YILDIRIM BEYAZIT UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

EARLY DETECTION OF LUNG CANCER

M.Sc. Thesis by

Shaymaa Abdulhafedh S. Shakir

Department of Computer Engineering

February, 2016

ANKARA

EARLY DETECTION OF LUNG CANCER

A Thesis Submitted to the

Graduate School of Natural and Applied Sciences of Yıldırım Beyazıt University

In Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Engineering, Department of Computer Engineering

by

Shaymaa Abdulhafedh S. Shakir

February, 2016

ANKARA

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"EARLY DETECTION OF LUNG CANCER"** completed by **Shaymaa Abdulhafedh S.Shakir** under supervision of **Prof. Dr. Fatih V. ÇELEBİ** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

.....

Prof. Dr. Fatih V. ÇELEBİ

Supervisor

.....	.....
Assoc. Prof. Dr. Süleyman TOSUN	**Assist. Prof. Dr. Hilal KAYA**
(Jury Member)	(Jury Member)

.....

Prof. Dr. Fatih V. ÇELEBİ

(Director)

Graduate School of Natural and Applied Sciences

ETHICAL DECLERATION

I have prepared this dissertation study in accordance with the Rules of Writing Thesis of Yıldırım Beyazıt University of Science and Technology Institute;

- Data I have presented in the thesis, information and documents that I obtained in the framework of academic and ethical rules,
- All information, documentation, assessment and results that I presented in accordance with scientific ethics and morals,
- I have gave references all the works that I were benefited in this dissertation by appropriate reference,
- I would not make any changes in the data that I were used,
- The work presented in this dissertation I would agree that the original,

I state, in the contrary case I declare that I accept the all rights losses that may arise against me.

EARLY DETECTION OF LUNG CANCER

ABSTRACT

The clinical diagnostics for lung cancer are mostly depended on physical and biochemical techniques. Imaging processes in computed tomography (CT) screening (low-dose computed tomography - LDCT) is convenient for discovering lung cancer in the early stages. CT scan images for this study were obtained from Ankara Atatürk Training and Research Hospital and also from the online international database of "The Lung Image Database Consortium (LIDC)".

Correct decision for diagnosis of lung cancer using CT scanning requires some processes to remove noise from image in enhancement stage. The noise removing process in this thesis have been proposed using a gradient magnitude in sobel filter. Finding edges on the images is a first step to detect nodule in the tissues. As well as following morphology operations to isolate background from the foreground is very important because background image represents the tissue of lung to locate a tumor on it, therefore foreground is unnecessary.

Second part in the thesis includes usage of watershed algorithm for segmentation of tumor from the tissue. Labeling nodular irregular area inside the tissue leads to over-segmentation on image by connected components and markers which have different values as intensity values and regional minimum represented in the foreground. Markers classify tumor area by labeling high intensity values, locating the region of interest in normal image for cutting random area from the tissue. Distinguishing the normal and abnormal images that needs to use statistical methods is depended on the cancer type. To get feature extraction of nodule shape provides certain parameters which is an essential step for classification processes.

Last part in this thesis focuses on classification processes on a dataset which includes values belong to five parameters that were taken from statistical method results. This dataset involves 306 images consisting of 153 normal images and 153 abnormal images. This database is implemented in nine classification algorithms and different results of accuracy performance were taken according to the theory of these

algorithms. This study aims detecting tumor area and getting high performance accuracy after classification were done to find out the best algorithm to help the doctor for the diagnosis.

Keywords: Lung cancer, early detection, sobel filter, morphology operations, watershed algorithm

AKCİĞER KANSERİ ERKEN TANI

ÖZET

Akciğer kanserinin klinik teşhis yöntemleri çoğunlukla fiziksel ve biyokimyasal tekniklere dayanır. Bilgisayarlı tomografi (BT) taramasındaki görüntüleme süreçleri (Düşük doz bilgisayarlı tomografi-DDBT), akciğer kanserinin erken evrelerde teşhisi için uygundur. Bu çalışma için BT tarama görüntüleri, Ankara Atatürk Eğitim ve Araştırma Hastanesi ve ayrıca "Akciğer Görüntüleri Veritabanı Konsorsiyumu (AGVK)" isimli uluslararası veritabanından elde edilmiştir.

BT taraması kullanılarak akciğer kanserinin teşhisinde doğru karar verilmesi için, doğruluğu artırma aşamasında görüntüdeki gürültünün giderilmesi gerekmektedir. Bu tezdeki gürültü giderme süreci için sobel filtrelemede gradyan büyüklüğünün kullanılması önerilmiştir. Ayrıca görüntüdeki arka planı ön plandan ayırmak üzere, önceki aşamaları izleyen morfoloji işlemlerinin kullanımı da çok önemlidir çünkü arka plan görüntüsü üzerinde tümör bulunan akciğer dokusunu temsil etmektedir, bu nedenle ön plan görüntüsü gerekli değildir.

Tezin ikinci bölümü tümörün dokudan bölütlenmesi için watershed algoritmasının kullanımını içermektedir. Doku içinde bulunan düzensiz nodüler alanın belirlenmesi farklı yoğunluk değerleri ve arka planda temsil edilen bölgesel minimum değerine sahip işaretleyicilerle ve bağlı bileşenlerle görüntünün aşırı bölütlenmesine neden olur. İşaretleyiciler, tümör alanını yüksek yoğunluk değerlerini belirleyerek ve dokudaki alanı kesmek üzere normal görüntüdeki ilgilenilen alanın yerini saptayarak tümörü sınıflandırır. İstatistiksel metotları kullanarak normal ve anormal görüntülerin ayrılması, kanser türüne bağlıdır. Nodül şeklinin özelliğinin çıkarılması, sınıflandırma süreçleri için önemli bir adım olan kesin parametreleri ortaya çıkarır.

Bu tezdeki son bölüm, istatistiksel metotların sonuçlarından alınan beş parametreye ait değerleri içeren veri seti üzerindeki sınıflandırma süreçlerine odaklanmaktadır. Bu veri seti, 153 adet normal ve 153 adet anormal görüntüyü içeren toplam 306 görüntüden oluşmaktadır. Bu veritabanı dokuz farklı sınıflandırma algoritmasına tabi tutulmuş ve bu algoritmaların çalışma teorilerine bağlı olarak farklı doğruluk

değerleri elde edilmiştir. Bu çalışma, tümörün belirlenmesi ve doktorun teşhisine yardımcı olmak üzere sınıflandırma yapılmasında en yüksek doğruluk performansını gösteren algoritmayı tespit etmeyi amaçlamaktadır.

Anahtar Kelimeler: Akciğer kanseri, erken teşhis, sobel filtre, morfoloji işlemleri, watershed algoritması

ACKNOWLEDGEMENTS

First of all, I would like to express my deepest sense of Gratitude to my supervisor **Prof. Dr. Fatih V. ÇELEBİ**, director of Graduate School of Natural and Applied Sciences, who offered his continuous constant advice also support and encouragement throughout the course of this thesis. I thank to him for the systematic guidance and great effort that he put into training me in the scientific field to improve my knowledge for this topic. I could not have imagined having a better advisor and mentor for my study.

I would also like to thank to **Assist. Prof. Dr. Hilal Kaya** for helping me to develop my background in biomedical image processing and give her best suggestions during my research. It would not have been possible without her help.

I would like to express my respects to all my teachers who have been teaching me in Computer Engineering Department, University of Yıldırım Beyazıt. I would like to thank to all those people who made this thesis possible and an unforgettable experience for me.

Last but not least, special thanks goes to my family, for giving birth to me at the first place and supporting me spiritually. They were always encouraging me with their best wishes. I also thank to my friends for their friendship and supports.

February, 2016 Shaymaa Abdulhafedh S. Shakir

CONTENTS

Page

ABBREVATIONS

LDCT	Low Dose Computed Tomography
CT	Computer Tomography
X-RAY	X-Radiation
PET	Positron Emission Tomography
MRI	Magnetic Resonance Imaging
NLST	National Lung Screening Trial
DNA	Deoxyribo Nucleic Acid
RNA	Ribo Nucleic Acid
SCLC	Small Cell Lung Cancer
NSCLC	Non Small Cell Lung Cancer
TNM	Tumor Node Metastases
PACS	Picture Archiving and Communication Systems
DICOM	Digital Imaging and Communications in Medicine
3D	Three Domination
2D	Two Domination
RSS	Random Sub Space
SVM	Support Vector Machines
ANN	Artificial Neural Network
Ediameter	Equivalent Diameter
ROI	Region Of Interest

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

1.1. Thesis Outline

In recent years, lung cancer is considered as a serious problem in the world as it causes death most commonly among other types of the cancer. Therefore large numbers of studies are mentioned to increase the survival rate by using different methods when tumor is detected in the lung tissues before spreading in the human body. Currently the application of low dose computed tomography (LDCT) is used for early stage lung cancer and this technique is proved to reduce the numbers of lung cancer based mortality [1].

Detection of tumor in small or large areas of tissues are generally performed in two phases: detection and analysis.

- In detection phase, discovering and labeling the abnormal nodule (tumor) from lung tissue is done.
  - Enhancement method to remove the noise from image to get good texture.
  - Segmentation of the tumor area from lung tissue.

- In analysis phase, segmenting the nodule (tumor) individually, then producing the specific feature information to reduce mistakes to diagnose.
  - Feature extraction of image for classification.
  - Classification methods to classify images in groups of normal and abnormal.

1.2. Related Works

At the beginning of this thesis presents a brief overview of advantages of the tumor detection in early stages. Review of cancer causes in general, types of lung cancer, how to be treated and how the tumors spread through the tissues. also explains the stages of the cancer and methods for lung screening is mentioned on CT scanning screens and why it is commonly used for diagnosing lung cancer in early stages is

also mentioned. The methods that were used in this thesis on the enhancement of the images for removing noise, morphology methods to prepare images for segmentation and detecting nodules on tissue of lung to get feature extraction of tumor and classifier. Demonstrates the steps of preprocessing, filtering images in sobel filter using gradient magnitude to remove noise from the images, processing, marking background and foreground of the images by morphology operations (dilation, erosion, opening, closing) to segment the images in threshold method and finally detecting tumor area on the image by watershed segmentation method. Detection processes of the tumor area using statistical methods to differentiate tumor feature from abnormal images, to cut area manually from the normal image for the purpose of classifying images to distinguish the normal and abnormal images.

1.3. Aim of the Work

The proposed system can be used to distinguish the normal and abnormal images by detecting nodules in the lung tissues. Therefore the main objective of this thesis is to help the doctor for diagnosing tumor area especially in early stages. In this study, a user interface is designed using MATLAB including filtering, segmentation, feature extraction and classification processes. This user interface provides the following main features as follows:

1.3.1. Filtering Process

Introduces enhancement method to remove noise from image which includes: sobel filter with gradient magnitude to focus on the edge detection of the image (normal and abnormal).

1.3.2. Segmentation Process

Labels foreground objects, computes opening operation (reconstruction) applying erosion method to reduce holes on image to get structured element. Also removes dark spots on image by combining regular closing operation and opening operation (reconstruction) to find the target area (tumor area) on CT image. The previous methods will be used to compute markers of background on CT images using threshold values to find the maximum value on the image that represents high

intensity values on the image (tumor). Implements watershed algorithm to segment tumor area from CT image and removes the background.

1.3.3. Feature Extraction Process

Uses statistics-based feature extraction methods to find the region of interest on image by using five parameters (area, perimeter, irregular, ediameter and eccentricity) to get results from random images (normal or abnormal).

1.3.4. Classification Process

Uses classification algorithms to classify the tumor area as normal and abnormal. At the end of the classification process, we've decided which classification algorithm gives better performance for this study.

CHAPTER TWO

MEDICAL BACKGROUND OF CANCER

The human body is composed of trillions of living cells. Normal body cells grow, divide into new cells and die in an orderly manner. During the early years of a person's life, normal cells divide faster to allow the person to grow. After the person becomes an adult, most cells divide only to replace worn-out or dying cells or to repair injuries. There are many kinds of cancer but all kinds of cancer begin when cells in a part of the body start to grow out of control [2].

Cells become cancer cells because of DNA damaging. DNA is in every cell and directs all its actions. In a normal cell when DNA gets damaged, the cell either repairs the damage or the cell dies. In cancer cells, the damaged DNA is not repaired but the cell doesn't die as it should. Instead, this cell goes on making new cells that the body does not need. These new cells all have the same damaged DNA as the first cell does. People can inherit damaged DNA, but most DNA damages are caused by mistakes that happen while the normal cell is reproducing or by something in our environment. Sometimes the cause of the DNA damage is something obvious like smoking but often no clear cause is found [3].

In most cases, the cancer cells form a tumor except for a few cancer types. Not all tumors are cancerous (malignant) [4]. Tumors that aren't cancer are called as benign. Benign tumors may also cause problems – they may grow in very large area and press on healthy organs and tissues. But they cannot grow into (invade) other tissues because they cannot invade, they also cannot spread to other parts of the body (metastasize). These tumors are almost never life threatening. Different types of cancer can behave very differently. For example, lung cancer and breast cancer are very different diseases. They grow at different rates and respond to different treatments that is why people with cancer need treatment that is aimed at their particular kind of cancer [5].

Some cancers, like leukemia, rarely form tumors. Instead, these cancer cells involve the blood and blood-forming organs and circulate through other tissues where they grow. Cancer cells often spread to other parts of the body, where they begin to grow and form new tumors that replace normal tissue. This process is called metastasis, it happens when the cancer cells get into the bloodstream or lymph vessels of the body. No matter where a cancer may spread, it is always named (and treated) based on the place where it started. For example, breast cancer that has spread to the liver is still breast cancer, not liver cancer. Likewise, prostate cancer that has spread to the bone is still prostate cancer, not bone cancer [6].

2.1. Types of lung cancer

There are several different types of lung cancer .In general; lung cancer is divided into two types:

- **Small cell lung cancer (SCLC):** approximately 10% to 15% of all lung cancers are small cell lung cancer (SCLC), named for the size of the cancer cells when examined under a microscope. SCLC often starts in the bronchi near the center of the chest. It tends to grow and spread quickly and it always spread to distant parts of the body before it is found.
- **Non-small cell lung cancer (NSCLC):** approximately 85% to 90% of lung cancers are non-small cell lung cancer (NSCLC). There are 3 main subtypes of NSCLC:

The cells in these subtypes differ in size, shape and chemical make-up when looked at under a microscope. But they are grouped together because the approach to treatment and prognosis (outlook) are similar [7]. They are discussed further in our document.

2.1.1 Lung Cancer - Non-Small Cell: Stages

Staging is a way of describing where the cancer is located, if or where it has spread, and whether it is affecting other parts of the body. Doctors use diagnostic tests to find out the cancer's stage, so staging may not be completed until all of the tests are finished. Knowing the stage helps the doctor to decide what kind of treatment is

fitted best and can help to predict a patient's prognosis, which has the chance of recovery. There are different stage descriptions for different types of cancer.

2.1.2 TNM (Tumor, Node, Metastases) staging system

The TNM system is mostly used cancer staging systems. This system has been accepted by the Union for International Cancer Control (UICC) and the American Joint Committee on Cancer (AJCC). In medical facilities are use the(TNM system). (T) it means the size and position of the tumor.(N) it means the cancer cells have spread into the lymph nodes. Finally (M) whether the tumor has spread anywhere else in the body – secondarily cancer or metastases[8].

2.1.3 Cancer stage grouping

The stage of NSCLC is described by a number, zero (0) through four (Roman numerals I through IV) and one extra stage called the hidden stage. One way to determine the staging of NSCLC is to find out whether the cancer can be completely removed by a surgeon. To completely remove the lung cancer, the surgeon must remove the cancer, along with the surrounding, healthy lung tissue.

- **Hidden stage**

  It means cancer cannot be seen by imaging or bronchoscopy. Cancer cells are found in sputum (coughed up from the lungs) or bronchial washing.

- **Stage 0**

  It means the cancer is "in place" and has not grown into nearby tissues and spread outside the lung.

- **Stage I**

  In stage one (I), lung cancer is a small tumor that has not spread to any lymph nodes, making it possible for a surgeon to completely remove it. Stage I is divided into two substages based on the size of the tumor:

  ❖ Stage IA tumors are less than 3 centimeters (cm) wide.

  ❖ Stage IB tumors are more than 3 cm but less than 5 cm wide.

- **Stage II**

Stage two (II) lung cancer is divided into two substages:

❖ Stage IIA lung cancer describes a tumor larger than 5 cm but less than 7 cm wide that has not spread to the nearby lymph nodes or a small tumor less than 5 cm wide that has spread to the nearby lymph nodes.

❖ Stage IIB lung cancer describes a tumor larger than 5 cm but less than 7 cm wide that has spread to the lymph nodes or a tumor more than 7 cm wide that may or may not have grown into nearby structures in the lung but has not spread to the lymph nodes.

- **Stage III**

Stage three (III) lung cancers are classified as either stage IIIA or IIIB. For many stage IIIA cancers and nearly all stage IIIB cancers, the surgeon on the tumor is difficult, and sometimes it is impossible to remove. For example, the lung cancer may have spread to the lymph nodes located in the center of the chest, which is outside the lung. Or, the tumor may have grown into nearby structures in the lung. In either situation, it is less likely that the surgeon can completely remove the cancer because removal of the cancer must be performed bit by bit.

- **Stage IV**

Stage four (IV) means the lung cancer has spread to more than one area in the other lung, the fluid surrounding the lung or the heart, or distant parts of the body through the bloodstream. Once released in the blood, cancer can spread anywhere in the body, but it is more likely to spread to the brain, bones, liver, and adrenal glands. It is divided into two substages:

❖ Stage IVA cancer has spread within the chest.

❖ Stage IVB has spread outside of the chest.

In general, surgery is not successful for most stage III or IV lung cancers. Lung cancer can also be impossible to remove if it has spread to the lymph nodes above the collarbone, or if the cancer has grown into vital structures within the chest, such as the heart, large blood vessels, or the main breathing tubes leading to the lungs.

The doctor will recommend other treatment options [9]. Figure 2.1 shows the lung cancer types and stages.

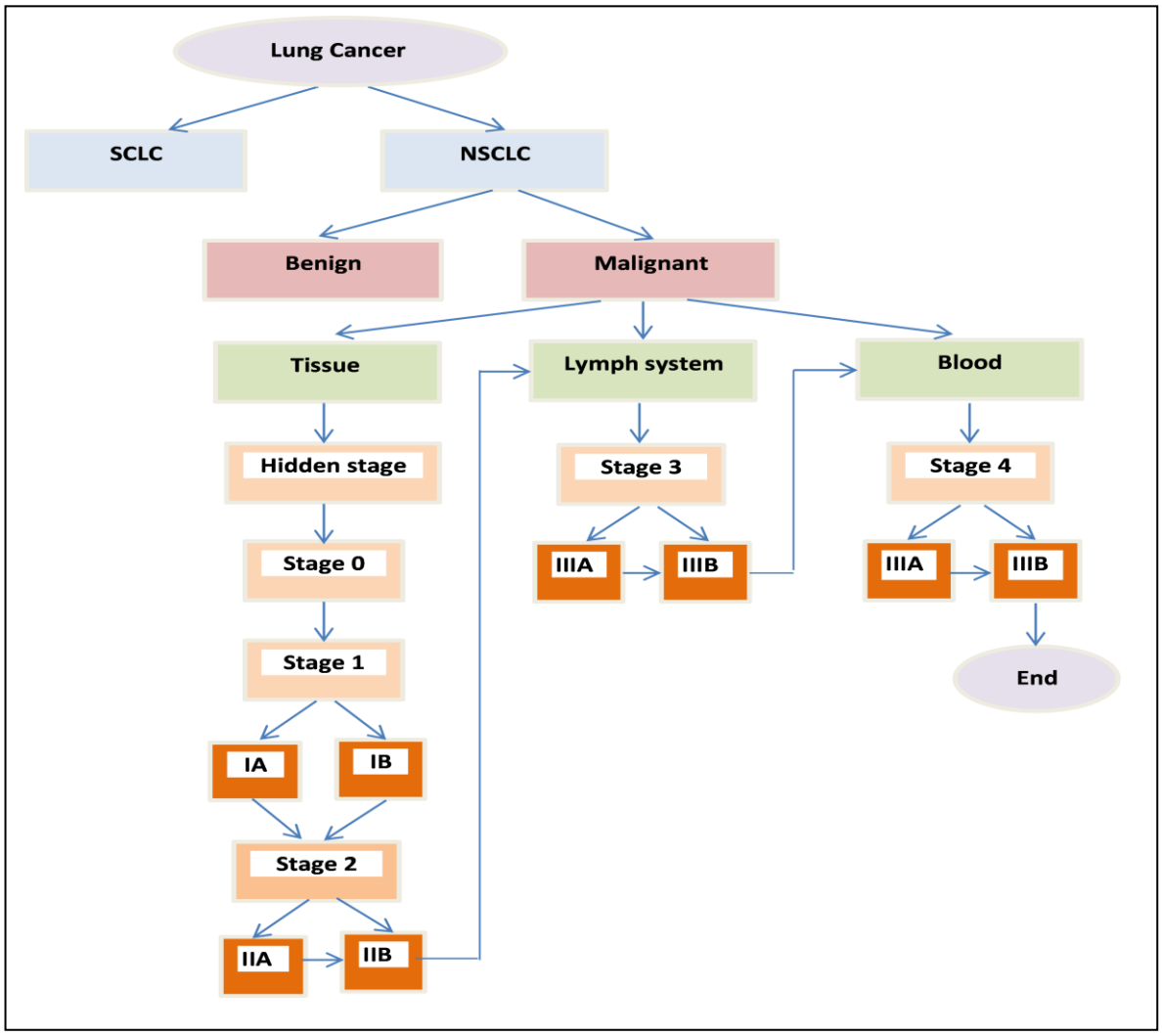


Figure 2.1 Structure of lung cancer from hidden stage to stage 4

2.2. Methods of diagnosis

2.2.1. Laboratory tests

- **DNA**

The diversity between transcript sequence in genes and their final product protein level and function. Translational and post-translational modifications regulate the protein expression in tumor cells in particular. Changing levels and functions of various proteins add another level of complexity in the regulation of cancer cells biology for sustaining proliferative signaling [10].

- **RNA**

It can serve as potential biomarkers for NSCLC with high accuracy, as its expression levels between NSCLC patients and healthy controls show significant differences [11].

- **Blood**

❖ Blood proteomic analysis have a great advantage over proteomics conducted in lung cancer tissue because blood samples are more readily accessible.

❖ Blood based tests for screening purposes or disease monitoring would be more suitable as they are minimally invasive, relatively have low cost and can be repeated as well.

❖ Blood should be examined as plasma rather than as serum and established standardized sample collection protocols [12].

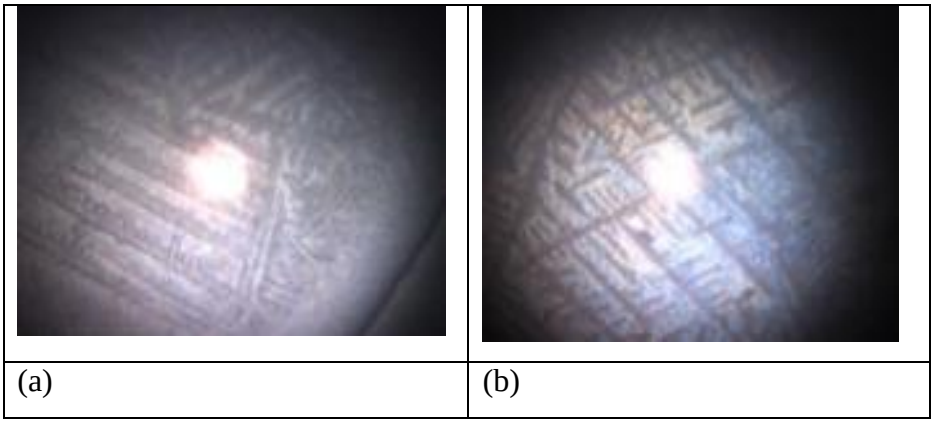
- **Serum**

❖ Serum based markers suitable for tumor detection are limited.

❖ Almost all of the serum markers currently in use are proteins and comprehensive approaches on proteomics have been applied [13].

- **Saliva**

Human saliva is a biofluid especially useful for early detection of various oral and systemic pathological condition .Used as discriminatory biomarkers to differentiate patients with early NSCLC from healthy control subjects. Figure 2.2 shows two surface enhanced Raman spectrum of the saliva samples [14].



(a) (b)

Figure 2.2 Saliva examination images (a) microscope image of the saliva sample from a healthy individual (b) microscope image of the saliva sample from a lung cancer patient.

- **Bronchoscope**

Bronchoalveolar is a diagnostic method that is suitable for the detection of asymptomatic lung cancer in patients that present suspicious masses in chest X-Ray or CT scan. Figure 2.3 shows cell in Bronchoalveolar test [15].

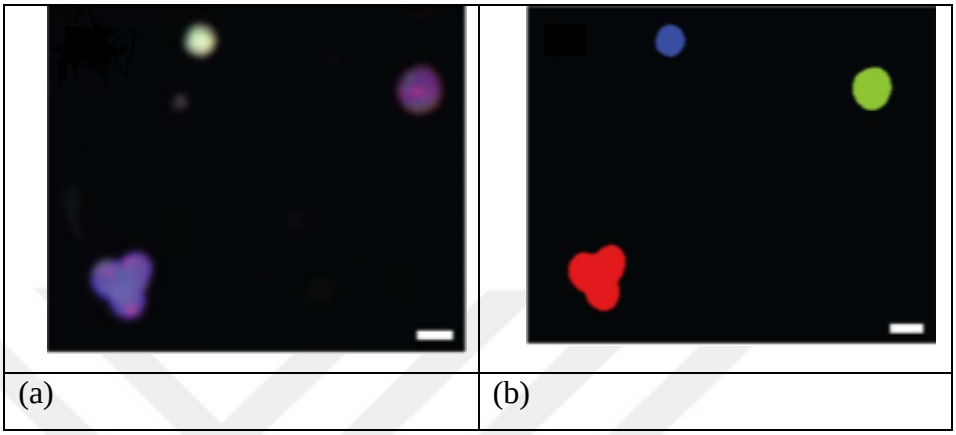


Figure 2.3 Illustration of the image analysis workflow (a) Color composite of the three acquired channels (scale bar (b) Result of the immunopheno typing step.

- **Sputum cytology**

Sputum Cytology is an economical, non-invasive and practical method for early lung cancer detection, which addresses some of the issues encountered in other modalities, such as cost, complexity, and radiation exposure. Figure2.4 shows sample of sputum cell and converted to gray cell [16].

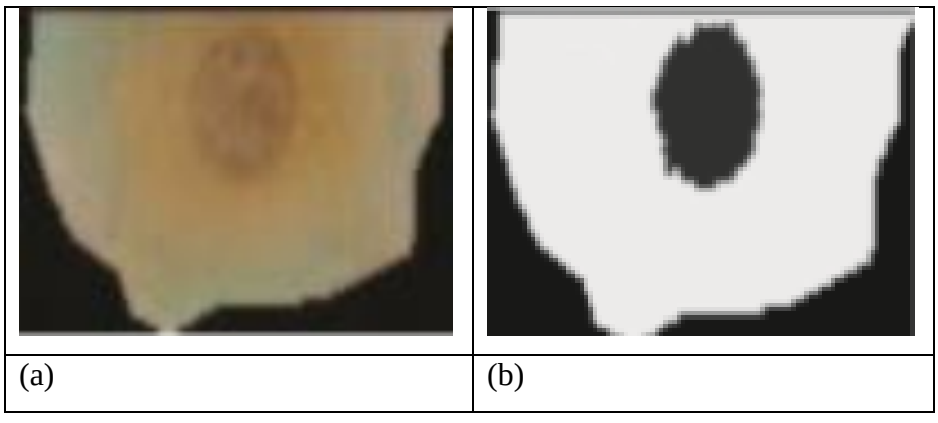


Figure 2.4 Sputum examination images (a) Sputum cells (b) Conversion to gray level.

2.2.2 Physical test

An examination of the body to check general signs of health, including checking for signs of disease, such as lumps or anything else that seems unusual. A history of the patient's health habits, including smoking and past jobs, illnesses and treatments will also be taken.

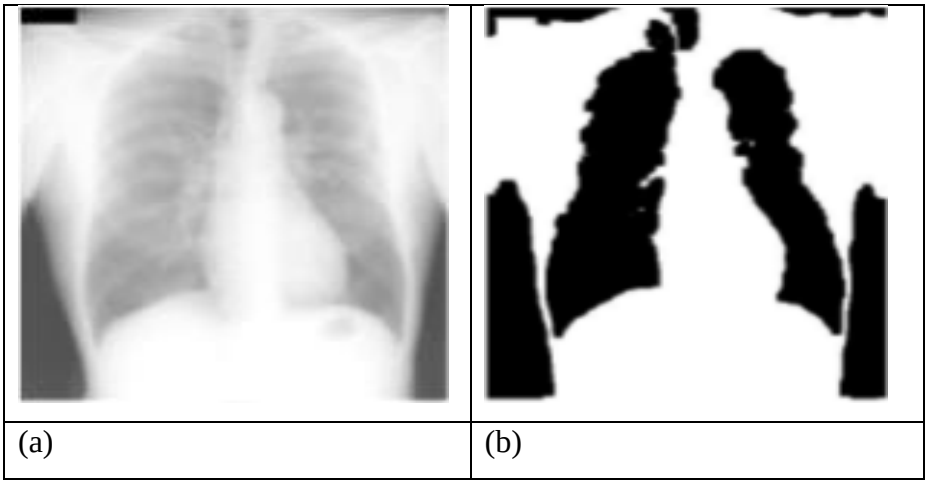
- **Imaging tests**

Imaging tests use x-rays, magnetic fields, sound waves, or radioactive substances to create pictures of the inside of body. Imaging tests have been done for a number of reasons before and after a diagnosis of lung cancer:

- ❖ Find a suspicious area, it may be cancerous.
- ❖ It gives hint how cancer can be spread.
- ❖ It helps doctor to determine if treatment has been effective.
- ❖ It gives information for possible signs of cancer coming back after treatment.

- **X-ray Chest**

It is considered as the first test of doctor. In a black and white x-ray image, a mass tissue looks like a gray shadow and sometimes becomes very hard to detect. If something suspicious is seen, the doctor may order more tests. Figure2.5 shows x-ray image and grey level of it [17].

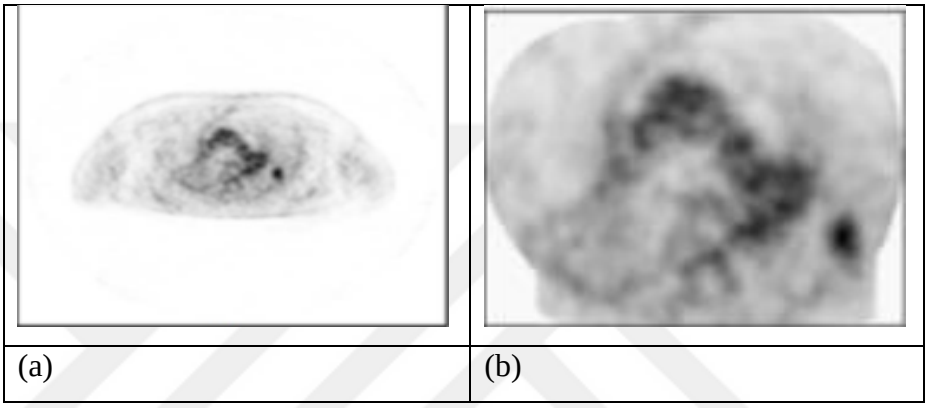


(a) (b)

Figure 2.5 X-ray imaging **(**a) x-ray image (b) Conversion to gray level.

- **PET scan**

It is very important test if tumor appears to have early stage lung cancer. Doctor can use this test to see if the cancer has spread to nearby lymph nodes or other areas, it has high contrast and reveals increased metabolism in structures with rapidly growing cancer cells, but their localization is limited by the low spatial resolution in PET images. It is helpful in getting a better idea whether an abnormal area on a chest x-ray or CT scan might be cancer. Figure 2.6 shows PET image and gray level of it [18].

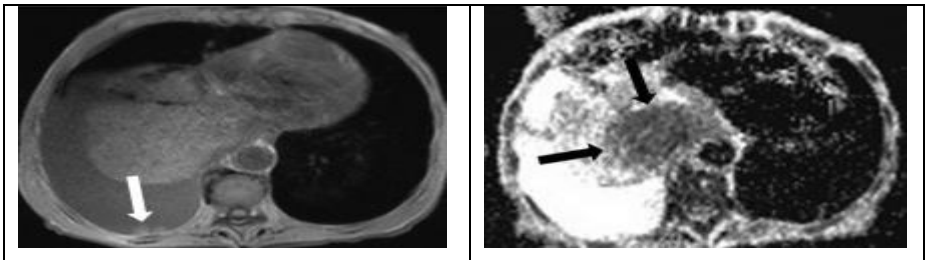


(a) (b)

Figure 2.6 PET imaging (a) PET image (b) Conversion to gray level.

- **MRI scans**

Scanning methods such as MRI scans and CT scans provide detailed images of soft tissues in the body. But MRI scans use radio waves and strong magnets instead of x-rays. The energy from the radio waves is absorbed and then released in a pattern formed by the type of body tissue and by certain diseases. A computer translates the pattern into a very detailed image of parts of the body. Figure 2.7 shows MRI image and gray level of it [19].

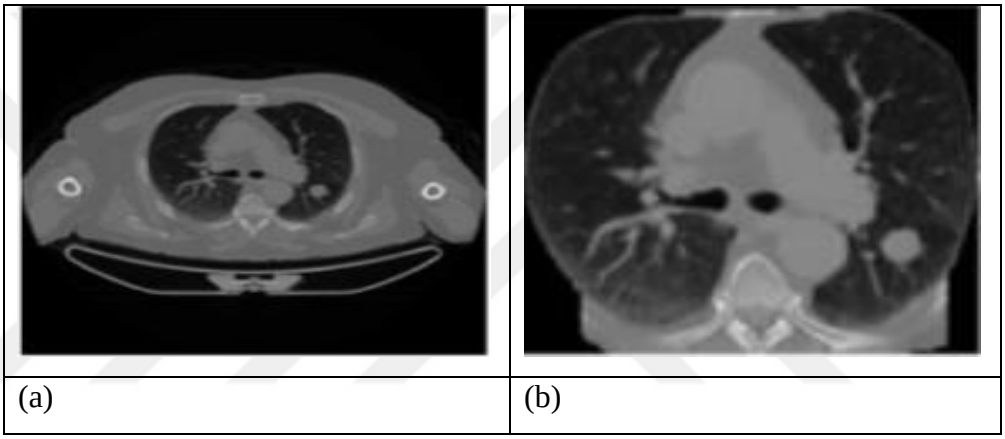


(a) (b)

Figure 2.7 MRI scanning images (a) MRI image (b) Conversion to gray level..

- **CT scan**

CT scan is more likely to show lung tumors than routine chest x-rays. It provides anatomical information about the size, shape and position of any lung tumors and can help to find enlarged lymph nodes that might contain cancer that has spread from the lung, it has relatively low soft tissue contrast causing difficulties in separating abnormalities from the surrounding tissues. CT scanner takes many pictures as it rotates around human body. A computer then combines these pictures into images of slices of the part of human body being studied. Unlike a regular x-ray, a CT scan creates detailed images of the soft tissues in the body. Figure 2.8 shows CT image and gray level of it [20].



(a) (b)

Figure 2.8 CT scanning images (a) CT image (b) Conversion to gray level.

2.2.3 Why CT scan is common for detecting lung cancer

Most of the CAD system are implemented on CT scan images because CT images are quickly obtained and do not damage the bones of the patient [21]. These kinds of images have very less noise effect so working on CT scan images is more preferred [22]. The CT scan images give 3D analysis of the internal body parts and organ analysis is easy because it is taken at different angles. The CT (computed tomography) devices are able to identify lung cancer at an earlier stage than other kinds [23]. For detection of lung cancer, CT has been widely used. In CT image, pulmonary nodule denotes lung cancer. So, to detect pulmonary nodule earlier from CT images contributes to improve survival rate. It provides ability to detect very small nodules improves with each new generation of CT scanner. The morphology,

biologic characteristics, and growth rates of small lung cancers has become available from CT lung cancer screening [24]. People who have been treated for lung cancer often have follow-up tests, including CT scans to see if the cancer has come back or spread. In recently years the National Lung Screening Trial (NLST) is American National Cancer Institute recommendation people have lung cancer to test themselves by chest CT exams could reduce death rates from lung cancer among those at high risk for the disease [25].

CHAPTER THREE

FUNDAMENTALS OF BIOMEDICAL IMAGE PROCESSING

In recent years, medical diagnostics depended on direct digital image processing systems become mostly important for health care issues. In addition to originally digital methods, such as Computed Tomography (CT) or Magnetic Resonance Imaging (MRI), initially analogue imaging modalities such as endoscopy or radiography are nowadays equipped with digital sensors [26].

Digital images consist of individual pixels which are separated as brightness or color values. They can be processed as target evaluated. They are founded at many places at the same time by means of convenient communication networks and protocols, such as Picture Archiving and Communication Systems (PACS) and the Digital Imaging and Communications in Medicine (DICOM) protocol, respectively. Based on digital imaging techniques, the entire spectrum of digital image processing is now applicable in medicine. Image processing covers four main areas [27]. Figure 3.1 shows modules of image processing.

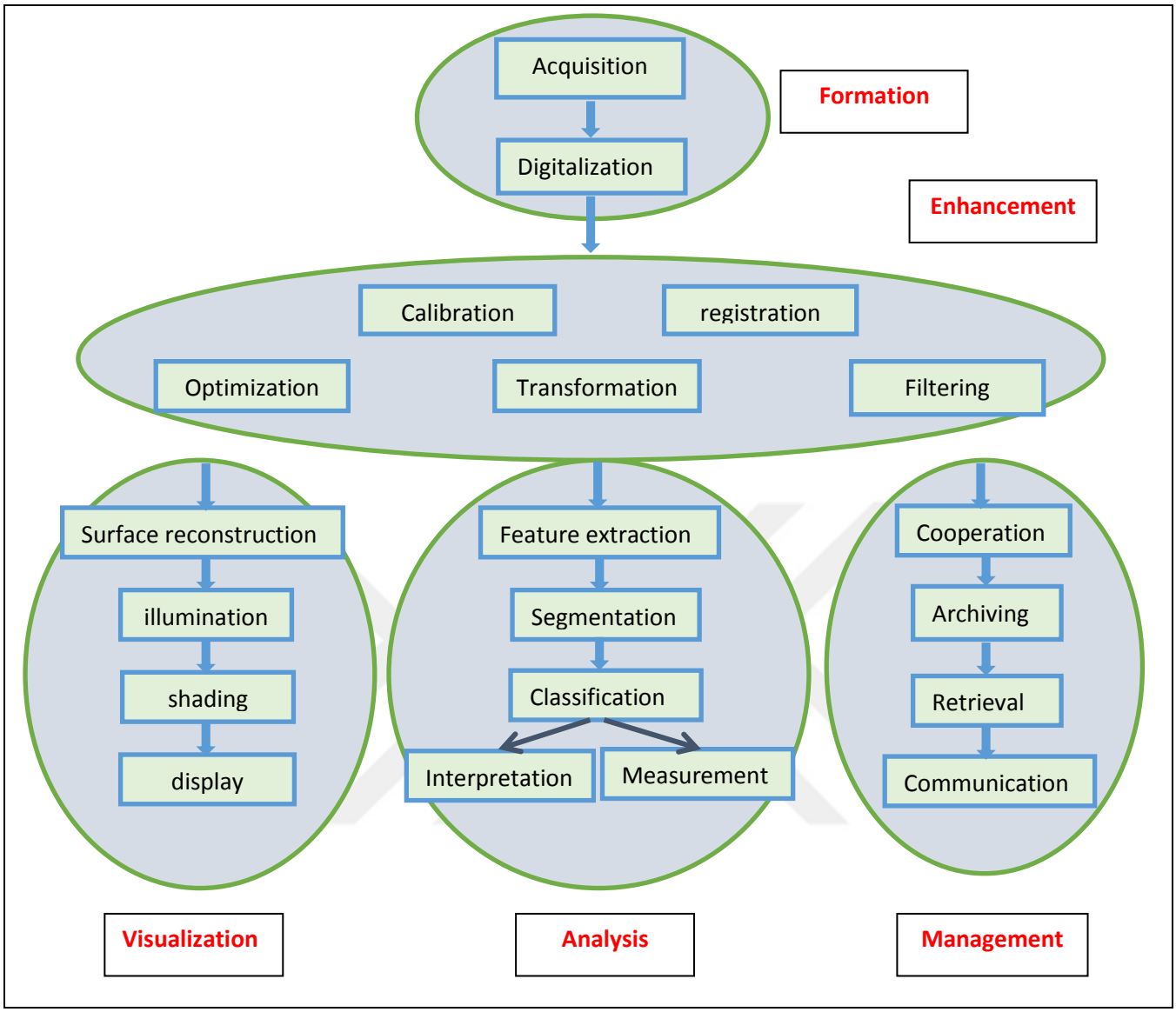


Figure 3.1 Modules of image processing

3.1. Steps of Image Processing

The popular term "biomedical image processing" means provided digital image processing for biomedical sciences. In general, digital image processing covers four major areas shown in Figure 3.1:

- **Image Formation:** It obtains the steps that are starting from capturing the image to forming a digital image matrix.
- **Image Enhancement:** Extraction of interested features in an image such as changing brightness.

- **Image Visualization:** It means all kinds of doctrinaire of this matrix resulting in an optimized output of the image.
- **Image Analysis:** It means all stages of processes are used for quantitative measurements, abstract interpretations of biomedical images. These stages need priority of knowledge on the nature and content of the images mostly combined into the algorithms on a high level of abstraction.
- **Image Management:** It consists all techniques that provides the efficient storage, communication, transmission, archiving and access (retrieval) of image data.

The contrast of image analysis assigning to high-level image processing, low-level processing denotes manual or automatic techniques. It can be realized without a priority knowledge on the specific content of images [27][28].

3.2. Biomedical Image Processing

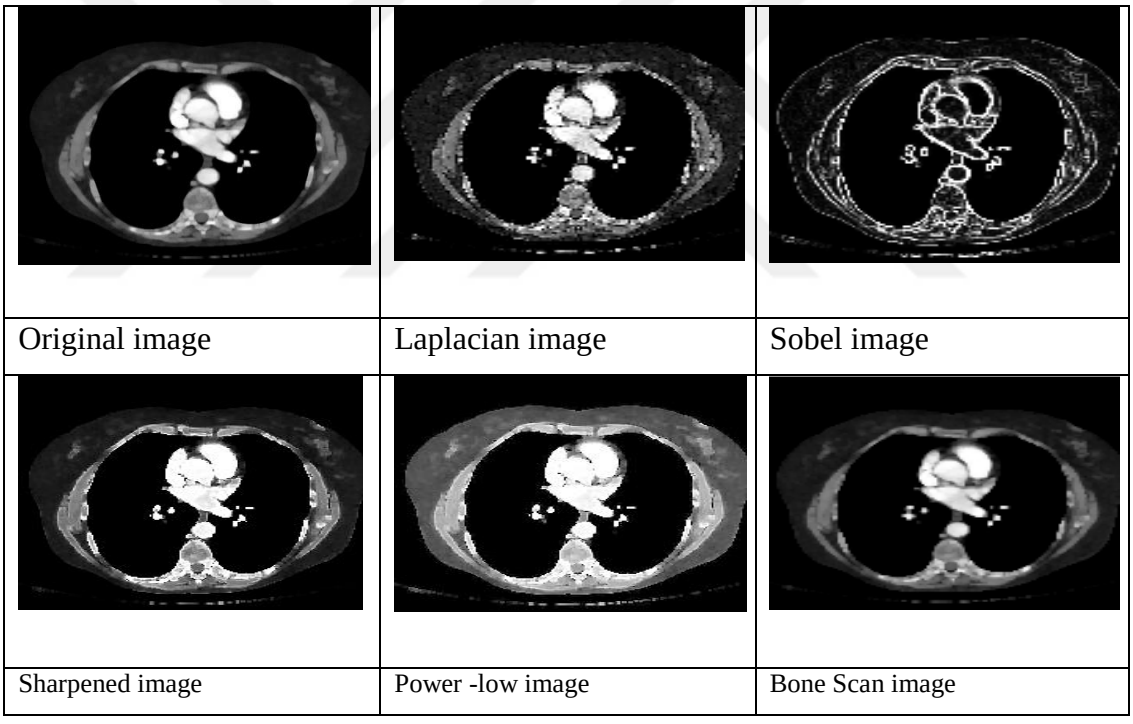
The main problem of biomedical images represented by high-level processing have complex nature, it is difficult to formulate medical as a priori knowledge, it can be easily processed into automatic algorithms of image processing. It refers to semantic gap that means difference between diagnostic image by the physician (high level) and the simple structure of discrete pixels, which is used in the program. In the medical domain, there are three main aspects bridging this gap:

- **Heterogeneity of images:** Medical images show living tissue, organs, or body parts.
- **Unknown delineation of objects:** Biological structures that are important parts for diagnoses therapeutically relative object is represented by the entire image, however cannot be discreted from the background. If the objects are notes in biomedical images, the segmentation is a problem because the border or shape itself is represented fuzzily or only partly, therefore in biomedical images depended mostly the texture level to extract the result of diagnoses.
- **Robustness of algorithms:** the properties of medical images depended of high-level processing, special requirements of reliability and robustness of medical procedures. If images do not processed correctly, classification

processes should be rejected. Then all images that have not been rejected must be evaluated correctly[29].

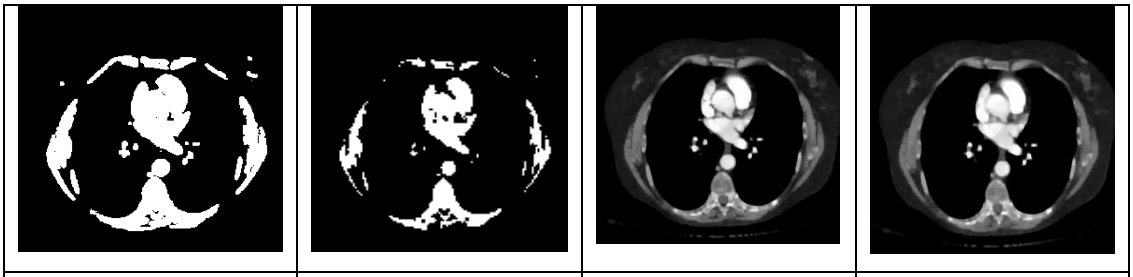
3.2.1 Image Enhancement methods

The procedures and algorithms are performed without a priori knowledge in Low-level methods of process. The specific content of an image, are widely applied to pre- or post-processing of medical images. There are some methods of histogram transforms, convolution and (morphological) filtering. Figure 3.2 shows enhancement in (Laplacian, Sobel, Sharpened, Power-low and Bone Scan filter). Figure 3.3 shows enhancement in morphology operations (dilation, erosion, opening, closing) [30].



Original image	Laplacian image	Sobel image
Sharpened image	Power -low image	Bone Scan image

Figure 3.2 Different enhancement methods

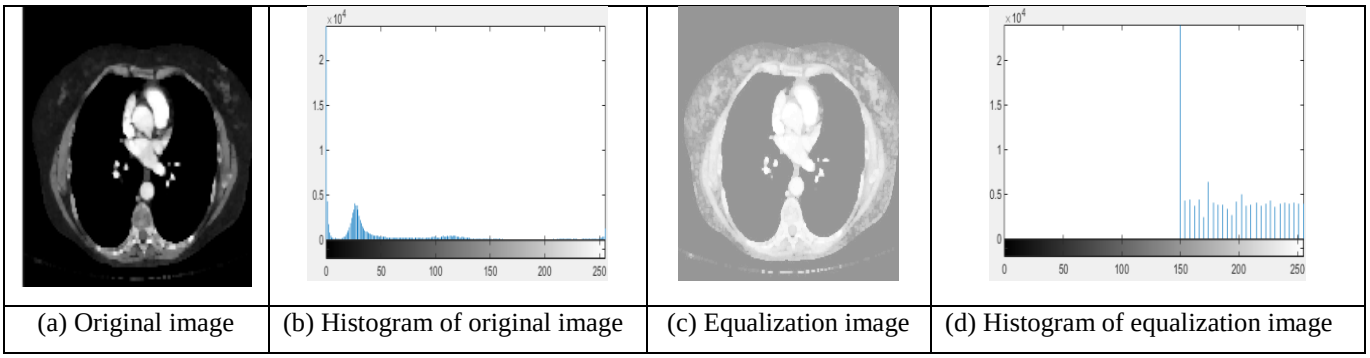


Dilation image	Erosion image	Opening image	Closing image

Figure 3.3 Morphology operations

3.2.2 Histogram Transforms and histogram equalization

Point operations are based on the histogram of the image. The pixel values have been modified depended on their positions in the image and their immediate neighborhood. However the types of transform refer to point operation. The histogram applies the frequency of pixel values is distributed on image. Grayscale image disregarding the certain positions where the gray scales occur in the image. Simple pixel transforms can be defined by using a histogram. Upper and lower bounds are located, after determining the histogram (lower bound to zero and the upper bound to 255 (maximal gray scale for 8 bit image). Enhanced contrast means distance between neighbored pixels of image is increased more than histogram of the initial image does not contain all possible gray scales. In enhancement contrast image becomes clear. Figure3.4 shows histogram transformation and histogram equalization [31].



(a) Original image	(b) Histogram of original image	(c) Equalization image	(d) Histogram of equalization image

Figure 3.4 Types of histogram images (a) Original image (b) Histogram of original image (c) Equalization image (d) Histogram of equalization image.

3.2.3 Convolution

The pixels in histogram transforms are combined with the values of their neighborhood during applied separate filter. The underlying mathematical operation, convolution that can be characterized is called templates. A template is a small, squared mask generally has odd lateral length [32]. The template is a reflect along two axes the name “convolution” is commonly used and position of one corner of the input image. The pixels' image under the mask are named kernel. Each pair of corresponding pixel values of template and kernel are multiplied and then summed up. The result has been registered at the position of the mask’s center pixel in the output image. Then, the template duty is shifted row by row and column by column to the next positions on the input image, until all the positions have been visited, and thus, the output image has been calculated completely.

The effect's filter determines the pixel values of the template. If template uses only positive value, it is weighted average is calculated in the local neighborhood of each pixel. As a result, the image appears with low noise (smoothing) also sharpness of edges is reduced [29].

3.2.4 Image Data Visualization

In medicine, the realistic visualization is performed on three dimensional images. These techniques have found some applications in medical research like diagnostics, treatment planning and therapy. In contrast to problems from the general area of computer graphics, the object is displayed in medical implementation are not included by format, mathematical expressions, but as an explicit set of voxel, therefore specific methods have been established for medical visualization [33].

3.3. Segmentation

Segmentation mostly means dividing an image into united area. It means emphasizing the pre-stage of classification. The main point in medical image processing for diagnosis is to distinguish healthy anatomical structures from pathological tissue. By definition, the result of segmentation is always on the regional level of abstraction. It is depended on level of feature extraction as an input

to the segmentation, the classification methods such as pixel, edge, texture and region based procedures [34].

3.3.1 Pixel-Based Segmentation

It has procedures of segmentation depended on two kinds of images, grayscale image and color image. The value of current pixel is regardless of its surroundings. Each pixel is considered only isolated from its neighborhood, uncertain that only connected segments are obtained. For this reason, it is necessary to apply post-processing stage [35].

3.3.2 Post-Processing

Segments obtained from pixel based analysis usually are discontinuous and highly noisy. However, post-processing is necessary. In mathematical morphology methods, effect of noisy structures can be reduced. Morphological opening removes spread parts from the segments, holes are closed by morphological closing. The connected components algorithm provides each separated segment with a unique reference number. After morphological post-processing and connected components analysis, cells are separated and colored (labeled) differently according to their segment numbers [36].

3.3.3 Edge-Based Segmentation

This type of segmentation is depended on finding the object that is closed outline in the image. Edge segmentation process is only used for such problems, if object that is clearly in image can find the boundaries easily in the biological tissue.

Generally, the image processing chain for edge-based segmentation consists of edge extraction and edge completion [37]. Edge extraction is getting edge feature extraction, for example creating with the Sobel filter. The next step is binarization process that obtains only edge pixels and non-edge pixels. Morphological filtering decreases noise and artifacts, finally a skeleton of the edge is computed. The main tasks of the edge segmentation are tracing and closing of binary contours.Figure3.5 shows edge of tumor area [38].

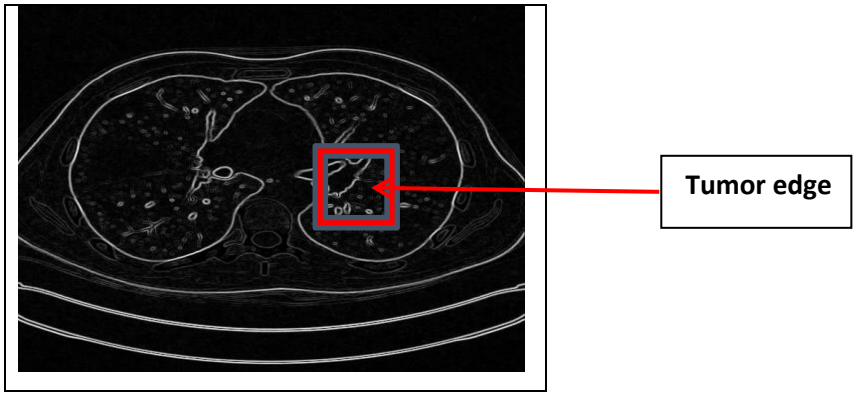


Figure 3.5 Edge detection for segmentation

3.3.4 Region-Based Segmentation

The idea of region-based segmentation is to represent only connected segments so the morphology methods are avoided. All approaches are based on a certain distance or similarity measure to guide the assignment of neighbored pixels or regions. Large numbers of methods are used like in grayscale image compare to calculate the mean parameter for gray value on image. For example, pixel clustering, which has been already introduced for segmentation, is an unsupervised classification process. The aim of isolating target is divided into similar groups. If classification is used for identical the target, the model reference should be available from which the ground truth of classification can be created. The features of these samples are then used for parameterization and optimization of the classifier. Figure3.6 shows tumor area segmentation [39].

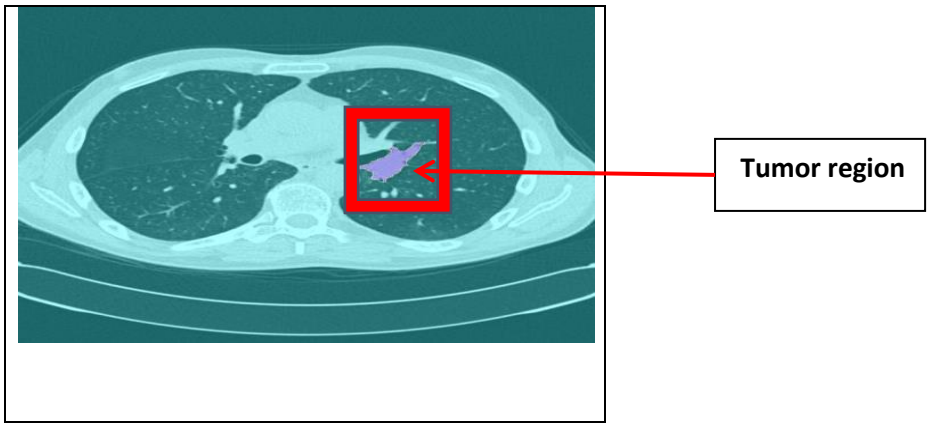


Figure 3.6 Region segmentation

3.4. Feature Extraction

Feature extraction is the first stage of image analysis depended on intensity value of image applied after segmentation and classification. Therefore, the task of feature extraction is to describe properties to emphasize image information on the particular level by algorithm. Then, information provided on other levels must be suppressed. Data reduction is important stage to determinate properties. Segmentation and feature extraction have been done at various levels of abstractions gradually, before classification is eventually performed at a high level of abstraction. Before classification, the feature extraction is a base on the region level to performance well [40].

3.4.1 Data Level

Collective information of all pixels are based on feature extraction. Therefore, all transforms manipulating the whole matrix of an image at once can be regarded for data feature extraction. Fourier transform is a most famous example of a data feature transform is describing two dimensional images in terms of frequencies, according to their amplitude and phase [41].

3.4.2 Edge Level

Edge-based features are defined as local contrast, a big difference between grayscale image and color image values of adjacent pixels.

3.4.3 Texture Level

Texture analysis attempts to quantify objectively the homogeneity in a heterogeneous but at least subjectively periodic structure.

- Structural approaches that are based on texture primitives (textone, texture element) and their rules of combinations,
- Statistical approaches that describe texture by a set of empirical parameters.

3.4.4 Region Level

Regional features are used primarily for object classification and identification. They are normally calculated for each segment after the segmentation process.

- Localization-descriptive measurements such as size, position, and orientation of the major axis.
- Delineation-descriptive measures such as shape, convexity and length of the border[42].

3.5. Classification

The classification process refers to all connected regions, which are getting from the segmentation step to particularly specified classes of targets. Region features are important steps before the classification processes. In addition, another feature extraction step is executing segmentation between the classification processes. There are lots of algorithms that have been used for classification of the data after feature extraction like neural network, Weka,…etc. [43] [44].

CHAPTER FOUR

FILTERING AN ENHANCEMENT AND SEGMENTATION

Preprocessing

4.1. Sobel Filter:

The sobel operator or sometimes called as sobel filter, is used in image processing to compute the edge detection algorithm for creating some images to be emphasized and translated. The idea of sobel filter is technically to describe a gradient operator to compute approximation on the gradient of image using intensity function at each point of the image. The result of the sobel operator is corresponding gradient vector or the norm of the vector. The sobel operator is the base to compute the image in small separable integer values filtering it in the vertical and horizontal direction [46]. Sobel filter is used in our study in order to make the small size tumors more visible because Sobel filter is a suitable method to find a small size tumor in CT scan which is not clear [47].

There are lots of edge detection methods that suppose the edge occurrence as a discontinuity in the intensity function or a very steep intensity gradient in the image. By using the assumption, if one takes the derivative of the intensity value across the image and finds points where the derivative is maximum, then the edge could be located. The gradient is a vector, it used for measuring the rapid pixel value that is changing with distance in the x and y direction. Thus, the components of the gradient may be found using the following approximation:

$$\frac{\partial f(x,y)}{\partial} = \Delta x = \frac{f(x+dx,y) - f(x,y)}{dx} \quad (1)$$

$$\frac{\partial f(x,y)}{\partial} = \Delta y = \frac{f(x,y+dx) - f(x,y)}{dy} \quad (2)$$

where dx and dyrepresent the distance along$x$ and y directions respectively. In discrete images, one can consider dx and dy in numbers of pixels between two points. $dx = dy = 1$(pixel spacing) is the point at which pixel coordinates are (i, j) , thus,

$$\Delta x = f\ (i+1, j)\ - f\ (i-1, j) \tag{3}$$

$$\Delta y = f\ (i, j-1)\ - f\ (i, j+1) \tag{4}$$

In order to detect the presence of a gradient discontinuity, one could calculate the change in the gradient at (i, j)[47]. This can be done by finding the following magnitude measure:

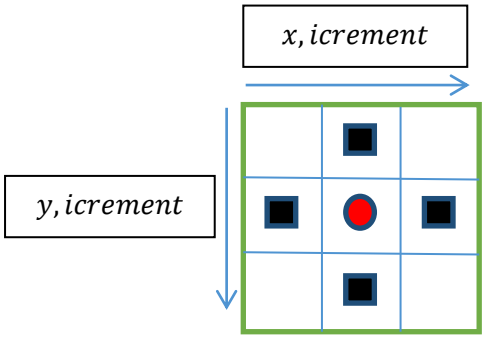
$$M = \sqrt{\Delta x^2 + \Delta y^2} \tag{5}$$

and the gradient direction θ is given by

$$\theta = \tan^{-1}\frac{dy}{dx} \tag{6}$$

❖ **An example: how to calculate gradient and magnitude in an image**

We can demonstrate the idea of this example by supposing we have an I image 3×3 mask and in the center of this image we have y pixel representing the rows and x pixel representing the columns. We can calculate the neighborhood surrounding the center of image's pixel in horizontal y and vertical x direction.



●	Center of pixel
■	$I(x, y-1)$neighborhood of image by gradient, (north pixel of the center pixel)
■	$I(x, y+1)$neighborhood of image by gradient, (south pixel of the center pixel)
■	$I(x+1, y)$neighborhood of image by gradient,(east pixel of the center pixel)
■	$I(x-1, y)$neighborhood of image by gradient,(west pixel of the center pixel)

We need these values to compute the intensity changes of the image.

$dy = I(x, y-1) - I(x, y+1)$, the change between the north pixel and the south pixel,

$dx = I(x+1, y) - I(x-1, y)$, the change between east pixel and west pixel and then $I(x, y)$ can be calculated by using the equation$I(x, y) = \frac{dy}{dx}$.

How to calculate the gradient orientation and the gradient magnitude can be explained as follows;

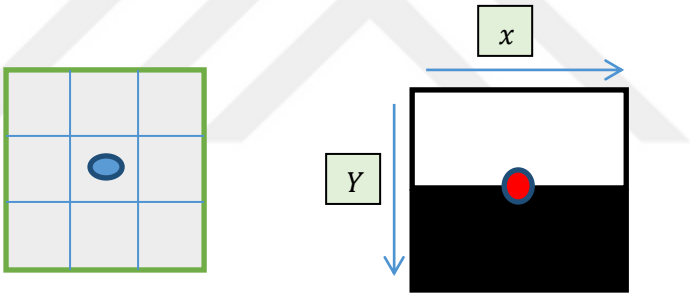


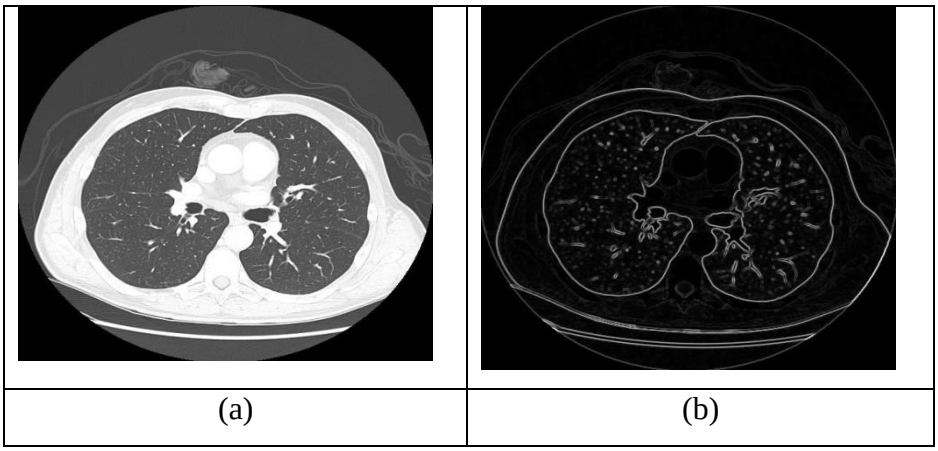
Figure 4.1 Calculate the gradient orientation and the gradient magnitude

These steps can be explained as follows,

1. Read 3D (original image) CT scan grayscale,
2. Convert 3D (original image) to 2D grayscale,
3. Find the center pixel value of the image, calculate sobel mask for x-direction and y- direction, sum the value of pixels (i, j) and its neighbours, Figure 4.1 shows implementation of sobel filter on the image:

$$\Delta x = f(i+1, j) - f(i-1, j) \quad (3)$$

$$\Delta y = f(i, j-1) - f(i, j+1) \quad (4$$



(a) (b)

Figure 4.2 Representation filtering process **(**a) Original image (b) Sobel image.

4.1.1 Area edge detection gradient

Gradient is one of the multiplying definitions conceptually to find a direct changes in image intensity or color. The gradient method in sobel filtering detects the edges of image by looking for the maximum and minimum values in the first derivative (dy/dx) [48].

To find a small size tumor which is not clear in the original image, the gradient method in sobel filter detects the edges of image by looking for the maximum and minimum values in the second derivative.

$$M = \sqrt{\Delta x^2 + \Delta y^2} \quad (5)$$

In this study, we used CT scan image (grayscale), however we focused about intensity of this image. Figure 4.2 shows the gradient magnitude of CT scan image.

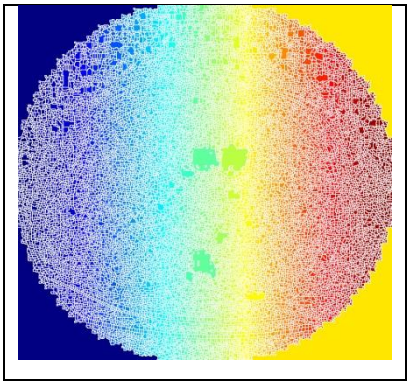


Figure 4.3 Implementation of gradient magnitude

4.2. Processes

Segmentation plays an important role in medical imaging as it helps extracting the organ of interest. For the diagnosis of the lung cancer, it is very necessary to segment the chest images and extract the lungs in the preprocessing step, especially in early stages and further the nodules are segmented with the different methods when tumor does not appear in image [50].

4.2.1 Morphology

Morphology is a branch in biology which is studied about the size, shape and structure of animals, plants, microorganisms and the relationships between the parts comprising them [51].

Mathematical Morphology is a tool for extracting image components that are useful in the representation and description of region shape. This morphology method that was started to develop in the late 1960s, stands as a relatively separate part of image analysis. Mathematical morphology uses the concept of mathematical set theory for extracting meaning from the image. This method that is based on Set theory is a unified and powerful approach to numerous image processing problems. In binary images, the set elements are members of the 2-D integer space – Z^2 where each element *(x,y)* is a coordinate of a black (or white) pixel in the image [52].

4.2.2 Binary dilation and erosion

The description of a binary image is a set of black and white pixels format. Assume that white pixel is a background; black pixel is the aim of the process. The dilation and erosion are primary morphological operations. Opening, closing is more complex than the dilation and erosion.

- **Dilation**

Dilation is an extended image that is in the same shape as the original, but has a different size. It makes figure bigger than the original by stretching or shrinking it [53]. Dilation increases the holes and enlarges the width of maximum regions, so it can remove negative impulsive noises but does little on positive ones [54].

A is an image, B is a structuring element inside the image A. It checks across the image in a way such as convolution, this process is defined as dilation operation. It consists of two main inputs for the dilation operator [55], one is the image which is dilated and the other is a set of regulate points that is recognized as a structuring element and also known as a kernel. The effect of the dilation execution on the input image is fixed by this structuring element [56].

$$\boldsymbol{A} \oplus \boldsymbol{B} = \{\boldsymbol{Z} | (B)_z \cap \boldsymbol{A} = \varnothing\} \quad (7)$$

- **Erosion**

The execution shrinking or thinning of the object is defined as erosion. The structuring element which is decided to extend this operation [52] is used for decreasing the objects in the image peaks while expanding the width of minimum regions, so it take out positive noise but affecting negative impulsive noises little [53].

The erosion process is similar to dilation but the pixels are converted to 'white', not 'black'. It consists of two main inputs for the erosion operator [57] one is the image which is eroded and the other is a set of regulate points that are recognized as a structuring element and known also as a kernel. Erosion execution on the input image is fixed by this structuring element [58].

$$\boldsymbol{A} \ominus \boldsymbol{B} = \{\boldsymbol{Z} | (B)_z \subseteq \boldsymbol{A}\} \quad (8)$$

Erosion enlarges holes, break thin parts and shrink the object. Figure 4.6 shows structure element in erosion operation implementation [59].

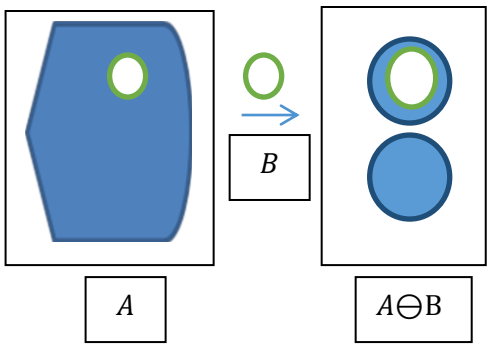


Figure 4.4 Structure element in erosion operation

4.3. Clustering

4.3.1 Opening and Closing Operations

❖ **The opening operator**

Dilation and erosion are the two main operations, producing more complex sequences. Opening and closing are the most important methods for morphological filtering [60]. An opening operation is known as erosion followed by a dilation. The same structuring element is used for both operations .The base of the two main inputs for opening operator are an image (open and a structuring element) .

Using the structure element B to do the open operation on the set A, expressed as $A \circ B$ definite as in (7);

$$f \circ b = (f \ominus b) \oplus b \qquad (9)$$

❖ **The closing operator**

Closing operator performs like opening operator, but reverse of it. It means dilation then erosion is applied and the same structuring element is used for both operations. The base of two main inputs for closing operator are an image (close and a structuring element).

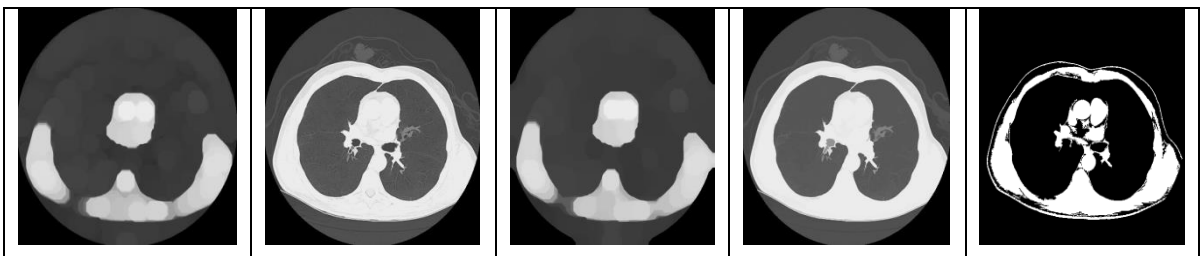
In binary image, at opening operation, bis used to erode f plainly first, then on the results obtained, b is used to do dilation operation. Also, using b to do closing operation on f, expressed as f●b, is definite as in (8) [61];

$$f \bullet b = (f \oplus b) \ominus b \tag{10}$$

4.4. Morphological Reconstruction

Distance transformation can transform each target of tumor area into a peak shape and the whole image of the target is expressed as a seat towering peaks. The process of CT scan has multiple peaks at the top of the first order peak. As much as possible to remove a peak is unnecessary to avoid a problem of follow-up image over-segmentation. This procedure needs morphology-based reconstruction transformation. It can consider two-dimensional image as a signal to level the peak of the image. Peaks in the image represented the high-frequency part, it helps to detect the tumor region, otherwise low-frequency part remove points that are undesirable.

The advantages of morphological reconstruction are that morphological reconstruction only needs to select a contrast parameter, the running speed of morphological reconstruction is almost constant and morphological reconstruction can maintain the shape of the border of the target (tumor) better [62]. Figure3 shows implementation the opening, closing operations in morphology and also reconstruction in of them.



Opening	Opening-by-reconstruction	Opening-closing	Opening-closing by reconstruction	maxima of opening-closing by reconstruction

Figure 4.5 Implementation of o morphology operation

4.5. Threshold

Threshold is one of the methods that is widely used for image segmentation. It is suitable for isolated foreground from the background by selecting a threshold value,gray image can be converted to binary image. The binary image contains vital data about the position and shape of the objects of interested foreground. It is used to reduce complexity data to make it simple for recognition and classification. The most common method has been used by converting a gray-level image to a binary image by selecting a single threshold value.Then in all the gray level values below to threshold will be classified as black (0 pixel), and those above threshold will be white (1 pixel).The segmentation problem is how to select the suitable value for the threshold. A frequent method used threshold by analyzing the histograms is depended the type of images required to be segmented. Threshold technique is important techniques in image segmentation. This technique can be expressed as in (1):

$$T = T\,(x, p(x,y), f(x,y)) \qquad (11)$$

where T is the threshold value and x, y are the coordinates of the threshold value point. $p(x,y), f(x,y)$are points the gray level image pixels. Threshold image$g(x,y)$ can be defined as in (9) [63]:

$$g(x,y) = 1 \;\; if \;\; f\,(x,y) > threshold \qquad (12)$$

$$g(x,y) = 0 \;\; if \;\; f\,(i,j) < threshold$$

4.5.1 Marker threshold

The target in gray scale image has different regions of the noise in thetarget nodule that is inside the area of the tissue. Tag region number related directly to the gray level of threshold. The maximum gray image's value is from the peak of the goal of the decision. Tumor area is represented by high intensity values in grayscale of CT

scan. Low gray value will be removed from the tag part of the region tissue. Through extracting the maximum gray performance, the target region of the markers is combined. The gray level of threshold experiment to find an empirical threshold selection, the result of selection of tumor area threshold have importance to take a decision that it is selected as the target particle size.If the gray threshold is high that means the threshold mark of small target areas markers have been removed and gray threshold is very low [64]. Figure 4.5 shows threshold method in opening, closing reconstruction operation.

Figure 4.6 Thresholded opening-closing by reconstruction

4.6. Watershed algorithm

The watershed algorithm consists of three basic segmentation approaches like threshold, edge detection and region based segmentation [65][66]. It provides more stable results than used gradient magnitude, morphological operations individually.

The watershed algorithm steps following by three points [67] that detecting points belong to regional minimum value, pointing the catchment basin / watershed lines and points at which water drop can more likely to pick than other points are specific to watershed.

The watershed lines or catchment basin are represented by the single minimum values which a drop of water falls with certainty. The gradient of the image is founded before watershed is applied.The feature that a pixel will be compared with the neighboring pixel, if it is similar, the pixels are added to shape a region. The exclusion pixel that is still edge of the region is found or the neighboring regions are over merged. At this step, a dam should be builded to avoid integration between two

different regions. After testing the result of all pixels only the top's dam is visible and resultant of the several regions segmentation in the test image.

Let $M1, M2, \dots, MR$ sets denoting the coordinates in the regional minima of an image $\mathbf{g}(\mathbf{x}, \mathbf{y})$, where $\mathbf{g}(\mathbf{x}, \mathbf{y})$, is the pixel value of coordinate $(\mathbf{x}, \mathbf{y})$. Let $\mathbf{C}(\mathbf{Mi})$, be the coordinates in the catchment basin associated with regional minimum M_i and let $(\mathbf{T}[\mathbf{n}]$ be the set of coordinates $(\mathbf{s}, \mathbf{t})$ for which $\mathbf{g}(\mathbf{s}, \mathbf{t}) < \mathbf{n}$ and is given by (1):

$$\mathbf{T}[\mathbf{n}] = \{(\mathbf{s}, \mathbf{t}) | \mathbf{g}(\mathbf{s}, \mathbf{t}) < \mathbf{n}\} \qquad (13)$$

Step1: The boundary value of the pixels is to be found and the minimum value is to be noted. The boundary of the pixels' values of $\mathbf{g}(\mathbf{x}, \mathbf{y})$ is founded. The minimum value is assigned to M_i. Flooding is done by initializing $= \mathbf{min} + \mathbf{1}$. Let $C_n(M_i)$ as the coordinates in the catchment basin associated with minimum (M_i) that are flooded at stage $\boldsymbol{n}$.

Step2: Compute catchment basins. Compute $C_n(M_i) = C(M_i) \cap T[n]$

Cn(Mi)=[1 for $(\mathbf{x}, \mathbf{y}) \in$ and also $(\mathbf{x}, \mathbf{y})$belong to $\mathbf{T}[\mathbf{n}]$ }
Cn(Mi)=[0 otherwise **}**

Step3: Derive the set of connected components**.** Derive the set of connected components in $\mathbf{T}[\mathbf{n}]\mathbf{denoting\ as}$ Q. For each connected component $\mathbf{q} \in$ Q$[\mathbf{n}]$, there are three conditions.

a. If connected component is empty, it represents a new minimum is encountered. If $\boldsymbol{q} \cap \boldsymbol{C}[\boldsymbol{n}-\mathbf{1}]$ is empty, connected component $\mathbf{q}$ is incorporated into $\boldsymbol{C}[\boldsymbol{n}-\mathbf{1}]$ to form $\boldsymbol{C}[\boldsymbol{n}]$ because it represents a new minimum is encountered.

b. If connected components contain at least one connected component it means connected components lay within the catchment basin of some regional minimum.
If $\boldsymbol{q} \cap \boldsymbol{C}[\boldsymbol{n}-\mathbf{1}]$ contains one connected component of $[\boldsymbol{n}-\mathbf{1}]$,connected component q is incorporated into $\boldsymbol{C}[\boldsymbol{n}-\mathbf{1}]$ to form $\boldsymbol{C}[\boldsymbol{n}]$ because it means $\boldsymbol{q}$ lies within the catchment basin of some regional minimum.

c. If connected components contains more than one connected component it represents all or part of a ridge separating two or more catchment basins and set them as "dam".

If $\boldsymbol{q} \cap \boldsymbol{C}[\boldsymbol{n}-\boldsymbol{1}]$ contains more than one connected component of $\boldsymbol{C}[\boldsymbol{n}-\boldsymbol{1}]$, it represents all or part of a ridge separating two or more catchment basins are encountered so that we have to find the points of ridge(s) and set them as "dam".

Step4: Construct dam for all possible catchment basins. Construct C[n] using the values obtained for $C_n(M_i)$ and $\boldsymbol{C}[\boldsymbol{n}]$ Set $\boldsymbol{n} = \boldsymbol{n} + \boldsymbol{1}$.

Step5: Repeat Step 3 and 4 until $\boldsymbol{n}$ reaches max + 1[68].

The main idea of watershed algorithm comes from geography that means the gray value of each point on image is equal to the altitude. Implementation processes are depended of simulated immersion process, simulation of precipitation, but precipitation during the simulation is difficult to convert to digital, simulated immersion process is better than it to performance. The gray pixel's value on image corresponding to the point that means the terrain altitude. In terrain, there are some basins (local minima value in the area of image),ridge is the watershed. The model of terrain at the beginning is immersed vertically in this lake, then the minimum portion basin opens holes, each hole water of basin is slowly immersed in uniform, when the basin has been filled by water, that is two or more basins water will be melt and the dam will be built between the two intersecting dykes, as the water level increasing by degrees [69]. The final step basin was fully flooded, not only dam was flooded, but all the basin is fully surrounded by dams, each dam can represent the watershed and the basin is the target discrete by the dam, so to achieve the purpose of the object segmentation. Figure 4.7 shows the schematic diagram of simulated [62]:

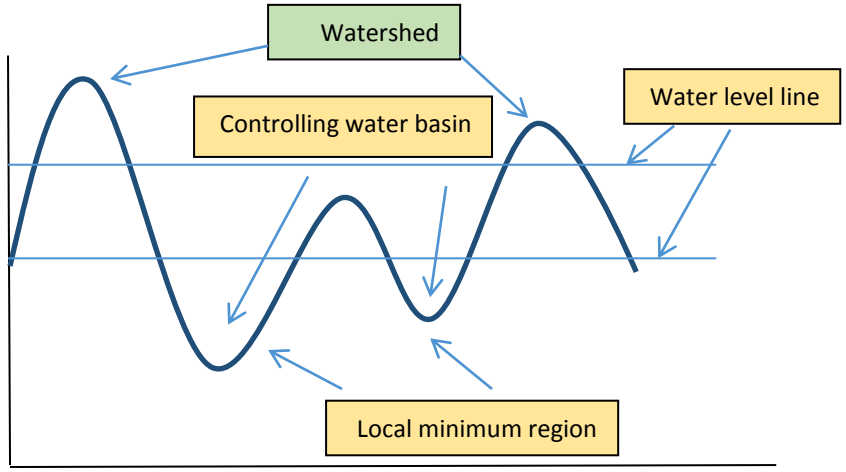


Figure 4.7 Schematic diagram of simulated drowning

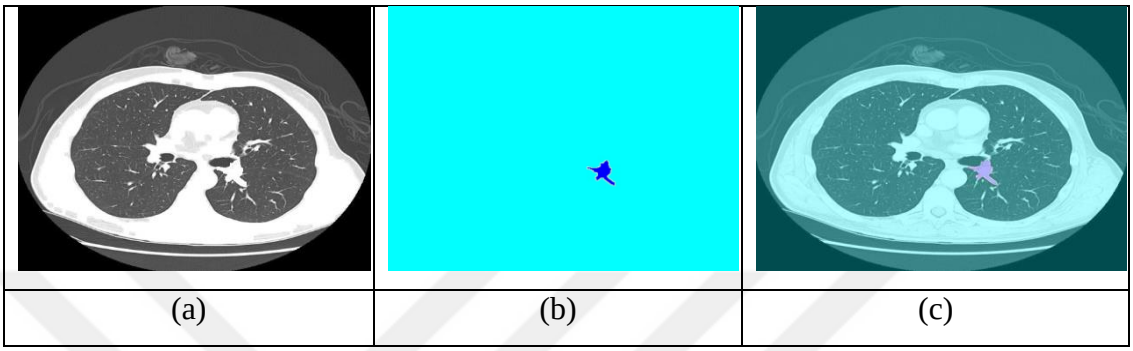
4.6.1 The Over-Segmentation

The use of the single watershed algorithm does not really allow good segmentation because far too many regions are detected. There are two main methods to limit this over- segmentation that are hierarchical watershed segmentation: It does not use in our work and Watershed by markers [70].

4.6.2 Marker Control Method

The application of this algorithm is widely affect the noisy and other local irregularities in the image [71]. This method leads to over-segmentation of regions as well as an extent in the segmented image itself, like a noisy image does not make doctor to diagnose the abnormal image. This is the major drawback of watershed algorithm [72] [73]. This problem can be avoided by using markers for segmentation. A marker is a connected component belonging to an image. The connected components by markers do not have same intensity values, treated as regional minima. Markers can be classified as internal (foreground) or external (background) depending on its location from region of interest [71]. Distance transform is a transformation for binary images. For some pixel connected components, the minimum distance from the background to each pixel is called distance transform processing [74].

After marking process on the reconstructed binary image, the marker watershed is segmented image, depended of maximum value of morphology method changes image marking standardization algorithm by extracting the local maximum mark to mask the images of non-local maximum region. Figure 4.8 shows the marker-based watershed segmentation.



(a) (b) (c)

Figure 4.8 Implementation watershed algorithm (a) Watershed ridge lines (b) Markers and object boundaries superimposed on original image (c) Colored watershed.

The results and analysis about the marker-based watershed segmentation algorithm, the CT scan of lung needs to achieve accurately target marker, after that process of marking, distance image processing have serious implications about extraction for tumor area and finally segmentation results.

4.6.3 Implementation of Marker Controlled Watershed Segmentation

- Input CT image that is converted from three dimensions to two dimensions (grayscale image).
- Compute the segmentation function in the image by using the Gradient Magnitude (sobel operator) for marking the dark regions of objects trying to segment.
- Morphological techniques called opening–by-reconstruction and closing-by-reconstruction to mark the foreground objects for cleaning up the image to get a suitable foreground marks by creating structure element (object).
- Threshold operation is used to compute the background marks of the image and remove dark pixels which are not available in object.

- After computing the background marks, the distance watershed transform calculate the boundary of object. Combining foreground, background markers and segmented object boundary on the original image.
- Label marks obtained in previous step to get computing the watershed-based segmentation to obtain color watershed on image.
- Visualize the locations of the foreground and background markers and label matrix on top of the original intensity image.

CHAPTER FIVE

FEATURE EXTRACTION FOR CLASSIFICATION

Image feature extraction is a very important stage of computer based systems. It uses some algorithm techniques to detect and isolate various desired portions or shapes (features) of an image. In this study, it has been performed after segmentation performance on lung tissue labeled as tumor region, or as a statistical method to provide certain parameters which is necessary for next step classification process to distinguish between normal (lung tissue)and abnormal (tumor or cancer) tissues in CT scan image. There are some statistical parameters used to diagnose the normal and abnormal tissues [75].

- **Area**

It is the scalar value that gives actual number of overall nodule pixels in the extracted ROI. Transformation function creates an array of ROI that contains pixels with 255 values.

$$Area = A = A_{i,j}, \mathrm{X\ ROI}[Area] = \mathrm{I}, \mathrm{Y\ ROI}[Area] = \mathrm{J} \quad (1)$$

Where, i, j are the pixels within the shape.ROI is the region of interest. X ROI][is a vector contains ROI x position, Y ROI[ ] is a vector contains ROI[ ] y position [76].

- **Perimeter**

It is a scalar value that gives actual number of the nodule pixels. It is the length of extracted ROI boundary. Transformation function creates an array of edges that contains pixels with 255 values that have at least one pixel which contains 0 values [76].

$$\mathrm{Perimeter} = \mathrm{P} = \mathrm{P}_{i,j}, \mathrm{X\ edg}e[\mathrm{P}] = \mathrm{i}, \mathrm{Y\ edg}e[\mathrm{P}] = \mathrm{j} \quad (2)$$

- **Irregularity Index**

Lung cancer is characterized partially by the irregularity in the tumor boundaries. For this analysis, the irregularities in the tumor are computed by an index:

$$\text{Irregularity} = \frac{4\pi * \text{Area}}{(\text{Perimeter})^2} \quad (3)$$

The metric value or roundness or circularity index or irregularity index (I) is equal to 1 only for circle and it is < 1 for any other shape. Here it has been assumed that more circularity of the object leads to the highness in the probability of being a nodule [77].

- **Equivalent Diameter**

Scalar that specifies the diameter of a circle with the same area as the region [78], [79] is computed by the formula in (4):

$$\text{ED} = \sqrt{\frac{4 * \text{Area}}{\pi}} \quad (4)$$

- **Eccentricity**

This metric value is also called as roundness or circularity. The major axis is the longest diameter and the minor axis is the shortest if the eccentricity is equal to onethat means shape is circular, otherwise the shape is irregular [80].

$$\text{Eccentricity} = \frac{\text{Length of Major Axis}}{\text{Length of Minor Axis}} \quad (5)$$

5.1. Statistical Parameters For Classification

The classification pre-process involves following steps:

- Creating dataset(training dataset) by building table in Microsoft Excel .
- Identifying class attribute and classes by using parameters from feature extraction.
- Choosing a suitable attributes for classification between normal cell and abnormal tumor.
- Using the model or algorithm to classify the dataset.

❖ **Data processing**

The lung cancer dataset which consists of 306 examples in each of two types of images: 153 normal and 153 abnormal. This data set is shown in Table (5.1). There are five attributes: area, perimeter, eccentricity, irregularity and equivalent diameter. The following set of rules might be learned from this dataset:

Results of Feature extraction

Table 5.1 Sample of feature extraction for classification

No	Image-name	area	perimeter	Irregularity	EDiameter	Eccentricity	tumor.type
1	Image 1	2229	203.727	0.6749	53.2734	1.0588	abnormal
2	Image 2	492	160.756	0.2392	25.0287	3.1703	abnormal
3	Image 3	157	64.285	0.4774	14.1386	4.3298	abnormal
4	Image 4	319	145.051	0.1905	20.1535	4.9903	abnormal
5	Image 5	263	73.8440	0.6061	18.2992	2.8710	abnormal
6	Image 6	231	134.655	0.1601	17.1499	3.4703	abnormal
7	Image 7	884	166.663	0.3999	33.5491	1.7842	abnormal
8	Image 8	382	104.486	0.4397	22.054	1.9174	abnormal
9	Image 9	164	60.711	0.5591	14.4503	3.603	abnormal
10	Image 10	2053	172.888	0.8631	51.1269	1.123	abnormal

Discussions

In this part, some specific explanations about details of twenty images(ten abnormal and ten normal) will be done by choosing sample values from the whole dataset.

A. Abnormal images quality

Area parameter of the tumor region is calculated with white pixel values, perimeter is calculated by the boundaries of tumor area, if boundary of tumor is irregular, the perimeter provided high values as the result. Irregularity parameters will give low values (shown in image4,image6, image2), also area parameters give low values, otherwise if boundary of tumor area is near to circular shape, the irregularity parameter gives high values like the area parameter, but the perimeter provided low values other images have irregular shapes (shown in image10, image1).

Consequently, ediameter depended on area value if it has high ediameter provided high value. The changeability of eccentricity value is affected by shape

of tumor, if like longitudinal the eccentricity gives high value as shown in image5, image6,image9, image2, image3, image4,otherwise if tumor shape as a circular shape gives an approximate value near to one shown in image1,image10or slightly higher as shown in image7, image8. Figure 5.1 shows sample of our data set related to abnormal images.

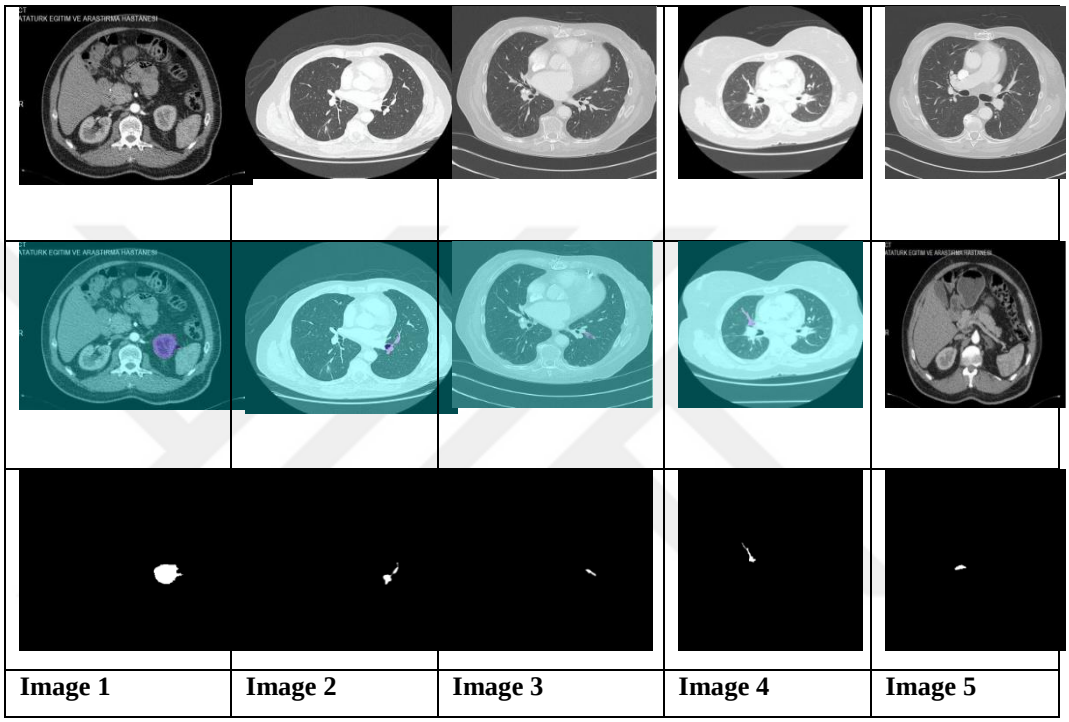
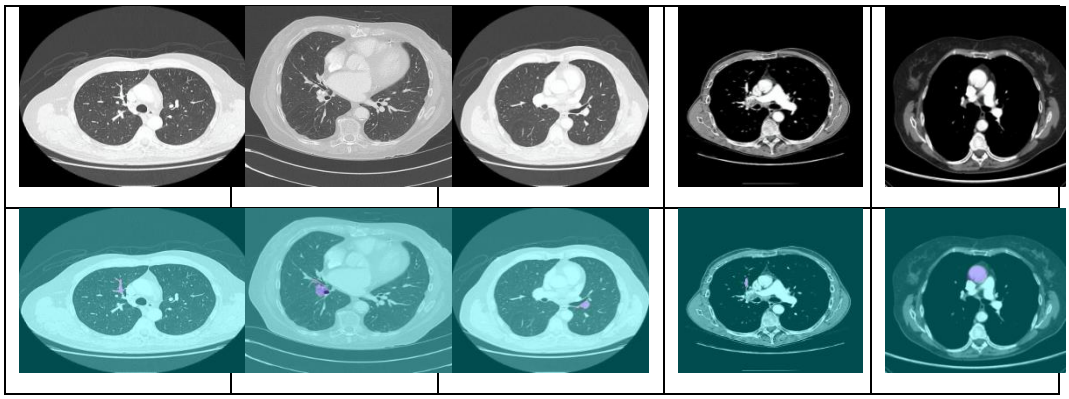


Figure 5.1 Sample of abnormal images



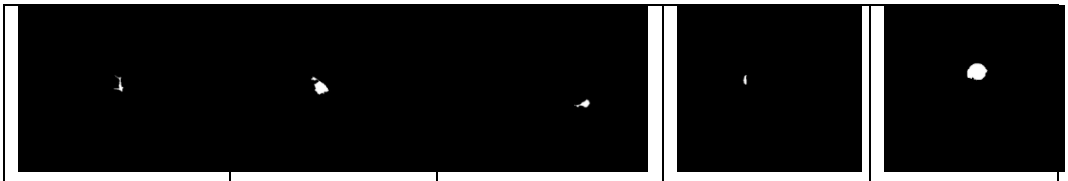


Image 6	Image 7	Image 8	Image 9	Image 10

Figure 5.1 Cont.

Results of Feature extraction

Table5.2 sample of feature extraction for classification

No	Image-name	area	perimeter	Irregularity	EDiameter	Eccentricity	tumor.type
1	image1	297	60.058	1.0347	19.4461	1.1731	normal
2	image2	842	111.623	0.8492	32.7424	1.8726	normal
3	image3	384	68.99	1.0138	22.1116	1.2516	normal
4	image4	417	76.442	0.8968	23.0422	1.6566	normal
5	image5	491	88.086	0.7952	25.0032	2.0893	normal
6	image6	162	44.932	1.0084	14.3619	1.5048	normal
7	image7	558	83.016	1.0175	26.6546	1.0978	normal
8	image8	705	105.164	0.8011	29.9605	2.0292	normal
9	image9	374	69.618	0.9697	21.8218	1.4137	normal
10	image10	750	96.207	1.0183	30.9019	1.1489	normal

In this part, specific explanations are done about normal images (healthy images) by cutting random regions mainly from tissues by the developed software in Matlab.

B. Normal images quality

In normal images, we can see area, parameter and ediameter are increasing or decreasing sequence because area shapes are approximately regular, there fore irregular parameters give high values. Area parameter region is calculated by the white pixel values. If area parameter has high values, the perimeter and ediameter give high values also, if we compare with abnormal tumor. Finally, eccentricity parameter gives low value because the shape is nearly rounded. Figure 5.2 sample of our dataset related to normal images.

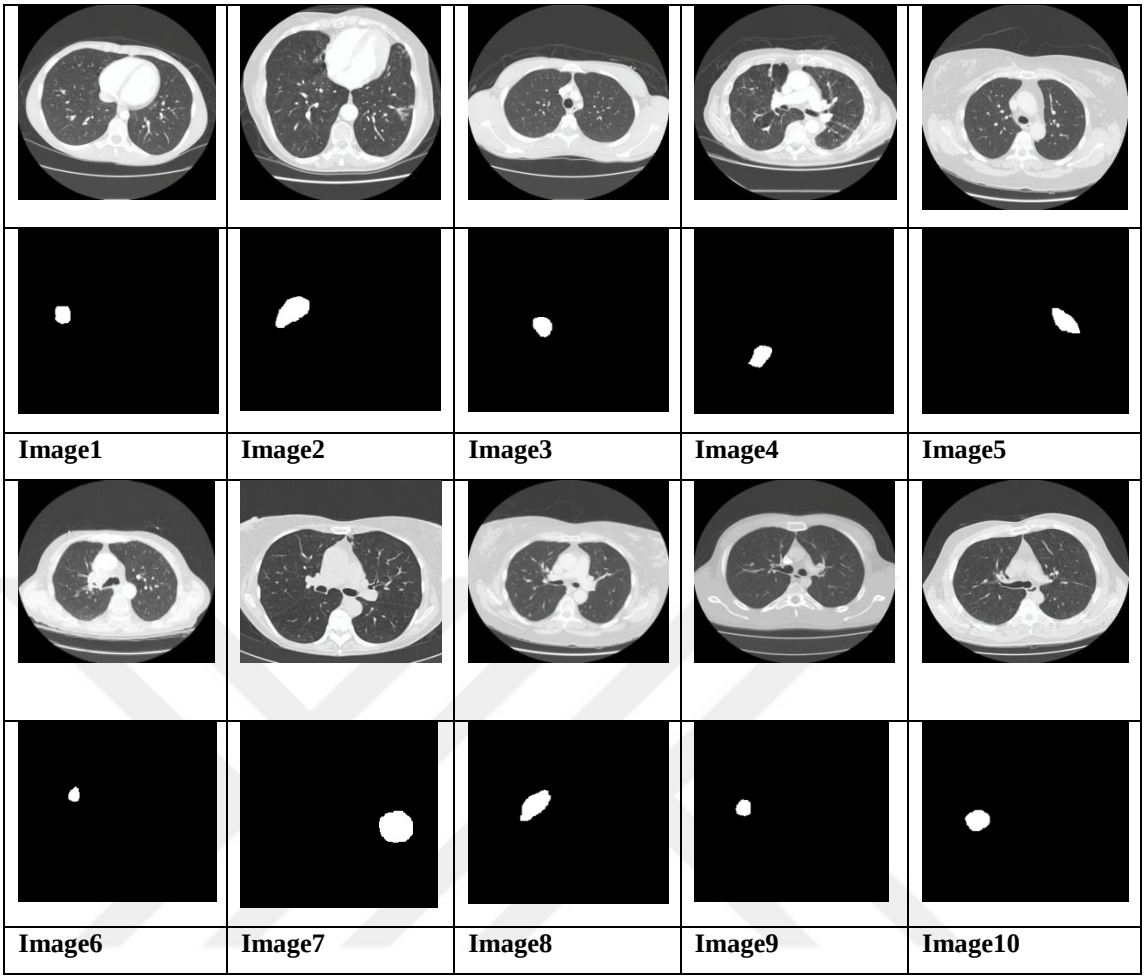


Figure 5.2 Samples of normal images

5.2. Steps Of Preparing Data For Classification

5.2.1 Data mining

Data mining is used to solve the problems by analyzing data already presented in the databases. It uses machine learning, statistical and visualization techniques, also defined as the application of algorithms to search the patterns and relationship found in the data (internal factors, external factors) or training data and testing data. Finding patterns in data can be done on semi_ automatic or automatic methods[81].
Data mining is used for extraction of hidden information from large databases, to help user to focus on the important information from database. Data mining is a way of technical implementation of the existing software and hardware platforms which is useful for enhancing the values of existing information resources. Figure 5.3 shows

the relation between machine learning, database, visualization and statistical methods [82].

Related Fields

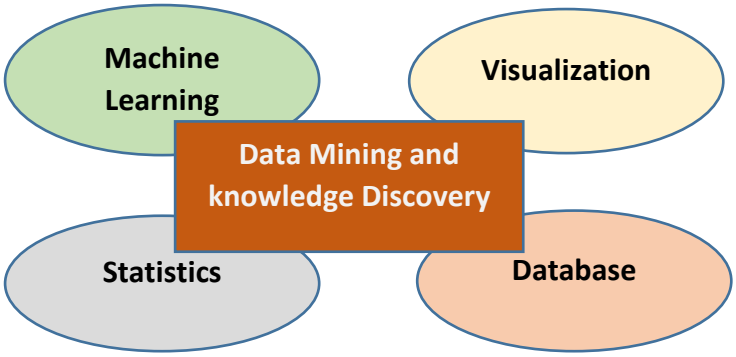


Figure 5.3 Relations between fields

Data mining steps in the knowledge discovery process are as follows:

- Data cleaning: Removing the noise and inconsistent data (improve data quality).
- Data integration: Collection of multiple data sources, reducing and avoiding redundancies in resulting dataset. This step is improving speed of data mining process.
- Data selection: Data relevant for analysis task or character from database for pre-processing.
- Data transformation: Unification and transformation of the data into appropriate forms for mining.
- Data mining: Using intelligent methods to extract patterns from data (the application of algorithms).
- Pattern evaluation: Identification of patterns that are interesting (data mining generates interesting patterns or evaluates patterns for desired results.
- ❖ Knowledge presentation: Visualization and knowledge representation techniques are used to present the extracted or mined knowledge to the end user.

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups

of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining)

[83]. Figure 5.4 shows the relation:

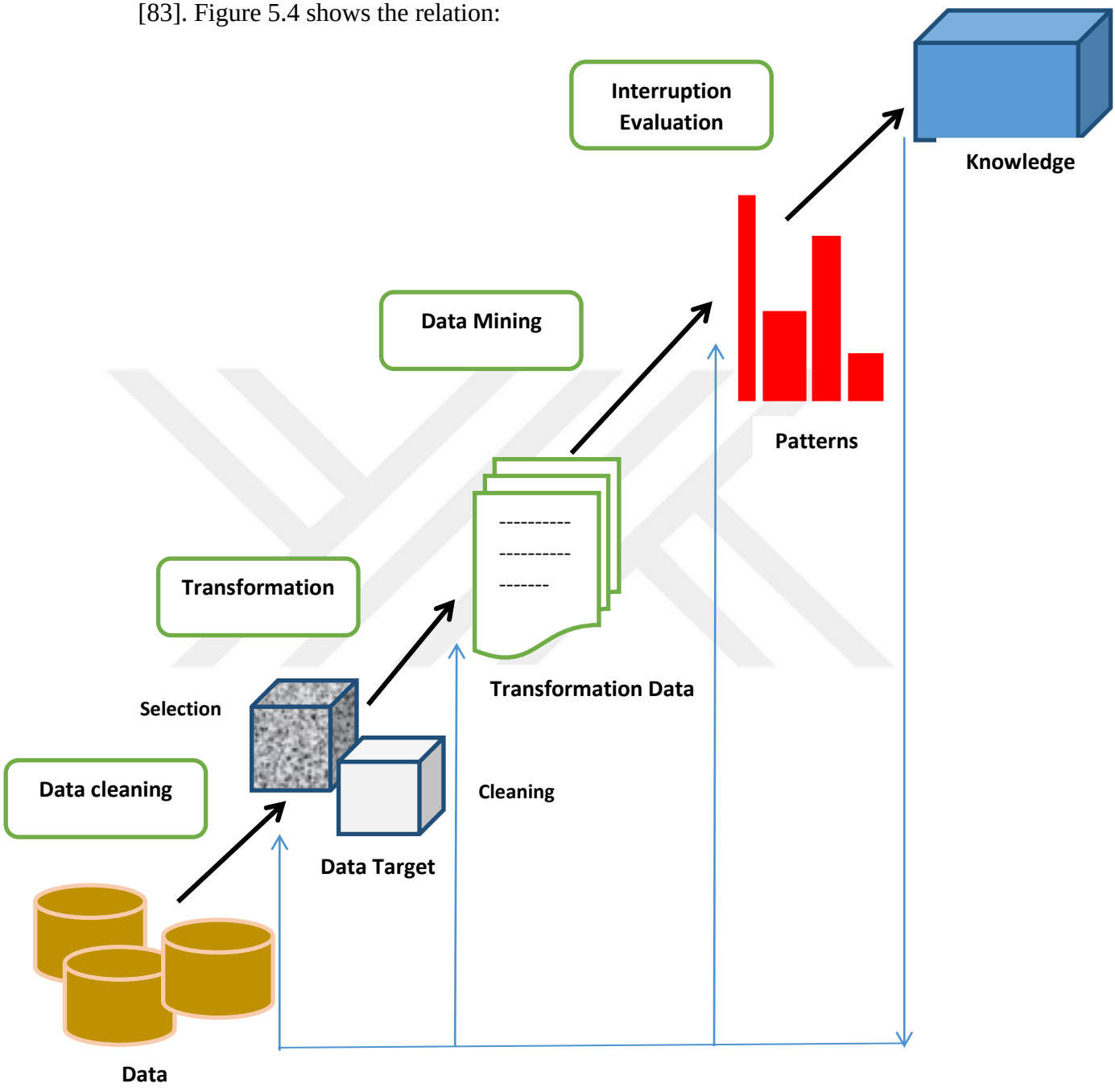


Figure 5.4 Relations between data mining steps

5.2.2. Machine learning

It is a method of teaching computers to make a prediction based on some data. It is the branch of artificial intelligence which improves programs by using the data[83].

❖ **What is machine learning**

Machine learning is the semi_automated extraction of knowledge from data. Machine learning is an algorithm that is a part of the data mining process. It consists of three component parts:

- **Knowledge from data**: It is started with a question that might be answerable using data.
- **Automated extraction:** It provides data manually and run some processes or algorithms.
- **Semi_automated:** It is practically machine learning required many smart decisions provided by a human [84].

Machine learning is divided into two types:

1. Supervised is a process on making prediction by using data, function is labeled as training data. It has a set of training examples. For instance, dataset of serious email messages, supervised learning task whether each email message is "spam" or not go to "inbox" because this specification turns to production.

2. Unsupervised is extracting structure from data or how to represent it. Function is describing hidden structure from unlabeled data. It means examples given by user to the learner are not labeled by the system, user is providing information and system will process without any error or reward signal to evaluate a potential solution [85].

5.3. Classification Algorithms

- **Adaboost**

It is Adaptive Boosting, a machine learning algorithm can be used itself as well with other learning algorithms to improve their performance. The main idea of Adaboost is taking many weak classifiers (each of which can be trained to detect a specific feature, with a success rate of only slightly over 50%). Adaboost duty distributes weights over all the classifiers depending on their performances [86]. The Adaboost algorithm extends to multiclass classification problems also suitable for recognition purposes and provides a strong result classifier for large

data sets. A disadvantage of this algorithm is that it's weak for a single classifier [87].

Input: Suppose of m examples $\langle (x_1, y_1), \dots\dots\dots (x_m, y_m)\rangle$ with labels $y_1 \in Y\{|$ $1, \dots\dots\dots, K\}$

weak linear algorithm **Weak Learn**

integer Tspecifyof iteration

Initialize D_1(i)= $1/m$for all$i$ **.**

Do for $i = 1,2, \dots\dots, \boldsymbol{T}$**.**

1. Call **Weak Learn** providing it with the distribution D_i

2. Get back a hypothesis $h \mid t : X1 + Y$**.**

3. Calculate the error of$h \mid t : \epsilon_t = \sum_{t|ht(xi)\ \#yi}^{n} D_t$(i)if $\epsilon_t > \frac{\mathbf{1}}{\mathbf{2}}$then set $\boldsymbol{T = t -}$ **1**and abort loop.

4. Set $\beta_i \in_t /(1 - \epsilon_t)$

5. Update distributed D_t:$D_t + 1(i) = \frac{D_t}{Z_t} \times \begin{cases} \beta_i \ if\ h_t(xi) = y_i \\ 1 \qquad \text{otherwise} \end{cases}$

where $Z_t$is a normalization constant (chosen so that D_{t+1} will be distribution)

Output the final hypothesis: $h_{f|n}(x) = arg\ max_{y \in Y} \sum_{t|ht(x)=y} \log \frac{\boldsymbol{log}}{\beta}$,[88].

- **Bagging**

It belongs to ensemble learning algorithms create individuals for its ensemble by training each classifier on a random as well redistribution of the training set, each classifier's training set is generated by randomly drawing with replacement [89]. Bagging is an unsuitable learning algorithm that can be used for small data set if small changes happened in the training data set, in this way it generates various classifiers. Generally, single classifier creates high test error. The series of classifiers make a lower test error than the individual classifier because the difference of classifiers usually compensates for errors of any individual classifier. A learning algorithm is grouping small changes in the training set leads to relatively large changes in accuracy[90].

Bagging = Bootstrap AGGregatING
$B$is the number of "bags" or base hypotheses
$L$is the base learning algorithm
Bagging(examples, B, L)

1. for $i \leftarrow 1$ to B
2. examples$s_i \leftarrow$ a bootstrap sample of examples
3. $h_i \leftarrow$ apply L to example s_i

4. ***return***$h_1,h_2,\ldots,h_B$ [91]

- **Random SubSpace**

It is ensemble learning algorithm composed of several classifiers. RSS is used for random decision forestswhich is generated from decision trees. Random SubSpace may be formatted from any classifier according to the structure of problem. Sometimes it may be applied in classification problems with single class[92].

1. Repeat for $b = 1,2, \ldots \ldots, B$
(a) Select an r-dimensional random subspace $\mathrm{X}^{\sim b}$ from

the original p-dimensional feature space X.

(b) Construct a classifier $\mathrm{C}^b(x)$ (with a decision boundary $\mathrm{C}^b(x) = 0\ in\ \mathrm{X}^{\sim b}$.

2. Combine classifiers $\mathrm{C}^b(x), b = 1,2, \ldots \ldots, B$ by simple majority voting to a final decision rule.

$\beta(x) = arg\ max_{y\in\{-1,1\}}\delta_{sgn(\mathrm{C}^b(x)),y}$, [93]

- **Random Forest**

Also it is an ensemble learning algorithm that utilizes structure for a multitude of decision tree sets training time that is suitable for a number of decision tree classifiers in diverse sub-samples of data set and use averaging to increase the predictive accuracy classifiers and also to control over-fitting [94]. Method creates a lot of decision trees and use them for classification as follows;

Create Random Subset

$$\boldsymbol{S1} = \begin{bmatrix} f_{A12} & f_{B12} & f_{C12} & C_{12} \\ f_{A15} & f_{A15} & f_{A15} & C_{12} \\ \vdots & & \vdots & \\ f_{A35} & f_{B35} & f_{C35} & C_{35} \end{bmatrix} \qquad \boldsymbol{S2} = \begin{bmatrix} f_{A2} & f_{B2} & f_{C2} & C_2 \\ f_{A6} & f_{A6} & f_{A6} & C_6 \\ \vdots & & \vdots & \\ f_{A20} & f_{B20} & f_{C20} & C_{20} \end{bmatrix}$$

$$\boldsymbol{Sm} = \begin{bmatrix} f_{A4} & f_{B4} & f_{C4} & C_4 \\ f_{A9} & f_{A9} & f_{A9} & C_9 \\ \vdots & & \vdots & \\ f_{A12} & f_{B12} & f_{C12} & C_{12} \end{bmatrix}$$

$S1$ (Decision Tree 1)

$S2$ (Decision Tree 2)

Sm (Decision Tree m) [95].

- **Naive Bayes**

It is a supervised learning method as well as a statistical method, prior knowledge and observed data can be combined for classification. It calculates explicit probabilities for hypothesis and it is robust to noise in input data. Naive Bayes requires linear data to work correctly.As a result it provides a useful perspective outputs for understanding probabilistic model. It captures uncertainty model by determining probabilities of the outputs and evaluating many learning algorithms[96].

- Bayesian classifiers use Bayes theorem, which says

$$p(C_j \backslash d) = p(d \backslash C_j)\, \frac{p(d \backslash C_j)\, p(C_j)}{p(d)}$$

- $p(C_j \backslash d)$= probability of instance d being in class C_j,
- $p(d \backslash C_j)$= probability of generating instance d given class C_j,
- $p(C_j)$ = probability of occurrence of class C_j,
- $p(d)$= probability of instance d occurring [97].

- **SVM**

Support vector machines are supervised learning models related to learning algorithms for analyzing data and also recognizing patterns will be used for classification process. The main idea of SVM takes a set of input data and for each given input, predicts which of two classes forms the input, making it a non-probabilistic binary linear classifier. Kernel function which maps the given data into a different space; the separations can be made even with very complex boundaries [98].

Binary Classification
Given training data (x_i, y_i) for $i = 1 \ldots . N$ with $x_i \in \mathrm{R}^d$ and $y_i \in \{-1,1\}$, learn a classifier f(x)

such that $\mathrm{f}(x_i) \begin{cases} \geq 0 & y_i = +1 \\ < 0 & y_i = -1 \end{cases}$
A linear classifier

$f(\mathrm{x}) = \mathrm{w}^T \mathrm{x} + \mathrm{b}$ [99]

- **ANN**

Neural networks is a branch of artificial neural networks. The architecture of it consists of three layers of input layer, hidden layer and the result display in output layer. It works as well for non-linear data because it is simulated on the observation of biological neurons and network of neurons. Wide range of input data for training makes neural network to work with higher accuracy, otherwise large set of similar data also small data set make system to be unfair of output results. However neural networks need large difference input data set for training and also long time to train system to give suitable results [100].

$$Y_{pj} = f\left[\sum_{i=0}^{n \blacksquare 1} W_i, X_i\right]$$

Calculate output

$W_{ij}(t \blacksquare 1) = W_{ij}(t) \blacksquare \eta \bar{\mathrm{O}}_{pj} o_{ij}$

$\bar{\mathrm{O}}_{pj} = K o_{pj}(1 \blacksquare o_{pj}) \sum_k \bar{\mathrm{O}}_{pk} W_{jk}$

hidden units

$\bar{\mathrm{O}}_{pj} = K o_{pj}(1 \blacksquare o_{pj}) (t_{pj} \blacksquare o_{pj})$

output units

$\bar{\mathrm{O}}_{pj} w_{ji} (t \blacksquare 1) = w_{ji} (t) \blacksquare \mu \bar{\mathrm{O}}_{pj} o_{ij} \blacksquare \alpha w_{ji} (t) \blacksquare w_{ji} (t \blacksquare 1)$ [101].

- **ThresholdSelector**

It is a branch of meta-classifier algorithm works by selecting a mid-point threshold on the probability output by a distribution classifier. The threshold is represented as a set that gives performance measure that is optimized. It measures the performance either on training data set, a hold-out set or using cross-validation. Furthermore, the probabilities returned by the base learner can

have their range expanded so that the output probabilities will reside between 0 and 1 [102].

- **Vote**

It provides weak classifiers that could be improved to achieve arbitrarily high accuracy, but that never means implementation of vote could always do the trick. It works by comparing the asymptotic misclassification error of the majority-vote classifier with the average individual error also can be implied binary classification with equal prior probabilities. The collection of weak classifiers must meet the minimum requirements of having an average true positive rate of at least 50% and an average false positive rate of at most 50% [103]. The result classification is shown in Table 5.3.

Results of classification

Table 5.3 Results of classification algorithms

No	Method	Correctly-Classified	Incorrectly-Classified	Time
1	AdaBoostM1	99.6732 %	0.3268 %	0.5 seconds
2	Bagging	99.6732 %	0.3268 %	0.03 seconds
3	RandomForest	99.6732 %	0.3268 %	0.16 seconds
4	SMO	98.0392 %	1.9608 %	0.09 seconds
5	MultilayerPerceptron	97.7124 %	2.2876 %	31.87 seconds
6	NaiveBayes	97.0588 %	2.9412 %	0.10 seconds
7	RandomSubSpace	96.0784 %	3.9216 %	0.05 seconds
8	ThresholdSelector	94.1176 %	5.8824 %	0.14 seconds
9	Vote	49.0196 %	50.9804 %	0.15 seconds

Discussions

Our data set consists of 306 images combining 153 abnormal images and 153 normal images. Learning algorithm (Adaboost, , Naive Bayes, SVM), ensemble learning

algorithms (Bagging, Random SubSpace, Random Forest), artificial intelligent (ANN) by using cross-validation were used.
All these algorithms need a huge data set to get good results, it increases performance of them because these algorithms belonged to artificial intelligent when user provides large data to the system, makes it to distinguish and recognize the difference among the value of data.

Adaboost, Bagging, Random Forest algorithms provide high accuracy classifier reach to (99.6732 %) for adaboost fixed on features of parameters that used between normal and abnormal images. It takes the hard part from data by increasing the probability of those data that means it is sensitive to noisy data. It works by taking high value over than 50%,otherwise the value less than 50% will be regarded then the adaboost algorithm will distribute the high value(weights) over all classifier, then gets good results and fast classifier in 5 seconds. In this way it decreases the mistakes made by the weak classifier.

Bagging is using multiple ensemble learning algorithms in statistics and also machine learning. The minimization of test error by creating single classifier for each training data set, reduce the errors of whole data set and achieve high predict performance in short time (3 second) because the algorithm collects the errors from each classifier in training data set lead to rise implementation, therefore it provides outliers data in small data set.

Random Forest uses iterative classifier from diverse values of parameters in data set built from plenty of decision trees, then gets average of high performance of classifier.

Those three algorithms are using multiple classifiers of data set at the same time lead to giving more opportunity to achieve reliable results.

SVM provides accuracy classifier reaches to 98.0392 %, analyses dataset and recognition for classifier divided into two categories positive data and negative

separation by a clear gap in training dataset depended to the values of parameters in dataset. In SVM algorithm, build on one category represents positive data implementation in 0.09 seconds.

ANN provides accuracy classifier reaches to 97.7124%. Neural network is sensitive to numbers of input data set, a huge data for training makes ANN to achieve high accuracy performance, otherwise the accuracy performance has been decreased in 31.87 seconds, it takes a longest time rather than other algorithms.

Naive Bayes provides accuracy classifier reaches to 97.0588 %, it uses many learning algorithms, then it gets results classifier of their represented in Naive Bayes algorithm in 10 seconds.

Random SubSpace provides accuracy classifier reaches to 96.0784 %less than Adaboost, Bagging, Random Forest, SVM and Naive Bayes because it uses several methods of classifiers like SVM, linear classifiers, nearest neighbours etc. also used other types of classifiers represented in Random SubSpace algorithm. All of these algorithms provide various accuracy of results in 0.05 seconds.

Threshold Selector provides accuracy classifier reaches to 94.1176 % depended on one type of classifier mid-point threshold set by considering the optimal performance in 0.14 seconds.

Vote provides accuracy classifier reaches to 49.0196 %. It works in weak classifier because it compares misclassification error of Vote classifier with single average. In this way, algorithm could not recognize the optimal performance without collective classifier results in 15 seconds.

CHAPTER SIX

CONCLUSIONS AND FUTURE WORKS

This chapter introduces conclusions and future works of this thesis with entitled early detection of lung cancer by using MATLAB platform that deals with following points are:

- Removing noise from abnormal images.
- Detecting edges of tumor areas using gradient magnitude.
- Marking the foreground objects on image in morphology operations (opening, closing, dilation and erosion).
- Computing background markers on image to observe the tumor area by thresholded opening-closing.
- Computing watershed transformation for segmentation (for detecting tumor area and removing background of the image.
- Isolating various desired portionsoftumor shape (feature extraction) of abnormal image and cutting area from normal image by usingstatistical methods.
- Evaluating the accuracy performance of the classification algorithms on the image.

6.1. Conclusion

This thesis presents sequences of methods to detect the tumor inside the tissue in medical images in the early stages. Early detection is a very important issue to improve the survival rate in lung cancer. By using some technical operations for edge detection of tumor in gradient magnitude of sobel filter and morphology operations for tumor segmentation steps. Feature extraction is necessary for preparing images to the classifier by using watershed algorithm. Therefore, this system package contains five sections.

The first section mentions briefly about types of lung cancer, also which type is common; explain kinds of tumors, stages of lung cancers from early to advance stages. This section also explains the methods of diagnosis and types of the scanning systems and discusses which is the better for diagnosing the lung cancer.

In the second section, the methods of biomedical imaging has been used in this thesis are explained.

In the third part, steps of removing the noise from images by using sobel filter in gradient magnitude (second derivative) to detect the edges of the tumor in small areas. Morphology operations (opening, closing, dilation and erosion) is important to mark the background and remove it from the image because background is unnecessary in this kind of studies. Computing background markers is important to label the tumor area by using thresholded (opening-closing), segmentation method to detect tumor area by using watershed transformation and removing the background image to focus on tumor area region.

In the fourth part, isolating various desired portions of tumor shape, also cutting randomly from tissue of lung on normal image by using statistical methods after segmentation has been done to get specific parameters to difference at the normal and the abnormal images. In this thesis, five parameters were used (Area, perimeter, eccentricity, irregularity and equivalent diameter) by using Matlab code and the results were saved on a table in Microsoft Excel to prepare the dataset (analysis data in data mining) for classification processes. Finally, classification algorithms were used in this step to distinguish the normal and the abnormal images, it is a very important stage in medicine to support the doctor for diagnosing the disease and getting correct decisions. For classification, nine algorithms were used (cross-validation) that provided different accuracy performance results which include:

- **Adaboost:** It is a machine learning algorithm, also its performance can be increased if it is used with learning algorithms, provided accuracy rate of 99.6732%, because it collects week classifier and distributes high value.

- **Bagging**: It is an ensemble learning algorithm, provided accuracy rate of 99.6732% in this study.This algorithm combines errors from dataset as a result share high value.
- **Random Forest:** It is an ensemble learning algorithm, provided accuracy rate of 99.6732% in this study. This algorithm takes average in multitude decision trees from each classifier.

Three previous algorithms get performances depend on ensemble learning algorithms.

- **SVM:** It is a supervised learning model related with learning algorithms, provided accuracy rate of 98.0392% in this study. It divided input data in two classes (positive, negative) by gab, results were given in the positive part.
- **ANN:** It is a branch of artificial neural networks, provided accuracy rate of 97.7124%**.** It works on three layers and results are displayed in the output layer.
- **Naive Bayes**: It is a learning method that can provide higher performances when they are used with statistical methods. This method provided accuracy rate of 97.0588%. It collects results from many learning methods represented on it .
- **Random SubSpace**: It is an ensemble learning algorithm and provided accuracy rate of 96.0784%. This method uses difference classifiers such as svm or others. Therefore, its performance accuracy was lower than previous algorithms.
- **Threshold Selector:** It is a branch of meta-classifier algorithm that works by selecting a mid-point threshold. This method provided accuracy rate of 94.1176%. It uses one classifier, therefore it provided lower performance accuracy than previous algorithms.
- **Vote:** This method was a weak classifier that provided accuracy rate of 49.0196%. It compares misclassification error with single average.

6.2. Future Work

Nowadays the early detection of lung cancer is a critical problem because it causes death through men in a high rate and secondly critical for women after breast cancer. Therefore, it is necessary to find a good method for diagnosis so there are lots of issues to be developed for this thesis are as follows:

- DNA sequence alignment contains all kind of gene in human body, therefore to discover lung cancer before it hits human body. So DNA parts that causes the lung cancer should be detected.
- Data mining methods are necessary because DNA is so vast and complex, it needs filleting data to remove noise from it to focus on gene which causes the lung cancer.
- Getting (AFM)microscope images to show the details of DNA sequence alignment.
- Improving segmentation method to differentiate different gene.
- Improving classification algorithms in artificial intelligence.

REFERENCES

[1]The National Lung Screening Trial Research Team, 2011 "*Reduced lung cancer mortality with low-dose computed tomographic screening*" N. Engl.J.Med, vol. 365, pp. 395–409.

[2]American Cancer Society, 2014"*Lung Cancer (Non-Small Cell)*",

[3]Hongdong Li, Guini Hong and ZhengGuo, IEEE 2014 "*Reversal DNA methylation patterns for cancer diagnosis* ", International Conference on Systems Biology (lSB).

[4]Cancer Causes, (11/02/2015) Retrieved from The Truth About Cancer, http://thetruthaboutcancer.com/benign-malignant-tumors-difference.

[5]Medical Advice, Diagnosis or Treatment, (05/01/2016), Retrieved from "http://www.webmd.com/a-to-z-guides/benign-tumors-causes-treatments".

[6]Advanced Cancer,(2014), Retrieved from "http://www.cancer.org/treatment/understandingyourdiagnosis/advancedcancer/advanced-cancer-what-is-metastatic".

[7]Cancer research UK,(202/04/2014), Retrieved from "http://www.cancerresearchuk.org/about-cancer/type/lung-cancer/about/types-of-lung-cancer".

[8]Cancer research UK, (06/09/2015), Retrieved from,"http://www.cancer.gov/about-cancer/diagnosis-staging/staging/staging-fact-sheet ".

[9]Cancer research UK, (29/10/2014), Retrieved from" http://www. cancerresearchuk.org/about-cancer/what-is-cancer/how-cancer-can-spread",

[10]Jiann-Der Lee1, Chung-Hsien Huang, Neng-Wei Wang, Chin-Song Lu2, 2011, "*Automatic DNA sequencing for electrophoresis gels using image processing algorithms*", J. Biomedical Science and Engineering.

[11]Heller G, Weinzierl M and Noll C, 2012 " *Genome-wide miRNA expression profiling identifies miR-9-3 and miR-193a as targets for DNA methylation in none small cell lung cancers"*. Clin Cancer Res .

[12]Mayo Clinic,(22/02/2014), Retrieved from"http://www.mayoclinic.org/diseases-conditions/cancer/in-depth/cancer-diagnosis/art-20046459".

[13]H. Sagha and M. Sorushniya , 2003,"Comprehensive reference laboratory equipment and diagnostic products", Tehran, ketabmir.

[14]Yan Wang, ZijianCui , Yan Wang ,WenxinZheng , Anyu Chen, ,IEEE 2012,"*A Feasibility Study of Early Detection of Lung Cancer by Saliva Test Using Surface Enhanced Raman Scattering* " International Conference on BioMedical Engineering and Informatics.

[15]Thomas Pengo, Arrate Muñoz-Barrutía and Carlos Ortiz-de-Solórzano, EEE 2014"*A Novel Automated Microscopy Platform for Multiresolution Multispectral Early Detection of Lung Cancer Cells in Bronchoalveolar Lavage Samples*" SYSTEMS JOURNAL, VOL. 8, NO. 3, SEPTEMBER.

[16]Christian Donner, NaoufelWerghi, FatmaTaher , Hussain Al-Ahmad , IEEE 2012" *Cell Extraction from Sputum Images for Early lung Cancer Detection*",.

[17]Medicine Net,(08/02/2016), Retrieved from" http://www.medicinenet.com/ script/main/mobileart.asp?articlekey=406&page=9",.

[18]Cancer Net,(08/05/2015), "Retrieved from "http://www.cancer.net/cancer-types/lung-cancer-non-small-cell/diagnosis.

[19]Biomed Center,(24/11/2011), Retrieved "http://bmccancer.com/articles/10.1186.

[20]American Cancer Society,(05/01/2016), Retrieved from "http://www.cancer.org/cancer/lungcancer-non-smallcell/detailedguide/non-small-cell-lung-cancer-diagnosis".

[21] JaspinderKaur M.E Research Scholar ECE ,UIETPanjab, 2014,"*An automatic CAD system for early detection of lung tumor using back propagation network*" , IEEE International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom).

[22] Keisuke YOKOTA, Shinya MAEDA, Hyoungseop KIM, JooKooi TAN, Seiji ISHIKAWA, , 2014, "*Automatic Detection of GGO Regions on CT Images in LIDC Dataset Based on Statistical Features* ", Kitakyushu, Japan, December 3-6.

[23] Heber MacMahon, John H. M. Austin, Gordon Gamsu, Christian J. Herold, James R. Jett, David P. Naidich, et al., , 2005, "*Guidelines for Management of Small Pulmonary Nodules Detected on CT Scans*": A Statement from the Fleischner Society, Radiology, Vol. 237, 395-400.

[24]RadiologyInfo.org,(22/03/2013), Retrieved from "http://www. radiologyinfo.org. radiologyinfo.org/en/info.cfm?pg=screening-lung".

[25]AmericanCancerSociety(08/02/2014)"http://www.cancer.org/cancer/lungcancer-non-smallcell/moreinformation/lungcancerpreventionandearlydetection/lung-cancer- prevention-and-early-detection-guidelines".

[26]WebMD,(09/09/2014), Retrieved from "http://www.webmd.com/a-to-z-guides/magnetic-resonance-imaging-mri".

[27]Bankman IN (ed), 2008,Handbook of Medical Image Processing and Analysis. 2nd ed. New York: Academic Press.

[28]TANMANYONRUN(26/10/2011)http://tanmayonrun.blogspot.com.tr/describe-fundamental-steps-of-digital.html".

[29]Thomas M. Deserno,2011,"*Biomedical Image processing*" ch1,SpringerHeidelbergDordrechtLondonNewYork.

[30] K.Sreedhar1 and B.Panlal2 ,2012" *ENHANCEMENT OF IMAGES USING MORPHOLOGICAL TRANSFORMATIONS*" ,International Journal of Computer Science & Information Technology (IJCSIT) Vol 4, No 1.

[31] Omprakash Patel, Yogendra P. S. Maravi and Sanjeev Sharma, October 2013, " *A COMPARATIVE STUDY OF HISTOGRAM EQUALIZATION BASED IMAGE ENHANCEMENT TECHNIQUES FOR BRIGHTNESS PRESERVATION AND CONTRAST ENHANCEMENT*", Signal & Image Processing : An International Journal (SIPIJ) Vol.4, No.5.

[32]Rafael C. Gonzalez, Richard E. Woods, 2002," *Digital Image Processing*" chapter2,Second Edition.

[33]Robin N. Strickland University of Arizona Tucson, Arizona, Chapter5,Image-Processing Techniques for Tumor Detection,2002.

[34]Nisar Ahmed Memon, Anwar MajidMirza, and S.A.M. Gilani, 2008 " *Deficiencies of Lung Segmentation Techniques using CT Scan Images for CAD*", International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:2, No:8.

[35]Nisar Ahmed Memon, Anwar MajidMirza, and S.A.M. Gilani, 2006" *Segmentation of Lungs from CT Scan Images for Early Diagnosis of Lung Cancer*", World Academy of Science, Engineering and Technology 20.

[36]Hoifung Poon, Colin Cherry, Kristina Toutanova, June 2009, "*Unsupervised Morphological Segmentation with Log-Linear Models*", Human Language Technologies: Annual Conference of the North American Chapter of the ACL, pages 209–217..

[37]N. Senthilkumaran1 and R. Rajesh2, May 2009" *Edge Detection Techniques for Image Segmentation – A Survey of Soft Computing Approaches*", International Journal of Recent Trends in Engineering, Vol. 1, No. 2.

[38]G.T. Shrivakshan1 , 1 Research scholar, Bharathiar University, Coimbatore, Tamilnadu, September 2012, " *A Comparison of various Edge Detection Techniques*

used in Image Processing ", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 1.

[39]Ku. Vasundhara H. Lokhande, January 2014, " *Study of Region Base Segmentation Method*", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 1,.

[40]Liang-Chi Chiu, Tian-Sheuan Chang, Jiun-Yen Chen, and Nelson Yen-Chung Chang, IEEE AUGUST 2013" *Fast SIFT Design for Real-Time Visual Feature Extraction*", VOL. 22, NO. 8.

[41]G. Y. Chen and B. K´egl, IEEE 2007,"*Feature Extraction Using Radon, Wavelet and Fourier Transform*", 1-4244-0991-8/07/$25.00.

[42]Athira Krishnan, Sreekumar K, 2014, "*A Survey on Image Segmentation and Feature Extraction Methods for Acute Myelogenous Leukemia Detection in Blood Microscopic Images"*,(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (6).

[43]Yoon Kim, 2014, "*Convolutional Neural Networks for Sentence Classification*",Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1746–1751.

[44]PayalP. Dhakate1,Suvarna Patil2 , K. Rajeswari3, August 2014" *Preprocessing and Classification in WEKA Using Different Classifiers*",Int. Journal of Engineering Research and Applications.

[45]Dong ping Tian, China, July 2013" A Review on Image Feature Extraction and Representation Techniques", International Journal of Multimedia and Ubiquitous Engineering Vol. 8, No. 4.

[46]Sobel Operator ,(05/10/2014), Retrieved from "https://www.youtub.com /watch?v=sLWRJVBuVtw".

[47]Hassan Badry Mohamed A. El-Owny , December 2013" *Edge Detection in Images based on Approximation Theory* ", International Journal.

[48]O. R. Vincent , O Folorunso , 2009,"*A Descriptive Algorithm for Sobel Image Edge Detection*", Nigeria,Proceedings of Informing Science & IT Education Conference (InSITE),.

[49]Edge Detection with Gradients, (05/09/2012), Retrieved from "https://www. youtube.com/watch?v=j7r3C-otk-U".

[50]G. Deep, L. Kaur, and S. Gupta, Jan 2013*" Lung Nodule Segmentation in CT Images using Rotation Invariant Local Binary Pattern*", ACEEE Int. J. on Signal & Image Processing, Vol. 4, No. 1.

[51]ENCYCLOPAEDIABIRTANNICA,(06/10/2015),"http://global.britannica.com/science/morphology-biology.

[52]Amalorpavam.G, Harish Naik T, JyotiKumari, Suresha M, February 2013,"*ANALYSISOF DIGITAL IMAGES USING MORPHLOGICAL OPERATIONS*", International Journal of Computer Science & Information Technology (IJCSIT).

[53] C. R. González and E.Woods, 1992 "*Digital Image Processing*". Englewood Cliffs, NJ: Prentice Hall.

[54] K.Sreedhar , B.Panlal, Feb 2012 "*ENHANCEMENT OF IMAGES USING MORPHOLOGICAL TRANSFORMATIONS*", International Journal of Computer Science & Information Technology (IJCSIT) .

[55]H.Heijman, 1994,"*Morphological image operators*", Advances in Electronics and Electron Physics. Academic Press.

[56]R. Haralick, S. Sternberg, and X. Zhuang, IEEE July 1987, "*Image analysis using mathematical morphology*", Transactions on Pattern Analysis and Machine Intelligence, vol. 9, no. 4, pp. 532.550.

[57]Serra.J, (1994)"Mathematical Morphology and Its Applications to Image Processing", Kluwer Academic Publishers.

[58]A.M.Raid1, W.M.Khedr2, M.A.El-dosuky1 and Mona Aoud1,June 2014 "*IMAGE RESTORATION BASED ON MORPHOLOGICAL OPERATIONS* ",International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol. 4, No.3.

[59]Morphology Processing,(21/10/2012), Retrieved from "https://www. youtube.com/watch?v=gmi4ah7YAi0".

[60]L.Vincent, September 1992 "*Morphological area openings and closings for greyscale images*", in Proc. Shape in Picture '92, NATO Workshop, Driebergen, The Netherlands, Springer-Verlag.

[61]A.M.Raid, W.M.Khedr, M.A.El-dosuky1 and Mona Aoud, June 2014 "*IMAGE RESTORATION BASED ON MORPHOLOGICAL OPERATIONS"*, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol. 4, No.3.

[62]Wei Zhang,DalingJiang, IEEE 2011,"The Marker-Based Watershed Segmentation Algorithm of Ore Image".

[63]Salem Saleh Al-amri, N.V. Kalyankar and KhamitkarS.D, MAY 2010,"*Image Segmentation by Using ThershodTechniques*",JOURNAL OF COMPUTING, VOLUME 2, ISSUE 5.

[64]Wei Zhang, Hong Jiang, Zhang Ren, 2007, "*Adaptive image segmentation algorithm based on multi-threshold value Pattern Recognition and Simulation*", ,26 (8) 71-73.

[65]Kostas Haris, Serafim N. Efstratiadis, NicosMaglaveras, and Aggelos K. Katsaggelos, IEEE December 1998. "*Hybrid Image Segmentation Using Watersheds and Fast Region Merging*", Transactions On Image Processing, Vol. 7, No. 12.

[66]S. Beucher, "The Watershed Transformation Applied to Image Segmentation", Centre de MorphologieMathematique, Ecolos des Mines de Paris, France.

[67]R.C Gonzalez and R.E. Woods, Chapter 9& 10; ,2008, "Digital Image Processing‟, third edition, PHI Publications.

[68]Research Scholar and Professor and Chairman, May-2014,"*BIO-MEDICAL IMAGE SEGMENTATION USING MARKER CONTROLLED WATERSHED ALGORITHM: A CASE STUDY"*, International Journal of Research in Engineering and Technology.

[69]Xiying Liu, ShuQuan Wu, XiangMinXu., 2003, "Improved Watershed Algorithm Based On Medical Image Segmentation". Microelectronics Technology.,31 (4) :39-42.

[70]M. Ali HAMDI, December 2011, "*Modified Algorithm marker-controlled watershed transform for Image segmentation Based on CurveletThreshold*" Canadian Journal on Image Processing and Computer Vision.pp, 88-91. Vol. 2 No. 8.

[71]Amanpreetkaur, AshishVerma, September 2013, "*The Marker-Based Watershed Segmentation –A Review*" International Journal of Engineering and Innovative Technology (IJEIT), Volume 3, Issue 3.

[72]Ravi S and A M Khan, 11 -12th Dec 2013 ,"*Morphological Operations for Image Processing: Understanding and its Applications*", Proc. 2nd National Conference on VLSI, Signal processing & Communications NCVSComs..

[73]K. Parvati, B. S. PrakasaRao and M. Mariya Das , 2008,"Image Segmentation Using Gray-Scale Morphology and Marker-Controlled Watershed Transformation" Article ID 384346.

[74]R.C Gonzalez and R.E. Woods, Chapter 9& 10; 2008 ,"*Digital Image Processing*‟, third edition, PHI Publications.

[75]K. P. Aarthy and U. S. Ragupathy, July 2012 "*Detection of lung nodule using multiscale wavelets and support vector machine,"* International Journal of Soft Computing and Engineering (IJSCE), vol. 2, issue 3.

[76]Suzuki K., February 2005 ,"*False-positive Reduction in Computer-aided Diagnostic Scheme for Detecting Nodules in Chest Radiographs*",Academic Radiology, Volume 13, Number 10, pp.10-15.

[77]S. A. Patil and V. R. Udpi, April 2010 ,"*Chest x-ray features extraction for lung cancer classification,*" Journal of Scientific and Industrial Research, vol. 69, pp. 271-277.

[78]R. C. Gozalez and R. E. Woods, Digital Image Processing Using Matlab, 2nd ed, Gatesmark, USA, 2002, ch. 12, pp. 642-654.

[79]A. K. Jain,1992, "*Fundamental of Digital Image Processing*", 1st ed., Englewood Cliffs: Prentice Hall, , ch. 7, pp. 241-250.

[80]VijayA.Gajdhane ,Deshpande L.M., (Sep – Oct. 2014),"*Detection of Lung Cancer Stages on CT scan Images by Using Various Image Processing Techniques*",IOSR Journal of Computer Engineering (IOSR-JCE), Volume 16, Issue 5, Ver. III.

[81]Ian H. Witten, Eibe, 2005, "Data mining Practical Machine Learning Tools and Techniques, Second Edition".

[82]George M. Marakas, 2005 "*Modern Data Warehousing Mining, and Visualization*", Pearson Education, New Delhi.

[83]Data Mining KDD Process,(22/05/2015),"https;//www.youtube.com "./watch?v=a4M3GdI5UFY".

[84]Machine Learning , (07/04/2015), Retrieved from"https://www.youtube.com /watch?v=elojMnjn4kk ".

[85]Machine Learning, (03/05/2015), " https://www.youtube.com/watch?v=NP2-M3qRqAU ".

[86] Anton Ericson, April 2013, " *Object Recognition Using Digitally Generated Images as Training Data* ", Examensarbete 30 hp.

[87]Freund, Y., Schapire, R. E., 1997 "*A decision-theoretic generalization of on-line learning and an application to boosting*", Journal of computer and system science, 55, 119-139.

[88]Yoav Freund, Robert E. Schapire," Experiments with a New Boosting Algorithm", January 22, 1996.

[89]Bagging Classifier,(24/08/1999), Retrieved from "https://www.cs.cmu.edu/afs/cs/project/jair/pub/volume11/opitz99a-html/node3.html".

[90]D. Opitz and R. Mach, 1999 ,"*Popular Ensemble Methods: An Empirical Study*", Journal of Artificial Intelligence Research, 11,169-198.

[91]Dettling ,MBag 2003,"*Boosing for tumor classification with gene regression : a statistical data in preparation*".

[92]Ho Tin, "*The Random Subspace Method for Constructing Decision Forests*", IEEE, Transactions on Pattern Analysis and Machine Intelligence, 20(8), 832–844, 1998.

[93]Marina Skurichina and Robert P. W. Duin, 2002" *Bagging, Boosting and the Random Subspace Method for Linear Classifiers*", Pattern Analysis & Applications (2002)5:121–135, Springer-Verlag London Limited.

[94]Bagging Classifier ,(09/10/2010), Retrieved from"https://www.cs.cmu.edu /afs/cs/project/jair/pub/volume11/opitz99a-html/node3.html" .

[95]Random Forest algorithm ,(04/04/2014),Retrieved from"https://www.youtube. "com/watch?v=loNcrMjYh64" ,.

[96]Christian Donner, NaoufelWerghi, FatmaTaher , Hussain Al-Ahmad , 2012 IEEE ," *Cell Extraction from Sputum Images for Early lung Cancer Detection*".

[97]Naive Bayes Classifier ,(10/06/2013),Retrieved from " https://www.youtube.com /watch?v=XcwH9JGfZOU".

[98]Mr. Vijay A. Gajdhane , Deshpande L.M, (Sep – Oct. 2014)," *Detection of Lung Cancer Stages on CT scan Images by Using Various Image Processing Techniques* ", IOSR Journal of Computer Engineering (IOSR-JCE), Volume 16, Issue 5, Ver. III.

[99]S. V. N. Vishwanathan , M. NarasimhaMurty. Geometric SVM, 2002, "*A fast and intuitive SVM algorithm*". Institute of Science, Bangalore, November Submitted to ICPR.

[100]Ada, Rajneet Kaur, June 2013 ," *Early Detection and Prediction of Lung Cancer Survival using Neural Network Classifier* ",International journal of Application or Innovation of Engineering &management, Volume 2, Issue 6.

[101]Fiona Nielsen 4i, 2001," *Neural Networks – algorithms and applications*", Neural Networks – algorithms and applications.

[102]Andrea Addis, GiulianoArmano, Eloisa Vargiu, June 2010" *Experimental Assessment of a Threshold Selection Algorithm for Tuning Classifiers in the Field of Hierarchical Text Categorization*", Experimental Evaluation of Algorithms for Solving Problems with Combinatorial Explosion Bologna.

[103]arxiv.org,(07/06/2013), Retrieved from "http://arxiv.org/abs/1307.6522".

RESUME

Personal Information	
Name and Surname	Shaymaa Shakir
Date of Birth	1978
Birth Place	Baghdad
Nationality	Republic of Iraq
Address	Ankara / Çankaya KEKLİKPINARI,No: 5/11
E-mail	shprogrmmer90@gmail.com
Educational Information	
High School	Alharery School .
Undergraduate	**University of Al-mustansiriya** Baghdad, Iraq. B.Sc. Degree / Computer Science Department
Language Skill	Arabic (mother language) ,second language English.