2020-03-20

IMAGE PROCESSING FOR LUNG DISEASES CLASSIFICATION

Meshesha, Kasanesh

BAHIR DAR UNIVERSITY
BAHIR DAR INSTITUTE OF TECHNOLOGY
SCHOOL OF RESEARCH AND POSTGRADUATE STUDIES
FACULTY OF COMPUTING

USING IMAGE PROCESSING FOR LUNG DISEASES CLASSIFICATION

Kasanesh Meshesha Alitah

Bahir Dar, Ethiopia

October 19, 2017

IMAGE PROCESSING FOR LUNG DISEASES CLASSIFICATION

Kasanesh Meshesha Alitah

A Thesis submitted to the school of Research and Graduate Studies of Bahir Dar Institute of Technology, BDU in partial fulfillment of the requirements for the degree of Masters in the Information Technology in the School of Computing

Advisor Name: Gebeyehu Belay (PhD)

Bahir Dar, Ethiopia

October 19, 2017

DECLARATION

I, the undersigned, declare that the thesis comprises my own work. In compliance with internationally accepted practices, I have acknowledged and refereed all materials used in this work. I understand that non-adherence to the principles of academic honesty and integrity, misrepresentation/ fabrication of any idea/data/fact/source will constitute sufficient ground for disciplinary action by the University and can also evoke penal action from the sources which have not been properly cited or acknowledged.

Name of the student _____ Signature _____

Date of submission: _____

Place: Bahir Dar

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

Advisor Name:_____

Advisor€s Signature: _____

Bahir Dar University

Bahir Dar Institute of Technology-

School of Research and Graduate Studies

Faculty of Computing

THESIS APPROVAL SHEET

Student:

Kasanesh Meshesaha _____

Name Signature Date

The following graduate faculty members certify that this student has succes presented the necessary written final thesis and oral presentation for partial fulfillme the thesis requirements for the Degree of Master of Science in Information Technol

Approved By:

Advisor:

Gebeyehu Belay (PhD) _____

Name Signature Date

External Examiner:

Dereje Teferi (PhD) _____

Name Signature Date

Internal Examiner:

PrasadKrishna (PhD) _____

Name Signature Date

Chair Holder

Prof. Bandaru R.K.Rao _____

Name Signature Date

Faculty Dean:

Dawed Neseru _____

Name Signature Date

DEDICATION

To myfamilies

ACKNOWLEDGEMENTS

First and for most I would like to thank the Almighty God for supporting me in all walks of my life including this course work. It could have not been possible for me to pass all the miserable ups and downs of life and multidirectional challenges if I was not completely relied upon him.

Next, my heartfelt gratitude goes to my advisor Dr. Gebeyehu Belay, first, for accepting me to advise at late times, and second, for his timely commendable and extraordinary constructive comments that really gave life to this thesis. I thank my previous advisor Dr. Bandaru Ramaiah too for his unconditional challenges that gave me unparalleled lessons on how to treat life on a different level and direction than the usual.

I also sincerely acknowledge Bahir Dar University for providing me the chance, the required finance for the study as well as the time in under taking this study. Felegehiwot Referal Hospital, Gamby Teaching Hospital and Adinas General Hospital have special place in my heart for giving me technical advice and providing me all the required information available on hand.

My husband Tilahun Ayalew and my son Kinfemichael Tilahun should also be praised for their understanding, patience, extraordinary love and encouragement during the course work.

Finally, my sincere thanks also goes to my beloved friends and colleagues especially Ato Abriham Debasu, who has been on my side in all the times in advising and supporting me during the course work and encouraged me to succeed in my academic endeavor.

ABSTRACT

The diagnosis of tuberculosis and lung cancer is difficult as symptoms of both diseases are similar. A missed or wrong diagnosis of lung cancer or TB by clinician can lead to delays in diagnosis and treatment and hence progression of the disease. This indicates that lung cancer is often misdiagnosed as pulmonary tuberculosis, and vice versa in most cases. Itschallenges and problems are big concerns in most developing countries, including Ethiopia.

Research Issue: As the problem of lung disease is raising over time, unlike the Ethiopian context which stuck on a diagnostic radiologist, globally, the means and techniques for detecting same have also been increasing. However, approaches integrating two or more lung disease together are rare.

Methods: In order to achieve the objective of the research, image processing based lung disease classification techniques using MATLAB proposed and defined. Accordingly, a digital image analysis technique based on morphological and Texture features was developed to classify the two lung diseases. Sample lung images taken from three hospitals and the internet, and on average 100 images taken from each; Normal Lung, Lung TB and Lung Cancer.

Finding: Approaches of KNN, Naïve Bayes and Neural Network classifiers on each classification parameters of morphology, texture and the combination of the two are compared. To evaluate accuracy of the classifier, 70% of the data set used for training and the remaining 30% for testing. The classification system is supervised corresponding to the predefined classes of the lung image. It is found that the classification performance of KNN is better than Naïve Bayes and ANN classifier. It is also identified that the discrimination power of texture feature is better than morphology feature, but when two of the features are used together the classification accuracy is greater. Of all the classification approaches, the best classification performance is obtained using KNN (specificity of 90%, and Sensitivity of 86.67% for Lung TB and 83.33% for Lung Cancer). The accuracy obtained from this approach is 86.67%.

Conclusion/Originality: The finding of this study revealed that the two major and ever deadly lung diseases can be classified more accurately from x-ray image than a radiologist can do. This will pave the way in treating the two diseases before progression and saves the lives of many in developing countries like Ethiopia.

Key words: Lung TB, Lung Cancer, image classification

TABLE OF CONTENTS

Contents Pages

LIST OF ABBREVATIONS

AFB	Acid Fast Bacilli
AGH	Adinas General Hospital
AIDS	Acquired Immuno Deficiency Syndrome
ANN	Artificial Neural Network
CAD	Computer Aided Diagnosis
COPD	Chronic Obstructive Pulmonary Disease
DFOV	Display Field OF View
FHRH	Felegehiwot Referal Hospital
GLCM	Gray-Level Co-occurrence Matrix
GTH	Gamby Teaching Hospital
HBC	High Burden Countries
HIS	Hue Saturation and Intensity
HIV	Human Immune Virus
JPEG	Joint Photographers Expert Groups
KNN	K-Nearest Neighbor
LC	Lung Cancer
LDRS	Lung Disease Classification system
MATLAB	matrix laboratory
MC	Montgomery County
MDR	Multidrug resistant

MLP	Multi layer Prceptron
MRI	Magnetic Resounance
NAA	Nucleic Acid Amplification
NSCLC	Non-Small-Cell Lung Carcinoma
RGB	Red Green and Blue
RIO	Region Of Interest
SCLC	Small-Cell Lung Carcinoma
TB	Tuberculosis
WHO	World Health Organization
ZN	Ziehl-Neelsen

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Background

Lung disease refers to many disorders affecting the lung such as asthma, Chronic Obstructive Pulmonary Disease (COPD), tuberculosis, influenza, lung cancer, pneumonia and other breathing problems. Lung disease signs and symptoms can differ by the type of the affected disease. Common signs are; trouble breathing, shortness of breath, feeling like not getting enough air, decreased ability to exercise, a cough that won't go away, coughing up blood or mucus, and pain or discomfort when breathing in or out [1][2]. Such diseases are caused by infection.

Tuberculosis (TB) is an infection disease and the most common over the world. It is commonly caused by bacteria, which is known as Mycobacterium tuberculosis and mostly affects the lungs of human. TB is spread through the air from everyone and everywhere to the other healthy people. By coughing, sneezing, spitting felt in patients, TB bacteria disperses widely into the air. In every year, third of population of the world has got Mycobacterium TB bacteria at a rate of one percent of population with new infection [3].

Tuberculosis has been present in humans since antiquity. Tubercular decay has been found in the spines of mummies from 3000 to 2400 BC [4]. Hippocrates identified phthisis (a Greek term for TB-460 BC) as the most widespread disease of the times involving coughing up blood and fever, which was almost always fatal. The bacillus causing TB was identified and described in 1882 by Robert Koch. This author also found

the similarity of bovine and human TB through a classification of cow milk as pasteurized and unpasteurized, and he knew a glycerin extract for tubercle bacilli as a remedy for TB.

Mycobacterium tuberculosis is the causative agent of TB, that one of the world's most devastating human pathogens, which causes more than 2 million deaths annually. In addition, an estimated 2 billion people are latently infected with M. tuberculosis. Ethiopia is ranked the 7th among TB burden shouldering countries in the world [5].

A person with untreated pulmonary TB is estimated, on average, to infect 10-15 persons annually. A primary infection due to Mycobacterium tuberculosis may actively develop into clinical TB, pass as in apparent infection, or remain latent in the individual for months or years depending on the various host and environmental factors. Overt TB, thus, could result from a reactivated latent infection or from a recent primary infection or (secondary) re-infection. It has been observed that the transmission of Mycobacterium tuberculosis is favored by dusty environment, poor ventilation, little sun light, malnourished, background of alcohol and drug abuse, overcrowding, relative virulence of the strain, the intensity of exposure to an infectious TB case (closeness and duration), and the susceptibility and immune status of the exposed individual [6, 7].

Lung cancer is a disease of abnormal cells multiplying and growing into a tumor. The mortality rate of lung cancer is the highest among all other types of cancer. Lung cancer is one of the most serious cancers in the world, with the smallest survival rate after the diagnosis, with a gradual increase in the number of deaths every year. Survival from lung cancer is directly related to its growth at its detection time. However, people do have a

higher chance of survival if the cancer can be detected in the early stages [8]. Cancer cells can be carried away from the lungs in blood, or lymph fluid that surrounds lung tissue. Lymph flows through lymphatic vessels, which drain into lymph nodes located in the lungs and in the centre of the chest. Lung cancer often spreads first toward the centre of the chest because the natural flow of lymph out of the lungs is toward the centre of the chest. Lung cancer can be divided into two main groups—non-small cell lung cancer and small cell lung cancer. These assigned groups of lung cancer types are dependent on their cellular characteristics [9].

Furthermore, the etiology of lung cancer has been associated with smoking, occupational exposure to arsenates, nitrosamines, asbestos, and aromatics, and indoor exposures to radon, and to fumes from fires or cooking stoves. And Outdoor air pollutions also substantially contribute to the burden of lung cancers in urban dwellers. Inflammation processes have long been linked to cancer development. Among intrinsic lung diseases with inflammatory components, chronic obstructive pulmonary disease (COPD), asthma, and pulmonary fibrosis have been linked to lung cancers. Tuberculosis with more than 80% of the cases primarily affecting the lungs entails a chronic inflammatory process. Coexistence of tuberculosis and lung cancers is not uncommon clinically [10].

Higher prevalence of tuberculosis and overlap of its clinical presentation and radiological features with lung cancer creates a scenario where a significant number of early lung cancer patients may get wrongly labeled as tuberculosis [11]. Early diagnosis and immediate initiation of treatment are essential for an effective TB control. Delay in diagnosis is significant to both disease prognosis at the individual level and transmission

within the community. Most transmissions occur between the onset of cough and initiation of treatment. The diagnosis of pulmonary TB depends on clinical suspicion, response to treatment, chest radiographs, staining for acid-fast bacilli (AFB), culture for TB, and nucleic acid amplification (NAA) [12].

Despite many advances in the diagnosis of TB in recent years, sputum smear testing using the Ziehl-Nielsen stain (ZN) is still the basic tool for TB diagnosis and monitoring because it is a quick, simple, and low cost test that can be reproduced in any setting and used to detect infectious cases in the community, a task that constitutes the cornerstone of TB diagnosis and monitoring [13]. The culture has always been considered to be the gold standard technique for the diagnosis of TB. The result may be negative in some smear-positive patients owing to the loss of viability of the bacilli or the process used to decontaminate the sample. Likewise, false positive results may arise because of contamination of specimens in the laboratory. Despite these limitations, culture still plays a key role in the diagnosis and management of TB [14].

In the case of lung cancer, one of the most important and difficult tasks the radiologist has to carry out consists of the detection and diagnosis of cancerous lung nodules from chest radiographs. Some of these lesions may not be detected since they may be hidden away by the underlying anatomical structure, or the poor-quality of the images or the subjective and variable decision criteria used by radiologists [15].

The clinical importance of chest radiographs, combined with their complicated nature, explains the interest to develop computer algorithms to assist radiologists in reading chest images [15]. To provide accurate diagnosis, nowadays, computer-aided diagnosis (CAD)

becomes one of the major research areas in medical imaging. Basically, computer-aided diagnoses are processes which give a lot of information that helps physicians understand medical images so that the accuracy of medical diagnosis could be improved and the time taken in reading an image by traditional methods could be decreased. Using image processing, researchers have now focused on developing algorithms that detect many types of diseases, such weakening of brain arteries, retinal fundus, lung cancer and pulmonary nodules, breast cancer, kidney diseases, and coronary artery diseases, to mention some that are helping radiologist in their decision-making. The CAD algorithm is provided with functions that automatically analyses acquired image and provides an automatic diagnosis to identify the suspected regions from images [16].

However, medical image processing needs continuous enhancements in terms of techniques and applications to help improve quality of services in health care industry [17]. Accordingly, this study is meant to design an image processing technique enable to detect the two killer lung diseases more accurately from x-ray image.

1.2. Statement of the Problem

Manually physicians diagnose lung diseases by simply observing x-ray, CT and MRI images. In this regard, one of the most important and difficult tasks the radiologists have to face is the detection and diagnosis of abnormalities from chest radiographs by naked eye. Some of these injuries is hardly detected because of their complicated nature. According to studies, radiologists fail to diagnose small lung nodules in as many as 30% of positive cases [18].

In recent years, globally, the CAD system for lung disease diagnosis has given a due attention and increasing over time. Accordingly, various researches have been carried out on the classification of lung diseases, including lung cancer and lung TB, based on CT scan and x-ray images by applying different image processing techniques.

Due to better clarity, low noise and distortion than x-ray images, CT images are usually preferred by researchers as data input in most of the literatures [19]. Accordingly, an automated system for nodule detection and classification [20], a system that identifies stages of lung cancer [9], a system that classifies lung diseases using Naïve bayes and decision tree classifier [1] and a GLCM and ANN based system that classifies stages of lung diseases [21] and many more works developed from a CT image using different processing techniques

However, x-ray is more generally available than CT images and thus initial diagnosis for TB and lung cancer are now widely performed by physicians mainly based on chest-x-ray images [1]. Though relatively small in number, there are also similar works that make use of x- ray images for lung disease classification. A lung segmentation method that identifies stage of lung TB [22], a texture analysis system that identify interstitial lung disease, and an image processing system that identifies lung TB using supported vector machine [23] are some to mention.

From what we have observed, the depth and width of many researches vary each other. The very close literature we came across during our review that resembles our proposed system is the work of Patil S.A. and Kuchanur M.P. (2012). This system classifies lung

cancers into malignant (SC, NSC) and benign (TB) from x-ray images using GLCM and ANN classifiers with 49 images from each and finally arrive at an accuracy level of 83%.

Despite continuous improvements in medical image processing, there is a lack of an integrated approach to support lung disease treatment, and most of the image processing techniques so far gave much attention to lung cancer, and usually make use of CT scan images (which are expensive and are not available everywhere). From all the available vast number of works on medical diagnosis, image processing techniques that classify lung TB and Lung Cancer at a time are rare, and if available less accurate and needs improvement. Besides, the approach in the Ethiopian context which is still sticking on the traditional method of lung disease detection, an x-ray imaging and a diagnostic radiologist, demands efficient detection system.

Therefore, instead of dedicating the classification technique on either of the two lung diseases, and at the same time considering the Ethiopian real situation, this study is supposed bridge the gap by employing a method of classifier designed to detect the two major lung diseases; lung TB and lung cancer from an x-ray image. To enhance all the previous works and obtain more accurate results, we acquired reasonable number of lung images, more features are extracted and comparisons are made among different classifiers. Hence, we try to answer the following research question:

Question 1; What are the common features of lung TB and lung cancer, and their distinct features?

Question 2; How to develop an effective lung disease diagnosis model?

Question3; What are the basic techniques to define using features of diseases from x-ray image?

1.3. Objective of the Study

1.3.1. General Objective

The general objective of this study is to develop lung disease Classification system for the detection of Lung TB and Lung cancer.

1.3.2. Specific Objectives

The specific objectives of the research are defined as follows

- To diagnosis lung diseases like; lung TB and lung cancer
- To introduce x-ray classification techniques for lung disease diagnosis performance.
- To implement dynamic feature extraction to improve accuracy of lung disease detection.
- To synthesis an x-ray image limitation and challenge for the better lung disease classification

1.4. Scope and Limitation

The system classifies occurrences only on the lung using only on two diseases, lung TB and Lung Cancer. Besides, accuracy level of the result is totally dependent on the quality of the source image, which is usually difficult in many healthcare giving institutes since most health facilities are less equipped with modern and quality equipments related

to x-ray imaging. Besides, the number of features, image acquisition environment and similar factors during the imaging process may affect the result. Accordingly, factors that probably set back the accuracy level of the system should be investigated for further improvement.

1.5. Significance of the Study

This thesis finding would be essential promoting good and convenient methods for classifying lung diseases type. Knowledge and experience of a physician can be one asset. However, as they have very close symptoms, without having image classification system to achieve the desired performance in diagnosis and treatment of the lung patient is impossible. Using image processing techniques for lung diseases x-ray image classification support is to achieve a better result so the diseases can be treated early and least costly. The study finding also essential to the radiologist/physicians in delivering a proper result on the given case and using the extra effort for some other productive purpose that enable to improve patient case. The patient is also benefit in getting the required quality medical service timely and at a better price to finally enjoy a healthier and happier life than otherwise.

1.6. Thesis Organization

The thesis is organized to in five chapters. The first chapter is an introduction and describes background, source issues, objectives of the thesis, methodologies as well as its significance. In the second part of the thesis, more emphasis is given to related

literatures on lung TB and lung cancer. Besides, image processing works on the two lung diseases reviewed.

The third part of the thesis focused on the design of disease classification, especially on features and classifiers. On the fourth chapter, the results from the experimentation described and discussed in detail. The final part of the study concludes the thesis and recommendation on future improvements.

2. LITERATRE RIVIEW

2.1.Introduction

The lungs are two spongy organs found in the chest. They are responsible for delivering oxygen to the bloodstream. When breathed in, air moves into the lungs causing them to expand. The air can then come very close to blood that is traveling in small vessels called capillaries. When breathed out, substances that are not needed, like carbon dioxide are exhaled. The lungs are specially designed to place blood in close contact with as much air as possible, so their tissues are very delicate [24].

It is part of a complex apparatus, expanding and relaxing thousands of times each day to bring in oxygen and expel carbon dioxide. Lung disease can result from problems in any part of this system and according to webmd, fall on one of the following types [25]; Lung Diseases Affecting the Airways, Lung Diseases Affecting the Air Sacs (Alveoli), Lung Diseases Affecting Blood Vessels, Lung Diseases Affecting the Pleura and Lung Diseases Affecting the Chest Wall

2.2.Pulmonary Tuberculosis

Tuberculosis (TB) is caused by bacteria (Mycobacterium tuberculosis) that most often affect the lungs. Tuberculosis is curable and preventable. TB is spread from person to person through the air. When people with lung TB cough, sneeze or spit, they propel the TB germs into the air. A person needs to inhale only a few of these germs to become infected[18].

About one-third of the world's population has latent TB, which means people have been infected by TB bacteria but are not (yet) ill with the disease and cannot transmit the disease. People infected with TB bacteria have a 10% lifetime risk of falling ill with TB. However, persons with compromised immune systems, such as people living with HIV, malnutrition or diabetes, or people who use tobacco, have a much higher risk of falling ill [ibid].

2.2.1. Lung TB Diagnosis Methods

Before clinicians can diagnose TB disease for a patient, they must think of the possibility of this disease when they see a patient with symptoms of TB or abnormal chest x-ray findings. Because TB is not as common as it was many years ago, many clinicians do not consider the possibility of TB when making diagnoses for patients who have symptoms. When this happens, the diagnosis of TB may be delayed or even overlooked, and the patient will remain ill and possibly infectious. Such a TB diagnosis can be categorized into four, which includes the medical history, the tuberculin skin test, the chest x-ray and the bacteriologic examination.

2.2.1.1. The Medical History

a. Exposure to TB: One important part of the medical history is asking a patient about his/her exposure to TB. Patients should be asked whether they have spent time with someone who has infectious TB or someone with TB-like symptoms. Anyone who has been exposed to TB may have LTBI. Many people become infected with *M. tuberculosis* without knowing the patient's status and condition

The risk of being exposed to TB is higher for some occupations (for example, certain health care workers) and in some residential facilities (for example, nursing homes or correctional facilities).

b. Symptoms of TB disease: Another important part of the medical history is checking for symptoms of TB disease. Although, people with TB disease may or may not have symptoms, most patients with TB disease have one or more symptoms that led them to seek medical care.

c. Previous TB infection or TB disease: During the medical history, the clinician can get valuable information from a patient, whether he/she has ever been diagnosed with or treated for TB infection or disease.

d. Risk factors for developing TB disease: A fourth part of the medical history is checking for risk factors for developing TB disease like: HIV infection, Low body weight (10% or more below an ideal), Diabetes mellitus, Chronic renal failure, Certain types of cancer, Solid organ transplant

2.2.1.2. The Tuberculin Skin Test

Patients with symptoms of TB disease are often given a tuberculin skin test to detect exposure to and infection with TB. However, 20% of the patients found to have TB disease have a negative tuberculin skin test reaction. For this reason, patients with symptoms of TB disease should always be evaluated for TB disease, regardless of their skin test results.

2.2.1.3. The Chest X-ray

The chest x-ray is useful for diagnosing TB disease. About 85% of TB patients have pulmonary TB. Usually, when a person has TB disease in the lungs, the chest x-ray appears abnormal. It may show infiltrates (collections of fluid and cells in the tissues of the lung) or cavities (hollow spaces within the lung that may contain many tubercle bacilli).

2.2.1.4. The Bacteriologic Examination

The next step in diagnosing TB disease is the bacteriologic examination. This is done in a laboratory that specifically deals with *M. tuberculosis* and other mycobacteria (a mycobacteriology laboratory). There are four parts to a bacteriologic examination.

a. Obtaining a specimen
b. Examining the specimen under a microscope
c. Culturing the specimen
d. Conducting drug susceptibility testing

2.3. Lung Cancer

Lung cancer refers to growth of malignant cells in the tissue of the lung which occurs when cells in the lung begin to grow out of control and can invade nearby tissues or spread throughout the body. Large collections of cancer cells are called tumors. If left untreated, this growth can spread beyond the lung by the process of metastasis into nearby tissue or other parts of the body. Most cancers that start in the lung, known as

primary lung cancers, are carcinomas. The two main types are small-cell lung carcinoma (SCLC) and non-small-cell lung carcinoma (NSCLC). Lung cancer, the leading cause of cancer death among men and the second among women, is responsible for 1.3 million deaths worldwide annually [24]. Treatment options are those commonly associated with other cancers and include surgical resection, chemotherapy, radiation therapy, and palliative therapy. Common symptoms of lung cancer include shortness of breath, chronic cough, weight loss, and fatigue [26].

2.3.1. Lung Cancer Diagnosis Methods

Imaging tests: Performing a chest radiograph is one of the first investigative steps if a person reports symptoms that may suggest lung cancer. This may reveal an obvious mass, widening of the mediastinum (suggestive of spread to lymph nodes there), atelectasis (collapse), consolidation (pneumonia), or pleural effusion. CT imaging is typically used to provide more information about the type and extent of disease. Bronchoscopy or CT-guided biopsy is often used to sample the tumor for histopathology [27].

Lung cancer often appears as a solitary pulmonary nodule on a chest radiograph. However, the differential diagnosis is wide. Many other diseases can also give this appearance, including metastatic cancer, hamartomas, and infectious granulomas such as tuberculosis. The definitive diagnosis of lung cancer is based on histological examination of the suspicious tissue in the context of the clinical and radiological features. A PET scan and an MRI scan of the brain are often done to examine other areas of the body where lung cancer can spread [28].

Sputum cytology: If there is a serious cough and are producing sputum, looking at the sputum under the microscope can sometimes reveal the presence of lung cancer cells.

Tissue sample (biopsy): While all of the remaining tests are important pieces of the puzzle, a biopsy is the only way to know for sure if there is a cancer. A biopsy takes a sample of the suspicious area, which is then examined under a microscope for the presence of cancer cells. In addition, the biopsy is necessary to determine the type of lung cancer and if cancer cells are present in the lymph nodes [28].

2.4. Prevalence of Lung TB and Lung Cancer in Ethiopia

Ethiopia is one of the 22 high burden countries (HBCs) and TB remains one of the leading causes of mortality. According to the 2014 WHO report (figure 2.1), the prevalence and incidence of all forms of TB are 211 and 224 per 100,000 of the population, respectively. Excluding HIV related deaths, in 2013 TB mortality was estimated to be 32 per 100,000 of the population. About 13% of all new TB cases are also HIV infected. Moreover, Ethiopia is one of the high TB/HIV and multidrug resistant TB (MDR TB) burden countries. Among TB patients with known HIV status, about 11% were HIV co infected. According to the recent national TB drug resistance surveillance report, 2.3% of new TB cases and 17.8% of previously treated TB cases were estimated to have MDR.

Figure 2.1 TB update 2014, March 2015

On the other hand, Lung cancer is the most frequent cancer death among men in the world with an estimated age-adjusted mortality rate of 23.0 per 100,000 in the year 2008. Furthermore, smoking is the leading cause of cancer and was approximated to contribute to about 21% of all deaths from cancer worldwide. In Sub Saharan Africa, the prevalence of smoking was 28% among male and 8% among female population aged 15 years and above in 1995. Despite this relatively high prevalence in this region, data on lung cancer prevalence is unavailable in Ethiopia [25].

2.5. The Mimic Between Lung TB and Lung Cancer

Lung cancer (LC) is the most deadly type of cancer and represents a major public health problem worldwide. It is the leading cause of cancer-related death in the world, with 1.3 million deaths annually. Similarly, another major cause of morbidity and mortality, especially in developing countries, is tuberculosis. In low incident countries with high

incidence of lung cancer and varying clinical presentations, TB often gets misdiagnosed with the result of delayed treatment start and unnecessary diagnostic procedures. On the other hand, many early lung cancers have been treated wrongly as pulmonary TB because the clinical and the radiological features for both conditions are similar and leading to delay in the correct diagnosis as well as exposure to inappropriate medication. Several factors are responsible for this situation in developing countries, including lack of awareness, inadequate infrastructure and socio-economic factors [29] [30][31].

There are many similarities between Lung Cancer and TB like; they are very common, have high prevalence, involve lung parenchyma and above all, characterized by similar symptoms. But, there are many differences between these two entities like they have different etiologies (pulmonary tuberculosis is infectious while lung cancer is the non-infectious disease), different consequences, and altogether different management. Delay in the diagnosis and treatment of lung cancer results in poorer outcome and lower survival [32]. According to studies, delay in diagnosis of lung cancer was significantly high in patients who had received anti-tubercular treatment for current symptoms compared with those who did not. This indicates that lung cancer is often misdiagnosed as pulmonary tuberculosis, and these patients are presumptively given ATT, hence causing a significant delay in diagnosing cancer. The majority of lung cancers (< 80%) are diagnosed at an advanced stage, i.e. stage III and IV [ibid].

2.6. Image Processing Based Diagnosis Techniques

Image processing has become an active research topic in recent years. Researchers have focused on developing algorithms using image processing to detect many types of

diseases using computer aided diagnosis, weakening of brain arteries, retinal fundus, lung cancer and pulmonary nodules, breast cancer, kidney diseases, and coronary artery disease [16].

Improving the quality of CAD diagnosis, increase therapy success by early detection of a disease, avoid unnecessary biopsies, reduce radiotherapist interpretation time, eliminates the need of repeated visits of patient to a doctor, increases accuracy of diagnosis, and improves the reliability of diagnosis. Although CAD is improving overtime, its accuracy has been an issue of concern in some areas. But, if new techniques and advanced methods like a number of classifiers are attached with the system then it can produce far more accurate results in diagnosis [ibid].

Physicians are usually unable to diagnose accurate lungs disease by only viewing x-ray, CT image, MRI image etc. This signifies the importance of using image processing to diagnose lung diseases. A study by [Sagar N. Vidhate, V. S. Dhongde (2013)] used texture analysis systems to diagnose lung disease from microscopic images of patients which are affected with interstitial lung disease (ILD) using high-resolution computed tomography (HRCT) data. This system segments the right and left lung in to three different sections and makes an analysis of texture patterns and then a classifier is trained to distinguish between emphysema and no emphysema tissue. In the texture analysis and classification values of standard deviation, entropy and texture index are considered and fuzzy logic is used in the classification of lungs disease. A HRCT image gives accuracy of 70-80 percent but by using microscopic images increased to an

accuracy level to 90% [33]. The image processing techniques developed by many authors on lung TB and lung cancer independently are described below;

2.6.1. Lung TB

Despite the existence of an effective and affordable cure, tuberculosis (TB) remains one of the world€s major health care challenges. Mortality and morbidity rates are only slightly lower than those of the well-known HIV/AIDS epidemic, but TB has received less attention of the media and public. One of the reasons for this has been the decline of TB in high-income countries. TB is diagnosed using a combination of clinical symptoms, chest radiography, and sputum examination. The typical symptoms associated with TB are fever, weight loss, night sweats, and coughing. Manually the detection of TB cavities, which is done by just looking at the X-rays/CT images by physicians/technicians by looking at the images by the naked eye, there are more chances for wrong prediction of the intensity of the cavities. Because of this wrong prediction of the cavities, the physicians may not prescribe the correct dosage of medicine. They may prescribe high or low dosage of medicine. If the dosage is too high it leads into various harmful effects such as causing other diseases. If the dosage is too low the patient cannot easily recover from the disease soon. So the accurate detection of the cavities must be done for the accurate prescription of medicine with the correct dosage to get rid of the disease completely [23].

So the automatic detection of tuberculosis from X-ray may be helpful in the rural area where an expert radiologist is not always available. Developing a CAD system for diagnosing TB is a challenging task. It includes; segmentation, feature extraction and

classification. In recent years, due to the complexity of developing fully automated CAD systems for x-ray analysis, research has concentrated on developing solutions for specific sub problems. The segmentation of the lung field is a typical task that any CAD system needs to support, for a proper evaluation of CXRs. In the segmentation step, shape methods were used to segment the lung field correctly. In general, segmentation in medical images has to cope with poor contrast, acquisition noise due to hardware constraints, and anatomical shape variations. Depending on the lung segmentation, different features were extracted for the further analysis. The extracted features are input to the classifier, which then classifies a given input image into either normal or abnormal. Here first extract the lung region using a region based active contour segmentation method. For this lung region, compute a set of texture and shape features, which enable the x-rays to be classified as normal or abnormal using a binary classifier [5][6].

Detecting cavities from chest-x-ray is an efficient method for diagnosing the TB. Region based active contour segmentation is used for segmenting lung field, and the extracted features are classified using supported vector machine as normal and abnormal. The Montgomery County (MC) Data set contains 138 posterioanterior cxrs, among which 80 cxrs are normal and 58 cxrs are abnormal with manifestations of TB are used. All images of the MC set are in 12-bit grayscale, captured with a Eureka stationary-x-ray machine (CR). The abnormal cxrs cover a wide range of-related abnormalities, including effusions and miliary patterns. TB can be detected from Chest-x-ray images by using image processing methods like segmentation, Feature Extraction and classification. Existing diagnostics method such as sputum staining has become less reliable in high

population. So this method can be helpful in rural areas. Increasing the features selected and using another segmentation method may get more accurate [ibid] result.

Rachna H. B., M. S. Mallikarjuna Swamy (2013) developed an algorithm based image processing for identification of TB bacteria in sputum. The method is based on Otsu thresholding and k-means clustering approach. The performance of clustering and thresholding algorithms for segmenting TB bacilli in tissue sections was compared. Authors developed a segmentation algorithm to automate the process of detection of TB using digital microscopic images of different subjects. A performance comparison of clustering and thresholding algorithms for segmenting TB bacilli in stained tissue slide images was carried out [34].

Adgaonkar A. et al (2014) also made use neural network based classifiers image processing techniques for an automatic identification of TB bacilli and found a good output. The proposed system performed 93.5% sensitivity for identifying individual bacilli. The technique involved segmentation followed by an identification procedure. The segmentation allowed the elimination of a great amount of unwanted objects, and therefore, only those characterized to have a similar color as that of the bacilli were retained [35].

On the other hand, in Adi K. et al (2013), research, an algorithm developed to identify and count the number of tuberculosis is also another interesting approach that focused on microscope imaging. Color segmentation done by way of extracting the saturation channel of NTSC (Luminance, Hue, saturation) color model. Feature extraction for bacteria shape identification process was using two parameters eccentricity and

compactness. The training and object classification was using Support Vector Machine algorithm [36]. They proved that Support Vector Machine is good to be applied in detecting and counting the number of tuberculosis bacteria

Ziehl-Neelsen (ZN) stained acid-fast bacilli (AFB) in digital images is detected using innovative computational algorism by P. Sadaphal, J. Rao, G. W. Comstock, M. K. Kennedy, F. Beg [37]. Automated, multi-stage, color-based Bayesian segmentation identified possible „TB objects€, removed artifacts by shape comparison and labeled objects as „definite€, „possible€ or „non-TB€, bypassing photo micrographic calibration [37].

Management of tuberculosis cavities that are clearly benign or malignant is straightforward. The difficulty is in the evaluation and management of the indeterminate nodule and the goal was to correctly diagnose indeterminate tuberculosis cavities, allowing curative resection of early-stage malignant tuberculosis cavities and avoiding the morbidity and mortality of surgery for benign tuberculosis cavities. The proposed technique was successful in detecting tiny cavities on lung x-ray images [38].

A paper by [Manisha R K, Palanisamy K S. (2016)] detailed an automated approach for lung TB diagnosis and makes use of Chest radiography for same. The lung region is extracted using Graph cut lung segmentation method for identifying the ribs and clavicles, which are needed for the diagnosis. The Graph cut segmentation method provides better accuracy and then the classification is performed between normal and abnormal x-ray patterns. Finally, the research found that the automated approach provides better performance than the manual diagnosis of Bhen and the stages are identified using classification algorithm [22].

2.6.2. Lung Cancer

Lung cancer seems to be the common cause of death among people throughout the world. Early detection of lung cancer can increase the chance of survival among people. The overall 5-year survival rate for lung cancer patients increases from 14 to 49% if the disease is detected in time. One of the most important and difficult tasks the radiologist has to carry out consists of the detection and diagnosis of cancerous lung nodules from chest radiographs. Some of these lesions may not be detected since they may be camouflaged (hidden away) by the underlying anatomical structure, or the low-quality of the images or the subjective (one-sided) and variable decision criteria used by radiologists. Previous studies showed that radiologists fail to diagnose small lung nodules in as many as 30% of positive cases [18]. Hence, a lung cancer detection system using image processing has been used to classify the present of lung cancer in images. In studies of this kind, MATLAB was used in all the procedures. To obtain more accurate results stages were divided in to three: Image Enhancement stage, Image Segmentation stage and Features Extraction stage [39].

Al-Fahoum A. et al (2014) developed an automated intelligent system for nodule detection and classification could read the DICOM CT images and applied some advanced image processing principles to facilitate the segmentation and detection of mass lesions. The small-sized cancer areas, presumably when they are biologically early in their evolution, are amenable to surgical cure. The proposed system was also able to display the size of the detected cancer, to replace the manually process of measurements that is taken by the radiologist to measure the width (transverse) and the length (anterior

posterior) distances of the cancer area. The system discussed in this study displays each detected area boundary to simplify the detection of region boundary that is subjected to the observer variations. Low contrast cancer areas that have advanced stages in the disease may have calcification, necrosis and cavitations; low contrast regions constitute a challenge and a source of error for radiologist [20].

G. Vijaya, A. Suhasini, and R. Priya (2014) and Vanitha Lakshmi. K (2013) also proposed an automatic cancer detection system of which the features focused on a hybrid approach called neuro - fuzzy algorithm. The segmentation was achieved through a series of techniques including, thresholding, median filtering, closing, and labeling. Lung region was extracted from the original CT image. From the lung region, the ROIs were obtained. The nodules were evaluated based on the features such as size of area, circularity, skewness, kurtosis and mean and then subjected to classification to classify the odules. Neural fuzzy model was designed to extract suitable diagnosis rules, and classified the true nodules from the ROIs [40] [41].

Faleh H. Mahmood, Wafaa A. Abbas and S. M. Ali (2014) worked on a semi - automatic segmentation algorithm for lung's tumor detection and extraction that showed a good performance. The extracted tumor area from the CT slice was measured by a method based on the Display Field OF View (DFOV). To provide physicians with volumetric data, the lung CT images have been pre - processed by enhancing their contrast to make them ready for segmentation by implementing the k - means classification algorithm. As the segmentation is performed on lung region, the tumor features have been determined and isolated by performing the seeded region growing algorithm. The tumor areas of the

image slices have been calculated and used to determine the tumor volume by stacking the extracted tumors on top of one another.

Vijay A.Gajdhane and Deepak L.M(2014) and Khin Mya, Mya Tun and Aung Soe Khaing (2014)also proposed a system that identifies stage of lung cancer from a CT scan images. The region of interest, i.e., tumor is identified accurately from the original image. Gabor filter and watershed segmentation gives best results for pre-processing stage. From the extracted region of interest, three features were extracted i.e., area, perimeter and eccentricity. These three features helped to identify the stage of lung cancer. The results indicated that the tumors were of different dimensions. By measuring the dimensions of the tumor the lung cancer stage can be detected accurately. Furthermore,for classification purpose, Support Vector Machines were an attractive approach to data modeling. They combined generalization control with a technique to address the curse of dimensionality. The kernel mapping provided a unifying framework for most of the commonly employed model architectures, enabling comparisons to be performed. In classification problems generalization control was obtained by maximizing the margin, which corresponds to minimization of the weight vector in a canonical framework. In feature extraction, physical dimensional measures and gray level co-occurrence matrix (GLCM) method were used. Artificial neural network (ANN) was applied for classification of disease stage [9][21].

In Khin M.M., Aung S.K.(2014) study, median filter used for image pre-processing and Otsu€s thresholding method for segmentation. In feature extraction, physical dimensional measures and gray level co-occurrence matrix (GLCM) method are used. Artificial

neural network (ANN) is applied for classification of disease stages from CT scan image [21].

Image processing has become an active research topic in recent years. Researchers have focused on developing an algorithm using image processing to detect the different types of cancer in its early stage. Various preprocessing steps using image processing have been proposed. But high accuracy, sensitivity and specificity are important. To achieve this, preprocessing of the acquired original image is needed. Santosh S. , Yogesh S. and Ritu V.(2016) study CT images used and the preprocessing using image histogram equalization, thresholding, filtering followed by feature extraction to reduce the process complications as well as improve accuracy [19].

CAD systems provide fast and reliable diagnosis for medical images. Essa, Magdy, Nourhan Zayed, and Mahmoud Fakhr (2015) developed an automatic system that classifies each lung into normal or cancer. In this work, using 70 CT images, Wiener filtering on the original CT images was applied as a preprocessing step. Histogram analysis combined with thresholding and morphological operations used to segment the lung regions and extract each lung separately, and Amplitude Modulation Frequency Modulation (AM-FM) method used to extract features from ROIs. The significant AM-FM features selected using Partial Least Squares Regression (PLSR), and finally, Knearest neighbour (KNN), support vector machine (SVM), Naive Bayes, and Linear classifiers used with the selected AM-FM features for classification. The performance of each classifier in terms of accuracy, sensitivity, and specificity evaluated [12].

Generally, from all the researchers observed so far we have never come across to a system that classifies Lung TB and Lung Cancer at a time except the work of PATIL S.A. and Kuchanur M. B(2012)[7]. The diagnostic results obtained from the system was as high as 83% accuracy, and the classification is achieved using training data sets of 147. Classification accuracy is improved as the numbers of training samples are increased. This study also concluded that, back-propagation algorithm of ANN is a good choice for classification of cancer and TB images.

Digital image processing techniques have been used in developing CAD systems for locating suspected nodules but too many false positive (FP) chest radiographs are made. These FP's include rib crossings, rib vessel crossings, vessel-vessel crossings and end-on vessels Accordingly, the challenge to solve for early diagnosis of lung cancer is associated with the reduction of the number of FP classifications while maintaining a high degree of true positive (TP) diagnoses, i.e., sensitivity. Several methods have been proposed to reduce the number of FP while maintaining a high sensitivity by define and specify a good feature space which will discriminate between nodules and non-nodules, malignant and benign[15].

Computed Tomography (CT) is efficient than x-ray. However, the latter is more generally available worldwide. Thus initial diagnosis for TB and lung cancer, now performed by medical doctors, is mainly based on chest X-rays images[1]. Therefore, to focus the image processing improvement on the least cost and widely used medical imaging technique, i.e., x-ray, is important for developing countries like Ethiopia where accessibility of other imaging is minimal

3. METHODOLOGY

In order to achieve objective of the study, the methodology used in this thesis work is experimental. Image processing based lung disease classification techniques using x-ray image X-ray images required for the study are collected from three hospitals and databases from the internet. To remove noises and artifacts from the acquired image, median filtering used in the preprocessing phase. Once the images are enhanced, to separate region of interest from the back ground image, image segmentation is carried out using k-means clustering. Consequently, morphology and texture features (nine feature sets in total) are extracted. Three classifiers; Naïve Bayes, ANN and KNN are trained using the features extracted from 70% (210 images) of the image data. The classification accuracy of individual classifier is tested using the features extracted from 30% (90 images) of the image data for comparison.

Finally, the results obtained from the process summarized in a confusion matrix and bar graphs. Accuracy and sensitivity analysis results used to describe inter and/or intra classifier performance.

3.1. The Implementation Tool

To process and classify the x-ray image, MATLAB R2013a application is used in the study. MATLAB is a dynamic and advanced application for such an image data. It is an interactive environment used by millions of users in science and engineering fields.

Therefore, this application is a capable and scalable tool to analyze and define a proper parameter.

MATLAB provides multi-platform environment. This tool is used for prototyping, data analysis, and visualization with built-in support for matrices and matrix operations. Also, it is loaded with graphic capabilities, as well as a friendly programming language and development environment. It enables to test algorithms immediately without recompilation, simplifies common programming/debugging tasks, and also has a very quick learning curve for using it in image processing [19].

Lung disease refers to many disorders affecting the lung. As described in section 2.1, lung disease can be classified into various categories depending on the part of the lung affected by the diseases. In the case of this research, the focus is the lung disease affecting the air sacs (alv`eoli); mainly lung TB and lung cancer. The classification of lung disease design based on the extracted features using image analysis technique is described in this chapter.

3.2. Lung Disease Classification Model

The task of classification occurs in a wide range of human activities. The problem of classification is concerned with the construction of a procedure that can be applied to differentiate events, in which they can be assigned to one of a set of predefined classes based on observed attributes or features. Image analysis or computer vision is used in the classification of lung disease. The classes are the feature or attributes computed from

lung disease images. These observed features of lung disease used to decide the class or the type of lung disease as to be Lung TB or Lung Cancer

Images of lung disease used in the study are taken from three hospitals and online from http://www.chestx-ray.com/, https://openi.nlm.nih.gov/ and https://wiki.cancerimagingarchive.net. Upon processing, the final output of the process gives the type of lung disease of the given x-ray image. Hence, in this research the main interest is to differentiate the type of lung disease (lung TB or Lung Cancer) by using image analysis technique. This is because, as described previously (in section 2.1 above), there are many similarities between Lung Cancer and TB. Both are very common, have high prevalence, involve lung parenchyma and above all, characterized by similar symptoms. However, there are many variations between these two entities like they have different etiologies, consequences, and altogether management. Delay in differentiating these two closely linked diseases in diagnosis may results in poorer outcome and lower survival. Accordingly, identifying the type of lung disease where it belongs will make the disease to be simple to cure

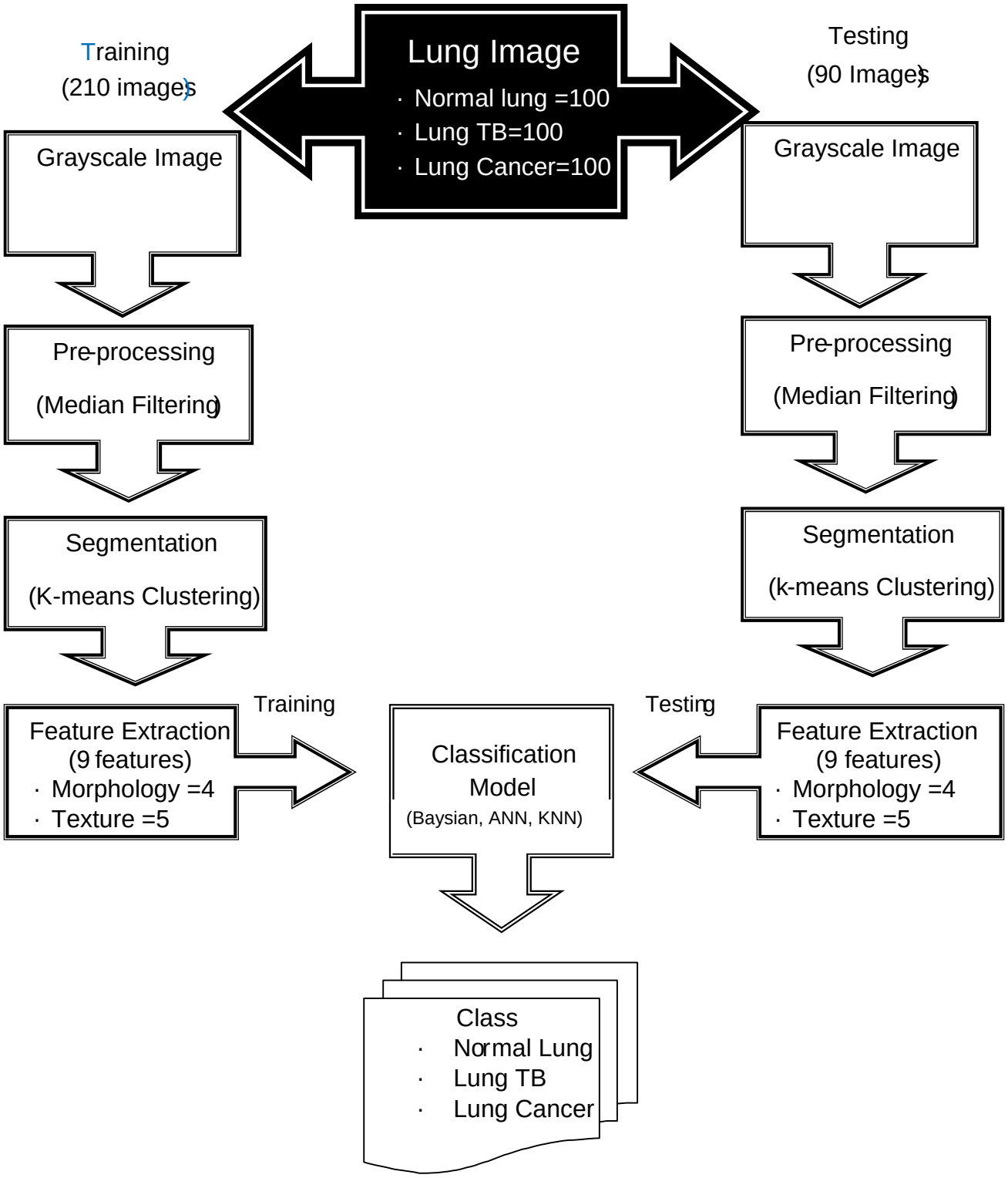

Figure 3-1: Lung Disease Classification Model

Lung disease classification comprises of: Image acquisition, image enhancement or pre processing, image segmentation, feature extraction, and finally classification. The lung disease classification model (figure 3.1) is also proposed in such order.

The 300 x-ray lung images acquired from different sources are resized to 360x360 windows and reduced to grayscale. The total images are partitioned using 70:30 ratio, 70% for the training of classifiers and 30% to test the accuracy. The next step after image acquisition is to pre-process the images so that noises and artifacts are removed from the original image. In our case this is done by median filtering. Once the images are free from noises, the segmentation process proceeds and the ROI are separated from the background images using K-means clustering and the images are ready for further process of feature extraction.

Consequently, features of morphology area (area, quivD, perimeter, and roundness) and GLCM (contrast, energy, homogeneity, correlation, and entropy) are extracted from the images. In the training line, this features used to train the three classifiers of our interest of comparison (Naïve Bayes, ANN and KNN) and finally used as a knowledge base. In the case of testing, the process is the same, and the features extracted from 30% of the images used as an input to finally produce the three classes of (Normal Lung, Lung TB and Lung Cancer) as an output and test accuracy of the classifiers under consideration.

3.3. Image Acquisition

The first step in the process of developing an image classification system is to acquire x ray image of our interest. Lung- x-ray images have higher noise compared to CT

(computed tomography) images; hence lot of noise is usually observed [9]. To improve the contrast, clarity and separate the background noise, it is required to pre-process the images. Hence, various techniques like smoothing, enhancement are applied to get the image in the required form. For this study, lung disease images were acquired from; http://www.chestxray.com/ (for normal lung Images), https://openi.nlm.nih.gov/ (for lung TB Images), and https://wiki.cancerimagingarchive.net (for Lung Cancer Images). Besides to these sources certain number of images are collected from; Felege Hiwot Referal Hospital, Gamby Teaching General Hospital and Adinas General Hospital.

The number of images taken from the three lung image categories and the sources they are acquired from are shown in Table 3-1.

Table 3-1: Lung Images taken from each type by Sources

No.	Lung Image Type and source					No. of images
	Lung type	FHRH	AGH	GTH	Internet	
1	Normal Lung	15	10	10	65	100
2	Lung TB	30	20	20	30	100
3	Lung Cancer	1	-	-	99	100
	Total	46	30	30	194	300

3.4. Image Processing

Image processing is, mainly, the manipulation of images in different ways so that the desired output of interest is obtained. For that to happen, the image required from different sources has to pass through various linked and rigorous steps.

3.4.1. Pre-processing

Image pre-processing, resample the chest radiographs by increasing the grey scale contrast and improve chest-x-ray image quality. Image pre-processing step involves addition of noise and removal of noise to suppress unwanted distortions and enhance the feature of the image for further processing. The reason for the need of image pre processing includes: noise reduction, contrast enhancement and elimination of acquisition-specific artifacts [22].

3.4.1.1. Edge Detection

Edges in images are areas with strong intensity. Edge detection of an image significantly reduces the amount of data and filters out useless information, while preserving the important structural properties in an image. Most edge detection methods work on assumption that the edge occurs where there is a discontinuity in the intensity function or a very steep intensity gradient in the image. Using this assumption, if one take the derivative of the intensity value across the image and find points where the derivative is maximum, then the edge could be located. The gradient is a vector, whose components measure how rapid pixel value are changing with distance in the x and y direction. There

are many methods of detecting edges; the majority of different methods may be grouped into Gradient and Lapalacian.

The gradient method detects the edges by looking for the maximum and minimum in the first derivative of the image and the Laplacian method searches for zero crossings in the second derivative of the image to find edges. The Sobel operator is an example of the gradient method a and popular edge detection method considered in this work. The Sobel operator is a discrete differentiation operator, computing an approximation of the gradient of the image intensity function. There exists a function, edge.m, which is in the image toolbox. In the edge function, the sobel method uses the derivative approximation to find edges. Therefore, it returns edges at those points where the gradient of the considered image is maximum [43].

The following advantages of Sobel edge detector justify its superiority over other edge detection techniques:

Edge Orientation: The geometry of the operator determines a characteristic direction in which it is most sensitive to edges. Operators can be optimized to look for horizontal, vertical, or diagonal edges.

Noise Environment: Edge detection is difficult in noisy images, since both the noise and the edges contain high frequency content. Attempts to reduce the noise result in blurred and distorted edges. Operators used on noisy images are typically larger in scope, so they can average enough data to discount localized noisy pixels. This results in less accurate localization of the detected edges.

Edge Structure: Not all edges involve a step change in intensity. Effects such as refraction or poor focus can result in objects with boundaries defined by a gradual change in intensity. The operator is chosen to be responsive to such a gradual change in those cases. Newer wavelet-based techniques actually characterize the nature of the transition for each

3.4.1.2. Median Filtering

The median is usually taken from a template centered on the point of interest. Given the arrangement of pixels, the pixel values are arranged into a vector format. The vector is then sorted into ascending order. The median is the central component of the sorted vector. The median has a well-known ability to remove noises from an image. It has also a practical advantage, owing to its ability to retain edges while suppressing the noise contamination [4].

3.4.1.3. Boundary Tracing in Lung Disease Image

A radial search technique is used for detecting lung disease borders in clinical x-ray images. First, it includes two rounds of radial search based on the same center. The first round search is independent, and the second round search is knowledge-based tracking. This algorithm provides an accurate way of detecting the border of a mole in an intensity image.

The radial search algorithm operates on an intensity image, so first the pixels of the image are converted to form an intensity mole image. Taking the centre of the intensity mole image as initial point, the algorithm casts n radial lines emanating from

this point at equal angles of (360/n) degrees. Radial search technique finds the border points through an independent search, for tracking the border based upon its nearest neighbor border point. Line by line the algorithm searches for a border point along each of the radial lines.

Figure 3-2 Traced lung Images (own experiment)

3.4.2. Image Segmentation

Image segmentation is an essential process for most image analysis subsequent tasks. In particular, many of the existing techniques for image description and classification depends highly on the segmentation results. The goal of segmentation is to simplify and/or change the representation of the image into something that is more meaningful and easier to analyze. Image segmentation is typically used to locate objects and boundaries

(lines, curves, etc.) in images. More precisely, image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain visual characteristics [45].

3.4.2.1. K-means Clustering

Cluster analysis is the assignment of a set of observations into subsets so that observations in the same cluster are similar in some sense. Clustering is a method of unsupervised learning, and a common technique for statistical data analysis used in many fields, including machine learning, data mining, pattern recognition, image analysis, information retrieval, and bioinformatics. In this work, k-means clustering is done using Euclidean distance measure for performing image segmentation. Compared to hierarchical clustering, k-means clustering is found to be simple and efficient. K-means clustering creates one set of clusters that partitions the data into similar groups, whereas in Hierarchical it finds successive clusters using previously established clusters. K-means is more efficient because it just needs to do distance calculation, whereas in hierarchical need to do full inverse distance weight where efficiency will get reduced. K-means clustering is an iterative technique that is used to partition an image into K clusters. This algorithm aims at minimizing an objective function, in this case a Euclidean distance measure.

3.4.3. Feature Extraction

Feature extraction is an important stage that uses algorithms and techniques to detect and isolate various desired portions or shapes of a given image. When the input data to an

algorithm is too large to be processed and it is suspected to be notoriously redundant, then the input data will be transformed into a reduced representation set of features [9].

Feature extraction is the method by which unique features of lung injury images are extracted. This method reduces the complexity in classification problems. The purpose of feature extraction is to reduce the original data set by measuring certain properties, or features, that distinguish one input pattern from another. In this study the two classification parameters are considered Morphology and Gray Level Co-occurrence Matrix (GLCM). The overall features considered from each are shown in Fig 3.3

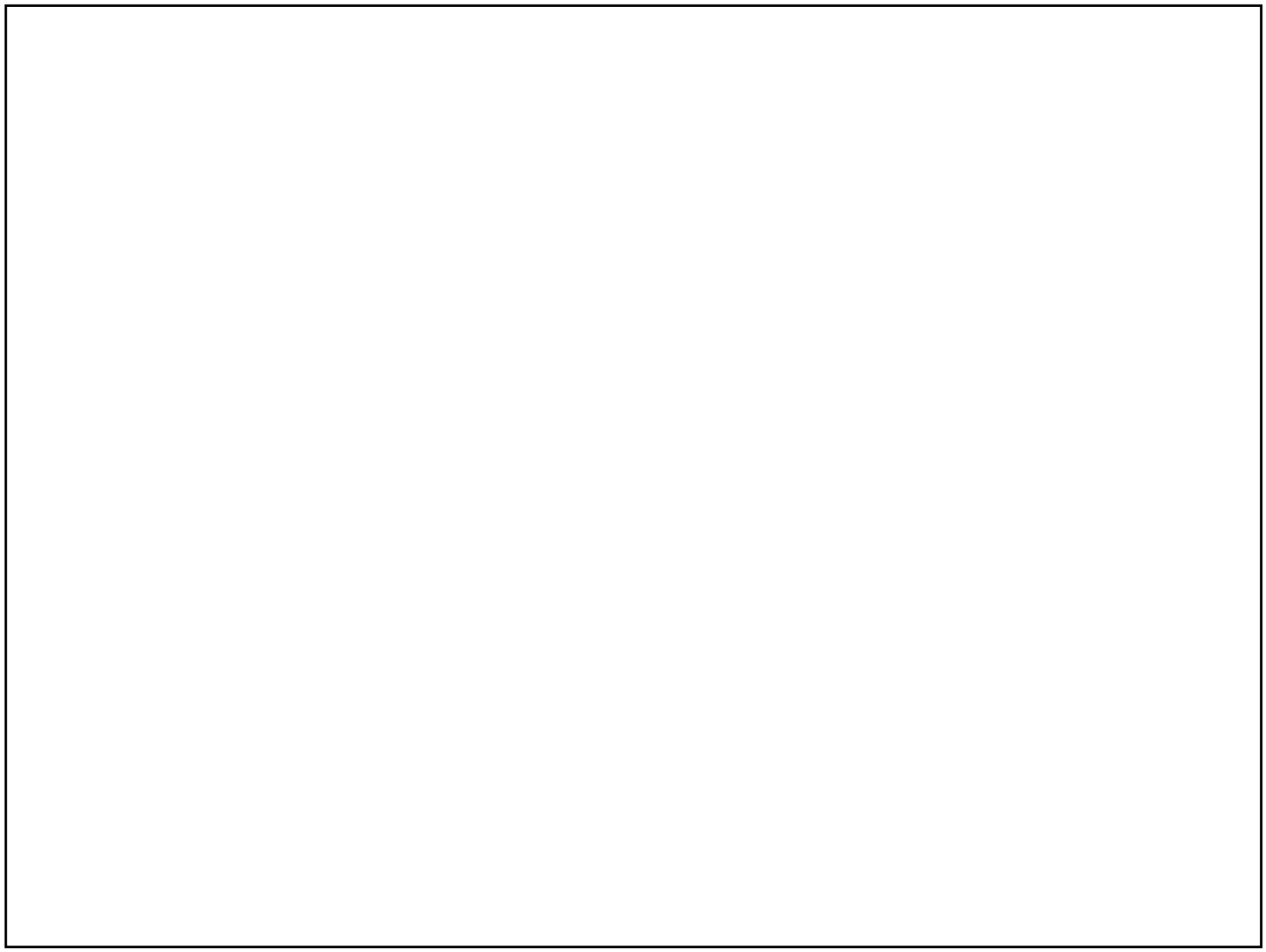

Figure 3.3: Morphology and Texture Features

3.4.3.1. Morphology Features

Morphology is the geometric property of images. In our case it is the size and shape characteristics of lung disease x-ray image. It can be obtained from the analysis of binary images. From the given morphology of lung disease images, Obj-area, EquivD, Perimeter, and Roundness features extracted.

3.4.3.2. Gray-Level Co-Occurrence Matrix

The properties of an image texture are detected indirectly by using the Gray-level co-occurrence matrix (GLCM) from which special indexes called image indicators† are exploited The GLCM is a square matrix whose size is equal to the number of gray levels which the starting image has been reduced in. It is a two dimensional histogram of gray levels for a pair of pixels, which are separated by a fixed spatial relationship. Once the window of comparison ends scanning the image, the statistical measures begin to extract the characteristics of the matrix. The next step to determine the texture features is to express GLCM's terms as probabilities, in order to achieve that goal selected statistics are applied by iterating through the matrix. The probability describes how often one gray tone will appear in a specified spatial relationship to another gray tone on the image. So the terms are divided in all possible combinations within the matrix of im selected direction [46]. Then it s considered the normalization equ follows:

..... eq.1

Where C(i, j) the value in cell (i, j), P(i, j) the probability, N the number o

The GLCM is just the tool to extract the texture features. In this study: contrast-info, correlation-info, energy-info, entropy-info and homogeneity-info are considered

3.4.4. Classification Model

In image classification, the primary objective is to categorize objects in the scene (lung disease Images) from a set of measurements of the objects. The measured values are the features of the pattern. A set of similar objects or patterns possessing more or less identical features are said to belong to a certain category called classes.

In this thesis work, five texture, four morphological and six color features are considered for the classification of lung disease image. Hence, three categories or classes of the feature values of each lung disease available in all images in the sample (N=300). In the training process, the class values were provided because supervised learning is used. In order to test the classification accuracy of the system, feature data sets not in the training data set used. In the classification process, the total data set partitioned to 70% for training and 30 % for testing. The class labels corresponding to the type of lung disease are categorical data. Hence, representing these values required using a corresponding name to simplify the representation that is appropriate to the pattern classifier program. Since there are three classes that correspond to the predefined type of lung disease, which correspond to three outputs neurons used in the output vector; first, second and third (i.e. normal, lung TB and lung cancer).

3.4.5. Training

The other major components of classification in lung disease are learning and testing processes. In the training process, to classify lung disease image, first each lung disease image is taken from different sources and labeled with the type of lung disease. Lung TB, for instance, labeled as lung TB category or class. Then, features are extracted from tagged images by using image analysis as described in the previous section. Features that are used as input classifier extracted and then the system is trained by the classifier from the available information source. Hence, the stored feature acts as a knowledge base, which is used to test the accuracy of the classifier.

3.4.6. Testing

The feature extraction of the testing process is done in the same way as the feature extraction of the training process. The extracted features are used as input to the pattern classifier. The pattern classifier uses the knowledge obtained in the training process to test the classification accuracy of the system.

3.5. Classifier

A pattern classifier is software that is used to train, test and analyze a problem based on the training and testing model of the classification algorithm. In this study we select three classifiers to analyze their generalization capability; Naïve Bayes, Artificial Neural Network (ANN) and K-Nearest Neighbor (KNN).

3.5.1. Naive Bayes Classifier

Naive Bayes classifier is a family of simple probabilistic classifiers based on applying Bayes' theorem with strong independence assumptions between features. Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. This assumption of class independence allows the Naive Bayes classifier to better estimate the parameters required for accurate classification while using less training data than many other classifiers.

For some types of probability models, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without accepting Bayesian probability or using any Bayesian methods.

Despite their naive design and apparently oversimplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations. An advantage of naive Bayes is that it only requires a small number of training data to estimate the parameters necessary for classification. Figure 3.4 shows the decision tree produced by the classifier for the lung disease.

Figure 3-4: Decision tree of Naive Bayes Classifier

3.5.2. Artificial Neural Network (ANN) Classifier

In this classifier the purpose is to select appropriate topology and learning algorithm that best fits to the proposed classification models. Its basic components are learning Paradigm, architecture, learning algorithm, and activation function:

Learning Paradigm: It is a model of the environment in which the neural network operates to learn or train the system. Based on the type of learning process, the networks can be supervised or unsupervised. In supervised learning, the desired output is available for all of the samples needed to be trained. In unsupervised learning, the system must

determine the classes structure mainly the optimal numbers of classes and their properties.

Architecture: It is the topology of the network that describes the pattern of connection between neurons. In this case, we use feed forward multi layer prceptron (MLP) model. A feed-forward multilayer ANN is a topology in which neurons are arranged in layer. The first layer gets the input data from the environment and the last layer generates output. Layers other than input and output layers are called hidden layers. The i^{th} layer gets input from each of the neurons in the immediate preceding layer ($i-1$) weighted by some weight factor for $1 < i \leq N$. A feed forward multilayer perceptron with 2 hidden layers is shown in Figure 3.2. In this case $N = 4$ and is called 3 layer perceptron.

Figure 3-5: Multilayer Perceptron Model with 2 hidden Layers (Source: own)

Learning algorithm: Learning in ANN indicates the methods used to determine the adaptation of weights between the connections of two neurons. MLP feed-forward network uses back propagation learning algorithm for adapting its weights to a sequence of training samples during a learning phase. A back propagation algorithm is a supervised learning algorithm that propagates classification errors from the output layers back toward to the input layers and modify the weight to minimize the total error.

It is the processing logic that computes the neuron€s final output state. Back propagation requires continuous and differentiable activation function. Hence, sigmoid function was used in order to provide smooth control of the input and output relationship.

3.5.3. K-Nearest Neighborhood Classifier (KNN)

The k-NN is a non-parametric method used for classification and regression. In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples. In the classification phase, k is a user-defined constant, and an unlabeled vector (a query or test point) is classified by assigning the label which is most frequent among the k training samples nearest to that query point.

A commonly used distance metric for continuous variables is Euclidean distance. For discrete variables, such as for text classification, another metric such as the overlap metric can be used. Often, the classification accuracy of k-NN can be improved significantly if the distance metric is learned with specialized algorithms such as Large Margin Nearest Neighbor or Neighbourhood components analysis. A drawback of the basic "majority voting" classification occurs when the class distribution is skewed. That is, examples of a more frequent class tend to dominate the prediction of the new example, because they tend to

be common among the k nearest neighbors due to their large number. One way to overcome this problem is to weight the classification, taking into account the distance from the test point to each of its k nearest neighbors [47].

After image acquisition, to remove noises and artifacts from the original image, median filtering used as a second step. Once the image is enhanced, to separate region of interest from the back ground image, image segmentation is carried out using k-means clustering as a third step. Finally, feature extraction of morphology and GLCM features

3.6. Performance Measurements

Classification performance is measured by the powers it can correctly predict true negatives (Specificity), true positives (Sensitivity) and its accuracy. Accordingly, to select the best classifier among classifiers considered in our model, these measurements are used and results are calculated using the following formulas,

$$\text{Sensitivity}(\%) = \frac{}{} \times 100\%,$$

$$\text{Specificity}(\%) = \frac{}{+} \times 100\%, \quad \text{eq. ••••••••.} \quad 2$$

$$\text{Accuracy}(\%) = \frac{+}{+ \quad + \quad +} \times 100\%,$$

Where, TP, TF, FP, and FN stand for True Positive, True Negative, False Positive and False Negative respectively [48].

Accordingly, table of experimental results by classifiers and using three scenarios contains the following arrangement in the confusion matrix.

	Normal Lung	Lung TB	Lung Cancer	Row Total
Normal Lung	True Normal Lung (TN)	Normal Lung, but classified as TB (FNTB)	Normal Lung, but classified as Cancer (FNLC)	TN+FNTB+FNLC
Lung TB	TB, but classified as Normal Lung (FPNL)	True Lung TB (TPTB)	TB, but classified as Cancer (FPLC)	FPNL+TPTB+FPLC
Lung Cancer	Cancer, but classified as Normal Lung (FPNL)	Cancer, but classified as TB (FPTB)	True Lung Cancer (TPLC)	FPNL+FPTB+TPLC
Column Total	TN+FPNL	TPTB+FPTB	TPLC+FPLC	Total Test Data
Sensitivity/ Specificity	(————————)x 1 0 0	(————————)x100	(————————)x100	

Table 3.2: lung disease classification Confusion matrix

Where, TN, TPTB, TPTC, FPNL, FNTB and FNLC are; True Negative, True Positive for Lung TB, True Positive For Lung Cancer, False Positive for Normal Lung, False Negative For Lung TB and False Negative for Lung Cancer respectively.

4. RESULTS AND DISCUSSION

In image processing, factors such as blurriness, unnatural colors, noise, and artifacts often affect the quality of the acquired image, which is not adequate for further processing. Therefore, the preprocessing of the image is done by median filtering. The preprocessed images are given as input to the segmentation process and then the ROIs are separated from the background image using K-means clustering. The output obtained in this process also used as an input to feature extraction process. The purpose of feature extraction is to reduce original data set by measuring certain features that distinguish one region of interest from another. Then, four morphology and five GLCM (texture) features extracted from the segmented image, and used as input to classifiers to eventually evaluate accuracy performance.

Accordingly, the evaluation of the testing instances using Naïve Bayes, ANN and KNN classifiers described in detail as follows;

4.1. Experimental Results

In the previous section, morphological and GLCM features are described. Nine features (four morphology and five GLCM features) identified. These features used to classify the three lung images, normal lung and the two lung diseases (TB and lung cancer).

Therefore, experimental scenarios designed to test the classification performance by taking the extracted features of lung disease images. In order to get a more accurate result, collected data are classified and tested using three different algorithms, namely,

ANN, KNN and Naive Bayes. A total of 300 lung images collected from three hospitals and the internet on a proportional basis from each of the three lung images: normal lung, and the predefined lung diseases (Lung TB and Lung Cancer) for which are 100 sample images from each.

The two basic phases of pattern classification, which are (i) the training phase, to obtain the desired result that data is repeatedly presented to the classifier; (ii) Testing phase, the trained system applied to data that it has never seen to check the performance of the classification. From the total data set of each lung images type (both diseased and normal) 70%, which is 210 used for training and 30% or 90 image data used for testing. In general, the classifier is given input features based on the scenario of the designed experiment and hence produce output features. In this study, as there are three predefined lung images, the output classes are also three. The data are normalized with mean 0 and variance 1.

4.1.1. Naïve Bayes Classifier

Naïve Bayes classifier is a statistical classifier, which is probability distribution. It classifies an object to the class to which it is most likely to belong to the observed features. Three experimental scenarios designed to test the classification performance based on the extracted set of features, morphological, GLCM (texture) and a combination of the two features. From the available 300 lung images (100 from each of the three types) 70% (70 images from each) used to build training, and the remaining 30% (30 images from each) used for testing purpose.

4.1.1.1. Morphology Features

In this experimentation, four morphological features: Area, EquivD, Perimeter, and Roundness are used from the data as input to the classifier. Output classes are three that correspond to the three predefined lung image types. The classification result of Naïve Bayes classifier using the selected morphology feature is shown in Table 4-1.

Table 4-1: Naïve Bayes Classifier Using Morphology Features

	Normal Lung	Lung TB	Lung Cancer	Total
Normal Lung	16	7	7	30
Lung TB	4	15	11	30
Lung Cancer	5	10	15	30
Total	25	32	33	90
Sensitivity/ Specificity	53.33%	50.00%	50.00%	

Classified	Misclassified
46	44
51.11%	48.89%

When the trained system is tested for accuracy by Naïve Bayes classifier using Morphology feature, of the test instances from Normal Lung, Lung TB and Lung cancer, 16, 15, and 15 respectively (46 in total) are correctly classified while the remaining 14, 15, and 15 for Normal lung, Lung TB and Lung Cancer respectively (44 in total) are misclassified.

Table 4-1, the confusion matrix, shows the number of test examples whose actual class is the row heading and whose predicted class is the column heading. Figures in the diagonal

from the left of the table are instances that are correctly classified where as the remaining are misclassified.

In accordance, the overall classification of Naïve Bayes classifier on the selected morphological feature showed that from the total 90 instances, 46 (51.11%) are correctly classified and 44 (48.89%) are misclassified.

Analysis of the Result

The result of Naïve Bayesian classification using morphology feature showed that the classification accuracy of Normal lung, Lung TB and Lung Cancer is 16 (53.33%), 15 (50.00%) and 15 (50.00%) respectively. Lung TB 15 (50%) and Lung Cancer 15 (50.0%) are more misclassified with equal percentage than the normal lung 14 (46.66%). This shows that there is a strong morphology relationship between the two lung diseases. Accordingly, the analysis of Naïve Bayes classification using morphology feature shows the strong correlation of lung TB and Lung Cancer. This is indicated as a bar chart in Figure 4-1 below.

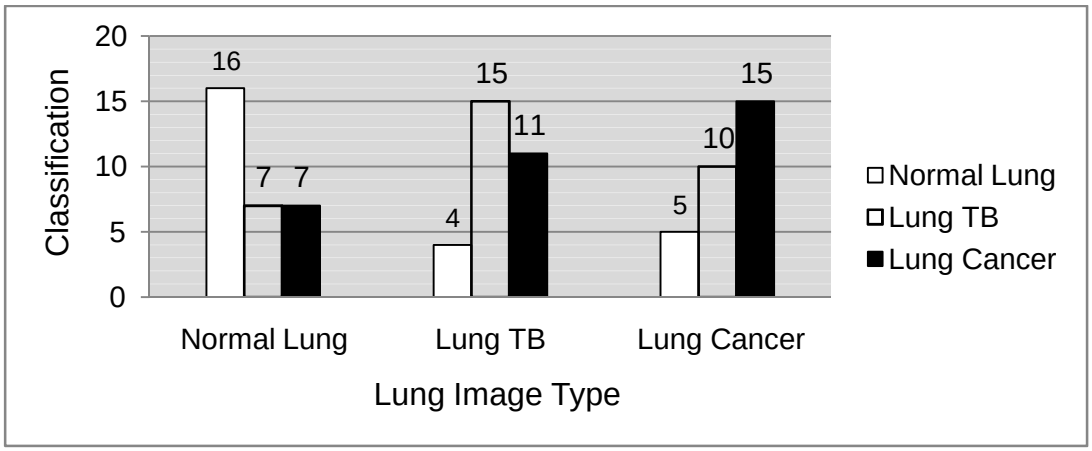

Figure 4-1: Naïve Bayes Classification using morphology Features

4.1.1.2. Texture Features

The GLCM is a tool for image feature extraction by mapping the grey level co-occurrence probabilities based on spatial relations of pixels in different angular directions. The GLCM based features extracted from the three image types are: contrast, energy, homogeneity, correlation, and entropy.

Table 4-2: Naïve Bayes Classifier Using Texture Features

	Normal Lung	Lung TB	Lung Cancer	Total
Normal Lung	20	4	6	30
Lung TB	2	18	10	30
Lung Cancer	4	9	17	30
Total	26	31	33	90
Sensitivity/ Specificity	66.67%	60.00%	56.67%	

Classified	Misclassified
55	35
61.11%	38.89%

As it is shown in Table 4-2, Naïve Bayes classifier of the texture features, 55 (61.11%) are correctly classified and 35 (38.89%) are misclassified.

Analysis of the Result

The result of Naïve Bayesian classification using texture feature showed that the accuracy of Normal lung, Lung TB and Lung Cancer are 20 (66.67%), 18 (60%) and 16 (56.67%) respectively.

Lung Cancer images, 13 (43.33%) are more misclassified than lung TB images, 12 (40%) and lung cancer images are more misclassified than normal lung images, 10 (33.33%). Besides, there is a significant misclassification among each type of lung diseases and normal lung. Normal lung to Lung TB 4 (13.33%), Normal Lung to Lung Cancer 6 (20%), and Lung TB to Lung Cancer 10 (33.33%). This means that there is sharing of texture among the two lung diseases, but compared to morphological features it is less strong. The chart in Figure 4-2 depicted such a behavior of lung diseases.

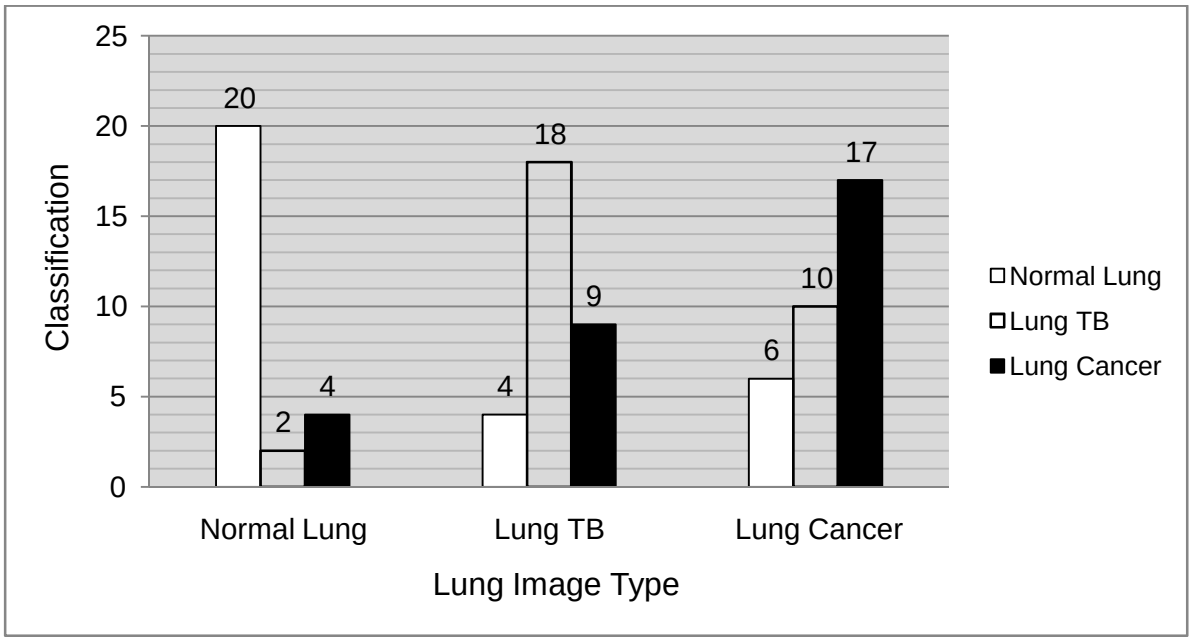

Figure 4-2: Bayes Classification using texture Feature

4.1.1.3. Combined Image Features

In this scenario, the classification input features are nine corresponding to the four morphological features and five texture features. There are also three output classes. The classification result is shown in Table 4-3.

Table 4-3: Naïve Bayes Classifier Using Morphology and Texture Features

	Normal Lung	Lung TB	Lung Cancer	Total
Normal Lung	21	5	4	30
Lung TB	1	19	10	30
Lung Cancer	4	8	18	30
Total	26	32	32	90
Sensitivity/ Specificity	70.00%	63.33%	60.00%	

Classified	Misclassified
58	32
64.44%	35.56%

As it is shown in Table 4-3, Naïve Bayes classifier using morphology and texture features showed, 58 (64.44%) are accurately classified while the remaining 32 (35.56%) are misclassified.

Analysis of the Result

The result of Naïve Bayesian classification using morphology and texture features showed that Normal lung, Lung TB and Lung Cancer are 21 (70%), 19 (63.33%) and 18 (60%) respectively. In this case,

In this case, Lung Cancer images, 12 (43.33%), are more misclassified than Lung TB images, 11 (36.67%), and lung cancer images are more misclassified than normal lung images, 9 (30%). Besides, there is a significant misclassification among each, Lung TB to Lung Cancer 10 (33.33%) and lung Cancer to lung TB (8 26.66%) As shown in the figure 4-3, the combined features accuracy is much better than individual.

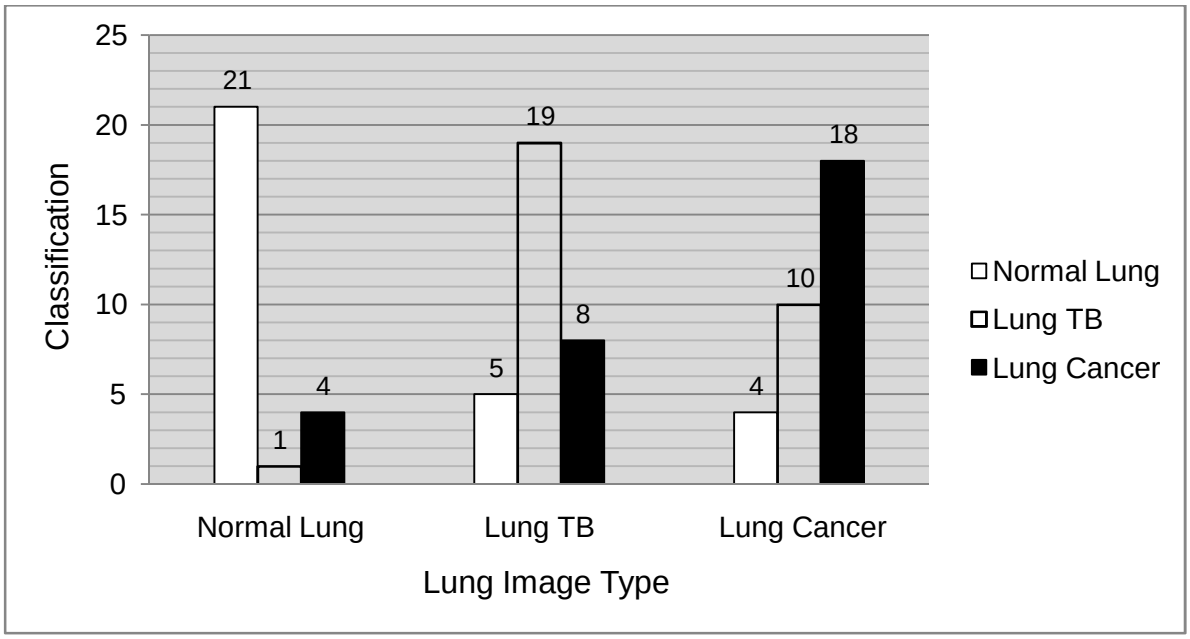

Figure 4-3: Bayes Classification Using Combined Features

4.1.2. Neural Network Classifier

A Neural Network is an adaptable system that can learn relationships through repeated presentation of data, and it is capable of generalizing to new or previously unseen data [34]. As described above (section 3.4.2), Feed forward multilayer perceptron (MLP) model used with back propagation learning rule, which is based on supervised learning. A three layer (2 hidden and 1 output layers) network is used, and the activation function of the hidden layers is tan hyperbolic function. The network is trained to output 1, 2, and 3 accordingly of the output vector as described in figure 3.5. When the network is trained, the neuron number of the input layer depends on the selected features as indicated in the experimentation scenarios. The neuron numbers of hidden layers are eight for both first hidden and second hidden layer. Trial and error approach is used to find a suitable number of the hidden layer that provided good classification of the data

input to the neural network. The neuron number of the output layer is three based on the number of predefined type of lung disease image data.

During training, the connection weights of the neural network are initialized with some random values. The training samples in the training set are input to the neural network classifier in random order and the connection weights adjusted according to the error backpropagation learning rule.

Similar to the Naïve Bayes classification, the experiment is conducted under three scenarios, morphology, texture and combined features.

4.1.2.1. Morphology Features

As mentioned in section 4.1.2, the numbers of neurons in the hidden layers are eight for the first and the second hidden layer. Table 4-4 shows the classification output as classified and misclassified of the testing image data.

Table 4-4: Artificial Neural Network Classifier Using Morphology Features

	Normal Lung	Lung TB	Lung Cancer	Total
Normal Lung	19	6	5	30
Lung TB	5	17	8	30
Lung Cancer	4	10	16	30
Total	28	33	29	90
Sensitivity/ Specificity	63.33%	56.67%	53.33%	

Classified	Misclassified
52	38
57.78%	42.22%

As it is shown in Table 4-4, ANN classifier using morphology feature, 52 (57.78%) are correctly classified and 38 (42.22%) are misclassified.

Analysis of the Result

The result of Artificial Neural Network (ANN) classification using morphology feature shows that the classification accuracies of Normal Lung, Lung TB and Lung Cancer are 19 (63.33%), 17 (56.67%) and 16 (53.33%) respectively. Normal lung is more misclassified to Lung TB 6 (20%), Lung TB more misclassified to Lung Cancer 8 (26.67%) and Lung Cancer is more misclassified to Lung TB 10 (33.33%). It means that there is a morphological relationship between Lung TB and Lung Cancer. In general, the morphological classification pattern of Naïve Bayes and Artificial Neural Network classifiers are similar, however, the performance accuracy is increased in Artificial Neural Network case (the correct classification for ANN was 57.78% while it is 51.11% for Naïve Bayes). The analysis of ANN classification using morphology feature that shows the correlation of each type lung diseases indicated in Figure 4-4.

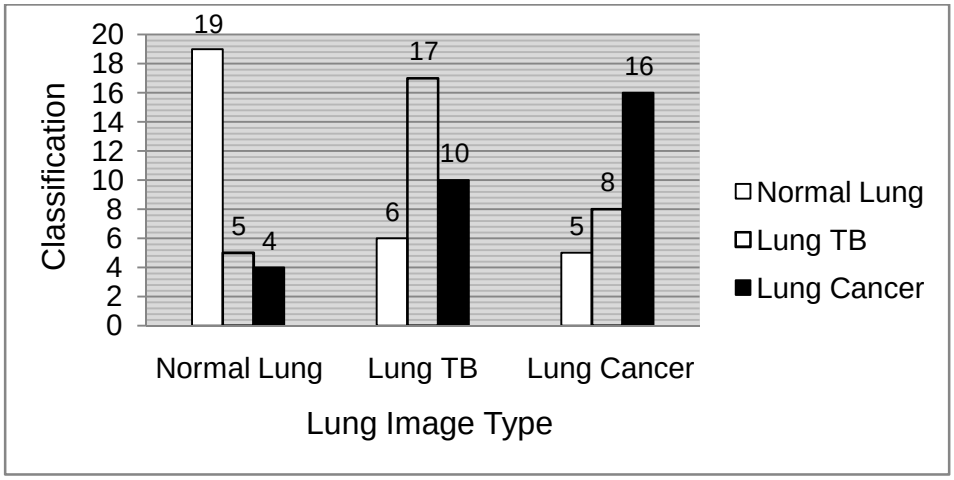

Figure 4-4: ANN Classification Using Morphology Feature

4.1.2.2. Texture Features

As it is indicated in Table 4-5, the result of ANN classifier based on the texture features 60 (66.67%) are accurately classified and 30 (33.33%) instances are misclassified.

Table 4-5: ANN Classifier Using Texture Features

	Normal Lung	Lung TB	Lung Cancer	Total
Normal Lung	21	5	4	30
Lung TB	1	19	10	30
Lung Cancer	2	8	20	30
Total	24	32	34	90
Sensitivity/ Specificity	70.00%	63.33%	66.67%	

Classified	Misclassified
60	30
66.67%	33.33%

Analysis of the Result

The classification accuracy of ANN classifier using texture feature of Normal Lung, Lung TB and Lung Cancer is 21 (70%), 19 (63.33%), and 20(66.67%) respectively. Lung TB is more misclassified to Lung Cancer, 10 (33.33%) and lung Cancer is more misclassified to Lung TB, 8 (26.67%). This indicates that there is a significant misclassification among lung diseases showing that there is a sharing of texture from one lung disease type to another. The analysis of ANN classification using texture feature shows the correlation of the three types of lung images as indicated in Figure 4-5.

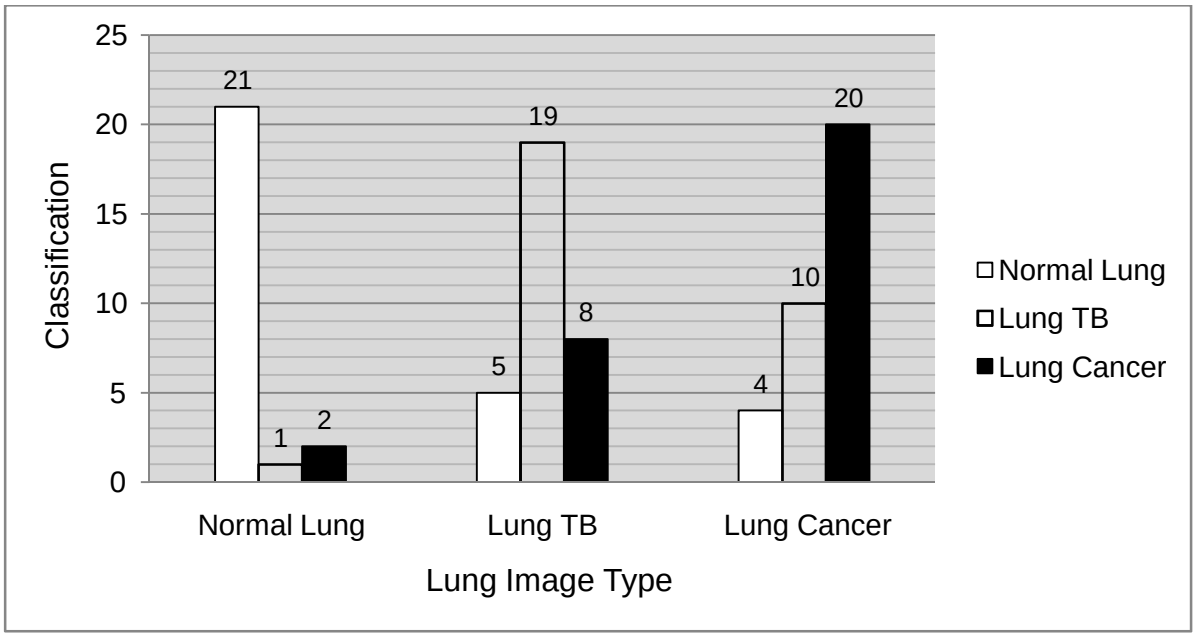

Figure 4-5: ANN Classification Using Texture Feature

4.1.2.3. Combined Features

Using combined features (morphology and texture), the classification input features are nine corresponding to the four morphological features and five texture features. Output classes are also three. The classification result is shown in Table 4-6.

Table 4-6: ANN Classifier using Combined Features

	Normal Lung	Lung TB	Lung Cancer	Total
Normal Lung	23	3	4	30
Lung TB	2	22	6	30
Lung Cancer	1	9	20	30
Total	26	34	30	90
Sensitivity/ Specificity	76.67%	73.33%	66.67%	

Classified	Misclassified
65	25
72.22%	27.78%

As it is indicated in Table 4.6, the result of ANN classifier using morphology and texture features, 65 (72.22%) are correctly classified and 25 (27.78%) are misclassified.

Analysis of the Result

The classification accuracy of ANN classifier using combined features of Normal Lung, Lung TB and Lung Cancer is 23 (76.67%), 22 (73.33%), and 20 (66.67%) respectively. Lung TB is more misclassified to Lung Cancer, 6 (20%) and lung Cancer is more misclassified to Lung TB, 9 (30%). Compared to the combined features in Naïve Bayes, ANN has lesser misclassification among each type lung disease. ANN classification using combined features is shown in Figure 4.6.

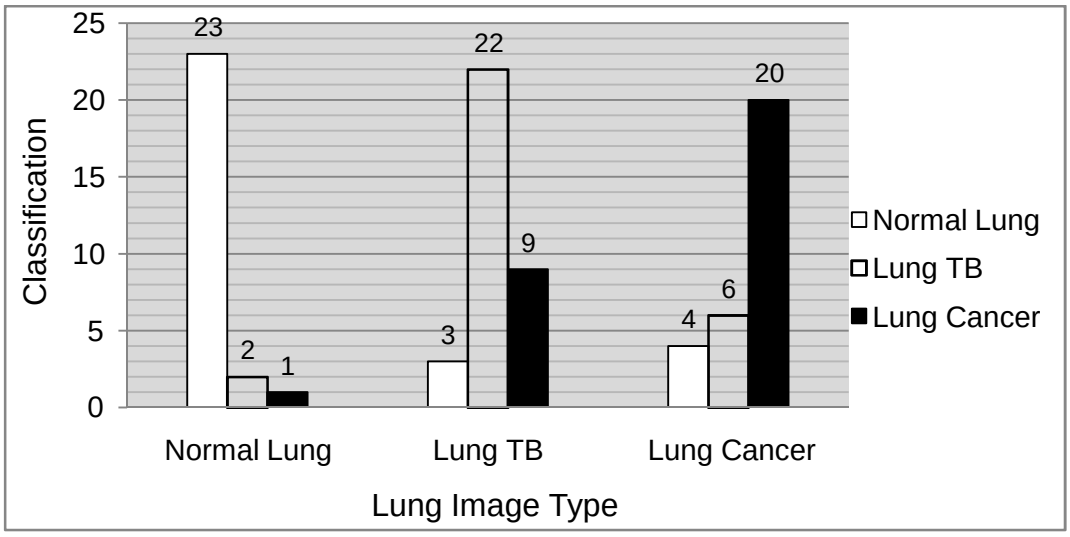

Figure 4.6: ANN Classification of Combined Features

4.1.3. KNN Classifier

In KNN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor. This is a type of instance-based learning, where the function is only approximated locally and all computation is deferred until classification.

The KNN algorithm is among the simplest of all machine learning algorithms. The training phase of the algorithm consists only of storing feature vectors and class labels of the training samples. In the classification phase, k is a user-defined constant, and an unlabeled vector (a query or test point) is classified by assigning the label which is most frequent among the k training samples nearest to that query point. A commonly used distance metric for continuous variables is Euclidean distance.

The comparison of the test data which contain only the input variables to that reference set the distance of the unknown to K nearest neighbors determines its class assignment by either averaging the class numbers of the K nearest reference points or by obtaining a majority vote for them. The k-nearest neighbor classifier for the lung disease data, where k = 9, as the input feature vectors are 9

Figure 4-7: KNN Classification model

4.1.3.1. Morphology Features

As in the previous cases, the four morphological features together are used as input and three output classes that correspond to the three predefined lung image types (Table 4.7).

Table 4-7: The KNN Classifier Using Morphology Features

	Normal Lung	Lung TB	Lung Cancer	Total
Normal Lung	24	4	2	30
Lung TB	2	22	6	30
Lung Cancer	2	7	21	30
Total	28	33	29	90
Sensitivity/ Specificity	80.00%	73.33%	70.00%	

Classified	Misclassified
67	23
74.44%	25.56%

As it is indicated in Table 4-7, the result of KNN classifier using morphology features 67 (74.44%) are correctly classified and 23 (25.56%) are misclassified.

Analysis of the Result

The KNN classification using morphology feature revealed that the classification accuracies of Normal Lung, Lung TB and Lung Cancer are 24 (80%), 22 (73.33), and 21 (70%) respectively. In this case, similar to other classifiers using morphology feature, Lung TB is more misclassified to lung Cancer 6 (20%) and Lung Cancer is more misclassified to lung TB 7 (23.33%) of which the result is significant in both cases. Hence, morphologically lung TB and lung Cancer images are closely related. As per the result, compared to KNN, the morphological classification patterns of Naïve Bayes and Artificial Neural Network have lower performance. The correlation of Normal Lung, lung TB and lung Cancer among each other using morphology feature by KNN showed in figure 4-8.

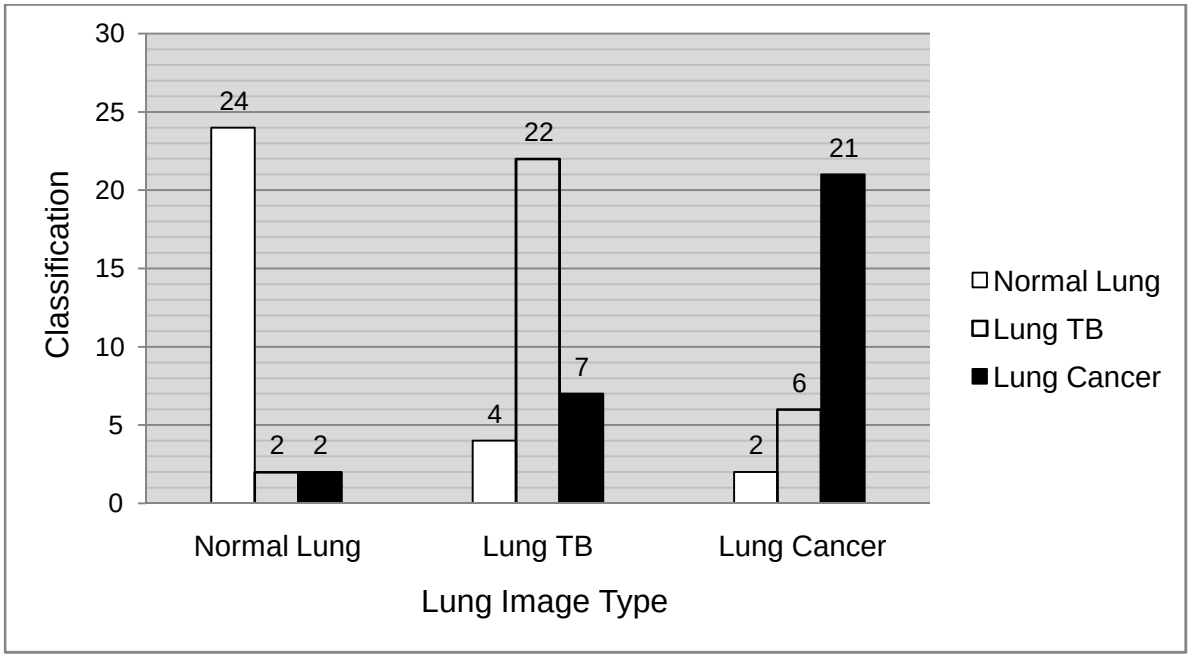

Figure 4-8: KNN Classification Using Morphology Feature

4.1.3.2. Texture Features

As indicated in Table 4-8, the KNN classifier based on the texture features, 76 (84.44%) are correctly classified and 14 (15.56%) are misclassified.

Table 4-8: KNN Classifier Using texture Features

	Normal Lung	Lung TB	Lung Cancer	Total
Normal Lung	26	2	2	30
Lung TB	1	26	3	30
Lung Cancer	1	5	24	30
Total	28	33	29	90
Sensitivity/ Specificity	86.67%	86.67%	80.00%	

Classified	Misclassified
76	14
84.44%	15.56%

Analysis of the Result

The result of KNN classification using texture feature showed that the classification accuracies of Normal Lung, Lung TB and Lung Cancer are 26 (86.67%), 26 (86.67%), 24 (80%) respectively. Normal Lung equally misclassified to both Lung TB and Lung Cancer 2 (3.33%), Lung TB is more misclassified to Lung Cancer 3 (10%) and Lung Cancer is more misclassified to Lung TB 5 (16.67%). Accordingly, there is a significant misclassification among Lung TB and Lung Cancer while it is not for the normal lung. This shows that there is a sharing of texture from lung TB to Lung cancer and the reverse.

The analysis of KNN classification using texture feature that shows the correlation of the three types of lung images is depicted in Figure 4-9.

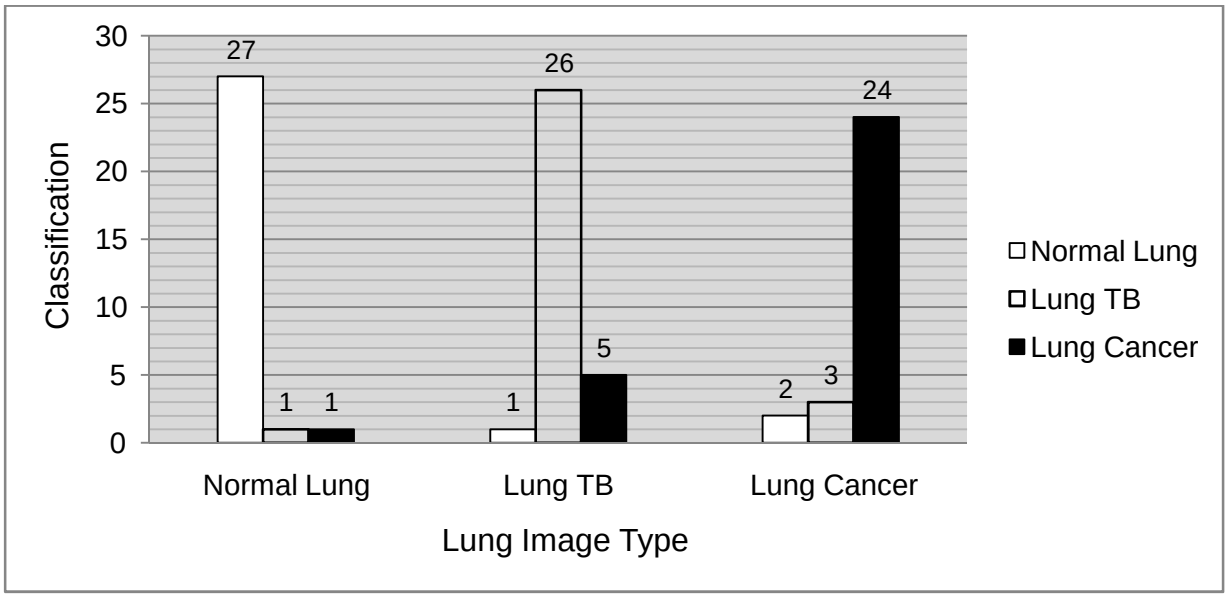

Figure 4-9: KNN Classification Using Texture Feature

4.1.3.3. Combined Features

Using combined features, the classification inputs are nine, corresponding to the four morphological and five texture features. There are also three output classes. The classification result is shown in Table 4-9.

Table 4-9: KNN Classifier using combined features

	Normal Lung	Lung TB	Lung Cancer	Total
Normal Lung	27	1	1	30
Lung TB	1	26	3	30
Lung Cancer	1	4	25	30
Total	29	32	29	90
Sensitivity/ Specificity	90.0%	86.67%	83.33%	

Classified	Misclassified
78	12
86.67%	13.33%

As it is indicated in Table 4-10, the result of KNN classifier using both morphology and features of the test data, 78 (86.67%) are accurately classified and 12 (13.33) are incorrectly classified.

The result of KNN classification using combined features showed that the classification accuracy of Normal Lung, Lung TB and Lung Cancer are 27 (90%), 26 (86.67%), 25 (83.33%) respectively. Normal Lung is more misclassified to Lung TB 2 (6.67%), Lung TB is more misclassified to Lung Cancer 4 (10%) and Lung Cancer is more misclassified to Lung TB 4 (13.33%). The analysis of KNN classification using combined features that shows the correlation of the three types of lung images is depicted in Figure 4-10.

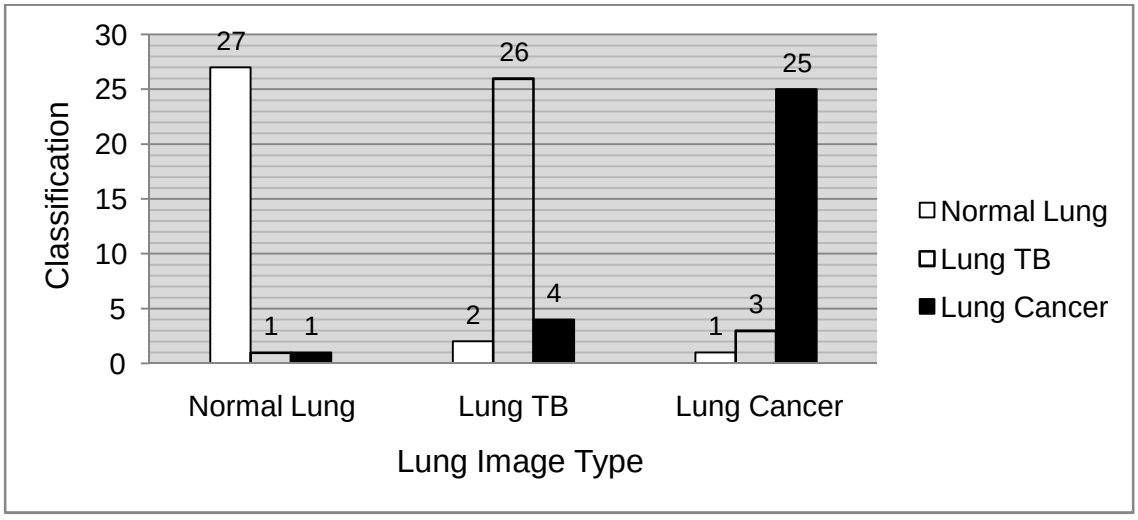

Figure 4-10: KNN Classification Using Combined Feature

4.1.4. Performance Comparison of Classifiers

From all the previous tables (confusion matrix), accuracy and sensitivity results summarized in table-4.10, and these measurements judge the most effective one among classifiers considered in the study

Table 4-10: Accuracy and Sensitivity by Classifier

Evaluation / Test / Classifier	Accuracy	Sensitivity		Specificity
		Lung TB	Lung Cancer	
Naïve Bayes	74.44%	63.33%	60.00%	70.00%
ANN	84.44%	73.33%	66.67%	76.67%
KNN	86.67%	86.67%	83.33%	90.00%

As shown in the Table, KNN gives the best classification accuracy of 86.67%. Its sensitivity is 86.67% for lung TB and 83.33% for Lung Cancer while its specificity is 90%. On the other hand, Naïve Bayes is the least classifier achieving only 74.44% accuracy. Its sensitivity is 63.33% and 60.00% for Lung TB and Lung Cancer respectively with a specificity of 70%.

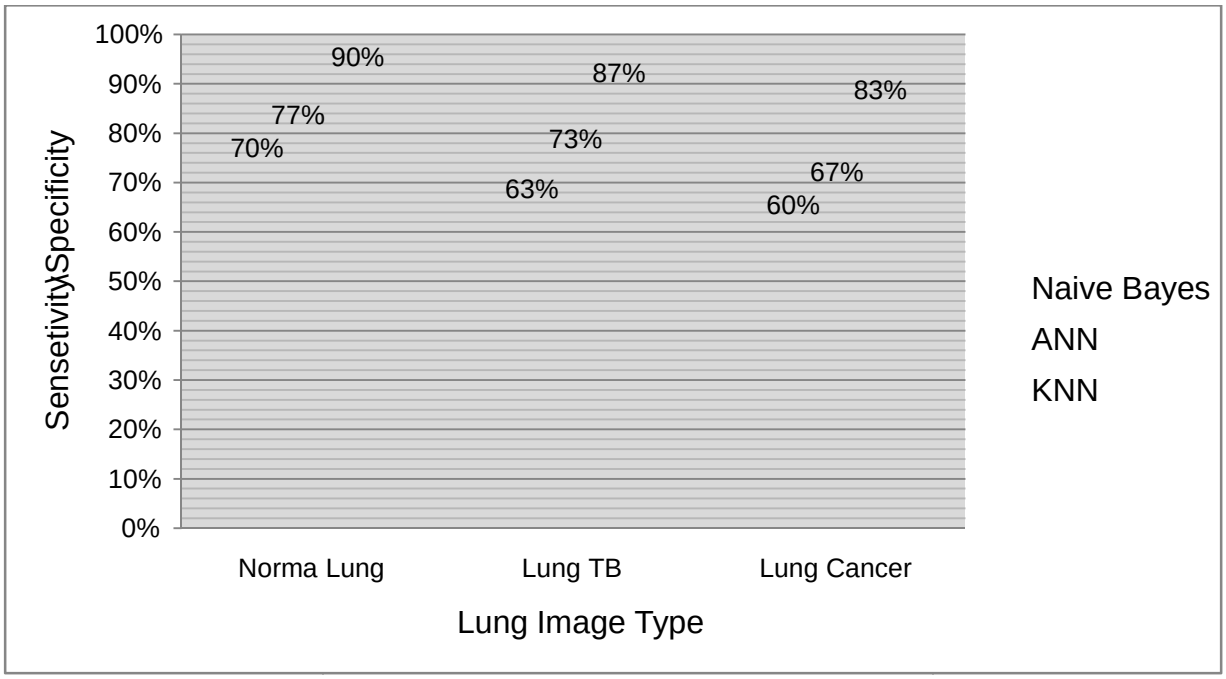

Figure 4-11: Sensitivity by Lung disease

As can be seen in Figure 4-11, sensitivity/specificity results are much higher for KNN in all the lung image types. When we look at the individual result by lung image type, KNN identify best True Negatives and it is more sensitive to Lung TB than Lung Cancer.

To summarize, the experiment conducted under three scenarios, first, by using feature sets of morphology then by using texture feature and, finally, by combining the two together. After an independent experimentation of these scenarios on each of the classifiers, the experimental results are then compared to each classifier using different measurements. Accordingly, the final result revealed that texture have more discriminating power than morphology features, and the classification performance of KNN is the best followed by Artificial Neural Network.

5. CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

TB remains one of the leading causes of mortality in Ethiopia. According to the 2014 WHO report, the prevalence and incidence of all forms of TB are 211 and 224 per 100,000 of the population, respectively. Moreover, Ethiopia is one of the high multidrug resistant TB (MDR TB) burden countries. According to the recent national TB drug resistance surveillance report, 2.3% of new TB cases and 17.9% of previously treated TB cases are estimated to have MDR. Lung cancer is also the most frequent cancer death among men in the world and also a problem in Ethiopia. And the diagnosis of tuberculosis and lung cancer can be difficult as symptoms of both diseases are similar. A missed or wrong diagnosis of the two diseases by clinician can lead to delays in treatment and hence progression of the disease. This means that lung cancer is often misdiagnosed as pulmonary tuberculosis and vice versa, hence causing significant delay in diagnosing both. Therefore, we proposed and analyzed the classification problem of lung disease under different categories, which include, morphological and GLCM features extracted from a lung images taken from Normal lung, Lung TB and Lung Cancer by using image analysis techniques. These selected features used as input to the classification model.

In the study we have found that Lung TB and Lung Cancer have close morphological features, where as the texture feature plays an important role in discriminating the two lung diseases. After partitioning the acquired images reasonably for training and testing, the preprocessing, segmentation and feature extraction are carried out with appropriate

techniques and, finally, the extracted features tested using three classifiers so that appropriate classifier selected using its accuracy and sensitivity result.

Accordingly, after the experiment is conducted under three scenarios of the features data, Morphology, texture, and by combining all together, the result showed that the two lung diseases classified more accurately using KNN than ANN or Naïve Bayes classifier.

5.2 Recommendations

It is true that there are progresses in the healthcare system of Ethiopia in recent years. However, there are also lots of assignments waiting ahead that still require an immediate attention. For instance, cancer is little known and little awareness to TB cases in the community, people die of lung cancer while they are treated for lung TB, and the prevalence of MDR TB is increasing overtime.

The image analysis for the classification of the type of lung disease can be further scrutinized. The work can also be seen in depth and research by the different image processing techniques. In light with this, the following recommendations are made for further research and improvements.

- Identification of more lung diseases like chronic obstructive pulmonary (COPD) disease and pneumonia by incorporating more features
- Lung disease Co-occurrence classification using image analysis

REFERENCES

[1]. C.Bhuvaneswari, P.Aruna, D.Loganathan (2014), Classification of Lung Diseases by Image Processing Techniques Using Computed Tomography Images.

[2].U.S. Department of Health and Human Services (2010), Office on Women€s Health, Content last updated November 29,

[3].Wai Yan Nyein Naing, Zaw Z. Htike (2014), Advances in automatic tuberculosis detection in chest x-ray images,

[4] Nobel Foundation (2016), The Nobel Prize in Physiology or Medicine, 1905 (accessed on April 14).

[5] Fekadu Alemu Atire (2015), Assessment the Prevalence of Pulmonary Tuberculosis Patients at Yirga Cheffe Health Center from 2008-2013, Ethiopia Clinical Medicine Research. Vol. 4, No. 2, pp. 42-48.

[6] Bergmire-S. et al (1996), Tuberculosis outbreak in a Texas prison, 1994; 11: 172-85

[7].Chigbu, L.N. and Iroegbu, C.U. (2010), Incidence and Spread of Mycobacterium tuberculosis associated Infection among Aba Federal Prison Inmates in Nigeria. J Health Popul Nutr; 28: 323-32.

[8].Ilya Levner, Hong Zhangm (2007), Classification driven Watershed segmentation,IEEE TRANSACTIONS ON IMAGEPROCESSING VOL. 16, NO. 5.

[9] Vijay A.Gajdhane, Deshpande L.M. IOSR (2014). Detection of Lung Cancer Stages on CT scan Images by Using Various Image Processing Techniques, Journal of Computer Engineering, Volume 16, Issue 5, PP 25-28

[10] Yang-Hao Yu, et al (2011), Increased Lung Cancer Risk among Patients with Pulmonary Tuberculosis: A Population Cohort Study, Journal of Thoracic Oncology, Volume 6, Issue 1, Pages 32-37

[11] B Selvaraj, K Senthil Kumaran, G P Sekar (2015), International Journal of Recent Trends in Science And Technology, , Volume 14, Issue 2,

[12] A. Konstantinos (2010), Testing for tuberculosis, Aust. Prescriber 33, 12-18

[13] Mohammad Rahbar, Masood Hajia (2007), Value of gastric lavage for diagnosis of pulmonary tuberculosis, Pak. J. Med. Sci. 23,513.

[14] E. Lee, R. Holzman (2002), Evolution and current use of the Tuberculin test, Clin. Infect. Dis. 34 365-370.

[15]. Patil S.A. and M. B. Kuchanur (2012), Lung Cancer Classification Using Image Processing,.

[16] Sheenum M, Himanshu M. and Shelza (2012), Automatic Diagnosis Systems Using Image Processing-A systematic Study, International Journal of Computer Science and Information Technology & Security Vol. 2, No.2.

[17]. G.Anil Kumar, NistalaV.E.S.Murthy (2014), Analysis of Medical Image Processing and its Applications in Healthcare Industry,

[18] http://www.who.int/mediacentre/factsheets/fs104/en/, Tuberculosis Fact sheet, No.104, accessed on April 15, 2016

[19] Santosh S. , Yogesh S. and Ritu V.(2016), An Evaluation of Features Extraction from Lung CT Images for the Classification Stage of Malignancy, Journal of Computer Engineering, Special Issue, pp 37-38

[20] Amjed S. Al-Fahoum, Eslam B. Jaber, Mohammed A. Al-Jarrah (2014), Automated detection of lung cancer using statistical and morphological image processing techniques Journal of Biomedical Graphics and Computing, Vol. 4, No. 2.

[21] Khin Mya Mya Tun and Aung Soe Khaing (2014), Feature Extraction and Classification of Lung Cancer Nodule using Image Processing Techniques, International Journal of Engineering Research & Technology, Vol. 3 Issue 3.

[22] Manisha RK, Palanisamy KS.(2016), Computer-aided diagnosis of tuberculosis using chest radiographs, pp 101-2019

[23] Hrudya Das and Ajay Nath (2015), An Efficient Detection of Tuberculosis from Chest X-rays, International Journal of Advance Research in Computer Science and Management Studies, Volume 3, Issue 5.

[24] Carolyn Vachani (2016), All about Small Cell Lung Cancer, http://www.oncolink.org/, accessed on April 10,

[25] http://www.webmd.com/lung/lung-diseases-overview, accessed on April 15, 2016

[26] http://www.english.cancercare.no, Lung Cancer in Ethiopia, accessed on April 4, 2016

[27] Volker Winkler and et al (2011), Lung Cancer, Predicting lung cancer deaths from smoking prevalence data, pp 1707-1717

[28] http://www.mayoclinic.org/ Lung Cancer Tests and Diagnosis, accessed on April 15, 2016

[29] J Bras Pneumol (2013), Pulmonary tuberculosis and lung cancer: simultaneous and sequential occurrence,.;39(4):484-489

[30] B Selvaraj1, K Senthil Kumaran, Kumaran (2015), Lung cancer misdiagnosed as sputum negative TB can we avoid this pitfall with the investigations available today International Journal of Recent Trends in Science And Technology, Volume 14, Issue 2

[31] Tuberculosis mimicking lung cancer, Respiratory Medicine Case Reports (2015), Volume 16, Pages 45-47

[32] MLB Bhatt, Surya Kant, and Ravi Bhaskar (2012), Pulmonary tuberculosis as differential diagnosis of lung cancer, South Asian J Cancer, 1(1): 36-42.

[33] Sagar N. Vidhate and V. S. Dhongde (2013), Describe texture analysis systems to identify interstitial lung disease (ILD).

[34] Rachna H. B., M. S. Mallikarjuna Swamy(2013), Detection of Tuberculosis Bacilli using Image Processing Techniques International Journal of Soft Computing and Engineering (IJSCE), Volume-3, Issue-4. 2013

[35] Adgaonkar A. et.al(2014), Identification of Tuberculosis bacilli using Image Processing, International Journal of Computer Applications (IJCA) (0975 – 8887), International Conference on Electronics & Computing Technologies,

[36] Adi K et.al (2013), Tuberculosis Identification in the Ziehl Neelsen sputum sample in NTSC channel and Support Vector Machine, International Journal of InnovativeResearch in Science, Engineering and Technology Vol. 2, Issue 9.

[37] P. Sadaphal, J. Rao, G. W. Comstock, M. F. Beg (2008), Image processing techniques for identifying Mycobacterium tuberculosis in Ziehl-Neelsen stains, The International Journal of Tuberculosis and Lung Disease,12(5):579582.

[38] Chandrika V., Parvathi C.S., and P. Bhaskar (2012), Novel Image Enhancement for Pulmonary Tuberculosis Analysis, International Journal of Science and Applied Information Technology, Volume 1, No.4, 102-106.

[39] Gajdhane V. and Deshpande L.M. (2014), Detection of Lung Cancer Nodule on Computed Tomography Images by Using Image Processing International Journal of Application or Innovation in Engineering & Management (IJAIEM) Volume 3, Issue 7.

[40] Varalakshmi.K (2013), Classification of Lung Cancer Nodules using a Hybrid Approach, Journal of Emerging Trends in Computing and Information Sciences, Vol. 4, No. 1.

[41] G. Vijaya, A. Suhasini, R. Priya (2014), Automatic Detection of Cancer in CT Imaging International Journal of Research in Engineering and Technology, Volume: 03 Special Issue: 07

[42] [Eman Magdy et al. (2015)], Automatic Classification of Normal and Cancer Lung CT Images Using Multiscale AM-FM Features, International Journal of Biomedical Imaging, Volume 2015, pp. 1-7

[43] O. R. Vincent and O. Folorunso (2009), A Descriptive Algorithm for Sobel Image Edge Detection, Proceedings of Informing Science & IT Education Conference, 100-104

[44] Mark S. Nixon and Alberto S. Aguado (2008), Feature Extraction and Image Processing, Second edition,

[45] Mokhled S. AL-TARAWNEH (2012), Lung Cancer Detection Using Image Processing Techniques, Leonardo Electronic Journal of Practices and Technologies, Issue 20, p. 147-158

[46] F. R. Renzetti et al. (2011), Use of a gray level co-occurrence matrix, pp. 45-51;

[47] D. Coomans and D.L. Massart (1982), Alternative k-nearest neighbour rules in supervised pattern recognition, Part 1, pp. 27-5

[48] http://www.chestx-ray.com/, accessed on April 2016

[49] https://openi.nlm.nih.gov/gridquery.php?q=tuberculosis%20chest, accessed on April 2016

[50] https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI, accessed on April 2016

APPENDIX

GUI

//

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

bw2  = handles.bw2;
imgvec = edu_imgresize(bw2);
% axes(handles.axes14);
```

```
% plotchar(imgvec);
% handles.imgvec = imgvec;
% guidata(hObject, handles);
% GLCM2 = graycomatrix(bw2);
% allst = graycoprops(GLCM2,'all');
% contrastInfo = allst.Contrast;
% [contrastInfo,PS] = mapstd(contrastInfo);
% display(contrastInfo)
% energyInfo=  allst.Energy;
% % display(energyInfo)
% homogeneityInfo = allst.Homogeneity;
% % display(homogeneityInfo)
% correlationInfo = allst.Correlation;
% % display(correlationInfo)
%total=contrastInfo+energyInfo+correlationInfo;
% display(total);
I2 = im2double(bw2);
% skewnessInfo=skewness(I2(:));
% skewnessinf = abs(skewnessInfo);
% display(skewnessInfo)
C=imfill(bw2,'holes');
[Label,Total]=bwlabel(C,8);
 num=4;
[row, col] = find(Label==num);
 sx=min(col)-0.5;
sy=min(row)-0.5;
breadth=max(col)-min(col)+1;
len=max(row)-min(row)+1;
BBox=[sx sy breadth len];
%display(BBox);
 Obj_area=numel(row);
% display(Obj_area);
%------------------
X=mean(col);
Y=mean(row);
Centroid=[X Y];
% display(Centroid);
 BW=bwboundaries(Label==num);
c=cell2mat(BW(1));
Perimeter=0;
for i=1:size(c,1)-1
Perimeter=Perimeter+sqrt((c(i,1)-c(i+1,1)).^2+(c(i,2)-c(i+1,2)).^2);
end
% display(Perimeter);
 EquivD=sqrt(4*(Obj_area)/pi);
% display(EquivD);
 Roundness=(4*Obj_area*pi)/Perimeter.^2;
 [Roundness,PS] = mapstd(Roundness);
display(Roundness)
display(Roundness);
% data=[Roundness,EquivD,Perimeter,correlationInfo,homogeneityInfo,energyInfo,contrastInfo,Obj_area];
% display(data);
%I=imread('37.jpg');
S = handles.S;
rgb = im2double(S);
im=imresize(S, [360 360]);
```

```
imgGray1 = rgb2gray(S);
% axes(handles.axes17);
% imshow(imgGray1)
GLCM2 = graycomatrix(imgGray1);
allst = graycoprops(GLCM2,'all');
contrastInfo = allst.Contrast;
[contrastInfo,PS] = mapstd(contrastInfo);
display(contrastInfo)
energyInfo =  allst.Energy;
% display(energyInfo)
homogeneityInfo = allst.Homogeneity;
% display(homogeneityInfo)
correlationInfo = allst.Correlation
entropyinfo = entropy(imgGray1)
display(entropyinfo)
r = rgb(:, :, 1);
g = rgb(:, :, 2);
b = rgb(:, :, 3);

% Implement the conversion equations.
num = 0.5*((r- g) + (r - b));
den = sqrt((r- g).^2 + (r- b).*(g - b));
theta = acos(num./(den + eps));

H = theta;
H(b > g) = 2*pi- H(b > g);
H = H/(2*pi);

num = min(min(r, g), b);
den = r + g + b;
den(den == 0) = eps;
S = 1- 3.* num./den;

H(S ==0) = 0;

I = (r + g + b)/3;

% Combine all three results into an hsi image.
%cat used for concatination
hsi = cat(3, H, S, I);
% axes(handles.axes15);
% title('HSI Image');
% imshow(hsi);



redMean = mean2(rgb(:,:,1));
% display(redMean)
greenMean =mean2(rgb(:,:,2));
% display(greenMean)
blueMean = mean2(rgb(:,:,3));
% display(blueMean)
%------------------------------
HueMean = mean2(hsi(:,:,1));
% display(HueMean)
```

```
SatuMean = mean2(hsi(:,:,2));
% display(SatuMean)
InteMean=mean2(hsi(:,:,3));
% display(InteMean)
data1=[redMean,greenMean,blueMean,HueMean,SatuMean,InteMean];
% display(data1)
%*************************************************************************************
******************
Featuredata=[Roundness EquivD Perimeter correlationInfo entropyInfo homogeneityInfo energyInfo
contrastInfo Obj_area redMean greenMean blueMean HueMean SatuMean InteMean];
%*************************************************************************************
**********************
% display(Featuredata)
[lungdataset,PS] = mapstd(Featuredata);
display(lungdataset)
save lungdata.mat lungdataset
set(handles.txt_msg,'String','Features are Extracted');

% display(Featuredata)
//

function mnuLoadImage_Callback(hObject, eventdata, handles)
% hObject    handle to mnuLoadImage (see GCBO)
% eventdata  reserved to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
[filename, pathname] = uigetfile({'*.bmp';'*.jpg';'*.gif';'*.*'}, 'Pick an Image File');
S = imread([pathname,filename]);
axes(handles.axes1);
imshow(S);

handles.S = S;
guidata(hObject, handles);

% --------------------------------------------------------------------
function mnuExit_Callback(hObject, eventdata, handles)
% hObject    handle to mnuExit (see GCBO)
% eventdata  reserved to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
selection = questdlg('Do You Want to QUIT.....',...
    'Close Request Function',...
    'Yes','No','Yes');
  switch selection,
    case 'Yes',
      close all
    case 'No'
    return
  end
```