidate Nodule in the 60 slice; b, c, d. Similar Structures of Nodules through 61, 62, 63; e. Candidate Nodule in 64 Slice, f. The Normal Image after 64 Slice

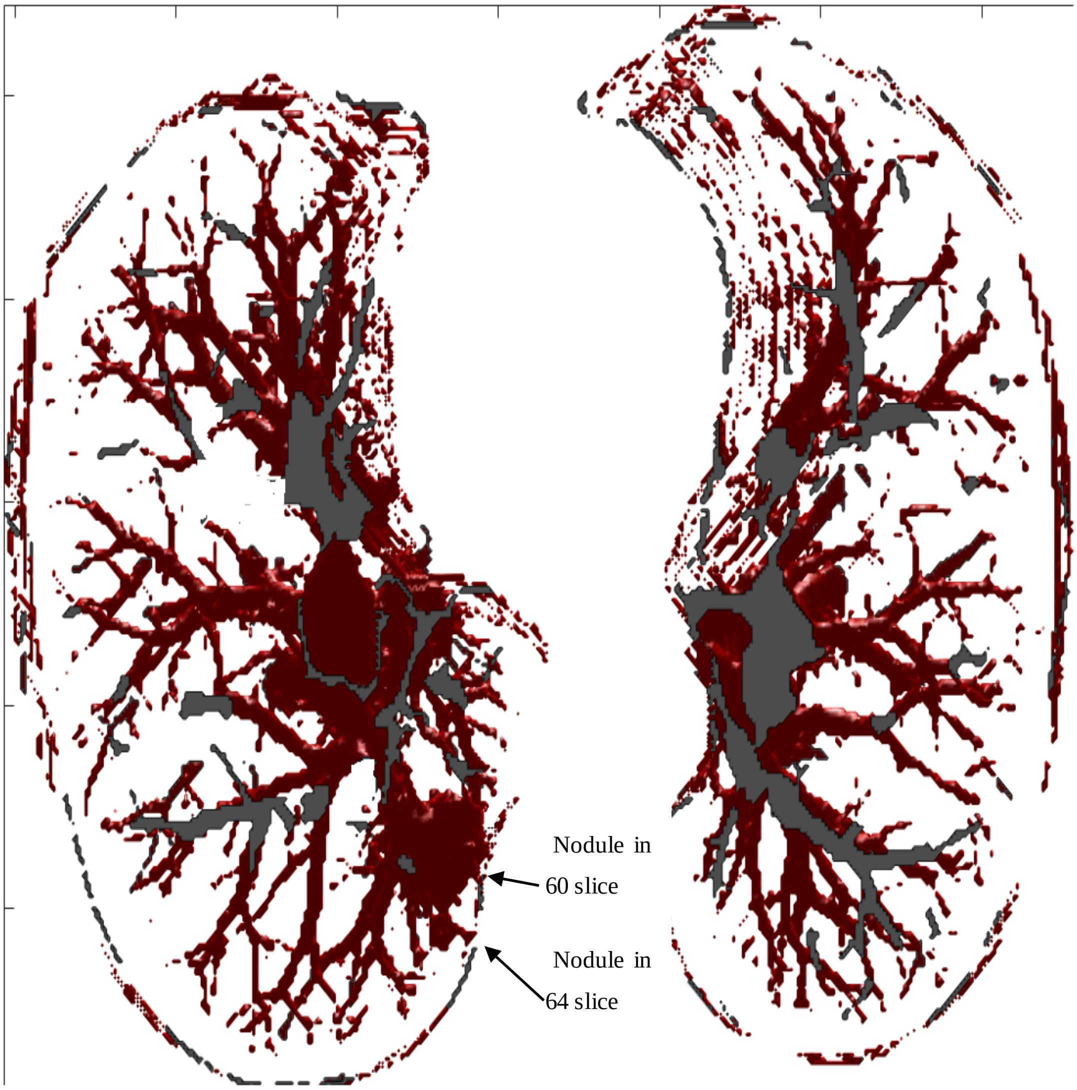


Figure 6-2: 3D Clustering of CT lungs with Referral to Two Diagnosed and Attached Nodules in 60 and 64 Slices

In addition, in this study, one of the ways used to validate the results is the centre pixels of nodules, which corresponded, with the centres of actual nodules in the radiologists' evaluation; Figure 6-3 shows the attached nodules and another nodule in the 46th slice with their centre pixels.

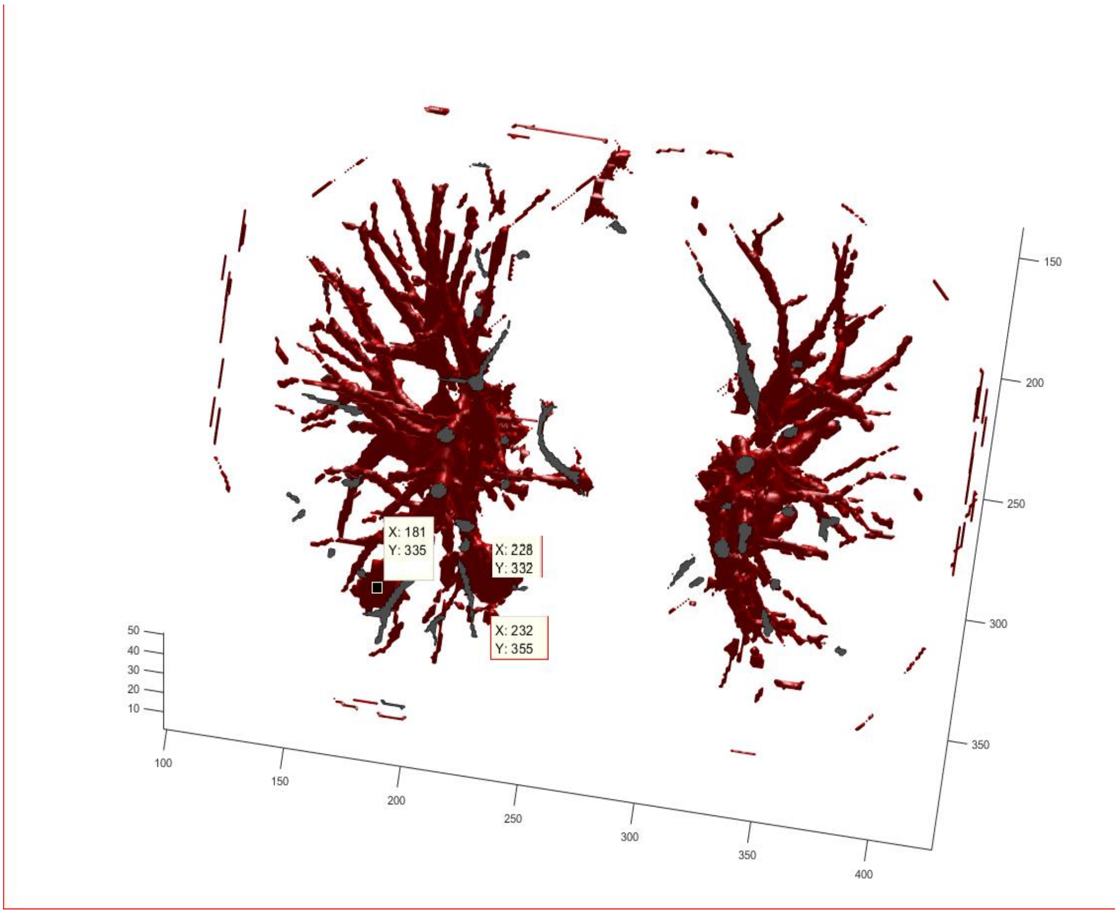


Figure 6-3: 3D Clustering of Attached Nodules in 60 and 64 slices and Another One in the 46 slice with Centre Pixels Corresponding Radiologists' Assessment

6.6 3D Clustering Evaluation

Although the 3D graphic is visually appealing, it is also necessary to quantify the quality of results. To evaluate the accuracy of the proposed 3D clustering and the case study result, another 3D clustering method (3D-DBSCAN) and 3D plot.ly application were employed in this research. Both evaluate the proposed 3D clustering method results in terms of the description and detection accuracy, matching the system output and actual nodules centre pixel, which were reported in the experts' assessment (Bankman and Morcovescu 2002) and (Papademetris and Joshi 2006).

6.6.1 DBSCAN Method

The DBSCAN method (Density-based spatial clustering of applications with noise) is one of the most effective and common clustering methods based on regions density (Sander et al. 1998), (Daszykowski and Walczak 2010)and (Schubert et al. 2017). This algorithm collects a set of nearby points to be one cluster while it marks other points as outlier pixels that lie in isolation in low-density areas. Clustering of this method relies on two effective parameters are Epsilon and MinPts which deal with core pixels in the data. These parameters and core pixels are defined below.

Core pixel: it is a point, which is surrounded by neighbours of clusters exceed the MinPts neighbour's number within a distance is equalled the radius of Epsilon.

Epsilon (ε): it is the maximum radius between the Core pixel and other pixels.

MinPts: it represents the smallest pixels number that forms a density region surrounding or close to a core pixel.

This algorithm has been developed to cluster in 3D and evaluate the results of previous 3D clustering and case study. In this study, the 3D-DBSCAN algorithm finds the connected objects to core pixels according to the ε (eps) and MinPts parameters. At the same time, the algorithm ignores all points that do not represent core pixels or the points that are not included by the core pixel's neighbours. This method assigns all nearby clusters to core pixel if the cluster is within ε (eps) neighbour while other points are assigned as noise. There are many advantages to the algorithm, but the most important is that the method does not need the previous determination of cluster numbers in the data as opposed to k-means. Secondly, this algorithm also discovers randomly shaped clusters and reduces the effect of single-link according to MinPts parameter, and finally, a domain expert can specify the MinPts and ε parameters values, when the data is fully understood.

In this study, very important tasks were achieved thanks to the valuable advantages of DBSCAN. This method is modified to cluster in 3D and to enable the evaluation of the proposed method in this work. The 3D-DBSCAN method was used to visually evaluate the case study and compare the results with the proposed method clustering outcomes by looking at centre pixel locations. Figure 6-4 displays the algorithm clusters in 3D, showing the structure of the linked nodules in different levels with centre pixels that show the true location of the nodule. In this study, the 3D-DBSCAN algorithm evaluates the connection between attached nodules in the previous clustering and the precise outcomes of the proposed clustering method validating by matching in the centres' pixels as shown in Figure 6-5. In Figure 6-6, DBSCAN clusters the linked nodules between 60th and 64th slices to show the spread of the nodule by a thin line to form small nodule that is considered isolated.

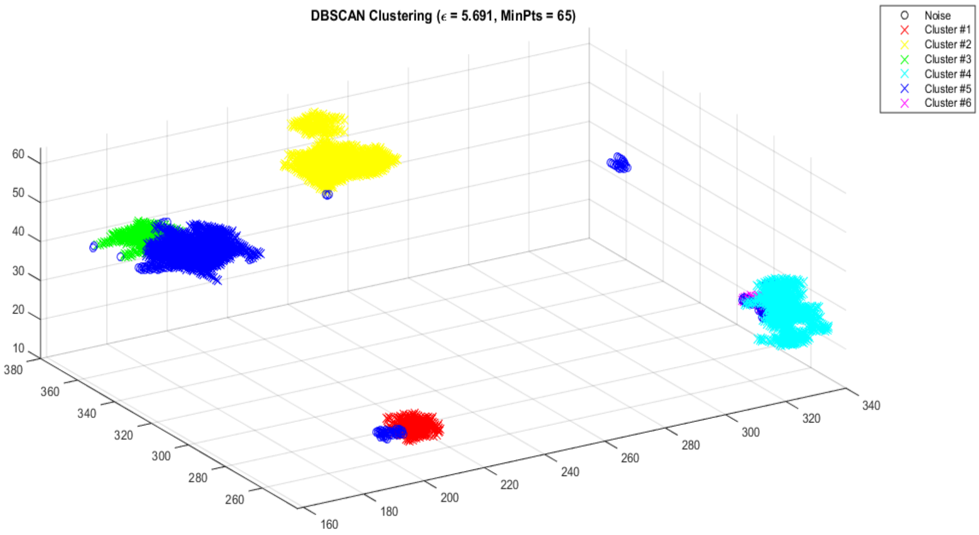


Figure 6-4: 3D-DBSCAN Method Clusters the Linked Nodules in Yellow and the Remaining Nodules with Different Colours of the Same Patient

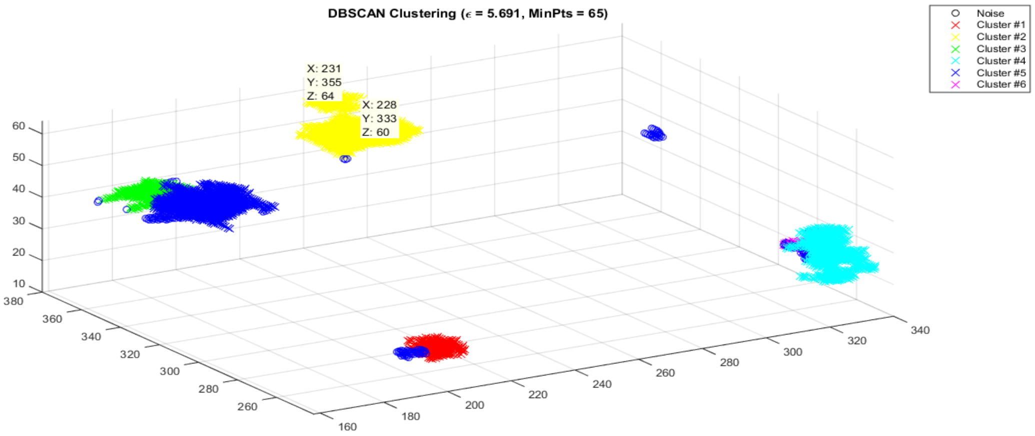


Figure 6-5: 3D-DBSCAN Method Clusters the Linked Nodules with Yellow Colour, Centres Pixels, Which Were Reported in Radiologist Evaluation, and the Remaining Nodules with Different Colours

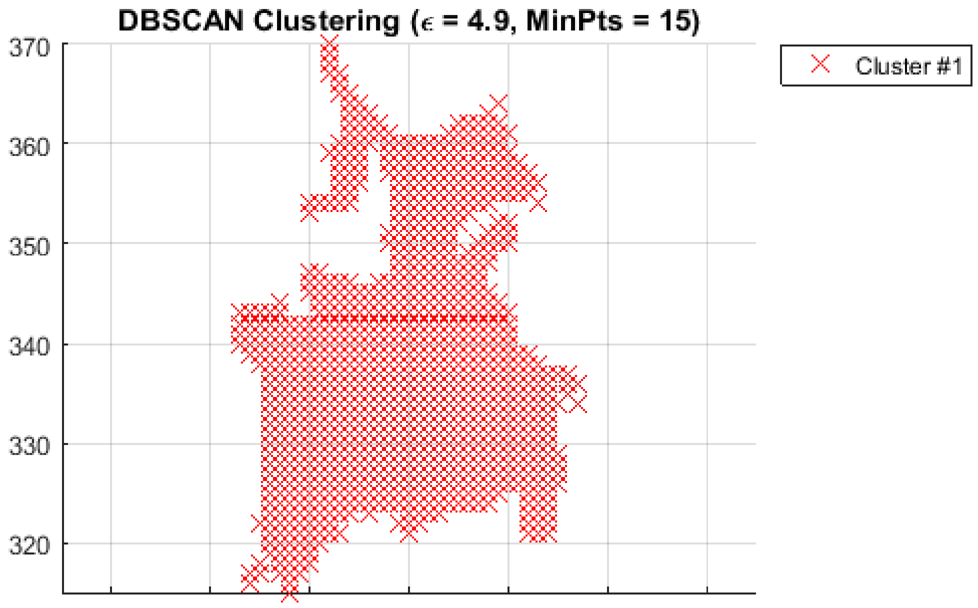


Figure 6-6: DBSCAN Method Just Clusters the Linked Nodules and Proves One Nodule

6.6.2 Plot.ly Application

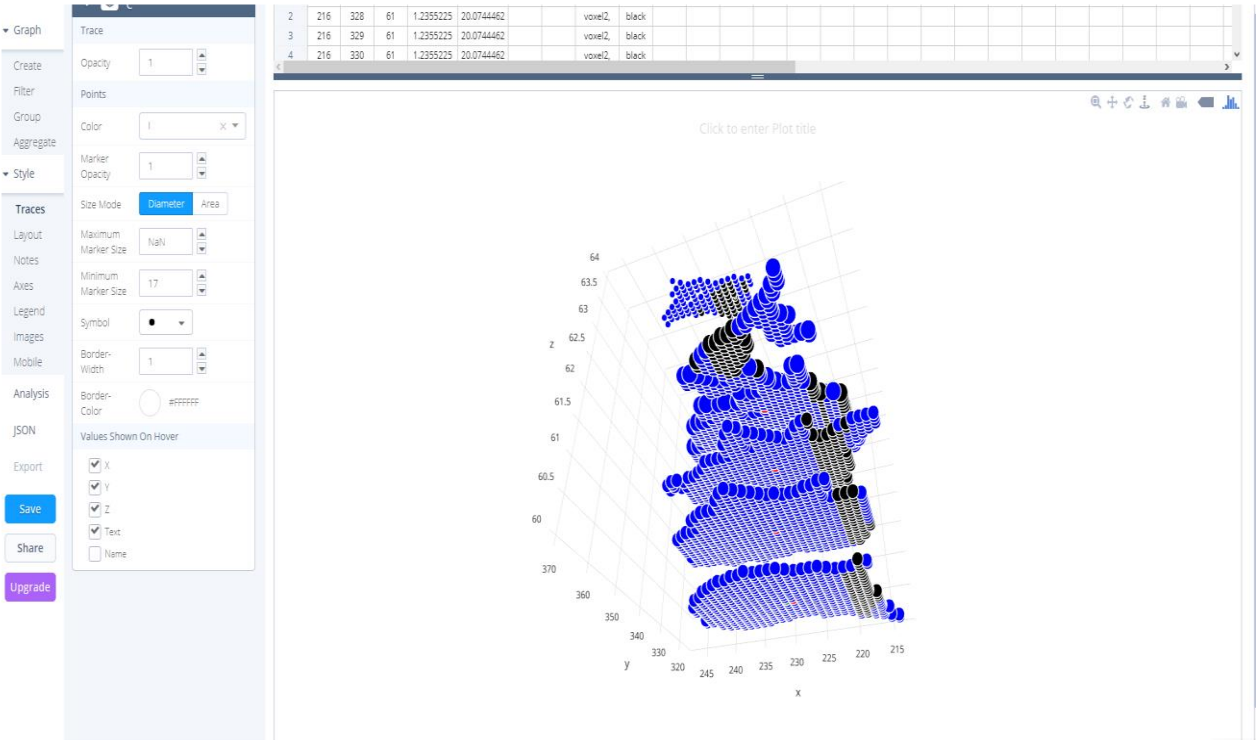
Another 3D application is also used for evaluating the case study outcomes. It is called plot.ly that is available on the site[2] and protected; therefore, it is difficult to know its algorithm. This application plots the location of pixels as coloured bubbles of the candidate nodule and similar structures at different levels as shown in Figure 6-7 and Figure 6-8. The plot of the application contains the area and location of the structure of nodules, which are clustered in the previous 2D study from consecutive images with centre pixels. The application inputs are vectors consisting of location values of clustered nodules and their similar structures that are extended through the consecutive slices. Also, the third dimension of the pixels represents the slice number that contains the candidate nodule structures. The pixel size is required in the application and measured by the pixel spacing,

---

[2] https://plot.ly/create/

which calculated by Eq.6-3, times the thickness of the image. In addition, the area pixels are labelled with two colours: red and green for the nodule centre pixels of the proposed method in this study and actual nodules that were assessed by radiologists respectively. Therefore, sometimes the bubbles of centre points overlap because their convergent values as shown in Figure 6-9. Besides, the bubbles with black colours illustrate the overlap regions and connectivity between clusters through slices 60 to 64 while the blue bubbles form the remaining pixels of the cluster areas.

Generally, the plot application and 3D-DBSCAN method efficiently evaluate the clustering of the 3D proposed method and case study when they could accurately describe the regions of the connection between linked nodules in CT lung images. By the evaluation of the results of previous techniques, the nodule that is detected in the slice 60 and unified with similar structures through slices 61, 62 and 63, actually ends in the slice 64. Consequently, the nodule in slice 64 is no other nodule but an extension of the 60 slice nodule.



2	216	328	61	1.2355225	20.0744462			voxel2,	black
3	216	329	61	1.2355225	20.0744462			voxel2,	black
4	216	330	61	1.2355225	20.0744462			voxel2,	black

Trace

Opacity 1

Points

Color

Marker Opacity 1

Size Mode Diameter Area

Maximum Marker Size NaN

Minimum Marker Size 17

Symbol ●

Border-Width 1

Border-Color #FFFFFF

Values Shown On Hover

☑ X
☑ Y
☑ Z
☑ Text
☐ Name

Figure 6-7: 3D Plot of the Areas of Nodules with Similar Structures Between Different Levels from Slice 60 to 64 (going upwards)

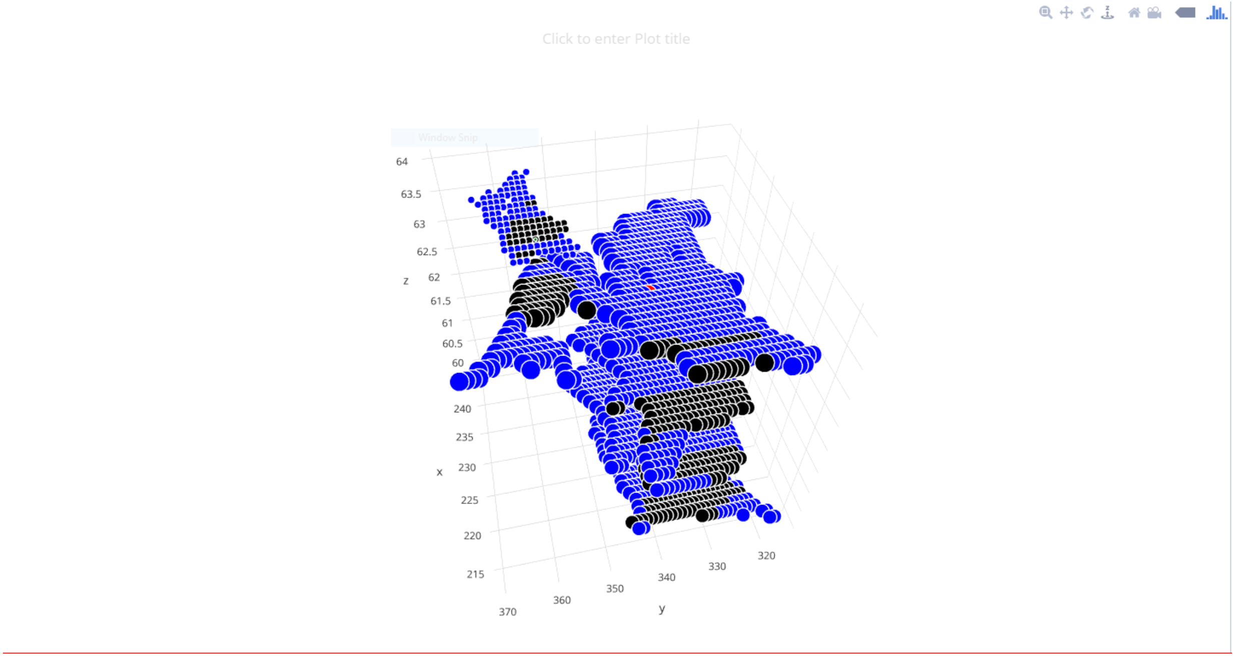
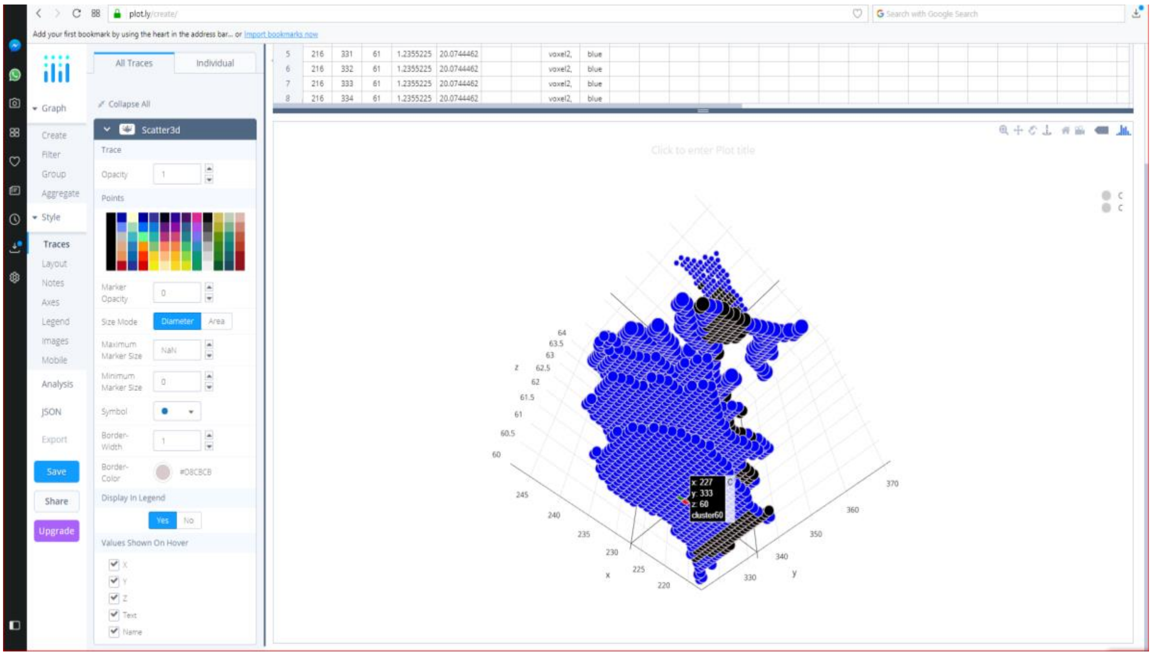


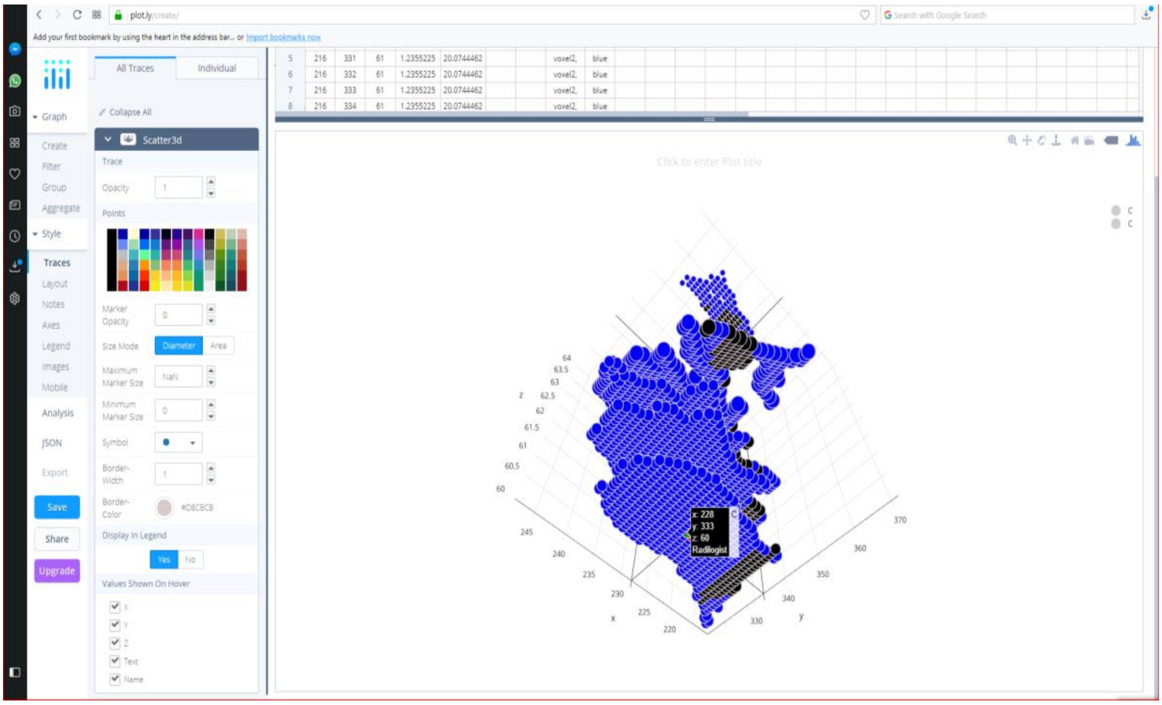
Figure 6-8: 3D Plot from Different Angle to Figure 6-7, Black Areas (Bubbles) Represent the Overlap Between Them to Show the Connected Regions



5	216	331	61	1.2355225	20.0744462			voxel2,	blue
6	216	332	61	1.2355225	20.0744462			voxel2,	blue
7	216	333	61	1.2355225	20.0744462			voxel2,	blue
8	216	334	61	1.2355225	20.0744462			voxel2,	blue

Opacity: 1
Marker Opacity: 0
Size Mode: Diameter Area
Maximum Marker Size: NaN
Minimum Marker Size: 0
Symbol: ●
Border-Width: 1
Border-Color: #D8CBCB
Display In Legend: Yes No
☑ X
☑ Y
☑ Z
☑ Text
☑ Name

-A-



5	216	331	61	1.2355225	20.0744462			voxel2,	blue
6	216	332	61	1.2355225	20.0744462			voxel2,	blue
7	216	333	61	1.2355225	20.0744462			voxel2,	blue
8	216	334	61	1.2355225	20.0744462			voxel2,	blue

-B-

Figure 6-9: (A, B) Show the Overlap between Centre Pixels of Clustered and Actual Nodule in 60 Slice with Two Colours Red and Green Respectively

6.7 Measuring the Volume

One of the most important characteristics of a nodule is its volume. This value is a vital parameter that describes the nodule development in the lung. A tumour's size determines the cancer risk progress that is necessary to be known for the specialist to select appropriate treatment management (treatment, surgery and follow up) of a tumour. A radiologist who relies on the Max software, which is common in a clinical setting to estimate the size of the candidate nodule, performs the task. The software groups the boundaries within unified regions of interest (ROI) corresponding to physical nodules. The volumetric size is computed by filling the area inside the boundary (excepting the boundary itself) and then multiplying the voxel size and nodule voxel number to produce the volume using Eq.6-5. The voxel size is the product of times x-size and y-size of the nodule voxel and the thickness value for the image slice using Eq.6-4. The x-size and y-size values represent a pixel spacing against two pixels centre in the medical image and are obtained through the squared ratio of the field of view of CT image to the image array size (512×512) using Eq.6-3 shown in Figure 6-10. The image parameters for the data in this work are: pixel spacing = 0.703/0.703, thickness=2.5 and voxel size =1.235.

$$\text{Pixel spacing}(\text{x}-\text{size},\text{y}-\text{size}) = \left(\frac{(\text{Field of view(FOV)})}{(\text{Matrix Size})}\right)^2 \qquad 6\text{-}3$$

$$= \left(\frac{360mm}{512}\right)^2 = 0.703^2 = 0.4943$$

$$\text{voxel size} = \text{pixel spacing}(\text{x}-\text{size}) \times \text{pixel spacing}(\text{y}-\text{size}) \times \text{thickness} \qquad 6\text{-}4$$

$$= \quad 0.703 \times 0.703 \times 2.5 = 1.235$$

$$\text{Nodule volume} = \text{voxel size} \times \text{nodul voxels number} \qquad 6\text{-}5$$

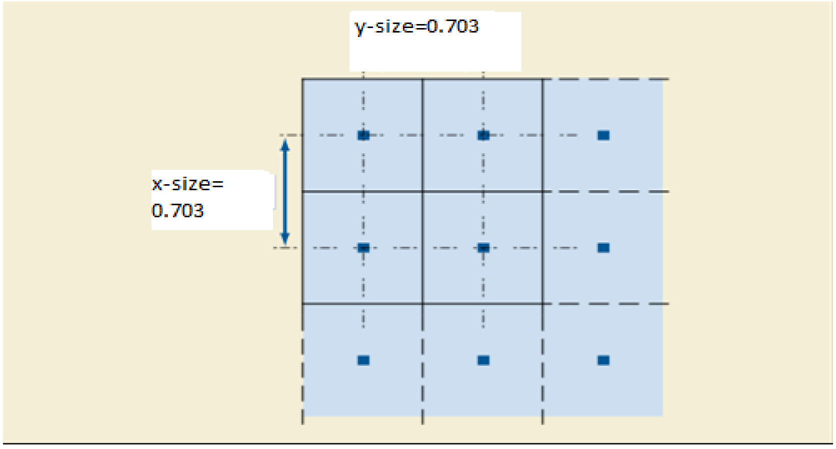


Figure 6-10: Pixel Spacing Values against of Two Pixels Centres in Vertical and Horizontal of 2D Pixels Matrix (Treichel et al. 2012)

6.8 Nodule Volume Measuring Methods

In other works of pulmonary nodules detection, the sizes of a tumour in 2D and 3D images are usually calculated in two ways: counting the object pixels to represent the nodule's area in the 2D image and manually adding up the areas of the attached structures in different levels to produce the nodule volume value in a 3D image. Although there are existing attempts to measure the nodule volumes, they are still computed manually in 3D images. In 2D images, the area of the nodule is computed with image features in the previous chapter.

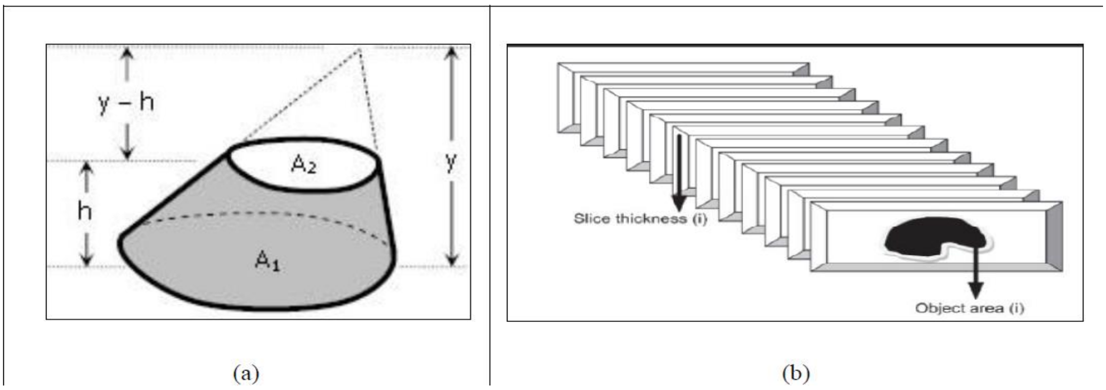
6.8.1 Manually Nodule Volume Calculation

Previous works calculate the nodule's volume in 3D depending on manual measurement. In this way, researchers usually use segmentation of consecutive nodule regions extended through a number of slices. The convex area, which represents the number of the actual pixels of hull area in the image and is a parameter is calculated in the Matlab, is measured individually for each region where the nodule structures are located, slice by slice (Mahmood, Abbas, and Ali 2014) and (Hr and Chitharanjan 2013). To measure the volume of the whole nodule the frustum formula is applied to areas collected from 2D images. Although this formula measures regular shapes such as circle and pyramid, it is used to measure the nodule volume that usually comprises the irregular areas. After that, the nodule areas and the thickness value of CT images are considered as inputs of the Frustum Formula (V) to produce the total volume using Eq.6-6.

$$V = \sum(\frac{h}{3}(A_1 + A_2 + \frac{(A_1 \times A_2)}{2}) \qquad 6\text{-}6$$

Where height h = slice thickness + slice separation, A1 and A2 denote the areas of the two consecutive slices having a tumour.

All the calculated areas have been converted their unit from pixels to mm as in section 6.8.2. Figure 6-11 shows a) the frustum model and b) the number of slices that include nodules and its sequential structures. For example, Figure 6-12 shows ten slices of segmented and clustered lung images in two columns with red arrows pointing to the nodule. For each slice, the convex area is measured for each candidate nodule in the clustered images. Also, one column shows the total manual volume produced by the frustum formula using Eq.6-6 converted to mm using Eq.6-7.



(a) (b)

Figure 6-11: (a) Frustum Model, (b) CT Slices Contain Nodule's Structures (Mahmood, Abbas, and Ali 2014)

6.8.2 Converting the Area Unit from Pixel to MM

In computer-aided detection systems, the object area consists of the number of pixels that form the object. The measuring unit is usually the pixel while the actual unit used by the radiologist to describe nodule volume is mm. The pixel is the smallest element and is a square in the image having a size mm measured using Eq. 6-4. To convert the nodule area unit to an actual unit in mm, the number of area voxels is multiplied by the pixel spacing value using Eq. 6-7.

$$\text{Nodule Area } (\text{mm}^2) = \text{ number of area voxels } \times \text{pixel spacing} \qquad 6\text{-}7$$

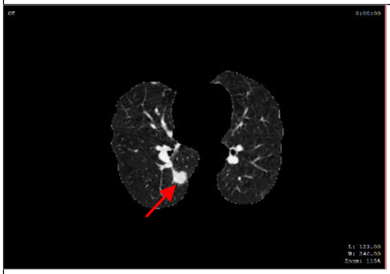
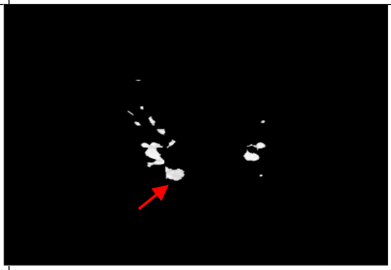
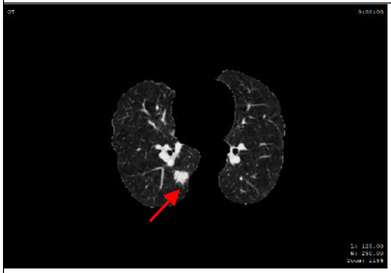
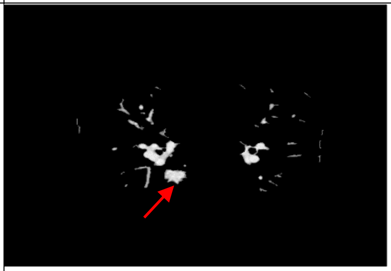
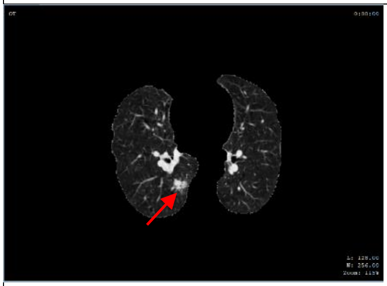
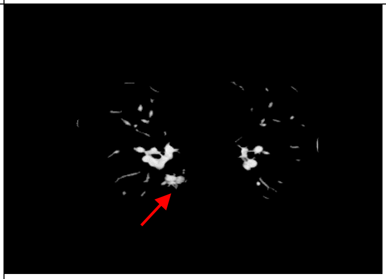
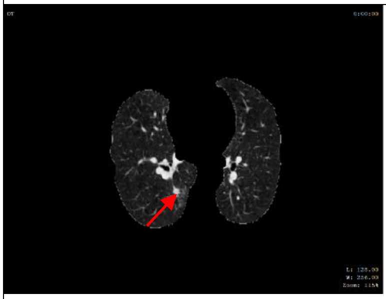
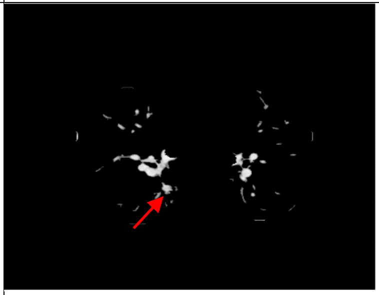
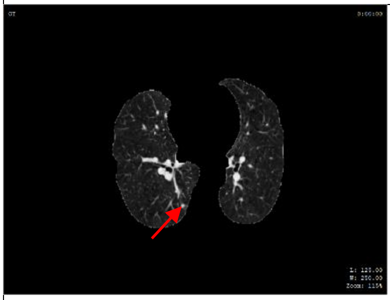
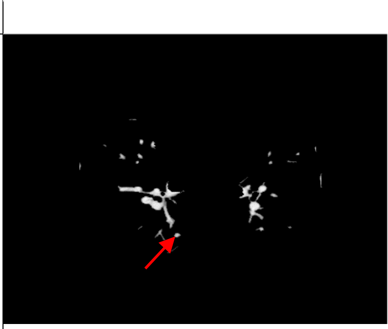
Pixel Spacing=0.4943	Voxel size=1.235			
2D segmented CT lung	clustering	Convex Area in pixels	Convex Area in mm	Manual volume
		497	247.377	45872.1534
		609	303.314	
		520	261.422	
		204	126.849	
		54	27.186	
		1954.5	966.10935	

Figure 6-12: Segmented and Clustered CT Lung Images for Nodule Starts from Slices 60 to 64 with Column for Convex Areas in Pixels and MM of Nodules, Manually Calculated Volume

6.8.3 Automated Nodule Volume Measurement

The idea of automatically measuring the nodule volume starts from the manual method of volume calculation employed in the current researchers. In this study, the automated nodule volume was measured by exploiting the DBSCAN method characteristics in the clustering. The method is developed to cluster in 3D whole classes that represent nodules and false positives in the classification results of the previous 2D detection from CT lung images. The classification outcomes that are usually a binary output (0, 1) where class 0 is considered vessels and the class 1 is nodules and false positives. The 3D-DBSCAN method clusters class 1 location (nodules and false positives). For each cluster, the radius and the number of pixels (area) are calculated to identify the Epsilon and MinPts as parameters of DBSCAN later. The 3D-DBSCAN method, which clusters the attached and high-density regions of the nodule location in the different levels, considers the maximum radius and the pixels that form the smallest density among class 1 clusters as Epsilon and MinPts parameters respectively in the clustering. Due to the connection between the cluster's structures in consecutive slides, which the DBSCAN method has performed in 3D clustering, the nodules have become accessible for processing purposes. This allows one to perform the calculation on each individual cluster easily. One of these calculations is to implement the trapezoidal method that measures the area of numeric data and irregular shapes. This method is applied to each a cluster in the 3D-DBSCAN figure directly by the trapezoidal integration function in Matlab (Davies and Hicks 1981). This function plots curve for numeric data and the area under the curve is divided by the trapezoidal method to intervals of the areas of attached trapezoids among curve's pixels. The integration of the intervals is measured by the function using Eq. 6-8. Then the summation of integrated areas under the curve represents the area of an irregular region. In this study, the function integrates all the nodule structures into a single cluster. The volume of this irregular shape (cluster), which consists of structures spread out into different levels, is equal to the summation of its cross-sectional structure areas multiplied by the thickness value using Eq.6-9. For each cluster, the obtained areas by trapezoidal integration of the cluster's surfaces, are summed and multiplied by the thickness of the CT image slice to produce the total volume for the nodule automatically. This process has been repeated for all clusters in the 3D-DBSCAN figure. Figure 6-14 and Figure 6-15 show the 3D-DBSCAN clustering of the classified objects to class 1 for the whole patient's stack of images with five nodules in the different slices. In the figures, aside list of numbered and marked clusters have a coloured cross along with their volume. Also, circular shape in the list indicates the noise structures of the classification outcomes in the figure. The coloured structures inside the figure represent the nodules and false positives of this scan. The five nodules and their characteristics were reported in the radiologists' report to be detected with location and volume value in the 3D-DBSCAN method. In general, the figures show the attached nodules in slices 60 and 64, as one cluster that has a single volume value is more evidence to confirm that the nodules are part of a single body. In addition, the nodules in slices 46, 60, 64 and12 that are identified in Figure 6-15, Figure 6-16 with their centre pixels match the experts' evaluation in their location and number.

Figure 6-17 is three parts and shows: (a) 3D-DBSCAN clusters from slices 60 to 64 that contain the attached nodules with their volume values, (b) the centre pixels of linked nodules are identified while (c) displays an example of attached nodules illustrating how 3D-DBSCAN clusters the class. The volume measuring results proved again the two nodules discussed in section 6.5, is still part of a single cluster and has a one-volume value. Figure 6-13 shows the pseudo-code of the proposed algorithm.

$$\text{Trapz function(Area)} = \frac{1}{2}\sum_{n=1}^{N}(x_{n+1} - x_n)[f(x_n) + f(x_{n+1})] \qquad 6\text{-}8$$

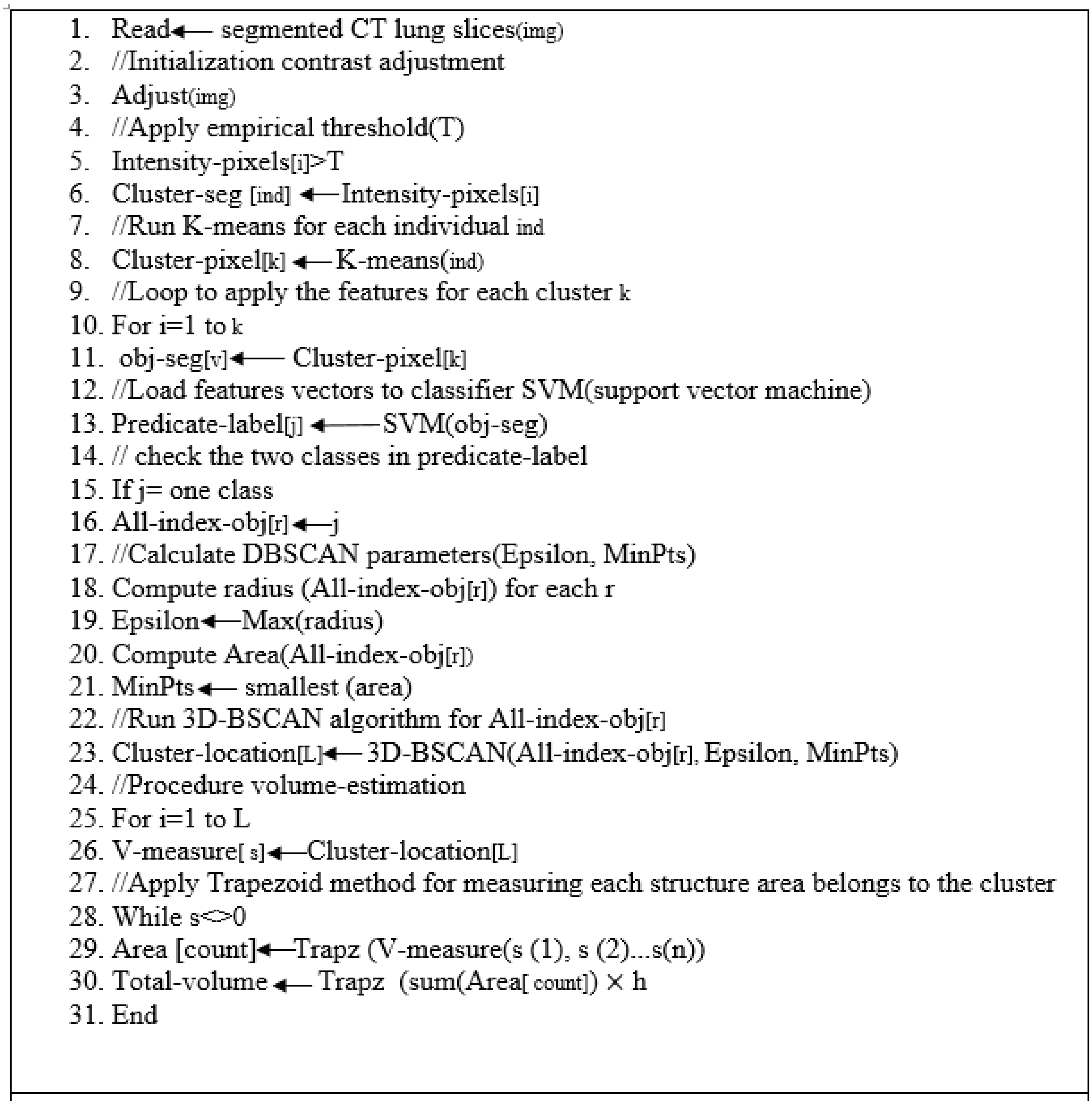
Where $(x_{n+1} - x_n)$ represents the spacing for each successive pair of pixels, N is the number of pixels of the region.

$$\text{Trapz function (Volume)} = \mathrm{h} \times (\mathrm{Area}_1 + \mathrm{Area}_2 + \mathrm{Area}_3 \ldots + \mathrm{Area}_\mathrm{n}) \qquad 6\text{-}9$$

Where h is the thickness of CT images slice.

6.8.4 3D-DBSCAN Performance in Automated Volume Measuring and Evaluation of 3D Clustering

The DBSCAN method has proved the efficiency in 3D clustering of nodules and false positive that have resulted from the classification stage and dealt with the noise in the best way. In the volume-measuring task, 3D-DBSCAN enabled to cluster the location of class 1 for processing each cluster individually. The ability to process the cluster allows Trapz function to be applied and measure the volume of nodules automatically. Furthermore, the 3D clustering of class 1 by the DBSCAN method within MinPts range and Epsilon distance parameters could display the extended nodules in consecutive layers and identify the connection between their structures. Also, the single volume value of the attached nodules in case study supports the merging of the nodules again into one body with a single volume. Moreover, the 3D-DBSCAN method could reduce the false positive rate through active clustering by excluding noise from the data. Furthermore, generally speaking, the variety of 3D-DBSCAN figure structures, in particular, the object patterns that reflect nodule shapes could likely be the nodules that are less than 3mm, of which only their number was given in the radiologists’ report.



```
Read ← segmented CT lung slices(img)
//Initialization contrast adjustment
Adjust(img)
//Apply empirical threshold(T)
Intensity-pixels[i]>T
Cluster-seg [ind] ←Intensity-pixels[i]
//Run K-means for each individual ind
Cluster-pixel[k] ←K-means(ind)
//Loop to apply the features for each cluster k
For i=1 to k
 obj-seg[v]← Cluster-pixel[k]
//Load features vectors to classifier SVM(support vector machine)
Predicate-label[j] ←SVM(obj-seg)
// check the two classes in predicate-label
If j= one class
All-index-obj[r]←j
//Calculate DBSCAN parameters(Epsilon, MinPts)
Compute radius (All-index-obj[r]) for each r
Epsilon←Max(radius)
Compute Area(All-index-obj[r])
MinPts← smallest (area)
//Run 3D-BSCAN algorithm for All-index-obj[r]
Cluster-location[L]←3D-BSCAN(All-index-obj[r], Epsilon, MinPts)
//Procedure volume-estimation
For i=1 to L
V-measure[ s]←Cluster-location[L]
//Apply Trapezoid method for measuring each structure area belongs to the cluster
While s<>0
Area [count]←Trapz (V-measure(s (1), s (2)...s(n))
Total-volume← Trapz  (sum(Area[ count]) × h
End
```

Figure 6-13: Pseudo-Code for Automatic Nodule Volume

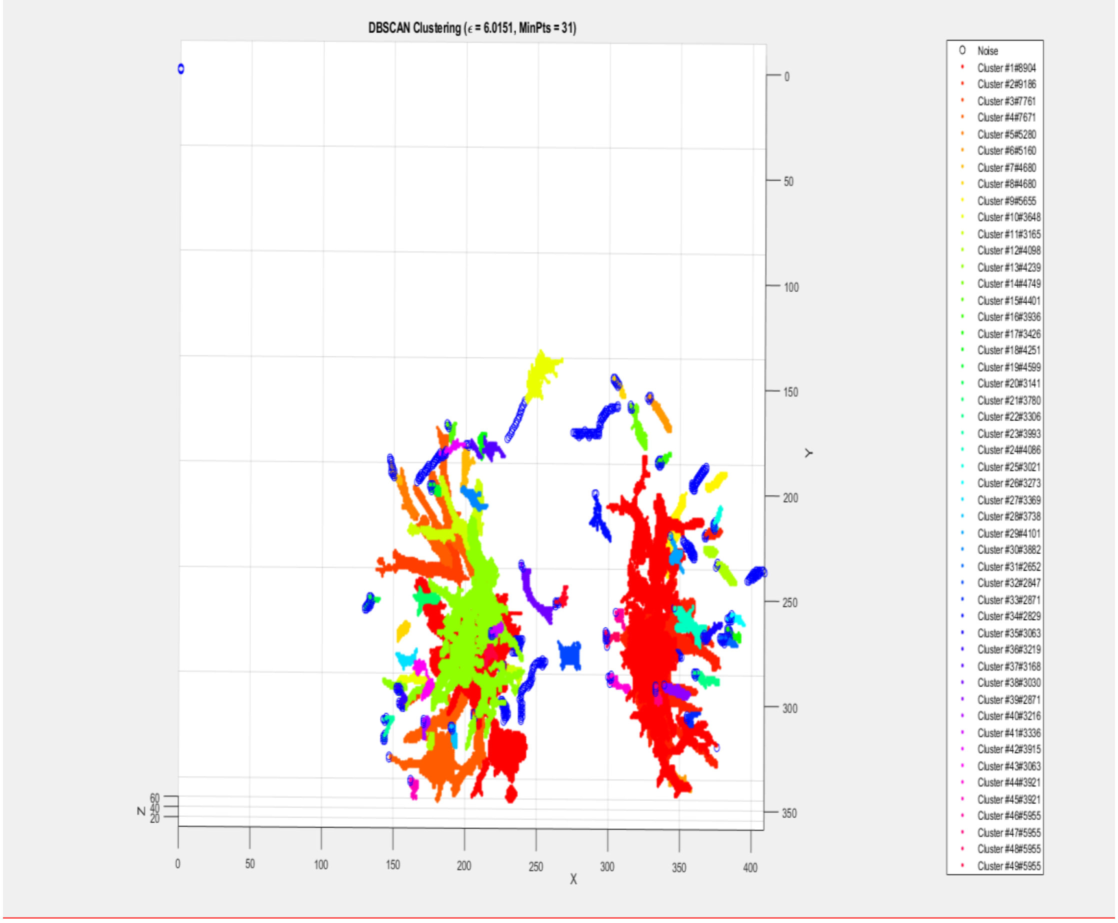


Figure 6-14: Nodule and False Positives Clustered by 3-DBSCAN and Measuring the Volume for each Cluster

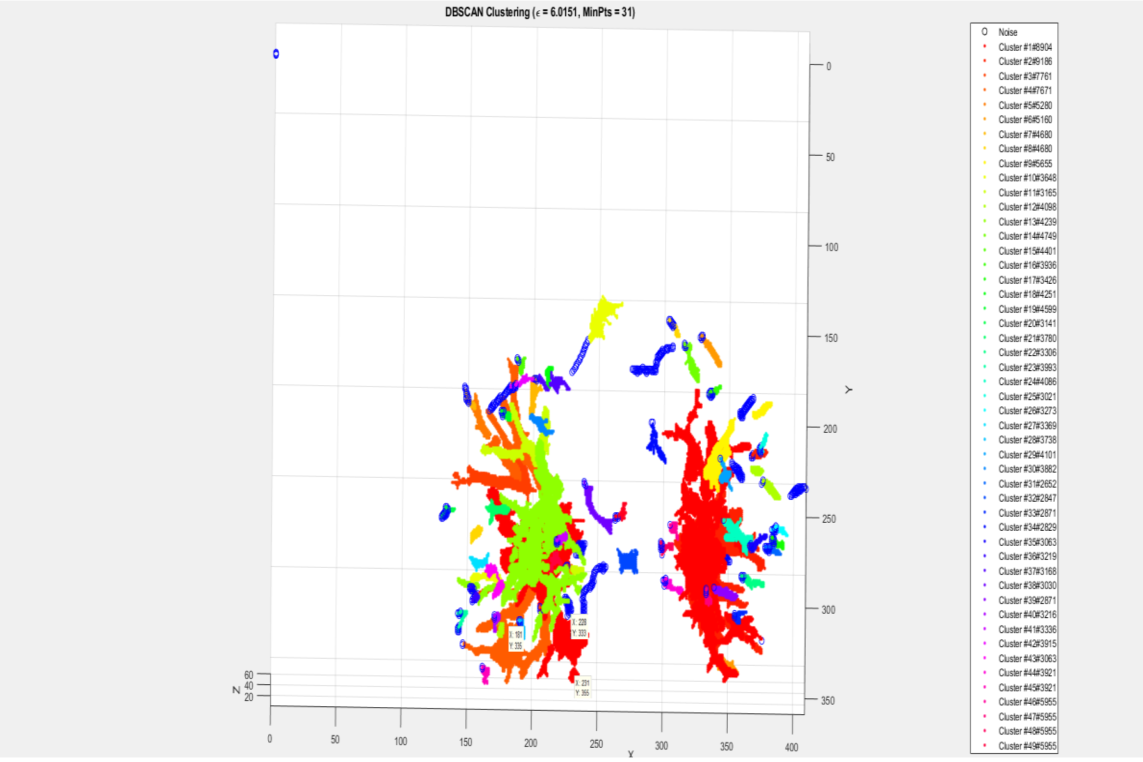


Figure 6-15: The Centre Pixels for Nodules in Slices 46, 60 and 64

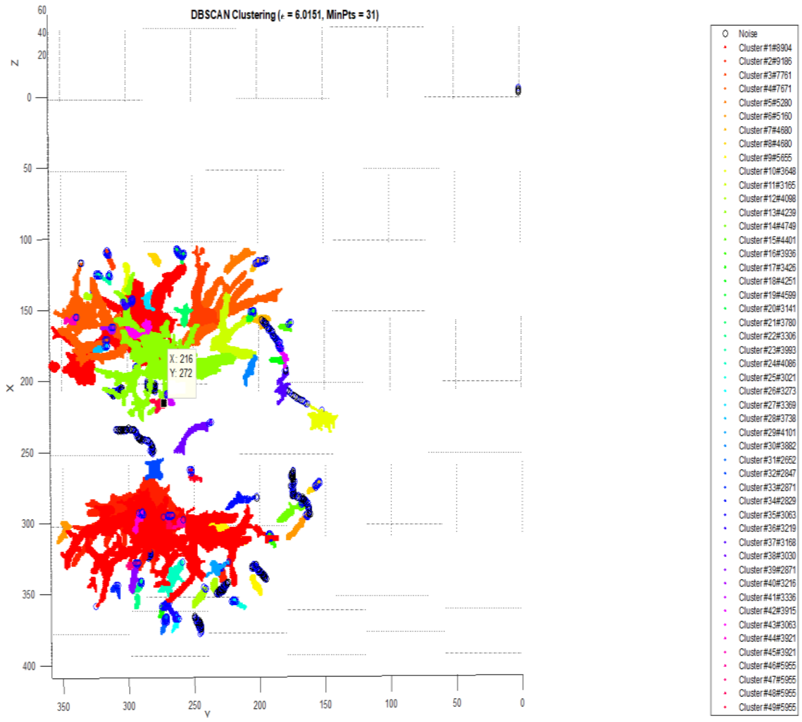
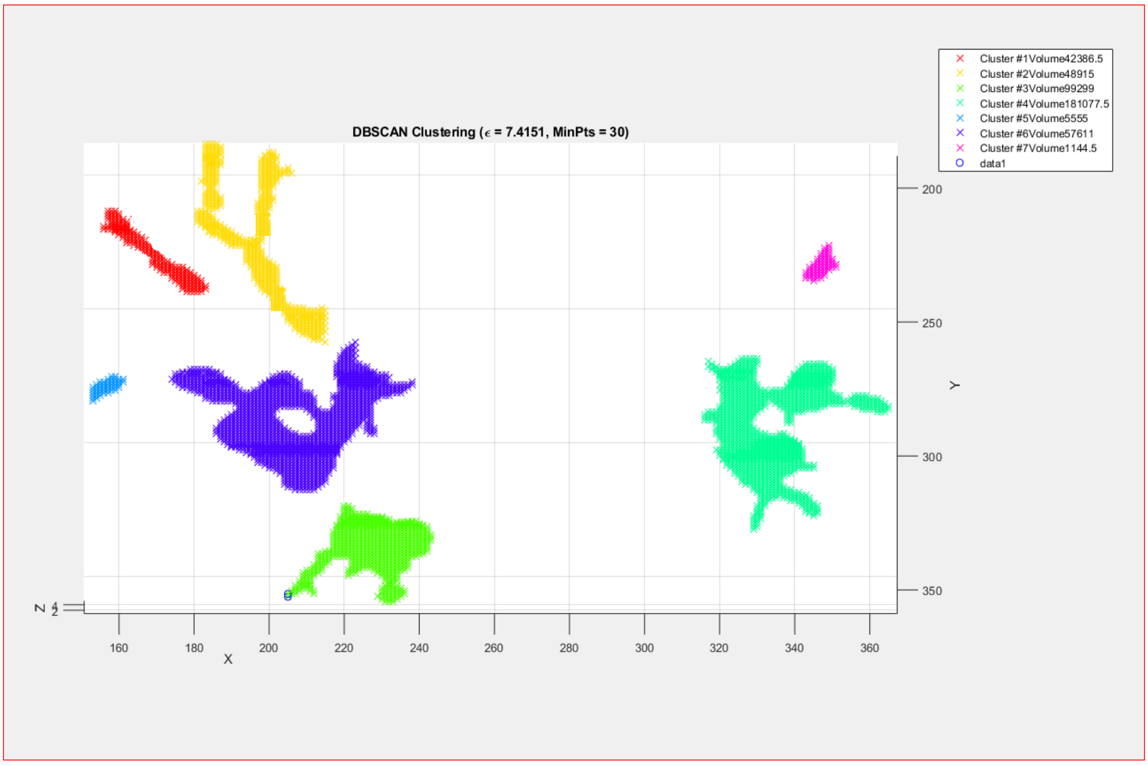
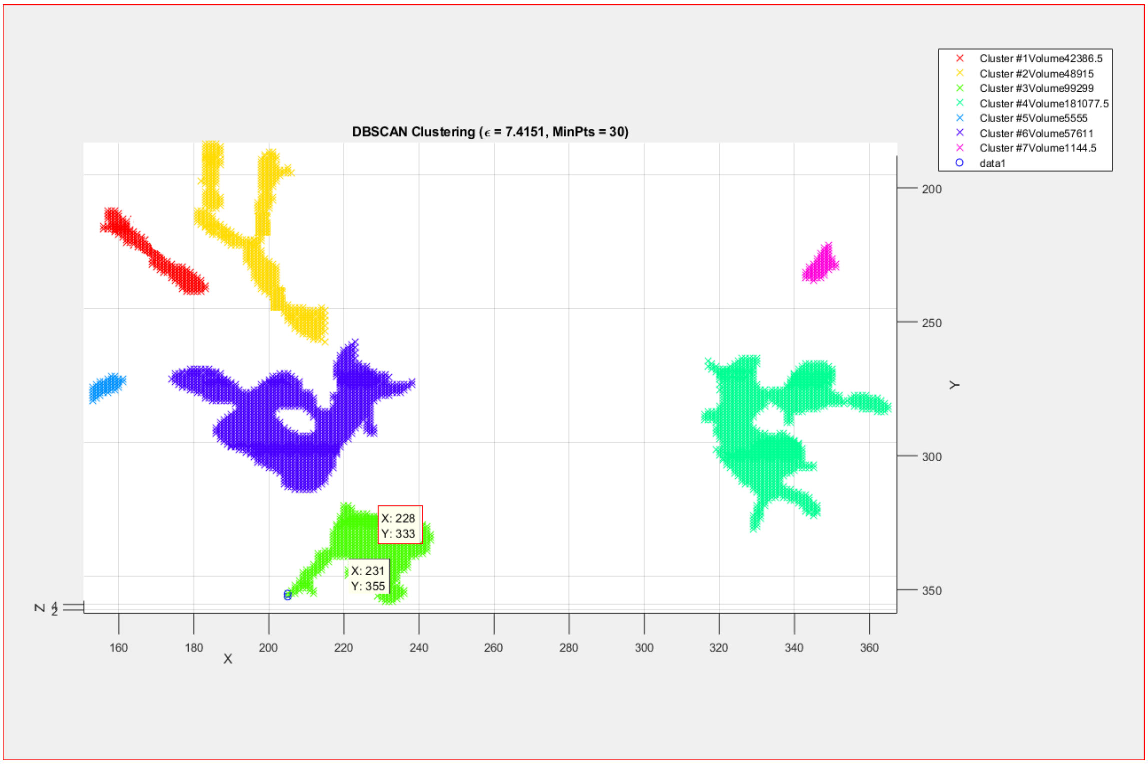


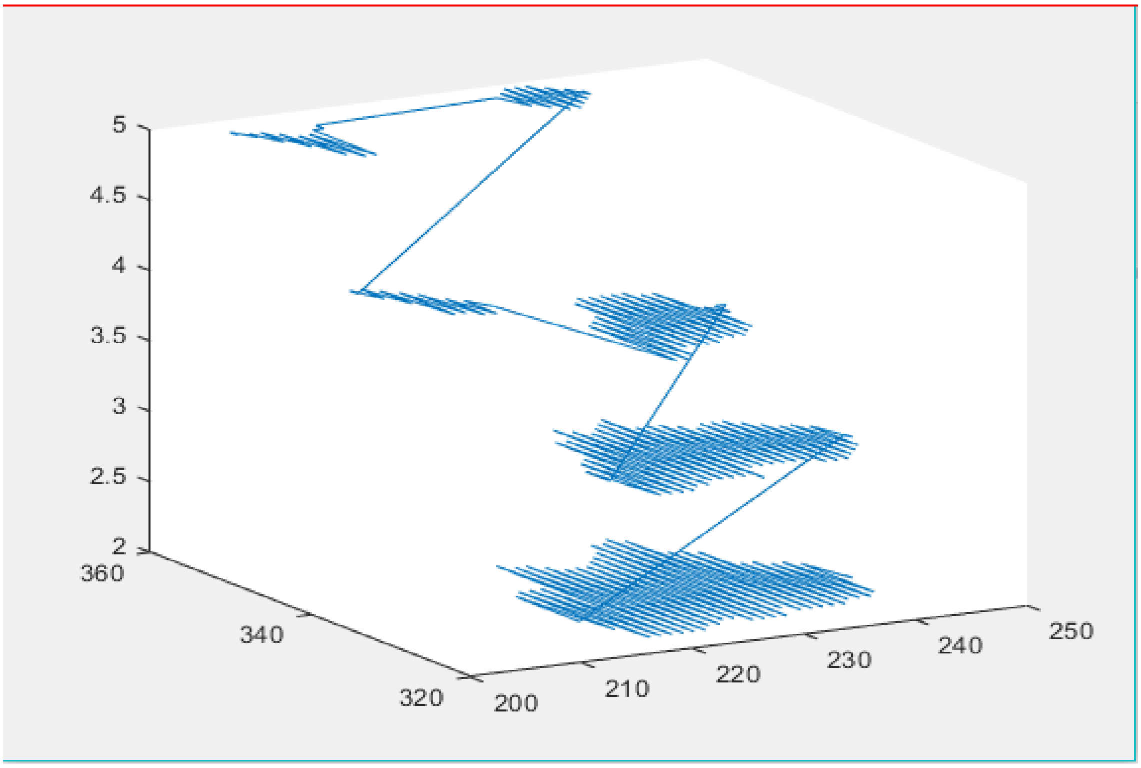
Figure 6-16: the Centre Pixel of Nodule in the Slice 12


DBSCAN Clustering (ϵ = 7.4151, MinPts = 30)
Cluster #1Volume42386.5
Cluster #2Volume48915
Cluster #3Volume99299
Cluster #4Volume181077.5
Cluster #5Volume5555
Cluster #6Volume57611
Cluster #7Volume1144.5
data1
X
160
180
200
220
240
260
280
300
320
340
360
Y
200
250
300
350
Z

-A-


DBSCAN Clustering (ϵ = 7.4151, MinPts = 30)
Cluster #1Volume42386.5
Cluster #2Volume48915
Cluster #3Volume99299
Cluster #4Volume181077.5
Cluster #5Volume5555
Cluster #6Volume57611
Cluster #7Volume1144.5
data1
X: 228
Y: 333
X: 231
Y: 355
X
160
180
200
220
240
260
280
300
320
340
360
Y
200
250
300
350
Z

-B-



-C-

Figure 6-17: A. 3D- DBSCAN Clustering of 2D Classification Results, B. Identifying the Attached Nodules Location in DBSCAN Figure, C. The Cluster of Attached Nodules are Clustered in B in a 3D View

6.9 Comparison

The comparison of the performance of our CAD system with approaches outcomes of other works is only significant if the other publications report about the performance measures in terms of detection accuracy, and a measured volume of the nodule. Three studies that show detection capability using the LIDC database in which reported detection meets the requirements of CADs, and are the closest to this study shown in Table 6-1.

The automated system proposed by Javed et al. (2016) was tested on lung CT scan images, and SVM classifier is applied to have the proposed system for which the sensitivity of large nodules classification is 93.8% analysed by 10-fold cross-validation. The overall sensitivity of the system is 91.65%, and accuracy is 96.22%.

An algorithm to detect the lung nodules was proposed by (Senthil Kumar et al., 2017). The algorithm manually measures the 3D nodule volume for 34 scans taken over a time span of 6, 9 or 12 months per patient.

In the work of (Oseas et al. 2014), an automated detection methodology is used, which is evaluated on 140 exams taken from the LIDC dataset of solitary nodules in CT lung scans. The nodules were described by shape and textures and classified by a support vector machine. The proposed system reaches a sensitivity of 85.91% and an accuracy of 97.55%.

The proposed method automatically detects 180 pulmonary nodules greater than 3mm and describes their characteristics (diameter, location and volume). The proposed system achieved an accuracy of 98% and sensitivity of 100% for four classifiers (Bayes, LR, MLP, and SVM) and a novel algorithm in 3D also to automatically measure the nodule volume and matches or is close to radiologists' assessments that were reported in reports.

Table 6-1: Other Works Compared with this Study in Terms of Accuracy and Measurements the Nodule Characteristics Automatically.

Papers	Pre-processing	Classifiers	Accuracy	Volume
(Oseas et al. 2014)	Not informed	SVM	97.55%	Not informed
(Javaid et al. 2016)	Lung segmentation by threshold and morphological	SVM	96.22%	Not informed
(Senthil Kumar, Ganesh, and Umamaheswari 2017)	Lung segmentation using Auto K-means and Morphological	Not informed	Not informed	Informed
Proposed Method	**Lung Segmentation by Fast Fuzzy C-Means and Morphological**	**SVM Bayesian Logistic Regression MLP**	**98%**	**Informed**

6.10 Summary

In this chapter, 3D clustering plays a vital role in describing, detecting and evaluating nodule patterns, types and sizes from CT lung images, which surpassed the 2D clustering in terms of accuracy and precision. The 3D clustering was able to show the nodule's extension and their attachments through the slices efficiently. The methods that are used in this chapter innovates a descriptive form that exhibits the nodule interior, complicated cases that are suspected have been missed by radiologists during the diagnosis. An illustration, identifying the actual two nodules that are reported by the radiologist, as one tumour by proposed 3D clustering method, has made. Also, the matching between the outcomes of proposed 3D clustering and other methods (3D-DBSCAN and 3D application), which are used for evaluation and comparison purposes, supported against each other to test the reliability of the outcome. Also, efficient automatic measuring of nodule size in both 2D (area) and 3D volume was conducted to improve the diagnosis performance and helps the radiologist to choose the suitable treatment by determining the nodule growth. Also, there is addressing of computer-aided detection systems requirements such as reduction of false positives by identifying the 3D-DBSCAN of noise into classification results that included the false positives. Furthermore, the speed in the detection and description of nodules is between 8-10 minutes for whole stack images and is better compared with the standard diagnosis at CT centres. All these processes were explained showing algorithm and flowchart and were supported by figures demonstrating the effectiveness of the method's performance regarding automation and accuracy. Furthermore, the method's ability in clustering the whole stack allows a possible detection of the nodules less than 3mm, which were, reported only their numbers in the radiologists' report for each case. The probability in detecting for these nodules has been expected within the rate of false positives, which resulted from the classification stage and are clustered by 3D-DBSCAN.

CHAPTER 7: SYSTEM DESIGN AND IMPLEMENTATION

7.1 Introduction

The main objective of this chapter is to offer and display some insight into the implementation phase of the research based on the methods and algorithms discussed in the previous chapters. Most of these image-processing methods have been employed to meet the requirements of the research needs and all techniques and algorithms presented are implemented using MATLAB (R2017b). The system could be implemented using any powerful programming language, but at this stage, MATLAB can achieve this purpose in an easier and faster way. Graphical user interfaces (GUI) have been designed as communication forms usable by both radiologists and users. It is worth mentioning that these interfaces make the system goal oriented without facing the difficulties of a design mechanism.

This chapter is organised as follows: Section presents the purpose of the development of the

7.2 Main Menu

This section reviews the essential stages for implementing the proposed system. These stages contain the loading and enhancing of the original image (CT lungs), the separating of the lungs from the background, the 2D clustering of segmented lungs, features extraction and classification of resulted objects, clustering of the nodule in 3D, and measuring the nodule volume. Figure 7-1 shows the system implementation stages, which are performed from the main menu.

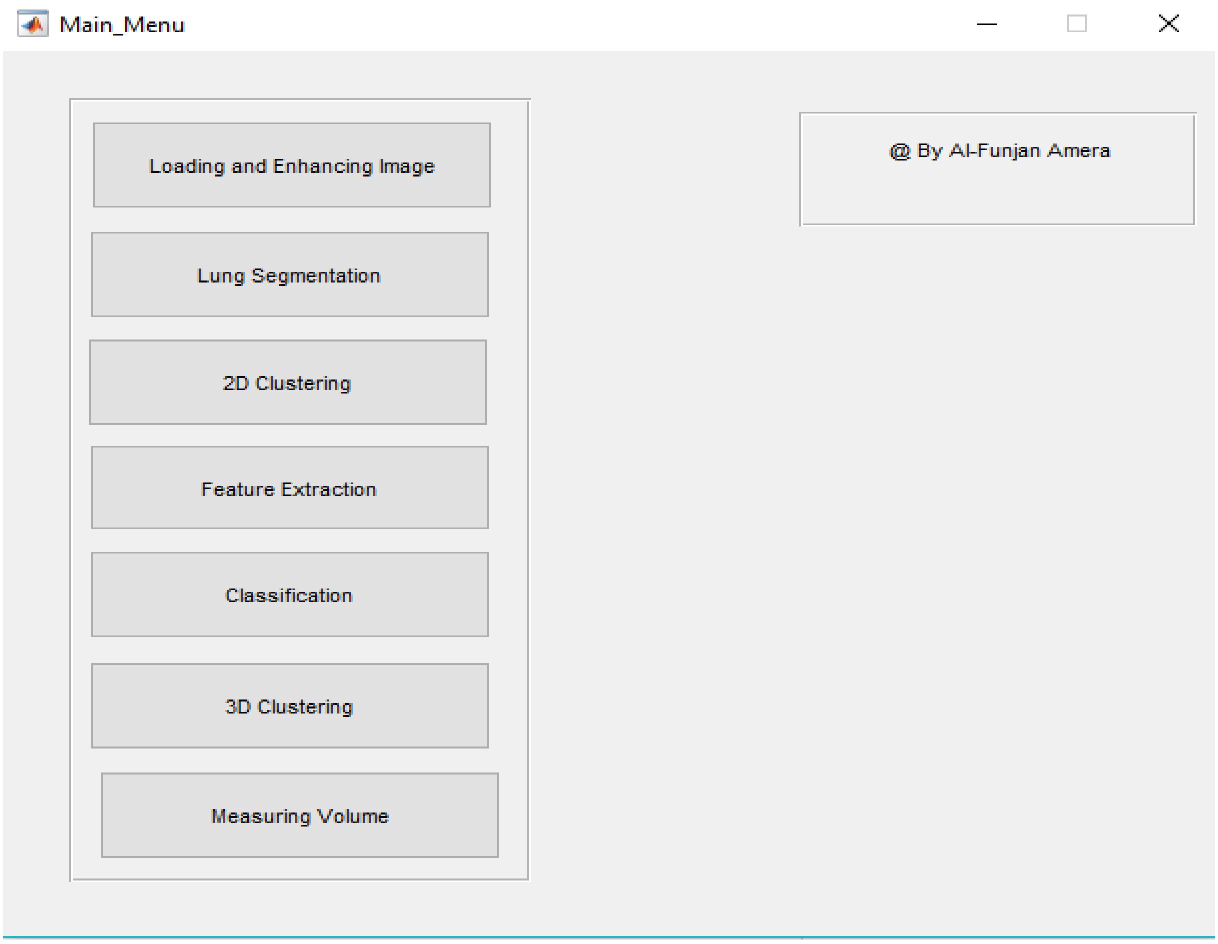


Figure 7-1: Main Menu of Proposed System

7.3 Loading and Enhancing Image

Loading and enhancing of the original image by Gaussian filter and Adjustment function processing is the first and important stage in the proposed system. In this stage, the original image is loaded and enhanced by enhancement processing that removes the noise and increases the image contrast as pre-processing of image. Figure 7-2 shows the implementation steps where the right image reviews the original image(CT lung image) while the middle and left images represent the same original image after applying Gaussian filter and Adjustment function respectively.

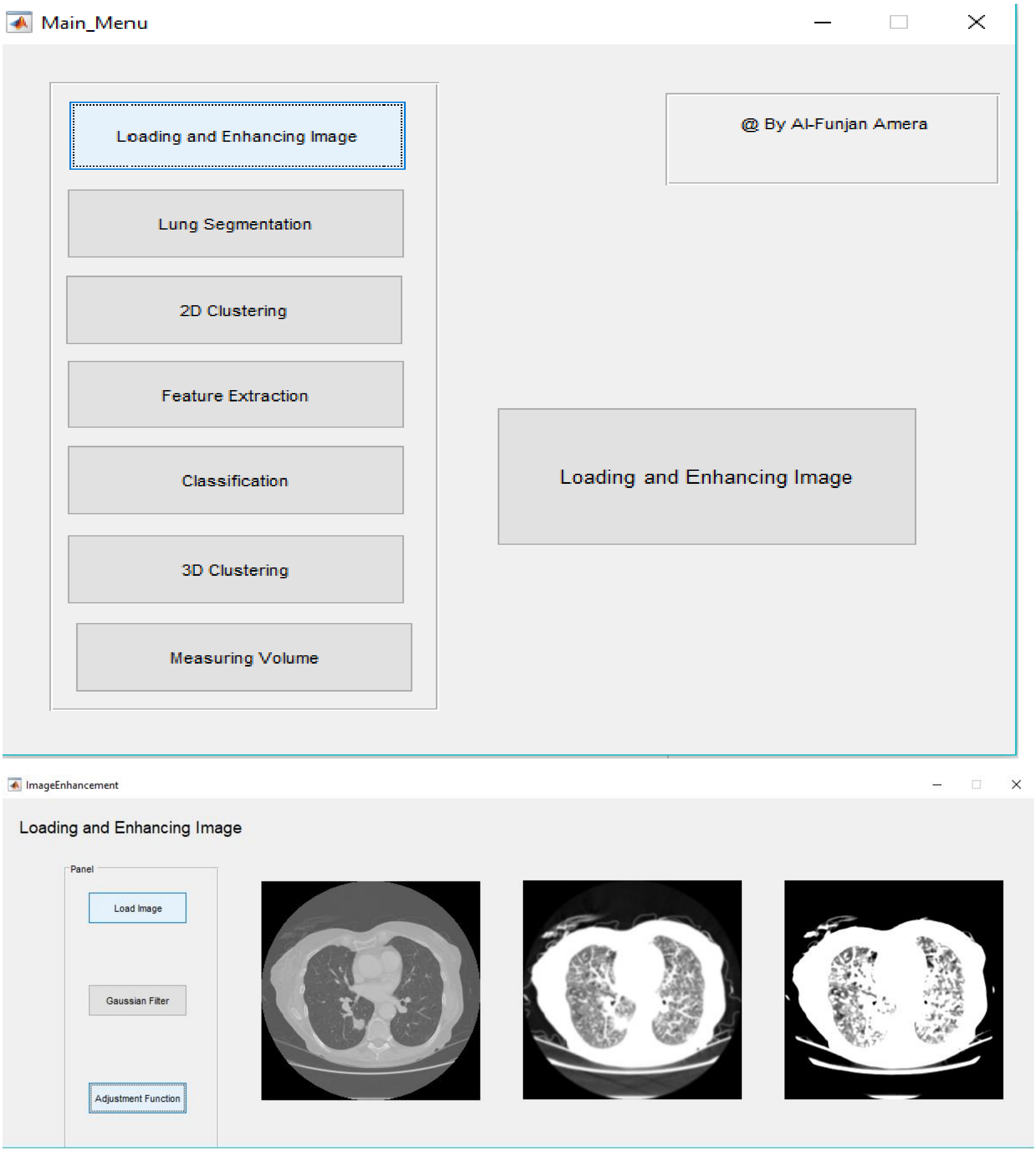


Figure 7-2: Loading and Enhancing Image

7.4 Lung Segmentation

The GUI of the lung segmentation stage displays the segmented lungs of CT image after enhancing the original image, which is done by applying the fast- fuzzy c-means method to segment the lungs from the background. The used method separates the foreground (lungs) and excludes the background as a step for reducing the research area of the nodule as shown in Figure 7-3.

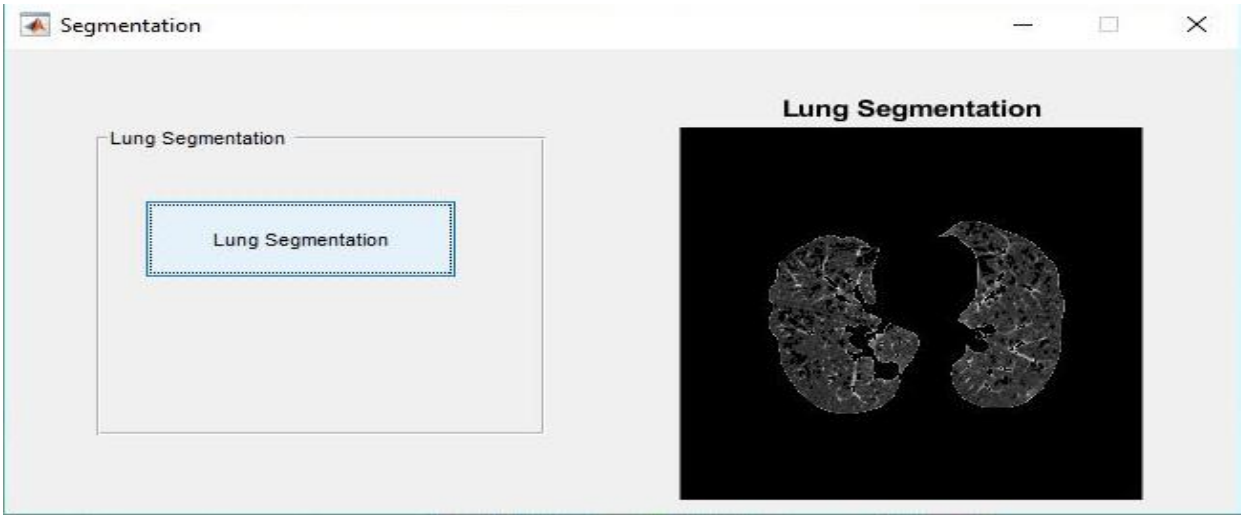
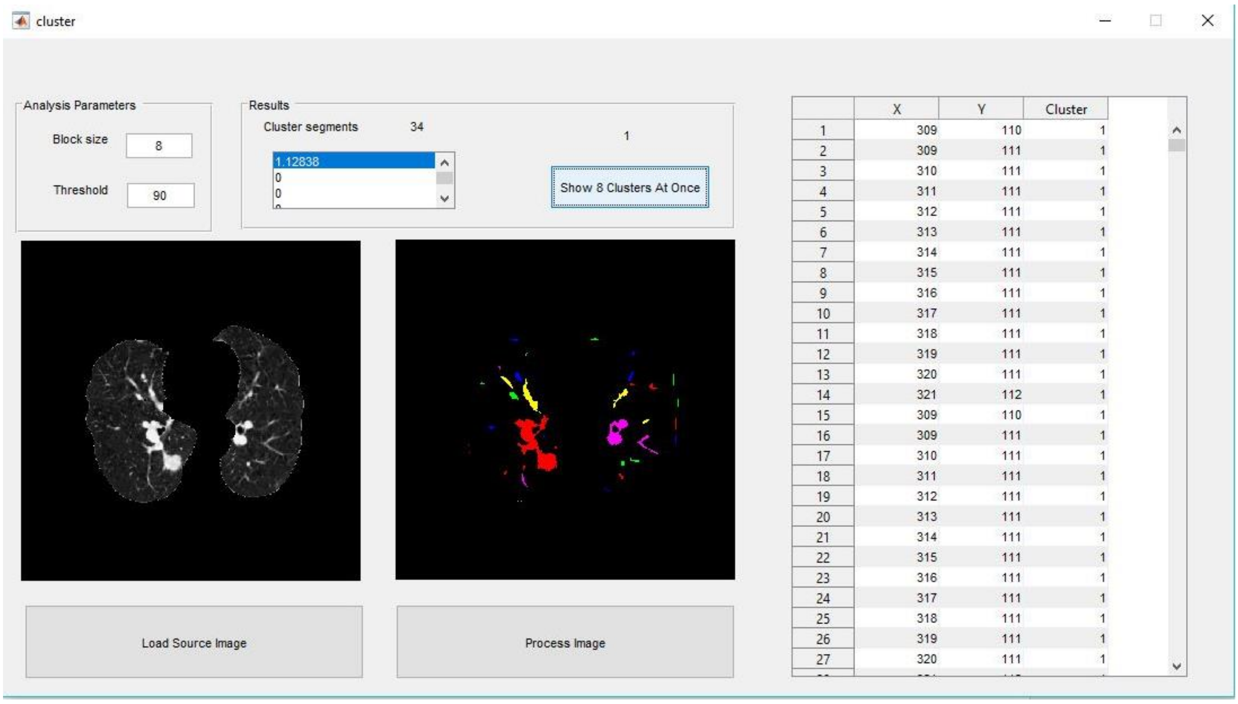


Figure 7-3: Lung Segmentation Stage

7.5 2D-Clustering

The 2D Clustering is an additional reduction of the region of interest (ROI), which is resulted from the previous stage based on the intensity values of the nodules. Consequently, the list of candidate of nodules for the highest intensity values are taken and neglect the lowest values, which are intuitively considered the normal tissue. It is called 2D clustering because of applying for each image individually. The GUI of this stage involves some experimental parameters such as threshold, blocks size to divide the image data, and has two options for loading and clustering the segmented lungs image (Load source image and Process image). In addition, there is a window of pixels location of each cluster, as shown in Figure 7-4.



Block size 8

Threshold 90

Cluster segments 34

1

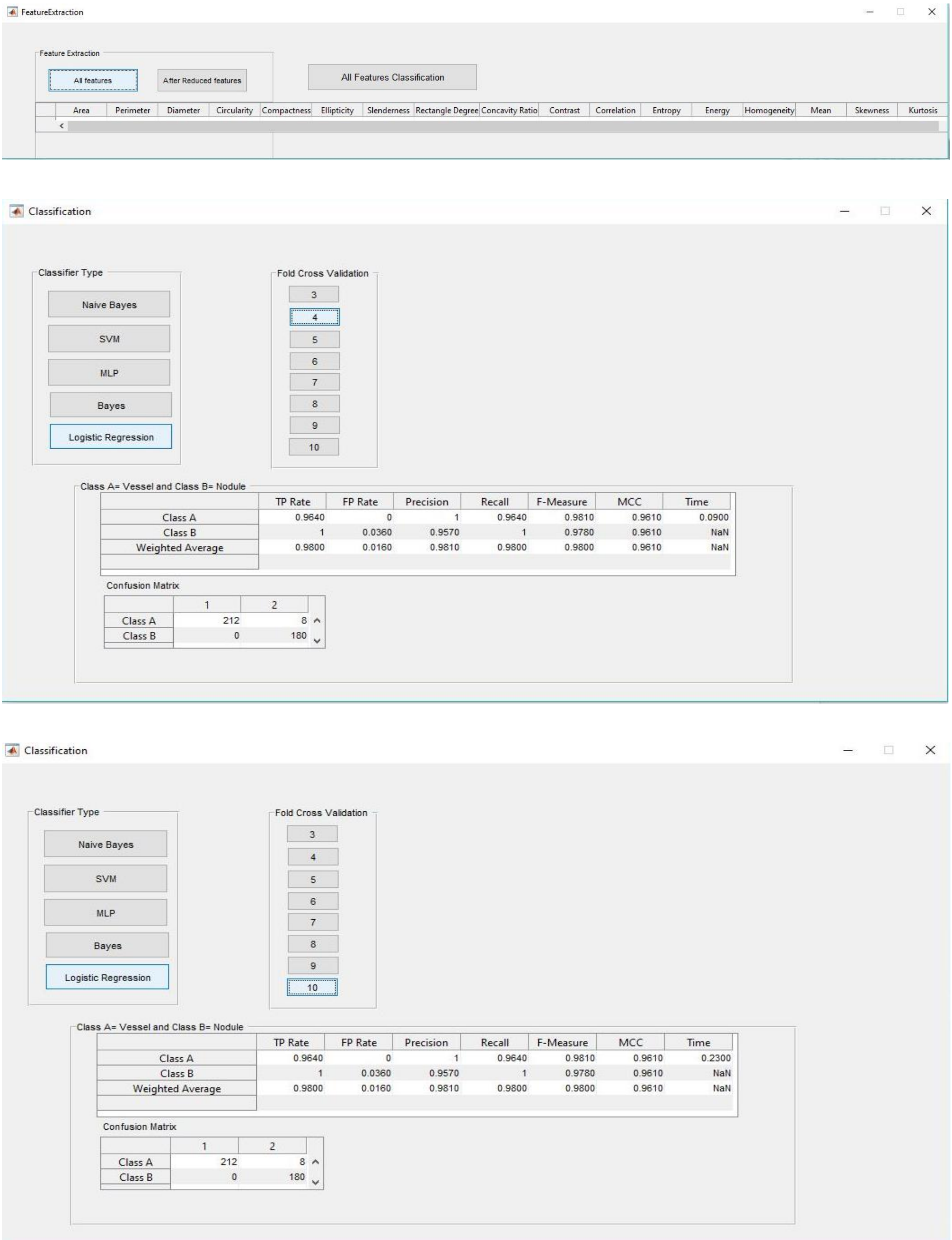
1.12838
0
0

	X	Y	Cluster
1	309	110	1
2	309	111	1
3	310	111	1
4	311	111	1
5	312	111	1
6	313	111	1
7	314	111	1
8	315	111	1
9	316	111	1
10	317	111	1
11	318	111	1
12	319	111	1
13	320	111	1
14	321	112	1
15	309	110	1
16	309	111	1
17	310	111	1
18	311	111	1
19	312	111	1
20	313	111	1
21	314	111	1
22	315	111	1
23	316	111	1
24	317	111	1
25	318	111	1
26	319	111	1
27	320	111	1

Figure 7-4: 2D-Clustering

7.6 Feature Extraction and Classification

The GUI of features extraction includes two options with a gridview of the names of the parameters that are based in the classification. The first option ‘All features’ displays evaluation measurements results of the classification for 19 features with execution time. The second option ‘After Reduced features’ excludes three features. Figure 7-5 shows an interface of results for logistic regression classifier in 4 and 10 folds-cross validation; these results have indicated the same evaluation measurements values of two options with a different execution time which is certainly less when the features are saved into 16 features only which are illustrated by Figure 7-6. GUI of classification results shows many classifiers, which are implemented with the 8-folds start from 3 fold, such as Naive Bayes, Bayesian, MLP and SVM. In addition, the results reflect the evaluation measurements of confusion matrix; accuracy, sensitivity and specificity in the list. Also, the execution time is calculated for each classifier with 4 and10 folds before and after features reduction.



FeatureExtraction

Feature Extraction

	Area	Perimeter	Diameter	Circularity	Compactness	Ellipticity	Slenderness	Rectangle Degree	Concavity Ratio	Contrast	Correlation	Entropy	Energy	Homogeneity	Mean	Skewness	Kurtosis

Classification

Classifier Type

Fold Cross Validation

Class A= Vessel and Class B= Nodule

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	Time
Class A	0.9640	0	1	0.9640	0.9810	0.9610	0.0900
Class B	1	0.0360	0.9570	1	0.9780	0.9610	NaN
Weighted Average	0.9800	0.0160	0.9810	0.9800	0.9800	0.9610	NaN

Confusion Matrix

	1	2
Class A	212	8
Class B	0	180

Classification

Classifier Type

Fold Cross Validation

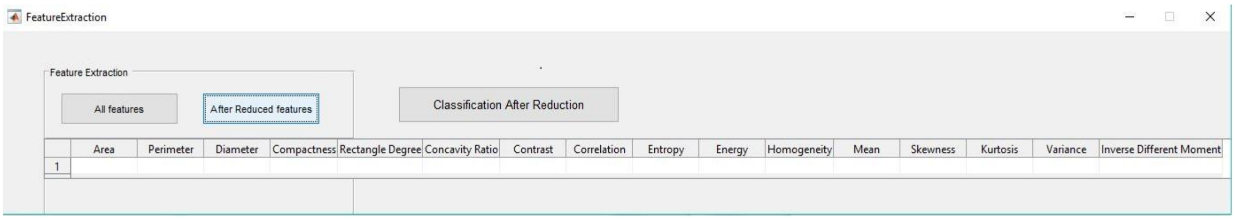
Class A= Vessel and Class B= Nodule

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	Time
Class A	0.9640	0	1	0.9640	0.9810	0.9610	0.2300
Class B	1	0.0360	0.9570	1	0.9780	0.9610	NaN
Weighted Average	0.9800	0.0160	0.9810	0.9800	0.9800	0.9610	NaN

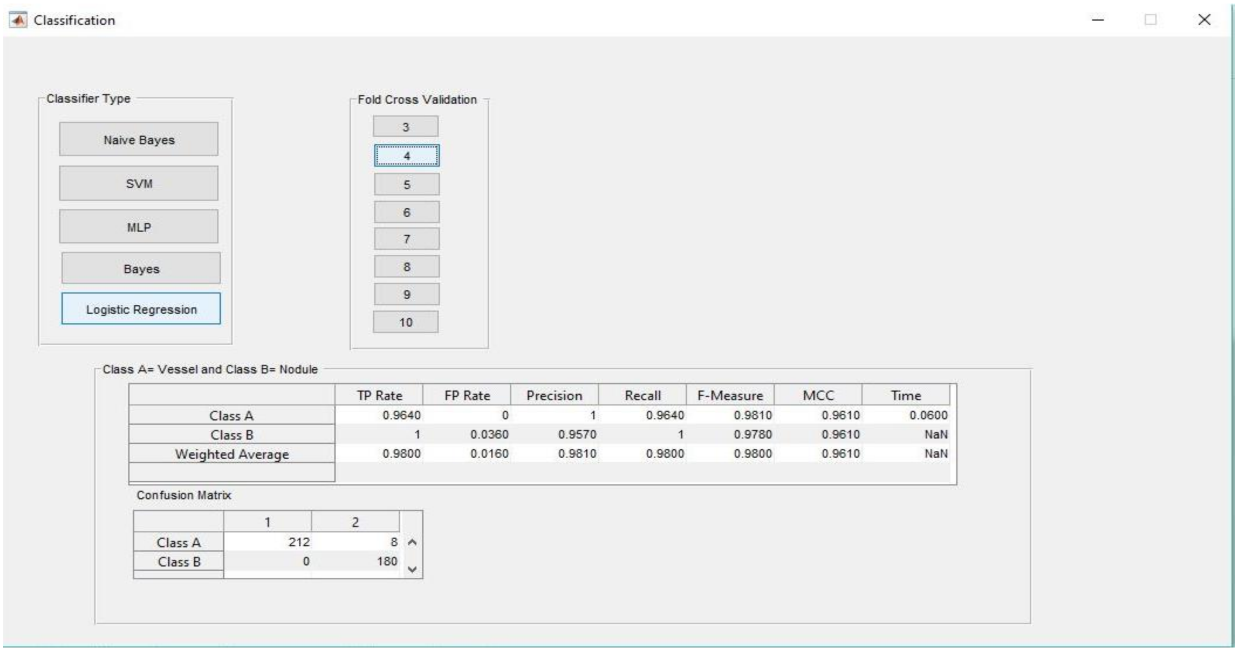
Confusion Matrix

	1	2
Class A	212	8
Class B	0	180

Figure 7-5: Results and Execution Time before Features Reduction for 4,10-Cross Validation



	Area	Perimeter	Diameter	Compactness	Rectangle Degree	Concavity Ratio	Contrast	Correlation	Entropy	Energy	Homogeneity	Mean	Skewness	Kurtosis	Variance	Inverse Different Moment
1																

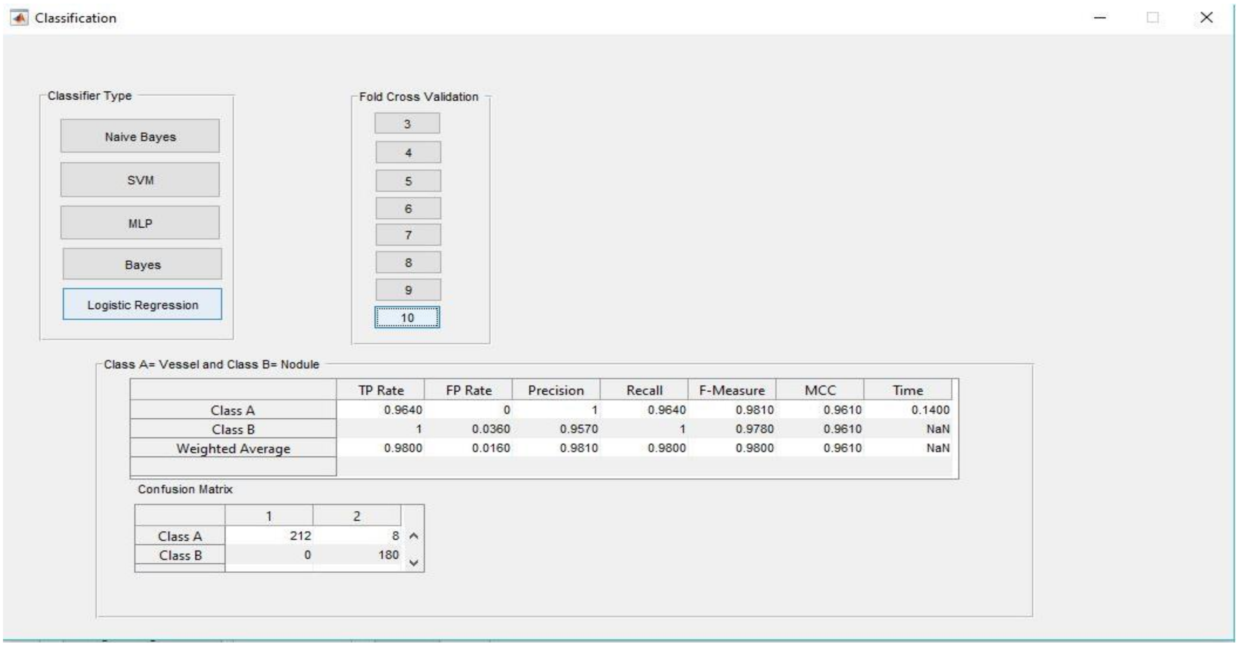


Class A= Vessel and Class B= Nodule

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	Time
Class A	0.9640	0	1	0.9640	0.9810	0.9610	0.0600
Class B	1	0.0360	0.9570	1	0.9780	0.9610	NaN
Weighted Average	0.9800	0.0160	0.9810	0.9800	0.9800	0.9610	NaN

Confusion Matrix

	1	2
Class A	212	8
Class B	0	180



Class A= Vessel and Class B= Nodule

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	Time
Class A	0.9640	0	1	0.9640	0.9810	0.9610	0.1400
Class B	1	0.0360	0.9570	1	0.9780	0.9610	NaN
Weighted Average	0.9800	0.0160	0.9810	0.9800	0.9800	0.9610	NaN

Confusion Matrix

	1	2
Class A	212	8
Class B	0	180

Figure 7-6: Results and Execution Time after Features Reduction for 4,10-Cross Validation

7.7 3D-Clustering

The extended nodules can be viewed by the 3D clustering of the stack of CT images to reform each nodule through multi-views. This provides a 3D vision of a nodule and describes many characteristics, such as the location and shape. In Figure 7-7, a range of images is chosen by the option ‘Load images range’ to prepare them for clustering by option ‘3D clustering’. Furthermore, two sides of clustered images in 3D are shown in two windows in the GUI.

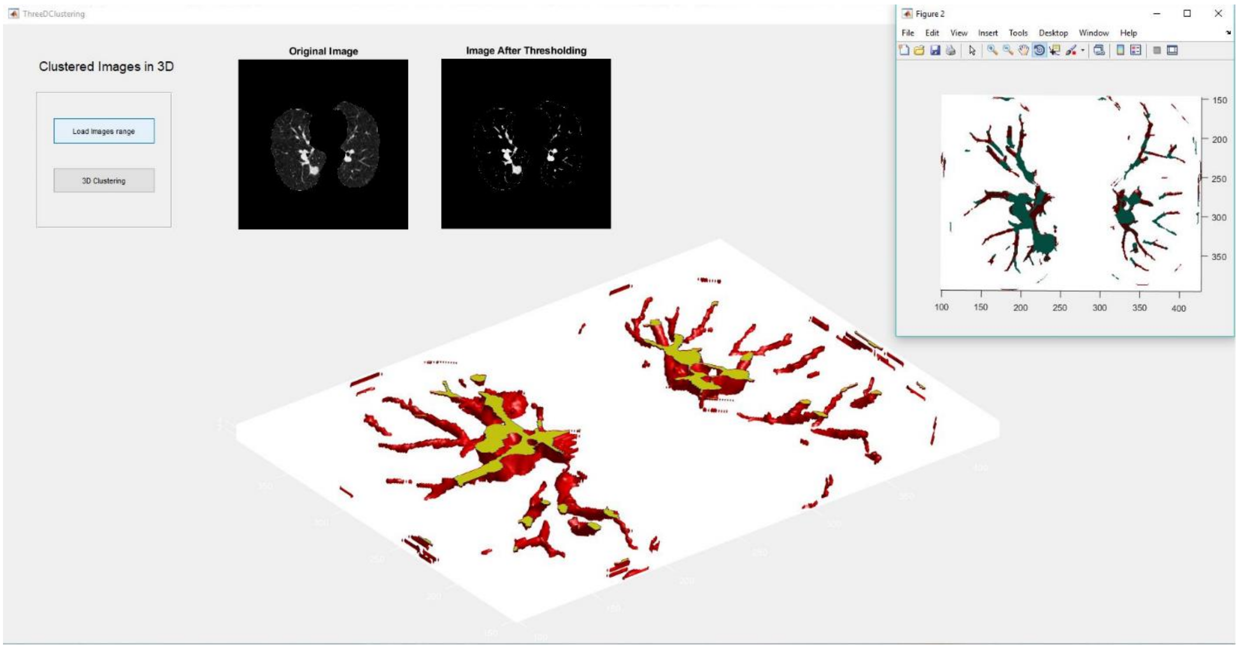


Figure 7-7: 3D-Clustering

7.8 Measuring Volume

In this stage, several images are selected by ‘Images Range’ option. Then pressing the ‘Measuring Volume’ option to calculate the volume of the structures that are clustered by the 3D-DBSCAN method, using the trapezoid method. In addition, Figure 7-8 reviews the body of the nodules as a surface figure while the volume values are stated in the side list.

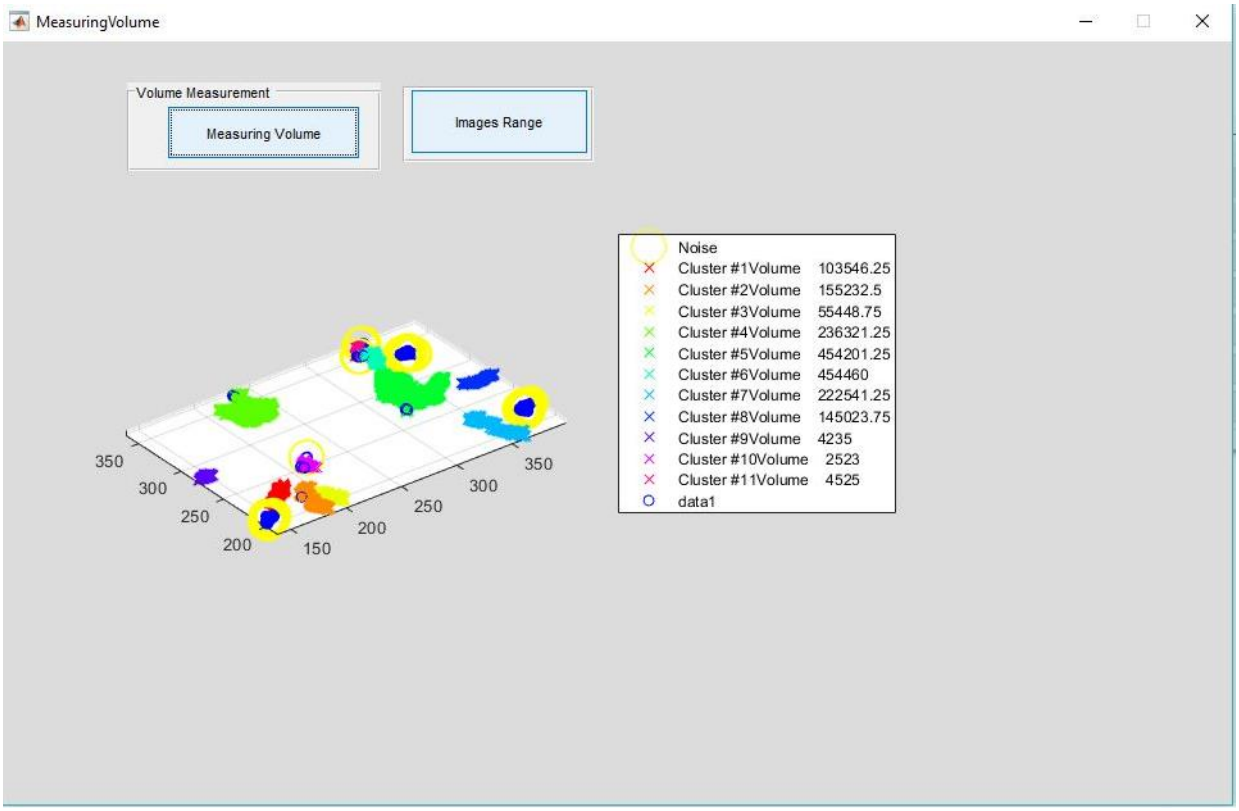


Figure 7-8: Measuring Volume Stage

7.9 Summary

The chapter demonstrates a collection of the proposed approaches and algorithms of the constructed system within a unified framework. To be more familiar for the user, the system has been divided into various stages that explain the processing and analysis of images in more details. Each stage is implemented into a platform package. The software has proved that the proposed system is applicable and friendly. Also, it is considered an opportunity to evaluate and test techniques used in the system of CT lung images.

CHAPTER 8: CONCLUSION AND FUTURE WORKS

8.1 Introduction

This chapter summarises the study that has been developed for the PhD programme. It discusses the obtained results as new achievements in the pulmonary nodules detection systems. Also, it included recommendations and suggestions for future works could be approached from this work. The methodology and techniques, which are performed to detect and describe the nodule in CT lung images, are reviewed.

8.2 Review of Methodology

The methodology applied research to process the aim through pursuing the activities of work, developing the clustering algorithms to segment, detect and measure the volume of the nodule within framework presents bettering understanding and actual development in this field. The fundamental knowledge of literature review identified the challenges and problem domain to determine the required tasks and suitable to develop the sought approach for achieving the target. This methodology manages a transition from common techniques, and simple application to development achievement based on a knowledge and understanding of the problem contribute to discovering the new approach for this study. Aim, objectives, and research question were distinguished in the research beginning, to form the research base, which is led by the iterative development of work stages. Also, it is supported and evaluated by the framework and the appropriate tools. The evaluation tools continually assess the performance to be compared with other works.

A novel approach to constructing a fully automated system is successfully performed in this work. The developed methodology is validated, evaluated and applied to 180 nodules from LIDC-IDRI dataset. It is essential to assert that the LIDC-IDRI is hard complex and various database. This dataset is established by a partnership between the Image Database Resource Initiative and the Lung Image Database Consortium. Therefore, it has diverse exams were acquired from different tomography protocols which increase the detection difficulties of computer-aided detection. The following characteristics may hinder the active detection: the various contrast between tomography devices, different spacing among the slices and annotation that is made by radiologists where every factor affects a stage of the computer-aided detection system. The successful results emphasise that the work's methodology has an efficient performance. With this system, proper segmentation of lung of image background by applying fast fuzzy c-means in pre-processing stage and evaluated with the adequate metric that concentrates on the lungs edge at segmentation, which is often attached to a nodule. Besides, a new algorithm based the image properties to cluster the irregular patterns of vessels and nodules in CT lung images are implemented in this study. In the classification stage, the highest rates of accuracies are achieved by four classifiers (Bayesian, Logistic Regression, Multilayer Perceptron and Support Vector Machine) are 98% within two folds of Cross Validation (4 and10 fold). The system tests 180 nodules are taken from 40 exams assessed by four experts have reported characteristics (volume, diameter, and centre pixel and slice number) for the nodules that are greater than 3mm and just the number of nodules that are less than 3mm.

The methodology succeeds in identifying a complicated case, could be invisible in 2D images, and was incorrectly diagnosed by the experts. This case is detected through efficient 3D clustering by K-means. 3D clustering is evaluated by another method for clustering and application to a 3D plot. An accepted description is received from 3D methods enable the system to measure the nodule volume automatically for the first time. Furthermore, the 3D method offers the detection potential of nodules that are less than 3mm.

Generally, many challenges are addressed in this work, such as full automation in the detection nodule and measuring volume of a tumour, enhancement in segmentation and high accuracy in the classification. The system is beneficial to the radiologist and ensures efficient reading of scan and early detection of lung cancer (Narad et al., 2015).

8.3 Conclusions

The proposed system has demonstrated its efficiency in the early detection and description of pulmonary nodules, which have a size greater than 3mm in computed tomography images (CT). The conclusions of this work can be summarised as follows.

- Fast fuzzy c-mean could succeed in segmenting the lungs from the background to be evaluated by qualitative (radiologists) and quantitative (Hausdorff metric) measures.
- One of the most complicated problems in this field is a homogeneity characteristic between the nodules and vessels. Hence, the IIBC has proved to be an efficient method for isolating vessels from nodules accurately.
- The high accuracy of the four machine learning classifiers, which are applied in the classification stage, has confirmed the precision separation results of nodules and vessels in the previous stage.
- The 3D clustering has given a clear and deep view of the nodules, which leads to discovering the complicated cases that are missed by radiologists.
- The proposed 3D-DBSCAN method has emphasized the evaluation of standard 3D clustering, which is used to describe the shape and extension of nodules, and has made possible with trapezoid method to measure the nodule volume automatically. This is considered the key contribution of this study.

8.4 Research Limitations

This research adopts a developed approach has taken the aim and objectives standpoint to address the problems of the nodule description and detection systems. Identifying of problem domain guides the direction of development of techniques from 2D to 3D to provide further information about the nodule regarding detection and description accuracy. Against this, there are some limitations encountered the research such as the complex type of noise which accompanies the CT image because of the patient's motion through the scan. Moreover, this hinders pre-processing tasks in the lung segmentation and avoids the volume-measuring algorithm to validate more nodules that is because the algorithm needs to cluster the whole images stack. The insufficient radiologists' assessment of nodules that are less 3mm leads to losing a chance to detect them by the 3D algorithm.

8.5 Future Works

In this study, several recommendations could be interesting points for other researchers in this area.

1- Implementing the proposed system on other organs of the body such as (brain, liver and kidney)

2- Employing the fuzzy system to generate membership function in 2D clustering which computes the empirical parameters automatically instead of experimental.

3- Using the ensemble method in the classification stage by relying on the voting decision of multi classifiers as an alternative tool to the five classifiers that are implemented dependently to ensure a higher accuracy and to mitigate the impact of the weak classifier (naïve Bayes) on reduction of false positive.

4- Applying the deep learning neural network with multilayers by including each stage in one layer or more. This can deal with huge datasets of CT images and reduce the number of classifiers into one classifier and it can deal with pixels directly without needing to extract features.

5- Evaluating nodule volume measurement using a phantom scale which represents an excellent tool of ground truth for testing the system activity as quantitative and qualitative evaluation (Ravenel et al., 2008) and (Prionas et al., 2011).

8.6 Summary

This chapter mentions reviewing methodology, research limitations, conclusions and achievements the study and recommendations summarise future works that contribute to this work development. The future perspectives support constructing integrated detection system corresponding computer-aided detection requirements to be based on the practice.

References

Abakar, Khalid A.A., and Chongwen Yu. 2014. "Performance of SVM Based on PUK Kernel in Comparison to SVM Based on RBF Kernel in Prediction of Yarn Tenacity." *Indian Journal of Fibre and Textile Research* 39 (1):55–59. https://doi.org/10.7763/IJIMT.2013.V4.427.

Ahmed, Mohamed N, Sameh M Yamany, Nevin Mohamed, Aly A Farag, and Thomas Moriarty. 2002. "A Modified Fuzzy C-Means Algorithm for Bias Field Estimation and Segmentation of MRI Data." *IEEE Transactions on Medical Imaging* 21 (3). IEEE:193–99.

Al-fahoum, Amjed S, Eslam B Jaber, and Mohammed A Al-jarrah. 2014. "Automated Detection of Lung Cancer Using Statistical and Morphological Image Processing Techniques." *Journal of Biomedical Graphics and Computing* 4 (2):33–42. https://doi.org/10.5430/jbgc.v4n2p33.

AL-Hashimi, M Y, and XiangJun Wang. 2013. "Trend of Leukemia in Ninawa/Iraq." *Clinical and Experimental Medical Sciences* 1 (8):353–62.

Al-Rahim, Yousif A, M B Ch, and / Cm. 2007. "Lung Cancer in a Sample of Iraqi Patients." *Al-Kindy Col Med J* 44 (11):53–59.

Al-Waeli, Ali Majeed Hasan. 2017. "AN AUTOMATED SYSTEM FOR THE CLASSIFICATION AND SEGMENTATION OF BRAIN TUMOURS IN MRI IMAGES BASED ON THE MODIFIED GREY LEVEL CO-OCCURRENCE MATRIX." University of Salford.

Anand, Rajaraman, and D Ullman Jeffrey. 2011. "Mining of Massive Datasets." *2011-01-03]. Http://Infolab. Stanford*. https://doi.org/10.1017/CBO9781139924801.

Annangi, Pavan, Sheshadri Thiruvenkadam, A Raja, Hao Xu, XiWen Sun, and Ling Mao. 2010. "A Region Based Active Contour Method for X-Ray Lung Segmentation Using Prior Shape and Low Level Features." In *Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium On*, 892–95.

Argenti, Fabrizio, Luciano Alparone, and Giuliano Benelli. 1990. “Fast Algorithms for Texture Analysis Using Co-Occurrence Matrices.” In *IEE Proceedings F (Radar and Signal Processing)*, 137:443–48.

Armato, Samuel G, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, et al. 2011. “The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans.” *Medical Physics* 38 (2). Wiley Online Library:915–31.

Ayash, Eng Mohannad M. 2014. “Research Methodologies in Computer Science and Information Systems.” *Retrieved November* 28:2014.

Badura, Pawel, and Ewa Pietka. 2014. “Soft Computing Approach to 3D Lung Nodule Segmentation in CT.” *Computers in Biology and Medicine* 53. Elsevier:230–43.

Bankman, Isaac N, and Serban Morcovescu. 2002. “Handbook of Medical Imaging. Processing and Analysis.” *Medical Physics* 29 (1). Wiley Online Library:107.

Biniaz, Abbas, Ataollah Abbasi, and Mosa Shamsi. 2012. “Fast FCM Algorithm for Brain MR Image Segmentation.” In *6th International Conference on Fuzzy Information and Engineering*, 1–8.

Birry, Rania Ahmed Kadry. 2013. “Automated Classification in Digital Images of Osteogenic Differentiated Stem Cells.” University of Salford.

Blume, H, and B M Hemminger. 1997. “Image Presentation in Digital Radiology: Perspectives on the Emerging DICOM Display Function Standard and Its Application.” *Radiographics* 17 (3):769–77. https://doi.org/10.1148/radiographics.17.3.9153711.

Burger, W., and M.J. Burge. 2009. “Principles of Digital Image Processing.” *Undergraduate Topics in Computer Science*. https://doi.org/10.1007/978-1-84800-191-6.

Campadelli, Paola, Elena Casiraghi, and Diana Artioli. 2006. “A Fully Automated Method for Lung Nodule Detection from Postero-Anterior Chest Radiographs.” *IEEE Transactions on Medical Imaging* 25 (12). IEEE:1588–1603.

Cascio, Donato, Rosario Magro, Francesco Fauci, Marius Iacomi, and Giuseppe Raso. 2012. “Automatic Detection of Lung Nodules in CT Datasets Based on Stable 3D Mass--Spring Models.” *Computers in Biology and Medicine* 42 (11). Elsevier:1098–1109.

Celebi, M. Emre, Hassan A. Kingravi, and Patricio A. Vela. 2013. “A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm.” *Expert Systems with Applications* 40 (1). Elsevier Ltd:200–210. https://doi.org/10.1016/j.eswa.2012.07.021.

Chaudhuri, Bidyut Baran, and Nirupam Sarkar. 1995. “Texture Segmentation Using Fractal Dimension.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 17 (1). IEEE:72–77.

Chen, Chi-hau. 2015. *Handbook of Pattern Recognition and Computer Vision*. World Scientific.

Chen, Hui, Jing Zhang, Yan Xu, Budong Chen, and Kuan Zhang. 2012. “Performance Comparison of Artificial Neural Network and Logistic Regression Model for Differentiating Lung Nodules on CT Scans.” *Expert Systems with Applications* 39 (13). Elsevier Ltd:11503–9. https://doi.org/10.1016/j.eswa.2012.04.001.

Chen, Kan, Bin Li, Lian-Fang Tian, and Jing Zhang. 2013. “Segmentation of Pulmonary Nodules Using Fuzzy Clustering Based on Coefficient of Curvature.” In *Image and Graphics (ICIG), 2013 Seventh International Conference On*, 225–30.

Choi, Wook-Jin, and Tae-Sun Choi. 2014. “Automated Pulmonary Nodule Detection Based on Three-Dimensional Shape-Based Feature Descriptor.” *Computer Methods and Programs in Biomedicine* 113 (1). Elsevier:37–54.

Clark, Kenneth, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, et al. 2013. “The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository.” *Journal of Digital Imaging* 26 (6).

Springer:1045–57.

Daszykowski, M., and B. Walczak. 2010. “Density-Based Clustering Methods.” *Comprehensive Chemometrics* 2:635–54. https://doi.org/10.1016/B978-044452701-1.00067-3.

Davies, Henri Gwyn, and Gordon Allen Hicks. 1981. *Mathematics II*. Springer.

Deserno, Thomas M. 2011. *Fundamentals of Biomedical Image Processing*. https://doi.org/10.1007/978-3-642-15816-2.

Domingos, Pedro, and Michael Pazzani. 1997. “On the Optimality of the Simple Bayesian Classifier under Zero-One Loss.” *Machine Learning* 29 (2–3). Springer:103–30.

Dougherty, Edward R, and Roberto A Lotufo. 2003. *Hands-on Morphological Image Processing*. Vol. 59. SPIE press.

Dougherty, Geoff. 2009. *Digital Image Processing for Medical Applications*. Cambridge University Press.

Draper, Norman R, and Harry Smith. 2014. *Applied Regression Analysis*. Vol. 326. John Wiley & Sons.

Dubitzky, Werner, Martin Granzow, and Daniel P Berrar. 2007. *Fundamentals of Data Mining in Genomics and Proteomics*. Springer Science & Business Media.

Eadie, Leila H, Paul Taylor, and Adam P Gibson. 2012. “A Systematic Review of Computer-Assisted Diagnosis in Diagnostic Cancer Imaging.” *European Journal of Radiology* 81 (1). Elsevier:e70--e76.

El-Baz, Ayman, Ahmed Elnakib, Mohamed Abou El-Ghar, Georgy Gimel’farb, Robert Falk, and Aly Farag. 2013. “Automatic Detection of 2D and 3D Lung Nodules in Chest Spiral CT Scans.” *International Journal of Biomedical Imaging* 2013. Hindawi.

Engel, K., R. Westermann, and T. Ertl. 1999. “Isosurface Extraction Techniques for Web-Based Volume Visualization.” *Proceedings Visualization ’99 (Cat. No.99CB37067)* Vi:139–47. https://doi.org/10.1109/VISUAL.1999.809878.

Eskandarian, Parinaz, and Jamshid Bagherzadeh. 2015. “Computer-Aided Detection of Pulmonary Nodules Based on SVM in Thoracic CT Images.” In *Information and Knowledge Technology (IKT), 2015 7th Conference On*, 1–6.

Farag, Amal, Hossam Abdelmunim, James Graham, Aly A Farag, Cambron Carter, Salwa Elshazly, Mohamed S El-Mogy, Sabry El-Mogy, and Robert Falk. 2012. “An AAM-Based Detection Approach of Lung Nodules from LDCT Scans.” In *Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium On*, 1040–43.

Fetita, Catalin I, Francoise Preteux, Catherine Beigelman-Aubry, and Philippe Grenier. 2003. “3D Automated Lung Nodule Segmentation in HRCT.” In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 626–34.

Fu, King-Sun, and J K Mui. 1981. “A Survey on Image Segmentation.” *Pattern Recognition* 13 (1). Elsevier:3–16.

Gevers, Theo, and A W N Smeulders. 1997. “Combining Region Splitting and Edge Detection through Guided Delaunay Image Subdivision.” In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference On*, 1021–26.

Ghosh, Jayanta K, Mohan Delampady, and Tapas Samanta. 2007. *An Introduction to Bayesian Analysis: Theory and Methods*. Springer Science & Business Media.

Gong, Jing, Ji-Yu Liu, Xi-Wen Sun, Bin Zheng, and Sheng-Dong Nie. 2018. “Computer-Aided Diagnosis of Lung Cancer: The Effect of Training Data Sets on Classification Accuracy of Lung Nodules.” *Physics in Medicine & Biology* 63 (3). IOP Publishing:35036.

Graupe, Daniel. 2013. *Principles of Artificial Neural Networks*. Vol. 7. World Scientific.

Günther, Frauke, and Stefan Fritsch. 2010. “Neuralnet: Training of Neural Networks.” *The R Journal* 2 (1):30–38.

Guyon, Isabelle, and André Elisseeff. 2003. “An Introduction to Variable and Feature

Selection.” *Journal of Machine Learning Research* 3 (Mar):1157–82.

Hagan, Martin T, Howard B Demuth, Mark H Beale, and others. 1996. *Neural Network Design*. Vol. 20. Pws Pub. Boston.

Hamad, Arin H, Hozheen O Muhammad, and Sandar P Yaba. 2014. “De-Noising of Medical Images by Using Some Filters.” *Int. Journal Of Biotechnology Research, 2014*.

Hamel, Lutz H. 2011. *Knowledge Discovery with Support Vector Machines*. Vol. 3. John Wiley & Sons.

Hamidian, Sardar, Berkman Sahiner, Nicholas Petrick, and Aria Pezeshk. 2017. “3D Convolutional Neural Network for Automatic Detection of Lung Nodules in Chest CT.” In *Medical Imaging 2017: Computer-Aided Diagnosis*, 10134:1013409.

Han, Jiawei, Jian Pei, and Micheline Kamber. 2011. *Data Mining: Concepts and Techniques*. Elsevier.

Hara, T, M Hirose, X Zhou, H Fujita, T Kiryu, R Yokoyama, and H Hoshi. 2006. “Nodule Detection in 3D Chest CT Images Using 2nd Order Autocorrelation Features.” In *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of The*, 6247–49.

Haralick, Robert M, Karthikeyan Shanmugam, and others. 1973. “Textural Features for Image Classification.” *IEEE Transactions on Systems, Man, and Cybernetics*, no. 6. Ieee:610–21.

Hr, Shally, and K Chitharanjan. 2013. “Tumor Volume Calculation Of” 4 (08):1126–32.

Hsieh, Fushing Y, Daniel A Bloch, and Michael D Larsen. 1998. “A Simple Method of Sample Size Calculation for Linear and Logistic Regression.” *Statistics in Medicine* 17 (14):1623–34.

Hwang, Jenq-Neng, and Yu Hen Hu. 2001. *Handbook of Neural Network Signal Processing*. CRC press.

Javaid, Muzzamil, Moazzam Javid, Muhammad Zia Ur Rehman, and Syed Irtiza Ali Shah. 2016. “A Novel Approach to CAD System for the Detection of Lung Nodules in CT Images.” *Computer Methods and Programs in Biomedicine* 135. Elsevier Ireland Ltd:125–39. https://doi.org/10.1016/j.cmpb.2016.07.031.

Javed, Umer, M Mohsin Riaz, Muhammad Rizwan Khokher, Abdul Ghafoor, and Tanveer A Cheema. 2013. “Fuzzy Logic and Local Features Based Medical Image Segmentation.” In *Image Processing (ICIP), 2013 20th IEEE International Conference On*, 1148–52.

Jiang, Jianmin, P Trundle, and Jinchang Ren. 2010. “Medical Image Analysis with Artificial Neural Networks.” *Computerized Medical Imaging and Graphics* 34 (8). Elsevier:617–31.

Jirapatnakul, Artit C, Yury D Mulman, Anthony P Reeves, David F Yankelevitz, and Claudia I Henschke. 2011. “Segmentation of Juxtapleural Pulmonary Nodules Using a Robust Surface Estimate.” *Journal of Biomedical Imaging* 2011. Hindawi Publishing Corp.:15.

Keshani, Mohsen, Zohreh Azimifar, Reza Boostani, and Alireza Shakibafar. 2010. “Lung Nodule Segmentation Using Active Contour Modeling.” In *Machine Vision and Image Processing (MVIP), 2010 6th Iranian*, 1–6.

Keshani, Mohsen, Zohreh Azimifar, Farshad Tajeripour, and Reza Boostani. 2013. “Lung Nodule Segmentation and Recognition Using SVM Classifier and Active Contour Modeling: A Complete Intelligent System.” *Computers in Biology and Medicine* 43 (4). Elsevier:287–300.

Koss, John E, F D Newman, T K Johnson, and D L Kirch. 1999. “Abdominal Organ Segmentation Using Texture Transforms and a Hopfield Neural Network.” *IEEE Transactions on Medical Imaging* 18 (7). IEEE:640–48.

Kumar, Nalin, and M Nachamai. 2012. “Noise Removal and Filtering Techniques Used in Medical Images.” *Indian Journal of Computer Science and Engineering* 3 (1):146–53.

Kuruvilla, Jinsa, and K. Gunavathi. 2014. “Lung Cancer Classification Using Neural Networks for CT Images.” *Computer Methods and Programs in Biomedicine* 113 (1). Elsevier Ireland Ltd:202–9. https://doi.org/10.1016/j.cmpb.2013.10.011.

Larose, Daniel T. 2005. *Introduction to Data Mining*. Wiley Online Library.

Lederlin, M, M P Revel, A Khalil, G Ferretti, B Milleron, and F Laurent. 2013. “Management Strategy of Pulmonary Nodule in 2013.” *Diagnostic and Interventional Imaging* 94 (11). Elsevier Masson SAS:1081–94. https://doi.org/10.1016/j.diii.2013.05.007.

Lekutai, Gaviphat. 1997. “Adaptive Self-Tuning Neuro Wavelet Network Controllers.” Virginia Tech.

Lewis, David D. 1998. “Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval.” In *European Conference on Machine Learning*, 4–15.

Li, Tao, Shenghuo Zhu, and Mitsunori Ogihara. 2006. “Using Discriminant Analysis for Multi-Class Classification: An Experimental Investigation.” *Knowledge and Information Systems* 10 (4). Springer:453–72.

Li, Zhao, Bo Ye, Minwei Bao, Binbin Xu, Qinyi Chen, Sida Liu, Yudong Han, et al. 2015. “Radiologic Predictors for Clinical Stage IA Lung Adenocarcinoma with Ground Glass Components: A Multi-Center Study of Long-Term Outcomes.” *PLoS ONE* 10 (9):1–12. https://doi.org/10.1371/journal.pone.0136616.

Liao, Shu Hsien. 2005. “Expert System Methodologies and Applications-a Decade Review from 1995 to 2004.” *Expert Systems with Applications* 28 (1):93–103. https://doi.org/10.1016/j.eswa.2004.08.003.

Ma, Ling, Xiabi Liu, Li Song, Yu Liu, Chunwu Zhou, Xinming Zhao, and Yanfeng Zhao. 2014. “A New Classifier Fusion Method Based on Confusion Matrix and Classification Confidence for Recognizing Common CT Imaging Signs of Lung Diseases.” In *Medical Imaging 2014: Computer-Aided Diagnosis*, 9035:90351H.

Ma, Ling, Xiabi Liu, Li Song, Chunwu Zhou, Xinming Zhao, and Yanfeng Zhao. 2015. “A

New Classifier Fusion Method Based on Historical and On-Line Classification Reliability for Recognizing Common CT Imaging Signs of Lung Diseases.” *Computerized Medical Imaging and Graphics* 40. Elsevier Ltd:39–48. https://doi.org/10.1016/j.compmedimag.2014.10.001.

Ma, Zhen, João Manuel R S Tavares, and R M Natal Jorge. 2009. “A Review on the Current Segmentation Algorithms for Medical Images.” In *Proceedings of the 1st International Conference on Imaging Theory and Applications (IMAGAPP)*.

Magalhães Barros Netto, Stelmo, Aristófanes Corrĉa Silva, Rodolfo Acatauassú Nunes, and Marcelo Gattass. 2012. “Automatic Segmentation of Lung Nodules with Growing Neural Gas and Support Vector Machine.” *Computers in Biology and Medicine* 42 (11):1110–21. https://doi.org/10.1016/j.compbiomed.2012.09.003.

Mahmood, Faleh H, Wafaa A Abbas, and S M Ali. 2014. “Estimating the Lung Tumor Size in CT Images Using Image Segmentation Techniques.” *International Journal of Emerging Technology and Advanced Engineering* 4 (7):509–17.

Maintz, Twan. 2005. “Digital and Medical Image Processing.” *Universiteit Utrecht*.

Mansoor, Awais, Ulas Bagci, Brent Foster, Ziyue Xu, Georgios Z Papadakis, Les R Folio, Jayaram K Udupa, and Daniel J Mollura. 2015. “Segmentation and Image Analysis of Abnormal Lungs at CT: Current Approaches, Challenges, and Future Trends.” *RadioGraphics* 35 (4). Radiological Society of North America:1056–76.

McCallum, Andrew, Kamal Nigam, and others. 1998. “A Comparison of Event Models for Naive Bayes Text Classification.” In *AAAI-98 Workshop on Learning for Text Categorization*, 752:41–48.

Mingqiang, Yang, Kpalma Kidiyo, and Ronsin Joseph. 2008. *A Survey of Shape Feature Extraction Techniques*. *Pattern Recognition Techniques, Technology and Applications*. https://doi.org/10.5772/6237.

Montero, Raul S, and Ernesto Bribiesca. 2009. “State of the Art of Compactness and Circularity Measures.” *International Mathematical Forum* 4 (27):1305–35. http://www.m-hikari.com/imf-password2009/25-28-2009/bribiescaIMF25-28-

2009.pdf.

Motoyama, Sadako, Takeshi Kondo, Masayoshi Sarai, Atsushi Sugiura, Hiroto Harigaya, Takahisa Sato, Kaori Inoue, et al. 2007. “Multislice Computed Tomographic Characteristics of Coronary Lesions in Acute Coronary Syndromes.” *Journal of the American College of Cardiology* 50 (4). Journal of the American College of Cardiology:319–26.

Mustra, M., K. Delac, and M. Grgic. 2008. “Overview of the DICOM Standard.” *2008 50th International Symposium ELMAR* 1 (September):39–44.

Naresh, Prashant, and Rajashree Shettar. 2014. “Image Processing and Classification Techniques for Early Detection of Lung Cancer for Preventive Health Care: A Survey.” *International Journal on Recent Trends in Engineering & Technology* 11 (1). Association of Computer Electronics and Electrical Engineers (ACEEE):595.

Negnevitsky, Michael. 2005. *Artificial Intelligence: A Guide to Intelligent Systems*. Pearson Education.

Ney, Hermann. 1992. “A Comparative Study of Two Search Strategies for Connected Word Recognition: Dynamic Programming and Heuristic Search.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14 (5). IEEE:586–95.

Ng, Andrew Y, and Michael I Jordan. 2002. “On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes.” In *Advances in Neural Information Processing Systems*, 841–48.

Nie, Shengdong, Lihong Li, Yuanjun Wang, Chaofan He, Feng Ji, and Jianmei Liang. 2012. “A Segmentation Method for Sub-Solid Pulmonary Nodules Based on Fuzzy c-Means Clustering.” In *Biomedical Engineering and Informatics (BMEI), 2012 5th International Conference On*, 169–72.

Oseas, Antonio, De Carvalho Filho, Wener Borges, De Sampaio, Aristófanes Corrêa, Anselmo Cardoso, De Paiva, Rodolfo Acatauassú, and Marcelo Gattass. 2014. “Artificial Intelligence in Medicine Automatic Detection of Solitary Lung Nodules Using Quality Threshold Clustering , Genetic Algorithm and Diversity Index.”

Artificial Intelligence In Medicine 60 (3). Elsevier B.V.:165–77. https://doi.org/10.1016/j.artmed.2013.11.002.

Ozekes, Serhat, and Onur Osman. 2010. "Computerized Lung Nodule Detection Using 3D Feature Extraction and Learning Based Algorithms." *Journal of Medical Systems* 34 (2). Springer:185–94.

Ozekes, Serhat, Onur Osman, and Osman N Ucan. 2008. "Nodule Detection in a Lung Region That's Segmented with Using Genetic Cellular Neural Networks and 3D Template Matching with Fuzzy Rule Based Thresholding." *Korean Journal of Radiology* 9 (1):1–9.

Özkan, Co\cskun, and Filiz Sunar Erbek. 2003. "The Comparison of Activation Functions for Multispectral Landsat TM Image Classification." *Photogrammetric Engineering & Remote Sensing* 69 (11). American Society for Photogrammetry and Remote Sensing:1225–34.

Pal, Nikhil R, and Sankar K Pal. 1993. "A Review on Image Segmentation Techniques." *Pattern Recognition* 26 (9). Elsevier:1277–94.

Panchal, Gaurang, Amit Ganatra, Y P Kosta, and Devyani Panchal. 2011. "Behaviour Analysis of Multilayer Perceptronswith Multiple Hidden Neurons and Hidden Layers." *International Journal of Computer Theory and Engineering* 3 (2). IACSIT Press:332.

Papademetris, Xenophon, and Alark Joshi. 2006. "An Introduction to Programming for Medical Image Analysis with the Visualization Toolkit." *Yale University*, 283. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.122.7477&rep=rep1&type=pdf.

Peter, V Joseph, and M Karnan. 2013. "Medical Image Analysis Using Unsupervised and Supervised Classification Techniques," no. 5:40–45.

Pham, Dzung L, Chenyang Xu, and Jerry L et al. Prince. 2000. "Current Methods in Medical Image Segmentation." *Annual Review of Biomedical Engineering* 2 (1). Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA:315–37.

ping Tian, Dong, and others. 2013. “A Review on Image Feature Extraction and Representation Techniques.” *International Journal of Multimedia and Ubiquitous Engineering* 8 (4). Citeseer:385–96.

Pratt, William K. 1991. “Morphological Image Processing.” *Digital Image Processing: PIKS Inside, Third Edition*. Wiley Online Library, 401–41.

Pratt, William K. 2001. *Processing Digital Image Processing*. *Image Rochester NY.* Vol. 5. https://doi.org/10.1016/S0146-664X(78)80023-9.

Ramesh, Venkataraman, Robert L. Glass, and Iris Vessey. 2004. “Research in Computer Science: An Empirical Study.” *Journal of Systems and Software* 70 (1–2):165–76. https://doi.org/10.1016/S0164-1212(03)00015-3.

Rebouc, Pedro Pedrosa, Roger Moura Sarmento, Paulo C Cortez, and Victor Hugo De. 2015. “Adaptive Crisp Active Contour Method for Segmentation and Reconstruction of 3d Lung Structures.” *International Journal of Computer Applications* 111 (4). Foundation of Computer Science.

Rebouças Filho, Pedro Pedrosa, Paulo César Cortez, John Hebert da Silva Félix, Tarique da Silveira Cavalcante, and Marcelo Alcantara Holanda. 2013. “Adaptive 2D Crisp Active Contour Model Applied to Lung Segmentation in CT Images of the Thorax of Healthy Volunteers and Patients with Pulmonary Emphysema.” *Revista Brasileira de Engenharia Biomédica* 29 (4). SciELO Brasil:363–76.

Rikxoort, Eva M. Van, and Bram Van Ginneken. 2013. “Automated Segmentation of Pulmonary Structures in Thoracic Computed Tomography Scans: A Review.” *Physics in Medicine and Biology* 58 (17). https://doi.org/10.1088/0031-9155/58/17/R187.

Rikxoort, Eva M. Van, Bartjan De Hoop, Saskia Van De Vorst, Mathias Prokop, and Bram Van Ginneken. 2009. “Automatic Segmentation of Pulmonary Segments from Volumetric Chest CT Scans.” *IEEE Transactions on Medical Imaging* 28 (4):621–30. https://doi.org/10.1109/TMI.2008.2008968.

Rojas, Raúl. 2013. *Neural Networks: A Systematic Introduction*. Springer Science & Business Media.

Ross, Timothy J, and others. 2004. *Fuzzy Logic with Engineering Applications*. Vol. 2. Wiley Online Library.

Roy, Tanushree Sinha, Neeraj Sirohi, and Arti Patle. 2015. "Classification of Lung Image and Nodule Detection Using Fuzzy Inference System." In *Computing, Communication & Automation (ICCCA), 2015 International Conference On*, 1204–7.

Russ, John C. 2016. *The Image Processing Handbook*. CRC press.

Saad, Mohd Nizam, and Hamzaini Abdul Hamid. 2014. "Image Segmentation for Lung Region in Chest X-Ray Images Using Edge Detection and Morphology." *4th IEEE International Conference on Control Systems, Computing and Engineering*, no. November:28–30. https://doi.org/10.1109/ICCSCE.2014.7072687.

Saeys, Yvan, Iñaki Inza, and Pedro Larrañaga. 2007. "A Review of Feature Selection Techniques in Bioinformatics." *Bioinformatics* 23 (19). Oxford University Press:2507–17.

Salim, Elsayed I, Abdul Rahman Jazieh, and Malcolm A Moore. 2011. "Lung Cancer Incidence in the Arab League Countries: Risk Factors and Control." *Asian Pac J Cancer Prev* 12 (9):17–34.

Sander, Jörg, Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. 1998. "Density-Based Clustering in Spatial Databases: The Algorithm Gdbscan and Its Applications." *Data Mining and Knowledge Discovery* 2 (2). Springer:169–94.

Santos, Alex Martins, Antonio Oseas de Carvalho Filho, Aristófanes Corrêa Silva, Anselmo Cardoso de Paiva, Rodolfo Acatauassú Nunes, and Marcelo Gattass. 2014. "Automatic Detection of Small Lung Nodules in 3D CT Data Using Gaussian Mixture Models, Tsallis Entropy and SVM." *Engineering Applications of Artificial Intelligence* 36. Elsevier:27–39.

Schoot, Rens Van de, David Kaplan, Jaap Denissen, Jens B. Asendorpf, Franz J. Neyer, and Marcel A.G. van Aken. 2014. “A Gentle Introduction to Bayesian Analysis: Applications to Developmental Research.” *Child Development* 85 (3):842–60. https://doi.org/10.1111/cdev.12169.

Schubert, Erich, Jörg Sander, Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. 2017. “DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN.” 42 (3 OP-ACM Transactions on Database Systems. Jul2017, Vol. 42 Issue 3, p1-21. 21p.):1. https://doi.org/10.1145/3068335.

Senthil Kumar, T. K., E. N. Ganesh, and R. Umamaheswari. 2017. “Lung Nodule Volume Growth Analysis and Visualization through Auto-Cluster k-Means Segmentation and Centroid/Shape Variance Based False Nodule Elimination.” *Biomedical Research (India)* 28 (5):1927–34.

Sharma, Avisha, and Sanyam Anand. 2013. “An Efficient Technique of De-Noising Medical Images Using Neural Network and Fuzzy -A Review,” no. 4:66–68.

Sieren, Jessica C., Yoshiharu Ohno, Hisanobu Koyama, Kazuro Sugimura, and Geoffrey McLennan. 2010. “Recent Technological and Application Developments in Computed Tomography and Magnetic Resonance Imaging for Improved Pulmonary Nodule Detection and Lung Cancer Staging.” *Journal of Magnetic Resonance Imaging* 32 (6):1353–69. https://doi.org/10.1002/jmri.22383.

Sivakumar, S, and C Chandrasekar. 2012. “Lungs Image Segmentation through Weighted FCM.” In *Recent Advances in Computing and Software Systems (RACSS), 2012 International Conference On*, 109–13.

Sluimer, Ingrid Christine. 2005. “Automated Image Analysis of the Pathological Lung in CT.” Utrecht University.

Suganya, R, and R Shanthi. 2012. “Fuzzy C-Means Algorithm-a Review.” *International Journal of Scientific and Research Publications* 2 (11):1.

Sun, Tao, Jingjing Wang, Xia Li, Pingxin Lv, Fen Liu, Yanxia Luo, Qi Gao, Huiping Zhu, and Xiuhua Guo. 2013. “Comparative Evaluation of Support Vector Machines for Computer Aided Diagnosis of Lung Cancer in CT Based on a Multi-Dimensional Data

Set.” *Computer Methods and Programs in Biomedicine* 111 (2). Elsevier Ireland Ltd:519–24. https://doi.org/10.1016/j.cmpb.2013.04.016.

Taha, Abdel Aziz, and Allan Hanbury. 2015. “Metrics for Evaluating 3D Medical Image Segmentation: Analysis, Selection, and Tool.” *BMC Medical Imaging* 15 (1). BioMed Central:29.

Talakoub, Omid, Javad Alirezaie, and Paul Babyn. 2007. “LUNG SEGMENTATION IN PULMONARY CT IMAGES USING WAVELET TRANSFORM Department of Electrical and Computer Engineering , Ryerson University , Canada Department of Systems Design Engineering , University of Waterloo , Canada Department of Medical Imaging Unive.” *Imaging*, 453–56.

Tibshirani, Robert. 2011. “Regression Shrinkage and Selection via the Lasso: A Retrospective.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73 (3). Wiley Online Library:273–82.

Tirz\"\ite, Madara, M\=aris Bukovskis, Gunta Strazda, Normunds Jurka, and Immanuels Taivans. 2018. “Detection of Lung Cancer with Electronic Nose and Logistic Regression Analysis.” *Journal of Breath Research* 13 (1). IOP Publishing:16006.

Treichel, Thomas, Michael Gessat, Torsten Prietzel, and Oliver Burgert. 2012. “DICOM for Implantations - Overview and Application.” *Journal of Digital Imaging* 25 (3):352–58. https://doi.org/10.1007/s10278-011-9416-8.

Tseng, Lin-Yu, and Li-Chin Huang. 2009. “An Adaptive Thresholding Method for Automatic Lung Segmentation in CT Images.” In *AFRICON, 2009. AFRICON’09.*, 1–5.

Valente, Igor Rafael S., Paulo César Cortez, Edson Cavalcanti Neto, José Marques Soares, Victor Hugo C. de Albuquerque, and João Manuel R.S. Tavares. 2016. “Automatic 3D Pulmonary Nodule Detection in CT Images: A Survey.” *Computer Methods and Programs in Biomedicine* 124. Elsevier Ireland Ltd:91–107. https://doi.org/10.1016/j.cmpb.2015.10.006.

Varios, P Marchand, and O T Holland. 2003. *Graphics and GUIs*. *New York*. Vol. 38.

http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Graphics+and+GUIs+with+MATLAB#0.

Wang, Jiahui, Feng Li, and Qiang Li. 2009. “Automated Segmentation of Lungs with Severe Interstitial Lung Disease in CT.” *Medical Physics* 36 (10). Wiley Online Library:4592–99.

Wang, Jinke, and Yuanzhi Cheng. 2015. “A New Pulmonary Nodules Computer-Aided Detection System in Chest CT Images Based on Adaptive Fuzzy C-Means Technology.” In *Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2015 7th International Conference On*, 1:514–17.

Wang, Qingzhu, Wenwei Kang, Chunming Wu, and Bin Wang. 2013. “Computer-Aided Detection of Lung Nodules by SVM Based on 3D Matrix Patterns.” *Clinical Imaging* 37 (1). Elsevier:62–69.

Wang, Yang, Xiaoqian Che, and Siliang Ma. 2012. “Nonlinear Filtering Based on 3D Wavelet Transform for MRI Denoising.” *EURASIP Journal on Advances in Signal Processing* 2012 (1). Springer:40.

Wang, Ying, Yong Fan, Priyanka Bhatt, and Christos Davatzikos. 2010. “High-Dimensional Pattern Regression Using Machine Learning: From Medical Images to Continuous Clinical Variables.” *Neuroimage* 50 (4). Elsevier:1519–35.

Wang, Zhiling, Andrea Guerriero, and Marco De Sario. 1996. “Comparison of Several Approaches for the Segmentation of Texture Images.” *Pattern Recognition Letters* 17 (5). Elsevier:509–21.

Washko, George R., Grace Parraga, and Harvey O. Coxson. 2012. “Quantitative Pulmonary Imaging Using Computed Tomography and Magnetic Resonance Imaging.” *Respirology* 17 (3):432–44. https://doi.org/10.1111/j.1440-1843.2011.02117.x.

Way, Ted W, Lubomir M Hadjiiski, Berkman Sahiner, Heang-Ping Chan, Philip N Cascade, Ella A Kazerooni, Naama Bogot, and Chuan Zhou. 2006. “Computer-Aided Diagnosis of Pulmonary Nodules on CT Scans: Segmentation and Classification Using 3D Active Contours.” *Medical Physics* 33 (7Part1). Wiley Online

Library:2323–37.

Wilson, Joseph N, and Gerhard X Ritter. 2000. *Handbook of Computer Vision Algorithms in Image Algebra*. CRC press.

Witten, Ian H, Eibe Frank, Mark A Hall, and Christopher J Pal. 2016. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

Xu, Lei, Adam Krzyzak, and Ching Y Suen. 1992. “Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition.” *IEEE Transactions on Systems, Man, and Cybernetics* 22 (3). IEEE:418–35.

Yang, Changbo, Ming Dong, and Farshad Fotouhi. 2005. “Image Content Annotation Using Bayesian Framework and Complement Components Analysis.” In *Image Processing, 2005. ICIP 2005. IEEE International Conference On*, 1:I--1193.

Yang, Xiaofeng, Srini Tridandapani, Jonathan J. Beitler, David S. Yu, Emi J. Yoshida, Walter J. Curran, and Tian Liu. 2012. “Ultrasound GLCM Texture Analysis of Radiation-Induced Parotid-Gland Injury in Head-and-Neck Cancer Radiotherapy: An in Vivo Study of Late Toxicity.” *Medical Physics* 39 (9):5732–39. https://doi.org/10.1118/1.4747526.

Ying, Wei, Jia Tong, and Lin Ming-Xiu. 2011. “Autonomous Detection of Solitary Pulmonary Nodules on CT Images for Computer-Aided Diagnosis.” *Proceedings of the 2011 Chinese Control and Decision Conference, CCDC 2011*, 4054–59. https://doi.org/10.1109/CCDC.2011.5968933.

Yong, Yang, Zheng Chongxun, and Lin Pan. 2004. “A Novel Fuzzy C-Means Clustering Algorithm for Image Thresholding.” *Measurement Science Review* 4 (1). Citeseer:11–19.

Zhang, Dengsheng, and Guojun Lu. 2004. “Review of Shape Representation and Description Techniques.” *Pattern Recognition* 37 (1). Elsevier:1–19.

Zhang, Xiangwei, Geoffrey McLennan, Eric A Hoffman, and Milan Sonka. 2005. “3D Segmentation of Non-Isolated Pulmonary Nodules in High Resolution CT Images.”

In *Medical Imaging 2005: Image Processing*, 5747:1438–46.

Zhao, Chunjiang, Wenkang Shi, and Yong Deng. 2005. “A New Hausdorff Distance for Image Matching.” *Pattern Recognition Letters* 26 (5). Elsevier:581–86.

Zhou, Hailing, Dmitry B. Goldgof, Samuel Hawkins, Lei Wei, Ying Liu, Doug Creighton, Robert J. Gillies, Lawrence O. Hall, and Saeid Nahavandi. 2016. “A Robust Approach for Automated Lung Segmentation in Thoracic CT.” *Proceedings - 2015 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2015*, 2267–72. https://doi.org/10.1109/SMC.2015.396.

Zhou, Kaijun, and Guirong Weng. 2015. “The Method of Detecting Pulmonary Nodules by Snake Model.” In *Control, Automation and Information Sciences (ICCAIS), 2015 International Conference On*, 174–78.