MACHINE LEARNING APPROACHES FOR THE DETECTION OF LUNG CANCER USING MRI IMAGES

MD. MOKHLESUR RAHMAN (SN. 0419140015)

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Science and Engineering

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

MILITARY INSTITUTE OF SCIENCE AND TECHNOLOGY

DHAKA, BANGLADESH

JAN 2024

MACHINE LEARNING APPROACHES FOR THE DETECTION OF LUNG CANCER USING MRI IMAGES

M.Sc. Engineering Thesis

By

MD. MOKHLESUR RAHMAN (SN. 0419140015)

Approved as to style and content by the Board of Examination on 12 February 2024:

Dr. T. M. Shahriar Sazzad Assistant Professor of Computer Science and Engineering MIST, Dhaka	Chairman (Supervisor) Board of Examination
Dr. Md. Hasanul Kabir Professor of Computer Science and Engineering IUT, Dhaka.	Member (External) Board of Examination
Lt Col Muhammad Nazrul Islam, PhD Associate Professor of Computer Science and Engineering MIST, Dhaka	Member Board of Examination
Brig Gen Md Towhidul Islam, PBGM, BGBMS, ndc, afwc, psc Senior Instructor of Computer Science and Engineering MIST, Dhaka	Head of the Dept Member (Ex-officio)

Department of Computer Science and Engineering, MIST, Dhaka

MACHINE LEARNING APPROACHES FOR THE DETECTION OF LUNG CANCER USING MRI IMAGES

DECLARATION

I therefore declare that this thesis is my unique work and authored entirely by myself. I have properly credited all sources of material used in the thesis. This thesis has never been presented for a degree or diploma at any university or institute before (in whole or in part). All sources and support obtained in preparing this thesis have been acknowledged and/or cited in the reference section.

_____

Md. Mokhlesur Rahman

Department of Computer Science and Engineering, MIST, Dhaka.

ABSTRACT

MACHINE LEARNING APPROACHES FOR THE DETECTION OF LUNG CANCER USING MRI IMAGES

Lung cancer ranks as the second most prevalent form of cancer worldwide, resulting in thousands of deaths annually. Nevertheless, the mortality rate can be mitigated by enhancing early detection and successful treatment, thereby bolstering the survival prospects of patients. There are different types of electronic modalities, e.g., CT/PET Scan, MRI, X-Ray etc. for lung diagnosis. With the advancement of technologies MRI is being used widely for lung cancer detection. But, the interpretation of MRI image is totally expert dependent and time consuming. An automated computerized approach can make lung cancer identification easier and more reliable. This study describes a fully automated technique for lung cancer detection using lung MRI and following two different approaches, i.e., conventional image processing approach and machine learning approach. The proposed conventional image processing method pro-vided an accuracy of 96.28%. However, CNN and SVM were used in machine learning approach and the classification accuracy were 96.55% and 90.5% respectively.

সারসংক্ষেপ

MACHINE LEARNING APPROACHES FOR THE DETECTION OF LUNG CANCER USING MRI IMAGES

পৃথিবীতে যত ধরনের ক্যান্সার আছে তন্মধ্যে ফুসফুসের ক্যান্সার দ্বিতীয় বৃহত্তম ক্যান্সার, যার কারনে প্রতি বছর হাজার হাজার মানুষের মৃত্যু হয়। যদিও, প্রাথমিক অবস্থায় ফুসফুসের ক্যান্সার সনাক্তকরণ এবং সফল চিকিৎসার মাধ্যমে মৃত্যুর হার হ্রাস করা যেতে পারে এবং রোগীদের বেঁচে থাকার সম্ভাবনাকে বাড়ানো সম্ভব। ফুসফুসের রোগ নির্ণয়ের জন্য বিভিন্ন ধরনের ইলেকট্রনিক পদ্ধতি রয়েছে, যেমন, সিটি/পিইটি স্ক্যান, এমআরআই, এক্স-রে ইত্যাদি। প্রযুক্তির অগ্রগতির সাথে সাথে ফুসফুসের ক্যান্সার সনাক্তকরণের জন্য এমআরআই ব্যাপকভাবে ব্যবহৃত হচ্ছে। কিন্তু, ফুসফুসের এমআরআই-এর মাধ্যমে রোগ নির্ণয় সম্পূর্ণরূপে বিশেষজ্ঞ নির্ভর এবং সময়সাপেক্ষ। একটি স্বয়ংক্রিয় কম্পিউটারাইজড পদ্ধতি ফুসফুসের ক্যান্সার সনাক্তকরণকে সহজ এবং আরও নির্ভরযোগ্য করে তুলতে পারে। এই গবেষণাটি ফুসফুসের এমআরআই ব্যবহার করে, স্বয়ংক্রিয়ভাবে ফুসফুসের ক্যান্সার সনাক্তকরণের কৌশল বর্ণনা করে যা, দুটি ভিন্ন পদ্ধতি অনুসরণ করে, যেমন, Conventional Image Processing পদ্ধতি এবং মেশিন লার্নিং পদ্ধতি। প্রথমত, প্রস্তাবিত Conventional Image Processing পদ্ধতি ৯৬.২৮% এর নির্ভুলতা প্রদান করেছে। দ্বিতীয়ত, মেশিন লার্নিং পদ্ধতিতে CNN এবং SVM ব্যবহার করা হয়েছিল এবং শ্রেণীবিভাগের যথার্থতা ছিল যথাক্রমে ৯৬.৫৫% এবং ৯০.৫%।

ACKNOWLEDGEMENTS

Alhamdulliah, all the praises to ALLAH who is the source of all strength. He has enabled me to carry on with and finish this research work. This journey has been an enlightening and enriching experience, and I am humbled and grateful for the support and contributions of numerous individuals who made this endeavor possible.

Above all, I express sincere gratitude to my esteemed supervisor, *Dr. T.M. Shahriar Sazzad, Assiatant Professor, Department of Computer Science and Engineering (CSE), Military Institute of Science and Technology (MIST)*, whose expertise, guidance, and encouragement have been invaluable throughout this research. His stimulating discussions and collaboration have been instrumental in shaping the direction of this thesis. I really want to offer my heartfelt gratitude, and deep admiration for his ongoing oversight, comments, and direction as this thesis progressed.

I am indebted to the faculty members of Department of CSE, whose commitment to education and research has been a constant source of inspiration. Their dedication to nurturing academic curiosity has played a pivotal role in my personal and intellectual growth.

I also wish to recognize *Dr. Md. Sadequel Islam Talukder* for generously supplying us with MRI images of the lung, along with pertinent information on features to aid in the identification of lung nodules.

I am compelled to express my deep appreciation for the steadfast support of my wife and family members, who have served as my unwavering pillars of strength during this arduous journey. Their encouragement, patience, and unwavering belief in my capabilities have consistently fueled my motivation.

Last but not least, I wish to express my deepest appreciation to the countless anonymous scholars whose research laid the foundation for this thesis. I offer my heartfelt thanks to all those who have directly or indirectly played a role in the completion of this thesis. May this work contribute meaningfully to the body of knowledge in the field and inspire further research and exploration.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATION

ML : Machine Learning

IP : Image Processing

CAD : Computer Aided Diagnosis

MRI : Magnetic Resonance Imaging

CT : Computed Tomography

PET : Positron Emission Tomography

US : Ultrasound

LDCT : Low Dose Computed Tomographic

ROI : Region of Interests

CNN : Convolutional Neural Network

SVM : Support Vector Machine

NSCLC : Non-Small Cell Lung Cancer

SCLC : Small Cell Lung Cancer

ACS : American Cancer Society

3D : 3-Dimensional

HSV : Hue, Saturation, Value

RGB : Red, Green, Blue

GLCM : Gray Level Co-occurrence Matrix

SCM : Structural Co-occurrence Matrix

MLP : Multi-Layer Perceptron

KNN : k-Nearest Neighbor

RBF : Radial Base Function

DICOM : Digital Imaging and Communications in Medicine

CHAPTER 1

INTRODUCTION

1.1 Background

Digital Image Processing (DIP) and Machine Learning (ML) are fundamental disciplines in contemporary Computer-Aided Diagnosis (CAD) Systems. The utilization of ML and MRI image analysis for identifying lung cancer has emerged as a promising domain in medical research. Given that lung cancer is among the most prevalent and fatal types of cancer globally, early detection significantly impacts patient outcomes. Traditional methods for lung cancer detection rely on invasive procedures, such as biopsies, which can be uncomfortable for patientsand may carry potential risks. On the other hand, visual inspection of MRI or Computed Tomography (CT) images requires substantial amount of time for analysis and often sufferfrom limitations in accuracy. By harnessing the power of machine learning algorithms andthe detailed imaging provided by MRI scans, researchers aim to develop non-invasive andaccurate methods for early detection and diagnosis. By integrating ML techniques with DIP, intricate features and patterns that may elude human perception can be extracted. This integration facilitates the creation of predictive models, aiding radiologists in early-stage lung cancer identification using images acquired from electronic modalities like MRI. This thesis seeks to investigate the capabilities of ML algorithms in scrutinizing MRI images to enhance lung cancer detection, with the goal of promoting timelier interventions and improving patient outcomes.

1.2 Motivation

Cancer remains a significant contributor to global mortality, with nearly 10 million deaths recorded in 2020. Among these, 18% were attributed to lung cancer, solidifying its position as the foremost fatal cancer worldwide (Ferlay et al. 2020). One in three people is diagnosed with cancer in their lifetime (American Cancer 2021). In 2020, breast cancer emerged as the most prevalent type of cancer with 2.26 million reported cases, closely followed by lung cancer with 2.21 million cases. Truly Lung Cancer is a disease which deserves all out efforts for quick and easy identification and treatment options. Therefore, it is essential to diagnose Lung Cancer at the initial stage for appropriate medication and treatments of the patients. Early detection of Lung Cancer will assist medical expert to provide proper medication which can increase the survivability rate of the patients significantly. Diagnosing lung cancer can be achieved through biopsy or various imaging technologies such as X-ray, Ultrasound (US), Computerized Tomography (CT), and Magnetic Resonance Imaging (MRI). Among these diagnostic tools, MRI is commonly employed for identifying abnormal cell growth. Nonetheless, its effectiveness relies on skilled experts, and subpar image quality can lead to inconsistent results. To address these challenges, leveraging a computerized approach could prove advantageous. IP techniques can improve image quality and subsequently segment the image to identify lung nodules. Basing on various parameters the extent of cancer can be identified. ML algorithms may further be used to automatically classify lung nodules.

1.3 Problem Statement

According to the research study (Tiwari 2016), lung cancer detection using IP techniques primarily focused on CT images. Similarly, a review of related literature (Biederer et al. 2017) corroborates this finding. However, there is a growing trend in MRI-based lung cancer screening, driven by the hypothesis that MRI may offer superior specificity for early detection compared to Low Dose Computed Tomography (LDCT). Besides MRI is also suggested for the patients who require regular checkup of the cancerous nodule for its radiation free characteristics. Hence, for detection of lung cancer using MRI through a

combination of appropriately selected DIP and ML approach is likely to be useful for Lung Cancer identification.

1.4 Thesis Objectives

From a broader viewpoint, this study will encompass Medical IP and ML concerning the diagnosis of lung cancer. The objectives of this research are outlined as follows:

(1) To explore how IP techniques are applied on MRI images to detect lung cancer.

(2) To propose an efficient technique for the detection of lung cancer using MRI through machine learning approaches.

(3) To compare the performance of existing IP techniques with proposed machine learning approach.

(4) As a result, this research is to propose a modified automated Lung Cancer Detection approach using MRI.

1.5 Methodological Overview

The research will be carried out as per the following steps:

(1) A systematic review will be conducted on research centered around the detection of lung cancer utilizing MRI images.

(2) Limitations of current techniques will be brought out from the literature review.

(3) A proficient method will be proposed for detecting lung cancer utilizing MRI images. Following steps will be conducted:

   (a) Image Enhancement and filter operations will be carried out on test data for visual perception and to remove noises.

   (b) Grayscale image segmentation will be carried out as the test images will be

grayscale images. Appropriate lower and upper boundary will be set to check against over and under-segmentation.

(c) For identification and classification two different approaches will be carried out which are as follows:

i. Identification of Lung Cancer Region of Interests (ROIs) using hand crafted features, typically used in the radiology laboratory.

ii. Classification and identification Lung Cancer using suitable Machine Learning Approach.

(d) Appropriate adjustments will be carried out to the above-mentioned steps to obtain an acceptable accuracy.

(e) Based on the outcomes of the tests, a refined version of the approach will be proposed to enhance the accuracy of lung cancer detection using MRI images.

1.6 Thesis Scope

The scope of this thesis is restricted to the utilization of MRI images exclusively. However, it's worth noting that MRI can diagnose a broad spectrum of injuries and diseases, encompassing soft tissues such as the lung, brain, and spinal cord, among others. But, it is not a perfect tool for imaging the organs which changesshapes, like lungs which expand and shrinks. In Bangladesh, till today doctors are not using lung MRI. However, with the latest techniques and advancements in technology, MRI can serve as a suitable alternative to lung CT scans. So, large scale data are not available for the said research. Limited onlineavailable MRI images could be collected the specialist and researchers.

1.7 Thesis Organization

The remaining chapters of this research are organized as follows:

In Chapter 2, the anatomy of the lung, lung cancer, and its various types are briefly introduced. Thereafter various screening techniques are described along with acceptability

of MRI as a screening technique. Staging of lung cancer has also been narrated to know about its growth and severity at different time period.

In Chapter 3, relevant literature on lung cancer detection using IP Techniques and Machine Learning algorithms is discussed.

Chapter 4 offers an overview of the research methodology. It outlines the research method, approach, and activities undertaken to conduct this study.

Chapter 5 discusses conventional IP steps along with the IPsteps followed in this research. Feature extraction phase has been described along with parameters used for feature extraction.

In Chapter 6, machine learning concepts are discussed, including detailed explanations of Convolutional Neural Network (CNN) and Support Vector Machine (SVM), which have been utilized for the identification of Lung Cancer from MRI images.

Chapter 7 discusses the experimental results of the thesis drawing comparison between conventional IP and machine learning approaches.

Chapter 8 presents the 'Conclusion'. Here, this research concludes the outcome and the significance of this work with the limitations. It narrates the future scope of the work as well.

CHAPTER 2

REVIEW OF EXISTING WORKS ON LUNG CANCER

This chapter begins with the descriptions of different parts of lung and how cancer is formed in lungs. Then, it describes different types of lung cancer along with their stages. Thereafter, Screening techniques using different electronic diagnostic tools have been described along with acceptability of MRI. Existing related research works using IP and ML approaches have been narrated subsequently. Finally, research gaps found from existing works are highlighted at the end of this chapter.

2.1 Lung Anatomy

The lungs are crucial organs of the human respiratory system, comprising the lower respiratory tract. They originate from the trachea and divide into two bronchial tubes for each lung on the left and right sides. The right lung is relatively larger, divided into three lobes, while the left lung is divided into two lobes (ACS 2022b). Lobes are balloon like subsection of ling filled with sponge like tissue (ACS 2022b). The bronchial tubes further branch into smaller air passages known as bronchioles, which ultimately terminate in tiny air sacs called alveoli. The human body contains approximately 600 million alveoli (Ratini 2021). Alveoli are enveloped by a network of minuscule blood vessels known as capillaries. The exchange of oxygen and carbon dioxide occurs between the lungs and the bloodstream via these capillaries.

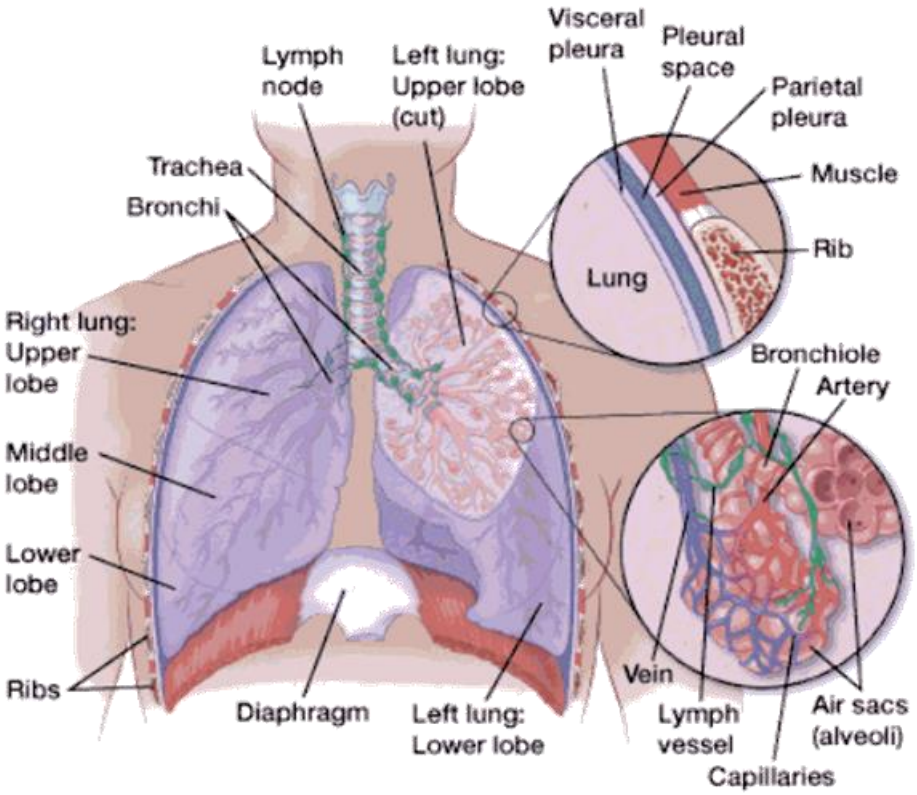

Fig. 2.1. Different parts of Lungs, Adapted from (ACS 2022b)

2.2 Lung Cancer

Smallest unit of a living body is cell. In a growing body cell division process takes place to increase the number of cells. In an adult body many mature cells do not grow or make copies of their cells with an exception to skin and blood cells, as they continue dividing all the time (Cancer Research 2020). When cells become abnormal or aged, they typically undergo cell death, and the body generates new cells to replace them. However, cancer begins when there is a disruption in the cell division process, leading to uncontrolled cell proliferation. In this scenario, the old or abnormal cells fail to undergo programmed cell death as they normally would (ACS 2022b). This phenomenon may take place on any parts of the body forming a lump or growth known as tumors. A tumor becomes cancer when it is malignant, otherwise it is benign and limited to a particular part of the body. Cancer cells have the capability to spread to other areas of the body, disrupting normal bodily functions.

2.3 Classification of Lung Cancer

There are essentially three primary types of lung cancer, which are:

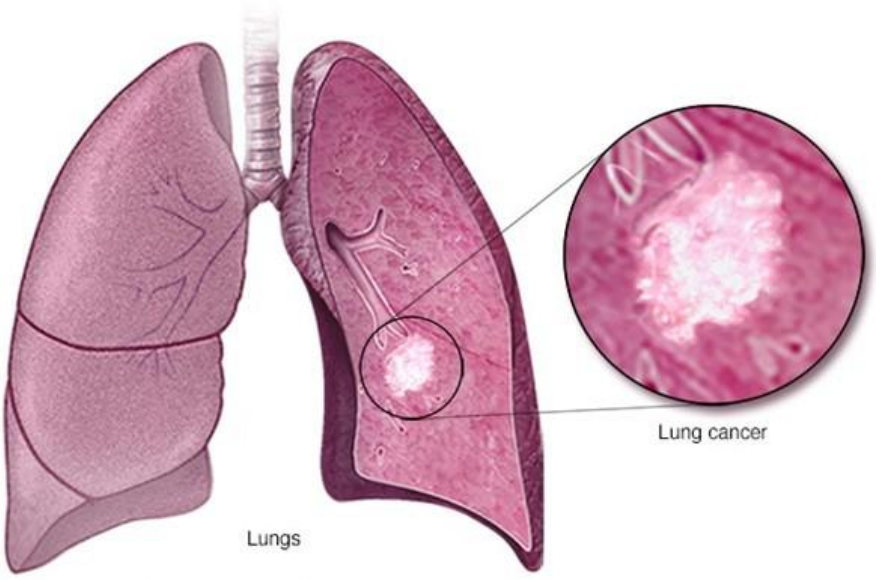

Fig. 2.2. Cancerous part of Lung, Adapted from (Mayo Clinic 2022)

(a) Non-Small Cell Lung Cancer (NSCLC).

(b) Small Cell Lung Cancer (SCLC).

(c) Bronchial Carcinoids.

A brief description of the types mentioned above are described below:

2.3.1 Non-Small Cell Lung Cancer (NSCLC)

This type is regarded as one of the most prevalent forms of lung cancer. According to the American Cancer Society (ACS), 80 to 85 percent of lung cancer cases are classified as NSCLC (Maurie2022). This type of cancer develops slowly than SCLC and causes few or no symptoms until it has grown to an advanced stage. NSCLC are of mainly three types:

(a) Adenocarcinoma of the lung is the most prevalent form of lung cancer. Typically, it is situated in the outer regions of the lung, within glands responsible for mucus secretion. Both smokers and non-smokers, as well as men and women, are equally susceptible to this type of cancer (Mayo Clinic 2022).

(b) Squamous cell carcinoma commonly originates in the bronchi near the center of the lungs, where the larger bronchi intersect the trachea. This type of cancer is typically exclusive to smokers (Mayo Clinic 2022).

(c) Large cell (undifferentiated) carcinoma may manifest in any area of the lung. It has a tendency to proliferate and metastasize rapidly, making treatment plan more challenging (ACS 2022b).

2.3.2 Small Cell Lung Cancer (SCLC)

SCLC, also known as oat-cell cancer, constitutes 10 to 15 percent of all lung cancers. Typically linked to tobacco smoking, this cancer type grows rapidly and spreads faster than any other form of cancer (Johns Hopkins, 2021). There are two distinct types of SCLC: small cell carcinoma (oat-cell cancer) and combined small cell carcinoma.

2.3.3 Bronchial Carcinoids

Carcinoid tumors are very uncommon, fewer than 5 percent of all lung tumors (Johns Hop- kins 2021). They typically tend to grow slower than other type of tumors. They are classified as typical or atypical carcinoids.

2.4 Lung Cancer Screening

The screening process for lung cancer often begins with patients describing their symptoms to healthcare professionals. Common symptoms that may indicate lung issues include chronic cough, shortness of breath, wheezing, persistent mucus production, coughing up blood, and chronic chest pain (American Lung 2022). Lung cancer screening involves conducting imaging tests and biopsies of the lung. Imaging tests utilize various techniques such as x-rays, magnetic resonance imaging (MRI), ultrasound, or positron emission tomography (PET) scans to capture images of the internal organs of the human body. Biopsy is used as a confirmation tool of lung cancer looking at the sample of lung cells. Screening test are carried out in the pathology laboratory to provide medication and to analyze the condition of the lung cancer by physician.

2.4.1 Traditional Microscopic Analysis

Biopsy is the typical and most traditional way of lung cancer screening. It is also used as the confirmatory tool for cancer detection in any part of human body. To do biopsy a small sample of body tissue is extracted and examined under the microscope. Tissue samples are extracted by a surgeon, and histopathologists examine these samples in their laboratory to identify the presence of cancer cells and symptoms. They scrutinize biopsy specimen slides using a microscope with various magnification levels, such as 10x, 20x, 40x, 100x, 200x, and 400x, to count and distinguish different types of tissues. Typically, sample tissues extracted for biopsy is sliced into very thin segments (5-20 microns), later placed onglass slides for analysis. Examining of slides in the lab is more informative than other modalities of cancer screening. This is due to the fact that the process represents tissue architecture of the body parts more clearly (Gurcan et al. 2009). To date, this procedure remains the gold standard for confirming the presence of a disease, despite its limitations (Kelsey et al. 2010). For instance, it demands a considerable amount of time for processing a single slide (typically 19-27 minutes for analysis using a microscope by experts), and experts encounter variability in observations both within and between individuals (Vander Kwast et al. 2010). Furthermore, instead of analyzing a large number of biopsy slides for a specific test (ranging from 1000 to 2500 slides for a single test), experts typically select slides randomly for analysis (Kothari et al. 2013). However, with the advent of digital scanners and IP techniques, digital histopathology has progressed significantly. It has streamlined the biopsy process, reducing both time and effort required, while also mitigating human error and subjectivity (Lamprecht, Sabatini, and Carpenter 2007).

2.4.2 Use of Radiology

Radiology is a medical specialty that utilizes imaging technology for disease diagnosis and treatment. Medical professionals specializing in this field are referred to as radiologists. There are differenttypes of scanning machines used by the radiologists to get images of internal organs of human body. These are Ultrasound, X-ray, CT (Computed Tomography), MRI (Magnetic Resonance Imaging), PET (Positron Emission

Tomography). Among all these scanning techniques, ultrasound is the most popular due to its affordability, portability, and lack of radiation effects (Kiruthika and Ramya 2014). However, its limitation lies in its ability to only process large and mature tissues (Skodras et al. 2009). Considering the limitations of US, X-ray is a viable diagnosis tool which is also cheap and widely available. So usually, it is used as the first-hand diagnosis technique. CT is used as an advanced diagnosis system. It provides detail diagnosis of any internal organ. Often, CT is combined with PET for enhanced visualization of specific areas of the body. MRI, on the other hand, employs powerful magnets and radio waves to generate detailed images of internal organs. It is an effective diagnosis tool to form images of soft tissues. As MRI technique is free of radiation, it is an appropriate diagnosis tool for the pediatric patients and the patients who require frequent diagnosis.

2.4.2.1 Chest X-ray

A chest X-ray is frequently the initial test ordered by doctors to evaluate suspicious areas of the lung. It uses high-energy electromagnetic radiation to get picture of lungs and surrounding tissues. X-ray sheet can show up changes in the lungs, although the changes can be due to cancer or for any other complicacies. X-ray along with acute symptoms or further investigation help doctors to diagnose lung cancer.

2.4.2.2 Computed Tomography (CT) Scan

A CT scanner utilizes X-rays to produce detailed cross-sectional images of internal body parts. It takes many pictures from different angles and then a computer is used to combine them together to make a 3-dimensional (3D) image of the lung. Doctors are more likely to detect lung tumors using CT scans compared to chest X-rays. CT scans aid in measuring the size, shape, and location of lung tumors. Additionally, they assist in identifying masses in the adrenal glands, liver, brain, and other organs that may indicate the spread of lung cancer. (ACS 2022a).

Fig. 2.3. CT Scanner, Adapted from (Cancer Research 2022a)

2.4.2.3 Magnetic Resonance Imaging (MRI)

Similar to CT scans, MRI scanners are employed to obtain detailed images of soft tissues within the body. Unlike X-ray and CT, it uses magnetism and radio waves. So, this diagnostic process is free from any sort of radiation hazards. That is why it is suggested for the patients who need to go through regular checks of internal organ, pregnant women and neonatal patients. MRI scans are frequently utilized to search for potential spread of lung cancer to the brain or spinal cord (ACS 2022a).

Fig. 2.4. MRI Scanner, Adapted from (Cancer Research 2022b)

2.4.2.4 PET (Positron Emission Tomography) Scan

A PET scan is a form of nuclear medicine imaging that involves the use of radioactive material similar to glucose, known as radiotracers. The most common radiotracer is FDG (F-18 fluoro-deoxy-glucose)which is injected into the blood. Cancer cells tend to absorb glucose at a higher rate. A special camera detects gamma ray emissions from the radiotracer. Thus, a PET scan helps to identify cancerous cells before tumors are formed. Frequently, a PET scan is combined with a CT scan. This combined approach helps to pinpoint tumors more precisely and may offer more accurate diagnoses compared to conducting the two scans separately (RSNA 2023).

Fig. 2.5. PET/CT Imaging Machine, Adapted from (Temple Health 2023)

2.4.3 Acceptability of MRI as a Screening Technique

CT or PET/CT is extensively and routinely utilized in the investigation of lung malignancies. The ongoing technical advancements, such as stronger gradients, parallel imaging, and shorter echo time, have rendered lung MRI an intriguing alternative to CT (Biederer et al. 2017). (Hirsch et al. 2020) opined that Lung MRI has the potential to replace up to 90% of CT examinations without compromising diagnostic accuracy. Furthermore, the advent of new TSE (turbo spin-echo) and T1-W ultrashort sequences allows for the production of CT-like images with high spatial resolution. (Hochhegger et al. 2011) suggested that MRI has the capability to detect and stage lung cancer, presenting

a potentially excellent alternative to CT or PET/CT in the investigation of lung malignancies. (Cieszanowski et al. 2016) conducted a comparative analysis between MRI and CT techniques for identifying and measuring the size of Pulmonary Nodules. It concluded that MRI exhibits high sensitivity in detecting lung nodules. The availability of high-performance gradient systems, alongside phased-array receiver coils and optimized imaging sequences, has rendered MRI of the lung feasible. Under optimal conditions, involving successful breath-holds with reliable gating or triggering, 90% of all 3-mm nodules can be accurately diagnosed, with nodules 5 mm and larger detected with 100% sensitivity (Wang et al. 2014). Again, lack of ionizing radiation has made it an appropriate alternative for the patients requiring frequent follow-up of pulmonary nodules and pediatrics.

2.4.4 Medical Image File Formats

Medical images serve some special purposes of record keeping and treating patents. Medical images consist series of images depicting projection of an anatomical volume of body parts. Even it may contain patient's personal information and about his diagnostic process. Pixel depth, photometric interpretation, metadata, and pixel data are among the typical elements of medical image formatting. Radiological images such as x-ray, CT, and MR images are generally interpreted in grayscale. Conversely, nuclear medicine images like PET are commonly displayed using a color map. CT and MRI images are acquired in a specialized digital format known as DICOM (Digital Imaging and Communications in Medicine) (.dcm, .dicom) format. Each CT or MRI scan comprises multiple images in the DICOM format. DICOM is a widely utilized medical imaging format that ensures the high quality of the images.

2.5 Related Works Using Conventional Image Processing Techniques for Lung Cancer Identification

IP technique has occupied a vast area in the field of research and industry. It has played a vital role in providing numerous solutions where images are used. Health sector also uses

IP technique for diagnosis of diseases and determining its stage to apply appropriate treatment. Automated IP technique expedite the whole diagnosis process and act as a tool to reduce the probability of human error. IP techniques are integral to the identification and staging of Lung Cancer. CT or PET/CT and MRI are extensively and routinely employed in the investigation of lung malignancies and other diseases. IP steps varies basing on the approaches followed by different researchers. Each step has different methods which are used by different researchers in numerous ways. This is due to the fact that there is no universal IP technique for medical imaging. Typically, IP technique includes followingtwo phases:

(a) Phase one or Preprocessing

(b) Phase two or Postprocessing

2.5.1 Preprocessing

Image preprocessing represents the initial phase of IP technique. Its objective is to enhance the quality of the image to a degree where it can be effectively analyzed. It is done by suppressing outlier and undesired data element. Generally, preprocessing considers following steps:

(a) Enhance Contrast

(b) Noise Removal

(c) Segmentation

2.5.1.1 Enhance Contrast

Contrast refers to the degree of variation between darker and lighter portions within an image. Contrast Enhancement is a technique to amplify the visible difference among different structure within an image. It plays a crucial role in IP techniques as it enhances the visibility of important information contained within the image. It is widely used in medical IP. Contrast enhancement helps better segmentation of Lung MRI and allows easier identification and classification of cancer tissues. Image enhancement serves the following purposes:

(a) Enhanced segmentation of lung MRI images in subsequent steps utilizing semi-automatic and automatic methodologies.

(b) Efficient identification of lung nodules and classification of cancer tissues.

There are different ways and means of contrast enhancement. Among which histogram processing is frequently used which includes following two approaches:

(a) Normalization (a procedure altering the range of pixel intensity values).

(b) Histogram Equalization (a widely employed technique in IP for contrast enhancement utilizing the image's histogram).

The histogram equalization technique finds wide application in medical IP. But it has some drawbacks as well. Sometimes, this approach often produces unrealistic effects (Entwistle 2004), (Entwistle 2005). Histogram equalization has two different categories:

(a) Local Histogram Equalization

(b) Global Histogram Equalization

MRI images are usually grayscale images. That's why working directly with the intensity of a grayscale image can be challenging (Nishu 2012). One of the major disadvantages of using grayscale images is that only mature follicles can be identified.

2.5.1.2 Noise Removal

Noise in the image is a big barrier for image segmentation and classification afterwards. It is very important to remove unwanted noise or unwanted regions from the images at the beginning of any preprocessing stage. It is a general criterion that in every IP approach at least one of the noise removal techniques is incorporated. Median filter is commonly used for both semi-automatic and automatic approaches for cancer cell identification (Sertel et al. 2009; Bapure 2012; Liu et al. 2004; Prabhpreet Kaur, G. Singh, and Parminder Kaur 2020). (Bari et al. 2019) also conducted a review of research literature on Lung Cancer detection to identify the IP techniques employed. They summarized that Gabor Filter, Weiner Filter, Fast Fourier Transformation, etc., are predominantly utilized in the image

preprocessing phase. (Kannan and Naveen 2020) used mean and median filters in preprocessing stage for different types of lung images captured using CT, MRI and X-Ray. (Sazzad et al. 2019) proposed an automated system to detect brain tumor from MRI, where they used median filter for noise removal. (Majib, Sazzad, and Rahman 2020) designed a model for brain tumor detection from MRI using V channel of HSV image where they incorporated morphological operation for noise removal. In this study, morphological operations were applied instead of using a median filter to eliminate undesirable regions and aid in the segmentation and identification of regions of interest.

2.5.1.3 Segmentation

Typically, segmentation constitutes the second phase of any automatic or semi-automatic IP method. Here, the image undergoes division into distinct components. Fundamentally, segmentation aims to partition a digital image into regions that hold significance or exhibit perceptual similarity. Various methodologies, including edge-based approaches, have been explored by researchers for image segmentation (Picut et al. 2008), threshold-based approaches (Kelsey et al. 2010; Sertel et al. 2009) and by using mathematical morphology (Bapure 2012; Skodras et al. 2009). (Bari et al. 2019) It can be summarized that techniques such as Thresholding, Morphological, Marker-Controlled Watershed, and Watershed Segmentation are employed during the image segmentation phase. Otsu's segmentation, based on thresholding, was applied to the filtered image composed of green and blue channels (Majib, Sazzad, and Rahman 2020). Feature extraction involved utilizing shape and size attributes such as region area, circularity, solidity, roundness, radius, and diameter.

2.5.2 Postprocessing

Postprocessing includes identification and classification of images. Supervised learning/-Classification and unsupervised learning/Clustering are used to explore the ROI from segmented images. Several features play a crucial role in identifying Regions of Interest (ROIs). (Asuntha et al. 2016) utilized Particle Swarm Optimization (PSO), Genetic Optimization, and the SVM algorithm for feature selection and classification. (Keerthana,

Thamilselvan, and Sathiaseelan 2016) suggested employing the decision tree algorithm, asserting its superiority in classifying Lung nodules with MRI over SVM, Naive Bayes, and CART algorithms. (Rodrigues et al. 2018) introduced a Structural Co-occurrence Matrix (SCM)-based approach for classifying pulmonary nodules as malignant or benign. (Shravya and Rajesh 2019) proposed eight combinations of SCM and classifier, with the most successful being SCM extractor paired with a Mean filter and MNN (Morphological Neural Network) classifier. (Sazzad et al. 2019) enhanced brain tumor identification by adding and complementing the green and blue channels of RGB images, followed by applying Otsu's method and transforming the images into binary form. Tumor regions were then detected using features such as area, roundness, diameter, and solidity. (Bari et al., 2019) concluded that structural features including area, centroid, perimeter, orientation, projection, aspect ratio, and Euler number are crucial for feature extraction. (Kannan and Naveen 2020) concluded that K-Means Clustering algorithm shows a better result comparing to Otsu's segmentation algorithm in segmenting the lung tumors irrespective of the type of images, e.g., CT, MRI and X-Ray image.

2.6 Related Works Using ML for Lung Cancer Identification

ML techniques are becoming more prevalent in the realms of image-based diagnosis, disease prognosis, and risk assessment. Over the time several ML and deep learning methods and algorithmshave evolved to determine lung cancer nodules from lung images (Rajalaxmi et al. 2022). Relating to this, (De Bruijne 2016) identified three main challenges:

(a) Deal with variations in imaging protocols.

(b) Learn from ambiguous or weakly labeled data.

(c) Interpreting and assessing the outcomes and findings.

(G. A. P. Singh and Gupta 2019) proposed a five-stage model as following for identifying Lung nodules as benign or malignant:

(a) Acquisition of image.

(b) Preprocessing and segmentation.

(c) Feature selection.

(d) Identification.

(e) Evaluation of performance.

In image segmentation, the watershed segmentation technique outperformed the thresholding technique. Texture features were extracted using two methods: the gray-level co-occurrence matrix (GLCM) and statistical parameters. Subsequently, seven distinct classifiers were employed: k-nearest neighbor (KNN), support vector machine (SVM), decision tree, multinomial naive Bayes, stochastic gradient descent random forest, and multi-layer perceptron (MLP) classifier; for performance evaluation and found MLP classifier showed highest accuracy of 88.55%. To streamline the intensive IP workload, (Jiang et al. 2017) introduced a lung nodule detection scheme that relies on multi-group patches extracted from lung images. They used Frangi filter to enhance vessel image and sub- tract the same from the original image. Lastly, Convolutional Neural Network (CNN) is employed to determine whether the nodule is cancerous. The researchers achieved a sensitivity of 94%. (Devarapalli, Kalluri, and Dondeti 2019) outlined the utilization of various IP methods such as the median-wiener filter during the preprocessing phase. For classification purposes, models including the Back Propagation model, Support Vector Machine (SVM), Forward Neural Networks, and CNN are utilized to identify the presence of cancerous nodules. (Kadam 2022) proposed a typical IP method to determine malignant and benign lung nodule using SVM hyper plane algorithm. They used Discrete Wavelet Transform (DWT) techniques for image segmentation and GLCM matrix for extracting features like entropy, co-relation, energy, contrast and dissimilarities. (Hussain et al. 2022) applied various ML algorithms on both normal and enhanced lung cancer images. They extracted GLCM features after applying various image enhancement methods and found much higher accuracy using SVM, RBF and polynomial kernels.

(Abdullah, Abdulazeez, and Sallow 2021) investigated the classification accuracy of SVM, KNN and CNN algorithms using WEKA Tool. Their findings indicate that SVM achieved the highest accuracy, with 95.56%, followed by CNN and KNN, which achieved accuracies of 92.11% and 88.40%, respectively. (Tiwari 2016) conducted a review of research works on Lung Cancer detection and found that Back Propagation Neural Network act as the best classifier working on CT images followed by Genetic Algorithm, SVM, SVM-KNN, Neuro-Fuzzy and Bayesian classification technique. (Fernandes et al., 2022) introduced a lung nodule classification method combining CNN with the extra tree feature selection algorithm. This model attained an accuracy of approximately 93.74%, a sensitivity of 94.02%, and a specificity of 93.07% using the LUNA16 dataset. (Thamilselvan and Sathiaseelan 2016a) presented an enhanced KNN technique for identifying and categorizing lung cancer from MRI images of the lung. The method involves four steps of KNN, which include calculating based on Euclidean distance, defining the k value, assigning the majority class, and determining the minimum distance. (Thamilselvan and Sathiaseelan 2016b) proposed utilizing PCA (Principal Component Analysis) for feature reduction and employing a modified Classification and Regression Tree (CART) approach to identify and classify MRI images of the lung, demonstrating reduced processing time and improved accuracy. After a detailed survey (Abdullah, Abdulazeez, and Sallow 2021) concluded that deep learning techniques for identifying lung nodules obtained better accuracy than different classical ML techniques where, the highest accuracy was approximately 99% using multi-resolution patch-based CNNs. (Makde et al. 2018) introduced a Deep Learning method utilizing CNN for tumor detection in lung nodules using CT images and brain MRI. The study applied the framework on AlexNet and ZFNet architectures, achieving a classification accuracy of over 97%. To implement Deep Learning in CAD, requires huge amount of high-quality labeled image as training samples. (Mohammed and Çinar, 2021) employed Transfer Learning and data augmentation techniques to address the challenge of limited training data and overfitting. They used four pre-trained models, namely AlexNet, ResNet18, GoogleNet and ResNet50 to classify Benign and Malignant lung cancer. Where AlexNet

achieved better results and GoogleNet showed comparatively poor performance.

2.7 Summary

Basing on the above discussion the chapter can be summarized as following. So far, only a few research works have been conducted for identifying lung nodules using MRI images. Calibration of parameters are required for identifying lung nodules from different batches of images. Human intervention is essential for segmenting ROI appropriately. In preprocessing step, enhance contrast is generally performed by histogram equalization. Addition- ally, median and mean filter are commonly used to remove noise. Besides Gabor filter, Weiner filter, Gaussian filter, Laplacian filter, Sobel filter etc. are also used. Edge-based and threshold-based methodologies are frequently employed during the image segmentation phase. Structural attributes such as area, centroid, circularity, solidity, roundness, radius, diameter, perimeter, aspect ratio, etc., are utilized for feature extraction. Commonly employed classifiers for lung nodule classification include CNN, KNN, SVM, CART, decision tree, multinomial naive Bayes, etc. Additionally, pre-trained models like AlexNet, ResNet18, GoogleNet, and ResNet50 have been utilized for distinguishing between benign and malignant lung cancer nodules.

Table 2.1: Existing Work on Lung Cancer Identification

Ser	Existing Work	Steps Used	Identification
1.	(Bari et al. 2019)	• Image pre-processing involves the utilization of techniques such as Gabor Filter, Wiener Filter, Fast Fourier Transformation, etc. • The image segmentation phase employs methods like Thresholding, Morphological operations, Marker-Controlled Watershed, and Watershed Segmentation. • Features extraction relies on structural attributes including area, centroid, perimeter, orientation, projection, aspect ratio, Euler number, etc. • Classification entails the application of algorithms such as KNN, ANN, SVM, and CNN.	Lung Cancer (Review research)
2.	(Kannan and Naveen 2020)	• Mean and median filters were used in the pre-processing stage. • K-Means clustering algorithm shows a better result comparing to Otsu's segmentation algorithm in segmenting the lung tumors irrespective of the type of images. • SNR, MSE, PSNR were used as performance evaluation parameters.	Lung Cancer (CT/MRI)

3.	(Thamilselvan and Sathiaseelan 2016b)	• Used PCA for reduction of features. • EKNN (Enhanced K-Nearest Neighbor) incorporates four steps of KNN: calculation based on Euclidean distance, determination of the k value, assignment of the majority class, and identification of the minimum distance. • Used modified CART method to identify and classify MRI of Lung images which takes less processing time and shows better accuracy.	Lung Cancer (MRI)
4.	(Asuntha et al. 2016)	• Employs CT, MRI, and Ultrasound images as input. • Utilizes the Gabor filter for image enhancement. • Implements a superpixel segmentation algorithm during the segmentation stage. • Utilizes Particle Swarm Optimization (PSO), Genetic Optimization, and the SVM algorithm for feature selection and classification.	Lung Cancer (CT/M-RI/ USG Image)

5.	(Rodrigues et al. 2018)	• Employed a Structural Co-occurrence Matrix (SCM) based approach to differentiate pulmonary nodules into malignant and benign categories. • Tested eight combinations of SCM and classifiers, each incorporating a filter and specific input image type. • Achieved the most favorable outcome with the SCM extractor combined with the Mean filter.	Lung Cancer (Grayscale and Hounsfield Units).
6.	(Shravya and Rajesh 2019)	• Suggested eight combinations of SCM and classifier, each employing a filter and grayscale image as input type. • Best result was achieved by SCM extractor combined with Mean filter and MNN classifier.	Lung Cancer (Grayscale image)
7.	(Sazzad et al. 2019)	• Used contrast stretching for image enhancement and median filter for noise removal. • In segmentation stage green and blue channel of the RGB image was added and complemented. Following this, Otsu's method was applied to the complemented image, resulting in its transformation into a binary image. • Tumor regions were identified by utilizing features such as area, roundness, diameter, and solidity.	Brain Tumor (MRI)

8.	(Majib, Sazzad, and Rahman 2020)	• Used V channel of HSV image. Histogram equalization and median filter [3x3] is ap- plied for image enhancement. • Instead of using a filter operation, Morpho- logical operations are applied on reconverted RGB image. • Otsu's segmentation threshold-based approach, is applied to the filtered image containing only the green and blue channels. • Feature extraction involves utilizing shape and size attributes such as region area, circularity, solidity, roundness, radius, and diameter.	Brain Tumor (MRI)
9.	(Jiang et al. 2017)	• Used Frangi filter to enhance vessel image and subtract the same from the original image. • CNN is used to classify lung nodule into cancerous and non-cancerous.	Lung Cancer (CT image)
10.	(Devarapalli, Kalluri, and Dondeti 2019)	• Median-wiener filter is used in the pre-processing stage. • To classify cancerous and non-cancerous cells, the Back Propagation model, Support Vector Machine (SVM), Forward Neural Networks, and Convolutional Neural Networks (CNN) are employed.	Lung Cancer (CT image)

11.	(Makde et al. 2018)	• Deep Learning technique using CNN using CT and MRI of brain. • Implemented the framework on AlexNet and ZFNet architectures.	Brain Tumor (CT and MRI)
12.	(Mohammed And Çinar2021)	• Used Transfer Learning and data augmentation technique. • Used four pre-trained models, namely AlexNet, ResNet18, GoogleNet and ResNet50 to classify Benign and Malignant lung cancer.	Lung Cancer (CT image)
13.	(Tiwari 2016)	• Found that Back Propagation Neural Net-work as the best classifier working on CT images followed by Genetic Algorithm, SVM, SVM-KNN, Neuro-Fuzzy andBayesian classification technique.	Lung Cancer (CT image)
14.	(Keerthana, Thamilselvan, andSathiaseelan 2016)	• Found decision tree algorithm to work better than SVM, Naive Bayes, CART algorithms.	(MRI)

CHAPTER 3

METHODOLOGY

3.1 Introduction

This chapter offers a synopsis of the dataset and delineates the methodologies utilized in this research endeavor. It starts with the research methodology followed by research approach where it outlines the whole research work using a diagram. Then, research activities are described, starting with the description of the data collection, followed by development and validation of proposed approach. Finally, it describes the proposed technique and ends witha summary of the chapter.

3.2 Research Methodology

All existing research endeavors can be classified into three categories: diagnostic research, qualitative research, and quantitative research (Kothari et al. 2013). Different types of re- search works follow different research methodology. There are numerous researches works that follow a combinational approach. The research methodology employed in this study combines elements of diagnostic and quantitative research.

3.3 Research Approach

The research methodology for this study encompassed several phases, as illustrated in Figure 3.1. These stages involved identifying the research problem, conducting a literature review, formulating the research methodology, identifying gaps and issues in existing approaches, collecting data, analyzing the data, and validating the proposed approach.

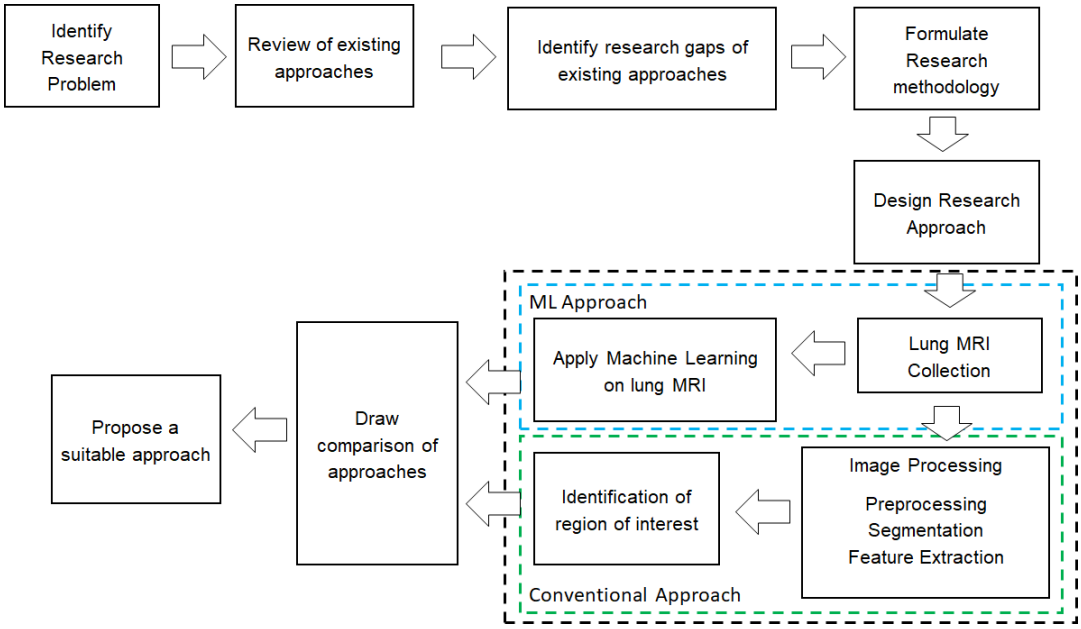

Fig. 3.1. Research Approach

3.4 Research Activity

Research activity as planned for identifying lung cancer is described below:

3.4.1 Data Collection

Usually, CT is used for diagnosis of lung disease. Lung continuously expands and shrinks for air inhalation and exhalation. Carrying out MRI of an organ whose shape changes continuously is difficult, hence searching for lung MRI image in different hospital and diagnostic centers can be a challenge. In this scenario, Lung MRI found in the open-source information system was the only source of its collection. However, images and data were validated by a medical expert, Md. Sadequel Islam, affiliated with Talukder Diagnostic Center in Mymensingh, Bangladesh. Dr. Islam, a renowned radiologist, formerly held the position of Associate Professor at Dinajpur Medical College, Bangladesh, and was nominated as the domain expert for this research study.

3.4.2 Development of Proposed Approach

A multitude of research works exist to date, which can be classified into categories such as diagnostic research, qualitative research, and quantitative research (Kothari et al. 2013).

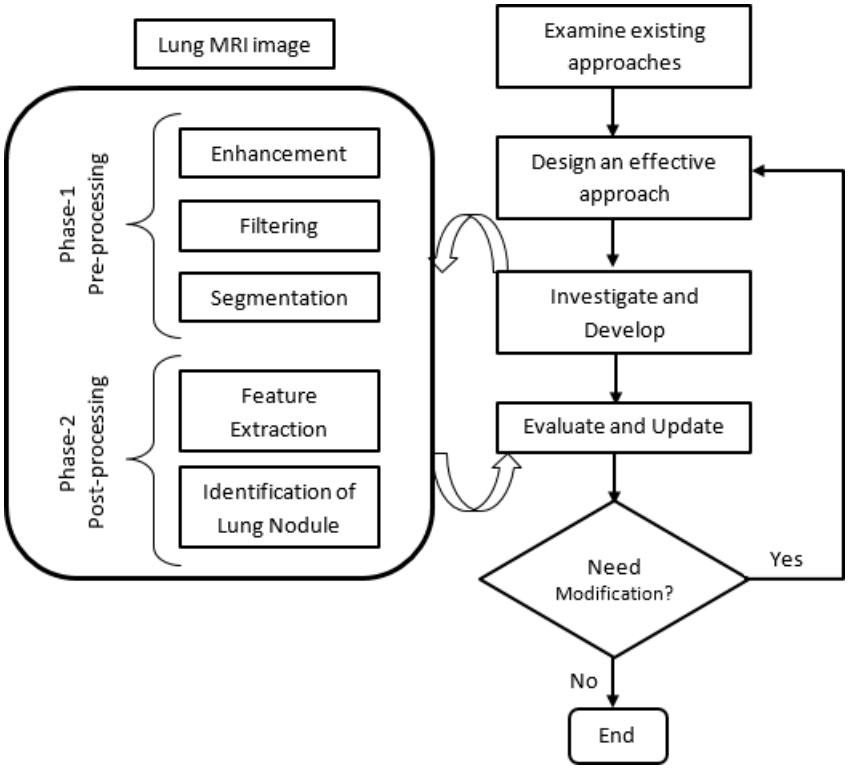

Fig. 3.2. Research Activity Plan with steps for the proposed approach

Each of these research types employs distinct methodologies. The methodology adopted in this study integrates elements of diagnostic research and a quantitative approach. A comprehensive description of the proposed approach in this study can be found in the following chapter.

3.4.3 Dataset Used in this Study

As lung cancer research using MRI is limited is Bangladesh, hence, it is a challenge to obtain lung MRI images locally. A total of 543 lung MRI images were collected from a radiologist, of which 379 MRIs contained lung cancer and rest 164 MRIs were of normal lung. All images are in JPEG format with a bit depth of 8 bits, and the dimensions of the test image were 500 x 500 pixels.

3.4.4 Validating the Proposed Approach

For an IP expert who does not have a medical background, need to get assistance of a medical expert to validate the results/outcome. This assistance can be provided in the

following manner:

- Validate the test results with a medical expert immediately following the completion of the IP phases.
- Acquire two identical image datasets: one serving as the test dataset, and the other containing regions marked by experts. Subsequently, validate the test results by comparing them with the marked regions.

Figure 3.3 illustrates an example of a matched identical pair consisting of labeled and unlabeled images, utilized as test data in this research investigation.

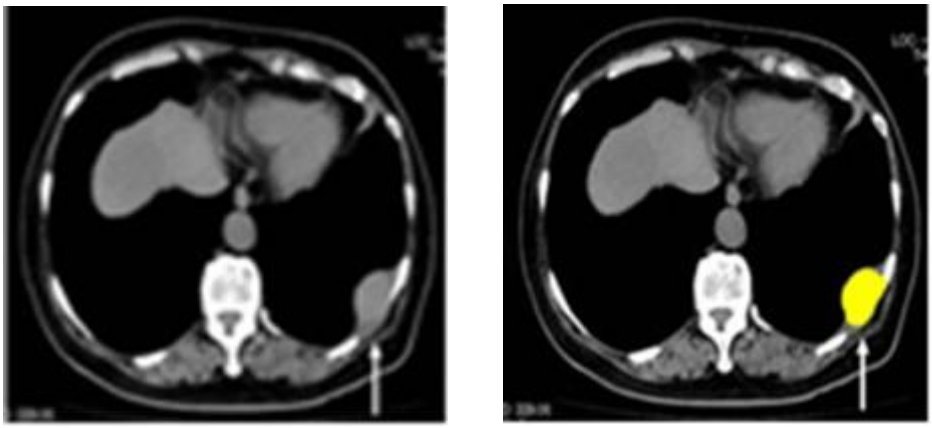

Fig. 3.3. Left: MRI of Lung Tumour (Test Image). Right: MRI of Lung Tumour (marked by experts). The images are identical, collected from radiologist.

Table 3.1 displays the random grouping of image datasets utilized in this research endeavor. The datasets are categorized into a training dataset, comprising test images paired with corresponding expert-tagged images. This randomly selected test dataset is employed for developing the proposed method, while an unmarked test dataset is utilized to assess the performance of the developed method. Validation is conducted by comparing the outcomes of the approach proposed in this study with regions annotated by experts in one image.

Table 3.1: Partition of the datasets into training and test dataset

Data	Images labelled by Experts	Test images
Data used for Training	A	A
Data used for Testing	B	B

3.5 Proposed Technique

This research study adheres to a research methodology plan (Figure 3.2), which outlines two distinct phases. A review of the recent works state that multiple IP steps can be incorporated into each phase. Different types of algorithms are currently available for each individual step. The primary objective is to maintain simplicity in IP and assess the effectiveness of basic IP steps for a given dataset. Secondary priority involves experimenting with various existing algorithms to identify those best suited for the dataset. Should existing approaches fail to yield satisfactory results, new or modified approaches should be proposed. The third and final priority is to compare computer-generated results with human-analyzed results to ensure validation and accuracy.

Drawing from a comprehensive literature review, this research study incorporates the most pertinent and widely utilized techniques. Furthermore, it introduces modified or novel techniques as necessary at each stage, culminating in a tailored and meticulously viable approach for the test image datasets utilized in this research investigation.

3.6 Summary

The methodology employed in this research is a fusion of diagnostic and quantitative research methodologies. Basically, this work draws a comparison between conventional IP approach and ML approach in identifying lung nodules. After making the com-parison the paper also recommends the best suited approach for lung cancer identification. In doing so necessary research activities has been identified and described using a diagram. The difficulties in performing the research with respect to dataset has also been found out. MRI of lung is yet to be a widely used technology for lung diagnosis, as such, collection of lung MRI image is a big challenge in this research work. The researcher had to depend on the online open sources for data collection of lung MRI. Output or result of the approaches are validated by a radiologist who had expertise in reading lung MRI image and finding the presence of lung nodules.

CHAPTER 4

IMAGE PROCESSING STEPS FOR LUNG CANCER DETECTION

This chapter describes the IP part of lung MRI to detect lung module. The process starts with preprocessing which includes intensity correction and elimination of noise from the image. The process is followed by segmentation, feature extraction and identification of lung nodule. Finally, radiologist gets a binary image of original lung MRI which can be used for further investigation.

4.1 Introduction

The early detection of lung cancer is crucial for improving patient survival rates. Detecting lung cancer at an early stage provides patients with more treatment options and significantly increases the likelihood of successful treatment. Delayed or unavailability of cancer treatment reduces survival chances, increases treatment-related problems and the cost of treatment. IP plays an important role to aid automatic detection of lung cancer without the intervention of a medical expert and can assist them to provide a suitable treatment plan. It augments the effort of a radiologist and reduces the possibility of human error. In this context, the IP steps involved in lung nodule detection play a critical role in the treatment of lung cancer patients.

4.2 Conventional Image Processing Approach

A conventional IP approach typically consists of number of steps: Enhancement, Filtering, Segmentation, Feature selection or Extraction and finally identification or Classification. All the above-mentioned steps can be divided in two different phases: preprocessing (Phase-1) and post-processing (Phase-2). Phase-1 or preprocessing includes: Enhancement, Filtering, Segmentation whereas Phase-2 or postprocessing includes Feature selection or Extraction followed by Identification or Classification.

There is no such "Gold standard" which defines IP steps sequentially and hence it is possible to incorporate IP steps the researchers want to incorporate. In this research, Filtering, Enhancement and Segmentation were incorporated as preprocessing steps, and Feature Extraction and Identification were incorporated as postprocessing steps. Figure 4.1 shows different steps of conventional IP approach.

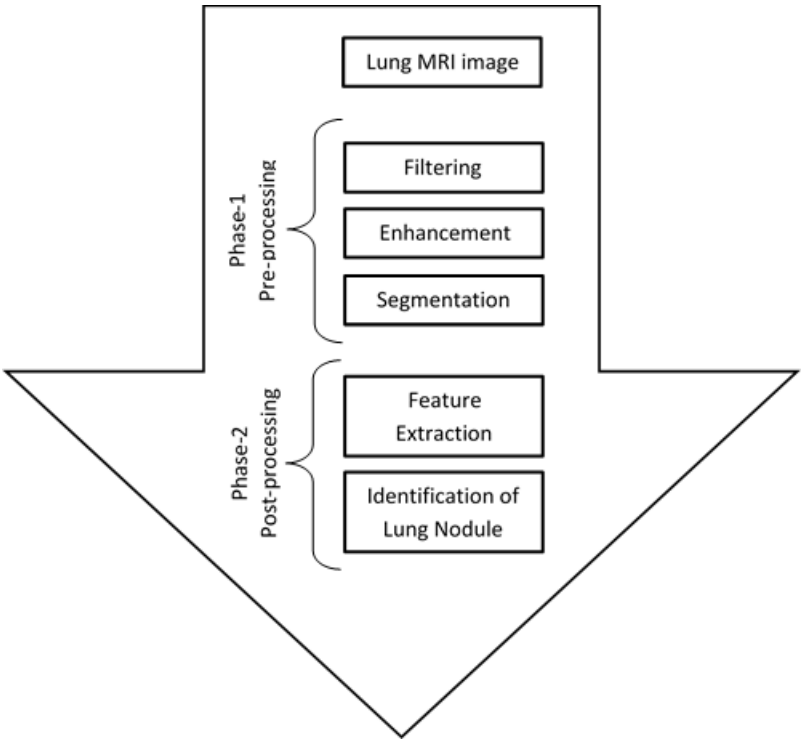

Fig. 4.1. Steps of Conventional Image Processing Approach

4.3 Preprocessing

In computer-aided IP, preprocessing is regarded as the initial step. It involves the elimination of unwanted or irrelevant areas and enhancing prominence by adjusting contrast (Lasker 2008), (Perumal and Velmurugan 2018), (Li and Gao 2013). For this research study, several steps have been integrated to detect or identify the Region of Interest (ROI). These steps include:

(a) Filter Operation

(b) Enhancement

(c) Segmentation

A filtering operation was implemented to eliminate as much unwanted region as possible from the captured MRI images, which was then followed by image enhancement and segmentation. In typical IP tasks, enhancement is often regarded as the initial step. However, in this study, a filtering operation was utilized as the first step instead of enhancement. This is due to the fact that it gives better result in identifying ROIs. The preprocessing methodology isshown in Figure 4.2.

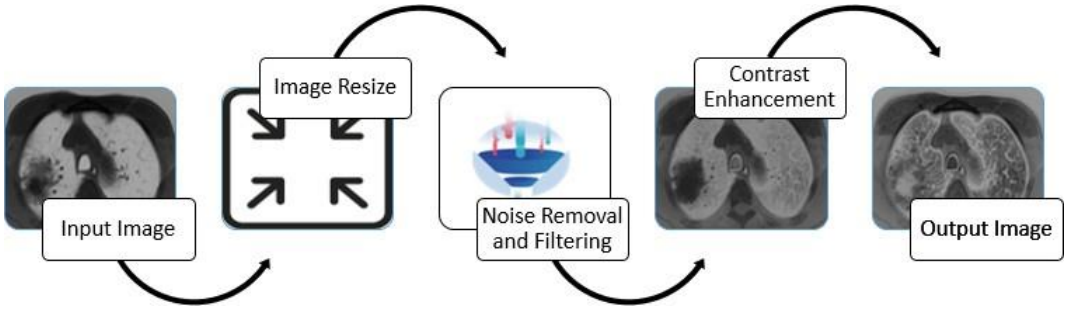

Fig. 4.2. Image Preprocessing Methodology

4.3.1 Filtering for Removal of Noise

Image noise refers to random fluctuations in brightness or color information within images generated by medical equipment or scanners. It is commonly viewed as an undesired outcome of image acquisition, representing the uncertainty of a signal caused

by random fluctuations. (M. Kaur and Behal 2013). When capturing images with an electronic scanner, unwanted noise can occur due to the weak signal from the scanner. Removing such unwanted regions captured by electronic scanners, particularly in MRI scans, is crucial. In this study, unwanted noises are the regions which cannot be considered as lung tumor and hence it is essential to remove them as much as possible. To do this a number of filter operations can be employed. Existing literature review indicates that a number of filter operations have been employed, which includes Mean, Gaussian, Laplacian, Wiener, Bilateral, Sobel, Prewitt, Spatial, low pass, and high pass filters etc. It is also possible to incorporate a number of filter operation sequentially. To achieve superior noise elimination, this study initially incorporated mathematical morphological operations, followed by the implementation of median filter operations to retain image edges.

4.3.1.1 Morphological Opening

Morphological operations generally involve taking a binary image and a structuring element as input, combining them using a set of operators (such as intersection, union, containment, complement). Unwanted regions can be eliminated using mathematical morphological operations, including erosion, dilation, morphological opening, morphological closing, bottom-hat, and top-hat filtering. Morphological opening incorporates erosion followed by dilation which entails to remove border pixels of an image region. This process aids in eliminating small objects and thin lines from an image while maintaining the shape and size of larger objects within the image. Bright features smaller than the structuring element cause a significant decrease in intensity, while larger features remain more or less unchanged in intensity. Figure 4.3 shows a sample MRI image and its version after applying morphological opening.

4.3.1.2 Median Filter

A median filter is a nonlinear digital filtering technique frequently employed to eliminate noise from an image or signal. It is highly favored for its effectiveness in noise removal while retaining edges and significant details in the image. The fundamental concept

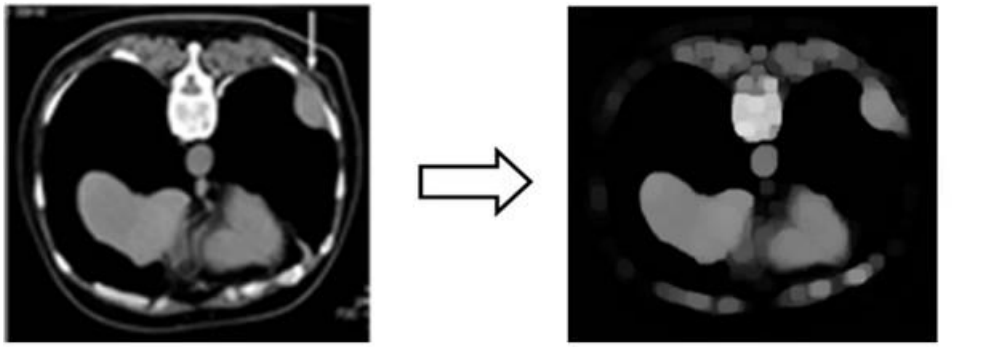

Fig. 4.3. Morphological Opening applied on Sample Lung MRI

behind the median filter involves iterating through the input signal, replacing each entry with the median value of its neighboring entries. This process occurs within a neighborhood pattern known as a "window," which moves entry by entry across the signal. Such noise reduction serves as a common preprocessing step to enhance the results of subsequent processing. Median filtering is widely adopted in digital IP due to its ability to preserve region edges. However, a notable drawback is its tendency to introduce a blurring effect, which can prolong processing time. Mitigating this effect involves selecting an appropriate clipping radius along with an optimal mask size. A commonly utilized kernel for the median filter is the 3x3 square kernel. Figure 4.4 shows a lung MRI image after applying Median filteron the output image of morphological opening.

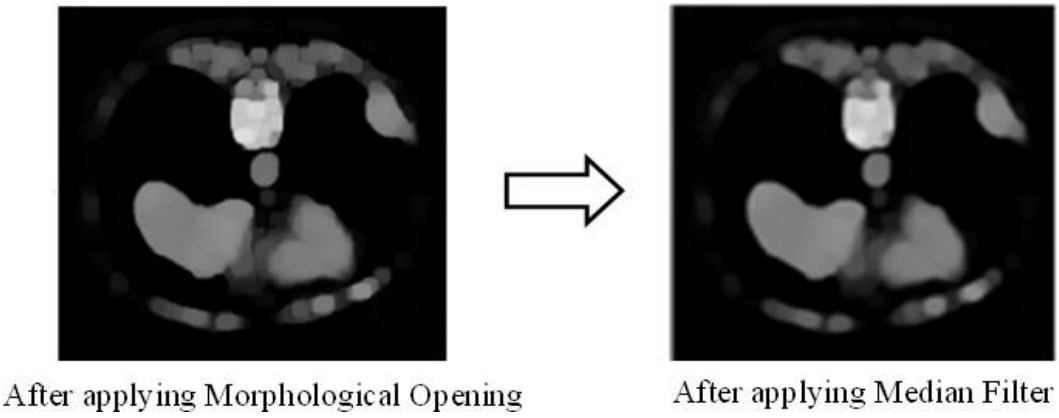

Fig. 4.4. Median Filter applied on the Lung MRI after applying Morphological Opening

4.3.2 Enhancement

Usually, the enhancement operation is carried out at the initial stage of most IP tasks. Although it is not essential but depending on the image quality (intensity variation issues) it can be incorporated. In this study, the image enhancement operation has been integrated after the filtering operation, as it significantly accentuates the results of the filtering process. Histogram processing is usually a popular approach of enhancement. Histogram processing consists of Histogram Equalization and Histogram Normalization. However, existing research work indicates that compared to histogram normalization, histogram equalization performs better (Majib, Sazzad, and Rahman 2020). In this study histogram equalization was incorporated.

Histogram equalization can be performed locally and globally. Local enhancement works on block by block, whereas global enhancement works on the whole picture. In many instances, the global histogram equalization approach proves ineffective in enhancing smaller regions (Sazzad et al. 2019). In this study, both local and global histogram equalization enhancement techniques were implemented, with the local enhancement operation demonstrating superior performance for the dataset utilized in this study. Figure 4.5 shows the results of applying enhancement operation on filtered lung MRI.

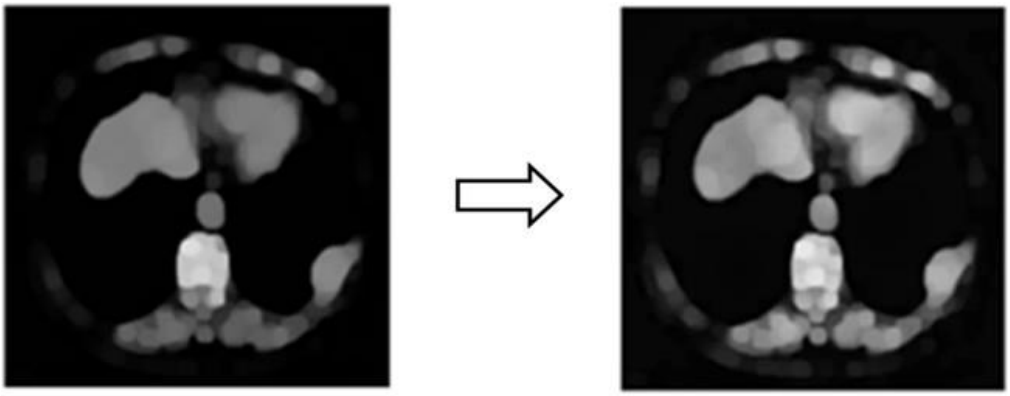

Fig. 4.5. Enhancement Operation applied on Filtered Lung MRI

4.3.3 Segmentation

Segmentation is the last step of Phase-1 and it is a vital step. If we do not get a good segmented region, it is very hard and a tedious task for any algorithm to identify or

analyze an image with precise accuracy. The primary objective of image segmentation is to simplify images to facilitate easier analysis. Segmentations are of two types: color segmentation and grey scale segmentation. There are numerous grayscale and color image segmentation approaches, with the following being the primary types of image segmentation techniques:

(a) Binary or Threshold based Segmentation

(b) Edge-Based Segmentation

(c) Region-Based Segmentation

(d) Watershed Segmentation

(e) Clustering-Based Segmentation

(f) Neural Networks for Segmentation

(g) Graph-cut

(h) Normalized-cut etc.

MRI images are grey scale image and hence color segmentation is not a viable choice for this study images. Among the various grayscale segmentation approaches, the threshold-based segmentation method stands out as the simplest, widely adopted, and popular technique for image segmentation. It converts grayscale images into binary images, where black pixels represent the background and white pixels represent the foreground, and vice versa. This conversion is achieved using a single parameter known as the intensity threshold. In a single-pass operation, each pixel in the image is compared to the intensity threshold. Pixels with intensity values higher than the threshold are set to white, while those with lower intensity values are set to black. For example, if the limit is set to 120 [intensity range 0-255], the pixel values greater than 120 will be converted to 1 and the pixel values less than 120 will be converted to 0. This technique is simple yet highly effective for partitioning or segmenting an image into background and foreground. It is useful when the required object has a higher intensity than the background objects which

are not necessary to retain. The technique works better if the image has very little noise. Based on different threshold values, the thresholding segmentation technique can be categorized into the following categories:

(a) Simple Thresholding.

(b) Otsu's Binarization.

(c) Adaptive Thresholding.

For simple thresholding, a constant threshold value is used to perform image segmentation which does not guarantee an appropriate value for segmentation. In this scenario, pixel intensity below the threshold value is converted to black, while those above the threshold value are converted to white. In this study, Otsu thresholding was employed instead of Simple Thresholding, due to the fact that, in case of Simple Thresholding user needs to set the threshold level which make the process semiautomatic and unreliable. Additionally, it may cause over and under segmentation. Over-segmentation causes loss of number of required regions and under-segmentation causes many unnecessary regions to exists.

Otsu threshold-based approach can be used using single threshold level and multi-threshold level. In case of Otsu single thresholding average of foreground and background peaks are set as threshold level and hence, it may cause over and under segmentation. Additionally, Otsu single thresholding can't be used on images that are not bimodal. To mitigate these limitations, it is always a viable choice to use Otsu Multi-thresholding approach. This approach categorizes the pixels of an input image into various classes, with each class determined by the intensity of the gray levels within the image. It reduces over and under seg- mentation. Interestingly, from literature review it is found that Otsu's Multi Thresholding approach has not been used till today. Figure 4.6 shows the results of applying segmentation operation on enhanced lung MRI.

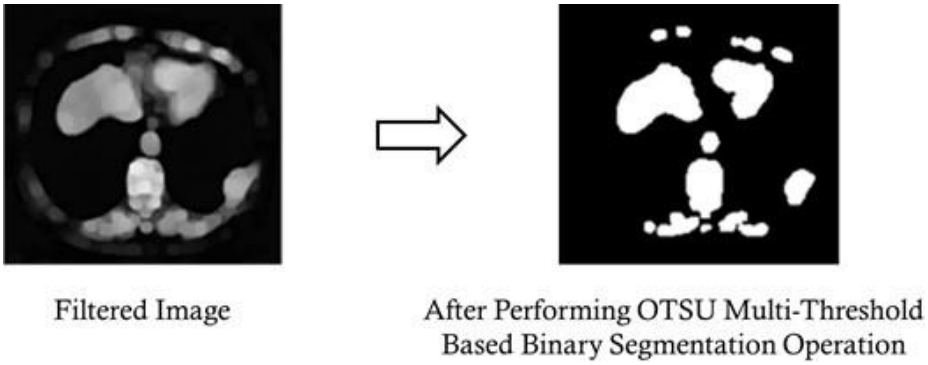

Fig. 4.6. Segmentation Operation Applied on Enhanced Lung MRI

4.4 Postprocessing

Phase-2 of the conventional IP technique consists of feature extraction and classification or identification steps. In post-processing features are extracted to identify the ROI (i.e., lung tumor or lung cancer region in the lung MRI), and the task may end with/without classification.

4.4.1 Feature Extraction

Feature selection involves the process of selecting the most relevant features, closely tied to dimensionality reduction. Its aim is to identify important features in the dataset for further operations while discarding irrelevant and redundant ones. By reducing the dimensionality of the data, feature selection enables data mining algorithms to operate more efficiently and quickly. There are two methods for selecting necessary features from a large feature set. The first method, known as feature extraction, involves considering the original feature space and mapping the most effective features to a lower-dimensional subspace. The second method, referred to as feature choice or band choice, entails selecting a small subset of features that adequately represent the classes.

Conventional supervised feature selection methods employ scoring functions or metrics to assess various feature subsets and retain only those relevant to the decision class of the data being analyzed. However, in numerous data mining scenarios, decision class labels are frequently unknown or incomplete, underscoring the significance of unsupervised feature selection. Unsupervised learning does not involve decision class labels.

In lung cancer analysis, a radiologist leverages their observations and expertise, considering a variety of features such as circularity, diameter, area, compactness, major axis length, minor axis length, and more. In this study, same features were incorporated for its identification of lung ROIs.

4.4.2 Identification and Classification

Classification, involves categorizing these identified objects or patterns into predefined groups or classes. All the lung MRI images were labeled by a radiologist into cancerous or non-cancerous lung image. Following IP approach cancerous regions were extracted using different feature characteristics. Thereby, RIOs were not available in non-cancerous lung MRI, on the contrary, lung tumor regions or ROIs were available in cancerous lung MRI. The accuracy of the conventional IP approach was determined by counting the number of correctly identified cancerous regions and dividing it by the total number of cancerous regions in all the lung MRI images. SVM which is a binary classifier was also employed to classify the lung MRI which gave False Positive (FP) and False Negative (FN) results.

Figure 4.7 shows the results of applying feature extraction operation on segmented lung MRI. In the figure, the picture on the left is the segmented image on which feature selection criterion are applied. After that other region are eliminated keeping only the lung tumor which is shown in the middle image. The rightmost image shows the identified region which has been colored and put on the original image.

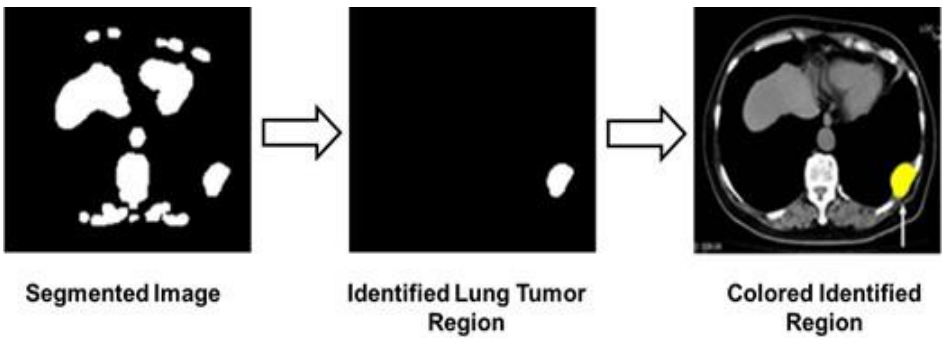

Fig. 4.7. Identified Lung Cancer Region is Colored and Placed over the Original Lung MRI

4.4.3 Summary

A conventional IP generally consists of preprocessing (Phase-1) and postprocessing (Phase- 2). Preprocessing includes enhancement, filtering and segmentation whereas postprocessing includes feature extraction and classification. As general practice preprocessing starts with enhancement of the input image. In this research study filtering has been used as the first step of the IP approach. There are different types of filters which are used for eliminating noise. Some of the researchers used more than one filter one after another. In this research work morphological opening which includes erosion followed by dilation has been used as the first step of the IP. Morphological opening is a process that eliminates small objects and thin lines from an image while maintaining the shape and size of larger objects. Thereafter, median filter which has the capability to preserve the region edges is applied on the lung MRI. Histogram processing which includes histogram equalization and histogram normalization is popular method used for enhancement of input image. Histogram equalization can be performed locally and globally. Since the global histogram equalization approach often fails to enhance smaller regions, in this research work local enhancement operation has been used to enhance the lung MRI. Between single threshold level and multi-threshold level of Otsu segmentation, Otsu Multi-threshold-based approach has been used in segmentation phase. It reduces over and under segmentation. Circularity, diameter, area, compactness, major axis length, minor axis length, and other features are among the key parameters utilized by radiologists to identify lung nodules. Same features have been used for identifying ROI in lung MRI. The accuracy of the system is determined by calculating the proportion of correctly identified lung nodules using the conventional IP technique, relative to the total number of lung nodules present in all the MRIs.

CHAPTER 5

MACHINE LEARNING STEPS FOR LUNG CANCER DETECTION

5.1 Introduction

Early detection of lung cancer is paramount for effective treatment and improved chances of survival. In recent years, ML techniques have shown great potential in detecting and diagnosing lung cancer from medical images such as MRI and CT scans. This chapter starts with basic description of ML concepts. It describes supervised and unsupervised learning and applicable technique for lung nodule detection using lung MRI. Thereafter it describes the architecture, functionality and usability of various ML algorithms for identifying lung cancer. It furnishes a comprehensive depiction of each step, encompassing various techniques and algorithms employed at each stage.

5.2 Classification Technique in Image Processing

Classification in IP involves categorizing an image or its components into various classes based on specific criteria or features. It finds applications in object recognition, face recognition, and medical image analysis. In medical image analysis, classification assists in identifying different tumor or lesion types in medical images, aiding in diagnosis and treatment planning. Various classification techniques are employed, including linear classifiers like logistic regression and support vector machines (SVMs), as well as non-linear classifiers such as decision trees and neural networks.

5.2.1 Supervised Learning

Supervised Learning is a machine learning paradigm where a model is trained on labeled data, where the desired output is known for each input. In the domain of lung cancer detection, supervised learning can be applied to train a model to detect lung nodules, which are small masses or lesions in the lung that may indicate the presence of cancer.

In supervised learning for lung nodule detection, the model is trained on a dataset of medical images, such as lung MRI or CT scans, where the images are labeled to indicate the presence or absence of lung nodules. The images are preprocessed to enhance features that may be indicative of nodules, such as texture and shape, and these features are used as inputs to the model. Several types of models can be used for lung nodule detection, such as CNNs, SVMs, and random forests. CNNs have demonstrated notable success in this application owing to their capability to directly learn intricate features from the image data.

The model's performance is assessed using metrics like sensitivity, specificity, and accuracy, and further refinement can be achieved through techniques like cross-validation and hyperparameter tuning. Supervised learning for lung nodule detection has exhibited promising outcomes in early lung cancer detection, often achieving high sensitivity and specificity in various studies. However, a limitation lies in the necessity for ample labeled data to effectively train the model, a process that can be both time-consuming and costly.

5.2.2 Unsupervised Learning

Unsupervised Learning is a machine learning approach wherein a model is trained on unlabeled data without predefined desired outputs. In the context of lung cancer detection, unsupervised learning can be employed to identify lung nodules, which are small masses or lesions in the lung potentially indicative of cancer. In unsupervised learning for lung nodule detection, the model is trained on a dataset of medical images, such as lung MRI or CT scans, lacking labels indicating the presence or absence of lung nodules. The images are preprocessed to enhance features that may be indicative of nodules, such as texture and shape, and these features are used as inputs to the model. Various

unsupervised learning models can be applied to lung nodule detection, including clustering algorithms such as k-means clustering, Gaussian mixture models, and hierarchical clustering. These algorithms group similar features together and can help to identify patterns or clusters that may be indicative of lung nodules.

The model's performance is assessed using metrics like clustering accuracy and silhouette coefficient, and further optimization can be achieved through techniques like cross-validation and hyperparameter tuning.

5.2.3 Applicable Technique for Lung Cancer Detection

Both supervised and unsupervised learning techniques offer their own advantages and limitations in the context of lung cancer detection. Supervised learning is a more applicable technique for lung cancer detection when there is a labeled dataset available, where each instance in the dataset is labeled with a class (i.e., cancer or non-cancer). In this case, the model can be trained to learn the relationship between the input features (e.g., medical image features) and the corresponding class labels. This approach has shown good performance in several studies and can help identify lung nodules with high accuracy and reliability (Shravya and Rajesh 2019).

On the other hand, unsupervised learning is more applicable when there is no labeled dataset available or when the dataset is small. In this case, the model can be trained to discern patterns or clusters within the input data without prior knowledge of the classes. While this approach may aid in identifying lung nodules that could potentially be overlooked by other detection methods, it might necessitate a substantial amount of unlabeled data and could exhibit lower accuracy compared to supervised learning.

In summary, the choice of the applicable technique for lung cancer detection depends on the availability of labeled data, the size of the dataset, and the specific requirements of the application.

5.3 CNN for Lung Cancer Detection

ML algorithms have exhibited promising outcomes in lung cancer detection. Among the most widely used ML algorithms for this purpose is the CNN. These algorithms utilize a blend of convolutional layers and pooling layers to extract features from the images. The extracted features are subsequently inputted into a fully connected layer, facilitating the final classification decision.

5.3.1 Architecture of CNN

CNNs, known as Convolutional Neural Networks, are a type of deep neural network frequently employed for tasks such as image recognition and classification, which involve analyzing visual data. They learn by processing training data, identifying patterns and features within the data pertinent to the task at hand. Diagrams illustrating the feature extraction and classification, as well as the layout of a CNN, are presented in Figure 5.1 and Figure 5.2, respectively.

CNNs consist of four primary types of layers: convolutional layers, pooling layers, fully-connected layers, and the output layer. In addition to these layers, several other crucial parameters play significant roles, including dropout, activation functions, and loss functions, which are elaborated below:

5.3.1.1 Convolutional layers

The convolutional layer serves as the fundamental component of a CNN, initiating the process of feature extraction from the input image. This layer applies a series of filters to the input image, performing the mathematical operation of convolution. This operation involves sliding a filter of size MxM across the input image, computing the dot product between the filter and the corresponding image patch at each position. The resulting output, known as a feature map, encapsulates information regarding the image's characteristics such as corners, edges, textures, or other visual patterns. These feature maps are subsequently passed to subsequent layers to capture additional features of the input image. The input image may traverse through multiple convolutional layers, each equipped with a set of filters tasked with extracting distinct features from the image.

5.3.1.2 Pooling layers

Following the Convolutional layer, a Pooling layer is typically employed. This layer serves to downsize the convolved feature map, thereby reducing computational complexity and extracting salient features. Various types of Pooling operations are performed, such as Max pooling, Average pooling, and Sum pooling, depending on the chosen method. These operations essentially condense the features generated by the convolutional layer. Max pooling, a prevalent technique in CNNs, retains the maximum value within each pool (a small segment of the feature map) while discarding the rest. Meanwhile, Average pooling computes the average of the elements, and Sum pooling calculates the total sum of the elements within a predefined image section. The Pooling layer acts as an intermediary between the Convolutional layer and the Fully connected layer.

5.3.1.3 Fully connected (FC) layers

Following several convolutional and pooling layers, the network typically concludes with one or more FC layers. These layers establish connections between every neuron from the preceding layer and those in the subsequent layer, resembling a conventional artificial neural network. The output of the final pooling layer is flattened into a vector and forwarded through one or more FC layers, where mathematical operations are performed on the input data to generate an output vector. Subsequently, this output vector is compared to the ground truth labels of the training data using a loss function, and the network's weights are adjusted via backpropagation to minimize the loss. FC layers are tasked with high-level reasoning and decision-making based on the extracted features.

5.3.1.4 Dropout

In order to regularize the network and mitigate overfitting, dropout layers can be incorporated. Overfitting arises when a model performs well with the training data but struggles with new data. To counteract overfitting, some neurons are omitted from the neural network during the training phase, typically dropping out around 30% of the nodes randomly from the network.

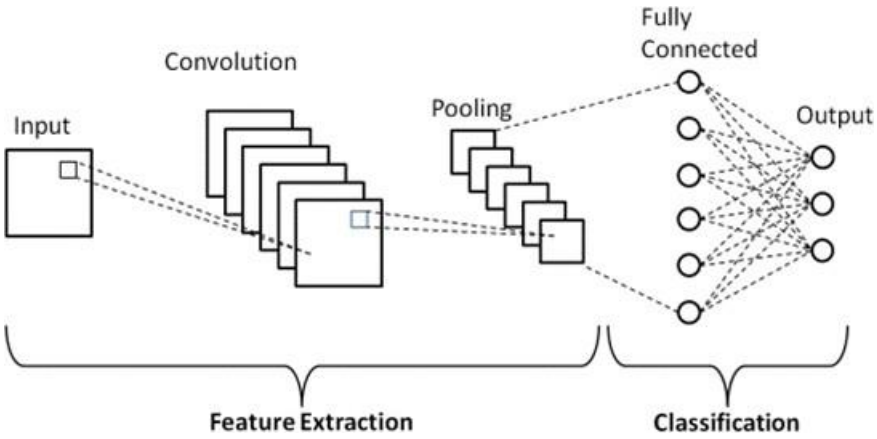

Fig. 5.1. Feature Extraction and Classification Diagram of CNN, Adapted from (Sharma, Gautam, and J. Singh 2023)

5.3.1.5 Activation Function

An activation function is typically applied element-wise to the output of the convolutional layer, serving as one of the most critical parameters in a CNN model. It dictates which information within the model should be activated in the forward direction and which should not. By introducing non-linearities, the activation function enables the network to comprehend more intricate relationships among features. Several commonly used activation functions include ReLU, Softmax, tanH, and Sigmoid functions, each with specific applications. For binary classification CNN models, sigmoid and softmax functions are often recommended, while softmax is typically employed for multiclass classification.

5.3.1.6 Loss Function

A loss function quantifies the disparity between the predicted output and the ground truth labels, commonly chosen as cross-entropy loss for classification tasks. Throughout training, the network's parameters are tuned to minimize this loss function, employing optimization techniques such as stochastic gradient descent (SGD) or its variants.

5.3.1.7 Output Layer

The final layer of the CNN is the output layer, responsible for generating the network's predictions. The number of neurons in this layer aligns with the number of classes in the classification task. For instance, in an image classification scenario with ten classes, the output layer would typically comprise ten neurons, each employing an appropriate activation function (such as softmax for multi-class classification) to produce class probabilities.

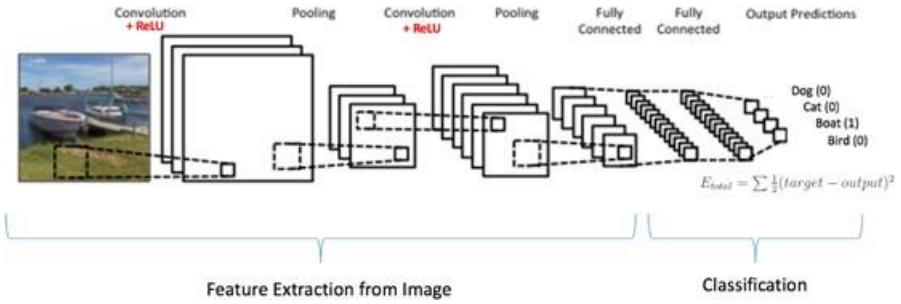

Fig. 5.2. Diagrammatic Layout of CNN, Adapted from (Pavan 2020)

5.3.2 Training

The training of a CNN is a crucial step in building an effective and accurate deep learning model for computer vision tasks. The complete network undergoes training utilizing a substantial dataset of labeled images. Throughout this process, the network acquires the capability to identify pertinent patterns and features within the input data concerning the designated task. By iteratively adjusting the weights in the network via backpropagation, the aim is to minimize loss and enhance the accuracy of the network's predictions. Over successive iterations, the CNN progressively discerns relevant features and patterns in the input data, eventually attaining high accuracy in image classification and identification. A concise outline of the CNN training procedure is outlined below:

5.3.2.1 Data Preparation

The initial stage of the training procedure involves collecting and preparing a substantial quantity of pertinent labeled images. Subsequently, the dataset is partitioned into training and validation sets, facilitating the training and validation of CNN performance throughout the training phase.

5.3.2.2 Architecture Design

Next, the architecture of the CNN is defined by specifying the number and type of layers where each layer applies specific operations to the input data, extracting increasingly complex features as the data flows through the network.

5.3.2.3 Initialization

The parameters of the network, encompassing the weights and biases of each layer, are initialized either randomly or with pre-trained weights sourced from another model.

5.3.2.4 Forward Propagation

Throughout the training process, the input images undergo successive processing via multiple layers employing convolution, activation functions (e.g., ReLU), pooling, and various other operations.

5.3.2.5 Loss Calculation

A loss function is employed to measure the disparity between the predicted output and the actual labels, with the objective of minimizing this loss throughout the training process.

5.3.2.6 Backpropagation

This step is crucial in the CNN training process as it entails calculating the gradients of the loss function concerning the network's parameters. These gradients denote the direction and magnitude of the parameter adjustments necessary to minimize the loss. Achieving this involves applying the chain rule of derivatives to propagate the gradients backward through the layers.

5.3.2.7 Parameter Updates

The gradients obtained from backpropagation are utilized to update the network’s parameters via an optimization algorithm. This updating process fine-tunes the weights and biases of the network iteratively to minimize the loss function.

5.3.2.8 Iterative Training

Steps starting from Forward Propagation to Parameter Updates are repeated multiple time for each epoch. The network gradually learns to better represent and classify the input data as the training progresses.

5.3.2.9 Validation and Evaluation

Periodically, the trained CNN is evaluated using the validation set to monitor its performance on unseen data. This helps identify potential overfitting, which occurs when the model performs well on the training set but poorly on new data, and allows for adjustments in the training process or model architecture.

5.3.2.10 Testing and Deployment

Once the training is complete, the trained CNN can be tested on a separate test set to assess its generalization performance. If the model performs satisfactorily, it can be deployed for real-world applications, where it can process new, unseen images and make accurate predictions.

5.3.3 Testing

Testing a CNN is an essential step in evaluating its performance and assessing its ability to generalize to unseen data. After training a CNN on a labeled dataset, the testing phase helps to measure the model's accuracy and robustness. Testing involves evaluating its performance on a separate set of data that was not used for training or validation. Below is a brief overview of the CNN testing process:

5.3.3.1 Test Dataset

During the testing process, a separate dataset, distinct from the training and validation sets, is employed. This dataset comprises labeled images that the model has not encountered during its training phase.

5.3.3.2 Preprocessing

Similar to the training phase, the test images may require preprocessing that were applied during training, before feeding them into the CNN.

5.3.3.3 Forward Propagation

The test images are forwarded through the trained CNN in a forward propagation step, resulting in predictions for each image.

5.3.3.4 Prediction Evaluation

The predicted outputs produced by the CNN are compared with the ground truth labels of the test dataset. Performance evaluation metrics like accuracy, precision, recall, and F1 score can then be calculated to assess the effectiveness of the model.

5.3.3.5 Performance Analysis

CNN's performance is assessed by analyzing the evaluation metrics. A high accuracy and consistent performance across different classes indicate that the model is generalizing well and making accurate predictions. On the other hand, poor performance may suggest issues like overfitting, underfitting, or biases in the dataset.

5.3.3.6 Fine-tuning and Hyperparameter Tuning

If the CNN's performance on the test dataset is unsatisfactory, further optimization may be required. Fine-tuning the model by adjusting hyperparameters, modifying the network architecture, or increasing the amount of training data can improve the CNN's performance.

5.3.3.7 Deployment and Real-world Testing

Once the CNN performs well on the test dataset, it can be deployed for real-world applications. Continuous monitoring and evaluation of the CNN's performance in real-world scenarios may be necessary to ensure its effectiveness over time.

5.3.4 Usability of CNN

CNNs have demonstrated significant value and effectiveness across various applications, particularly in the realm of computer vision. Their usability extends beyond some specific applications, with ongoing research continuously exploring new possibilities. CNNs have demonstrated their effectiveness in various domains, making them a versatile and widely adopted tool for visual data analysis and understanding. Some of them are: Image Classification, Object Detection, Semantic Segmentation, Feature Extraction, Face Recognition, Video Analysis,Transfer Learning etc.

5.4 SVM for Lung Cancer Detection

Another ML algorithm that has been used for lung cancer detection is SVM. It is a supervised machine learning algorithm commonly employed for both binary and multiclass classification tasks.

5.4.1 Architecture of SVM

The architecture of SVM for image classification is relatively straightforward. In the context of image classification, SVMs utilize a set of training data to create an optimal hyperplane that effectively separates the data into different classes. The algorithm can then use this hyperplane to classify new data. An overview of the SVM architecture for image classification is described below:

5.4.1.1 Preprocessing

The MRI images are preprocessed to enhance the relevant features and reduce noise. Common preprocessing steps include resizing, normalization, denoising, and image registrationetc.

5.4.1.2 Feature Selection

As with any machine learning-based image classification approach, the initial step involves extracting relevant features from the images. In traditional SVMs, feature extraction is typically done manually, where each image is transformed into a fixed-size

feature vector representing various characteristics of the image. In this step, relevant features are extracted from the preprocessed MRI images. Feature extraction methods could include techniques like histogram of gradients (HOG), texture analysis, wavelet transforms.

5.4.1.3 Data Representation

After the features are extracted, each image is represented as a feature vector in a high-dimensional space, where the dimensionality corresponds to the number of features extracted from each image.

5.4.1.4 Labeling and Data Preparation

Each image in the dataset is labeled with a class. SVM, typically a binary classifier, operates on datasets with only two classes.

5.4.1.5 Data Splitting

The dataset is partitioned into training, validation, and testing subsets to accurately assess the model's performance.

5.4.1.6 Kernel Trick (Optional)

In instances where the data is not linearly separable in the feature space, a kernel trick can be employed to map the data into a higher-dimensional space where linear separability is achieved. Popular kernel functions include the Radial Basis Function (RBF) kernel, polynomial kernel, and sigmoid kernel. It is important to note that traditional SVMs have some limitations when dealing with high-dimensional and complex datasets like images.

5.4.2 Learning

The SVM classifier is trained using the extracted features and their corresponding labels (cancer or non-cancer). It aims to find the optimal hyperplane that effectively separates the feature vectors of different classes while maximizing the margin between them. The margin represents the distance between the hyperplane and the nearest data points

(support vectors) of each class. The SVM algorithm is designed to identify the hyperplane that maximizes this margin. Hyperparameters (e.g., kernel type, regularization parameter) may need to be optimized to achieve better performance. Techniques like cross-validation can be employed to tune these hyperparameters.

5.4.3 Testing

After training, the SVM can be deployed to classify new, unseen images. Each new image undergoes feature extraction, similar to the training images. The SVM then assigns a class label based on the position of the feature vector relative to the learned hyperplane in the feature space. Performance evaluation metrics such as accuracy, precision, recall, and F1 score are computed on a separate testing dataset to assess how well the SVM generalizes to unseen data.

5.4.4 Usability of SVM

SVMs are popular across various machine learning applications owing to their versatility and effectiveness in both binary and multiclass classification tasks. They are particularly useful for small to medium-sized datasets with well-defined features, and their effectiveness can be enhanced by combining them with appropriate feature extraction techniques and parameter tuning. SVMs have demonstrated success in a wide array of domains, including text and image classification, bioinformatics, finance, remote sensing, and anomaly detection.

5.5 Summary

Early detection of lung cancer is the most essential requirement for successful cancer treatment and recovery. ML can play vital role in identifying lung nodules from medical images. Existing ML algorithms can be broadly categorized into two categories: supervised and unsupervised learning. Supervised learning performs well on labeled data, while unsupervised learning is more suitable for situations where the dataset is small and unlabeled. Two of the supervised ML algorithms: CNN and SVM have been discussed in details in this chapter considering their suitability for feature selection and image. CNN, a

type of deep neural network primarily employed for image recognition and classification, comprises four types of layers: convolutional layers, pooling layers, fully-connected layers, and output layer. In addition to these layers, other vital parameters include dropout, activation function, and loss function. At first the architecture of the CNN is designed specifying the number and type of layers which are initialized with weights and biases for each layer. The entire dataset is partitioned into three subsets for training, validation, and testing purposes. During the training phase, the CNN incrementally learns to identify features and patterns within the input data. Periodically, the trained CNN is evaluated using the dataset for validation and through backpropagation method parameters are updated to increase classification accuracy. In the testing phase, the CNN's performance is assessed on unseen data using evaluation metrics such as accuracy, precision, recall, and F1 score. If the CNN's performance on the test dataset is unsatisfactory, further optimization is pursued by modifying its architecture, adjusting hyperparameters, or increasing the training data. Similarly, SVM, as a supervised ML algorithm primarily functioning as a binary classifier, follows a similar methodology. The dataset is divided into training, validation, and testing sets, and a suitable SVM model for classification is constructed. Both CNN and SVM are versatile and effective ML algorithm which are widely used in different domains. CNN has demonstrated its effectiveness in the domains of computer vision and image classification.

CHAPTER 6

EXPERIMENTAL RESULTS

6.1 Introduction

This chapter presents the findings and analysis of the experimental results obtained from both the Conventional IP Approach and the Machine Learning Approach. The experiments were designed conforming to the research approach and research activities described in Chapter 3 to address the research objectives. The experimental results shed light on the outcomes of various tests and measurements, enabling a deeper exploration of the research questions and paving the way for meaningful discussions and conclusions in the subsequentchapter.

In this chapter the steps followed in Conventional IP Approach has been described first. The resulted images of using different filters and enhancement methods have been shown in different table. Then the results of ML Approach have been described.At the end of this chapter a comparison has been drawn between the findings of the two different approaches.

6.2 Preprocessing Stage

6.2.1 Filtering

Existing literature review indicates that different types of filters are used for noise elimination, of which Mean and Median filters are commonly used. In this study Morphological Opening has been used followed by Median filter for eliminating noises while preserving the edges. Figure 6.1 shows resultant images after performing different

filter operations. Kernel Size for Mean and Median filters were considered 3, 5 and 7, among which 7x7 showed the best results. Morphological operation, in particular morphological opening wasused with a diamond shaped structuring element of size 5x5 for as a fact that it provided thebest results.

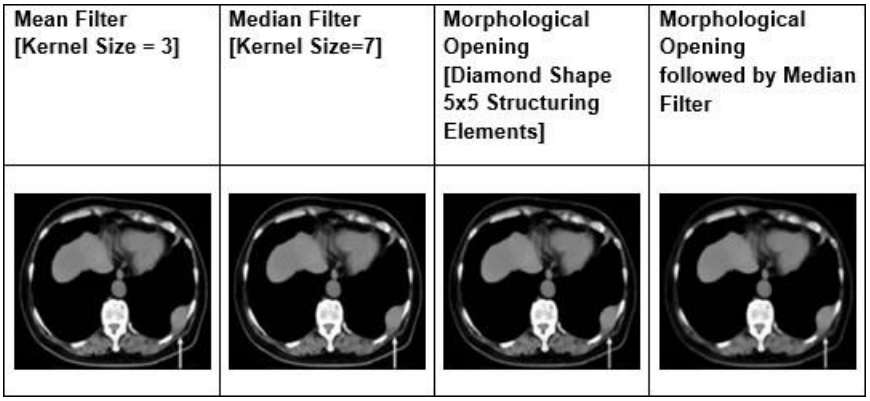

Fig. 6.1. Different Types of Filter Operation Applied on Lung MRI

6.2.2 Enhancement

In this study, two types of enhancement operations were explored: Histogram Equalization and Histogram Normalization. Specifically, local histogram equalization was implemented instead of global histogram equalization, yielding improved results. Figure 6.2 shows the resultant images of applying enhancement operation on the filtered lung MRI.

6.2.3 Segmentation

Typically, threshold-based segmentation is recommended for gray-scale images. In this research study Otsu multi threshold-based approach has been used in Segmentation stage. Resultant images are shown in Figure 6.3.

6.3 Postprocessing Stage

This stage consists of Feature Extraction and Identification of ROIs which are the lung nodules or cancerous region within the lung. Identified lung nodule from segmented image,

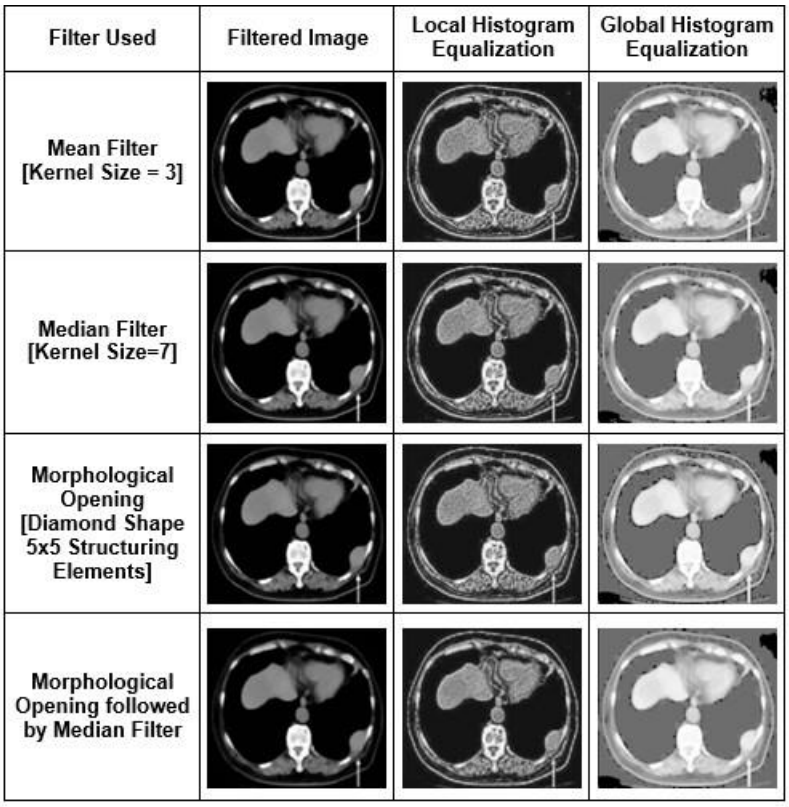

Fig. 6.2. Enhancement Operation Applied on Filtered Lung MRI

after applying different types of filters and enhancement technique is shown in Figure 6.3.

6.4 Results of Conventional Image Processing Approach

In conventional IP approach total 543 MRI images were used of which 164images were of normal lungs and 379 images contained cancerous regions. By following the mentioned IP approach, out of total 403 cancerous regions in 379 MRI images, 388 regions were identified correctly and the accuracy found as 96.28%. To the radiologists an accuracy of 95% and above is an acceptable limit and the result of the conventional IP approach is very well above the required accuracy range. The Precision (accuracy of positive predictions) and Recall (Sensitivity or True Positive Rate) of the IP method are 96.41% and 92.08% respectively. The value of F1 score is 0.942 which suggests that the IP method that has been followed for classifying cancerous and noncancerous region is good at correctly classifying while minimizingFP and FN. The Confusion Matrix and F1 score of the model are presented in Table 6.1.

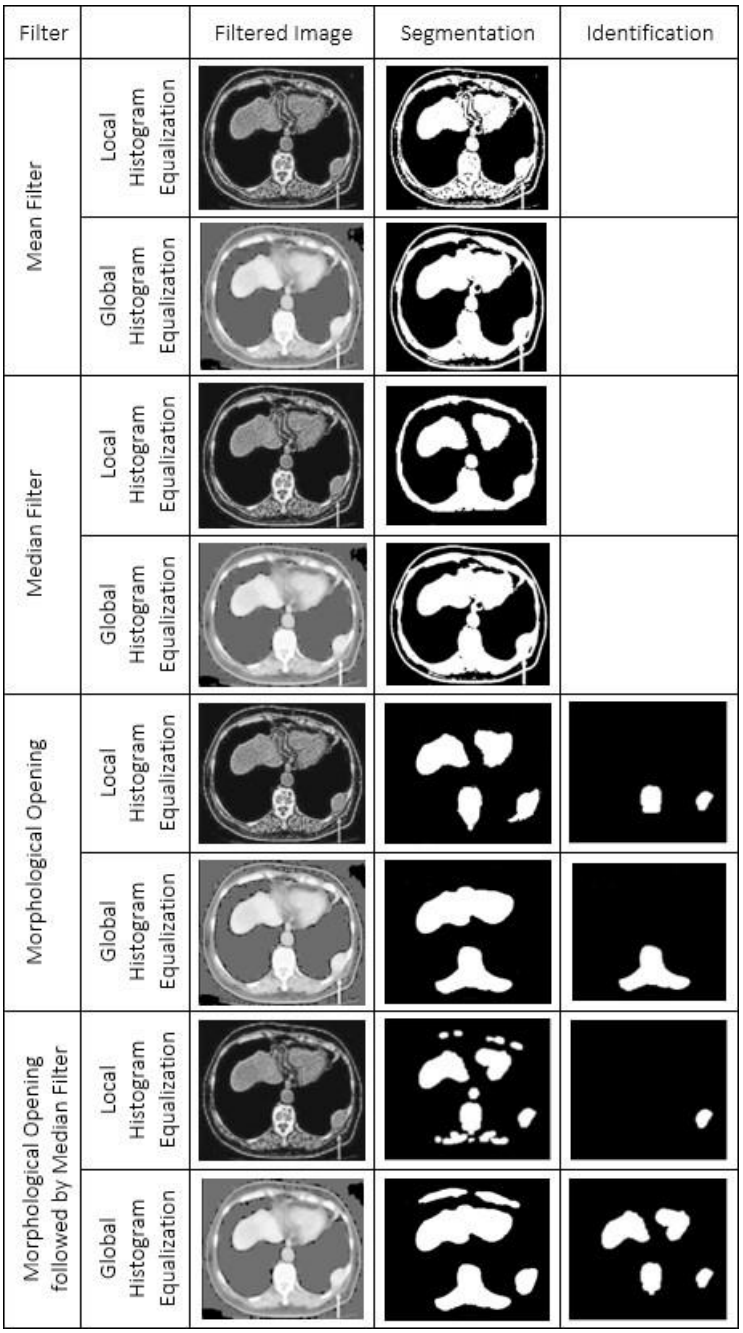

Fig. 6.3. Segmentation and Identification of Lung Nodules from Lung MRI

6.5 Machine Learning Approach

Conventional IP approach gives us satisfactory results. We also aimed to assess the performance of the machine learning algorithm in identifying lung cancer and subsequently make a comparison. In the literature review we found that Two ML algorithm: CNN and SVM have been used widely in lung cancer identification. As such,

Table 6.1: Confusion Matrix and Metrics

Metrics	Confusion Matrix				Scores		
	TP	**FP**	**TN**	**FN**	**Precision**	**Recall**	**F1 Score**
Value	349	13	151	30	96.41%	92.08%	0.942

we have picked up these two algorithms for using in this research. Classification steps are discussed below:

6.5.1 Data Augmentation

Typically, ML approach requires a good number of images whereas this study dataset is small (only 543). Data augmentation was carried out to create a dataset of 10,000 images. Thereafter the test was conducted in three phases. In 1st phase, among 10,000 images 60% were used for training and 30% for testing and 10% for validation. In 2nd and 3rd phase the data distribution were 70%, 20%, 10% and 80%, 10%, 10% respectively. In all the cases the accuracy was maximum in the 3rd phase where data distribution for training, test and validation were 80%, 10%, 10% respectively.

6.5.2 Convolutional Neural Network (CNN)

The CNN, or Convolutional Neural Network, stands out as the premier neural network model for image classification tasks. This feed-forward neural network architecture is specifically designed to excel in detecting and categorizing objects within images, consistently yielding superior performance in computer vision applications.

6.5.2.1 Architecture of CNN Model

VGG-16 has been used which is a 16-layer CNN model. The architecture of the model is shown in Figure 6.4. We have used mask size 7x7, filters 512. It's a 2-D image, so conv2D has been used, mask 7x7, filters 240. At the end flatten and dense layer have been used. In our case the algorithm showed an accuracy of 96.55% which is very well more than the acceptable range.

6.5.2.2 Training vs Validation Loss

Figure 6.5 shows Training vs validation loss for each epoch. Training loss is in blue color and validation loss in yellow color.

Layer (type)	Output Shape	Param #
vgg16 (Functional)	(None, 7, 7, 512)	14714688
conv2d_4 (Conv2D)	(None, 7, 7, 240)	6021360
dropout_4 (Dropout)	(None, 7, 7, 240)	0
conv2d_5 (Conv2D)	(None, 7, 7, 120)	1411320
dropout_5 (Dropout)	(None, 7, 7, 120)	0
conv2d_6 (Conv2D)	(None, 7, 7, 60)	352860
batch_normalization_1	(Batch (None, 7, 7, 60)	240
dropout_6 (Dropout)	(None, 7, 7, 60)	0
flatten_1 (Flatten)	(None, 2940)	0
dense_1 (Dense)	(None, 1)	2941
Total params: 22,503,409 Trainable params: 7,788,601 Non-trainable params: 14,714,808		

Fig. 6.4. Architecture of the Proposed CNN Model

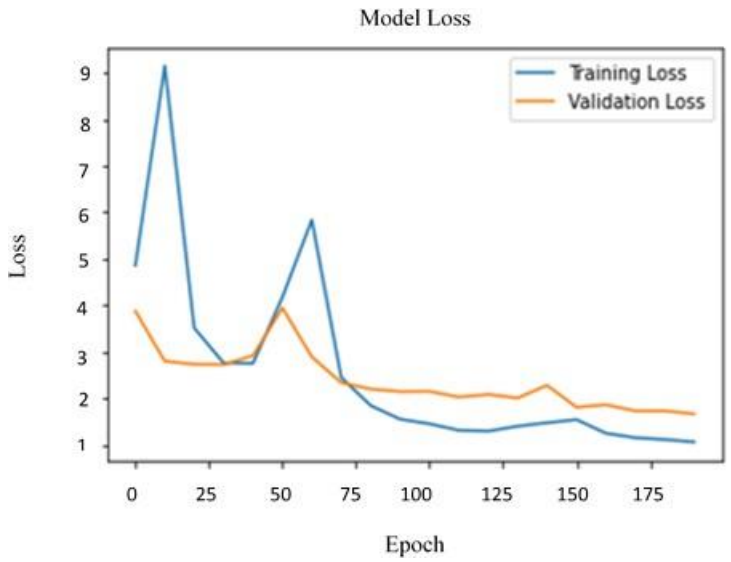

Fig. 6.5. Training vs Validation Loss for Each Epoch

6.5.2.3 Training vs Validation Accuracy

Figure 6.6 illustrates the training and validation accuracy graph. We find that Training accuracy is higher than the validation accuracy. It also proves that the model is working well.

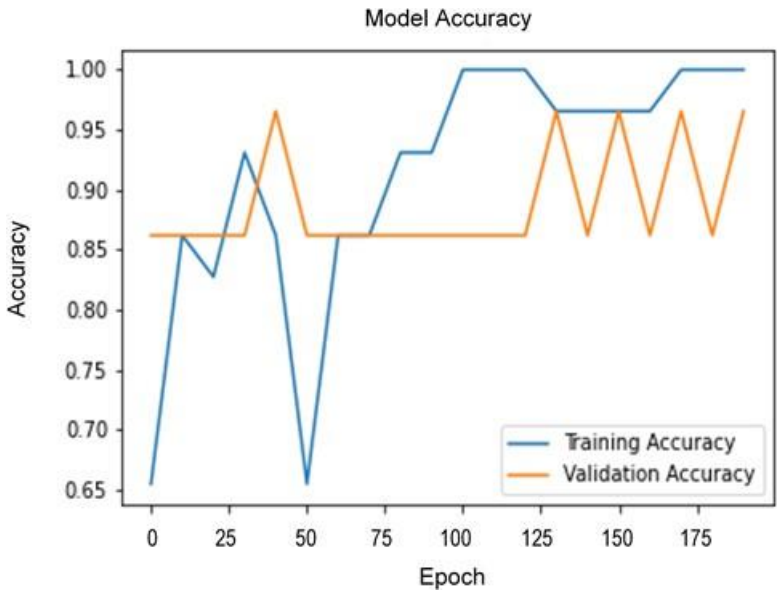

Fig. 6.6. Training Accuracy vs Validation Accuracy

6.5.3 Support Vector Machine (SVM)

SVM is a binary or a true class classifier. In our case it showed an accuracy of 90.5%. Training, Testing and F1 Score are 0.968, 0.92 and 0.92 respectively. Total number of Misclassified Samples and 208.

6.6 Comparative Results of Proposed Approach and Available Works

The dataset used in this research study has been applied on some prominent past researches and their accuracies were found out. A table showing comparative results of existing approaches with the proposed approach is shown at Table 6.2. In the table, top three research are on using Conventional IP approach which shows that none of existing research has the accuracy more than 94%. In case of Classification Approach the accuracy is below 95%. The proposed Conventional IP approach has the accuracy more than 95% and Classification Approach has an accuracy a bit higher than Conventional IP approach.

6.7 Summary

In ML approach two widely used ML algorithms namely CNN and SVM have been used. The accuracy of CNN and SVM are 96.55% and 90.5% respectively. CNN is a popular neural network algorithm which is used to perform computer vision task and it has proved its worthiness for identifying lung cancer from MRI image.

Table 6.2: Accuracy comparison of different approaches

Ser	Proposed by	Approach Used	Accuracy (approx.)
1.	(Kannan and Naveen 2020)	Conventional Image Processing	67%
2.	(Bari et al. 2019)	Conventional Image Processing	94%
3.	(Cieszanowski et al. 2016)	Conventional Image Processing	89.5%
4.	(Prabhpreet Kaur, G. Singh, and Parminder Kaur 2020)	ML Approach	92%
5.	(Keerthana, Thamilselvan, and Sathiaseelan 2016)	ML Approach	94.5%
6.	(Mohammed and Çinar 2021)	ML Approach	94.5%
7.	[Proposed]	Conventional Image Processing	96.28%
8.	[Proposed]	ML Approach	96.55%

CHAPTER 7

CONCLUSION

7.1 Introduction

This chapter outlines the key findings and contributions of the thesis, followed by an emphasis on the contributions, limitations, and potential avenues for future research.

7.2 Main Outcomes

Two different approaches have been used is this research work to identify lung cancer. Approach one was to identify lung cancer using conventional IP with modification in its usual sequence and combination of steps. In this approach the final validationwas done by pulmonologist. It was feasible to precisely detect regions of lung cancer with an acceptable accuracy rate of 96.28% employing an IP approach.

Approach two was to identify lung cancer using ML algorithm. In this approach CNN and SVM have been used on MRI image. In this approach the accuracy of CNN and SVM were 96.55% and 90.5% respectively.

The difference in accuracy between conventional and ML (CNN) approach is 0.27%. But, to run ML model high end computers are required which are not always available to the radiologists or pulmonologist. So, for the doctors using conventional IP approach is more convenient.

The proposed image analysis approach offers ease of use and reduced complexity in terms of time and space. In contrast, the CNN model utilized in this study exhibits greater time complexity while delivering a higher accuracy rate. To have better accuracy CNN model can be used. But, for this computer with high computing capacity are required. But, still there are possibilities to increase the accuracy. In that case we need support of the radiologists and pulmonologist. The accuracy of lung cancer identification can be improved, if they can minimize their inter-observationvariability.

7.3 Thesis Contributions

MRI is free of ionizing radiation. That's why it is an appropriate alternative for the pediatric patients and for the patients requiring frequent follow-up of lung disease. But identification of lung cancer is a specialist job. So, identification of lung cancer using semi-automated or automated process will ease the job of the pulmonologist in lung diagnosis. Adoption of our research findings will help them to increase the accuracy of their findings from lung MRI which will enhance the reliability of their treatment plan for lung cancer patients.

7.4 Limitations of the Thesis

During the research study, it became evident that:

1. No benchmark dataset exists that can be utilized for comparing existing approaches for the detection of lung cancer from MRI images.

2. A substantial volume of data, which is a prerequisite for the second approach (ML Approach), was not available.

7.5 Future Work

This research has been conducted for the identification of Non-Small Cell Lung Cancer (NSCLC). In future, further research can be carried out for identification of Small Cell Lung Cancer (SCLC) and Bronchial Carcinoid Tumors. Again, identification of lung cancerous cell into benign or malignant category along with its stage identification can also be incorporated.

REFERENCES

Abdullah, Dakhaz Mustafa, Adnan Mohsin Abdulazeez, and Amira Bibo Sallow (2021). 'Lung cancer prediction and classification based on correlation selection method using machine learning techniques.' In: *Qubahan Academic Journal* 1.2, pp. 141–149.

ACS (2022a). *Tests for Lung Cancer*. URL: https://www.cancer.org/cancer/types/lung-cancer/detection-diagnosis-staging/how-diagnosed.html.

– (2022b). *What Is Cancer?* URL: https://www.cancer.org/cancer/understanding-cancer/what-is-cancer.html.

American Cancer, Society (2021). *Cancer Information and Resources*. URL: https://www.cancer.org/.

American Lung, Association (2022). *Warning Signs of Lung Disease*. URL: https://www.lung.org/lung-health-diseases/warning-signs-of-lung-disease.

Asuntha, A, A Brindha, S Indirani, and Andy Srinivasan (2016). 'Lung cancer detection using SVM algorithm and optimization techniques.' In: *J. Chem. Pharm. Sci* 9.4, pp. 3198–3203.

Bapure, Kusuma (2012). 'Automated image analysis for nuclear morphometry using h&e and feulgen stains in prostate biopsies.' PhD thesis. University of Illinois at Chicago.

Bari, Mehwish, Adeel Ahmed, Muhammad Sabir, and Sajid Naveed (2019). 'Lung cancer detection using digital image processing techniques: A review.' In: *Mehran University Research Journal of Engineering & Technology* 38.2, pp. 351–360.

Biederer, Juergen et al. (2017). 'Screening for lung cancer: does MRI have a role?' In: *European journal of radiology* 86, pp. 353–360.

Cancer Research, UK (2020). *How cells and tissues grow*. URL: https://www.cancerresearchuk.org/about-cancer/what-is-cancer/how-cancer-starts/how-cells-and-tissues-grow.

– (2022a). *CT scan*. URL: https://www.cancerresearchuk.org/about-cancer/tests-and-scans/ct-scan.

– (2022b). *MRI scan*. URL: https://www.cancerresearchuk.org/about-cancer/tests-and-scans/mri-scan (visited on 03/06/2023).

Cieszanowski, Andrzej et al. (2016). 'MR imaging of pulmonary nodules: detection rate and accuracy of size estimation in comparison to computed tomography.' In: *PLoS One* 11.6, e0156272.

De Bruijne, Marleen (2016). *Machine learning approaches in medical image analysis: From detection to diagnosis*.

Devarapalli, Retz Mahima, Hemantha Kumar Kalluri, and Venkatesulu Dondeti (2019). 'Lung cancer detection of CT lung images.' In: *International Journal of Recent Technology and Engineering* 7.5S4, pp. 413–416.

Entwistle, A (2004). 'Moderated histogram equalization, an automatic means of enhancing the contrast in digital light micrographs reversibly.' In: *Journal of microscopy* 214.3, pp. 272–286.

– (2005). 'A method for the blind correction of the effects of attenuation and shading in light micrographs based upon moderated histogram equalization.' In: *Journal of microscopy* 219.3, pp. 141–156.

Ferlay, Jacques, Isabelle Soerjomataram, Les Mery, and Freddie Bray (2020). *All cancers*. URL: https://gco.iarc.fr/today/data/factsheets/cancers/39-All-cancers-fact-sheet.pdf.

Fernandes, Jessnia, Nikeeta Simoes, Dominic Vaz, Saurav Tiwari, Amrita Naik, and Damodar Reddy Edla (2022). 'Prediction of malignant lung nodules in CT scan images using cnn and feature selection algorithms.' In: *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*. IEEE, pp. 218–224.

Gurcan, Metin N, Laura E Boucheron, Ali Can, Anant Madabhushi, Nasir M Rajpoot, and Bulent Yener (2009). 'Histopathological image analysis: A review.' In: *IEEE reviews in biomedical engineering* 2, pp. 147–171.

Hirsch, Franz Wolfgang et al. (2020). 'The current status and further prospects for lung magnetic resonance imaging in pediatric radiology.' In: *Pediatric Radiology* 50, pp. 734–749.

Hochhegger, B et al. (2011). 'MRI in lung cancer: a pictorial essay.' In: *The British journal of radiology* 84.1003, pp. 661–668.

Hussain, Lal et al. (2022). 'Lung cancer prediction using robust machine learning and image enhancement methods on extracted gray-level co-occurrence matrix features.' In: *Applied Sciences* 12.13, p. 6517.

Jiang, Hongyang, He Ma, Wei Qian, Mengdi Gao, and Yan Li (2017). 'An automatic detection system of lung nodule based on multigroup patch-based deep learning network.' In: *IEEE journal of biomedical and health informatics* 22.4, pp. 1227–1237.

Johns Hopkins, Medicine (2021). *Lung cancer types*. URL: https://www.hopkinsmedicine.org/health/conditions-and-diseases/lung-cancer/lung-cancer-types.

Kadam, Krishna (2022). 'Use of Machine Learning to Detect Lung Cancer.' In: *International Journal of Software Innovation (IJSI)* 10.1, pp. 1–12.

Kannan, V and V Jagan Naveen (2020). 'Detection of lung cancer using image segmentation.' In: *International Journal of Electrical Engineering & Technology (IJEET)* 2.11, pp. 7–16.

Kaur, Manpreet and Sunny Behal (2013). 'Study of Image Denoising and Its Techniques.' In: *International Journal* 3.1.

Kaur, Prabhpreet, Gurvinder Singh, and Parminder Kaur (2020). 'Classification and validation of MRI brain tumor using optimised machine learning approach.' In: *ICDSMLA 2019: Proceedings of the 1st International Conference on Data Science, Machine Learning and Applications*. Springer, pp. 172–189.

Keerthana, P, P Thamilselvan, and JGR Sathiaseelan (2016). 'Detection of Lung Cancer in MR Images by using Enhanced Decision Tree Algorithm.' In: *Int. J. of Control Theory and Appl. (IJCTA* 9.27, pp. 267–273.

Kelsey, Thomas W, Benedicta Caserta, Luis Castillo, W Hamish B Wallace, and Francisco Cóppola Gonzálvez (2010). 'Proliferating cell nuclear antigen (PCNA) allows the automatic identification of follicles in microscopic images of human ovarian tissue.' In: *Pathology and Laboratory Medicine International*, pp. 99–105.

Kiruthika, V and MM Ramya (2014). 'Automatic segmentation of ovarian follicle using K-means clustering.' In: *2014 fifth international conference on signal and image processing*. IEEE, pp. 137–141.

Kothari, Sonal, John H Phan, Todd H Stokes, and May D Wang (2013). 'Pathology imaging informatics for quantitative analysis of whole-slide images.' In: *Journal of the American Medical Informatics Association* 20.6, pp. 1099–1108.

Lamprecht, Michael R, David M Sabatini, and Anne E Carpenter (2007). 'CellProfiler™: free, versatile software for automated biological image analysis.' In: *Biotechniques* 42.1, pp. 71–75.

Landini, Gabriel and IE Othman (2003). 'Estimation of tissue layer level by sequential morphological reconstruction.' In: *Journal of microscopy* 209.2, pp. 118–125.

Lasker, Joseph M (2008). *A digital-signal-processor-based optical tomographic system for dynamic imaging of joint diseases*. Columbia University.

Li, Qiang and Jinghuai Gao (2013). 'Contourlet based seismic reflection data non-local noise suppression.' In: *Journal of Applied Geophysics* 95, pp. 16–22.

Liu, Xioqiu, Jinglu Tan, Iyad Hatem, and Barry L Smith (2004). 'Image processing of hematoxylin and eosin-stained tissues for pathological evaluation.' In: *Toxicology mechanisms and methods* 14.5, pp. 301–307.

Majib, Mohammad Shahjahan, TM Shahriar Sazzad, and Md Mahbubur Rahman (2020). 'A framework to detect brain tumor cells using mri images.' In: *2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*. IEEE, pp. 1–5.

Makde, Vipin, Jenice Bhavsar, Swati Jain, and Priyanka Sharma (2018). 'Deep neural network based classification of tumourous and non-tumorous medical images.' In: *Information and Communication Technology for Intelligent Systems (ICTIS 2017)-Volume 2 2*. Springer, pp. 199–206.

Maurie, Markman (2022). *Lung cancer types*. URL: https://www.cancercenter.com/cancer-types/lung-cancer/types.

Mayo Clinic, Researchers (2022). *Lung cancer*. URL: https://www.mayoclinic.org/diseases-conditions/lung-cancer/symptoms-causes/syc-20374620.

Mohammed, Shivan HM and Ahmet Çinar (2021). 'Lung cancer classification with convolutional neural network architectures.' In: *Qubahan Academic Journal* 1.1, pp. 33–39.

Nishu, Sunil A (2012). 'Quantifying the defect visibility in digital images by proper color space selection.' In: *International journal of engineering research and applications* 2.3, pp. 1764–1767.

Pavan, Sanagapati (2020). *A Simple CNN Model Beginner Guide !!!!!!* URL: https://www.kaggle.com/code/pavansanagapati/a-simple-cnn-model-beginner-guide (visited on 03/06/2023).

Perumal, S and Thambusamy Velmurugan (2018). 'Preprocessing by contrast enhancement techniques for medical images.' In: *International Journal of Pure and Applied Mathematics* 118.18, pp. 3681–3688.

Picut, Catherine A, Cynthia L Swanson, Kathryn L Scully, Vern C Roseman, Regina F Parker, and Amera K Remick (2008). 'Ovarian follicle counts using proliferating cell nuclear antigen (PCNA) and semi-automated image analysis in rats.' In: *Toxicologic pathology* 36.5, pp. 674–679.

Rajalaxmi, RR, S Kavithra, E Gothai, P Natesan, and R Thamilselvan (2022). 'A Systematic Review Of Lung Cancer Prediction Using Machine Learning Algorithm.' In: *2022 International Conference on Computer Communication and Informatics (ICCCI)*. IEEE, pp. 1–7.

Ratini, Melinda (2021). *Respiratory System*. URL: https://www.webmd.com/lung/.

Rodrigues, Murillo B et al. (2018). 'Health of things algorithms for malignancy level classification of lung nodules.' In: *IEEE Access* 6, pp. 18592–18601.

Rosen, Daniel G, Xuelin Huang, Michael T Deavers, Anais Malpica, Elvio G Silva, and Jinsong Liu (2004). 'Validation of tissue microarray technology in ovarian carcinoma.' In: *Modern pathology* 17.7, pp. 790–797.

RSNA, Editorial Team (2023). *PET/CT*. URL: https://www.radiologyinfo.org/en/info/pet (visited on 03/06/2023).

Sazzad, TM Shahriar, KM Tanzibul Ahmmed, Misbah Ul Hoque, and Mahmuda Rahman (2019). 'Development of automated brain tumor identification using MRI images.' In: *2019 International conference on electrical, computer and communication engineering (ECCE)*. IEEE, pp. 1–4.

Sertel, Olcay, Umit V Catalyurek, Hiroyuki Shimada, and Metin N Gurcan (2009). 'Computer-aided prognosis of neuroblastoma: Detection of mitosis and karyorrhexis cells in digitized histological images.' In: *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, pp. 1433–1436.

Sharma, Aanchal, Rahul Gautam, and Jaspal Singh (2023). 'Deep learning for face mask detection: a survey.' In: *Multimedia Tools and Applications*, pp. 1–41.

Shravya, T and T Rajesh (2019). 'Intelligent Prediction of Lung Cancer Via MRI Images using Morphological Neural Network Analysis.' In: *Int. Res. J. of Eng. and Technol.(IRJET)* 6.9, pp. 110–117.

Singh, Gur Amrit Pal and PK Gupta (2019). 'Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans.' In: *Neural Computing and Applications* 31, pp. 6863–6877.

Skodras, Angelos, Stamatia Giannarou, Mark Fenwick, Stephen Franks, Jaroslav Stark, and Kate Hardy (2009). 'Object recognition in the ovary: quantification of oocytes from microscopic images.' In: *2009 16th International Conference on Digital Signal Processing*. IEEE, pp. 1–6.

Temple Health, Editorial Team (2023). *PET/CT Scan*. URL: https://www.templehealth.org/services/pet-ct-scan (visited on 03/06/2023).

Thamilselvan, P and JGR Sathiaseelan (2016a). ‘An enhanced k nearest neighbor method to detecting and classifying MRI lung cancer images for large amount data.’ In: *Int. J. Appl. Eng. Res* 11.6, pp. 4223–4229.

– (2016b). ‘Early Detection of Cancer in MRI Lung Images using Classification and Regression Tree (CART) Method.’ In: *Int. J. of Control Theory and Appl.(IJCTA)* 9.26, pp. 397–405.

Tiwari, Arvind Kumar (2016). ‘Prediction of lung cancer using image processing techniques: a review.’ In: *Advanced Computational Intelligence: An International Journal* 3.1, pp. 1–9.

Van der Kwast, Theodorus H et al. (2010). ‘Variability in diagnostic opinion among pathologists for single small atypical foci in prostate biopsies.’ In: *The American journal of surgical pathology* 34.2, pp. 169–177.

Wang, Yi-Xiang J, Gladys G Lo, Jing Yuan, Peder EZ Larson, and Xiaoliang Zhang (2014). ‘Magnetic resonance imaging for lung cancer screen.’ In: *Journal of thoracic disease* 6.9, p. 1340.

APPENDIX A

ALGORITHM AND SOURCE CODES

Matlab Code for Conventional Image Processing

```
clear all;
close all;
I1 = imread('C:\SELF\MSc Thesis\Test Img\Mri\12.jpg');
%figure(1), imshow(I1);

I2 = medfilt2(V, [5 5]); % medain flter
I3 = adapthisteq(I2); % Contrast-limited adaptive histogram equalization (CLAHE)
%I4 = imbilatfilt(I2); % bilateral filter
%I5 = imnlmfilt(I2); % Non-local means filter
I4 = imadjust(I3); % Contrast Strecting For Thresholding
gt = graythresh(I4); % computes a global threshold T from grayscale image, usingOtsu's
method
I5 = imbinarize(I4,gt);
[Gmag, Gdir] = imgradient(I5, 'sobel'); % Calculate the gradient magnitude anddirection,
specifying the sobel gradient operator.
I6 = Gmag; % gradient magnitude
I7 = imclearborder(I6); %% Clear bordersI8 =
imfill(I7,'holes'); %% Fills holes
gt = graythresh(I8); %% Thresholding using graythreshI9 =
imbinarize(I8,gt);

se1 = strel('disk',3);
se2 = strel('disk',3);

I10 = imerode(I9,se1); % Erode image
I11 = imdilate(I10,se2); % Dilate image
I12 = immultiply(im2double(I5),I11);

LB = 0; UB = 5000;
I13 = xor(bwareaopen(I12,LB), bwareaopen(I12,UB)); %% Removes pixelsgreater than
2100

subplot(3,4,1); imshow(I1);
subplot(3,4,2); imshow(I2), title('Median Filter'); subplot(3,4,3);
imshow(I3), title('CLAHE'); subplot(3,4,4); imshow(I4),
title('imadjust'); subplot(3,4,5); imshow(I5), title('imbinarize');
subplot(3,4,6); imshow(I6), title('Sobel Edge'); subplot(3,4,7);
imshow(I7), title('Clear borders'); subplot(3,4,8); imshow(I8),
title('Fills holes'); subplot(3,4,9); imshow(I9), title('imbinarize');
subplot(3,4,10); imshow(I10), title('Erode image');
subplot(3,4,11); imshow(I11), title('Dilate image');
subplot(3,4,12); imshow(I13), title('imbinarize');
```

Python Code for MRI Image Augmentation

```
# -*- coding: utf-8 -*- """AugmentMRI.ipynb

Automatically generated by Colaboratory.

Original file is located at
https://colab.research.google.com/drive/1iGwYTH62Yc1-
C6l_4KcbPVVie5jLoITL"""

from keras.preprocessing.image import ImageDataGenerator, array_to_img,img_to_array, load_img

datagen = ImageDataGenerator(
        rotation_range=40,
        width_shift_range=0.2,
        height_shift_range=0.2,
        shear_range=0.2,
        zoom_range=0.2,
        horizontal_flip=True,
        fill_mode='nearest')

img = load_img('surjoban (9).PNG') # this is a PIL image
x = img_to_array(img) # this is a Numpy array with shape (3, 150, 150)
x = x.reshape((1,) + x.shape) # this is a Numpy array with shape (1, 3, 150, 150)

# the .flow() command below generates batches of randomly transformed images# and
saves the results to the `preview/` directory

i = 0
for batch in datagen.flow(x, batch_size=1, save_to_dir='/content/kh',save_prefix='mri',
save_format='jpeg'):
    i += 1
    if i > 150:
        break # otherwise the generator would loop indefinitely

import shutil shutil.make_archive('kh',
'zip', 'kh')
```

Pyhton Code for Identification of Lung Cancer using CNN

```
# -*- coding: utf-8 -*-
"""Lung MRI Cancer Classification
Automatically generated by Colaboratory.
Original file is located at
https://colab.research.google.com/ drive/11eX-
FSCtk8EXuz5p2gqrT2RXBbio5xKK"""

from google.colab import drive
drive.mount('/content/drive')

# import the necessary packages

from tensorflow.keras.preprocessing.image import ImageDataGeneratorfrom
tensorflow.keras.preprocessing.image import img_to_array
from tensorflow.keras.preprocessing.image import load_imgimport
numpy as np
import cv2
import os
import matplotlib.pyplot as pltfrom
zipfile import ZipFile import
tensorflow as tf

#Extracting the dataset contents

zipFilePath='/content/drive/MyDrive/MIST related
works/MokhlessSir/dataMSir.zip'
zipFileObj=ZipFile(file=zipFilePath)
zipFileObj.extractall('/tmp/data')
train_datagen = ImageDataGenerator(rescale=1./255,
    validation_split=0.2) # set validation split
train_generator = train_datagen.flow_from_directory('/tmp/data',
    target_size=(224, 224), batch_size=128,
    class_mode='categorical', subset='training')
    # set as training data
validation_generator = train_datagen.flow_from_directory('/tmp/data',
    # same directory as training data target_size=(224, 224),
    batch_size=128, class_mode='categorical',
    subset='validation') # set as validation data

#Trying to save

checkpoint_filepath = '/content/drive/MyDrive/Colab
Notebooks/Regularizer/checkpoints/ckpt-{epoch:02d}_val_loss_{val_loss:.2f}'
model_checkpoint_callback_val_loss = tf.keras.callbacks.ModelCheckpoint(
    filepath=checkpoint_filepath,
```

```
    save_weights_only=True,
    monitor='val_loss', mode='min',
    save_best_only=True)

from keras.layers import Conv2D, MaxPooling2D, GlobalAveragePooling2D,
Dropout, Activation, Average,Dense from
keras.models import Model, Input
from tensorflow.keras.applications.densenet import DenseNet201from
tensorflow.keras.optimizers import RMSprop
oneInputLayer=Input(shape=(224,224,3),name="my Single Input Layer")
base_model = DenseNet201(include_top=False, weights='imagenet',input_tensor=oneInputLayer)
for layer in base_model.layers:
    layer.trainable = False
x=Conv2D(128,kernel_size=(3,3),activation='relu',padding='same')(base_model.ou tput)
x=Conv2D(128,kernel_size=(3,3),activation='relu',padding='same')(x)
x=Conv2D(128,kernel_size=(3,3),activation='relu',padding='same')(x)
x=Conv2D(128,kernel_size=(3,3),activation='relu',padding='same')(x)
x=Conv2D(128,kernel_size=(3,3),activation='relu',padding='same',name='VGG16
LastConvLayer')(x)
x=GlobalAveragePooling2D()(x)
# use global average pooling to take into account lesser intensity pixelsx =
Dense(128, activation='relu')(x)
x = Dense(64, activation='relu')(x)x =
Dense(32, activation='relu')(x)
x = Dense(2, activation='softmax',name='VGG16LastOutputLayer')(x)

model = Model(base_model.input, x)

model.compile(
    optimizer=RMSprop(lr=0.0001),
    loss='categorical_crossentropy',
    metrics=['accuracy']
)

model.summary()

history = model.fit(
    train_generator,
    epochs=30,
    validation_data=validation_generator
)

import matplotlib.pyplot as plt
plt.plot(history.history['accuracy'])
```

```
plt.plot(history.history['val_accuracy'])
plt.title('model accuracy')
plt.ylabel('accuracy')  plt.xlabel('epoch')
plt.legend(['train', 'val'], loc='upper left')
plt.show()

import matplotlib.pyplot as plt
plt.plot(history.history['loss'])
plt.plot(history.history['val_loss'])
plt.title('model loss') plt.ylabel('loss')
plt.xlabel('epoch')
plt.legend(['train', 'val'], loc='upper left')
plt.show()

# set validation split

test_generator = train_datagen.flow_from_directory('/tmp/data', # same
    directory as training data target_size=(224, 224),
    batch_size=128, class_mode='categorical',
    subset='validation')
from sklearn.metrics import confusion_matrix
testData=test_generator.next()
testX=testData[0]
testY=testData[1]
print(testX.shape)
print(testY.shape)
testHatY=model.predict(testX)
print(np.argmax(testHatY,axis=1))
matrix = confusion_matrix(np.argmax(testHatY,axis=1), np.argmax(testY,axis=1))print(matrix)
from sklearn.metrics import accuracy_score
testAccuracy = accuracy_score(np.argmax(testHatY,axis=1),np.argmax(testY,axis=1))
print(f'Test Accuracy: {testAccuracy}')

# Example 1:
(batch_size = 1, number of samples = 4)y_true = testY
y_pred = testHatY
bce = tf.keras.losses.CategoricalCrossentropy(from_logits=True)print('Test loss')
bce(y_true, y_pred).numpy()
```

Pyhton Code for Identification of Lung Cancer using SVM

Load Modules

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
```

Prepare/collect data

```
from google.colab import drive
drive.mount('/content/drive')
import os
path = os.listdir('/content/drive/MyDrive/Colab Notebooks/Dataset/Training')
classes = {'Noncancerous':0, 'Cancerous':1}
import cv2
X = []
Y = []
for cls in classes:
    pth = '/content/drive/MyDrive/Colab Notebooks/Dataset/Training/'+cls
    for j in os.listdir(pth):
        img = cv2.imread(pth+'/'+j, cv2.IMREAD_GRAYSCALE)
        img = cv2.resize(img, (200,200))
        X.append(img)
        Y.append(classes[cls])
X = np.array(X)
Y = np.array(Y)
X_updated = X.reshape(len(X), -1)
np.unique(Y)
pd.Series(Y).value_counts()
X.shape, X_updated.shape
```

Visualize data

```
plt.imshow(X[0], cmap='gray')
```

Prepare data

```
X_updated = X.reshape(len(X), -1)
X_updated.shape
```

Split Data

```
xtrain, xtest, ytrain, ytest = train_test_split(X_updated, Y, random_state=10,
test_size=.20)
xtrain.shape, xtest.shape
```

Feature Scaling

```
print(xtrain.max(), xtrain.min())
print(xtest.max(), xtest.min())
xtrain = xtrain/255
xtest = xtest/255
print(xtrain.max(), xtrain.min())
print(xtest.max(), xtest.min())
```

Feature Selection: PCA

```
from sklearn.decomposition import PCA
print(xtrain.shape, xtest.shape)
pca = PCA(.98)
# pca_train = pca.fit_transform(xtrain)
# pca_test = pca.transform(xtest)
pca_train = xtrain
pca_test = xtest
```

Train Model

```
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
import warnings
warnings.filterwarnings('ignore')
lg = LogisticRegression(C=0.1)
lg.fit(xtrain, ytrain)
sv = SVC()
sv.fit(xtrain, ytrain)
```

Evaluation

```
print("Training Score:", lg.score(xtrain, ytrain))
print("Testing Score:", lg.score(xtest, ytest))
print("Training Score:", sv.score(xtrain, ytrain))
print("Testing Score:", sv.score(xtest, ytest))
```

Prediction

```
pred = sv.predict(xtest)
misclassified=np.where(ytest!=pred)
misclassified
print("Total Misclassified Samples: ",len(misclassified[0]))
print(pred[36],ytest[36])
```

TEST MODEL

```
dec = {0:'Noncancerous', 1:'Cancerous'}
plt.figure(figsize=(12,8))
p = os.listdir('/content/drive/MyDrive/Colab Notebooks/Dataset/Testing/')
```

```
c=1
for i in os.listdir('/content/drive/MyDrive/Colab Notebooks/Dataset/Testing/Noncancerous/')[:9]:
    plt.subplot(3,3,c)

    img = cv2.imread('/content/drive/MyDrive/Colab Notebooks/Dataset/Testing/Noncancerous/'+i,0)
    img1 = cv2.resize(img, (200,200))
    img1 = img1.reshape(1,-1)/255
    p = sv.predict(img1)
    plt.title(dec[p[0]])
    plt.imshow(img, cmap='gray')
    plt.axis('off')
    c+=1

plt.figure(figsize=(12,8))
p = os.listdir('/content/drive/MyDrive/Colab Notebooks/Dataset/Testing/')
c=1
for i in os.listdir('/content/drive/MyDrive/Colab Notebooks/Dataset/Testing/Cancerous/')[:16]:
    plt.subplot(4,4,c)
    img = cv2.imread('/content/drive/MyDrive/Colab Notebooks/Dataset/Testing/Cancerous/'+i,0)
    img1 = cv2.resize(img, (200,200))
    img1 = img1.reshape(1,-1)/255
    p = sv.predict(img1)
    plt.title(dec[p[0]])
    plt.imshow(img, cmap='gray')
    plt.axis('off')
    c+=1
```

F1 score

```
from sklearn.metrics import f1_score

# Calculate the F1 score

f1 = f1_score(ytest, pred, average='weighted')  # 'weighted' accounts for class imbalance

# Print the F1 score

print(f"F1 Score: {f1}")
```