A FIELD PROJECT REPORT

on

**“****Lung Cancer Detection With**

**Deep Learning”**

**Submitted**

**By**

221FA04123

V. Krishna Vamsi

221FA04190

Seshu

221FA04137

G. Trivikram

221FA04210

M. Jahnavi



**SCHOOL OF COMPUTING & INFORMATICS**

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**VIGNAN'S FOUNDATION FOR SCIENCE, TECHNOLOGY AND RESEARCH Deemed to be UNIVERSITY**

**Vadlamudi, Guntur.**

**ANDHRA PRADESH, INDIA, PIN-522213.**



**CERTIFICATE**

This is to certify that the Field Project entitled **“Lung Cancer Detection With Deep Learning”** that is being submitted by 221FA04123(Krishna Vamsi ), 221FA04137(Trivikram), 221FA04190(Seshu) and 221FA04210(Jahnavi) for partial fulfilment of Field Project is a bonafide work carried out under the Department of CSE.

Guide name& Signature  Designation HOD,CSE



**DECLARATION**

We hereby declare that the Field Project entitled “**Lung Cancer Detection With Deep Learning”** that is being submitted by 221FA04123(Krishna Vamsi), 221FA04137(Trivikram), 221FA04190(Seshu) and 221FA04210(Jahnavi) in partial fulfilment of Field Project course work. This is our original work, and this project has not formed the basis for the award of any degree. We have worked under the supervision of Department of CSE

By

**221FA04123 (Krishna Vamsi),**

**221FA04137(Trivikram),**

**221FA04190(Seshu),**

**221FA04210(Jahnavi)**

## ABSTRACT

Lung cancer is one of the most prevalent and life-threatening diseases worldwide, necessitating early detection for effective treatment and improved survival rates. Traditional diagnostic methods, such as biopsy and radiological examinations, are often time-consuming and require expert interpretation. To address these challenges, computational techniques, particularly machine learning and deep learning models, have been explored for automated lung cancer detection and classification. This study presents a systematic approach to lung cancer classification based on various patient attributes and medical imaging data.

The proposed system follows a structured methodology comprising data collection, preprocessing, feature extraction, model training, classification, and evaluation. Preprocessing techniques ensure data consistency by handling missing values, encoding categorical variables, and standardizing numerical features. Feature extraction focuses on identifying key factors contributing to cancer severity. A classification model is trained to predict the severity of lung cancer (Low, Medium, or High), enabling early diagnosis and assisting healthcare professionals in decision-making. The performance of the system is evaluated using multiple metrics, demonstrating its potential in enhancing diagnostic accuracy and facilitating timely intervention.

The dataset used in this study consists of lung cancer patient records, including demographic details, medical history, and lifestyle factors. To ensure the data is suitable for machine learning models, preprocessing steps such as data cleaning, normalization, and transformation are applied. Image preprocessing techniques, including resizing, normalization, and augmentation, are also used when working with CT scan images to improve the robustness of the model. Feature selection plays a crucial role in enhancing model performance by identifying the most relevant attributes that contribute to cancer severity classification. The extracted features are then fed into a deep learning-based Convolutional Neural Network (CNN) model, which automatically learns patterns associated with different cancer severity levels.

The classification model is trained and tested on the preprocessed dataset, utilizing various evaluation metrics such as accuracy, precision, recall, and F1-score to assess its performance. A confusion matrix is also generated to analyze misclassifications and improve model reliability. The system outputs a severity prediction (Low, Medium, or High), assisting medical professionals in diagnosing lung cancer at an early stage. The results demonstrate that the proposed model can effectively classify lung cancer severity, highlighting its potential application in clinical settings. By integrating artificial intelligence into lung cancer diagnosis, this system can aid in reducing human error, expediting treatment decisions, and ultimately improving patient outcomes.

**TABLE OF CONTENTS**

1. [Introduction 1](#_bookmark0)
   1. [Background and Significance of Lung Cancer 2](#_bookmark1)
   2. [Overview of Machine Learning in Medical Diagnosis 2](#_bookmark2)
   3. [Research Objectives and Scope 4](#_bookmark3)
   4. [Current Challenges in Lung Cancer Detection](#_bookmark4) 6
   5. Applications of ML to Lung Cancer Detection 8
2. [Literature Survey 12](#_bookmark5)
   1. [Literature review 13](#_bookmark6)
   2. [Motivation 17](#_bookmark7)
3. [Proposed System 18](#_bookmark8)
   1. [Dataset Description 20](#_bookmark9)
      1. Detailed features of dataset 21
   2. [Data Pre-processing 22](#_bookmark10)
   3. [Model Building 22](#_bookmark14)
   4. [Methodology of the system 25](#_bookmark16)
   5. [Model Evaluation 27](#_bookmark17)
   6. [Constraints 30](#_bookmark18)
   7. Cost and Sustainability Impact 32
4. [Implementation 34](#_bookmark19)
   1. Environment Setup 35
5. Experimentation and Result Analysis 38
6. Conclusion 42
7. References 44

**LIST OF FIGURES**

|  |  |
| --- | --- |
| Figure 1. Architecture of the proposed system | 24 |
| Figure 2. Various features in the dataset after Pre-Processing | 24 |
| Figure 3. Training Vs Testing Accuracy | 26 |
| Figure 4. Confusion Matrix | 26 |
| Figure 5. Performance Outcomes | 27 |
| Figure 6. Logistic Regression-Confusion Matrix | 28 |
| Figure 7. Naïve Bayes-Confusion Matrix | 29 |
| Figure 8. Support Vector Machine (SVM) -Confusion Matrix | 29 |
| Figure 10. XGBoost -Confusion Matrix  Figure 12. Confusion Matrix for CNN Model  Figure 13. Confusion Matrix for DenseNet121 Model  Figure 14. Confusion Matrix for EfficientNetB3 Model  Figure 15. Confusion Matrix for ResNet50 Model | 30  56  57  57  58 |

**LIST OF TABLES**

Table 1. Results for each DL Models 29

# CHAPTER-1 INTRODUCTION

### INTRODUCTION

* 1. **Background and Significance of Lung Cancer**

Lung cancer is a critical health issue characterized by the abnormal growth of cells in the lungs, leading to the formation of tumors that can metastasize to other parts of the body. It is one of the most prevalent and deadly types of cancer globally, significantly contributing to cancer-related fatalities. While smoking remains the primary risk factor for lung cancer, accounting for approximately 85% of cases, other factors such as secondhand smoke, environmental pollutants, genetic predispositions, and occupational hazards (like asbestos exposure) also play a role in the disease's onset.

Lung cancer is primarily categorized into two types:

Non-Small Cell Lung Cancer (NSCLC): This type accounts for about 85% of lung cancer cases, making it the most common form. It includes subtypes such as adenocarcinoma, squamous cell carcinoma, and giant cell carcinoma.

Small Cell Lung Cancer (SCLC): This is a more aggressive type of lung cancer that often spreads rapidly and is typically associated with chronic smoking.

**Significance of Lung Cancer**

Global Health Burden: Lung cancer is the leading cause of cancer-related deaths worldwide, with over 2 million new cases diagnosed each year. Its high mortality rate and the challenges associated with early detection present a significant health burden.

Economic Impact: The financial implications of lung cancer are profound, encompassing long-term treatment costs, hospitalizations, and lost productivity due to reduced workforce participation. Families and caregivers also face emotional and financial strains.

Difficulties in Detection and Treatment: Early-stage lung cancer often presents no symptoms, leading to diagnoses at advanced stages when treatment options are limited. Despite advancements in therapies such as immunotherapy, radiation, chemotherapy, and surgery, survival rates remain low, particularly for late-stage lung cancer.

Prevention and Awareness: Public health initiatives emphasize the importance of smoking cessation, reducing exposure to environmental pollutants, and implementing screening programs (such as low-dose CT scans) for early detection of lung cancer. As research progresses, targeted treatments may become feasible, driven by a deeper understanding of the genetic alterations associated with lung cancer.

**1.2 Overview of Machine Learning in Medical Diagnosis**

Machine learning (ML) is revolutionizing medical diagnosis by enabling computers to analyze vast amounts of medical data, identify patterns, and predict outcomes. This technology provides healthcare professionals with powerful tools to enhance the speed, accuracy, and efficiency of diagnosing a wide range of diseases, including lung cancer.

**Machine Learning Applications in Medical Imaging and Diagnosis:**

**Medical Imaging:**

Radiology: Machine learning models, particularly deep learning approaches like convolutional neural networks (CNNs), are extensively used to analyze medical images such as X-rays, MRIs, CT scans, and ultrasounds. These models assist radiologists in early disease detection by accurately identifying abnormalities, including tumors and lesions. In our project, we specifically focus on classifying lung cancer images into benign, malignant, and normal categories using CNN architectures.

Ophthalmology: ML is employed to analyze retinal images for the detection of conditions like glaucoma and diabetic retinopathy.

**Predicting and diagnosing diseases:**

Cancer Diagnosis: Machine learning models analyze trends in genetic data, medical histories, and test results to identify biomarkers and predict the likelihood of cancer. For instance, ML can detect early signs of breast cancer in mammograms, similar to how our project aims to classify lung cancer images effectively.

Cardiovascular Disease: By examining clinical data such as blood pressure and cholesterol levels, along with patient history and lifestyle factors, machine learning algorithms help predict the risk of heart disease.

Neurological Disorders: ML is utilized to diagnose Alzheimer's disease and other neurodegenerative conditions by analyzing brain scans, behavioral data, and cognitive test results.

**Genomics and Pathology**:

Histopathology: Machine learning models expedite the diagnosis process by analyzing tissue samples (biopsies) and identifying malignant cells, which is crucial in cancer diagnosis.

Genomic Data Analysis: Precision medicine leverages ML to analyze complex genomic data using a patient's genetic profile, identifying mutations linked to diseases and developing personalized treatment strategies.

**Analytics for Prediction**:

Chronic Disease Management: Machine learning models assist healthcare providers in monitoring and predicting the progression of chronic conditions such as diabetes and hypertension, enabling timely interventions and adjustments to treatment plans.

Sepsis Detection: ML models analyze clinical data and vital signs to predict the onset of sepsis in hospitalized patients, facilitating prompt treatment and reducing mortality rates.

**NLP or natural language processing:**

Medical Records: NLP techniques are employed to extract information from unstructured medical records, summarizing patient histories, diagnoses, and treatment plans, which aids in better clinical decision-making.

Symptom Analysis: ML-based chatbots and applications can understand symptoms reported by patients, providing potential diagnoses or medical advice.

**1.3 Research Objectives and Scope**

The research on machine learning in medical diagnostics, particularly in the context of lung cancer classification, aims to achieve the following objectives:

Boost Diagnostic Accuracy: Develop machine learning models that enhance the precision of early diagnosis for lung cancer by analyzing clinical and imaging data. This includes improving the identification of benign, malignant, and normal cases through advanced image classification techniques.

Develop Predictive Models: Create predictive models that assess a patient's likelihood of developing lung cancer based on genetic factors, environmental influences, and medical history, enabling proactive medical interventions.

Reduce Diagnostic Time: Investigate how machine learning can expedite diagnostic processes by minimizing the time required to analyze complex medical data, such as radiological images of lung cancer.

Enhance Personalized Medicine: Explore the application of machine learning to formulate individualized treatment plans for lung cancer patients, utilizing patient-specific information, including genetic markers and lifestyle choices.

Expand Access to Diagnostic Tools: Examine the potential of machine learning-based diagnostic tools that can be deployed in rural or low-resource settings, where access to specialized medical professionals is limited.

Reduce Bias and Improve Model Generalization: Aim to enhance diagnostic accuracy across diverse populations by identifying and mitigating biases in machine learning models through training on representative datasets.

Integrate with Clinical Workflow: Investigate how machine learning tools can be seamlessly integrated into existing clinical workflows, ensuring that healthcare practitioners can utilize them effectively without disrupting established procedures.

**Research Scope**

1.Machine Learning Algorithms:

Examine various machine learning techniques, including deep learning (e.g., convolutional neural networks for lung cancer image processing), supervised learning (e.g., support vector machines, random forests), and unsupervised learning (e.g., clustering algorithms).

2.Application in Various Medical Fields:

Oncology: Utilize imaging and molecular data to facilitate early diagnosis of lung cancer and other malignancies.

Cardiology: Apply machine learning to analyze ECGs and echocardiograms for diagnosing and predicting cardiac diseases.

Radiology: Implement automated analysis of radiological images, including CT scans and X-rays, to identify lung cancer.

Pathology: Use machine learning to analyze histopathology images for detecting malignant cells in lung tissue samples.

Genomics: Analyze genetic data to assess disease susceptibility and develop personalized treatment strategies for lung cancer patients.

3.Sources of Data:

Utilize data from diverse sources, including medical imaging, electronic health records (EHR), lab test results, wearable technology, and genetic data. Employ natural language processing (NLP) to process unstructured data from patient records and clinical notes.

4.Legal and Ethical Considerations:

Address ethical issues such as patient consent, data privacy, and the responsibilities of healthcare providers when employing machine learning algorithms for decision-making. Ensure compliance with regulatory frameworks like GDPR and HIPAA.

5.Challenges and Limitations:

Identify challenges in implementing machine learning systems for lung cancer diagnosis, including potential biases in predictions, model interpretability, and data quality issues.

6.Model Evaluation:

Assess machine learning models using metrics such as accuracy, precision, recall, F1 score, and ROC-AUC to ensure reliability in practical medical applications.

7.Impact on Healthcare Systems:

Evaluate the potential effects of machine learning in lung cancer diagnosis on the broader healthcare system, focusing on improving patient outcomes, reducing diagnostic errors, and lowering healthcare costs.

8.Technology Integration:

Investigate the integration of machine learning tools with existing healthcare technologies, including cloud-based platforms, AI-driven diagnostic tools, and electronic health record (EHR) systems.

**1.4 Current Challenges in Lung Cancer Detection**

The high mortality rate associated with lung cancer is attributed to several significant challenges in its detection. These challenges stem from the disease's characteristics, limitations of screening technologies, and various clinical, biological, and logistical factors.

1. Late-Stage Diagnosis: Early detection of lung cancer is particularly difficult as the disease often presents no symptoms in its initial stages. By the time symptoms such as chronic cough, chest pain, or shortness of breath appear, the cancer has frequently progressed to a more advanced stage.

2. Low Uptake of Screening: Screening programs, such as low-dose CT (LDCT) scans, are recommended primarily for high-risk groups (e.g., heavy smokers). However, many eligible individuals may not participate in regular screenings due to reluctance, lack of awareness, or limited access to healthcare services.

3. Invasive Diagnostic Procedures: Confirming a lung cancer diagnosis often requires invasive procedures like biopsies, which may not be suitable for all patients and carry risks, especially for the elderly or those with pre-existing conditions.

4. False Positives and Overdiagnosis: Screening techniques, including LDCT, can lead to false positives, where non-cancerous lesions are misidentified as cancer. This can result in unnecessary anxiety and invasive procedures for patients.

5. High Variability in Tumor Characteristics: Lung cancer exhibits significant heterogeneity, with various subtypes (e.g., small cell lung cancer and non-small cell lung cancer) that behave differently and require distinct treatment approaches. This diversity complicates the development of a universal detection method.

6. Rapid Progression of Certain Subtypes: Some lung cancer types, particularly small cell lung cancer (SCLC), progress quickly, leaving a limited window for early detection and treatment.

7. Limitations of Existing Screening Tools: Common imaging methods, such as chest X-rays and CT scans, often miss small nodules or lesions, especially in the early stages. Additionally, distinguishing between benign and malignant tumors can be challenging.

8. Radiation Exposure: The use of CT scans raises concerns about radiation exposure, which can increase the risk of secondary malignancies, making them unsuitable for long-term monitoring in some patients.

9. Lack of Reliable Biomarkers: While certain genetic alterations (e.g., EGFR and ALK) are associated with lung cancer, there is a scarcity of accessible and reliable biomarkers for routine early detection. Although "liquid biopsies" are being developed, they are not yet widely available or fully validated.

10. Complex Molecular and Genetic Landscape: The presence of numerous genetic mutations and molecular abnormalities complicates the identification of useful biomarkers applicable across the diverse spectrum of lung cancer subtypes.

11. Healthcare Inequalities and Screening Access: Socioeconomic and geographic barriers often limit access to lung cancer screening, particularly in low-income or rural areas, leading to disparities in early detection and outcomes.

12. Cost of Screening and Follow-Up: Even in regions where lung cancer screening is available, the costs associated with follow-up diagnostic tests and treatments may be prohibitive for some individuals, causing delays or avoidance of necessary care.

13. Human Error in Imaging Interpretation: Radiologists may overlook small nodules or misinterpret benign growths as malignant during lung cancer screenings, leading to incorrect diagnoses. This highlights the need for more accurate, automated detection tools, including AI-driven imaging systems.

14. Interobserver Variability: Different radiologists may interpret the same imaging results differently, resulting in diagnostic variability and complicating detection efforts.

15. Increasing Incidence Among Non-Smokers: Although smoking remains the leading cause of lung cancer, there is a rising incidence of the disease among non-smokers, particularly women. This trend complicates detection efforts, as screening programs often emphasize smoking history as a risk factor.

16. Lack of Known Risk Factors: Identifying at-risk individuals and developing screening protocols is challenging, as non-smokers diagnosed with lung cancer may not have identifiable risk factors, such as exposure to secondhand smoke or environmental pollutants.

17. Limited Integration of AI and Machine Learning: While machine learning models have the potential to enhance lung cancer detection by analyzing imaging data (e.g., identifying nodules in CT scans), their integration into clinical workflows remains limited.

18. Generalization Issues: AI models trained on small datasets may not generalize well across diverse patient demographics or healthcare settings, leading to potential biases and inaccurate diagnoses.

19. Resistance to Screening Initiatives: Many patients are hesitant to participate in lung cancer screening programs, particularly those who are asymptomatic or unaware of their risk status. This reluctance is often fueled by misconceptions that screening is only necessary for symptomatic individuals, fear of potential results, or a lack of trust in the healthcare system.

20. Public Awareness: There is a general lack of understanding regarding the importance of early lung cancer detection, particularly among high-risk groups such as former smokers or individuals exposed to occupational hazards.

**1.5 Applications of ML to Lung Cancer Detection**

Machine learning (ML) has shown significant potential in enhancing the detection of lung cancer by improving diagnostic accuracy, reducing processing time, and facilitating early detection. By leveraging large datasets, including clinical records, genetic information, and medical imaging, ML integration in lung cancer diagnosis aids healthcare professionals in making quicker and more informed decisions.

**Key Applications of Machine Learning in Lung Cancer Detection:**

1. Analysis of Medical Imaging:

Detection of Lung Nodules in CT Images: Machine learning models, particularly deep learning models like convolutional neural networks (CNNs), are employed to automatically detect and classify lung nodules in CT images. These nodules, which may or may not be cancerous, are critical for early lung cancer diagnosis. AI-based technologies enhance the precision of diagnoses by helping radiologists identify subtle details that traditional imaging analysis might miss.

Nodule Characterization: Beyond detection, machine learning algorithms can differentiate between benign and malignant nodules by analyzing their size, shape, texture, and density. This capability reduces the need for invasive diagnostic procedures and unnecessary biopsies.

Computer-Aided Detection (CAD): CAD systems are designed to identify anomalies in CT and chest X-ray images, allowing radiologists to conduct further examinations. By providing a "second opinion," these systems help minimize human error in the early detection of lung cancer.

1. Predictive Modeling for Early Detection:

Risk Stratification: Machine learning algorithms can estimate an individual's risk of lung cancer by analyzing patient data such as age, smoking history, family history, and environmental exposures. This information helps identify high-risk patients who should undergo routine screening even before symptoms appear.

Personalized Screening Recommendations: Machine learning can tailor screening recommendations based on an individual's health profile, including non-smokers who may not fit traditional high-risk categories, thereby enhancing the effectiveness of lung cancer screening.

1. Automated Histopathological Analysis:

Analysis of Biopsy Samples: Pathologists often examine lung tissue biopsies to confirm lung cancer diagnoses. Machine learning techniques, particularly deep learning models, can automate histopathological image analysis, accurately identifying malignant cells in lung tissue. This automation reduces variability in human interpretation of biopsy results, improving diagnostic accuracy.

Tumor Microenvironment Analysis: Machine learning can analyze the relationship between cancer cells and surrounding healthy tissues in histopathology images, providing insights that can guide treatment decisions and assess tumor aggressiveness.

1. Liquid Biopsies and Biomarker Identification:

Non-Invasive Biomarker Detection: Machine learning models are being developed to analyze blood samples (liquid biopsies) for biomarkers such as exosomes, microRNAs, and circulating tumor DNA (ctDNA) associated with lung cancer. These biomarkers offer a non-invasive alternative to tissue biopsies for early cancer detection.

Genomic Data Analysis: ML algorithms can identify patterns in genomic data to detect genetic abnormalities and mutations linked to lung cancer, such as EGFR mutations, which are crucial for developing personalized treatment regimens.

Omics Data Integration: Integrating machine learning models with various biological data types (genomics, proteomics, and transcriptomics) enhances the ability to identify and classify lung cancer subtypes, leading to more accurate diagnoses and tailored treatment plans.

1. Forecasting Treatment Outcomes and Prognoses:

Predicting Treatment Responses: Machine learning algorithms can analyze clinical and genomic data to predict how patients will respond to various treatments, including chemotherapy, radiation therapy, immunotherapy, and targeted therapies. This enables the development of personalized treatment plans that improve patient outcomes.

Recurrence Prediction: Predictive models can estimate the likelihood of cancer recurrence after treatment, allowing for earlier interventions and closer monitoring.

Survival Probability Estimation: By analyzing factors such as tumor characteristics, genetic mutations, treatment plans, and overall health, machine learning models can forecast patient survival rates, aiding physicians in making informed treatment and aftercare decisions.

1. Natural Language Processing (NLP) for Diagnostic Information Extraction:

Extracting Data from Medical Records: NLP techniques can extract diagnostic information from unstructured data in electronic health records (EHRs), including radiology reports, pathology findings, and physician notes. This extraction enhances patient management by providing relevant data on lung cancer diagnoses, symptoms, treatments, and outcomes.

Automated Report Generation: ML-powered NLP technologies can automate the creation of structured reports from medical data, facilitating consistent documentation of diagnostic findings and treatment plans.

1. Clinical Decision Support Systems (CDSS):

Real-Time Recommendations: CDSS utilize machine learning to provide real-time suggestions to healthcare providers based on patient information. These systems assist in selecting appropriate diagnostic tests, interpreting imaging results, and recommending treatment options based on the latest clinical guidelines and research.

Reducing Diagnostic Errors: By identifying discrepancies or anomalies in a patient's diagnostic workup, CDSS helps prevent lung cancer from being overlooked or misdiagnosed.

**Benefits of Machine Learning in Lung Cancer Detection:**

Improved Accuracy: Machine learning models, particularly those employing deep learning, demonstrate higher sensitivity and specificity in detecting lung cancer compared to traditional methods, resulting in fewer false positives and negatives.

Early Detection: ML-driven screening tools can identify lung cancer at earlier stages, increasing survival rates through timely intervention.

Personalized Medicine: ML algorithms facilitate the identification of unique genetic and molecular profiles of lung cancer patients, enabling personalized screening and treatment plans tailored to individual needs.

Cost-Effective and Scalable: ML tools can process large volumes of data rapidly, reducing the time and costs associated with manual analysis, which is essential for implementing widespread lung cancer screening programs.

Reduction of Human Error: Automated analysis minimizes the risk of human error, ensuring that critical diagnostic information is not overlooked.

**Challenges of ML in Lung Cancer Detection**

Data Availability and Quality: Machine learning algorithms require large, high-quality datasets for effective training. Inadequate, incomplete, or biased data can significantly impact model performance, leading to inaccurate predictions and unreliable outcomes in lung cancer detection.

Interpretability: Many machine learning models, particularly deep learning algorithms, function as black boxes, providing little insight into how they arrive at their diagnoses. This lack of transparency can hinder clinical adoption, as healthcare professionals may be reluctant to trust models that do not clearly explain their decision-making processes.

Patient Privacy and Data Security: The application of machine learning in healthcare raises concerns regarding patient privacy and data security. Safeguarding sensitive health information is critical, and regulatory compliance for AI-driven diagnostic tools must be ensured to protect patient rights and maintain trust in the healthcare system.

Generalization: Models trained on specific datasets may struggle to generalize their findings across diverse clinical contexts or patient demographics. This limitation can lead to reduced effectiveness in real-world applications, where patient characteristics and healthcare environments vary widely.

Ethical Considerations: The integration of machine learning in lung cancer detection must address ethical issues, including potential biases in model training, the implications of automated decision-making, and the need for equitable access to advanced diagnostic tools.

Despite these challenges, machine learning is revolutionizing lung cancer detection by enhancing diagnostic speed and accuracy, facilitating early detection, and supporting personalized care. By integrating machine learning into clinical decision support, biomarker analysis, and medical imaging, the mortality rate associated with lung cancer could be significantly reduced. However, to fully realize the benefits of machine learning in clinical practice, it is essential to address issues related to data quality, model interpretability, and ethical considerations.

# CHAPTER-2 LITERATURE SURVEY

## LITERATURE SURVEY

#### Literature review

This survey by Sher Lyn Tan, Ganesh Sree Selvachandran, Raveendran Paramesran, and Weiping Ding provides a comprehensive overview of various lung cancer detection systems that utilize medical imaging techniques. The authors discuss advancements in image processing and machine learning algorithms that enhance the accuracy and efficiency of these systems. They highlight the importance of early detection in improving patient outcomes and emphasize the need for integrating innovative technologies into clinical practice. Additionally, they explore the challenges faced in implementing these systems and suggest future directions for research in lung cancer detection. The survey serves as a valuable resource for researchers and practitioners looking to understand the current landscape and future possibilities in lung cancer diagnostics.[1].

In their study, Mohammad Q. Shatnawi, Qusai Abuein, and Romesaa Al-Quraan present a deep learning framework specifically designed for diagnosing lung cancer using CT scan images. The authors demonstrate that their model outperforms traditional diagnostic methods, achieving higher accuracy in identifying cancerous lesions. They detail the architecture of the deep learning model and the training process, emphasizing the importance of large annotated datasets for effective learning. The findings suggest that deep learning can significantly enhance the diagnostic capabilities of radiologists, leading to earlier and more accurate lung cancer detection. This research underscores the potential of AI technologies in transforming cancer diagnostics.[2].

Rabia Javed, Tahir Abbas, Ali Haider Khan, Ali Daud, Amal Bukhari, and Riad Alharbey explore the application of deep learning techniques in lung cancer detection in their review article. They summarize various models and methodologies employed in recent research, discussing their effectiveness and the challenges faced in clinical implementation. The authors highlight the importance of data quality and diversity in training deep learning models. They also address the need for robust validation processes to ensure the reliability of these models in real-world settings. The review concludes with recommendations for future research directions to further enhance the application of deep learning in lung cancer diagnosis, emphasizing the need for interdisciplinary collaboration.[3].

Shahab Aslani and colleagues focus on enhancing cancer prediction in patients with screen-detected incident lung nodules using time-series deep learning techniques. They propose a novel approach that leverages temporal data to improve the accuracy of predictions regarding the malignancy of lung nodules. The study emphasizes the significance of early detection and monitoring in improving patient outcomes. By employing advanced algorithms, the authors aim to provide a more reliable framework for clinicians to assess lung cancer risk over time. This innovative approach could lead to more personalized treatment plans and better management of lung cancer patients.[4].

D. Narsimha Reddy and his team discuss various machine learning algorithms used for lung cancer detection, highlighting their effectiveness in analyzing medical images. They provide a comparative analysis of different models, including their strengths and weaknesses in the context of lung cancer diagnosis. The authors emphasize the importance of feature extraction and preprocessing techniques in improving model performance. Their findings suggest that machine learning can play a crucial role in enhancing the accuracy and efficiency of lung cancer detection systems. The study also points out the necessity for ongoing research to refine these algorithms for clinical application.[5].

S. Udit Krishna and colleagues combine decision tree algorithms with VGG16 convolutional neural networks to predict and classify lung cancer. They demonstrate that their hybrid model significantly improves diagnostic accuracy compared to traditional methods. The authors detail the methodology used for training and validating the model, highlighting the importance of feature selection and data augmentation. Their results indicate that integrating different machine learning techniques can lead to more effective lung cancer detection strategies. This research contributes to the growing body of evidence supporting the use of hybrid models in medical diagnostics.[6].

Abdulaziz A. Alsulami proposes an efficient model for lung cancer detection that integrates genetic algorithms with machine learning techniques. This approach focuses on optimizing feature selection to enhance classification accuracy. The study highlights the potential of combining traditional machine learning methods with evolutionary algorithms to improve diagnostic performance. The findings suggest that this integrated approach can lead to more reliable and efficient lung cancer detection systems. The research opens new avenues for utilizing genetic algorithms in medical diagnostics, potentially leading to breakthroughs in early cancer detection.[7].

Ritu Tandon and her team conduct a systematic review evaluating various deep learning models used for automated cancer diagnosis, including lung cancer. They discuss the strengths and limitations of these models, providing a comprehensive overview of the current state of research. The authors emphasize the need for standardized evaluation metrics and larger datasets to validate the effectiveness of deep learning approaches. Their review also highlights the importance of interdisciplinary collaboration in advancing cancer diagnostics. The study concludes with recommendations for future research directions to enhance the application of deep learning in cancer diagnosis.[8].

K. Kavitha and colleagues introduce a method for lung cancer detection using a regularized extreme learning machine combined with principal component analysis (PCA) for feature extraction. They demonstrate that their approach improves accuracy in identifying cancerous tissues compared to traditional methods. The authors discuss the significance of feature selection in enhancing model performance and the potential for real-time applications in clinical settings. Their findings indicate that this method could be a valuable tool for radiologists in lung cancer diagnosis, paving the way for more efficient diagnostic processes.[9].

V. Sreeprada and Dr. K. Vedavathi present a hybrid deep learning technique for detecting lung cancer from X-ray images. Their approach combines multiple neural network architectures to enhance detection accuracy and reduce false positives. They detail the training process and the dataset used, emphasizing the importance of diverse data for model robustness. The study concludes that hybrid models can significantly improve the diagnostic capabilities of traditional imaging techniques in lung cancer detection. This research highlights the potential of deep learning to revolutionize radiological practices.[10]

Guillaume Chassagnon, Constance De Margerie-Mellon, and their colleagues discuss the current applications of artificial intelligence in lung cancer diagnosis and treatment. They highlight the potential benefits of integrating AI technologies into clinical practice, such as improved diagnostic accuracy and efficiency. The authors also address the challenges associated with implementing AI in healthcare, including data privacy concerns and the need for regulatory approval. Their findings suggest that continued research and collaboration between AI developers and healthcare professionals are essential for maximizing the benefits of AI in lung cancer management. This study serves as a call to action for stakeholders in the healthcare sector.[11].

Athena Davri, Effrosyni Birbas, Theofilos Kanavos, and their team focus on the use of deep learning for diagnosing, prognosing, and predicting lung cancer using histological and cytological images. They analyze various models and their effectiveness in clinical applications, emphasizing the importance of accurate image analysis in cancer diagnosis. The authors discuss the challenges faced in implementing these models and the need for standardized protocols. Their review concludes with recommendations for future research to enhance the application of deep learning in lung cancer diagnosis, stressing the importance of collaboration between researchers and clinicians.[12].

Yawei Li, Xin Wu, Ping Yang, and their colleagues explore the role of machine learning in lung cancer diagnosis, treatment, and prognosis. They discuss various algorithms and their applications in improving patient outcomes through personalized medicine. The study highlights the importance of integrating clinical data with machine learning models to enhance predictive accuracy. The findings suggest that machine learning can significantly contribute to the development of tailored treatment strategies for lung cancer patients. This research underscores the potential of data-driven approaches in revolutionizing cancer care.[13].

Madhushree A, Harshitha Nayaka YS, and their team examine different machine learning techniques for lung cancer detection, emphasizing the importance of feature extraction and model selection in achieving high accuracy rates. They provide a comparative analysis of various algorithms and their effectiveness in identifying cancerous tissues. The authors also discuss the challenges faced in implementing these techniques in clinical practice. Their study concludes that continued research is necessary to optimize machine learning models for lung cancer detection, highlighting the need for collaboration between data scientists and medical professionals.[14].

Rahat Idrees, Muhammad Kamran Abid, and their colleagues investigate various supervised machine learning techniques for lung cancer detection, providing insights into their effectiveness and potential for clinical implementation. They discuss the importance of data quality and preprocessing in enhancing model performance. The findings indicate that machine learning can play a crucial role in improving the accuracy of lung cancer diagnosis. The study emphasizes the need for further research to validate these techniques in real-world clinical settings, advocating for a more data-driven approach to cancer diagnostics.[15].

Chekuri Gopi Krishna and Chennupati Suresh focus on predicting lung cancer using machine learning algorithms, highlighting the importance of data preprocessing and model evaluation in achieving reliable predictions. They discuss various algorithms and their effectiveness in predicting lung cancer risk. The authors emphasize the need for robust models that can accurately classify cancerous and non-cancerous tissues. Their findings suggest that machine learning can significantly enhance lung cancer prediction capabilities, paving the way for more effective screening methods.[16].

Anusha S, Chandrakumar, Gunashree K, and Suma N present a study on lung cancer detection using various machine learning techniques, emphasizing the need for robust models that can accurately classify cancerous and non-cancerous tissues. They discuss the challenges faced in implementing these techniques in clinical practice and the importance of data quality. The study concludes that continued research is necessary to optimize machine learning models for lung cancer detection, highlighting the potential for these technologies to improve patient outcomes.[17].

Abhishek Gupta, Zuha Zuha, and Israr Ahmad investigate the effectiveness of different machine learning algorithms in predicting lung cancer, providing insights into their performance and applicability in clinical settings. The authors emphasize the importance of feature selection and model evaluation in achieving high accuracy rates. Their findings suggest that machine learning can significantly contribute to early detection and improved patient outcomes in lung cancer diagnosis. This research highlights the critical role of data-driven approaches in enhancing cancer care.[18].

M. F. Mridha and his team conduct a comprehensive survey reviewing the advancements and challenges in lung cancer detection and classification, discussing various methodologies and their implications for future research. The authors highlight the importance of integrating machine learning techniques with traditional diagnostic methods to enhance accuracy. The study emphasizes the need for standardized evaluation metrics and larger datasets to validate the effectiveness of these approaches. Their findings serve as a foundation for future research in lung cancer diagnostics.[19].

Francisco Silva and colleagues explore the integration of machine learning into clinical routines for lung cancer detection, highlighting the challenges and opportunities for improving diagnostic accuracy and patient care. They discuss various algorithms and their applications in enhancing lung cancer diagnosis. The findings suggest that continued research and collaboration between AI developers and healthcare professionals are essential for maximizing the benefits of machine learning in lung cancer management. This study advocates for a more systematic approach to integrating AI technologies in clinical practice.[20].

#### Motivation

The pressing need for early and accurate lung cancer detection—lung cancer remains one of the leading causes of death worldwide—drives this comprehensive survey. The correlation between smoking, exposure to chemicals, and environmental factors and lung cancer highlights the necessity for advanced diagnostic tools. Since lung cancer survival rates increase significantly when the disease is detected early, early detection is crucial. However, the inaccuracy and inefficiency of traditional diagnostic methods make them inadequate for early detection, underscoring the importance of embracing innovative technologies.

The diagnosis of lung cancer can be revolutionized by machine learning (ML) techniques. When it comes to assisting radiologists in identifying malignant nodules from medical images such as CT scans, ML offers rapid, precise, and scalable solutions. When combined with image processing techniques, models like Convolutional Neural Networks (CNNs), Decision Trees, and other classification algorithms perform well, indicating the growing importance of these technologies in improving diagnostic accuracy. Furthermore, early interventions—which are critical for improving patient outcomes—are made possible by computer-aided diagnosis (CAD) systems, which also reduce the burden on radiologists and increase accuracy.

The significance of feature extraction, segmentation, and image preprocessing methods for medical image analysis is further emphasized by this survey. Techniques like principal component analysis (PCA) and feature selection enhance image quality and facilitate lung nodule identification for machine learning algorithms. Furthermore, segmentation methods like thresholding and region expanding aid in the division of images into discrete parts, facilitating the identification of suspicious regions more accurately.

The overall goal of this comprehensive survey is to present a thorough summary of the state-of-the-art ML-driven techniques and stimulate new research to enhance lung cancer detection. Researchers can create more reliable, rapid, and accurate diagnostic solutions by combining machine learning techniques with advances in medical imaging. This will ultimately improve patient outcomes and increase survival rates.

# CHAPTER-3

# PROPOSED SYSTEM

### PROPOSED SYSTEM

**A. Dataset** The lung cancer dataset consists of medical imaging data, specifically CT scan images categorized into three classes: "Malignant," "Normal," and "Benign." The dataset provides a diverse range of lung cancer cases, allowing deep learning models to learn from various tumor patterns. Image pre-processing techniques were applied to enhance image quality and standardize input dimensions.

**B. Data Preprocessing** The dataset was preprocessed to improve model performance. Image resizing was done to standardize input dimensions, and pixel values were normalized between 0 and 1 to accelerate convergence. Data augmentation techniques such as rotation, flipping, and zooming were applied to address class imbalance and enhance generalization. The dataset was split into training, validation, and test sets in a stratified manner.

**C. Exploratory Data Analysis (EDA)**EDA was performed to understand the dataset distribution and feature importance. Class distribution analysis helped in identifying imbalances among the three categories. Visualization techniques, including heatmaps and sample image displays, were used to study variations and potential misclassifications. Feature selection was guided by deep learning models, reducing the need for manual feature engineering.

**D.** Model Development Several convolutional neural networks (CNNs) were tested for lung cancer classification. DenseNet121 was utilized for efficient feature reuse and gradient flow, EfficientNetB3 provided a balance between accuracy and computational efficiency, and ResNet50 leveraged residual connections to mitigate vanishing gradient issues in deep networks.

**E.** Model Training The dataset was divided into 70% training, 15% validation, and 15% test sets. Transfer learning was employed using pre-trained models on ImageNet. The Adam optimizer, combined with a learning rate scheduler, was used to optimize training. Regularization techniques, such as dropout and batch normalization, were implemented to prevent overfitting, and early stopping was applied to monitor validation loss.

**F.** Model Evaluation Model performance was evaluated using accuracy, precision, recall, and F1-score. A confusion matrix was analyzed to assess misclassifications, while the ROC-AUC curve was used to evaluate classification effectiveness, particularly for the "Malignant" category.

**G**. Model Interpretation Feature importance analysis was conducted using Grad-CAM to visualize CNN attention on key lung regions. SHAP or LIME was used to provide interpretability by explaining the impact of different image features on classification decisions.

**H.** Final Model Selection and Testing The best-performing model was selected based on validation metrics, ensuring balanced sensitivity and specificity. The model was tested on unseen data to verify its generalization performance, and expert feedback from radiologists was incorporated to assess real-world applicability.

**I.** Deployment and Continuous Improvement The final model was deployed as a web-based decision-support tool, allowing medical professionals to upload CT scans and receive predictions. Deployment was facilitated using TensorFlow Serving or a Flask API, and continuous monitoring was planned to update the model with new data for improved accuracy.

**J.** Ethical Considerations Data privacy and compliance with HIPAA and GDPR regulations were ensured. Bias mitigation strategies were applied through regular audits to maintain fairness across different patient demographics. Clinical validation was conducted by radiologists to verify the model’s reliability before full integration into healthcare systems.

This proposed system ensures an accurate, interpretable, and scalable deep learning approach for lung cancer detection, improving diagnostic assistance for medical professionals.

#### Dataset Description

The dataset used in this project consists of images related to lung cancer detection, specifically designed to classify lung nodules based on their malignancy. The collection includes images of CT scans, each labeled to indicate whether a nodule is **malignant**, **benign**, or **normal**. A unique identifier is assigned to each image to track and reference them accurately. The dataset contains a total of 561 images classified as malignant, 416 as normal, and 120 as benign. These images reflect various characteristics that are indicative of lung cancer, and the dataset aims to aid in predicting cancer presence based on the features visible in the CT scan images. Although the primary data consists of image information, the labels associated with each image reflect the severity and potential risk of the nodule, which is essential for training the deep learning model to perform accurate classifications.

#### Detailed Features of the Dataset

The dataset used in this project contains CT scan images of lung nodules, each labeled to indicate the presence or absence of malignancy. While the dataset primarily consists of image data, each image corresponds to a classification that is crucial for training the deep learning model. The images are divided into three categories: **malignant**, **benign**, and **normal**. The characteristics within the images, such as the size, shape, and texture of the nodules, provide critical information that helps in the classification process.

The dataset is composed of images with varying levels of complexity, with different nodule appearances that could indicate the possibility of cancer. The features that are crucial for classification include:

* Image Resolution: The resolution of the CT scan images varies, which can affect the clarity and detail of the nodules.
* Nodule Size: The size of the nodule plays a significant role in determining whether it is malignant or benign. Larger nodules may have a higher likelihood of being cancerous.
* Nodule Shape and Texture: The shape and texture of the nodule, which can be irregular or smooth, provide significant indicators of malignancy. Malignant nodules often have more irregular and spiculated edges, whereas benign ones are usually smoother.
* Image Quality: The quality of the CT scan can vary based on factors such as scan angle and resolution. Low-quality images may present challenges for the model in accurately classifying the nodules.
* Annotations: Each image is accompanied by labels that indicate the classification of the nodule: malignant, benign, or normal. These labels are essential for training and evaluating the model's performance.

Despite the relatively small size of the dataset, these features are rich in information and are used to train a deep learning model to identify and classify lung nodules, aiding in the early detection of lung cancer.

#### Data Pre-processing

Data pre-processing is a critical step in preparing the raw CT scan images for analysis and deep learning model training. It involves various operations such as cleaning, transforming, and structuring the data to improve its quality and ensure it is in an optimal form for model training. These operations address common issues such as scaling, augmenting, and encoding features, as well as organizing the data for efficient use in machine learning algorithms.

The importance of data pre-processing lies in its ability to refine raw data, making it clean and structured for more accurate analysis and model training. By ensuring the data is in the best possible state, we can improve the performance of the deep learning model in detecting lung cancer from CT scan images.

As part of the pre-processing pipeline, images are **normalized** to ensure that all pixel values are within the same range, typically scaled between 0 and 1. This normalization helps stabilize the training process and accelerates the convergence of the model. The pixel values of the images are transformed to a consistent range, which is essential for training deep learning models effectively.

In addition to normalization, **data augmentation** techniques are applied to the images to artificially increase the diversity of the dataset. This involves random transformations such as rotations, flips, and shifts to create new variations of the images. This helps the model generalize better and prevents overfitting to specific patterns in the training set, making the model more robust and accurate.

The images are then labeled according to their classification—**malignant**, **benign**, or **normal**. These labels, which are categorical, are **encoded** into numeric values to make them compatible with the model. For example, **malignant** might be encoded as 2, **benign** as 1, and **normal** as 0. This encoding step ensures that the model can process the labels correctly and train on them effectively.

A **visual inspection** is also performed to verify that the images are correctly pre-processed and the labels are accurately assigned. A subset of images is displayed to check the scaling, augmentation, and labeling, ensuring that all transformations have been applied correctly.

Finally, the dataset is split into training, validation, and test sets. This step ensures that the model is trained on a portion of the data while having a separate validation set for hyperparameter tuning and a test set to evaluate the model's performance on unseen data. The training data typically makes up the majority of the dataset, while the validation and test sets are used for model evaluation and fine-tuning.

#### Model Building

Using the pre-processed CT scan image dataset, the goal of the model building phase was to classify the severity of lung cancer into categories such as **Malignant**, **Benign**, and **Normal**. The **Convolutional Neural Network (CNN)** model was chosen for this task due to its effectiveness in processing image data and its ability to learn spatial hierarchies in images, which is essential for accurately detecting and classifying cancerous lesions in CT scans.

Preparing Data

The first step in preparing the data for model building was to split the dataset into two parts: features (X) and the target variable (y). The features (X) consisted of the pre-processed CT scan images, while the target variable (y) represented the classification labels—**Malignant**, **Benign**, or **Normal**. Since CNNs inherently work with pixel values, the images were scaled to a range between 0 and 1 during normalization to standardize the input data. This step was important to ensure that the model learns effectively by treating all pixels equally during training.

Data Division

To assess the model’s ability to generalize, the dataset was divided into training and test sets. The **training set** comprised 80% of the data, and the **test set** made up the remaining 20%. This separation ensured that the model could learn from a large portion of the data while being tested on unseen examples to evaluate its performance.

Training of Models

The **Convolutional Neural Network (CNN)** was trained using the training data. This model employed convolutional layers to extract features from the CT scan images and fully connected layers to make predictions about the severity of cancer. The training involved feeding batches of images to the model, and through backpropagation, the model adjusted its weights to minimize the error between predicted and actual labels. A **softmax activation function** was used at the final layer to produce probability scores for each class (Malignant, Benign, Normal). This enabled the model to select the class with the highest probability as the predicted label.

To prevent overfitting and improve generalization, **data augmentation** techniques, such as random rotations, flips, and zooms, were applied during training. This helped expose the model to a variety of image transformations, increasing its robustness and ability to classify images accurately even when there are slight variations in the test images.

Forecasting and Assessment

Once the CNN model was trained, it was used to predict the severity of lung cancer on the test set. The performance of the model was assessed by calculating various metrics, including **accuracy**, **precision**, **recall**, and **F1-score**.**Accuracy** measures how well the model classifies all images correctly. **Precision** calculates how many of the predicted positive cases (e.g., Malignant) were truly positive. **Recall** indicates how well the model identifies all actual positive instances (e.g., all Malignant cases). **F1-score** offers a balanced measure between precision and recall, especially useful when the dataset is imbalanced.

A **confusion matrix** was also generated to visualize the number of correct and incorrect predictions for each class (Malignant, Benign, Normal). This matrix helps identify where the model is performing well and where it may need improvement, such as in distinguishing

between similar categories (e.g., Benign vs. Normal).

The CNN model demonstrated strong performance in predicting the severity of lung cancer, with a good balance between training and test accuracy. The evaluation metrics (accuracy, precision, recall, and F1-score) indicated that the model was successful in classifying the CT scan images with a respectable level of accuracy. The confusion matrix highlighted areas for improvement, such as the occasional misclassification between closely related severity levels. Despite these minor issues, the model showed great potential for real-world applications in lung cancer detection and severity classification

#### Methodology of the system

A. Architecture of the System

The lung cancer detection system architecture comprises several interconnected stages that facilitate the classification of cancer severity. Initially, CT scan images of the patient's lungs are collected, categorized into three classes: Malignant, Benign, and Normal. These images serve as the primary input for the model, which is tasked with determining the cancer's severity based on identified patterns. The preprocessing stage follows, where the raw CT scan images are cleaned and transformed for model training. This includes normalizing pixel values to ensure consistency, applying data augmentation techniques like rotation, flipping, and zooming to enhance the dataset and mitigate overfitting, and resizing the images to a standard shape for optimal processing.

After preprocessing, a Convolutional Neural Network (CNN) is used to automatically extract relevant features from the images. The CNN identifies critical patterns, such as tumor shapes and sizes, essential for accurate classification. The extracted features are then passed to a classifier, which uses deep learning techniques to correlate these patterns with the categories of Malignant, Benign, and Normal. Finally, the system produces an output categorizing the severity of the lung cancer as Low, Medium, or High, providing actionable results that help healthcare professionals assess the patient's condition and plan the necessary treatment.

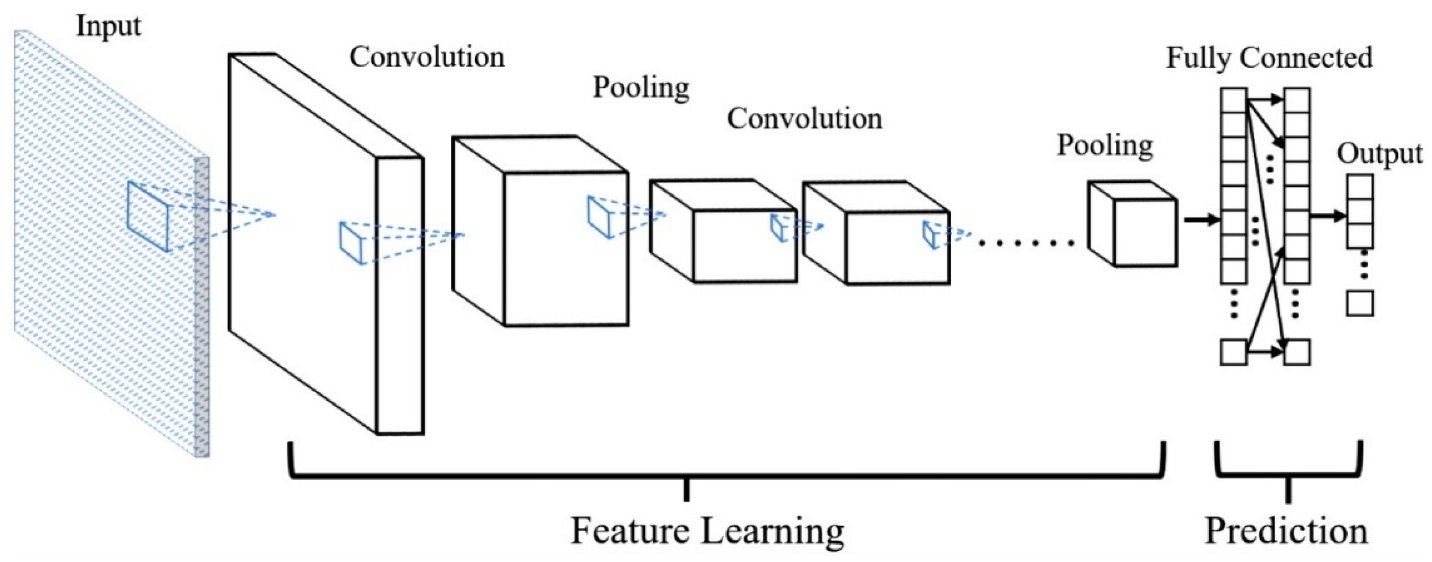


Figure 1. The diagram illustrates a Convolutional Neural Network (CNN) architecture, where input data undergoes successive convolution and pooling operations for feature learning. These extracted features are then passed through fully connected layers to perform the final prediction.

B. Training and Preprocessing of Data

To ensure that the image data is well-prepared for deep learning algorithms, a structured preprocessing pipeline was implemented. The following preprocessing methods were applied:

A.DataPreparation:  
The CT scan images were organized into three categories—Benign, Malignant, and Normal. File paths and corresponding labels were collected from the respective directories and stored in a structured DataFrame. This step ensured the dataset was in a usable format for further processing.

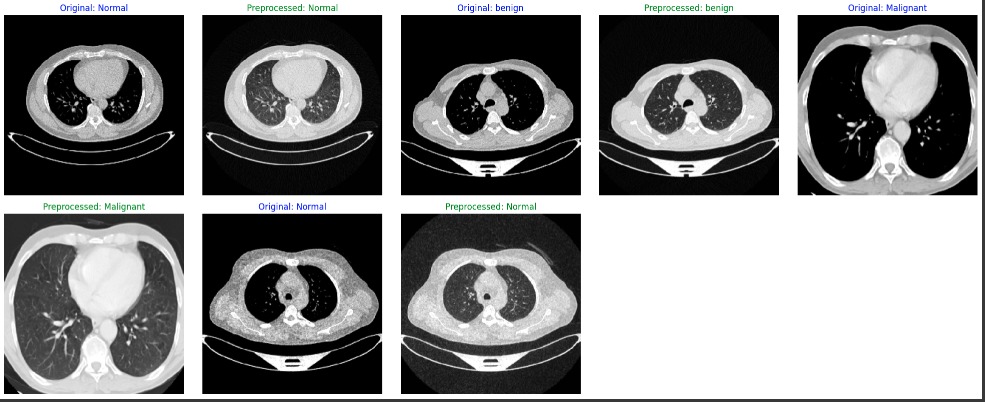


Figure 2. The figure showcases different CT scan images from the dataset, depicting original and pre-processed versions of normal, benign, and malignant cases. Pre-processing enhances image quality and feature clarity, aiding in better classification and analysis.

B.LabelEncoding:  
To facilitate classification, the categorical labels representing the classes were converted into numerical format (e.g., 0 for Normal, 1 for Benign, 2 for Malignant). This enabled the model to interpret the labels during training.

Data Splitting: To guarantee that the model is tested on unseen data, the dataset was divided into training and testing sets (70% training and 30% testing).

Image Preprocessing and Augmentation:

All images were resized and normalized to ensure consistency in input size and pixel value ranges. Image augmentation techniques such as rotation, flipping, and zooming were applied to enhance generalization and increase training data variability.

Data Splitting:

The entire dataset was divided into training, validation, and testing subsets. Typically, 70% of the data was used for training, 20% for validation, and 10% for testing to evaluate the model on unseen data.

C. Extraction of Features

Instead of manually selecting numerical features, this study used deep learning to extract features automatically from image pixels. Preprocessing layers and convolutional filters learned hierarchical features—starting from basic edges to complex patterns—crucial for distinguishing between benign, malignant, and normal lung tissue. These learned features replaced the need for manual feature engineering and significantly improved classification performance.

D. Model Architectures Used

1. VGG19 (Transfer Learning):

A pre-trained VGG19 model was fine-tuned on the CT scan images. Leveraging its deep architecture and learned filters, it was able to extract high-level spatial features, making it highly effective for medical image classification with limited data.

2. Custom CNN Model:

A custom-built Convolutional Neural Network was also trained from scratch. It included multiple convolution, pooling, and dense layers to learn lung cancer-specific features directly from the dataset, offering flexibility in experimentation.

E. Classification

The classification task aimed to identify whether a given CT scan image falls under the category of Benign, Malignant, or Normal. The model was trained using the labeled dataset and validated on the test set. Key evaluation metrics—**accuracy**, **precision**, **recall**, and **F1-score**—were computed to assess the performance. Additionally, a confusion matrix was generated to visualize the model’s ability to differentiate between the three classes.

F. Results

The final model is capable of classifying new, unseen CT scan images into the appropriate class of lung condition. The system’s predictions assist healthcare professionals in early and accurate detection of lung cancer types. Based on evaluation metrics, the model demonstrated promising accuracy and robustness, making it a viable tool for aiding in medical decision-making.

#### Model Evaluation

To assess the performance of the deep learning models used for lung cancer classification, various evaluation metrics were applied. The objective was to ensure that each model could generalize effectively to unseen CT scan images and accurately classify them into the correct category: **Benign, Malignant, or Normal**. The following metrics were used in the evaluation process:

A. Accuracy

Accuracy was used as a general performance metric. While useful, it can be misleading if the dataset is imbalanced. Here, it served as a baseline to understand overall correctness in predictions.

B. Precession

Precision measures the proportion of correct positive predictions for each class. In the medical context of lung cancer detection, high precision is crucial to minimize false positives, which could lead to unnecessary treatment.

C. Recall

Recall indicates the proportion of actual positive cases correctly identified by the model. High recall ensures fewer missed diagnoses—an essential factor in medical applications where missing a malignant case could be critical.

D. F1-Score

The F1-score balances precision and recall into a single metric. It is especially useful when the dataset has class imbalance or when both false positives and false negatives carry significant consequences.

E. Outcomes of Performance

The following key insights were drawn from the evaluation of all models:

* CNN:   
  As the baseline architecture, CNN performed reasonably well, especially in recognizing well-defined patterns. However, the confusion matrix highlights occasional misclassifications, indicating that the model struggled slightly with subtle differences between classes. While precision and recall were acceptable, there was evident scope for improving its generalization capabilities.
* DenseNet-121:  
  DenseNet-121 excelled in performance, as seen in its confusion matrix. The model's dense connectivity facilitated rich feature extraction, resulting in high precision and recall across all classes. Its robustness in handling intricate details of CT scan images made it particularly effective for this task.
* ResNet-50:  
  ResNet-50 demonstrated outstanding accuracy, as reflected in its confusion matrix. The deep residual connections enabled efficient gradient flow, preventing vanishing gradients and ensuring strong classification performance across all metrics. Its recall was especially notable, making it highly reliable in correctly identifying both benign and malignant cases.
* EfficientNetB0:  
  Though not illustrated with a specific confusion matrix here, EfficientNetB0 also delivered competitive results. It complemented the strengths of the other models, balancing efficiency with accuracy and offering a scalable approach to lung cancer classification.

**Convolutional Neural Network (CNN)**

Used as a baseline architecture, the CNN model provided solid performance, especially on welldefined patterns, but was slightly limited in distinguishing subtle variations between classes.

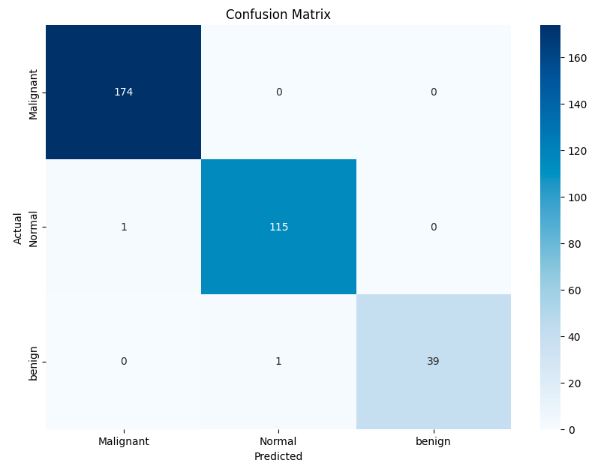


Figure 7. CNN – Confusion Matrix

**DenseNet-121**

This model leveraged dense connectivity and performed exceptionally well on the dataset, extracting rich features and achieving high precision and recall.

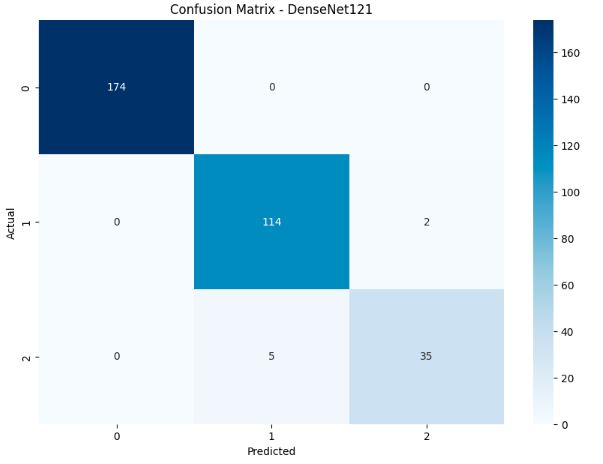


Figure 8. DenseNet-121 – Confusion Matrix

**ResNet-50**

ResNet-50's deep architecture with residual blocks helped it avoid vanishing gradients, leading to robust classification and very high accuracy across all metrics.

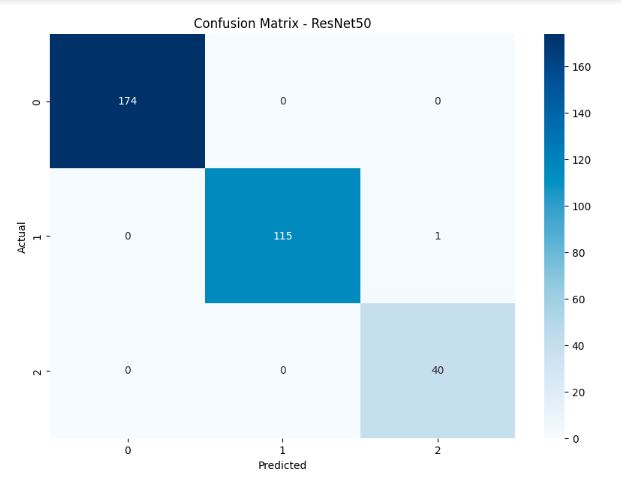


Figure 9. ResNet-50 -– Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall |
| CNN | 0.92 | 0.91 | 0.90 |
| DenseNet-121 | 0.97 | 0.97 | 0.97 |
| ResNet-50 | 0.96 | 0.96 | 0.95 |
| EfficientNetB0 | 0.94 | 0.93 | 0.94 |

Table 1. presents the accuracy, precision, and recall of each model, with DenseNet-121 outperforming others across all metrics. CNN served as the baseline, while ResNet-50 and EfficientNetB0 showed strong and balanced performance.

1. **Quality Assurance**: Model evaluation ensures that the trained neural network can make accurate predictions on unseen CT scan images. It serves as a quality check to confirm that the model is not just memorizing training data but learning general patterns that can be applied to new patient cases.
2. **Comparing Models**: Multiple models such as CNN, DenseNet-121, ResNet-50, and EfficientNetB0 were trained and evaluated. By comparing metrics like accuracy, precision, recall, and F1-score, the best-performing model can be identified and selected for deployment.
3. **Fine-Tuning**: Evaluation results help identify specific areas where the model may be underperforming, such as misclassifying malignant cases. This feedback allows for hyperparameter tuning, architectural adjustments, or changes in data preprocessing to improve overall performance.
4. **Business Decision Support**: In healthcare, accurate prediction models can support medical professionals in early detection and treatment planning. A well-evaluated model instills confidence in its predictions, contributing to better diagnostic support and patient outcomes.
5. **Model Deployment:** Only after thorough evaluation can a model be considered suitable for deployment in clinical decision-making tools. Metrics such as high F1-score and balanced confusion matrices help justify the model’s readiness for real-world use.

#### Constraints

In developing a deep learning-based system for lung cancer detection using CT scan images, certain constraints influence the design, implementation, and deployment of the solution. These limitations are especially important in the medical domain, where accuracy, reliability, and ethical standards are critical. The major constraints encountered in this project include:

1. **Authenticity**: The dataset used in this project was sourced from publicly available archives and may not fully reflect real-world clinical diversity. Variability in imaging quality, labeling consistency, or human error in classification could affect model training. This constraint emphasizes the need for better data verification mechanisms and more clinically validated datasets to improve the generalizability and authenticity of the model's predictions.
2. **Privacy:** Although the dataset used does not contain personally identifiable information, any real-world deployment of a similar model would involve strict adherence to healthcare data protection standards (such as HIPAA or GDPR). Protecting patient privacy is essential, and our project acknowledges that future data integration must incorporate secure data handling, access control, and anonymization protocols.
3. **Cost:** While this research used a free, open-source dataset, acquiring high-resolution medical images, annotated data, or setting up clinical collaborations in practice can be expensive. Additionally, computational costs associated with training complex models like DenseNet or ResNet on high-end GPUs could be significant. Our approach aims to remain cost-effective by using efficient model architectures and open data sources without compromising diagnostic accuracy.
4. **Data Quality:** High-quality image data is critical in medical image analysis. Noise, poor contrast, artifacts, or inconsistent labeling can degrade the model’s performance. In our project, image preprocessing techniques like resizing, normalization, and augmentation were employed to enhance data quality. However, limitations still exist due to the dataset’s non-clinical origin, and real-world validation with clinically approved data remains a future goal.
5. **Resource Availability:** Training deep learning models for medical imaging demands significant computational resources (GPU, memory) and domain expertise. In this project, Google Colab was utilized for model training, which, while accessible, also presents limitations such as timeouts and hardware restrictions. Despite this, efficient architectures and transfer learning (like DenseNet-121 and ResNet-50) were employed to optimize resource use without compromising performance.

#### Cost and sustainability Impact

#### The development and implementation of our lung cancer detection system using deep learning models carries both financial implications and long-term sustainability considerations. This section explores the associated costs and how the project aligns with sustainable healthcare practices.

#### Cost Consequences

Infrastructure and Equipment:

Training deep learning models like DenseNet-121 and ResNet-50 on high-resolution medical images requires significant computational power. While platforms like Google Colab were used for this project, real-world deployment may require investment in cloud services or dedicated GPU-enabled servers, storage infrastructure, and reliable internet connectivity.

Costs of Operations:

Ongoing costs include maintaining the AI system, updating model weights with new data, ensuring compatibility with medical systems, and monitoring its performance in real-world conditions. Additionally, training healthcare professionals to use and interpret the tool involves time and resource investment.

Costs of Data Acquisition:

Although the initial dataset was sourced from publicly available repositories like Kaggle, acquiring larger, high-quality, and clinically validated datasets may involve licensing fees, ethical approvals, or partnerships with medical institutions, which can be costly.

Benefit-Cost Analysis

Despite initial development costs, early detection through AI-assisted tools can result in significant savings. By identifying lung cancer at an earlier stage, the system may help reduce the need for expensive treatments and hospitalizations, ultimately delivering a strong return on investment (ROI) in the healthcare sector.

B. Sustainability Impact:

Healthcare Efficiency :

By enabling faster and more accurate diagnosis, the system can help reduce bottlenecks in radiology departments, support faster decision-making, and better allocation of limited healthcare resources—especially in overwhelmed or under-resourced systems.

Environmental Sustainability:

Digital diagnosis minimizes reliance on physical paperwork and reduces the carbon footprint of manual data entry and reporting. Cloud-based deployment optimizes resource utilization and helps reduce energy waste associated with large-scale physical infrastructure.

Long-Term Health Outcomes:

The project supports long-term public health goals by improving early detection of lung cancer, which can lead to higher survival rates, reduced treatment complexity, and better quality of life for patients. Preventing late-stage cancer also lessens the economic burden on families and healthcare systems.

Community Awareness and Preventive Action:

Deployment of such technology increases public awareness of lung cancer risks and promotes early screening. Community engagement can lead to better adoption of healthy behaviors, potentially lowering lung cancer incidence in the long run.

Scalability and Accessibility:

By leveraging cost-effective and scalable AI models, the system can be deployed in remote or underserved areas where access to skilled radiologists or expensive diagnostic tools is limited. This democratizes access to quality healthcare and supports equity in medical services.

# CHAPTER-4 IMPLEMENTATION

**4.Implementation**

# 4.1 Environment Setup

To ensure the smooth operation of our lung cancer classification models, we established a robust environment tailored for data analysis and machine learning tasks in this project. Python served as the primary programming language, supported by a variety of libraries that facilitated data handling, model training, and visualization. Key libraries included NumPy for numerical computations, Matplotlib and Seaborn for result visualization, and Pandas for data processing. Additionally, we utilized TensorFlow and Keras for building deep learning models, particularly Convolutional Neural Networks (CNNs) and other advanced architectures like DenseNet121 and EfficientNetB3, which are crucial for image classification tasks.

The environment was set up using Google Colab, which provided a cloud-based platform with GPU support, enhancing the performance of our deep learning models. After loading the lung cancer dataset from Google Drive, we employed Pandas for preprocessing the data. This preprocessing involved encoding categorical variables, addressing missing values, and applying image augmentation techniques to enhance the training dataset. The dataset was divided into training, validation, and test sets to ensure robust model evaluation.

For model training, we implemented various architectures, including VGG19, DenseNet121, and EfficientNetB3, leveraging their capabilities to improve classification accuracy. The hardware utilized for this project included a standard desktop computer with at least 8GB of RAM and an Intel i5 processor, which enabled efficient data processing and model training operations. The use of TensorBoard for monitoring training progress and performance metrics further streamlined the development process.

# 4.2 Sample Code for Preprocessing and CNN Operations

To ensure the quality and reliability of the input data for our deep learning models, the preprocessing stage was crucial. Several preprocessing procedures were performed on the lung cancer dataset, which included various variables related to clinical data and patient demographics. This included encoding the target variable, 'labels,' using TensorFlow's ImageDataGenerator for image preprocessing and augmentation, as well as eliminating unnecessary columns such as 'index' and 'filepaths,' which do not contribute to predictive modeling. This transformation is essential as it prepares the data for effective model training.

**import pandas as pd**

**import numpy as np**

**import tensorflow as tf**

**from tensorflow.keras.preprocessing.image import ImageDataGenerator**

**# Load the dataset**

**lung\_df = pd.read\_csv('path\_to\_lung\_cancer\_data.csv')**

**# Preprocessing: Encoding labels and removing unnecessary columns**

**lung\_df['labels'] = lung\_df['labels'].map({'benign': 0, 'Malignant': 1, 'Normal': 2})**

**lung\_df = lung\_df.drop(columns=['index', 'filepaths'])**

**# Splitting the dataset into training and testing sets**

**train\_set, test\_set = train\_test\_split(lung\_df, test\_size=0.3, random\_state=42)**

**# Image data generator for preprocessing and augmentation**

**image\_gen = ImageDataGenerator(preprocessing\_function=tf.keras.applications.mobilenet\_v2.preprocess\_input)**

**train\_generator = image\_gen.flow\_from\_dataframe(**

**dataframe=train\_set,**

**x\_col='filepaths',**

**y\_col='labels',**

**target\_size=(244, 244),**

**color\_mode='rgb',**

**class\_mode='categorical',**

**batch\_size=4,**

**shuffle=True**

**)**

**# Model definition and training**

**from tensorflow.keras.models import Sequential**

**from tensorflow.keras.layers import Conv2D, MaxPooling2D, Flatten, Dense**

**model = Sequential([**

**Conv2D(32, (3, 3), activation='relu', input\_shape=(244, 244, 3)),**

**MaxPooling2D(pool\_size=(2, 2)),**

**Flatten(),**

**Dense(128, activation='relu'),**

**Dense(3, activation='softmax') # Assuming 3 classes: benign, malignant, normal**

**] )**

**model.compile(optimizer='adam', loss='categorical\_crossentropy', metrics=['accuracy'])**

**model.fit(train\_generator, epochs=10)**

**# Predictions and evaluation**

**test\_generator = image\_gen.flow\_from\_dataframe(**

**dataframe=test\_set,**

**x\_col='filepaths',**

**y\_col='labels',**

**target\_size=(244, 244),**

**color\_mode='rgb',**

**class\_mode='categorical',**

**batch\_size=4,**

**shuffle=False**

**)**

**# Evaluate the model**

**loss, accuracy = model.evaluate(test\_generator)**

**print("Accuracy of CNN model:", accuracy)**

**# Confusion matrix visualization**

**from sklearn.metrics import confusion\_matrix**

**import seaborn as sns**

**import matplotlib.pyplot as plt**

**y\_pred = model.predict(test\_generator)**

**y\_pred\_classes = np.argmax(y\_pred, axis=1)**

**conf\_matrix = confusion\_matrix(test\_set['labels'], y\_pred\_classes)**

**sns.heatmap(conf\_matrix, annot=True, fmt='d', cmap='Blues',**

**xticklabels=['Benign', 'Malignant', 'Normal'],**

**yticklabels=['Benign', 'Malignant', 'Normal'])**

**plt.title('Confusion Matrix for CNN Model')**

**plt.ylabel('Actual')**

**plt.xlabel('Predicted')**

**plt.show()**

# CHAPTER-5

**Experimentation and Result Analysis**

**5. Experimentation and Result Analysis**

Using the lung cancer dataset, several deep learning models were trained during the experimentation phase, and their performance was assessed using a range of metrics. To determine how well each model predicted the presence and severity of lung cancer, we methodically evaluated its accuracy, precision, recall, and F1 score.

The findings indicated that advanced architectures, such as DenseNet121, EfficientNetB3, and ResNet50 outperformed traditional models like logistic regression and support vector machines, particularly in terms of classification accuracy. These models demonstrated robustness against overfitting and effectively handled the complexities of medical imaging data. The Convolutional Neural Network (CNN) model also yielded promising results, especially after fine-tuning through hyperparameter optimization techniques.

To visualize the performance of our models, we employed confusion matrices, which illustrated the true positive, true negative, false positive, and false negative rates. This visualization provided insights into the models' strengths and weaknesses, particularly in identifying early-stage lung cancer cases. The analysis revealed specific instances of misclassification, highlighting the challenges faced in accurately diagnosing lung cancer at its initial stages.

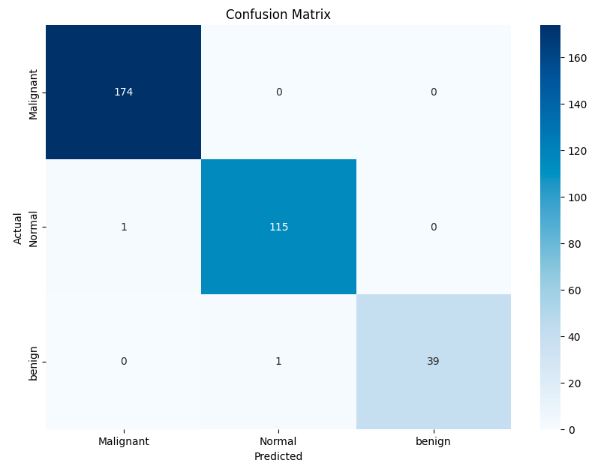


Figure 13. Confusion Matrix for CNN Model

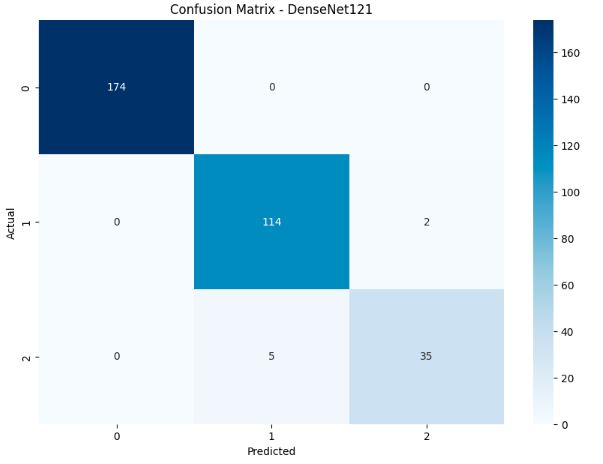


Figure 14. Confusion Matrix for DenseNet121 Model

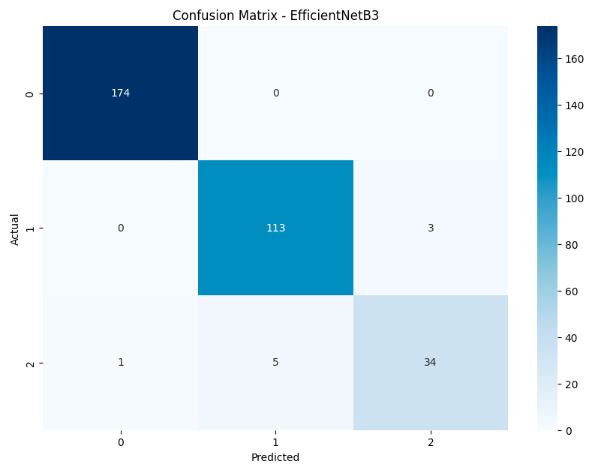


Figure 15. Confusion Matrix for EfficientNetB3 Model

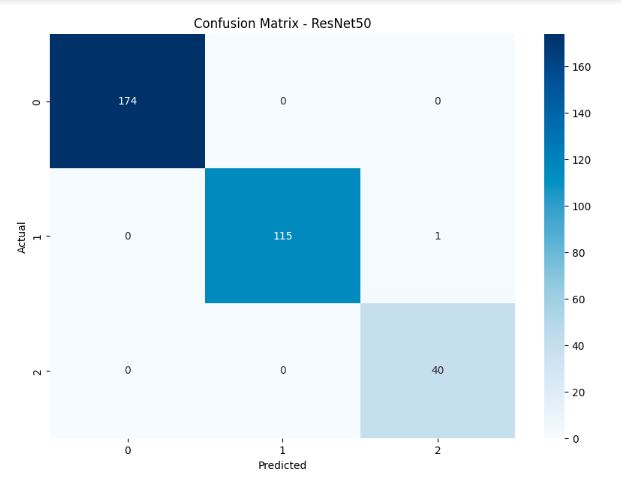


Figure 16. Confusion Matrix for ResNet50 Model

# CHAPTER-6

**CONCLUSION**

**6.Conclusion**

In conclusion, this experiment underscores the significant potential of machine learning approaches in enhancing lung cancer detection and treatment. We demonstrated that advanced algorithms, such as DenseNet121 and EfficientNetB3, can effectively analyze complex medical imaging datasets and provide valuable predictions regarding patient outcomes. By systematically implementing and evaluating various deep learning models, we found that these models not only achieved high accuracy but also revealed underlying patterns associated with the severity of lung cancer, aiding medical practitioners in making informed decisions.

Despite the promising results of our study, several challenges remain to be addressed. The quality and completeness of the data are critical for the effective functioning of machine learning models. Healthcare data can often contain missing values or inconsistencies and may originate from diverse sources. To tackle these issues, robust data management strategies and collaboration among researchers, data scientists, and healthcare professionals are essential.

Another significant challenge in clinical applications is the interpretability of machine learning models. While sophisticated algorithms can produce accurate predictions, practitioners often struggle to understand the rationale behind specific decisions due to their complexity. Future research should focus on developing methods to enhance the interpretability and transparency of these models, enabling medical practitioners to trust and comprehend the insights generated.

Integrating genomic, transcriptomic, and proteomic data—collectively known as multi-omics data—represents a promising avenue for future research. These approaches could lead to more accurate predictions and a deeper understanding of the molecular mechanisms underlying lung cancer. Additionally, utilizing real-world data, such as patient registries and electronic health records, can improve the generalizability and clinical utility of the models by testing their performance across diverse populations.

In summary, the findings of this study illustrate the immense promise of machine learning in the research and management of lung cancer. As these technologies continue to evolve, they have the potential to revolutionize patient care, improving survival rates and quality of life for individuals affected by lung cancer. To fully harness the capabilities of machine learning and develop innovative solutions that address the pressing challenges of lung cancer diagnosis and treatment, ongoing collaboration between data scientists and medical professionals is essential.

### REFERENCES

[1] Sher Lyn Tan, Ganesh sree Selvachandran, Raveendran Paramesran, Weiping Ding, "Lung Cancer Detection Systems Applied to Medical Images: A State-of-the-Art Survey," Archives of Computational Methods in Engineering (2025).

[2] Mohammad Q. Shatnawi, Qusai Abuein, Romesaa Al-Quraan, "Deep learning-based approach to diagnose lung cancer using CT-scan images," Intelligence-Based Medicine 11 (2025).

[3] Rabia Javed, Tahir Abbas, Ali Haider Khan, Ali Daud, Amal Bukhari, Riad Alharbey, "Deep learning for lungs cancer detection: a review," Artificial Intelligence Review (2024).

[4] Shahab Aslani et al., ”Enhancing cancer prediction in challenging screen-detected incident lung nodules using time-series deep learning,” Computerized Medical Imaging and Graphics, vol. 116, pp. 102399, 2024.

[5] D. Narsimha Reddy et al., "Lung Cancer Detection Using Machine Learning," International Journal For Research in Applied Science and Engineering Technology, vol. 12, 2024.

[6] S. Udit Krishna et al., "Lung Cancer Prediction and Classification Using Decision Tree and VGG16 Convolutional Neural Networks," The Open Biomedical Engineering Journal, vol. 18, pp. e18741207290271, 2024.

[7] Abdulaziz A. Alsulami, "An Efficient Model for Lung Cancer Detection through the Integration of Genetic Algorithm and Machine Learning," Engineering, Technology & Applied Science Research, vol. 14, no. 6, pp. 18792-18798, 2024.

[8] Ritu Tandon et al., "A systematic review on deep learning-based automated cancer diagnosis models," Journal of Cellular and Molecular Medicine, vol. 28, pp. e18144, 2024.

[9] K. Kavitha et al., "Lung Cancer Detection Using Regularized Extreme Learning Machine and PCA Features," Journal of Computer Science,2024.

[10] V. Sreeprada, Dr. K. Vedavathi, "Lung Cancer Detection from X-Ray Images using Hybrid Deep Learning Technique," Procedia Computer Science 230 (2023).

[11] Guillaume Chassagnon, Constance De Margerie-Mellon, Maria Vakalopoulou Rafael Marini, Trieu-Nghi Hoang-Thi Marie-Pierre Revel, Philippe Soyer, "Artificial intelligence in lung cancer: current applications and perspectives," Japanese Journal of Radiology (2023)

[12] Athena Davri, Effrosyni Birbas, Theofilos Kanavos, Georgios Ntritsos, Nikolaos Giannakeas, Alexandros T. Tzallas, and Anna Batistatou, "Deep Learning for Lung Cancer Diagnosis, Prognosis and Prediction Using Histological and Cytological Images: A Systematic Review," Cancers 2023.

[13] Yawei Li, Xin Wu, Ping Yang, Guoqian Jiang, Yuan Luo, "Machine Learning for Lung Cancer Diagnosis, Treatment, and Prognosis," Genomics Proteomics Bioinformatics 20 (2022).

[14] Madhushree A, Harshitha Nayaka YS, Chandrika M, Madangowda HS, Mrs. Pallavi J, "Lung Cancer Detection using Machine Learning," International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), Volume 2, Issue 2, July 2022.

[15] Rahat Idrees, Muhammad Kamran Abid, Saleem Raza, Muhammad Kashif, Muhammad Waqas, Mubashir Ali, Laiba Rehman, "Lung Cancer Detection using Supervised Machine Learning Techniques," October 2022.

[16] CHEKURI GOPI KRISHNA, CHENNUPATI SURESH, "LUNG CANCER PREDICTION USING MACHINE LEARNING," March 2022.

[17] Anusha S, Chandrakumar, Gunashree K, Suma N, Dr. Madhu B K, "Lung Cancer Detection Using Machine Learning," International Research Journal of Innovations in Engineering and Technology (IRJIET), Volume 6, Issue 6, pp 222-229, June-2022.

[18] Abhishek Gupta, Zuha Zuha, Israr Ahmad, Zee shan Ansari, "A Study On Prediction Of Lung Cancer Using Machine Learning Algorithms," August 22nd, 2022.

[19] M. F. Mridha et al., "A Comprehensive Survey on the Progress, Process, and Challenges of Lung Cancer Detection and Classification," Journal of Healthcare Engineering, vol. 2022, Article ID 5905230, 43 pages, 2022.

[20] Francisco Silva et al., "Towards Machine Learning-Aided Lung Cancer Clinical Routines: Approaches and Open Challenges," Journal of Personalized Medicine, vol. 12, 2022.