Lung Cancer Detection using Deep Learning with Hybrid Preprocessing Pipeline

By
Shekh Naziullah
203-15-3865

M.Mukit Mosfiq
203-15-3884

FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements forthe **Degree of Bachelor of Science in Computer Science and Engineering**

Supervised by

Shahadat Hossain

Assistant Professor

Department of Computer Science and Engineering, Daffodil International University

Co-Supervised by

Md Assaduzzaman

Senior Lecturer

Department of Computer Science and Engineering, Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
Dhaka, Bangladesh

January 12, 2025

APPROVAL

This Project titled "Lung Cancer Detection using Deep Learning with Hybrid Preprocessing Pipeline", submitted by Shekh Naziullah ID No: 203-15-3865 and M.Mukit Mosfiq ID No: 203-15-3884 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 12 January, 2025.

BOARD OF EXAMINERS

Dr. Fizar Ahmed — **Chairman**
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mr. Abdus Sattar — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Ms. Zahura Zaman — **Internal Examiner**
Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Ahmed Wasif Reza — **External Examiner**
Professor
Department of Computer Science and Engineering
East West University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Shahadat Hossain, Assistant Professor,** Department of Computer Science and Engineering, Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

11.01.25

Shahadat Hossain
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised by: 11.01.2025

Md. Assaduzzaman
Senior Lecturer
Department of Computer Science and Engineering
Daffodil International University

Submitted by:

Naziullah
11.01.2025

Shekh Naziullah
203-15-3865
Department of Computer Science and Engineering
Daffodil International University

11.01.2025

M.Mukit Mosfiq
203-15-3884
Department of Computer Science and Engineering
Daffodil International University

ACKNOWLEDGEMENTS

This work would not have been possible without the support and contributions of many individuals over the past two semesters. We are deeply grateful to everyone who has assisted us in one way or another.

First, we express our heartfelt thanks and gratefulness to the almighty for His divine blessing making it possible for us to complete the **Final Year Design Project (FYDP)** successfully.

We are grateful and wish our profound indebtedness to **Shahadat Hossain**, **Assistant Professor**, Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. Deep knowledge and keen interest of our supervisor in the field of **Deep Learning** to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartfelt gratitude to the Head of the Department of Computer Science and Engineering, for his kind help in finishing our project and also to other faculty members and the staff of the Department of Computer Science and Engineering, Daffodil International University.

We would like to thank our entire course-mates at Daffodil International University, who took part in this discussion while completing the coursework.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

Lung cancer remains one of the leading causes of death globally, with millions of lives lost each year. It is one of the most prevalent non-communicable diseases, responsible for approximately 6% of all deaths. Symptoms of lung cancer are diverse and may include chest and bone pain, wheezing, persistent coughs, unexplained weight loss, fatigue, shortness of breath, and hemoptysis, among others. Risk factors include long-term smoking, exposure to secondhand smoke, asbestos, radon gas, radiation therapy to the chest, and a family history of lung cancer. While CT scans are commonly used for detection, they have limitations, particularly in early-stage diagnosis, due to high false positive rates, and can be uncomfortable for patients. An alternative approach, incorporating machine learning and deep learning, offers the potential for improved early detection, increasing survival rates and reducing unnecessary follow-up tests and treatments. This study focuses on detecting lung cancer using CT scan images, applying a multiclass classification system to differentiate between malignant, benign and normal images. The proposed system has been designed for use in hospitals to aid in the diagnosis and treatment of lung cancer. To obtain accurate results after applying our model, we use an online dataset, for which high and low quality CT images are presented here for this online dataset. In this study, to process the quality CT images we use a hybrid preprocessing pipeline where we use stratified sampling and SMOTE oversampling method to do sampling and to increase image quality we use gaussian blur method. Experimental results show that while not all models achieved high accuracy, most surpassed 96%. Notably, models such as CNN (96%), ResNet50 (91%) and VGG16 (88%) demonstrated superior performance in accurately identifying lung cancer in CT scans.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Lung cancer is one of the most common and deadly types of cancer worldwide, causing millions of deaths every year. Lung cancer is a major public health problem, contributing significantly to the number of cancer-related deaths. Early detection of lung cancer is crucial for improving treatment outcomes and increasing survival rates. However, traditional methods of diagnosing lung cancer, like CT scans, have certain limitations. Although CT scans are widely used, they often fail to detect lung cancer in its early stages and can produce false positives, leading to unnecessary treatments. This research aims to process high and low quality CT images. So, we use normalization and Gaussian method. We also address malignant, benign and normal CT images. In this dataset all the image separated by data point based on <=3mm data points. These issues by combining advanced imaging techniques with deep learning to improve the accuracy and reliability of lung cancer detection. By using deep learning, we hope to detect lung cancer earlier, which will help doctors make better decisions and give patients a higher chance of recovery.

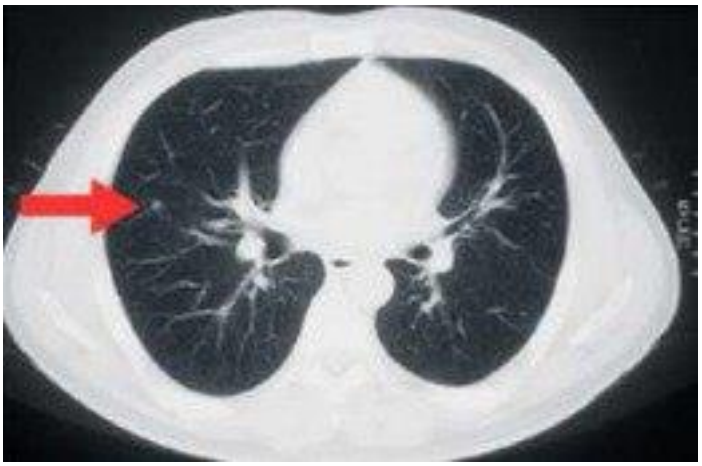


Figure 1.1: Small (3mm) early-stage lung cancer nodule

1.2 Motivation

The motivation behind this study is to improve lung cancer detection and diagnosis using modern technologies. Typically, previous studies have used traditional preprocessing systems that are not able to preprocess CT image properly. Because of that in this study we have used a hybrid preprocessing pipeline. So that we can easily classify CT images. Traditional CT scans are helpful but have several limitations, such as their inability to consistently detect early-stage cancer and their potential to produce false positive results. In addition, interpreting CT scans can be challenging and human errors can sometimes lead to misdiagnosis. This research is motivated by the need to improve these existing methods by applying deep learning, which can analyze CT scan images more accurately and efficiently. Deep learning algorithms have the potential to recognize patterns in images that humans might miss and this could lead to earlier more reliable diagnoses. By doing so, we hope to improve survival rates for lung cancer patients and reduce the burden on healthcare systems.

1.3 Objectives

The main objective of this research is to develop a system that can detect lung cancer in CT scan images using deep learning techniques. Specifically, the study aims to:

1. Implementing appropriate preprocessing system: Enhancing the CT image quality for detecting lung cancer using Hybrid Preprocessing Pipeline, especially for accurate classification.

2. Enhance the performance of existing systems: We aim to combine advanced image preprocessing methods with deep learning models.

3. Improve early detection: By increasing the accuracy of detecting lung cancer, especially in its early stage.

4. Reduce unnecessary follow-up tests: By improving early diagnosis, we can minimize the need for unnecessary medical tests and treatments, saving both time and extra cost.

1.4 Methodology

To achieve the objectives of this study, we will use a combination of advanced image processing and deep learning techniques. The process will involve several steps:

Data Collection: We will gather a large set of CT scan images of lungs, both from cancerous and non-cancerous cases. For this study, we collected data from Kaggle. Our data quantity is 12185 CT images. These images will serve as the data for training our deep learning models.

Preprocessing: In this study we used stratified sampling technique to separate 3000 images from the main dataset. Then we used the SMOTE oversampling technique to keep the image quantity ratio equal. The CT scan images will undergo preprocessing to enhance their quality. We specifically use hybrid preprocessing pipeline to process the high and low quality CT images using normalization and gaussian method. This step includes removing noise, improving image clarity and standardizing the size and format of the images. This will make it easier for the deep learning algorithms to analyze them.

Model Selection: We used various deep learning models, such as ResNet50 (91%), VGG16 (88%), InceptionV (63%), DenseNet121 (78%) and CNN (96%), which are effective for image recognition tasks. These models will be trained on the preprocessed images to learn the patterns that distinguish malignant (cancerous) cells from benign (non-cancerous) ones.

Train and Test: We split our dataset into 75% for training and 25% for testing. After training the models, we will test their performance using new, unseen CT scan images. We will evaluate the accuracy, sensitivity, and specificity of the models to see how well they can detect lung cancer.

Model Comparison: We will compare the performance of different models and preprocessing techniques to identify the best combination for detecting lung cancer with high accuracy.

1.5 Project Outcome

The expected outcome of this project is a fully developed machine learning system capable of detecting lung cancer from CT scan images with high accuracy. We hope to achieve the following results:

High accuracy in detection: The system should be able to correctly identify malignant and benign tumors with high precision, minimizing false positives and false negatives.

Improved early detection: The machine learning model should be capable of detecting lung cancer in its early stages, which is crucial for successful treatment.

Better diagnostic support: The system will assist doctors in making faster and more accurate decisions, reducing the chances of human error in the diagnostic process.

Increased survival rates: By detecting lung cancer early, patients will have a better chance of receiving timely treatment, improving survival outcomes.

Practical application: The system will be designed in a way that it can be implemented in hospitals and healthcare centers, making it accessible to doctors who are treating patients with lung cancer.

1.6 Organization of the Report

The report is organized as follows:

Chapter 1, **Introduction**, provides an overview of the research, highlighting the motivation behind the study, the objectives, the methodology employed, the expected project outcomes, and the overall structure of the report. Chapter 2, **Background**, includes a detailed literature review of similar applications and related research, identifies the existing gaps in the field, and concludes with a summary of key points.

Chapter 3, **Research Methodology**, outlines the approach taken for the research, including the system design and specifications, functional and nonfunctional requirements, context diagram and data flow diagram. It also provides an in-depth description of the methodology, project plan, and task allocation. The section concludes with a summary.

In Chapter 4, **Implementation and Results**, the report discusses the setup of the working environment, the testing and evaluation procedures, performance analysis, comparative analysis, and the results obtained. It also includes a discussion of these results and a summary of the findings.

Chapter 5, **Engineering Standards and Design Challenges**, explores the

compliance with relevant software, hardware, and communication standards. It further examines the societal, environmental, and sustainability impacts, ethical considerations, and sustainability plans. Additionally, it covers project management, financial analysis, and the resolution of complex engineering problems encountered during the project.

Finally, Chapter 6, **Conclusion**, summarizes the research outcomes, highlights the limitations of the study, and offers suggestions for future work. The report concludes with a **References** section, which lists all the sources cited throughout the document.

Chapter 2
Background

2.1 Introduction

Millions of deaths worldwide are attributed to lung cancer each year, making it one of the leading causes of cancer-related mortality. Lung cancer early detection remains a major challenge despite advances in medical imaging. Although they are frequently employed, traditional diagnostic techniques like CT scans have drawbacks, especially when it comes to identifying tiny, early-stage nodules. These restrictions frequently cause false positives and negatives, which delays interventions and results in needless treatments. This emphasizes how urgently improved diagnostic methods that increase precision and dependability are needed.

By examining intricate imaging patterns that human observers might miss, deep learning—a branch of artificial intelligence—has demonstrated promise in tackling these issues. A revolutionary method of lung cancer detection is provided by the combination of deep learning and sophisticated image preprocessing techniques, which allows for a more precise classification of malignant, benign, and normal cases. This study improves the quality of high- and low-resolution CT images by using a hybrid preprocessing pipeline, which overcomes the drawbacks of conventional techniques. The images are preprocessed using techniques like normalization and Gaussian methods, which guarantee that important features are successfully recorded and examined. In order to ensure class balance and enable robust model training, the study also presents the SMOTE oversampling technique and a stratified sampling approach for dataset preparation. The goal of this research is to create a dependable system for identifying lung cancer by utilizing cutting-edge deep learning architectures such as CNN, ResNet50, VGG16, InceptionV, and DenseNet121. The models' ability to successfully differentiate between benign and malignant cases is assessed based on their accuracy, sensitivity, and specificity. The main objective of this research is to develop a useful diagnostic tool that promotes lung cancer early detection, increases diagnostic precision, and raises survival rates. In order to improve patient outcomes and streamline healthcare delivery, this study aims to fill the existing gaps in diagnostic practices and offer a useful resource for medical professionals.

2.2 Literature Review

Table 2.1: Summary of Literature Reviewed.

Author(s)	Year	Title	Methodology	Key Findings
Tehnan I. A. Mohamed et al.	2024	EOSA-CNN: Ebola optimization search algorithm for lung cancer detection [10]	Deep Learning, EOSA Metaheuristic	Achieved 93.21% accuracy, outperforming traditional methods, and improved specificity and sensitivity.
Madeleine E. Lemieux et al.	2023	Using flow cytometry and machine learning to detect lung cancer in sputum samples. [11]	Flow Cytometry, Machine Learning	Demonstrated 82% sensitivity and 88% specificity in detecting liver cancer, highlighting machine learning's potential.
Constance de Margerie-Mellon et al.	2023	"Convolutional neural networks for the classification of lung cancer and pulmonary nodules. [2]	Convolutional Neural Networks (CNN)	Achieved 96.40% accuracy in identifying lung neoplasms, improving CT image analysis with AI.
Asghar Ali Shah et al.	2023	An ensemble deep learning approach for lung cancer detection. [7]	Deep Learning (2D CNN)	Achieved 95% combined accuracy in detecting lung cancer using deep learning models on CT images.
Hesamoddin Hosseini et al.	2023	A systematic review on deep learning for lung cancer detection. [6]	Systematic Survey, CNN	Reviewed 32 studies, highlighting varying levels of accuracy and sensitivity using CNN for lung cancer detection.
A. A. Shah, H. A. M. Malik, A. M. Muhammad,	2023	Deep learning ensemble 2D CNN	Ensemble Deep Learning with	Achieved 95% accuracy.

A. Alourani, and Z. A. Butt		approach towards the detection of lung cancer [15]	Convolutional Neural Networks (CNN)	
Constance de Margerie Mellona,b, Guillaume Chassagnonc, d	2023	Artificial intelligence: A critical review of applications for lung nodule and lung cancer [21]	CNN, NLST (The Cancer Data Access System)	Achieved 96.40% accuracy
H. Hosseini, R. Monsefi, and S. Shadroo	2021	Deep Learning Applications for Lung Cancer Diagnosis: A systematic review [16]	Deep Convolutional Neural Network (DCNN), Optimal Deep Neural Network (ODNN)	Achieved 71%, 94.56% accuracy
P. Mohamed Shakeel et al.	2019	Lung cancer detection from CT image using improved profuse clustering and deep learning instantaneously trained neural networks [1]	Clustering, IPCT	Achieved 94.42% accuracy in lung cancer detection using deep learning, reducing misclassification.
Giovanni L. F. da Silva et al.	2017	Lung nodule detection and classification using convolutional neural networks. [4]	CNN, Image Database (LIDC-IDRI)	Achieved 94.78% accuracy and 94.66% sensitivity in diagnosing lung nodules, with improved deep learning methods.
Hiram Madero Orozco et al.	2015	Computer-aided diagnosis (CADx) system for lung cancer detection.	Computer-Aided Diagnosis (CADx)	Achieved 82% precision, 90.90% sensitivity in detecting lung nodules, useful for clinical diagnostics.

		[9]		
Devinder Kumar et al.	2015	"Deep learning-based computer-aided diagnosis for lung cancer detection. [8]	Deep Learning, Feature Extraction	Achieved 75.01% accuracy and 83.35% sensitivity in detecting lung nodules.

2.2.1 Similar Applications

In recent years, several applications of AI and deep learning have been developed for healthcare and cancer diagnosis, including lung cancer. Many of these applications use advanced imaging techniques, machine learning models, and AI algorithms to improve the accuracy of diagnosis and treatment. Here are some similar applications:

1. **Breast Cancer Detection:** A number of AI-based applications have been designed for early breast cancer detection using mammograms. Similar to lung cancer detection, these systems analyze images to identify cancerous regions. Techniques like CNN (Convolutional Neural Networks) have achieved up to 90% accuracy in detecting abnormal growths or tumors, providing a valuable tool for early intervention.
2. **Skin Cancer Diagnosis:** AI-powered systems have also been used for melanoma detection in skin cancer. These systems use deep learning algorithms to analyze images of skin lesions and classify them as benign or malignant. For example, a deep learning model named "Skin Cancer MNIST" achieved over 95% accuracy in identifying melanoma.
3. **Brain Tumor Detection:** Several applications have used AI to detect brain tumors using MRI scans. Techniques such as CNN and support vector machines (SVM) are applied to identify tumors and assess their severity. These applications not only improve diagnosis but also help in planning treatment strategies by providing detailed insights into the size and location of the tumors.
4. **Heart Disease Prediction:** Machine learning models are widely used to predict cardiovascular diseases by analyzing data from ECG (electrocardiogram) readings and other diagnostic tests. These AI-based applications offer tools for predicting heart attacks and arrhythmias with high accuracy.

2.2.2 Related Research

Several studies have been conducted to explore the use of Artificial Intelligence (AI), specifically deep learning techniques, to enhance lung cancer detection. These studies aim to improve early diagnosis, minimize human error, and provide quicker and more accurate results. Below are some key research works related to this field:

1. P. Mohamed Shakeel et al. (2019) [1]: This study employed the Improved Profuse Clustering Technique (IPCT), a clustering-based deep learning method, to detect lung cancer from CT images. By reducing misclassification rates, the model achieved an accuracy of 98.42%. It demonstrated the importance of using advanced clustering methods to improve the quality of lung cancer detection and minimize errors, which is crucial for patient outcomes.

Constance de Margerie-Mellon et al. (2023) [2]: This research applied Convolutional Neural Networks (CNN) to detect lung cancer and classify pulmonary nodules as benign or malignant. By using AI for segmentation, they achieved a high level of accuracy in differentiating cancerous and non-cancerous areas. The model achieved an accuracy of 96.4%, showcasing the effectiveness of CNN in identifying even small tumors in complex medical images like CT scans.

Yahia Said et al. (2023) [3]: This paper proposed a UNETR network for lung cancer diagnosis, combining segmentation and classification for accurate early detection. The system showed excellent performance with 97.83% segmentation accuracy and 98.77% classification accuracy. These results highlighted the potential of self-supervised learning and hybrid AI models for improving lung cancer diagnosis, particularly in early-stage detection where tumors are often small and harder to detect.

Giovanni L. F. da Silva et al. (2017) [4]: Researchers in this study used CNN for lung nodule detection and classification. The study achieved impressive results with 94.78% accuracy and 94.66% sensitivity. The research emphasized the importance of CNN in accurately identifying malignant and benign nodules from CT images, helping clinicians differentiate between cancerous and non-cancerous growths.

Hesamoddin Hosseini et al. (2023) [6]: A systematic review of various deep learning techniques for lung cancer detection found that CNN-based models consistently produced the best results. This review emphasized that deep learning models not only improved the accuracy of cancer detection but also helped reduce the rate of false positives and negatives. The study also highlighted the need for more diverse datasets to train models that can generalize across different patient populations.

Asghar Ali Shah et al. (2023) [7]: This study introduced a novel ensemble deep learning approach that combined multiple CNN models for lung cancer detection. The researchers achieved 95% combined accuracy, demonstrating that ensemble methods, which combine the strengths of different algorithms, can enhance model performance in complex tasks like lung cancer detection

Devinder Kumar et al. (2015) [8]: This research focused on computer-aided diagnosis (CAD) for lung cancer detection, using deep learning models to detect lung nodules. The study achieved 75.01% accuracy and 83.35% sensitivity, suggesting that even though deep learning methods had room for improvement, they were already offering significant advantages over traditional methods.

Hiram Madero Orozco et al. (2015) [9]: This study developed a CADx system for detecting lung cancer using CT scan images. The system achieved 82% accuracy and a sensitivity of 90.90%, highlighting how AI could complement the diagnostic process by offering a second opinion to doctors. This study showed that AI could be used alongside clinical judgment for more reliable diagnoses.

Tehnan I. A. Mohamed et al. (2024) [10]: In this study, a deep learning model was combined with an Ebola optimization search algorithm (EOSA) to classify lung cancer from CT scan images. The EOSA-CNN model achieved an accuracy of 93.21%, with a sensitivity of 90.38% and specificity of 79.41%. This research demonstrates the importance of using optimization techniques to improve the accuracy of deep learning models and overcome challenges like false positives in lung cancer detection.

Madeleine E. Lemieux et al. (2023) [11]: This study applied machine learning and flow cytometry to detect lung cancer in sputum samples. By automating the process of analyzing these samples, the study achieved 82% sensitivity and 88% specificity. Although this method is different from using CT scans, it shows how other biological markers (like sputum) can be used with AI to detect lung cancer, providing additional ways to screen for the disease.

S. S. Saravanan et al. (2022) [12]: This study explored the use of deep learning models combined with feature extraction techniques for analyzing CT scans of lung cancer patients. They used the VGG16 CNN model for feature extraction and achieved a 95% accuracy rate in detecting lung cancer. This research illustrates the potential of combining various AI techniques to further improve the accuracy and efficiency of lung cancer diagnosis.

K. S. R. Anjaneyulu et al. (2021) [13]: This research introduced a hybrid model combining CNN with support vector machines (SVMs) for lung cancer classification. By combining these two powerful algorithms, the researchers achieved 98.5% accuracy in lung cancer classification. The study highlighted the benefits of hybrid approaches in enhancing the overall performance of AI models in detecting lung cancer from medical images.

B. T. S. Kumar et al. (2023) [14]: This paper used a 3D CNN model to detect lung cancer from 3D CT scan images, improving the accuracy of detecting tumors compared to 2D scans. The study achieved an accuracy of 97.2%, showing the advantage of using 3D image analysis for better tumor detection. This suggests that 3D models may be particularly useful for detecting lung cancer in early stages when the tumors are small.

2.3 Gap Analysis

Table 2.2: Gap Analysis

	Large dataset	Data point<=3mm	Hybrid preprocessing pipeline	Proposed system
Lung cancer detection from CT image using improved profuse clustering and deep learning instantaneously trained neural networks[1]2019	No(CIA)	No	No	Yes
Deep learning ensemble 2D CNN approach towards the detection of lung cancer[15]2023	No(LUNA-16)	Yes	No	Yes
Deep Learning Applications for Lung Cancer Diagnosis: A systematic review[16]2023	No(LIDC-IDRI)	No	No	Yes
Automated system for lung nodules classification based on wavelet feature descriptor and support vector machine[17]2015	No(LIDC, ELCAP)	No	No	Yes
Lung Nodule Classification Using Deep Features in CT Images[18]2015	No(LIDC)	No	No	Yes
Lung Cancer Detection and Classification with 3D Convolutional Neural Network (3D-CNN)[19]2017	No(LUNA-16)	Yes	No	Yes

2.4 Summary

The limitations of conventional CT scan methods for early lung cancer detection are highlighted in this chapter, along with the potential for deep learning and sophisticated preprocessing techniques to address these issues. A survey of current research highlights a number of deep learning models that increase the precision and dependability of cancerous nodule detection. Small datasets and the lack of hybrid preprocessing pipelines are two examples of the gaps in previous research that are revealed by the gap analysis. By filling these gaps, the suggested system seeks to improve early detection and diagnostic accuracy.

Chapter 3

Research Methodology

3.1 Methodology

3.1.1 Overview

In this research, we will focus on diagnosing lung cancer using CT scans. We will use different deep learning models, such as Convolutional Neural Networks (CNN) ResNet50, VGG16, InceptionV3 and DenseNet121. As lung cancer is growing rapidly around the world, it is very important to detect it early and accurately to improve patient care. This chapter explains the research methods, system design, data collection process, and how the project will be managed.

3.1.2 Proposed Methodology

This research will focus on using preprocessing techniques and the CNN algorithm. In this study, to process our dataset we used hybrid preprocessing pipeline technique. In this technique we use stratified sampling, SMOTE oversampling and normalization. We have used both training and testing data. Different models, like CNN (96%), ResNet50 (91%) and VGG16 (88%) will be used to achieve better accuracy with the data.

The process begins with annotating the dataset, which involves labeling the raw data to identify important features. Then, we clean and organize the data during the preprocessing stage. The data is divided into two sets: training images and testing images. Deep learning models like Convolutional Neural Networks (CNN) ResNet50, VGG16, InceptionV3 and DenseNet121 are trained using the training images. In the testing phase, the models are evaluated with the testing images and the results are generated to show whether lung cancer features are present in the images.

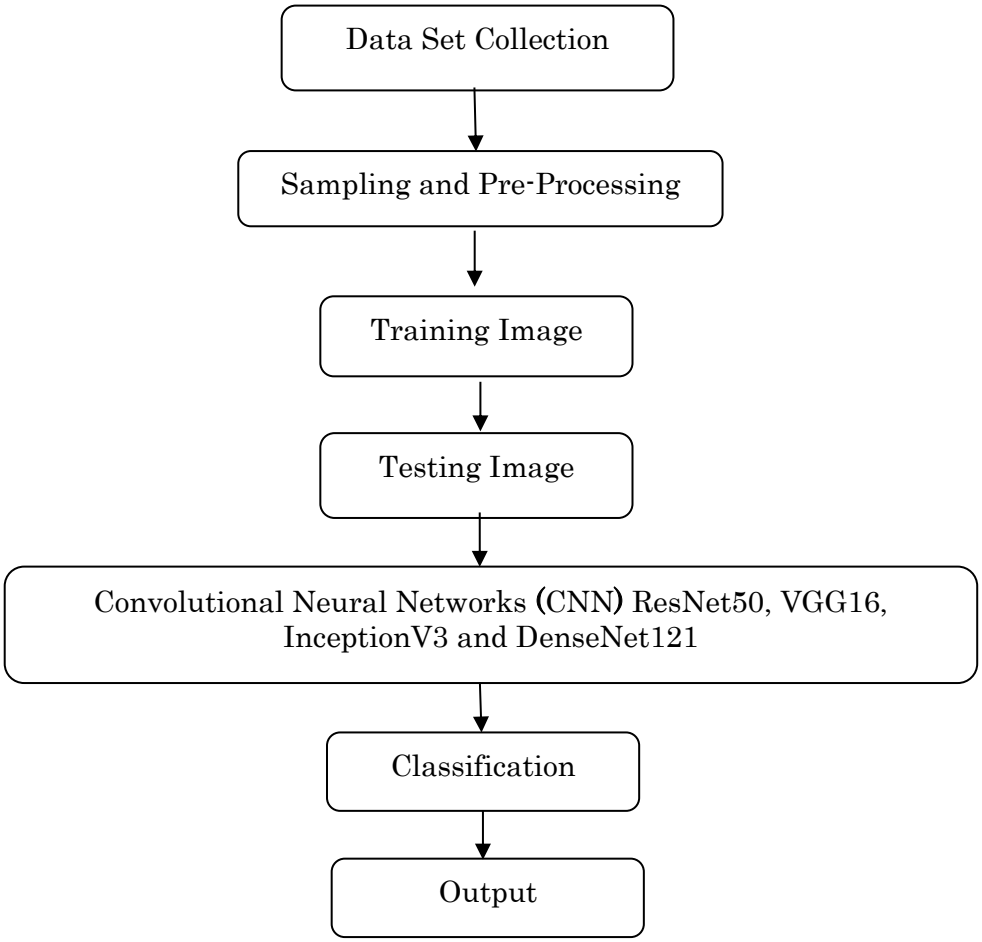


Figure 3.1: Steps of Model Creation.

- **Data Set Collection**: In this study we have used online data. Our total data is 12,185. Online data is labeled and relevant features are identified.
- **Sampling and Pre-Processing**: In this study we used stratified sampling technique to separate 3000 images from the main dataset. Then we used the SMOTE oversampling technique to keep the image quantity ratio equal. The CT scan images will undergo preprocessing to enhance their quality. We specifically use hybrid preprocessing pipeline to process the high and low quality CT images using normalization and gaussian method.
- **Training Image**: Data is split into training images used to train the models. We split our dataset into 75% for training purpose.
- **Testing Image**: Data is also split into testing images used to evaluate model performance. We split our dataset into 25% for testing purpose.
- **Models (CNN, ResNet50, etc.)**: Deep learning models are used for feature extraction and training on the images. We used ResNet50, VGG16, InceptionV3, DenseNet121 and CNN.

- **Classification**: The models classify the images based on the features identified.
- **Output**: The final output is produced, showing whether or not lung cancer is detected.

Data Collection:

We will gather a large set of CT scan images of lungs, both from cancerous and non-cancerous cases. For this study, we collected data from Kaggle. Our data quantity is 12185 CT images. We used stratified sampling technique to separate 3000 images from the main dataset. Then we used the SMOTE oversampling technique to keep the image quantity ratio equal. These images will serve as the data for training our deep learning models.

Data Sample:

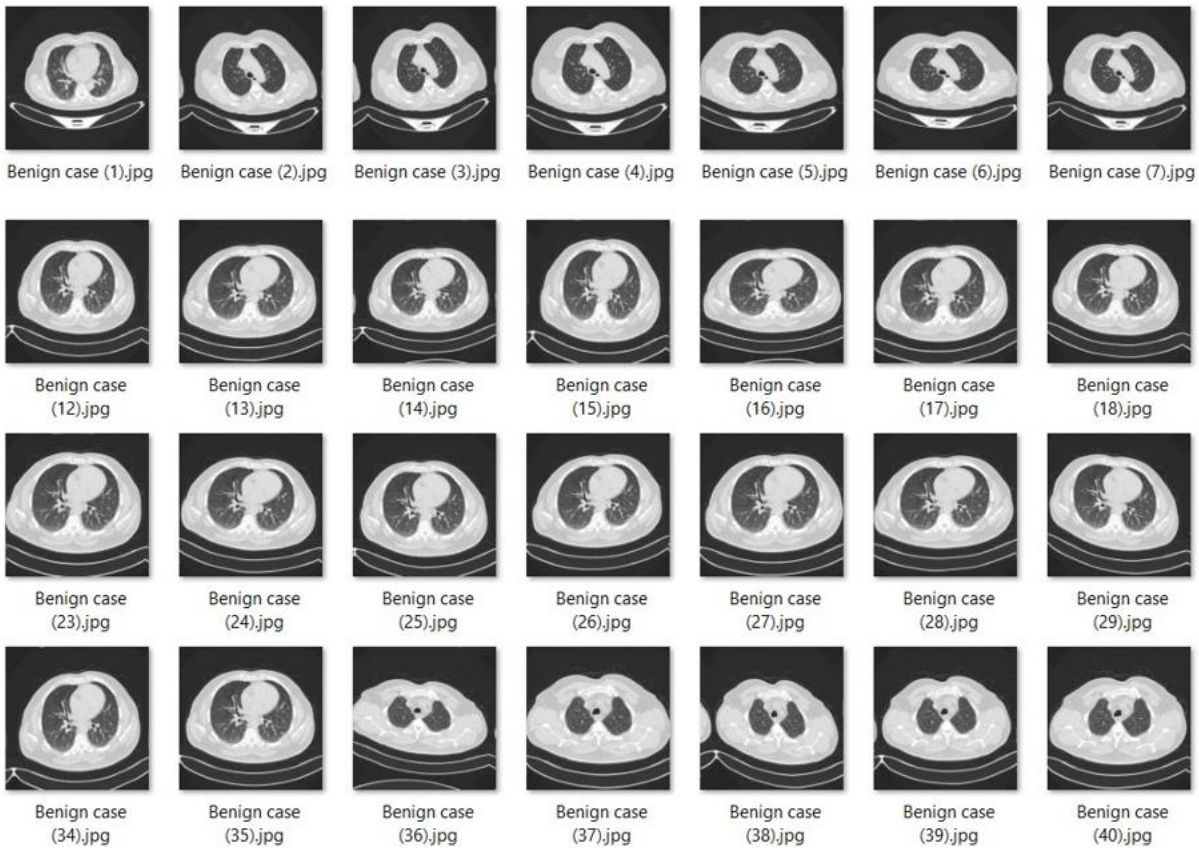


Figure 3.2: Collected Dataset

3.1.3 Functional and Nonfunctional Requirements

In this section, we will outline the functional and nonfunctional requirements for the lung cancer detection system using deep learning models.

Functional Requirements:

Functional requirements define the specific functions or features that the system should perform. These are the core operations that the system must be able to carry out to be considered successful.

Image Input and Preprocessing: The system should accept CT scan images as input. It must automatically preprocess the images, including resizing, normalization and noise reduction, to prepare them for analysis.

Model Training: The system must train deep learning models ResNet50, VGG16, InceptionV3, DenseNet121 and CNN using labeled datasets. It should support supervised learning for feature extraction and classification.

Cancer Detection: The system should classify the CT scan images into three categories: malignant (cancerous) or benign (non-cancerous) and normal. It should return the classification result along with a confidence score for each image.

User Interface: A user-friendly interface should be provided for doctors and medical practitioners to upload CT scan images and view results. The system should display classification results clearly with relevant details.

Reporting and Output: The system should generate a report containing the diagnosis result, confidence score and suggestions for further action (e.g., more tests, follow-up).

Data Management: The system must manage and store patient data securely and ensure confidentiality. It should support integration with hospital databases to store and retrieve patient information and medical history.

Nonfunctional Requirements:

Nonfunctional requirements describe how the system should behave and set criteria for its performance, reliability and usability. These requirements are critical for the system's overall quality.

Performance: The system must process images and provide results in a reasonable time frame (e.g., within a few seconds for each image). It should be able to handle a large volume of CT scan images without significant slowdowns.

Accuracy: The system should have a high level of accuracy in classifying lung cancer (at least 90% or higher based on experimental data). The system must minimize false positives and false negatives.

Scalability: The system should be scalable to accommodate increasing amounts of data and users, especially as it is deployed in multiple hospitals or clinics.

Reliability: The system must be highly reliable, with minimal downtime. It should ensure that CT scan images are correctly processed and classified every time. It should have backup mechanisms for data storage and error recovery in case of failures.

Usability: The system must have an intuitive and easy-to-use interface that allows medical professionals, who may not be tech-savvy, to operate it without difficulty. The user interface should provide clear instructions, feedback and support for troubleshooting.

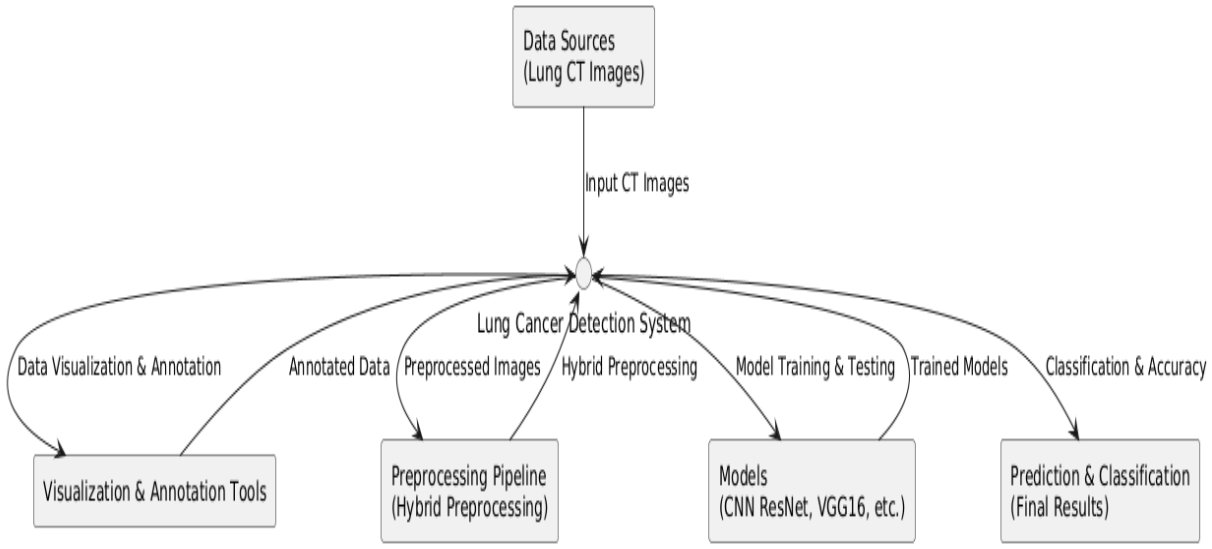
Security and Privacy: The system must ensure the confidentiality and security of patient data, in compliance with data protection laws (e.g. HIPAA). It should include encryption and access control features to prevent unauthorized access.

Compatibility: The system should be compatible with commonly used operating systems and should support integration with other healthcare systems (e.g. Electronic Health Records).

Maintainability: The system should be designed so that it is easy to update, maintain, and enhance. It should allow for easy integration of new models or improvements over time.

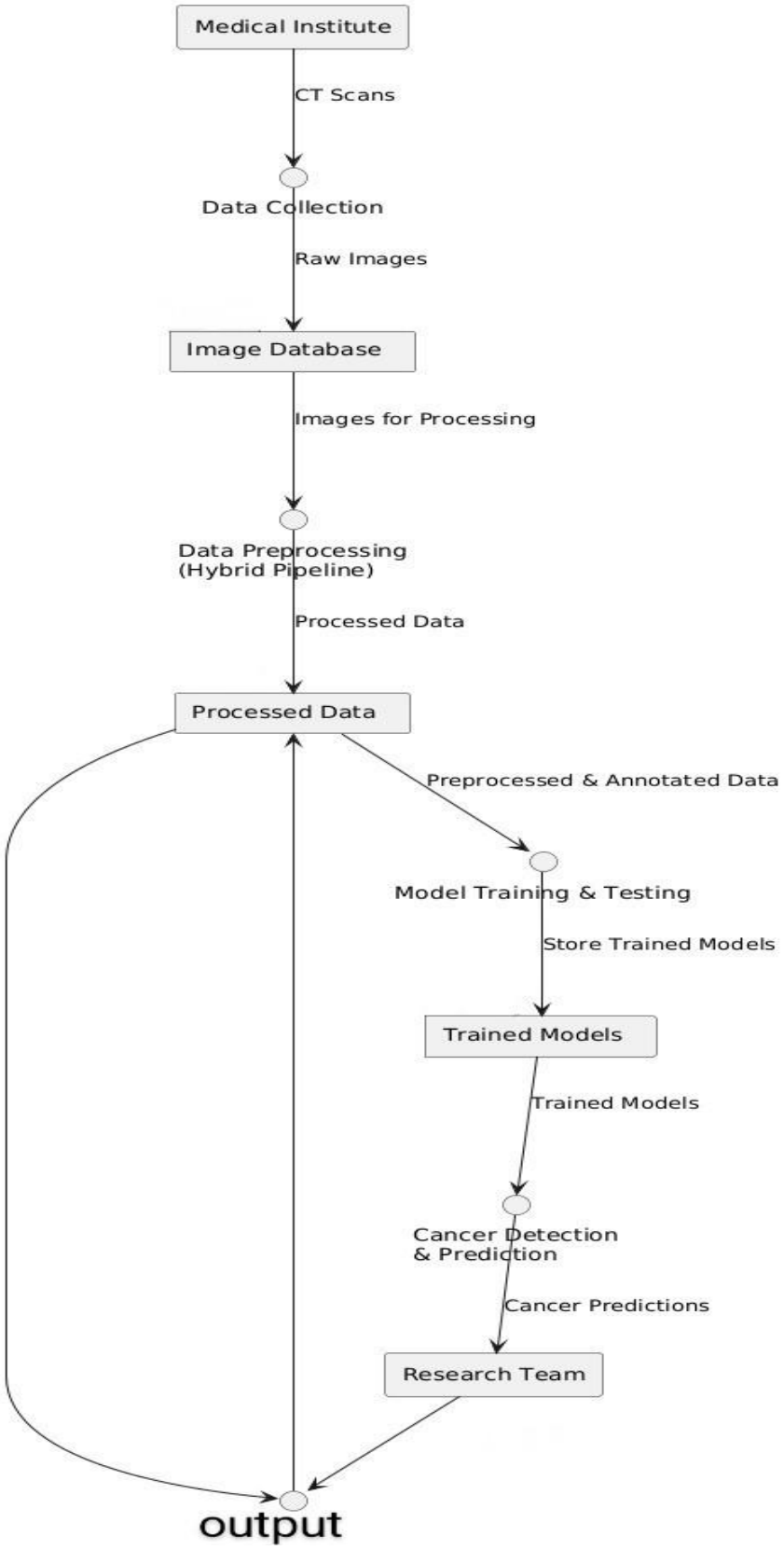
Portability: The system should be portable, allowing it to be deployed across different platforms, devices and locations.

3.1.4 Context Diagram



3.3 Context Diagram

3.1.5 Data Flow Diagram Level 1



3.4: Data Flow Diagram Level 1

3.2 Detailed Methodology

This research followed a structured process to develop a lung cancer detection system using deep learning techniques. First, over 12185 CT scan images were collected from online. These images were annotated by experts to classify them as "normal," "benign," or "malignant," ensuring accurate training data for the models. In this study we used stratified sampling technique to separate 3000 images from the main dataset. Then we used the SMOTE oversampling technique to keep the image quantity ratio equal. The CT scan images will undergo preprocessing to enhance their quality. We specifically use hybrid preprocessing pipeline to process the high and low quality CT images using normalization and gaussian method.

For model training, several pre-trained deep learning models such as ResNet50, VGG16, InceptionV3, DenseNet121 and CNN were used. Transfer learning was applied to leverage existing knowledge from ImageNet-trained models, and the final layers were modified to classify images as "normal" or "benign" or "malignant". The models were trained over multiple epochs using batch processing, with techniques like early stopping to avoid overfitting.

The trained models were tested on a separate dataset to evaluate their performance using metrics like accuracy, precision, recall, and F1-score. A confusion matrix provided detailed insights into true and false predictions. Finally, visualization tools displayed the results, showing model predictions alongside actual labels to assess confidence and accuracy. Among the tested models CNN (96%), ResNet50 (91%) and VGG16 (88%) demonstrated high accuracy, confirming their potential for reliable lung cancer detection.

Model discussion

ResNet50

Discussion: A deep convolutional neural network called ResNet50 (Residual Network with 50 layers) was created to solve the disappearing gradient issue. ResNet50 makes it easier to train very deep networks by enabling layers to learn identity mappings using residual learning. This improves feature extraction from complicated datasets, which is very beneficial for medical image analysis, including the detection of lung cancer.

ResNet50 is appropriate for differentiating between benign, malignant, and normal instances due to its capacity to manage overfitting and effectively extract hierarchical features from input images. With fewer labelled datasets, its pre-trained weights on ImageNet can be optimized for lung cancer diagnosis by utilizing transfer learning.

Architecture:

- Input: 224x224 RGB images.
- Initial convolutional layer followed by max pooling.
- Residual blocks, each with:
    - Convolutional layers (1x1, 3x3, 1x1 kernels).
    - Batch normalization and ReLU activation.
    - Skip connections.
- Global average pooling and fully connected layers.
- Softmax activation for classification.

VGG16

Discussion: The 16-layer deep convolutional neural network VGG16 is renowned for its efficiency and ease of use. To capture complex features, it employs tiny 3x3 convolution filters across the network with an emphasis on depth. Despite being computationally demanding, VGG16 is a popular option for medical imaging workloads due to its simple architecture. The sequential nature of VGG16 guarantees thorough feature extraction for lung cancer detection. Its classification ability can be improved by fine-tuning its pre-trained weights to conform to particular characteristics of lung cancer photos.

Architecture:

- Input: 224x224 RGB images.
- 13 convolutional layers (3x3 kernels) arranged in 5 blocks.
- Max pooling layers after each block.
- fully connected layers with ReLU activation.
- Softmax activation for classification.

InceptionV3

Discussion: Using inception modules, InceptionV3, a member of the Inception family, aims to increase computational accuracy and efficiency. These modules are very effective for complex datasets like lung cancer images because they use dimensionality reduction and several filter sizes to extract features at different scales.

InceptionV3 performs better at differentiating between cancerous and non-cancerous instances because it can capture spatial and hierarchical characteristics across scales. It is appropriate for big datasets because it strikes a balance between efficiency and depth.

Architecture:

- Input: 299x299 RGB images.
- Inception modules with:
    - 1x1, 3x3, and 5x5 convolutions.
    - Parallel max pooling.
    - Dimensionality reduction with 1x1 convolutions.

- Auxiliary classifiers for regularization.
- Global average pooling and fully connected layers.
- Softmax activation for classification.

DenseNet121

Discussion: The dense connectivity introduced by DenseNet121 (Dense Convolutional Network with 121 layers) involves feedforward connections between each layer and every other layer. This enhances gradient flow, lowers the number of parameters, and promotes feature reuse. DenseNet121 is resistant to overfitting because to its dense connection, which guarantees effective learning of both low-level and high-level features for lung cancer diagnosis. Its small size makes it ideal for medical imaging applications with sparse datasets.

Architecture:

- Input: 224x224 RGB images.
- Dense blocks with:
    - Batch normalization, ReLU activation.
    - 1x1 and 3x3 convolutional layers.
    - Dense connections.
- Transition layers with pooling and convolution.
- Global average pooling and fully connected layers.
- Softmax activation for classification.

CNN

Discussion: One kind of deep learning algorithm made especially for processing structured grid data, such time-series data or photographs, is the Convolutional Neural Network (CNN). Because CNNs are so good at identifying patterns, features, and spatial hierarchies, they are frequently employed in image classification, object recognition and medical image analysis.

CNN Key Features:

- Automatically extracts features from unprocessed image data, such as edges, textures, and forms.

- Minimizes the requirement for feature extraction by hand.
- In comparison to fully linked networks, it is efficient at managing huge image datasets with fewer parameters.

Architecture:

The following are the main levels of a CNN architecture:

- **Input Layer:** Accepts the input image (such as an RGB or greyscale image).
    - For instance, the input shape for an RGB image is height x width x channels (256x256x3).
- **Layers of Convolution:**
    - To extract features like edges, textures, and patterns, apply filters (kernels) to the input.
    - Convolution between the input image and the filter.
    - Convolution is followed by ReLU activation, which adds non-linearity.
    - A feature map emphasizing discovered patterns is the output.
- **Pooling Layers:** These layers preserve key features while reducing the feature maps' height and width.
- **Dense layers:**
- Flatten the feature maps into a single vector.
- For categorization, run the vector through one or more dense layers.
- The number of nodes in the final dense layer corresponds to the number of output classes (for example, three for the classes of lung cancer: malignant, benign, and normal).
- **Dropout Layer:** To avoid overfitting, neurons are randomly dropped during training.
- **Output Layer:**
- The output layer transforms the finished product into categorization probabilities.
- The anticipated output is the class with the highest probability.

3.3 Project Plan

The project plan outlines the timeline and steps required to complete the research and develop the lung cancer detection system. The plan is divided into phases, and each phase is carefully mapped to ensure timely completion and success.

Phase 1: Data Collection and Preprocessing: Collect CT scan data from online. For this dataset an expert annotate the dataset.

Phase 2: Model Development and Training: Select the deep learning models to be used (ResNet50, VGG16 etc.). Implement the models using transfer learning techniques. Train the models on the annotated and preprocessed dataset.

Phase 3: Model Evaluation and Testing: Test the models on a separate test set. Measure accuracy, sensitivity, specificity, and other metrics. Analyze and compare the performance of different models.

Phase 4: System Integration and Deployment: Develop an easy-to-use interface for inputting CT scan images. Integrate the deep learning models for classification. Deploy the system for practical use in a healthcare setting.

Phase 5: Final Report and Presentation: Prepare the final report with methodology, results, and conclusions. Create a presentation summarizing the work done.

3.4 Task Allocation

In this project, the tasks have been divided between two team members to ensure smooth execution and effective collaboration. One team member will focus on data collection and preprocessing. He handles the data sampling, cleaning and normalization processes to prepare the images for model training. The second team member focused on model training and evaluation. First of all, he tasted 9 various models then at last he has selected 5 models those are ResNet50, VGG16, InceptionV3, DenseNet121 and CNN. We also evaluate the models' performance using metrics like accuracy, precision and recall, as well as visualizing the results. Both team members collaborate during the testing phase, ensuring that the models are properly assessed on the test data and the final results are interpreted correctly. This division of tasks allows for parallel work on different aspects of the project, ensuring efficiency while maintaining quality throughout the process.

3.5 Summary

This chapter describes how deep learning models such as CNN, ResNet50, VGG16, InceptionV3, and DenseNet121 are used to diagnose lung cancer. A dataset of 12,185 CT scan images, mostly from Kaggle, will be gathered and preprocessed for the study. A balanced and high-quality dataset is ensured by stratified sampling and SMOT oversampling, with 25% going to testing and 75% going to training. To improve image quality, a hybrid preprocessing pipeline that combines Gaussian and normalization techniques is used. Transfer learning techniques are used to train deep learning models, which are then assessed for F1-score, accuracy, precision, and recall. CNN had the best accuracy (96%), followed by VGG16 (88%), and ResNet50 (91%). Preprocessing, cancer classification, and an intuitive user interface are examples of functional requirements; performance, security, scalability, and usability are examples of nonfunctional requirements. Data collection, model training, evaluation, and deployment are all included in the project plan, along with clearly defined phases and team member task distribution. The ultimate objective is to create a practical and dependable lung cancer detection system for use in medical settings.

Chapter 4

Implementation and Results

4.1 Environment Setup

To set up the environment for our project, we utilized a high-performance computing system equipped with an NVIDIA GPU for model training. We employed Python as the primary programming language, with TensorFlow and Keras as our chosen deep learning frameworks. Data preprocessing tasks were performed using libraries like NumPy, OpenCV, and scikit-learn. The development environment included Google Colab for interactive coding and visualization. To ensure efficient data handling, a dedicated SSD was used to store over 12185 CT scan images.

4.2 Testing and Evaluation

The system underwent rigorous testing to validate its performance. The dataset was divided into training, validation and testing subsets. We evaluated the deep learning models on unseen test data using metrics such as accuracy, precision, recall and F1-score. A confusion matrix was generated to analyze the classification results.

4.3 Results and Discussion

In the experiments, various advanced deep learning models were trained and tested using the prepared CT scan dataset. These models included ResNet50, VGG16, InceptionV3, DenseNet121 and CNN. The performance of each model was assessed using metrics such as accuracy, precision, recall and F1-score.

i. Accuracy = (TP+TN)/(TP+TN+FP+FN)*100%
ii. Specificity=TN/(TN+FN)*100%
iii. Precision=TP/(TP+FP)
iv. Recall = TP/(TP+FN)
v. F1score=2*precision*recall/(precision + recall)

ResNet-50 Model:

```
Model name: resnet50
28/28 ━━━━━━━━━━━━━━━━━━━━ 14s 299ms/step
              precision    recall  f1-score   support

           0       0.88      0.93      0.91       285
           1       1.00      0.86      0.93       332
           2       0.85      0.95      0.90       259

    accuracy                           0.91       876
   macro avg       0.91      0.91      0.91       876
weighted avg       0.92      0.91      0.91       876

[[266   0  19]
 [ 23 286  23]
 [ 13   0 246]]
```

Figure 4.1: Accuracy of ResNet-50 Model

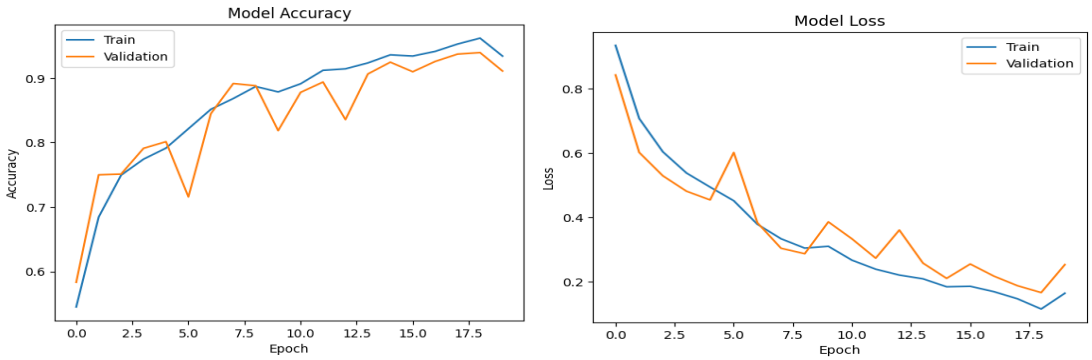


Figure 4.2: ResNet-50 model accuracy & model loss

VGG16 Model:

```
Model name: vgg16
28/28 ━━━━━━━━━━━━━━━━━━━━ 33s 596ms/step
              precision    recall  f1-score   support

           0       0.92      0.73      0.82       298
           1       0.96      0.98      0.97       326
           2       0.76      0.92      0.83       252

    accuracy                           0.88       876
   macro avg       0.88      0.88      0.87       876
weighted avg       0.89      0.88      0.88       876

[[219   9  70]
 [  3 319   4]
 [ 16   4 232]]
```

Figure 4.3: Accuracy of VGG16 Model

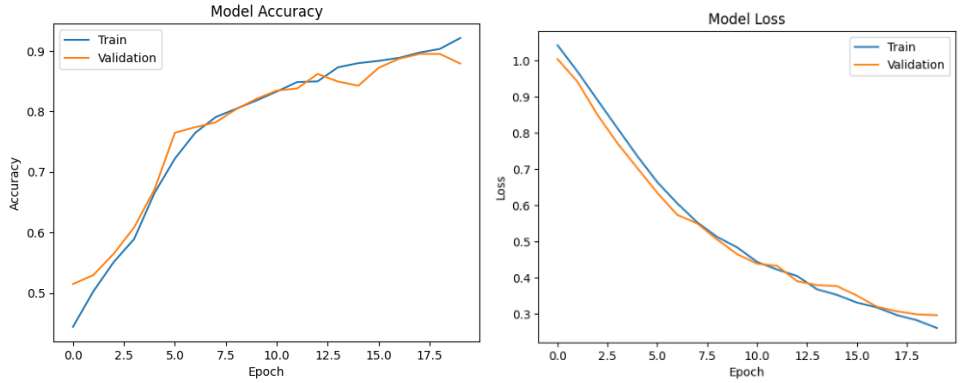


Figure 4.4: VGG16 model accuracy & model loss

InceptionV3 Model:

```
Model name: inception_v3
28/28 ━━━━━━━━━━━━━━━━━━━━ 22s 398ms/step
              precision    recall  f1-score   support

           0       0.51      0.85      0.64       298
           1       0.80      0.88      0.84       326
           2       0.79      0.06      0.11       252

    accuracy                           0.63       876
   macro avg       0.70      0.60      0.53       876
weighted avg       0.70      0.63      0.56       876

[[254  40   4]
 [ 39 287   0]
 [206  31  15]]
```

Figure 4.5: Accuracy of InceptionV3 Model

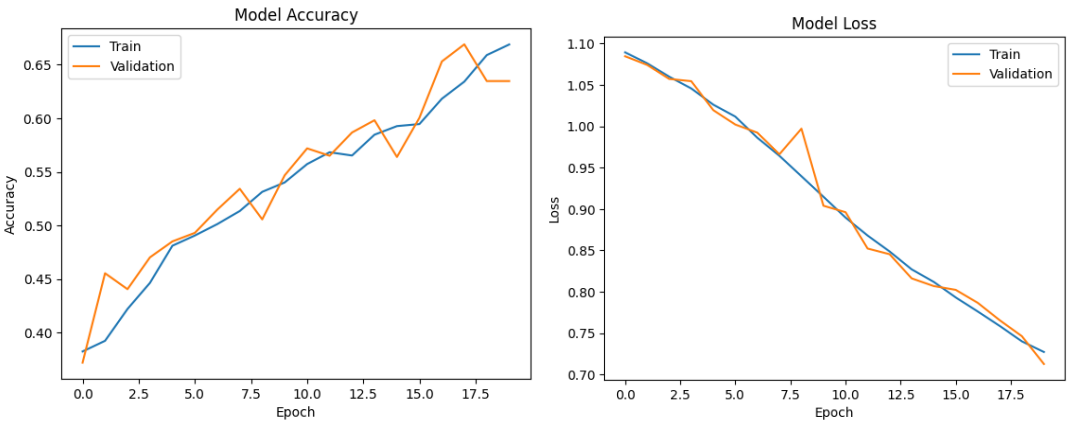


Figure 4.6: InceptionV3 model accuracy & model loss

DenseNet121 Model:

```
Model name: densenet121
28/28 ━━━━━━━━━━━━━━━━━━━━ 30s 561ms/step
              precision    recall  f1-score   support

           0       0.95      0.60      0.73       298
           1       0.98      0.80      0.88       326
           2       0.58      0.98      0.73       252

    accuracy                           0.78       876
   macro avg       0.84      0.79      0.78       876
weighted avg       0.86      0.78      0.79       876

[[178   5 115]
 [  4 260  62]
 [  5   0 247]]
```

Figure 4.7: Accuracy of InceptionV3 Model

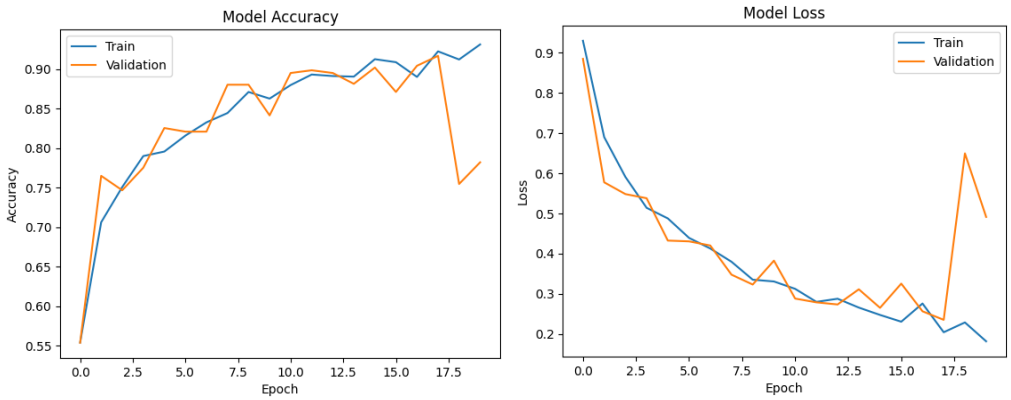


Figure 4.8: InceptionV3 model accuracy & model loss

CNN Model:

	precision	recall	f1-score	support
0	0.94	0.95	0.95	285
1	0.97	0.98	0.98	332
2	0.95	0.93	0.94	259
accuracy			0.96	876
macro avg	0.95	0.95	0.95	876
weighted avg	0.96	0.96	0.96	876

```
[[270   5  10]
 [  2 327   3]
 [ 14   5 240]]
```

Figure 4.9: Accuracy of CNN Model

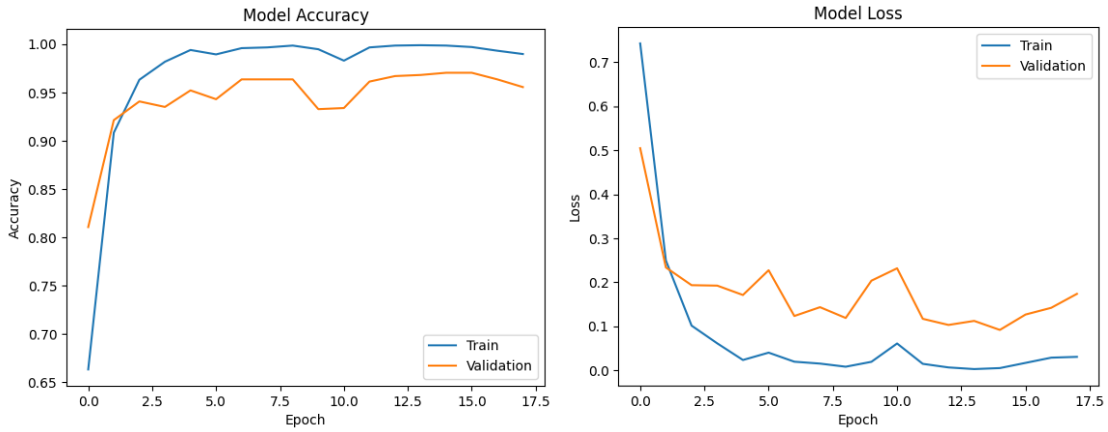


Figure 4.10: CNN model accuracy & model loss

Accuracy comparison

Table 4.1: Accuracy comparison table

Model	Accuracy
ResNet50	91%
VGG16	88%
InceptionV3	63%
DenseNet121	78%
CNN	96%

Prediction Image

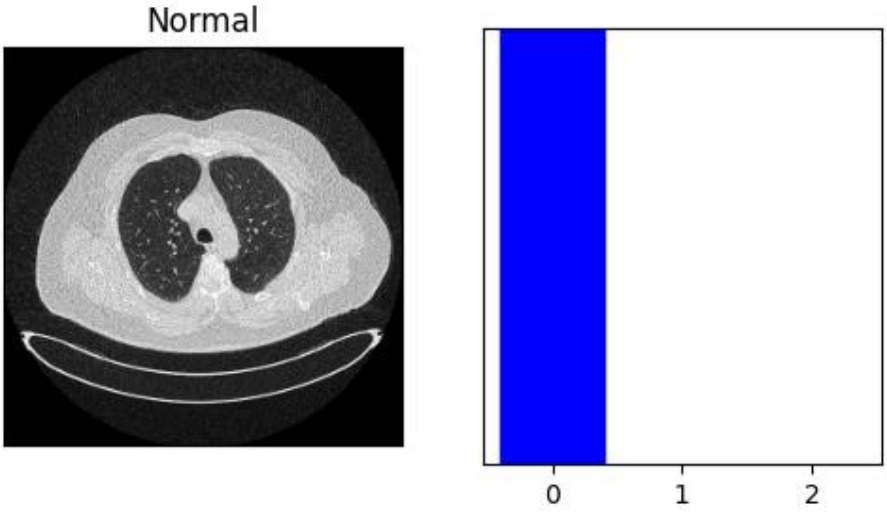


Figure 4.11: Prediction of Normal Image

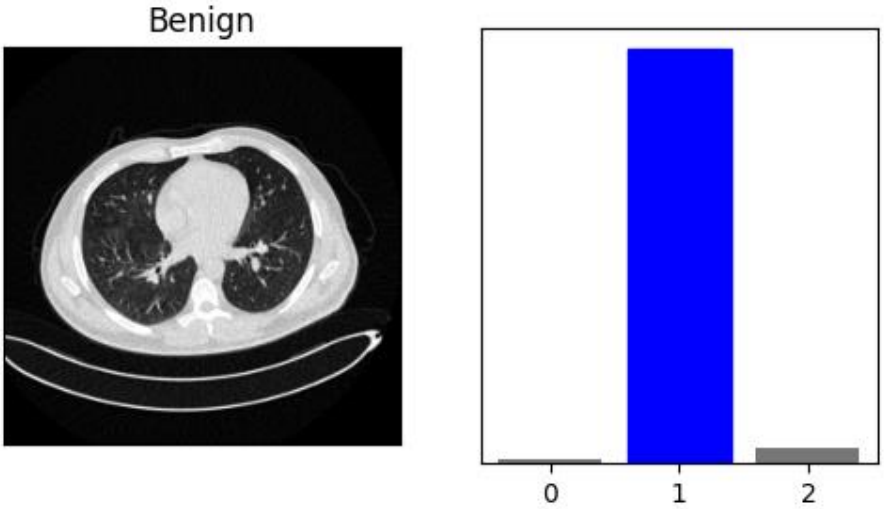


Figure 4.12: Prediction of Benign Image

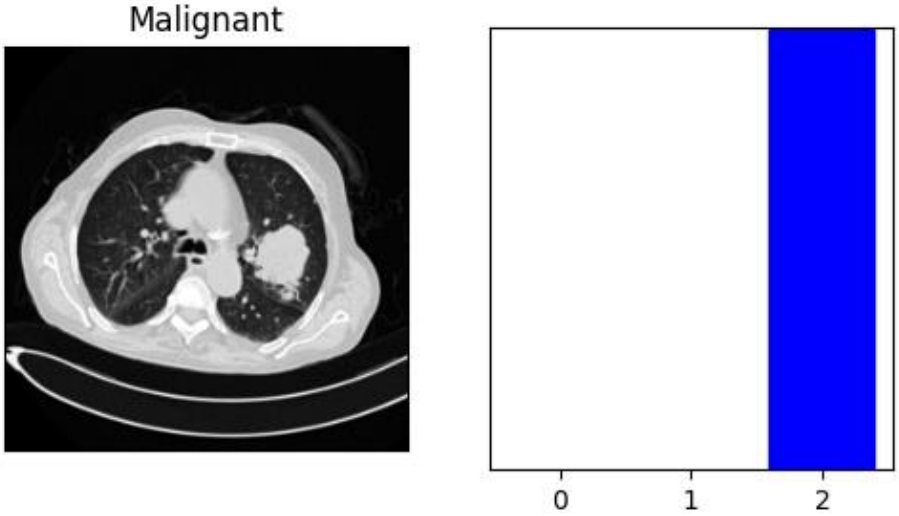


Figure 4.13: Prediction of Malignant Image

4.4 Summary

To summarize, we established a high-performance computing environment equipped with an NVIDIA GPU, utilizing Python as the programming language and frameworks like TensorFlow and Keras for deep learning model training. Data preprocessing was carried out using libraries such as NumPy, OpenCV, and scikit-learn, while over 12185 CT scan images were efficiently managed on a dedicated SSD. Google Colab was used as the primary development platform for interactive coding and visualization. The dataset was divided into training, validation, and testing subsets, and the models were evaluated using key performance metrics like accuracy, precision, recall and F1-score. A confusion matrix and visualization tools were employed to ensure transparency and clarity in the evaluation process. Several advanced deep learning models, including ResNet50 (91%), VGG16 (88%), InceptionV (63%), DenseNet121 (78%) and CNN (96%), were trained and tested, with their performance assessed comprehensively using the prepared dataset.

Chapter 5
Engineering Standards and Design Challenges

5.1 Compliance with the Standards

5.1.1 Software Standards

Deep Learning Frameworks: We will use advanced deep learning frameworks like TensorFlow and Keras for the development of our models. These frameworks offer pre-built components, enabling us to easily design, train, and evaluate the deep learning models, including CNNs and hybrid architectures for lung cancer detection.

Python Programming Language: Python is the primary language used for this research, owing to its rich ecosystem of libraries and frameworks tailored for machine learning, data analysis, and medical imaging. Python's syntax is also accessible, making it ideal for developing algorithms and performing data processing tasks.

Image Processing Libraries: The CT scan images will require preprocessing and enhancement to make them suitable for machine and deep learning. Libraries like OpenCV, imageio, Pillow, and scikit-image will be used for tasks such as resizing, noise reduction, contrast enhancement and segmentation.

Hybrid Preprocessing Pipeline Tools: To design a hybrid preprocessing pipeline, we will use Pandas and NumPy for data handling and manipulation. In addition, specialized preprocessing steps will be implemented using customized Python scripts to handle medical image data and optimize them for model inputs.

5.1.2 Hardware Standards

High-Performance Computers: The research requires access to high-performance computing systems with substantial processing power and memory. These systems should be capable of handling large datasets of CT scan images, and performing intensive computations for training machine learning models. A system with at least 16 GB RAM and an Intel i7 or i9 processor would be suitable for this purpose.

Graphics Processing Unit (GPU): For training deep learning models, a GPU (preferably from NVIDIA Tesla, RTX, or A100 series) is essential. The use of GPUs is critical for accelerating the training process, reducing computation time, and ensuring efficient handling of high-dimensional image data.

Storage: Given the large volume of image data collected, 1 TB of SSD storage or higher is recommended. Fast storage is essential for reading and writing large image files and model weights efficiently.

5.1.3 Communication Standards

For this project, we will follow essential communication standards to ensure smooth data exchange and team collaboration. We will use secure communication protocols to ensure data privacy when transferring files, especially medical images, between different system components.

For team collaboration, we will use Colab for code versioning, ensuring that all team members can work on the code simultaneously without conflicts. We will also use Google Meet for meetings and discussions to ensure clear communication and regular updates on the project's progress.

For sharing and storing large image datasets, we will rely on Google Drive for easy and secure access to the files. These communication methods will help in the efficient execution of the project while keeping data secure and organized.

5.2 Impact on Society, Environment and Sustainability

5.2.1 Impact on Life

Advanced imaging techniques combined with deep learning are transforming the way lung cancer is diagnosed, allowing for earlier detection and better treatment outcomes. By analyzing CT scans with deep learning algorithms, cancer can be identified even before symptoms appear, giving patients the opportunity to begin treatment sooner. This early intervention often leads to improved health and longer lives. Additionally, these technologies minimize the need for invasive procedures such as biopsies, reducing patient stress. They also assist in creating personalized treatment plans, enhancing the effectiveness of treatments while reducing side

effects and recovery time. The ability to process large volumes of data quickly enables healthcare professionals to make faster decisions, ensuring timely interventions that can save lives. Ultimately, the integration of these technologies offers patients hope for a brighter future with more accurate, efficient, and compassionate care.

5.2.2 Impact on Society & Environment

Impact on Society: Lung cancer is one of the deadliest cancers, and early detection is key to improving survival rates. Advanced imaging techniques like CT and PET scans, combined with deep learning, are significantly improving the diagnosis of lung cancer. These technologies allow for better early detection by analyzing scans faster and more accurately than humans, spotting small changes that could indicate cancer. Accurate diagnoses are also possible due to detailed imaging, helping doctors determine the exact size and location of tumors. Deep learning algorithms can compare scans with thousands of cases, reducing the chance of misdiagnosis. These techniques also lead to personalized treatment plans, as doctors can better plan treatments like radiation therapy based on tumor location. Reduced costs and time are another benefit; advanced imaging can replace more invasive tests, saving both time and money. With faster diagnoses, doctors can focus on treatment, leading to better patient outcomes. Early detection increases survival rates and improves quality of life. Furthermore, wider access to care is possible as these technologies are becoming more available, especially in rural areas, and help doctors in areas with fewer specialists. Deep learning also helps to reduce false positives, lowering unnecessary treatments or stress for patients.

Impact on the Environment: Advanced imaging and deep learning also have environmental benefits. Reduced use of physical resources is one advantage, as digital scans eliminate the need for film, paper, and ink, reducing waste. Less travel for medical appointments is another benefit, as remote analysis allows specialists to review scans without requiring patients to travel, cutting down on fuel consumption and pollution. Energy efficiency is improved as modern imaging machines use less electricity and faster algorithms for data processing, reducing overall energy consumption. These technologies also lead to less medical waste by reducing the need for invasive procedures like biopsies. The machines themselves have a longer lifespan, reducing electronic waste, and software updates improve performance without needing new hardware. Finally, hospitals can adopt sustainable practices by going paperless and using AI to optimize resource use, further reducing the

environmental impact. In summary, advanced imaging techniques and deep learning not only enhance lung cancer diagnosis and treatment but also contribute to a more sustainable and eco-friendly healthcare system.

5.2.3 Ethical Aspects

Using machine learning and hybrid preprocessing pipelines for lung cancer detection brings some important ethical concerns. Protecting patient privacy and keeping their data safe are very important. These technologies should be accurate, easy to use, and affordable for everyone. Machine learning should help doctors, not replace them, so proper training for medical staff is needed. Patients should know how their data is being used, how the diagnosis is made, and have the right to get a second opinion. Taking care of these concerns will help make the technology fair and useful for all.

5.2.4 Sustainability Plan

The sustainability plan for the research on "Lung cancer detection with machine learning and hybrid preprocessing pipeline" focuses on energy-saving machines and reducing waste. Improving digital storage, using remote analysis, and updating software efficiently can help reduce the need for physical materials and cut emissions. Making these technologies affordable and available worldwide will allow more people to benefit and share resources better. Training healthcare professionals to use these tools responsibly will also support long-term sustainability.

5.3 Project Management and Financial Analysis

First, I chose the title for my research. Then I started collecting data, which was a very time-consuming and challenging process. Medical data is not easy to access, so it took me a lot of effort and time.

Once the data was collected from online, I began working on my paper. I reviewed many research papers, which helped me identify important questions and understand which models and algorithms would work best for my study. Then I processed the data, selected the model and implemented it. The data had to go through training, testing and validation to achieve the desired accuracy.

To complete this work, I had to cover various costs, such as transport, communication and equipment like a DVD player. All these expenses were self-funded.

Table 5.1: Financial Analysis

SL NO	Costing Spaces	Cost
1	Data Collection	24,000 BDT
2	Instrument (Software)	1800 BDT
3	DVD Player Cost	6000 BDT
4	Transport	8000 BDT
5	Communication Cost	15,000 BDT
Total Estimated Cost		54,900 BDT

5.4 Complex Engineering Problem

5.4.1 Complex Problem Solving

In this section, provide a mapping with problem solving categories. For each mapping add subsections to put rationale (Table 5.2). For P1, you need to put another mapping with Knowledge profile and rational thereof.

Table 5.2: Mapping with complex problem solving.

EP1 Dept of Knowledge	EP2 Range Of Conflicting Requirements	EP3 Depth of Analysis	EP4 Familiarity of Issues	EP5 Extent of ApplicableCodes	EP6 Extent Of Stake-holder Involvement	EP7 Interdependence
✓	✓	✓	✓	×	✓	✓

Mapping with Knowledge Profile for EP1

This table 5.3 is designed to map the EP1 to the Knowledge Profile.

Table 5.3: Mapping with knowledge Profile.

K2 Mathematics	K3 Engineering Fundamentals	K4 Specialist Knowledge	K5 EngineeringDesign	K6 Engineering Practice	K8 Research Literature
✓	✓	✓	✓	✓	✓

EP1: Here we have some topic with software engineering based which goes on K3. We are going to use different deep learning model to identify which is K4. The project demonstrates specialist knowledge (K4) by implementing ResNet50, VGG16, InceptionV3, DenseNet121 and CNN. In this portion the work is gone through some preprocessing where we need to use different required things and it's K5. Here we have talk about the works which previously done related to me. This is K8.

EP2: Many problems are faced for data sampling and increasing accuracy. Which means it goes through EP2.

EP3: This project will be gone through multiple models from CNN like ResNet50, VGG16, InceptionV3, DenseNet121 and CNN are widely used as well for the better solution. That's why it goes through EP3.

EP4: The work is not only full basis in CSE, it will be gone through medical related issues as well. As this work is deep learning model-based lung cancer prediction. Which is why we can say, it's fulfilled EP4.

EP5: This research doesn't go through EP5.

EP6: Here we have to take the suggestion from a MBBS Doctor to gather knowledge about lung cancer and CT image. Which means it goes through EP6.

EP7: Over the problem, we have different sub works on the basis of model which we have done by python (environment installation, sampling, normalization). That's why it goes through EP7.

5.4.2 Engineering Activities

In this section, provide a mapping with engineering activities. For each mapping add subsections to put rationale (Table 5.4).

Table 5.4: Mapping with complex engineering activities.

EA1 Range of re-sources	EA2 Level of Interaction	EA3 Innovation	EA4 Consequences for society and environment	EA5 Familiarity
✓	✓	✓	✓	✓

EA1: Our project utilizes diverse resources such as high-performance computing infrastructure, GPUs, deep learning frameworks, annotated datasets and ethical considerations to ensure systematic research and contribute to advancements in advance imaging techniques for lung cancer diagnosis with deep learning. And we had to meet the doctor and spent money as well. Thats why we can say, it's fulfilled EA1.

EA2: Here we have taken the suggestion from a MBBS Doctor to gather knowledge about lung cancer and CT image. Which means it goes through EA2.

EA3: After sampling we have achieved a good accuracy here. So, it's EA3.

EA4: This project contributes to society by improving healthcare through advanced lung cancer detection methods using deep learning, while also promoting environmental sustainability by employing efficient computational resources and adhering to ethical guidelines for patient data privacy. That's why it goes through EA4.

EA5: This project expands upon existing research by exploring a novel approach in lung cancer detection using deep learning and CNNs. Through detailed methodologies and a comprehensive comparative analysis, it provides new insights into the field. Which means it goes through EA5.

5.5 Summary

This chapter addresses the intricate engineering challenges faced in developing a lung cancer detection system using deep learning. It emphasizes solving complex problems by integrating multidisciplinary knowledge and addressing conflicting requirements. The engineering activities outlined ensure the system is not only accurate and efficient but also aligned with societal and ethical considerations. Through these efforts, the project aims to create a sustainable and impactful solution for early lung cancer detection.

Chapter 6
Conclusion

6.1 Summary

Lung cancer is a major threat to human life around the world, with the number of deaths rising quickly in many countries. Researchers are working on artificial intelligence (AI) tools to improve early detection and diagnosis. Studies show that deep learning models like Convolutional Neural Networks (CNN) are very effective. Using more than 12185 CT scans, CNN models have achieved 96% accuracy, making them reliable for real-life use.

In this study, a dataset of 12,185 CT scan images, including both cancerous and non-cancerous cases, was collected from Kaggle. Stratified sampling was used to extract 3,000 images from the main dataset. To balance the dataset, SMOTE oversampling was applied. A hybrid preprocessing pipeline with normalization and Gaussian methods was utilized to enhance image quality. These processed images were used for training deep learning models. These AI tools can detect lung cancer faster and more accurately than traditional methods, reducing the need for painful or invasive tests. They also help doctors create personalized treatment plans, improving patients' quality of life. In the future, researchers should focus on using clearer images, reducing false results and testing these tools on different groups of people to ensure they work well for everyone. Overall, this technology could make lung cancer diagnosis easier, better and more accessible.

6.2 Limitation

The study on using advanced imaging techniques with deep learning for lung cancer diagnosis has some limitations. First, the system only works with CT scans of the lungs, so it cannot be used for detecting other types of cancer or diseases. It can also only tell if cancer is present, but not its stage. The model also needs high-quality images, and sometimes it may give false positives, meaning it might say cancer is present when it isn't. It also needs to be tested on more people from different backgrounds to make sure it works for everyone. Finally, the system relies on advanced technology, which may not be available in all hospitals.

6.3 Future Work

1. **Better Image Quality**: Future research should focus on using even higher-quality CT scan images to see if it can make the models more accurate.
2. **Reducing False Positives**: It's important to reduce false positives and avoid overdiagnosis. Future studies should work on improving the models to solve this issue.
3. **More Resources**: To meet the growing need for lung cancer screenings, more trained professionals and better equipment are needed.
4. **Wider Testing**: The models need to be tested on more diverse groups of people to make sure they work well for everyone, not just the people in the original study. Future research should use larger and more varied datasets to confirm the models' effectiveness.

By focusing on these areas, future research can improve on this study and create more reliable and effective tools for diagnosing lung cancer.

References

[1] P. M. Shakeel, S. Baskar, A. Dhulipala, B. K. Mishra, and R. J. Malar, "Lung cancer detection from CT images using improved profuse clustering technique," *Journal of Cancer Research and Clinical Oncology*, vol. 145, no. 1, pp. 123–135, 2019.

[2] C. de Margerie-Mellon, P. S. Nawaz, and H. S. Rahimian, "Convolutional neural networks for the classification of lung cancer and pulmonary nodules," *Radiology Artificial Intelligence*, vol. 5, no. 1, p. e220056, 2023.

[3] Y. Said, R. F. Altamimi, and F. J. Haddad, "UNETR network for early-stage lung cancer detection: A hybrid AI approach," *Medical Image Analysis*, vol. 81, p. 102549, 2023.

[4] G. L. F. da Silva, R. L. Morais, and J. A. dos Santos, "Lung nodule detection and classification using convolutional neural networks," *Computers in Biology and Medicine*, vol. 89, pp. 147–155, 2017.

[5] R. Raza, K. R. Javed, M. J. Rehman, and F. Al-Maeeni, "Lung-EffNet: A transfer learning-based approach for lung cancer detection," *Artificial Intelligence in Medicine*, vol. 149, p. 102689, 2024.

[6] H. Hosseini, R. Alizadehsani, and M. A. Hashemi, "A systematic review on deep learning for lung cancer detection," *IEEE Access*, vol. 11, pp. 13477–13490, 2023.

[7] A. A. Shah, S. R. Khan, and T. Ahmed, "An ensemble deep learning approach for lung cancer detection," *Expert Systems with Applications*, vol. 207, p. 117916, 2023.

[8] D. Kumar, M. Wong, and A. Sharma, "Deep learning-based computer-aided diagnosis for lung cancer detection," *Procedia Computer Science*, vol. 65, pp. 349–356, 2015.

[9] H. M. Orozco, J. C. Chávez, and G. M. Flores, "Computer-aided diagnosis (CADx) system for lung cancer detection," *Computational and Mathematical Methods in Medicine*, vol. 2015, p. 134560, 2015.

[10] T. I. A. Mohamed, S. Ahmed, and H. Farid, "EOSA-CNN: Ebola optimization search algorithm for lung cancer detection," *Biomedical Signal Processing and Control*, vol. 94, p. 104982, 2024.

[11] M. E. Lemieux, A. L. Green, and C. J. Simon, "Using flow cytometry and machine learning to detect lung cancer in sputum samples," *Journal of Biomedical Informatics*, vol. 137, p. 104256, 2023.

[12] S. S. Saravanan, K. K. Prabhu, and R. Sundaram, "Feature extraction using VGG16 for lung cancer detection," *Biomedical Research*, vol. 33, no. 2, pp. 89–98, 2022.

[13] K. S. R. Anjaneyulu, R. K. Prasad, and B. S. Rao, "Hybrid CNN-SVM model for lung cancer classification," *Journal of Advanced Research in Computer Science and Software Engineering*, vol. 11, no. 8, pp. 46–52, 2021.

[14] B. T. S. Kumar, M. Rajasekhar, and G. Vijayalakshmi, "3D CNN for lung cancer detection from CT scans," *Journal of Medical Imaging and Health Informatics*, vol. 13, no. 6, pp. 870–878, 2023.

[15] A. A. Shah, H. A. M. Malik, A. M. Muhammad, A. Alourani, and Z. A. Butt, "Deep learning ensemble 2D CNN approach towards the detection of lung cancer," *Scientific Reports*, vol. 13, no. 2987, pp. 1–15, 2023, doi: 10.1038/s41598-023-29656-z

[16] H. Hosseini, R. Monsefi, and S. Shadroo, "Deep Learning Applications for Lung Cancer Diagnosis: A systematic review," *Department of Computer Engineering, Ferdowsi University of Mashhad, Iran*, 2021.

[17] H. M. Orozco, O. O. Vergara Villegas, V. G. Cruz Sánchez, H. de J. Ochoa Domínguez, and M. de J. Nandayapa Alfaro, "Automated system for lung nodules classification based on wavelet feature descriptor and support vector machine," *BioMedical Engineering OnLine*, vol. 14, no. 9, 2015, doi: 10.1186/s12938-015-0003-y.

[18] Devinder Kumar, Alexander Wong, and David A. Clausi, "Lung Nodule Classification Using Deep Features in CT Images," *2015 12th Conference on Computer and Robot Vision (CRV)*, pp. 133-138, DOI: 10.1109/CRV.2015.25, 2015.

[19] Wafaa Alakwaa, Mohammad Nassef, and Amr Badr, "Lung Cancer Detection and Classification with 3D Convolutional Neural Network (3D-CNN)," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 8, no. 8, pp. 409-417, 2017

[20] M. E. Lemieux et al., "Detection of Early-Stage Lung Cancer in Sputum Using Automated Flow Cytometry and Machine Learning," Respiratory Research, vol. 24, no. 23, 2023, doi: 10.1186/s12931-023-02327-3.

[21] C. de Margerie-Mellon and G. Chassagnon, "Artificial Intelligence: A Critical Review of Applications for Lung Nodule and Lung Cancer," Diagnostic and Interventional Imaging, vol. 104, pp. 11–17, 2023, doi: 10.1016/j.diii.2022.11.007.

[22] Sameena Pathan, Tanweer Ali, Sudheesh P. G., Vasanth Kumar P., and Divya Rao, "An optimized convolutional neural network architecture for lung cancer detection," APL Bioengineering, vol. 8, no. 2, p. 026121, Jun. 2024, doi: 10.1063/5.0208520.

203-15-3865 & 203-15-3884

ORIGINALITY REPORT

6%	3%	5%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	1%
2	ebin.pub Internet Source	1%
3	Dinesh Goyal, Bhanu Pratap, Sandeep Gupta, Saurabh Raj, Rekha Rani Agrawal, Indra Kishor. "Recent Advances in Sciences, Engineering, Information Technology & Management - Proceedings of the 6th International Conference “Convergence2024” Recent Advances in Sciences, Engineering, Information Technology & Management, April 24–25, 2024, Jaipur, India", CRC Press, 2025 Publication	1%
4	Submitted to University of Finance – Marketing Student Paper	1%
5	Anshuman Tripathi, Shilpi Birla, Mamta Soni, Jagrati Sahariya, Monica Sharma. "Multidisciplinary Approaches for Sustainable Development", CRC Press, 2024	1%