Lung Cancer Prediction using Machine Learning and Imaging techniques

A Project Report Submitted in the
Partial Fulfillment of the Requirements
for the Award of the Degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted by

AKHIL KUMAR.S	**19881A0504**
KOTALA SAI KIRAN	**19881A0530**
SEETHA HARSHAVARDHAN	**19881A0544**

SUPERVISOR
Dr. M. A. Jabbar
HOD & Professor
Department of CSE (AI&ML)

Department of Computer Science And Engineering
VARDHAMAN COLLEGE OF ENGINEERING, HYDERABAD
An Autonomous Institute, Affiliated to JNTUH

May, 2022

VARDHAMAN COLLEGE OF ENGINEERING, HYDERABAD

An Autonomous Institute, Affiliated to JNTUH

Department of Computer Science And Engineering

CERTIFICATE

This is to certify that the project titled **Lung Cancer Prediction using Machine Learning and Imaging techniques** is carried out by

AKHIL KUMAR.S	**19881A0504**
KOTALA SAI KIRAN	**19881A0530**
SEETHA HARSHAVARDHAN	**19881A0544**

in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in during the year 2021-22.

Supervisor
Dr. M. A. Jabbar
HOD & Professor
Department of CSE (AI&ML)

HOD
Dr. Ramesh Karnati
HOD & Associate Professor
Department of CSE

Declaration

We, AKHIL KUMAR.S, KOTALA SAI KIRAN, SEETHA HARSHAVARDHAN, bearing Roll Numbers 19881A0504, 19881A0530, 19881A0544, respectively, hereby declare that this thesis entitled *Lung Cancer Prediction using Machine Learning and Imaging techniques* presents our original work carried out as a B.Tech student of Vardhaman College of Engineering, Hyderabad and to the best of our knowledge, contains no material previously published or written by another person. Any contribution made to this work by others, with whom I have worked at VCEH or elsewhere, is explicitly acknowledged. Works of other authors cited in this project have been duly acknowledged under the sections "Reference".

We are fully aware that in case of any non-compliance detected in future, the Senate of Vardhaman College of Engineering, Hyderabad may withdraw the degree awarded to us on the basis of the present dissertation.

AKHIL KUMAR.S
KOTALA SAI KIRAN
SEETHA HARSHAVARDHAN

Acknowledgement

The satisfaction that accompanies the successful completion of the task would be put incomplete without the mention of the people who made it possible, whose constant guidance and encouragement crown all the efforts with success.

We wish to express our deep sense of gratitude to **Dr. M. A. Jabbar**, HOD & Professor and Project Supervisor, **Department of CSE (AI&ML)**, Vardhaman College of Engineering, for his able guidance and useful suggestions, which helped us in completing the project in time.

We are particularly thankful to **Dr. Ramesh Karnati**, the Head of the Department, **Department of CSE**, his guidance, intense support and encouragement, which helped us to mould our project into a successful one.

We show gratitude to our honorable Principal **Dr. J.V.R. Ravindra**, for providing all facilities and support.

We avail this opportunity to express our deep sense of gratitude and heartful thanks to **Dr. Teegala Vijender Reddy**, Chairman and **Sri Teegala Upender Reddy**, Secretary of VCE, for providing a congenial atmosphere to complete this project successfully.

We also thank all the staff members of Computer Science and Engineering department for their valuable support and generous advice. Finally thanks to all our friends and family members for their continuous support and enthusiastic help.

AKHIL KUMAR.S

KOTALA SAI KIRAN

SEETHA HARSHAVARDHAN

Abstract

We're aiming to create a web application with Django. Where radiologists can go to the website and upload X-ray scanned images to determine whether the person has a brain tumour, lung disease, or covid disease, allowing radiologists to make speedy judgements. The Model-View-Controller (MVC) architectural paradigm is used by Django. Its purpose is to make building complex, database-driven websites easier. Django stresses component reusability and "pluggability," quick development.

The formation of aberrant cells in the brain, some of which may progress to cancer, is known as a brain tumour. Magnetic Resonance Imaging (MRI) scans are the most common tool for detecting brain tumours. Information on aberrant tissue growth in the brain is identified using MRI imaging. AI and deep learning are critical in finding and classifying COVID-19 cases utilising computer-assisted applications, which yields great results for identifying COVID-19 cases based on known symptoms such as fever, chills, dry cough, and a positive x-ray. Lung disorders are linked to a wide range of diseases that affect people all over the world. Lung disorders, often called respiratory diseases, are diseases that affect the airways and other components of the lungs. As a result, detecting lung disease early is important to saving lives. Early detection can improve the odds of a patient being cured and recovering.

A self-defined Artificial Neural Network (ANN) and a Convolution Neural Network (CNN) are used in the proposed study to detect the existence of brain tumours, lung illness, and covid disease, and their performance can be evaluated. The radiologist can use these predictions to make timely decisions.

Keywords:Deep Learning; Lung Cancer; Transfer Learning; VGG16; Machine Learning

Table of Contents

List of Figures

Abbreviations

Abbreviation		Description
VCE	:	Vardhaman College of Engineering
CNN	:	Convolutional neural network
VGG	:	Visual Geometry Group

Chapter 1

Introduction

1.1 Lung Cancer

Lung cancer incidence, adult smoking prevalence, estimated percent of radon tests at or above the United States Environmental Protection Agency action level, five-year survival, early diagnosis, surgery as part of the first course of treatment, lack of treatment, and screening among those at high risk are all included in the report.

For the years 2014-2018, the data on lung cancer incidence, stage, surgical therapy, and absence of treatment includes malignant lung and bronchus tumours. These figures are based on data submitted by the North American Association of Central Cancer Registries (NAACCR) in December 2020. In the United States, registries may also participate in the Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute or the National Program of Cancer Registries (NPCR) of the Centers for Disease Control and Prevention (CDC), or both. The state, province, or territory in which the register is located provides support for cancer registries.

Surgical removal of distant lymph glands or other tissue(s)/organ(s) beyond the primary site; surgery for lung cancer; radiation; chemotherapy; systemic hormonal agents; immunotherapy; other, including experimental, double-blind, and unproven; and Because a parameter was used this year, rates from earlier reports should not be compared.

The Lung Association polled state Medicaid programmes to acquire information on coverage of low dose CT scans for patients at high risk for lung cancer and analysed publicly available coverage policies to assess current lung cancer screening coverage in state Medicaid fee-for-service systems.

1.2 Proposed System features for Lung Cancer Detection

The most popular noninvasive imaging modalities for detecting and diagnosing lung nodules are positron emission tomography (PET), computed tomography (CT), low-dose computed tomography (LDCT), and contrast-enhanced computed tomography (CE-CT). PET scans are utilised to distinguish between cancerous and non-cancerous lung tumours. CT and LDCT scans, which allow for recreating the anatomy of the chest and detecting anatomic changes, can be used to detect nodules early. The CE-CT can be used to reconstruct the anatomy of the chest and examine the characteristics of the discovered nodule. In CT pictures, healthy lung tissues appear darker than other parts of the chest, such as the heart and liver.

In most cases, a biopsy is the only option to confirm a cancer diagnosis. Doctors examine cell samples under a microscope in the laboratory.

1.3 Feature Extraction

Image processing, particularly feature extraction using CNN, is a hot area in computer science research. In this suggested study, an experiment was done utilising scratch and pre-trained CNN models. The scratch model's results were unsatisfactory, while the pre-trained model performed admirably.

The VGG16 model (pre-trained) was fine-tuned to fit the experimental dataset utilised in this work as a feature extractor. This network model was created using a VGGNet with 16 layers. VGG16 outperformed VGG16, scratch model, and other deep learning models in an experimental trial, according to the results. The activation function for getting the output of convolution layers was a Binary Sigmoid, and the convolution section was separated into five consecutive max-pooling layers. The first and second subregions were developed using two convolution layers with depths of 64 and 128 respectively. Furthermore, the remaining three subregions were constructed using four consecutive convolution layers with depths of 224, 224, and 224, respectively.Following that, Pooling layers were used to reduce the learnable parameter. The proposed VGG16 model's final layer assisted in getting the feature vector, whereas the two hidden layers placed before the feature collection layer had 1024 and 512 neurons, respectively. For reducing overfitting during the fine-tuning model implementation. VGG16 models based on CNN give 4096 suitable characteristics.

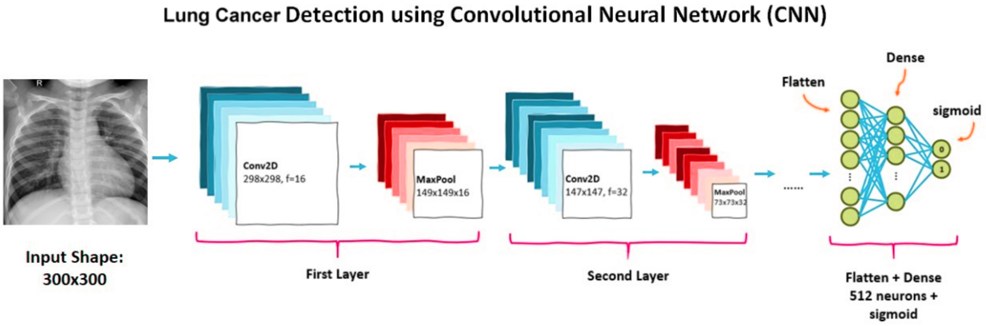


Figure 1.1: Feature Extraction

1.4 Understanding Lung Cancer

Lung cancer develops in the lungs. The lungs are two spongy organs in your chest that inhale oxygen and expel carbon dioxide. Lung cancer is the most common cancer in the world.

1.4.1 What is Lung Cancer?

Cancer is a condition in which the body's cells grow out of control. Lung cancer is a kind of cancer that begins in the lungs. Lung cancer starts in the lungs and can spread to the lymph nodes or other bodily organs, including the brain. Other organ cancers can potentially move to the lungs. Metastases are the spread of cancer cells from one organ to another.

Small cell and non-small cell lung tumors are the two primary forms of lung cancer (including adenocarcinoma and squamous cell carcinoma). These kinds of lung cancer develop and respond to treatment in various ways. Compared to small cell lung cancer, non-small cell lung cancer is more prevalent. Visit the Lung Cancer page of the National Cancer Institute for further information.external icon

1.4.2 What Are the Symptoms of Lung Cancer?

Lung cancer symptoms range from person to person. Some people experience lung-related problems. Some persons with lung cancer who have spread to other areas of their bodies (metastasized) experience symptoms unique to that area.

Some folks just exhibit general signs of being unwell. Most persons with lung cancer do not have symptoms until the disease has progressed. Symptoms of lung cancer include:

- Coughing that gets worse or doesn't go away.
- Chest pain.
- Shortness of breath.
- Wheezing.
- Coughing up blood.
- Feeling very tired all the time.
- Weight loss with no known cause.

Repeated bouts of pneumonia and swollen or enlarged lymph nodes (glands) within the chest in the region between the lungs are further alterations that can develop with lung cancer.

Other conditions might also cause same symptoms. If you have any of these symptoms, consult your doctor, who can assist you in determining the source.

1.4.3 What Are the Risk Factors for Lung Cancer?

Several risk factors have been discovered in studies that may affect your risks of getting lung cancer.

1.4.3.1 Smoking

People who smoke cigarettes are 15 to 30 times more likely than nonsmokers to get lung cancer or die from it. Lung cancer is increased by even a few cigarettes each day or smoking on occasion. The greater the danger, the longer a person smokes and the more cigarettes smoked per day.

People who quit smoking have a lower risk of lung cancer than those who continue to smoke, but they still have a larger risk than those who have never smoked. Lung cancer risk can be reduced by quitting smoking at any age. Cigarette smoking has been linked to cancer in practically every organ of the body. Cigarette smoking causes cancers of the mouth and throat, oesophagus, stomach, colon, rectum, liver, pancreas, voicebox (larynx), trachea, bronchus, kidney and renal pelvis, urinary bladder, and cervix, as well as AML.

1.4.3.2 Radon

In the United States, radon is the second largest cause of lung cancer after smoking. Radon is a gas that develops naturally in rocks, soil, and water. It is not visible, tasteable, or odorable. When radon enters a home or building through cracks or openings, it can become trapped and accumulate in the air. High radon levels are inhaled by those who live or work in these homes and businesses. Radon exposure over time can lead to lung cancer. According to the US Environmental Protection Agency (EPA), radon causes roughly 21,000 lung cancer deaths each year. People who smoke have a higher risk of lung cancer from radon exposure than those who do not. The Environmental Protection Agency (EPA) believes that more than 10% of radon-related lung cancer fatalities occur in adults who have never smoked cigarettes. In the United States, nearly one in every fifteen residences has elevated radon levels. Learn how to test your house for radon and how to lower a high radon level.

1.4.3.3 Exposure to asbestos

People who work with asbestos are several times more likely to develop lung cancer (for example, in mines, mills, textile companies, places where insulation is used, and shipyards). The risk of lung cancer is substantially higher among asbestos workers who also smoke. It's unclear how much low-level or short-term asbestos exposure increases the risk of lung cancer.

People who are exposed to a lot of asbestos are more likely to acquire mesothelioma, a cancer that develops in the pleura (the lining surrounding the lungs). See Malignant Mesothelioma for more information on this cancer.

Government regulations have significantly reduced the use of asbestos in commercial and industrial items in recent years.

1.4.4 Can Lung Cancer Be Prevented?

1.4.4.1 Stay away from tobacco

The best approach to lower your lung cancer risk is to quit smoking and avoid inhaling other people's smoke.

If you stop smoking before a cancer develops, your damaged lung tissue gradually starts to mend itself. Quitting smoking, regardless of your age or length of smoking, may reduce your risk of lung cancer and help you live longer.

1.4.4.2 Avoid radon exposure

Radon is an important cause of lung cancer. You can reduce your exposure to radon by having your home tested and treated, if needed.

1.4.4.3 Eat a healthy diet

A balanced diet rich in fruits and vegetables may also assist to lower your lung cancer risk. Some data suggests that eating a diet rich in fruits and vegetables can help smokers and nonsmokers avoid lung cancer. However, any beneficial benefit of fruits and vegetables on lung cancer risk would be dwarfed by the increased risk associated with smoking.

Attempts to minimise the risk of lung cancer in adults who presently or formerly smoked by giving them high doses of vitamins or vitamin-like medications have so far failed. In fact, several studies have suggested that taking beta-carotene supplements, a substance related to vitamin A, increases the risk of lung cancer in these patients.

Some people who get lung cancer do not have any clear risk factors. Although we know how to prevent most lung cancers, at this time we don't know how to prevent all of them.

1.5 What is lung cancer screening?

Screenings detect disease before symptoms appear. Screening aims to catch disease at its earliest and most treatable stage. A screening programme must meet a number of criteria in order to be generally acknowledged and recommended by medical practitioners, including minimising the number of deaths caused by the disease.

Lab tests to check blood and other bodily fluids, genetic testing to look for inherited genetic markers associated to disease, and imaging scans to take pictures of the inside of the body are all examples of screening tests. The general public usually has access to these tests. Individual screening needs, on the other hand, are determined by characteristics such as age, gender, and family history.

In lung cancer screening, individuals who have a high risk of developing lung cancer but no signs or symptoms of the disease undergo low-dose computed tomography (LDCT) scanning of the chest.

LDCT combines special x-ray equipment with sophisticated computers to produce multiple, cross-sectional images or pictures of the inside of the body. LDCT produces images of sufficient quality to detect many abnormalities while using up to 90 percent less ionizing radiation than a conventional chest CT scan.

In the past, doctors used chest x-ray and sputum cytology to check for lung cancer. A chest x-ray makes images of the heart, lungs, airways, blood vessels and the bones of the spine and chest. Sputum cytology is a lab test in which a sample of sputum (mucus that is coughed up from the lungs) is viewed under a microscope to check for cancer cells. However, the use of chest x-ray and sputum cytology, individually or in combination, has not resulted in a decreased risk of dying from lung cancer.

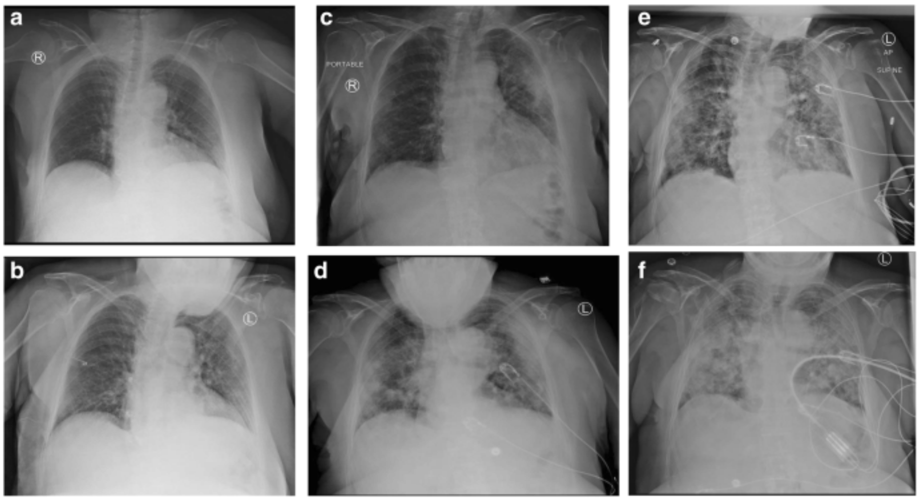


Figure 1.2: Lung Cancer Screening

1.5.1 What are the benefits and risks of lung cancer screening?

- Because CT scans can detect even very small nodules in the lungs, LDCT of the chest is especially effective for diagnosing lung cancer at its earliest, most treatable stage.
- CT is fast, which is important for patients who have trouble holding their breath.
- CT scanning is painless and noninvasive. LDCT does not require contrast material.
- No radiation remains in a patient's body after a CT exam.
- X-rays used in LDCT of the chest have no immediate side effects and do

not affect any metal parts in your body, such as pacemakers or artificial joints.

- LDCT scans of the chest produce images of high enough quality to detect many abnormalities while using up to 90 percent less ionizing radiation than a conventional chest CT scan.
- Studies prove that lung cancer screening with LDCT reduces the number of deaths from lung cancer in patients at high risk.
- When cancer is found with screening, it is often at an early stage. Patients can more often undergo minimally invasive surgery and have less lung tissue removed.

Chapter 2

Literature Survey

Janee et al. developed ”multi-stage lung cancer detection and prediction using multi-class SVM classifier,” which outlines a successful approach that use SVM for lung cancer detection, diagnosis, and prediction. The Gray Level Cooccurrence Method (GLCM) method is used to design an algorithm that employs image processing techniques such as image enhancement, detection and segmentation, detection, and feature extraction. For classification purposes, SVM is used. The binarization approach is used to make predictions. The UCI ML database is obtained, containing 500 infected and non-infected CT scans. Out of 130 photos of lung cancer caused over the world, the proposed approach identified 126 as contaminated. In India, lung cancer mortality are also on the rise.

For the evaluation and analysis of human perception and cognition, ML/AI employs complex and sophisticated algorithms. Regardless of user input, computer algorithms can reach cessations. Traditional technologies in Ai are distinguished by the intelligence of Ai algorithms for acquiring information, processing it, and deciding/finalizing accurate output. ML techniques are used to implement these algorithms. AI in healthcare is divided into two groups based on the type of data. ML algorithms are used to analyse structured data as well as pictures, genes, and biomarkers. Natural language processing (NLP) technologies are used to examine unstructured data such as prescription notes, medical publications, and journals.First, the gathered text is transformed to binary representation using NLP methods, and then this binary data is processed using machine learning techniques to create correct output and decisions.

Cancer, neurology, and cardiology are the most common medical research in which AI is used. As this disease has a higher fatality rate. Apart from these disorders, AI is being used in other fields of medicine for prediction, analysis, and treatment. SVM, NN, random forest, logistic regression, discriminant analysis, decision trees, linear regression, closest neighbour, naive bayes, and other well-known ML techniques are commonly used in the healthcare sector.

The following are the top AI algorithms that can be used in healthcare right now:

- The algorithm detecting variation in a tumor
- Classification of heart images
- Heart attack predicting algorithm
- More precise skincare cancer diagnosis with AI
- AI system for ICU
- Computers detecting breast cancer risk AI useful to diagnose breast cancer
- Smart algorithm predicting suicide risk
- Inpatients mortality can be predicted by AI

"A Computer-Aided Diagnosis System for Detection of Lung Cancer Nodules Using Extreme Learning Machine" was defined by M.Gomathi et al. A CAD model is created in this study for the analysis of cancer in CT scans. The detection of the region of interest in input CT scans is the most basic phase of CAD. The extraction of lung regions is preceded by lung region segmentation, and cancer nodules are detected using the Fuzzy Possibility CMean (FPCM) clustering algorithm. The diagnostic guidelines are formulated using a maximum Drawable Circle intensity value. Then, with the help of the ELM, these rules are put into action to learn (Extreme Learning Machine).

Shingo Kakeda and colleagues proposed a commercial CAD model for detecting lung nodules on chest radiographs. The CAD model, which includes an image server and EpiSight/XR software, is presented in this work. This approach is broken down into four basic components. To generate alternative images, the first complex automated structure is decreased. Multiple Gray Level thresholding approaches are used to detect nodule candidates. To distinguish between true and false-positive nodules, input features and difference images are used to extract features. For the decrease of false-positive nodules, previously extracted features are used in conjunction with a rule-based analysis and ANN. For testing, the model was applied to a database of 274 radiographs with 323 lung nodules. Out of 315, 235(75%) false-positive images were detected as a normal automatic structure and 155(49%) are detected as pulmonary vessels.

"Lung nodule detection on thoracic computed tomography images: Preliminary evaluation of a computer-aided diagnosis system," offered Metin N et al. The suggested approach consists of five steps. The first step is to segment regions using the k-means clustering algorithm. After segmenting the lung curves, suspicious regions are segregated from lung regions using pre-processing techniques, resulting in a binary picture with holes due to the segmentation process. Flood – the filled algorithm is used to fill these gaps as nodule-candidates are regarded solid objects. This system may include general regions and blood vessel-rich lung nodules. Rule-based classifiers using 2D and 3D features are utilised to distinguish this nodule.At last, the false positive objects are detected using Linear Discrimination Analysis (LDA). The proposed method was analyzed on a dataset that includes 1454CT images gathered from 34 diagnosed patients with 63 lung nodules.

Kazak Awai et al [10] proposed a system for assessing the impact of CAD on a radiologist's ability to detect lung nodules. For lung and intrapulmonary structure segmentation, the suggested method used image processing techniques such as gray-level threshold, 3D labelling techniques, and mathematical morphological techniques. On an input image, the Top-hat transformation approach is used to detect the smoothed image for segmentation of intrapulmonary structures. The principal prospective nodules are identified using a sieve filter, and then features of these pulmonary nodules are retrieved to distinguish actual nodules from false-positive nodules. Finally, an ANN is utilised to determine the probability of a region of interest based on an image feature. The adaptation of this approach improved the detection of CT scans by pulmonary nodule residents.

"Computer-aided diagnosis for lung CT utilising artificial life models" was proposed by Cheran et al. Various algorithms are used to create this CAD model. First, a 3D region growth algorithm is used to determine the ribcage region, and then the active contour technique is used to create a specific area for approaching ants, which are subsequently redistributed to create a specific and precise rebuilding of the vascular tree and pleura. Artificial life models are employed to renew the bronchial and vascular trees. It is assessed whether the previously constructed branches include nodules and whether the nodules are related to the pleura using active shape models. Cleaner method is created by using snakes and dot enhancement algorithm to restrict the nodules.

The Optimal Deep Neural Network (ODNN) and Linear Discriminate Analysis (LDA) based classification model for CT images was created by Lakshmanaprabu et al. To forecast lung cancer, LDR is used to classify lung nodules, and the Modified Gravitational Search Algorithm is used to optimise them. For the experimental analysis, a standard CT database with 50 low-dose lung cancer CT pictures was used. This model is compared to other models such as KNN, NN, DNN SVM, and so on, and the experimental analysis demonstrates that the constructed model performs better, with 94.56 percent accuracy, 96.2 percent sensitivity, and 94.2 percent specificity, respectively.

Worawate et al devised a method called ”Automatic Lung Cancer Prediction from Chest X-ray Images Using Deep Learning Approach,” in which the authors employed DensNet-121 (121 layers Convolutional neural network) in conjunction with transfer learning to identify images from the chest. To identify the nodules, the model is trained on two datasets: Chest X-ray 14 and JSRT. The model had an accuracy of around 74.436.01%, a sensitivity of about 74.68 15.33 percent, and a specificity of about 74.96 9.85 percent, respectively.

Christoph et al proposed a tomography lung cancer image-based prediction model. By fine-tuning pre-trained ResNet18, a CNN is used for feature extraction, and a multimodal features CNN is trained by the Cox model for hazard prediction. The Lung1 dataset, which can be acquired from ”The Cancer Imaging Archive” (TCIA), contains 422 NSCLC (Non-Small Cell Lung Carcinoma) images for 318 of 422 patients, is used for experimental research.

Jason et al created the Deep Screener algorithm, which is a type of deep learning approach. The Deep Screener uses low-dose CT scans to do an end-to-end automated lung cancer screening. The TCIA dataset, which contains 1449 low dose CT scans, is used for the experimental analysis. For lung cancer analysis, the model generated was compared to the grt123 method from Data Science Bowl 2017, and the outcome was too close to call. The suggested model accurately predicted 1359 out of 1449 CT scans with an AUC of 0.885 and an AUPRC of 0.837. Table 4 shows an overview of various lung cancer prediction and detection research publications, as well as the algorithms, datasets, and measurements used in those papers.

2.1 Prediction of Lung Cancer

In lung cancer, tumour histology is a key predictor of therapy response and outcome. Although tissue sample for pathologist assessment is the most reliable approach for histology classification, current improvements in deep learning for medical image analysis suggest that radiologic data could be useful in characterising disease features and risk stratification. We present a radiomics technique for predicting non-small cell lung cancer (NSCLC) tumour histology from non-invasive standard-of-care computed tomography (CT) data in this work. On a dataset of 311 early-stage NSCLC patients receiving surgical treatment at Massachusetts General Hospital (MGH), we trained and validated convolutional neural networks (CNNs), with a focus on the two most frequent histological types: adenocarcinoma (ADC) and Squamous Cell Carcinoma (SCC) (SCC).The CNNs were able to predict tumor histology with an AUC of 0.71($p = 0.018$). We also found that using machine learning classifiers such as k-nearest neighbors (kNN) and support vector machine (SVM) on CNN-derived quantitative radiomics features yielded comparable discriminative performance, with AUC of up to 0.71 ($p = 0.017$). Our best performing CNN functioned as a robust probabilistic classifier in heterogeneous test sets, with qualitatively interpretable visual explanations to its predictions. Deep learning based radiomics can identify histological phenotypes in lung cancer. It has the potential to augment existing approaches and serve as a corrective aid for diagnosticians.

Lung cancer is the leading cause of death from cancer worldwide, with 2.09 million new cases and 1.76 million deaths in 2001. In the early 2000s, four case-control studies from Japan found that combining chest radiographs with sputum cytology in lung cancer screening reduced mortality2. In contrast, two randomised controlled trials conducted between 1980 and 1990 found that chest radiograph screening was ineffective in reducing lung cancer mortality3,4. Compared to low-dose computed tomography, chest radiographs are more cost-effective, easier to access, and deliver a lower radiation dose (CT). Excessive false positive (FP) results are another drawback of chest CT. It according to one study, 96 percent of nodules detected by low-dose CT screening are FPs, resulting in unnecessary follow-up and invasive examinations5. In terms of sensitivity, chest radiography is inferior to chest CT, but it is superior in terms of specificity. Taking these factors into account, developing a computer-aided diagnosis (CAD) model for chest radiographs would be beneficial in terms of improving sensitivity while maintaining low FP results. Convolutional neural networks (CNN), a type of deep learning (DL)6,7, have recently been used to achieve dramatic, state-of-the-art improvements in r adiology8. DL-based models have also shown promise for nodule/mass detection on chest radiographs9–13, with sensitivities ranging from 0.51 to 0.84 and mean number of FP indications per image (mFPI) ranging from 0.02 to 0.34. In Furthermore, with these CAD models, radiologist performance for detecting nodules was better than without them9. It can be difficult for radiologists to detect nodules and distinguish between benign and malignant nodules in clinical practise. Because normal anatomical structures can mimic nodules, radiologists must pay close attention to the shape and marginal properties of nodules. Even skilled radiologists can misdiagnose14,15 because these issues are caused by the conditions rather than the ability of the radiologist. Detection and segmentation are the two main methods for detecting lesions using DL. A region-level classification is used for detection, while a pixel-level classification is used for segmentation. The detection method cannot provide as much detail as the segmentation method. In clinical settings, Making a correct diagnosis is more likely when the size of a lesion is classified at the pixel level. Because the shape can be used as a reference during detection, pixel-level classification makes it easier to track changes in lesion size and shape. When determining the effect of treatment, it is also possible to consider not only the long and short diameters, but also the area of the lesion16. There have been no studies using the segmentation method to detect pathologically proven lung cancer on chest radiographs to our knowledge.

The goal of this research was to train and validate a DL-based model capable of detecting lung cancer on chest radiographs using the segmentation method, as well as to assess the model's characteristics. Improve sensitivity while maintaining low FP results with this model.

The following points summarise the article's contributions:

- A deep learning-based model for detecting and segmenting lung cancer on chest radiographs was developed in this study

- Our data is of high quality because all of the nodules/masses were pathologically proven lung cancers that were annotated at the pixel level by two radiologists
- The segmentation method was more informative than the classification or detection methods, which is useful not only for lung cancer detection but also for treatment efficacy and follow-up

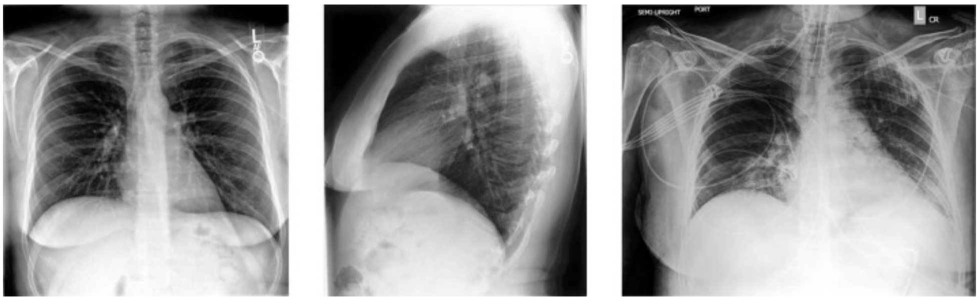


Figure 2.1: Lung CT Scan image

2.2 Related Work

Lung cancer is the leading cause of cancer death for both men and women in the United States, accounting for about 6% of all deaths each year. 1 Most patients with lung cancer who receive an initial diagnosis have advanced stage disease, making cure with current therapies unlikely. Individuals with early-stage disease, on the other hand, can be cured through surgical resection. Because of this disparity in outcome based on stage, Dr. Bach and Ms. Tate, Health Outcomes Research Group, Department of Epidemiology and Biostatistics, and Department of Medicine, Memorial Sloan-Kettering Cancer Center, New York, NY; Dr. Kelley, and the Center for Clinical Health Policy Research (Dr. McCrory), Duke University, Durham, NC; Durham Veterans Affairs Medical Center, Durham, NC e-mail: bachp@mskcc.org; Peter B. Bach, MD, MAPP, Health Outcomes Research Group, Memorial Sloan-Kettering Cancer Center, 1275 York Ave, Box 221, New York, NY 10021; diagnosis, there has been persistent interest in designing and testing methods for early detection of lung cancer. To place these studies in context, we articulate the critical elements of successful screening tests and describe the research methods that can be used to evaluate them. We then focus on lung cancer screening studies with chest x-ray (CXR), sputum cytology, and low-dose CT (LDCT) scanning. For the first two modalities, we review published randomized controlled trials (RCTs). For LDCT, we review published and ongoing studies, discuss the evidence that has fueled the recent debate over efficacy, and discuss future plans of assessment.

A Beneficial Screening Test's Components The value of a screening test is usually measured in two ways. First, the screening test must benefit people who have the illness, usually by increasing their life expectancy. To increase life expectancy, the test must be able to detect the disease at a point in its natural history when it can be altered through treatment, which is typically earlier than in sporadic practise. Patients live longer after an early stage diagnosis. lung cancer than they do after a diagnosis of more advanced stage lung cancer; as a result, it is widely assumed that early detection followed by definitive treatment will alter the disease's natural history and thus reduce lung cancer mortality. Second, the screening test should not be dangerous or painful, and it should not produce a large number of falsepositive results that cause anxiety or require invasive or dangerous follow-up tests. From a societal standpoint, the screening test should not harm the large proportion of the population who do not have the disease, either by consuming vast amounts of resources or by directly affecting the health-care system's capacity to provide for others. Screening Test Evaluation Research Methods For the evaluation of screening tests, three general methods have been advocated: randomised screening trials, population-based screening studies, and observational screening studies in select cohorts. RCTs have traditionally been at the top of the scientific community's hierarchy of study designs. Individuals are subjected to different intensities of screening in a randomised trial, and the outcome is diseasespecific mortality. This method was used to prove the efficacy of fecal-occult blood testing for colon cancer detection. 2–4 The Mayo Lung Project, for example, randomised people to receive regular screening (intervention arm) or routine care (control arm); the primary outcome of interest was lung cancer mortality. 5–7 The population-based study is an alternative highly indicative design.

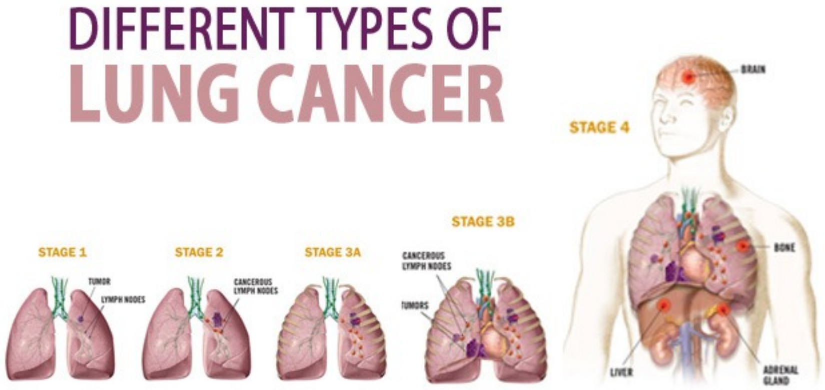


Figure 2.2: Types of Lung Cancer

For the evaluation of screening tests, three general methods have been advocated: randomised screening trials, population-based screening studies, and observational screening studies in select cohorts. RCTs have traditionally been at the top of the scientific community's hierarchy of study designs. Individuals are subjected to different intensities of screening in a randomised trial, and the outcome is diseasespecific mortality. This method was used to prove the efficacy of fecal-occult blood testing for colon cancer detection. 2–4 The Mayo Lung Project, for example, randomised people to receive regular screening (intervention arm) or routine care (control arm); the primary outcome of interest was lung cancer mortality. 5–7 The population-based study is an alternative highly indicative design. In All of these study designs have potential flaws, which readers should be aware of. Negative randomised trial results may be viewed with scepticism due to design flaws (such as insufficient power) or study conduct flaws (such as having excessive crossover between arms). For example, despite seven different RCTs, the efficacy of mammography in preventing breast cancer death is still unknown. Some researchers believe that the majority of mammography studies have significant design flaws that limit their ability to measure the impact of mammography. These researchers have stated that even the best-designed studies show no discernible impact on overall mortality10,11, a claim that has prompted an independent panel of US National Cancer Institute (NCI) advisers to revise their previous conclusions. All of these study designs have potential flaws, which readers should be aware of. Negative randomised trial results may be viewed with scepticism due to design flaws (such as insufficient power) or study conduct flaws (such as having excessive crossover between arms). For example, despite seven different RCTs, the efficacy of mammography in preventing breast cancer death is still unknown. Some researchers believe that the majority of mammography studies have significant design flaws that limit their ability to measure the impact of mammography. These researchers have stated that even the best-designed studies show no discernible impact on overall mortality10,11, a claim that has prompted an independent panel of US National Cancer Institute (NCI) advisers to revise their previous conclusions. 55,034 London men were randomly assigned to CXR every six months for three years, or CXR at the start and end of the three-year period. In both study arms, the follow-up rate exceeded 99 percent. In this study, the frequently screened group (132 cases) identified 36 more cases of lung cancer than the group that was only screened at the start and end of the study (96 cases). During the three-year study period, 44 percent of cancers detected in the screened group were resected, compared to only 29 percent in the control group. Lung cancer mortality was similar over the three-year study period: 62 people in the frequently screened group died of lung cancer, compared to 59 in the less frequently screened group. The hypothesis that CXR screening would be associated with a mortality benefit was not supported by this study or several observational studies, and widespread screening was not pursued. 19,20 In the 1970s, advances in CXR and sputum cytology technology prompted the NCI to revisit the issue. The National Cancer Institute supported the Cooperative Early Lung Cancer Group21, which carried out three randomised studies on sputum cytologic examination and CXR. In Czechoslovakia, a randomised study of CXR and sputum cytology was conducted at the same time. 22 The Johns Hopkins Lung Project and the Memorial Sloan-Kettering Cancer Center [MSKCC] Lung Cancer Screening Program were two NCI-funded RCTs that randomised subjects to CXR alone or CXR plus sputum cytology. There were no differences in the number of cancers detected, the fraction of ”resectable” cancers, or the lung cancer mortality rate in either study (Table 1). 23–25 The primary goal of these studies was to determine the incremental impact of sputum. cytology in a CXR-positive cohort As a result, the lack of a mortality benefit in these studies could be due to the insensitivity of sputum cytology in its current form. The Mayo Lung Project, which was funded by the National Cancer Institute, looked at the combined impact of CXR and sputum cytology for screening.

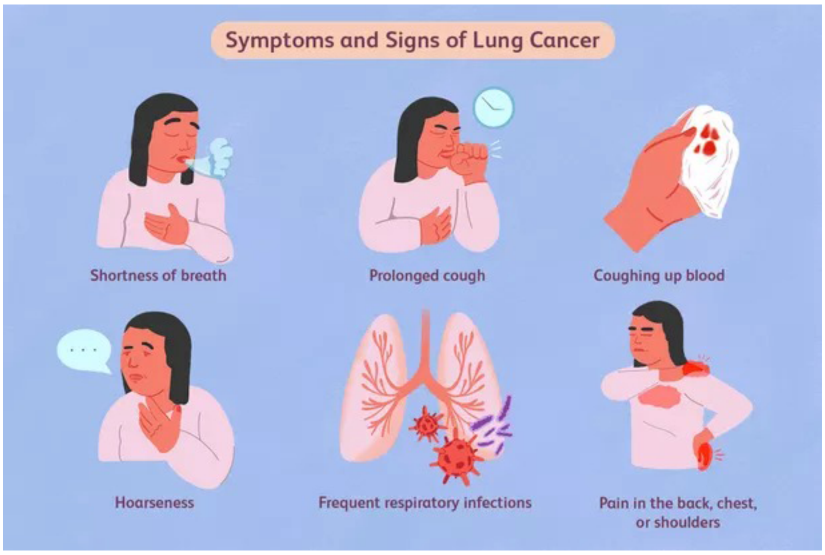

Symptoms and Signs of Lung Cancer
Shortness of breath
Prolonged cough
Coughing up blood
Hoarseness
Frequent respiratory infections
Pain in the back, chest, or shoulders

Figure 2.3: Symptoms and signs of Lung Cancer

5–7 This study enrolled 45-year-old male smokers who had smoked at least one pack per day for at least five years, had no evidence of lung cancer on initial examination, and had sufficient lung function to tolerate lobectomy between 1971 and 1976. 91 prevalent cases of lung cancer were detected on screening studies (0.83 percent) among the 10,933 subjects initially recruited, 51 cases by CXR, 17 cases by sputum cytology alone, and 15 cases by a combination of the two. of the two tests Individuals who had negative and satisfactory prevalence screens were then randomly assigned to one of two groups. The ”screening” group had CXRs and sputum cytology every four months for six years, while the ”control” group was given a recommendation to have a yearly CXR and sputum cytology examination at the start of the study (that being the standard Mayo Clinic recommendation at that time). The subjects were then tracked for 1 to 5.5 years (median, 3 years). During the first year, compliance in the screened group was at 85 percent, but by the end of the study, it had dropped to 75 percent. During the study, CXR was performed on more than half of the control group. The screened group had a higher rate of lung cancer detection (206 cases) compared to those in the control group (160 cases; Table 1). Only 32% of the cancers detected in the control group were localised, making surgical resection with curative intent possible. However, no significant differences in lung cancer-specific mortality rates were found in the first analysis: 3.2 per 1,000 person-years in the screened group and 3.0 per 1,000 person-years in the control group. Deaths from all causes were also comparable (Table 1). The Mayo Lung Project's researchers concluded that combining CXR and sputum cytology every four months does not reduce lung cancer mortality when compared to the annual screening recommendation. 6 An A 15-year analysis of the same cohort revealed a similar result: there was no discernible difference in lung cancer mortality rates between the screened (4.4 per 1,000 person-years) and the control groups (3.9 per 1,000 person-years). 6,26 In a concurrent study in Czechoslovakia, participants were randomly assigned to CXR and sputum cytology every 6 months or to initial screening followed by screening again after 3 years. After that, subjects in both arms were screened annually for another three years. At three years, the findings were similar to the Mayo study: the intervention group had detected more cancers than the control group (39 cases vs 27 cases), including a higher number of cases with early detection. disease stage (20 cases vs 10 cases; Table 1). However, there was no difference in the number of people diagnosed with late-stage cancer in the screened group versus the control group (19 patients vs 17 patients), nor in the number of people who died of lung cancer. 22,27 Conclusions: Sputum Cytology and CXR The findings of the five RCTs indicate that neither CXR nor sputum cytology meet the primary criteria for a useful screening test. Neither test appears to extend a person's life expectancy if they have the disease. Any of these studies did not go into enough detail to determine whether either test meets the second criterion—that it is not particularly painful or harmful. Despite the consistency, readers should be aware that There is still scepticism about the interpretation of the findings in these studies. One observation; The International Union Against Cancer (IUCN) was the first to provide it. 28 is that many of these studies may be invalid because they lacked a "noscreening" study arm, making true efficacy impossible to determine. Other authors have claimed that these studies' sample sizes were insufficient. The Mayo Lung Project and the Czechoslovakian studies were both powered to detect a 50% reduction in lung cancer mortality in the screened group versus the control group. 27 The power to detect a smaller but clinically significant effect, such as a 10% decrease in mortality, was much lower (only 0.21 and 0.16, respectively). 7,27,29–31 Case-control studies have suggested that CXR screening may still be beneficial, which has added to the debate. 32 Currently, The Prostate, Lung, Colorectal, and Ovarian Trial, sponsored by the National Cancer Institute, is a randomised trial aimed at determining the role of CXR screening in a low-risk population of both genders, but the results have yet to be released. 33,34 LDCT Scanning LDCT scanning is a technique for obtaining a low-resolution image of the entire thorax in a single breath-holding procedure with minimal radiation exposure. As a screening test, LDCT has a lot of supporters. LDCT has only been tested in observational studies with volunteer cohorts so far.

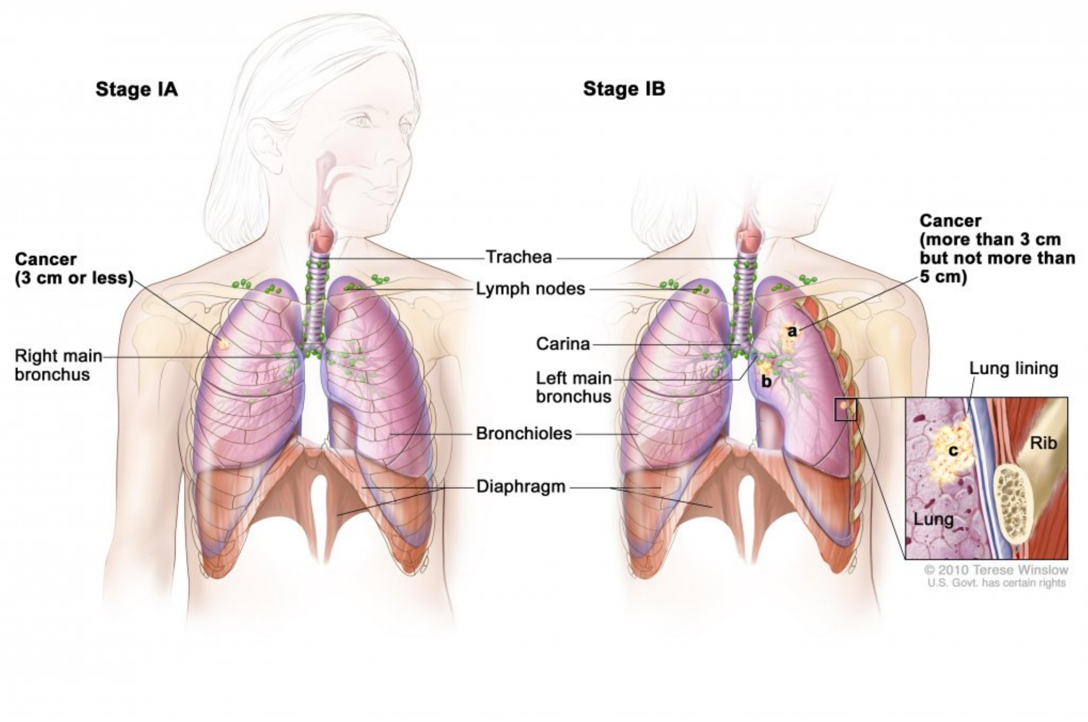


Figure 2.4: Stage 1A and 1B of Lung Cancer

LDCT Screening Research In 1996, a prevalence screening using a single-spiral CT scan was conducted in 5,483 people aged 40 to 74 in Matsumoto, Japan. 35 The majority of the subjects had gone through annual evaluations. CXRs and sputum cytologic screening were performed simultaneously with spiral CT scanning on 3,967 subjects. Sputum cytology was also performed on smokers. There were 59 subjects (1%) with noncancerous but suspicious lung lesions, 84 subjects (2%) with lesions suspicious for lung cancer, and 80 subjects (2%) with indeterminate small lung nodules among those who had abnormalities prompting further evaluation. There were 19 lung cancers discovered (8.5 percent of subjects with abnormal findings prompting further evaluation). These lesions were initially radiographically suspicious for lung cancer in 14 cases, benign but suspicious in three cases, and indeterminate in two cases. 18 of the 19 cases were surgically confirmed, while 1 case was diagnosed clinically. 16 of the 19 tumours (84 percent) were stage I, while the rest (3%) were stage II. Stage IV tumours made up 19 of the total. Four tumours were 1 cm in diameter, and 14 tumours were between 1 and 2 cm in diameter. On CXR, only one of the 17 tumours measuring 2 cm was visible. Sputum cytology was used to diagnose a lung cancer that was not visible on a CT scan. Overall, 0.48 percent of people were diagnosed with lung cancer, which was significantly higher than the 0.03 to 0.05 percent rate in the same area before spiral CT screening. At the American Society of Clinical Oncology's annual meeting in May 1999, a historical comparison of two screening strategies used by the Anti-Lung Cancer Association in Japan was updated. 36,37 The results of CXR and sputum cytology screenings from September 1975 were compared in the report. from September 1993 to December 1998 (a total of approximately 26,000 screenings) to the same screening strategy plus spiral CT scan. Prior to CT scanning, 43 patients with primary lung cancer were discovered, compared to 36 patients with primary lung cancer during the CT scan period. Furthermore, the proportion of stage IA tumours increased from 42 to 81 percent, and 5-year survival increased from 48 to 82 percent. Because of the potential for lead time and other biases, the results of this nonrandomized, historical comparison suggest that screening with CT scans can improve the ability to diagnose lung cancers at an earlier stage, but it does not constitute strong evidence of a mortality benefit. The Early Lung Cancer Action Project (ELCAP) conducted low-dose lung cancer screenings every year. A single-arm study of 1,000 smokers in New York City used spiral CT scans and CXR. 38 In 233 participants, one to six noncalcified lung nodules were found at baseline (”prevalence”); 27 nodules (2.7%) were malignant. Twenty of the 27 malignant nodules were not found on standard CXR (74 percent); however, no malignant nodules were found on CXR that were not also seen on CT scanning. The majority of these cancers (23 of 27; 85%) were stage I, and all but one nodule (97%) could be removed. Seven cancers were detected in the second year of screening (”incidence”), six of which were stage I. In the United States, only about 20% of sporadic lung cancer diagnoses are stage I. 39 These studies’ findings Table 3 summarises the results of these and related studies. everything There was a high rate of false-positive results in studies, as well as a preponderance of stage I lesions among cancers discovered. Lung cancer is the leading cause of death from cancer worldwide, with 2.09 million new cases and 1.76 million deaths in 2001. In the early 2000s, four case-control studies from Japan found that combining chest radiographs with sputum cytology in lung cancer screening reduced mortality2. In contrast, two randomised controlled trials conducted between 1980 and 1990 found that chest radiograph screening was ineffective in reducing lung cancer mortality3,4. Compared to low-dose computed tomography, chest radiographs are more cost-effective, easier to access, and deliver a lower radiation dose (CT). Excessive false positive (FP) results are another drawback of chest CT. It according to one study, 96 percent of nodules detected by low-dose CT screening are FPs, resulting in unnecessary follow-up and invasive examinations5. In terms of sensitivity, chest radiography is inferior to chest CT, but it is superior in terms of specificity. Taking these factors into account, developing a computer-aided diagnosis (CAD) model for chest radiographs would be beneficial in terms of improving sensitivity while maintaining low FP results. Convolutional neural networks (CNN), a type of deep learning (DL)6,7, have recently been used to achieve dramatic, state-of-the-art improvements in r adiology8. DL-based models have also shown promise for nodule/mass detection on chest radiographs9–13, with sensitivities ranging from 0.51 to 0.84 and mean number of FP indications per image (mFPI) ranging from 0.02 to 0.34. In Furthermore, with these CAD models, radiologist performance for detecting nodules was better than without them9. It can be difficult for radiologists to detect nodules and distinguish between benign and malignant nodules in clinical practise. Because normal anatomical structures can mimic nodules, radiologists must pay close attention to the shape and marginal properties of nodules. Even skilled radiologists can misdiagnose14,15 because these issues are caused by the conditions rather than the ability of the radiologist. Detection and segmentation are the two main methods for detecting lesions using DL. A region-level classification is used for detection, while a pixel-level classification is used for segmentation. The detection method cannot provide as much detail as the segmentation method. In clinical settings, Making a correct diagnosis is more likely when the size of a lesion is classified at the pixel level. Because the shape can be used as a reference during detection, pixel-level classification makes it easier to track changes in lesion size and shape. When determining the effect of treatment, it is also possible to consider not only the long and short diameters, but also the area of the lesion16. There have been no studies using the segmentation method to detect pathologically proven lung cancer on chest radiographs to our knowledge. The goal of this research was to train and validate a DL-based model capable of detecting lung cancer on chest radiographs using the segmentation method, as well as to assess the model's characteristics. Improve sensitivity while maintaining low FP results with this model.

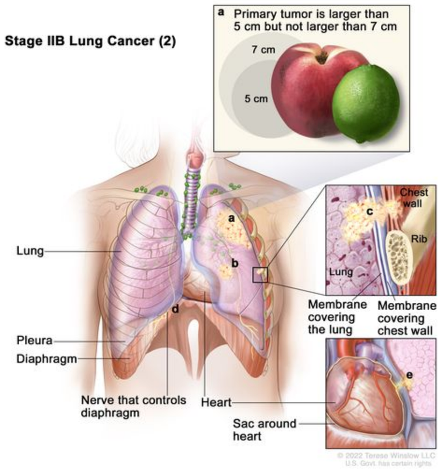

Stage IIB Lung Cancer (2)
a Primary tumor is larger than 5 cm but not larger than 7 cm
7 cm
5 cm
Chest wall
c
Rib
a
b
Lung
Lung
d
Membrane covering the lung
Membrane covering chest wall
Pleura
Diaphragm
Nerve that controls diaphragm
Heart
e
Sac around heart
© 2022 Terese Winslow LLC U.S. Govt. has certain rights

Figure 2.5: Stage IIB of Lung Cancer

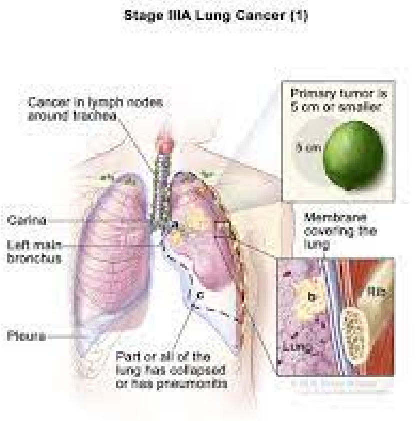

Stage IIIA Lung Cancer (1)
Cancer in lymph nodes around trachea
Primary tumor is 5 cm or smaller
5 cm
Carina
Left main bronchus
Membrane covering the lung
b
Rib
c
Lung
Pleura
Part or all of the lung has collapsed or has pneumonitis

Figure 2.6: Stage IIIB of Lung Cancer

The following points summarise the article's contributions:

- A deep learning-based model for detecting and segmenting lung cancer on chest radiographs was developed in this study.
- Our data is of high quality because all of the nodules/masses were pathologically proven lung cancers that were annotated at the pixel level by two radiologists.

The segmentation method was more informative than the classification or detection methods, which is useful not only for lung cancer detection but also for treatment efficacy and follow-up.

Chapter 3

Chest X-ray Dataset

3.1 Chest X-Ray Images

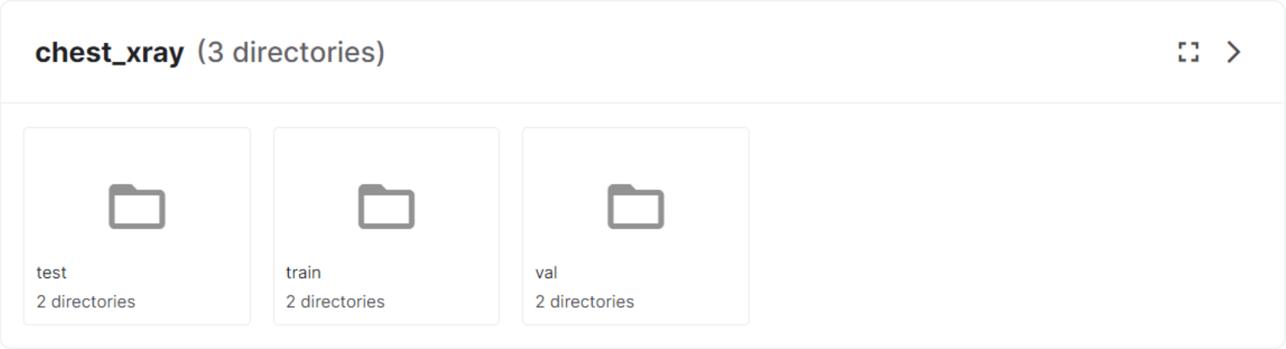
The data is divided into three folders (train, test, and val), with subfolders for each image category (Pneumonia/Normal). There are 5,863 JPEG X-Ray images in two categories (Pneumonia/Normal).

Anterior-posterior chest X-ray images were chosen from retrospective cohorts of children patients aged one to five years old at Guangzhou Women and Children's Medical Center in Guangzhou. All chest X-ray imaging was done as part of the patients' regular medical treatment.

All chest radiographs were originally examined for quality control, with any scans that were low quality or unreadable being removed. The photos' diagnosis were then graded by two experts before being approved for use in training the AI system.

3.1.1 Files Detail and Naming Convention

We used the same transfer learning architecture to diagnose paediatric pneumonia to test the generalizability of our AI system in the diagnosis of common diseases. According to the World Health Organization (WHO), pneumonia kills around 2 million children under the age of five every year, making it the top cause of paediatric death (Rudan et al., 2008). It kills more children than HIV/AIDS, malaria, and measles combined (Adegbola, 2012). According to the World Health Organization, nearly all cases of new-onset childhood clinical pneumonia (95 percent) occur in developing nations, mainly in Southeast Asia and Africa. The two most common causes of pneumonia, bacterial and viral infections, require completely different treatments (Mcluckie, 2009). While bacterial pneumonia necessitates an urgent referral for antibiotic treatment, viral pneumonia is treated with supportive care.



chest_xray (3 directories)

test
2 directories

train
2 directories

val
2 directories

Figure 3.1: Data Directives

Images are not in dcm format, the images are in jpg or png to fit the model

Data contain 1 chest cancer type which is Pneumonia, and 1 folder for the normal cell

Data folder is the main folder that contain all the step folders

inside Data folder are test , train , valid

test represent testing set

train represent training set

valid represent validation set

training set is 70%

testing set is 20%

validation set is 10%

Clinical decision support algorithms for medical imaging encounter issues in terms of reliability and interpretability. We develop a diagnostic tool for screening individuals with common curable blinding retinal disorders based on a deep-learning architecture. Our framework employs transfer learning, which allows us to train a neural network with a fraction of the data required by traditional methods. We exhibit performance comparable to that of human experts in diagnosing age-related macular degeneration and diabetic macular edoema using our approach on a collection of optical coherence tomography pictures. By emphasising the regions recognised by the neural network, we also provide a more clear and interpretable diagnostic. We also use chest X-ray pictures to demonstrate the wide applicability of our AI system for diagnosing juvenile pneumonia.This tool may eventually help to speed up the diagnosis and referral of certain curable illnesses, allowing for earlier treatment and better clinical outcomes."Deep learning-based categorization and referral of curable human diseases" describes and analyses a dataset of verified OCT and Chest X-Ray images. The OCT images are divided into two groups: a training set and a testing set of patients. OCT images are divided into two directories: Pneumonia, and NORMAL, and are labelled as (disease)-(randomized patient ID)-(image number by this patient).

Chapter 4

Proposed Methodology

4.1 Algorithm Details

We have implemented our model using VGG16 Convolutional Neural Network in Keras on chest x-ray dataset

4.1.1 Machine Learning

Machine Learning algorithms come in a wide range of types and are used to solve a wide range of problems. We'll go over some of the most popular models in data science research and production in the sections below. The probabilistic approach of Naive Bayes is used to determine the class of fresh input data. The Bayes Theorem is used in this categorization technique. It works effectively in specific areas while being rather simplistic. Before performing a classification, this technique demands that the data characteristics be independent (i.e., there should be no association between variables). The Support Vector Machine algorithm is based on the concept of determining the largest margin line that minimises classification error between classes.It's a powerful and flexible method that can handle linear and nonlinear classifications, regressions, and outlier detection with ease.Artificial neural networks (ANNs) are a classification and resilient algorithm. ANN has demonstrated amazing outcomes in the research community by replicating the way the brain operates. It is highly known for its ability to solve complex issues in a variety of fields. Making features suitable for these algorithms takes up the majority of the data scientist's effort because they influence the algorithm's outcome more than anything else. The purpose of feature engineering is to build up the data properly for the algorithm requirements and to improve the accuracy of the machine learning models. Many experts and data scientists are searching for and developing high-quality features using a variety of evaluation methodologies.

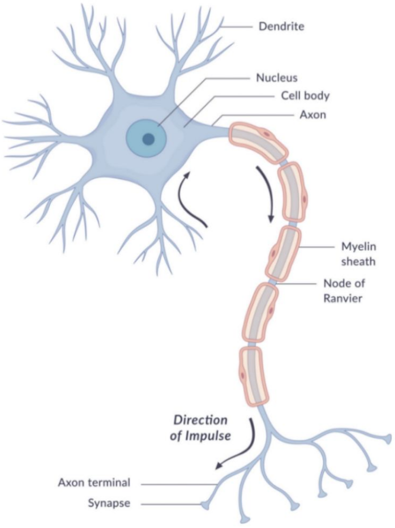


Figure 4.1: Biological neuron

4.1.2 Deep Learning

Deep learning is a type of machine learning that is inspired by how neurons and synapses in the brain transmit information. A perceptron, which was conceived by Frank Rosenblatt in 1957 and can be seen as a simple linear classifier, is the basic unit of a neural network. Biologically, depending on the intensity of the action potential, neuron transfers nerve impulses to neighbouring cells. The nucleus (cell body), dendrites, axon, and axon terminal are all parts of a neuron. Dendrites are analogous to signal receivers; nuclei are analogous to signal processors; and axon and axon terminals are analogous to signal transmitters to surrounding neurons. Synapse is the junction between one neuron's axon and another's dendrites.

The input layer refers to the first section of the neuron, from x1 to xm. The input signal or data from other neurons is received in this layer. The output layer is the final section, and it calculates the neuron's output, as the name implies. There is a nucleus in the middle that processes all of the info. A more complete model is shown in the second image, which incorporates bias b and the activation function. The bias acts as a threshold for neuron activation, while the activation function redistributes the function in a more perceptron-friendly shape. Values in the activation function can change before being input to the cell.To have values ranging from 0 to 1, for example, we use the activation function, which does the work for us.

To summarise, the neuron computes a ”weighted sum” of its inputs, compares it to a threshold b, and then converts the result into a specified range.

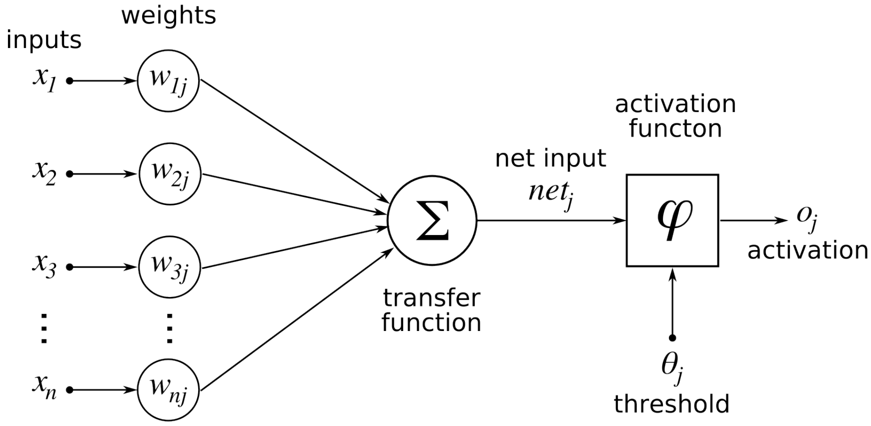


Figure 4.2: Artificial Neuron

4.1.3 Transfer Learning

- Transfer learning is a machine learning method where a model developed for a task is reused as the starting point for a model on a second task
- Transfer learning is an optimization that allows rapid progress or improved performance when modeling the second task.
- Transfer learning only works in deep learning if the model features learned from the first task are general.

Transfer learning is a machine learning technique in which a model created for one job is utilised as the basis for a model on a different task.

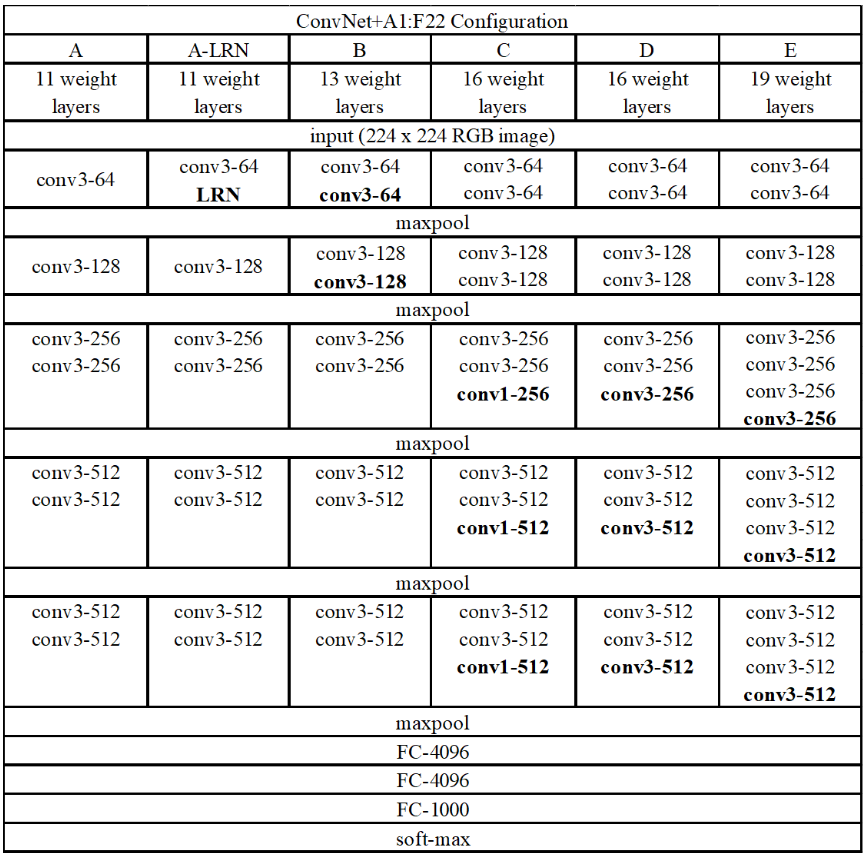
Given the vast compute and time resources required to develop neural network models on these problems, as well as the huge jumps in skill that they provide on related problems, it is a popular approach in deep learning where pre-trained models are used as the starting point on computer vision and natural language processing tasks.

You'll learn how to apply transfer learning to speed up training and increase the performance of your deep learning model in this post.

4.1.4 Convolutional Neural Networks

Artificial intelligence has made significant progress in closing the gap between human and computer capabilities. Researchers and hobbyists alike work on a variety of facets in the subject to achieve incredible results. The field of computer vision is one of several such disciplines.

The goal of this field is to enable machines to perceive the world in the same way that humans do, and to use that knowledge for a variety of tasks such as



ConvNet+A1:F22 Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 x 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Figure 4.3: Table - 1 : Transfer Learning

image and video recognition, image analysis and classification, media recreation, recommendation systems, natural language processing, and so on. Advancements in Computer Vision using Deep Learning have been built and developed through time, mostly through the use of a single algorithm – the Convolutional Neural Network.

A **Convolutional Neural Network (ConvNet/CNN)** is a Deep Learning system that can take an input image, assign relevance (learnable weights and

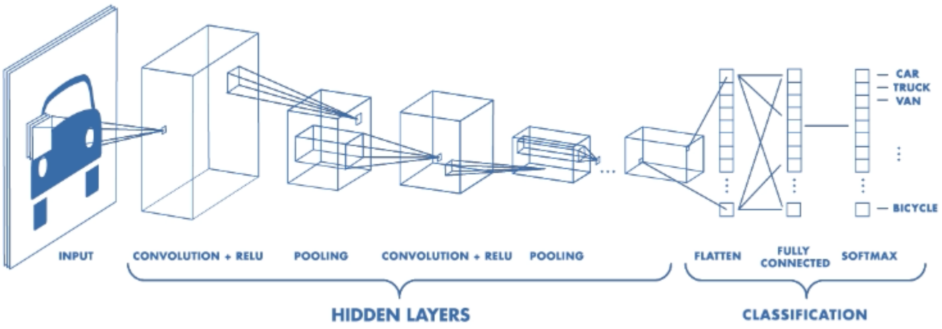


Figure 4.4: CNN Architecture

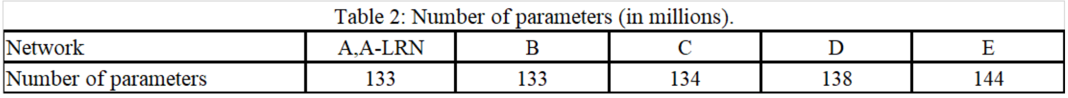


Table 2: Number of parameters (in millions).

Network	A,A-LRN	B	C	D	E
Number of parameters	133	133	134	138	144

Figure 4.5: Table - 2 : Transfer Learning

biases) to various aspects/objects in the image, and distinguish between them. In comparison to other classification algorithms, ConvNet requires substantially less pre-processing. While basic approaches require hand-engineering of filters, ConvNets can learn these filters/characteristics with enough training.

The architecture of a ConvNet is inspired by the organisation of the Visual Cortex and is akin to the connectivity pattern of Neurons in the Human Brain. Individual neurons can only respond to stimuli in a small area of the visual field called the Receptive Field.

4.1.5 VGG16 Model

VGG16 is a convolution neural net (CNN) architecture which was used to win ILSVR(Imagenet) competition in 2014. It is considered to be one of the excellent vision model architecture till date. Most unique thing about VGG16 is that instead of having a large number of hyper-parameter they focused on having convolution layers of 3x3 filter with a stride 1 and always used same padding and maxpool layer of 2x2 filter of stride 2. It follows this arrangement of convolution and max pool layers consistently throughout the whole architecture. In the end it has 2 FC(fully connected layers) followed by a softmax for output. The 16 in VGG16 refers to it has 16 layers that have weights. This network is a pretty large network and it has about 138 million (approx) parameters.

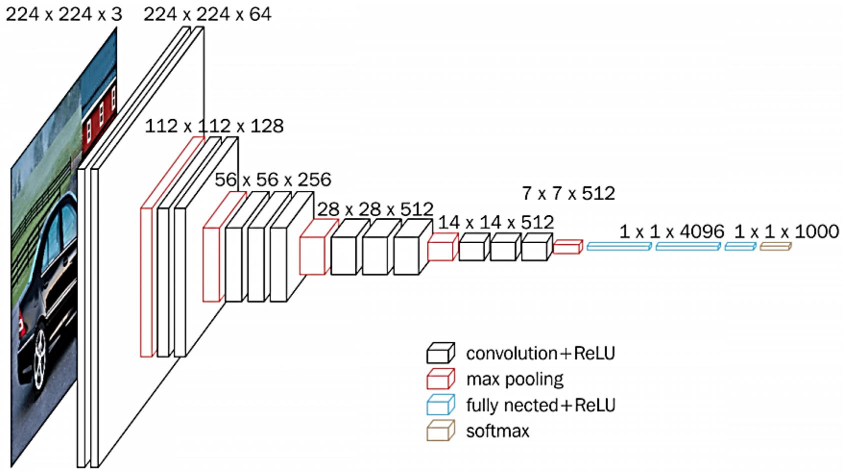
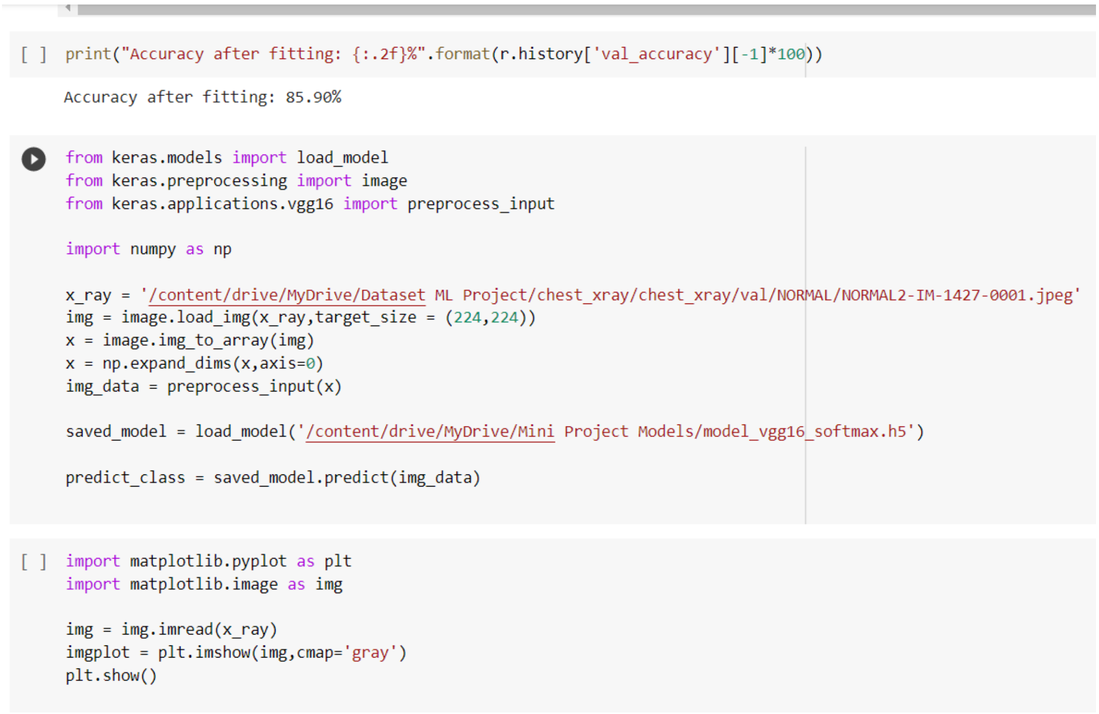


Figure 4.6: VGG16 Architecture



```
[ ] print("Accuracy after fitting: {:.2f}%".format(r.history['val_accuracy'][-1]*100))

    Accuracy after fitting: 85.90%
```

```
from keras.models import load_model
from keras.preprocessing import image
from keras.applications.vgg16 import preprocess_input

import numpy as np

x_ray = '/content/drive/MyDrive/Dataset ML Project/chest_xray/chest_xray/val/NORMAL/NORMAL2-IM-1427-0001.jpeg'
img = image.load_img(x_ray,target_size = (224,224))
x = image.img_to_array(img)
x = np.expand_dims(x,axis=0)
img_data = preprocess_input(x)

saved_model = load_model('/content/drive/MyDrive/Mini Project Models/model_vgg16_softmax.h5')

predict_class = saved_model.predict(img_data)
```

```
[ ] import matplotlib.pyplot as plt
    import matplotlib.image as img

    img = img.imread(x_ray)
    imgplot = plt.imshow(img,cmap='gray')
    plt.show()
```

Figure 4.7: Preprocessing

4.1.6 Implementation

4.1.6.1 Existing System

CNN is a type of deep neural network that focuses solely on data collecting and is not labelled. Visual imagery analysis is the most prevalent use. In comparison to other image classification algorithms, CNN requires very little preprocessing. But It is challenging to obtain precise results. Multiple photos for lung identification in a short time are not relevant.

4.1.6.2 PREPROCESSING

Although the most acceptable input data has been chosen, it must be preprocessed in order for the neural network to produce correct results. This helps the network learn more readily by reducing the amount of inputs. It clears the CT of unwanted impulses. images. Color photos are converted to grey-level coding.

4.1.6.3 Building the Model

A DenseNet architecture is suggested here for classifying lung cancer images. DenseNet is a densely connected convolutional neural network made up of numerous dense blocks with dense connection and transition layers. Unlike standard architectures, which only add L layer connections, a dense block of L layer introduces L(L + 1)/2 connections. Between any two layers, there is a direct

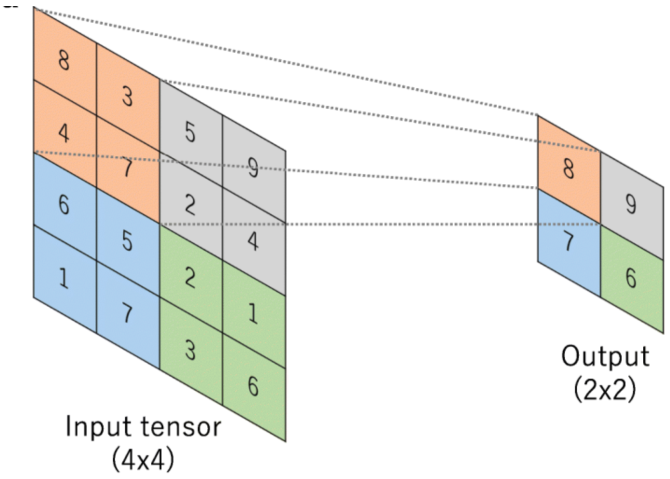


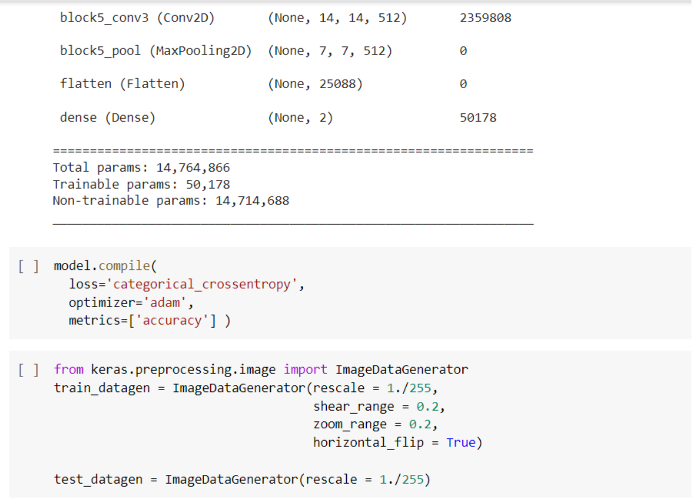
Figure 4.8: Max Pooling

connection. The input to each layer of the network is the sum of all previous layers' outputs, and the feature maps learned by this layer are immediately sent as input to all following layers. a thick block through which feature maps can be concatenated All of the previous feature-maps are included in the lth layer's l inputs.

4.1.7 Max - Pooling

Max Pooling is a type of convolution in which the Kernel collects the maximum value from the area it convolves. Max Pooling basically tells the Convolutional Neural Network that just that information will be carried forward if it is the most amplitude-wise accessible.

Using a 2*2 kernel and a stride of 2, max-pooling on a 4*4 channel: We're using a 2*2 Kernel to convolve with. The channel has four values 8,3,4,7 if we look at the first 2*2 set on which the kernel focuses. Max-Pooling selects "8" as the maximum value from the set. Examine your photograph. Assume your image is 28 * 28 pixels in size. If you zoom in to a receptive field of 5*5, you can see some features in this image. Max Pooling is recommended when you can extract some features. Max pooling is not recommended in the early stages of a Convolutional Neural Network because the Kernels are still extracting edges and gradients.Shift Invariance, Rotational Invariance, and Scale Invariance are all added via Max Pooling. Because we receive the maximum value from the 2 *2 image, a slight alteration or shift does not produce invariance. Shift invariance is the term for this. Max Pooling is also scale-invariant and slightly rotational.



```
 block5_conv3 (Conv2D)       (None, 14, 14, 512)       2359808

 block5_pool (MaxPooling2D)  (None, 7, 7, 512)         0

 flatten (Flatten)           (None, 25088)             0

 dense (Dense)               (None, 2)                 50178

=================================================================
Total params: 14,764,866
Trainable params: 50,178
Non-trainable params: 14,714,688
_________________________________________________________________
```

```
[ ] model.compile(
      loss='categorical_crossentropy',
      optimizer='adam',
      metrics=['accuracy'] )
```

```
[ ] from keras.preprocessing.image import ImageDataGenerator
    train_datagen = ImageDataGenerator(rescale = 1./255,
                                       shear_range = 0.2,
                                       zoom_range = 0.2,
                                       horizontal_flip = True)

    test_datagen = ImageDataGenerator(rescale = 1./255)
```

Figure 4.9: Compile the Model

4.1.8 Why Python ?

Python has become an unwritten standard in the burgeoning machine learning industry for constructing artificial intelligence algorithms that handle complicated problems.

Python was utilised for many purposes. To begin with, Python is simple to use. Most individuals are drawn to it because of its simple and accessible grammar. Python is the go-to language for most machine learning engineers due to the large number of tools and frameworks available. The following libraries are particularly useful for the artificial intelligence field.

TensorFlow is a free soft neural network application; Scikitware is a software library that contains many classification, clustering, and regression algorithms related to machine learning; Fastai is a high level software library that makes the machine learning and deep learning process easy for newcomers.Python supports cross-platform, cross-platform, and extensible operations. Python has proven to be extremely straightforward and well-suited for such tasks, as we previously mentionedlanguage operations which make it very portable the steps GPU have had in the whole creation of these algorithms.

4.1.9 Model Fitting

The ability of a machine learning model to generalise data comparable to that with which it was trained is measured by **model fitting**. When given unknown inputs, a good model fit refers to a model that accurately approximates

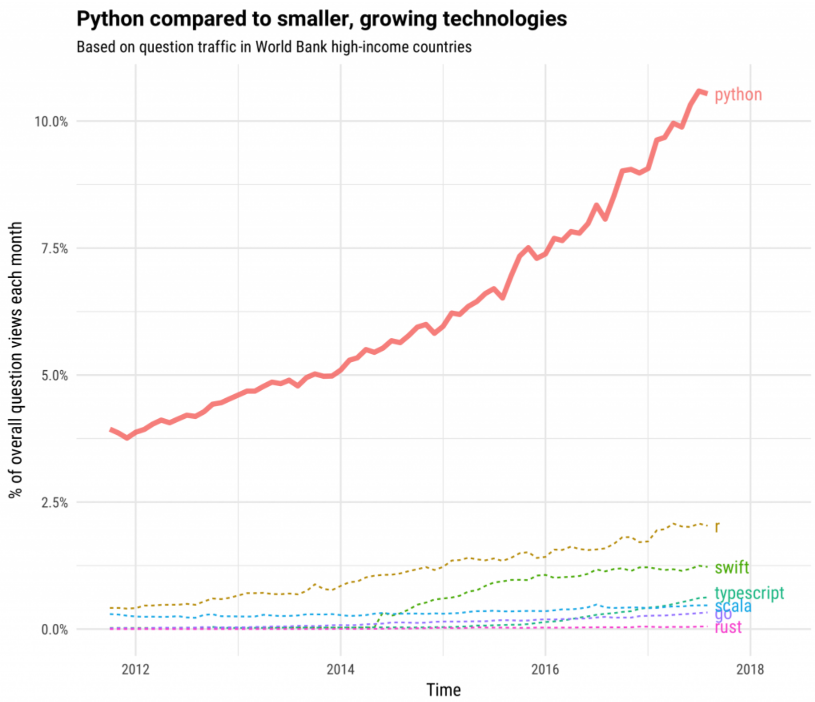


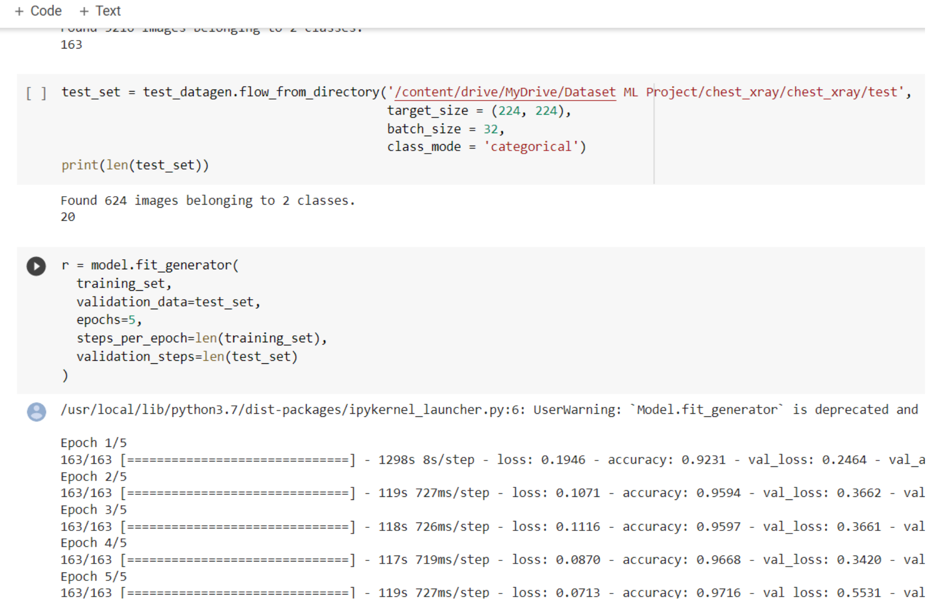
Figure 4.10: Growth in Python Users

the output.

Adjusting the parameters in the model to improve accuracy is referred to as fitting. To construct a machine learning model, an algorithm is performed on data for which the goal variable (”labelled” data) is known. The model’s outputs are then compared against the target variable’s actual, observed values to determine accuracy. The next stage is to tweak the algorithm’s standard parameters to reduce error and improve the model’s accuracy when determining the relationship between the variables.Fitting a model is instructing your algorithm to understand the relationship between predictors and outcomes so that future values of the outcome can be predicted. As a result, the best-fit model has a collection of parameters that best describes the problem at hand.

4.1.10 Plotting the Graphs of Accuracy and Loss

Validation, Training, and Testing The validation model will be used to test the model’s fitness, while the train data will be used to train it. Users can alter hyperparameters such as the number of layers in the network, the number of nodes per layer, the number of epochs, and so on after each run. These modifications are mainly made by trial and error, while visualisation tools like Matplotlib’s plots can assist in achieving ideal outcomes. At the parameter or hyperparameter level,



```
Found 5216 images belonging to 2 classes.
163
```

```
[ ] test_set = test_datagen.flow_from_directory('/content/drive/MyDrive/Dataset ML Project/chest_xray/chest_xray/test',
                                                target_size = (224, 224),
                                                batch_size = 32,
                                                class_mode = 'categorical')
    print(len(test_set))
```

```
Found 624 images belonging to 2 classes.
20
```

```
r = model.fit_generator(
  training_set,
  validation_data=test_set,
  epochs=5,
  steps_per_epoch=len(training_set),
  validation_steps=len(test_set)
)
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6: UserWarning: `Model.fit_generator` is deprecated and

Epoch 1/5
163/163 [==============================] - 1298s 8s/step - loss: 0.1946 - accuracy: 0.9231 - val_loss: 0.2464 - val_a
Epoch 2/5
163/163 [==============================] - 119s 727ms/step - loss: 0.1071 - accuracy: 0.9594 - val_loss: 0.3662 - val
Epoch 3/5
163/163 [==============================] - 118s 726ms/step - loss: 0.1116 - accuracy: 0.9597 - val_loss: 0.3661 - val
Epoch 4/5
163/163 [==============================] - 117s 719ms/step - loss: 0.0870 - accuracy: 0.9668 - val_loss: 0.3420 - val
Epoch 5/5
163/163 [==============================] - 119s 727ms/step - loss: 0.0713 - accuracy: 0.9716 - val loss: 0.5531 - val
```

Figure 4.11: Model Fitting

the Test Set must not be used in the training activity.

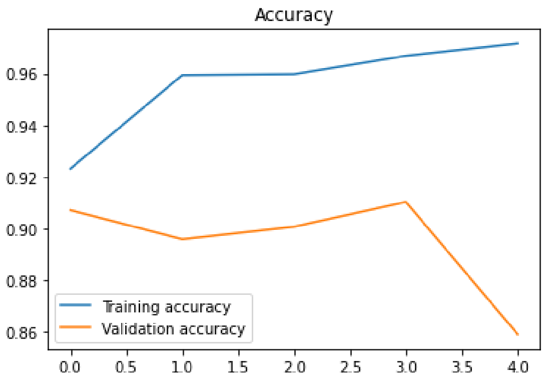


Figure 4.12: Validation Accuracy

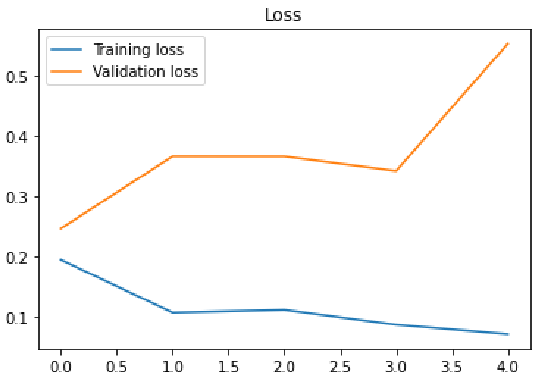


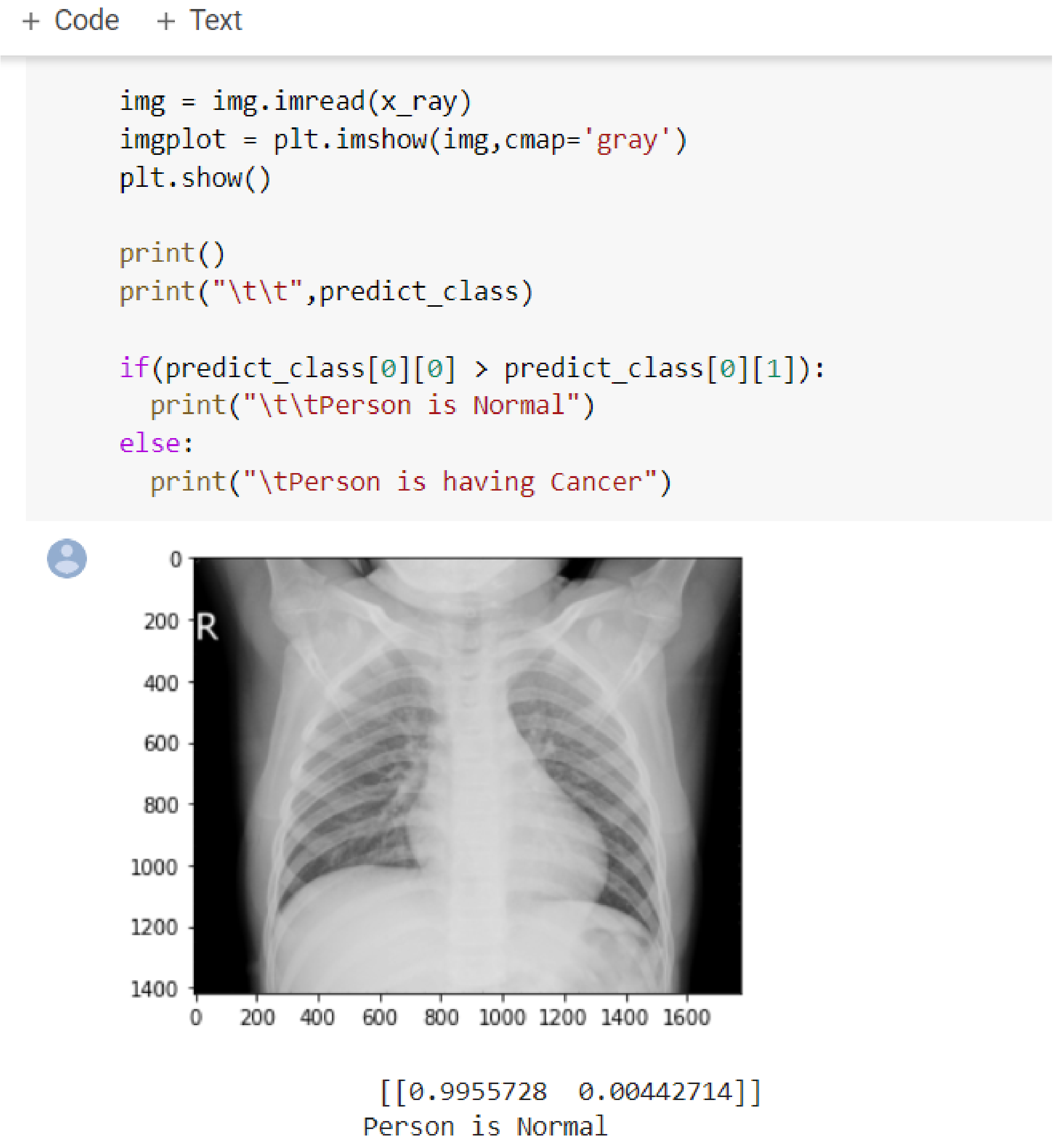
Figure 4.13: Validation Loss

The software below creates a Deep Learning Binary Classification Model. The information is divided into three categories:

- Training set
- Validation set
- Test set

4.1.11 Output

The proposed system use a deep learning based CNN architecture. Where the system improve the accuracy based on learning To improve the accuracy in terms of specificity and sensitivity.



```
img = img.imread(x_ray)
imgplot = plt.imshow(img,cmap='gray')
plt.show()

print()
print("\t\t",predict_class)

if(predict_class[0][0] > predict_class[0][1]):
  print("\t\tPerson is Normal")
else:
  print("\tPerson is having Cancer")
```

```
                [[0.9955728  0.00442714]]
                Person is Normal
```

Figure 4.14: Output of Normal Chest X-ray

Chapter 5

Architecture and UML diagrams

5.1 Architecture Diagram

An architectural diagram is a visual representation that shows how components of a software system are physically implemented. It depicts the overall structure of the software system, as well as the relationships, limitations, and boundaries that exist between each component.

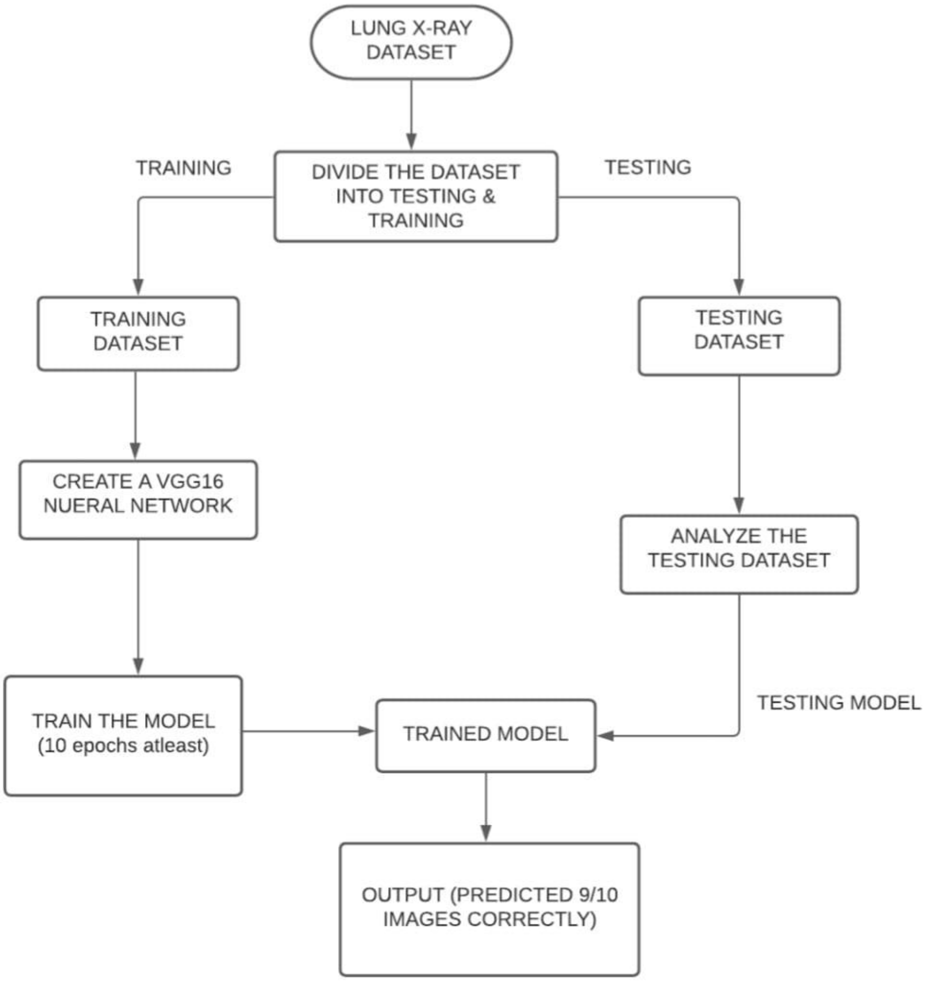

LUNG X-RAY DATASET
TRAINING
DIVIDE THE DATASET INTO TESTING & TRAINING
TESTING
TRAINING DATASET
TESTING DATASET
CREATE A VGG16 NUERAL NETWORK
ANALYZE THE TESTING DATASET
TESTING MODEL
TRAIN THE MODEL (10 epochs atleast)
TRAINED MODEL
OUTPUT (PREDICTED 9/10 IMAGES CORRECTLY)

Figure 5.1: Architecture Diagram of Proposed System

5.2 State chart Diagram

A state diagram is a type of diagram used to describe the behaviour of systems in computer science and related fields. State diagrams assume that the system being described has a finite number of states; this is not always the case, but it is often a reasonable assumption. State diagrams come in a variety of shapes and sizes, each with its own set of semantics.

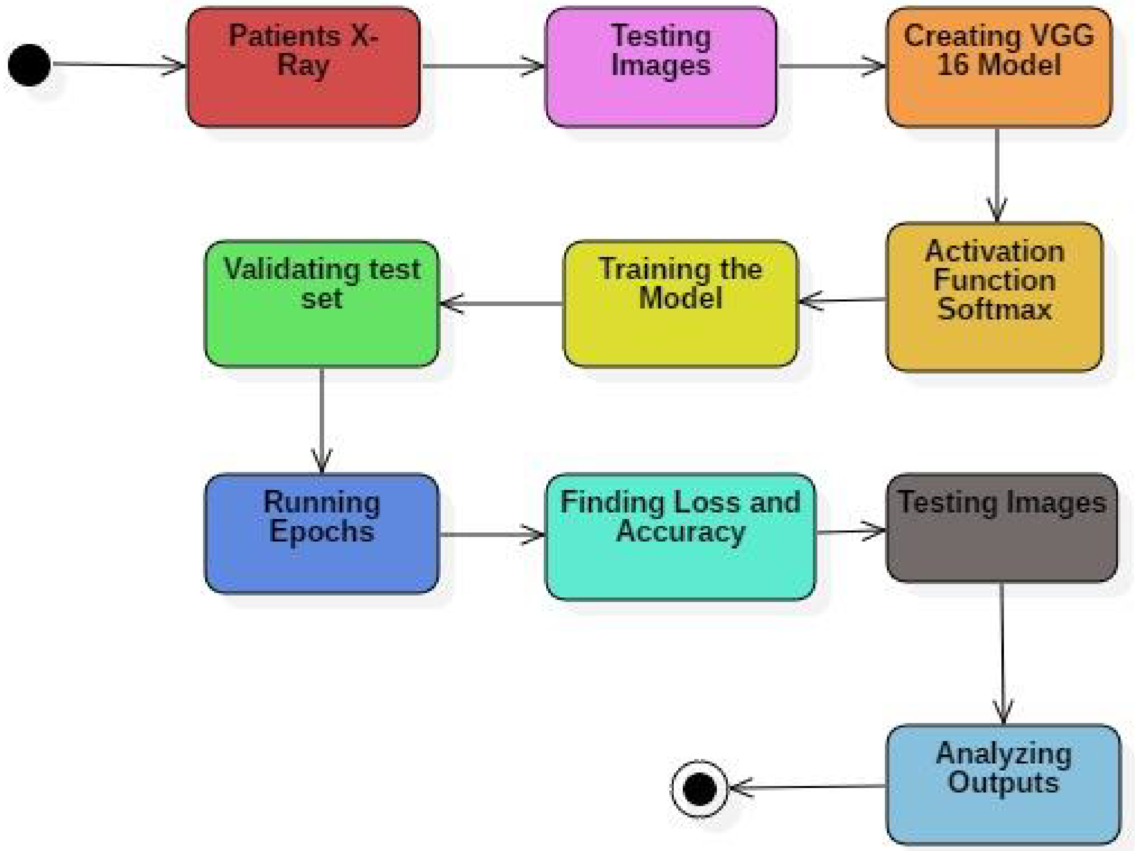


Figure 5.2: State chart Diagram

5.3 Use case Diagram

A use case diagram is a visual representation of how a user might interact with a system. A use case diagram depicts the system's various use cases and different types of users, and is frequently accompanied by other diagrams. Circles or ellipses are used to represent the use cases.

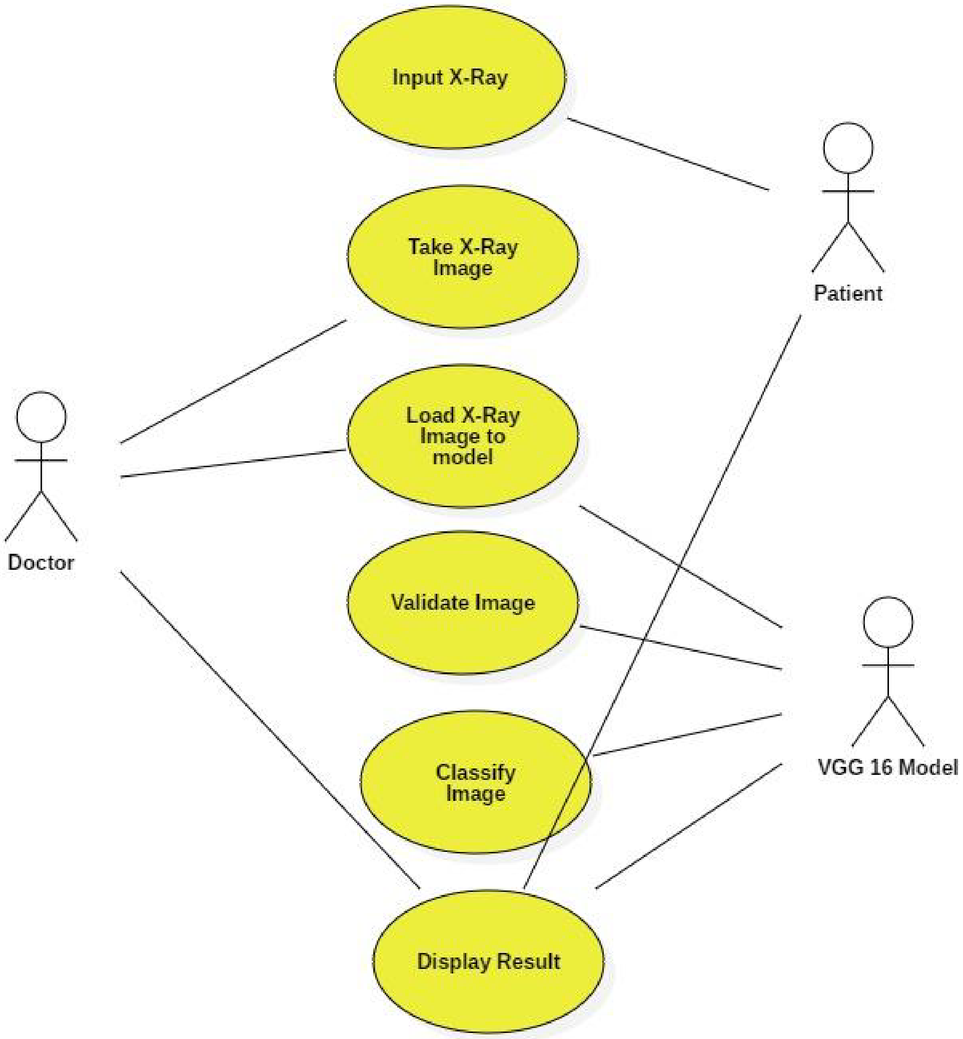


Figure 5.3: Use case Diagram

Chapter 6

Results

6.1 Results

The original data set is divided into two parts: a test set with 20% of the data and a training set with the remaining data. The train set is then split again, with 20% of the train set serving as a validation set and the remainder being used for training. The training set is 64 percent of the whole data set, the validation set is 16 percent, and the test set is 20 percent. The three-layered Neural networks are fed the training data set, with the first two layers each having four nodes and the output layer having only one node. The history object stores the model's loss and accuracy data for each epoch.

We used different activation functions to improve the accuracy of the model so we decided to apply 5 activation function and we noticed the changes in the accuracy of the model, out of five we got Binary sigmoid which is best fit activation funtion for our model.

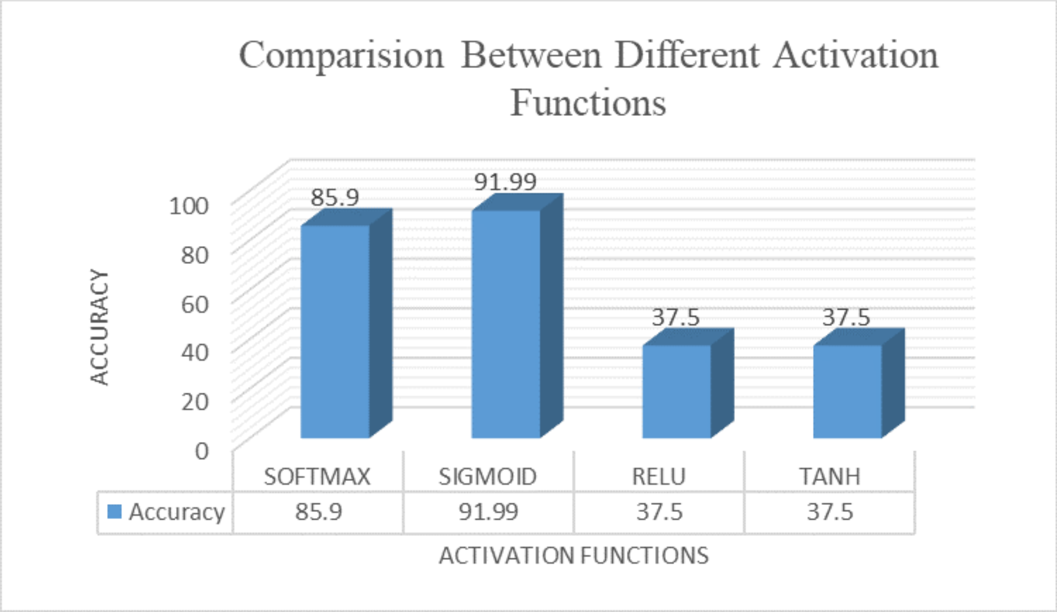
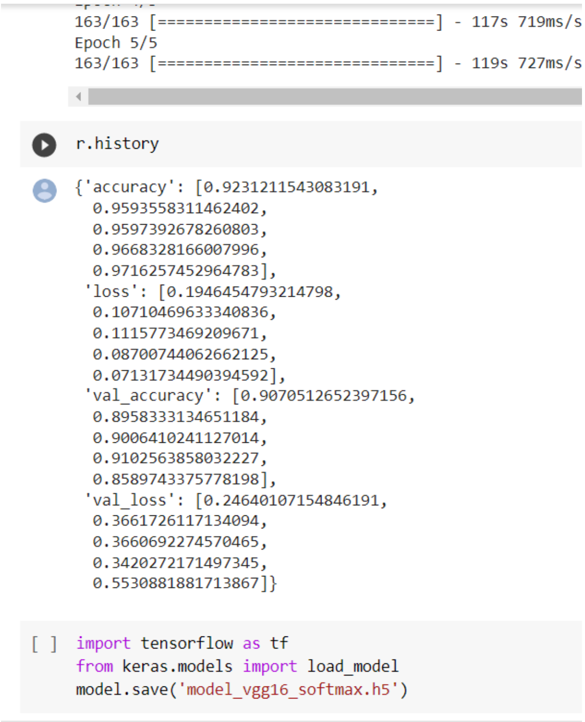


Figure 6.2: Comparision Between Different Activation Functions



```
163/163 [==============================] - 117s 719ms/s
Epoch 5/5
163/163 [==============================] - 119s 727ms/s
```

```
r.history
```

```
{'accuracy': [0.9231211543083191,
  0.9593558311462402,
  0.9597392678260803,
  0.9668328166007996,
  0.9716257452964783],
 'loss': [0.1946454793214798,
  0.10710469633340836,
  0.1115773469209671,
  0.08700744062662125,
  0.07131734490394592],
 'val_accuracy': [0.9070512652397156,
  0.8958333134651184,
  0.9006410241127014,
  0.9102563858032227,
  0.8589743375778198],
 'val_loss': [0.24640107154846191,
  0.3661726117134094,
  0.36606922274570465,
  0.3420272171497345,
  0.5530881881713867]}
```

```
[ ] import tensorflow as tf
    from keras.models import load_model
    model.save('model_vgg16_softmax.h5')
```

Figure 6.1: Result of Proposed Method

6.1.1 Conclusion

Using the DenseNet network, we created an automated lung cancer CT image classification model. The denseNet algorithm is then used to analyse and categorise the lung cancer datasets. Our model delivers improved classification results in CT image categorization of cancers, with a test accuracy of 92.91 percent, according to experimental results. Our innovative approach of lung cancer image classification will aid radiologists' treatment in the future, simplifying the steps of lung cancer diagnosis, improving the accuracy of lung cancer diagnosis, and lowering the rate. Furthermore, we will process the categorization of lung cancer using more high-quality lung cancer CT scans, significantly boosting the network's accuracy.

6.1.2 Future Scope

We can Futher Increase the Accuracy by increasing the dataset or applying new methodologies in Deep Learning as Technology increasing their is lot of scope, every month new research as are coming forward to increase the Model performance. We can also develop app using the model to predict cancer or not like we can develop an Android or Web application, there is lot of scope in this Domain and Futher can be improved

References

[1] S. S. Raoof, M. A. Jabbar, and S. A. Fathima, “Lung cancer prediction using machine learning: A comprehensive approach,” in *2020 2nd International conference on innovative mechanisms for industry applications (ICIMIA)*. IEEE, 2020, pp. 108–115.

[2] S. Janee Alam and A. Hossan, “Multi-stage lung cancer detection and prediction using multi-class svm classifier,” in *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, 2018.

[3] M. G. D. P. Thangaraj, “A computer aided diagnosis system for detection of lung cancer nodules using extreme learning machine,” *International Journal of Engineering Science and*.

[4] S. Kakeda, J. Moriya, H. Sato, T. Aoki, H. Watanabe, H. Nakata, N. Oda, S. Katsuragawa, K. Yamamoto, and K. Doi, “Improved detection of lung nodules on chest radiographs using a commercial computer-aided diagnosis system,” *American Journal of Roentgenology*, vol. 182, no. 2, pp. 505–510, 2004.

[5] M. N. Gurcan, B. Sahiner, N. Petrick, H.-P. Chan, E. A. Kazerooni, P. N. Cascade, and L. Hadjiiski, “Lung nodule detection on thoracic computed tomography images: Preliminary evaluation of a computer-aided diagnosis system,” *Medical Physics*, vol. 29, no. 11, pp. 2552–2558, 2002.

[6] K. Awai, K. Murao, A. Ozawa, M. Komi, H. Hayakawa, S. Hori, and Y. Nishimura, “Pulmonary nodules at chest ct: effect of computer-aided diagnosis on radiologists’ detection performance,” *Radiology*, vol. 230, no. 2, pp. 347–352, 2004.

[7] S. C. Cheran and G. Gargano, “Computer aided diagnosis for lung ct using artificial life models,” in *Seventh International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC’05)*. IEEE, 2005, pp. 4–pp.

[8] S. Lakshmanaprabu, S. N. Mohanty, K. Shankar, N. Arunkumar, and G. Ramirez, “Optimal deep learning model for classification of lung cancer on ct images,” *Future Generation Computer Systems*, vol. 92, pp. 374–382, 2019.

[9] W. Ausawalaithong, A. Thirach, S. Marukatat, and T. Wilaiprasitporn, “Automatic lung cancer prediction from chest x-ray images using the deep learning approach,” in *2018 11th Biomedical Engineering International Conference (BMEiCON)*. IEEE, 2018, pp. 1–5.