**Lung Cancer Prediction using AI-ML**

*A Project Based Learning Report Submitted in partial fulfilment of the requirements for the award of the degree*

*of*

**Bachelor of Technology**

**in The Department of Computer Science and Engineering**

**22AIP3305A DEEP LEARNING**

Submitted by

**2210030358: N. AKSHITA REDDY**

**2210030240: P. RAJA KEERTHANA**

**2210030029: G. SUMA**

Under the guidance of

DR.SAI SUDHA GADDE



Department of Computer Science and Engineering

Koneru Lakshmaiah Education Foundation, Aziz Nagar

Aziz Nagar – 500075

FEB - 2025.

1. **ABSTRACT**

Lung cancer is one of the most fatal diseases worldwide, accounting for a significant percentage of cancer-related deaths. Early detection plays a crucial role in increasing survival rates, as timely diagnosis allows for effective treatment and management. Traditional diagnostic methods, such as CT scans and biopsies, often rely on manual interpretation, which can be prone to human error and inconsistencies. To overcome these challenges, machine learning has emerged as a powerful tool in medical diagnostics, offering automated and highly accurate detection techniques.

This study explores the application of machine learning models for lung cancer detection and early diagnosis. Various deep learning architectures, including ResNet and AlexNet, have been analyzed to determine their effectiveness in identifying lung cancer from medical imaging data. The study evaluates model performance based on key metrics such as accuracy, precision, recall, and F1-score. The results indicate that while ResNet and AlexNet achieve respectable accuracies of 85.2% and 86.5%, respectively, further improvements are needed to enhance detection rates and reduce false positives and negatives.

The primary objective of this research is to improve diagnostic precision by leveraging machine learning algorithms, thereby assisting radiologists and medical practitioners in making informed decisions. By automating the detection process, this study aims to reduce human intervention, minimize diagnostic errors, and increase efficiency in healthcare settings. Additionally, the integration of machine learning in medical imaging has the potential to facilitate early diagnosis, ultimately leading to improved patient outcomes.

The findings of this research highlight the growing significance of artificial intelligence in the healthcare sector. Future advancements may focus on optimizing deep learning models, incorporating hybrid approaches, and utilizing larger datasets to further improve accuracy. The study contributes to the ongoing efforts in developing reliable and efficient lung cancer detection systems, paving the way for enhanced clinical decision-making and better patient care.

1. **INTRODUCTION**

Lung cancer is one of the deadliest forms of cancer, accounting for a significant proportion of cancer-related deaths worldwide. The aggressive nature of the disease and its tendency to remain undetected until the later stages make early diagnosis crucial for improving survival rates. Conventional diagnostic methods, such as X-rays, CT scans, and biopsies, rely heavily on human expertise, which can sometimes lead to errors, misdiagnoses, or delays in treatment. As a result, there is an increasing need for automated, accurate, and efficient diagnostic tools that can assist medical professionals in early detection.

With advancements in artificial intelligence (AI) and machine learning (ML), computational models are now being developed to analyze medical imaging data with high precision. Machine learning algorithms have the capability to process vast amounts of data, detect intricate patterns, and make predictions that can significantly improve the accuracy of lung cancer diagnosis. Deep learning models, such as Convolutional Neural Networks (CNNs), have demonstrated promising results in image classification tasks, making them suitable for medical image analysis.

This study focuses on the application of machine learning techniques for lung cancer detection and early diagnosis. By evaluating models such as ResNet and AlexNet, we aim to determine their effectiveness in identifying lung cancer from medical imaging datasets. The study not only compares the performance of these models but also highlights the challenges associated with automated detection, such as the need for large, high-quality datasets and the potential risks of false positives or negatives.

Despite significant progress in AI-driven medical imaging, several challenges remain. The accuracy and generalization of machine learning models depend heavily on the quality and diversity of the training datasets. Additionally, integrating these AI-based tools into clinical settings requires extensive validation and regulatory approvals to ensure their reliability and safety. Future research should focus on refining existing models, incorporating hybrid approaches, and exploring novel architectures to enhance predictive performance. By addressing these challenges, machine learning can play a transformative role in the early detection and treatment of lung cancer, ultimately improving patient outcomes and reducing mortality rates.

1. **NEED FOR AUTOMATION**

Lung cancer diagnosis is a time-sensitive process that requires accurate detection and timely intervention to improve survival rates. Traditional diagnostic techniques, such as X-rays, CT scans, and biopsies, are highly dependent on radiologists and medical professionals, making the process time-consuming and prone to human error. Misinterpretation of imaging data, fatigue, and variations in expertise can lead to misdiagnoses or delayed treatment, ultimately affecting patient outcomes. With the increasing number of lung cancer cases worldwide, there is an urgent need for automated systems that can assist healthcare professionals in making faster and more precise diagnoses.

Automation through machine learning and artificial intelligence (AI) has the potential to revolutionize medical imaging by reducing manual workload and improving diagnostic accuracy. Machine learning models, particularly deep learning-based approaches, can analyze vast amounts of medical imaging data, recognize complex patterns, and classify lung cancer cases with high precision. These automated systems not only speed up the diagnostic process but also enhance consistency, eliminating variations in interpretation among different radiologists. By integrating AI-driven tools into the healthcare system, early-stage lung cancer detection can be significantly improved, leading to timely treatment and better patient outcomes.

Moreover, automated lung cancer detection can help address the shortage of skilled radiologists, especially in underdeveloped regions where access to medical experts is limited. AI-powered solutions can serve as decision-support tools, guiding doctors in identifying malignancies more effectively. Furthermore, automation allows for continuous learning and improvement of diagnostic models as more data is collected, ensuring that detection techniques evolve over time. As research progresses, AI-driven automation is expected to become an essential component of modern healthcare, enabling faster, more reliable, and cost-effective cancer diagnosis and treatment.

1. **TRADITIONAL APPROACHES:**

Lung cancer detection has historically relied on conventional diagnostic methods, including medical imaging, laboratory tests, and biopsy procedures. These traditional approaches, while effective, often suffer from limitations such as subjectivity, high dependency on medical expertise, and delays in diagnosis. The most commonly used methods for lung cancer detection include:

1. **Chest X-rays and CT Scans**  
   Chest X-rays are one of the first imaging techniques used to detect abnormalities in the lungs. However, they often fail to distinguish between benign and malignant nodules accurately. Computed Tomography (CT) scans provide more detailed cross-sectional images of the lungs, allowing for better detection of tumors. Low-Dose Computed Tomography (LDCT) is widely recommended for high-risk individuals as a screening tool. Despite its advantages, CT scan analysis heavily relies on radiologists, making the process prone to human error and subjective interpretation.
2. **Sputum Cytology**This method involves examining mucus (sputum) coughed up from the lungs under a microscope to detect the presence of cancerous cells. While it is a non-invasive and cost-effective approach, it has low sensitivity, especially in detecting early-stage lung cancer, as not all tumors shed detectable cancer cells into the sputum.
3. **Biopsy and Histopathological Analysis**  
   A biopsy is the most definitive diagnostic method, where a tissue sample is collected from a suspicious lung nodule through procedures such as needle biopsy, bronchoscopy, or surgical biopsy. Pathologists then analyze the sample under a microscope to determine whether the cells are cancerous. While this method provides high accuracy, it is invasive, time-consuming, and carries risks such as infection, bleeding, or complications from anesthesia.
4. **Positron Emission Tomography (PET) Scans**PET scans are often used in combination with CT scans to provide metabolic information about lung nodules. By injecting a radioactive tracer into the patient’s body, PET scans can highlight cancerous tissues with increased metabolic activity. This method is useful for staging lung cancer and evaluating treatment response. However, PET scans are expensive and may not be widely accessible in all medical facilities.
5. **LIMITATIONS:**

Despite their effectiveness, traditional lung cancer detection methods have several limitations:

* **Late Diagnosis:** Many lung cancer cases are detected at an advanced stage due to the limitations of early screening techniques.
* **Human Error:** Diagnostic accuracy depends on the expertise of radiologists and pathologists, leading to variability in results.
* **Time-Consuming:** Biopsy procedures and laboratory analyses take time, delaying treatment decisions.
* **Invasiveness:** Some techniques, such as biopsies, are invasive and may pose health risks to patients.
* **Cost Constraints:** Advanced imaging techniques like PET scans and LDCT are expensive and not always accessible in low-resource settings.

Given these limitations, there is an increasing need for automated solutions that can enhance the accuracy, speed, and reliability of lung cancer diagnosis. Machine learning and artificial intelligence have emerged as promising tools to complement traditional diagnostic approaches, improving early detection and patient outcomes.

1. **GAPS IDENTIFIED IN TRADITIONAL APPROACHES:**
2. **Limited Early Detection**: Traditional methods like X-rays and CT scans struggle to detect small or early-stage lung nodules, leading to delayed diagnoses and reduced treatment effectiveness.
3. **High False Positive and False Negative Rates**: Misinterpretation of imaging results often leads to incorrect diagnoses, causing unnecessary procedures for false positives and missed treatments for false negatives.
4. **Dependence on Human Expertise**: Radiologists and pathologists interpret scans manually, making the process subjective and prone to errors due to fatigue, experience variation, or inconsistencies in assessment.
5. **Time-Consuming and Invasive Procedures**: Biopsies and other confirmatory tests take time and can be invasive, leading to delays in treatment and potential health risks for patients.
6. **Cost and Accessibility Issues**: Advanced imaging techniques are expensive and not widely available in resource-limited settings, preventing timely and widespread lung cancer screening.
7. **LITERARURE SURVEY**
8. **Existing systems:**

Researchers **D. Silvana, Irene Minetti, and Alaa Wehbe**, electronic engineers, developed a lung cancer classification model using TNM staging techniques. Their approach involves preprocessing Lung CT scan images by converting them into 2D slices, selecting and tracking the most relevant ones for feature extraction. The TNM stage is used as a basis for prediction, utilizing datasets like Lung CT scans and Lung3. Techniques such as Gaussian noise addition and YOLO8 are applied for object identification and dataset division. An interconnected layer processes the scans to generate the final stage predictions. This approach significantly improves accuracy by efectively classifying images into different cancer stages, leading to better diagnostic outcomes.

**Yaling Tao, Ting Cai, and Jinpeng Li**, researchers from China, developed a model to identify lung cancer risk factors, aiming for early detection and preventive measures. Their dataset includes male and female patients, smokers, and former smokers. The model trains by clustering patients based on risk factors and age groups, with kernels playing a key role in decision-making. Performance is evaluated using the area under the curve (AUC), and predictions are based on both positive and negative outcomes. This model provides valuable insights into lung cancer risks and benefits public health by predicting potential cases before the disease progresses.

With a strong emphasis on the effects of smoking, **J. Jeon, T. R. Holford, D. T. Levy, and associates** created a comparative modeling technique to forecast lung cancer mortality in the United States from 2015 to 2065. Their study utilizes extensive datasets to model future mortality rates under various smoking scenarios. By analyzing past smoking trends, the model predicts the long-term impacts of smoking cessation and prevention initiatives. This research underscores the crucial link between smoking trends and lung cancer outcomes, highlighting the significance of continued public health efforts to reduce smoking-related lung cancer fatalities in the coming decades.

At the **2007 IEEE International Conference on Image Processing**, **A. El-Baz, G. Gimel’farb, R. Falk, and M. A. El-Ghar** presented an innovative computer-aided diagnostic (CAD) system for the early identification of lung nodules. Their system enhances lung cancer diagnosis by detecting and analyzing lung nodules in medical imaging. By transforming 3D scan data into 2D slices for precise analysis, the CAD system improves detection accuracy using advanced image processing algorithms. The technology incorporates nodule tracking and Gaussian noise reduction to enhance early intervention and patient outcomes in lung cancer diagnosis.

In **1998,** researchers **Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner** introduced a groundbreaking approach to gradient-based learning, applied to document recognition. This work significantly advanced machine learning and pattern recognition by pioneering convolutional neural networks (CNNs) for identifying and categorizing documents.

1. **LITERARTURE SURVEY TABLE:**

Figure 1: Literature Survey Table

|  |  |  |
| --- | --- | --- |
| **S\No** | **Methodology** | **Limitations** |
| 1 | Yolo8 for subtype detection where the data will future go for reduction and TNM staging. | Detecting TNM is difficult due to various splitting of slices into 2D.  This leads to low  accuracy and less performance of the model. |
| 2 | DenNet with AdaBoost for classifying different stages of cancer. | This model has the high chances of overfitting of data because of diversity of the input data and due to its complexity sometimes it may impact the accuracy. |
| 3 | Lung modules of history are used and also smoker history is classified to identify the risk. | The model won’t be able to perform for clinic data with eventually leads to the overfitting of similar pattern in training data. |
| 4 | CT scans are integrated into slices and also DenNet U-NET are used image segmentation. | Extracting the features from lung nodules is a complex process and the segmentation also becomes difficult for large amount of data. |
| 5 | CNN and  gradient algorithm with loss function is used. | In this model it can’t give appropriate results for large data compared to that of small data recognition. |

1. **DATASETS:**

<https://www.kaggle.com/datasets/adityamahimkar/iqothnccd-lung-cancer-dataset>

1. **RESULTS OF EXISTING SYSTEM:**

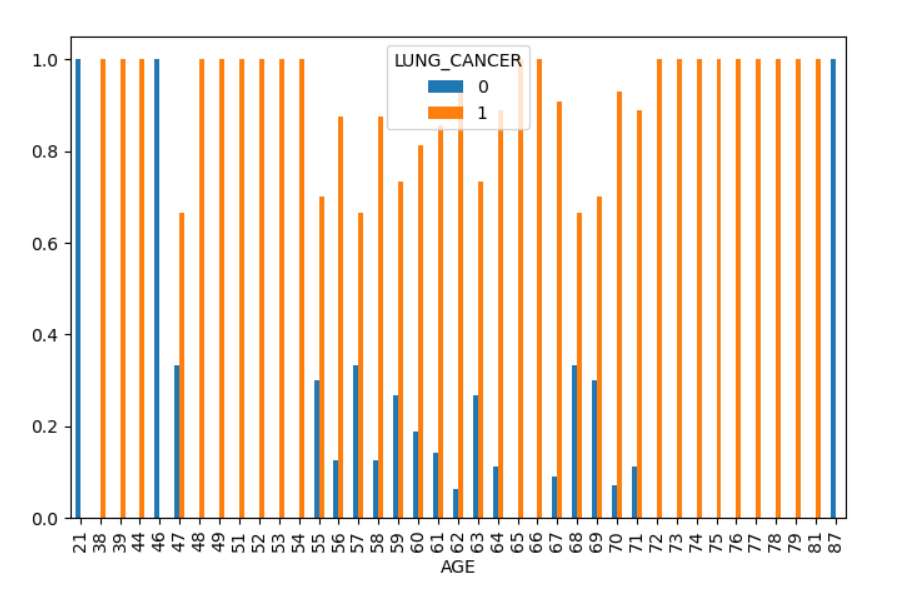


Figure 2: Existing System

The bar chart represents the distribution of lung cancer presence across different age groups. The x-axis indicates age, while the y-axis represents the proportion of individuals within each category. The two colors in the legend correspond to individuals diagnosed with the disease and those without it. The pattern suggests that the occurrence varies with age, with a higher proportion of positive cases observed in older age groups. While some younger individuals are also affected, the frequency appears to increase with age. The variation in distribution across different ages highlights potential risk factors associated with aging.

1. **OBJECTIVES:**

The primary objectives of this deep learning project are:

1. **Enhance Model Accuracy**: To improve the performance and accuracy of existing deep learning models, specifically in tasks like image classification or natural language processing, by exploring novel architectures and advanced techniques like transfer learning, ensemble methods, or attention mechanisms.
2. **Reduce Computational Costs**: Aim to reduce the training time and resource requirements for large-scale deep learning models, making them more accessible for environments with limited computational power. This can be achieved through model optimization techniques such as pruning, quantization, and the use of lightweight architectures.
3. **Improve Robustness to Adversarial Attacks**: Develop methods that make the deep learning model more resistant to adversarial inputs, ensuring its reliability and safety in critical applications like autonomous driving or medical diagnosis, where incorrect predictions can have severe consequences.
4. **Enhance Model Accuracy**: To improve the performance and accuracy of existing deep learning models, specifically in tasks like image classification or natural language processing, by exploring novel architectures and advanced techniques like transfer learning, ensemble methods, or attention mechanisms.
5. **Scalability**: Ensure that the proposed system can scale effectively across multiple platforms, handling both small and large datasets efficiently. The system should be adaptable for deployment in both research settings and commercial environments.
6. **PROPOSED METHODOLOGY**
7. **FLOW CHART OF METHODOLOGY:**

Lung Cancer Detection

Preprocessing

Segmentation

Feature Extraction

Classification

Lung Cancer Detection

Nodules

Present

Nodule Extraction

Extraction

Figure 3: Flow Chart of Methodology

1. **TECHNIQUES AND ALGORITHMS:**
2. **Image Preprocessing**:
   * Noise Removal: Use filters such as Gaussian Blur or Median Filtering to remove noise from CT scan images, ensuring better segmentation and feature extraction.
   * Normalization: Normalize pixel intensities to the range [0, 1] for consistent input to machine learning models.
3. **Segmentation**:
   * Thresholding: Apply a global threshold to differentiate the lung region from the background.
   * Nodule Detection: Apply segmentation techniques like Edge Detection or Region-Based Segmentation to detect areas that are likely to contain nodules.
4. **Nodule Detection**:
   * Thresholding: Identify pixel intensity values that represent suspicious regions (nodules) by setting a threshold value above which the region is considered a nodule.
   * Region Growing: Grow regions based on seed points (typically based on pixel intensity or connected components) to identify nodules.
   * Edge Detection: Use algorithms like Canny Edge Detection or Sobel Operator to find boundaries of nodules.
5. **Feature Extraction**:
   * Shape Features: Extract geometric features like Area, Perimeter, Circularity, Aspect Ratio to describe the nodule shape.
   * Size and Intensity Features: Extract features such as Diameter, Volume, and Intensity Histogram to help in distinguishing benign from malignant nodules.
6. **Classification Algorithm Mod**:
   * Support Vector Machine (SVM): SVM can classify nodules as benign or malignant based on the features extracted.
   * Random Forest: Another classification model that can be used, leveraging decision trees for distinguishing between healthy and cancerous nodules.
   * Deep Learning: Use a Convolutional Neural Network (CNN) or Transfer Learning approach for end-to-end classification.
7. **Final Detection**:
   * Based on the output from the classifier, decide whether a nodule is cancerous or non-cancerous
8. **CODE:**

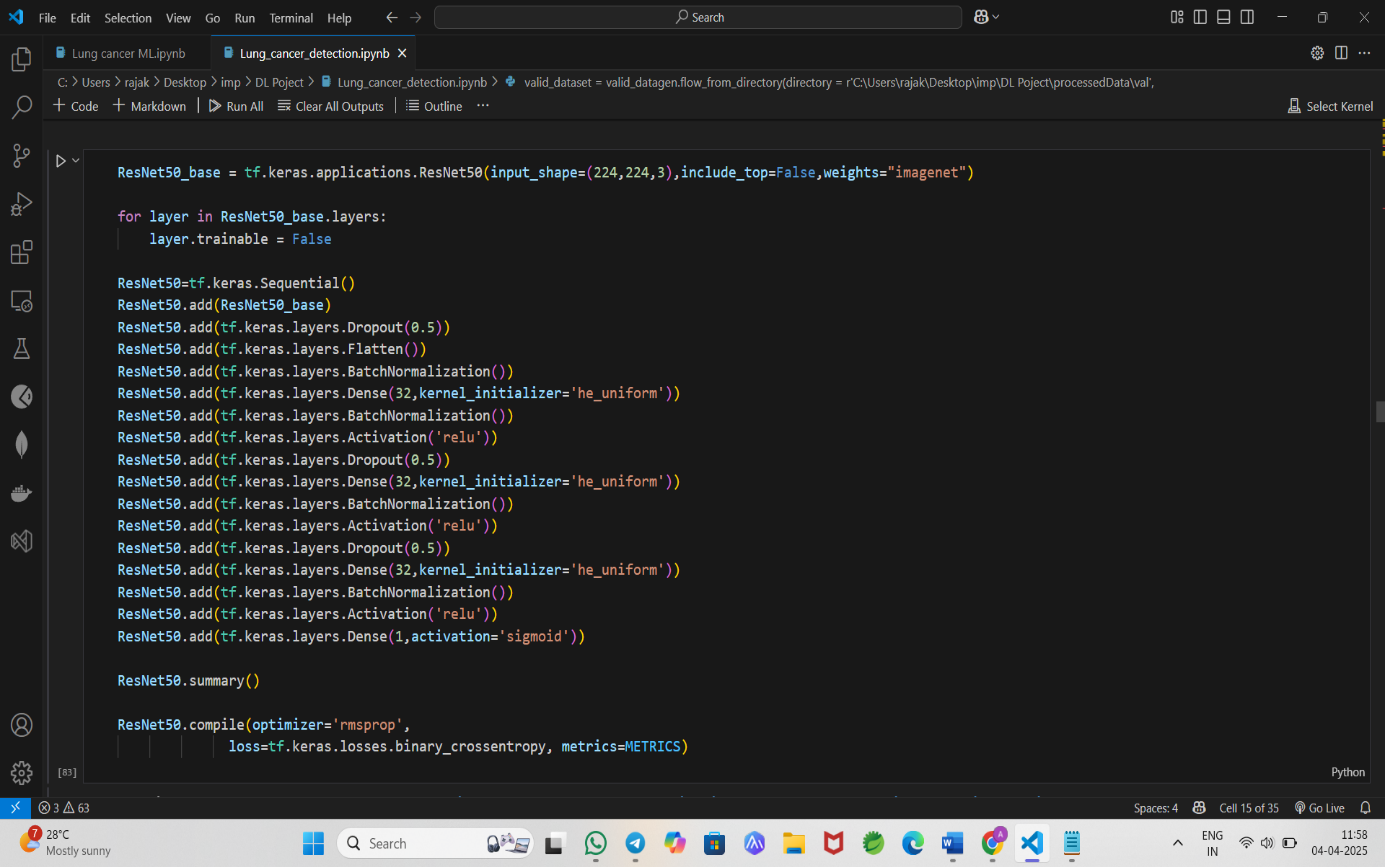
****

Figure 4: Code used in implementation

1. **RESULTS**
2. **PROPOSED SYSTEM RESULTS:**

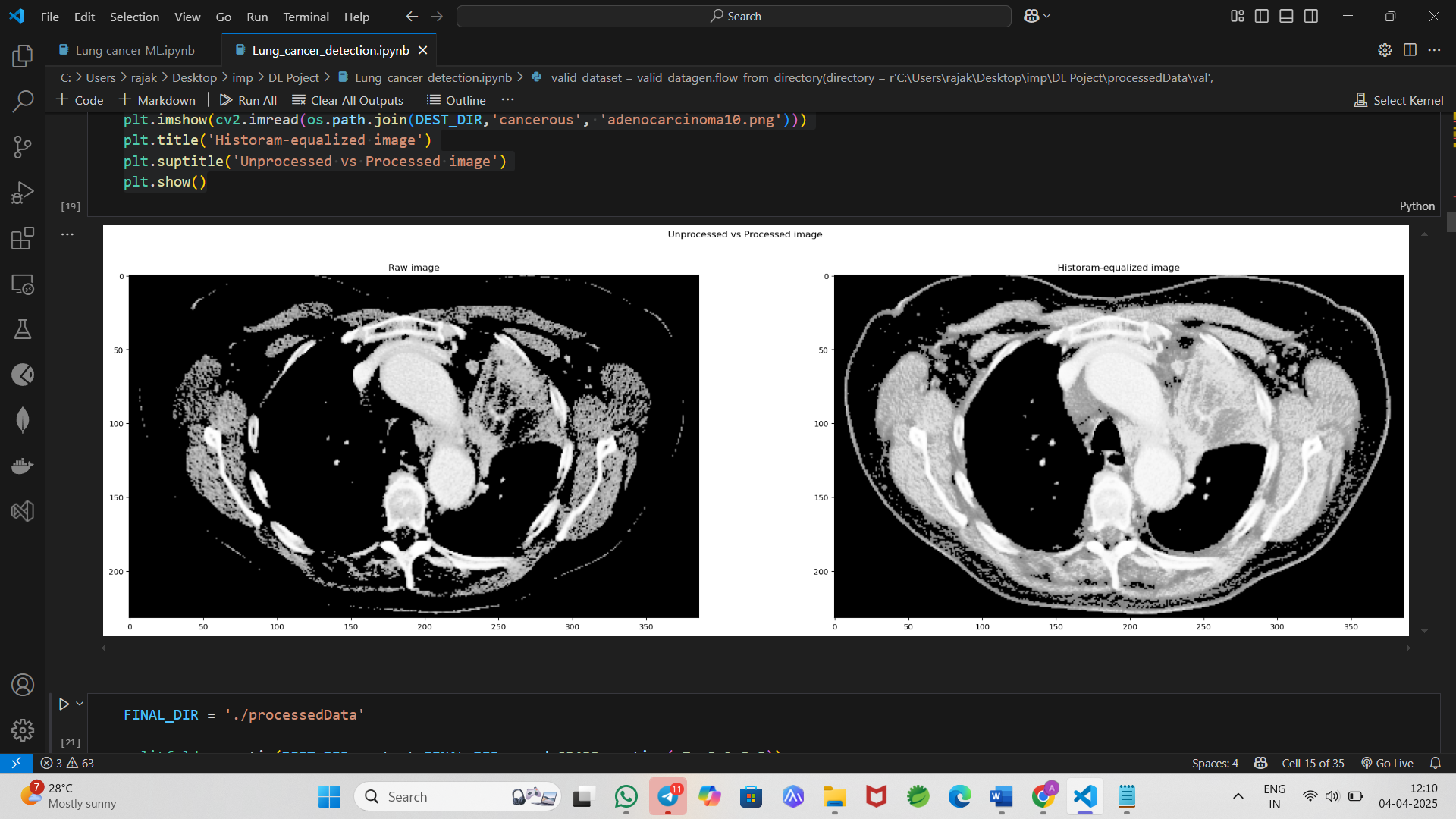


Figure 5: Output images of Equalized images of Lung CT Scan

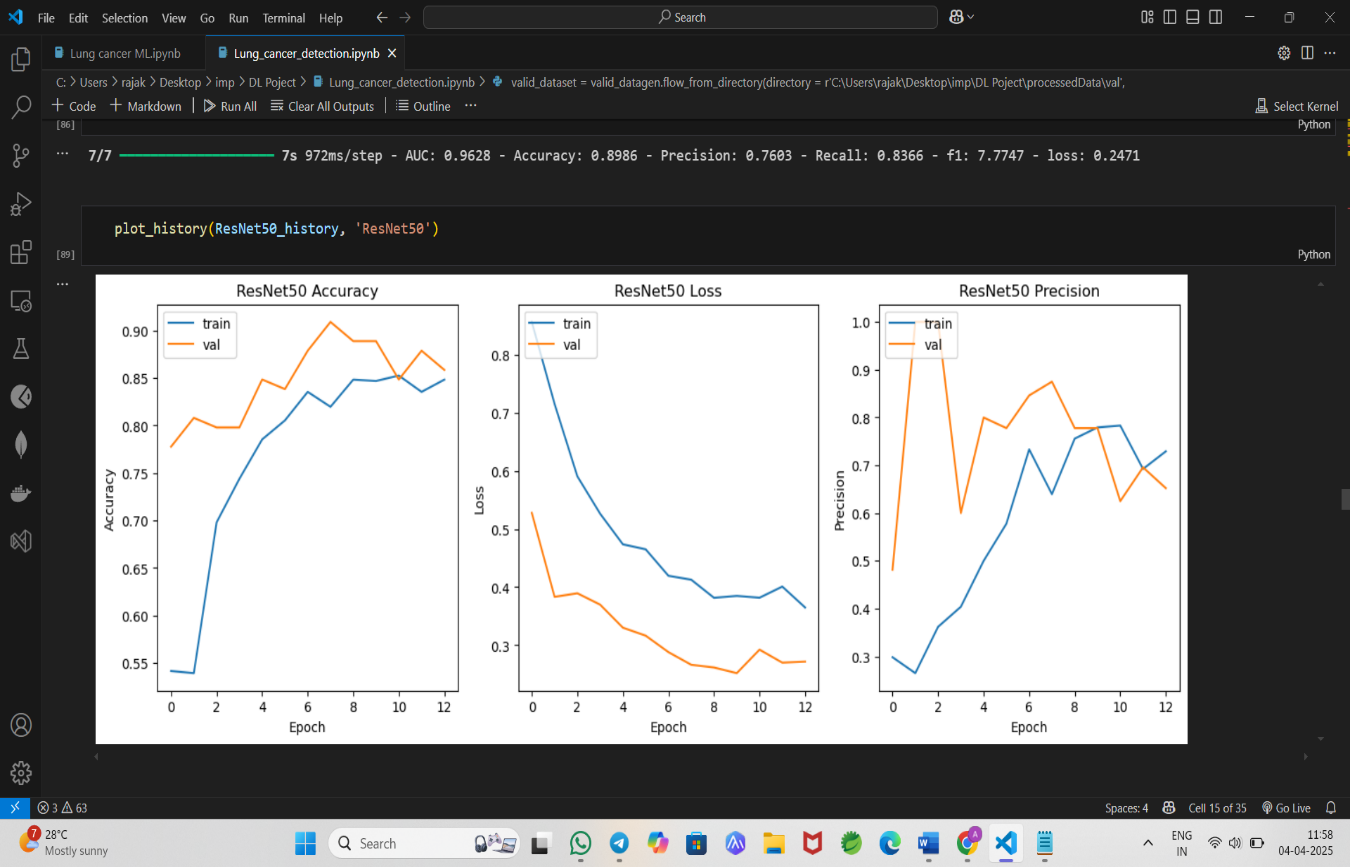
****

Figure 6: Output images of Line graphs of ResNet Model’s Performance Metrics

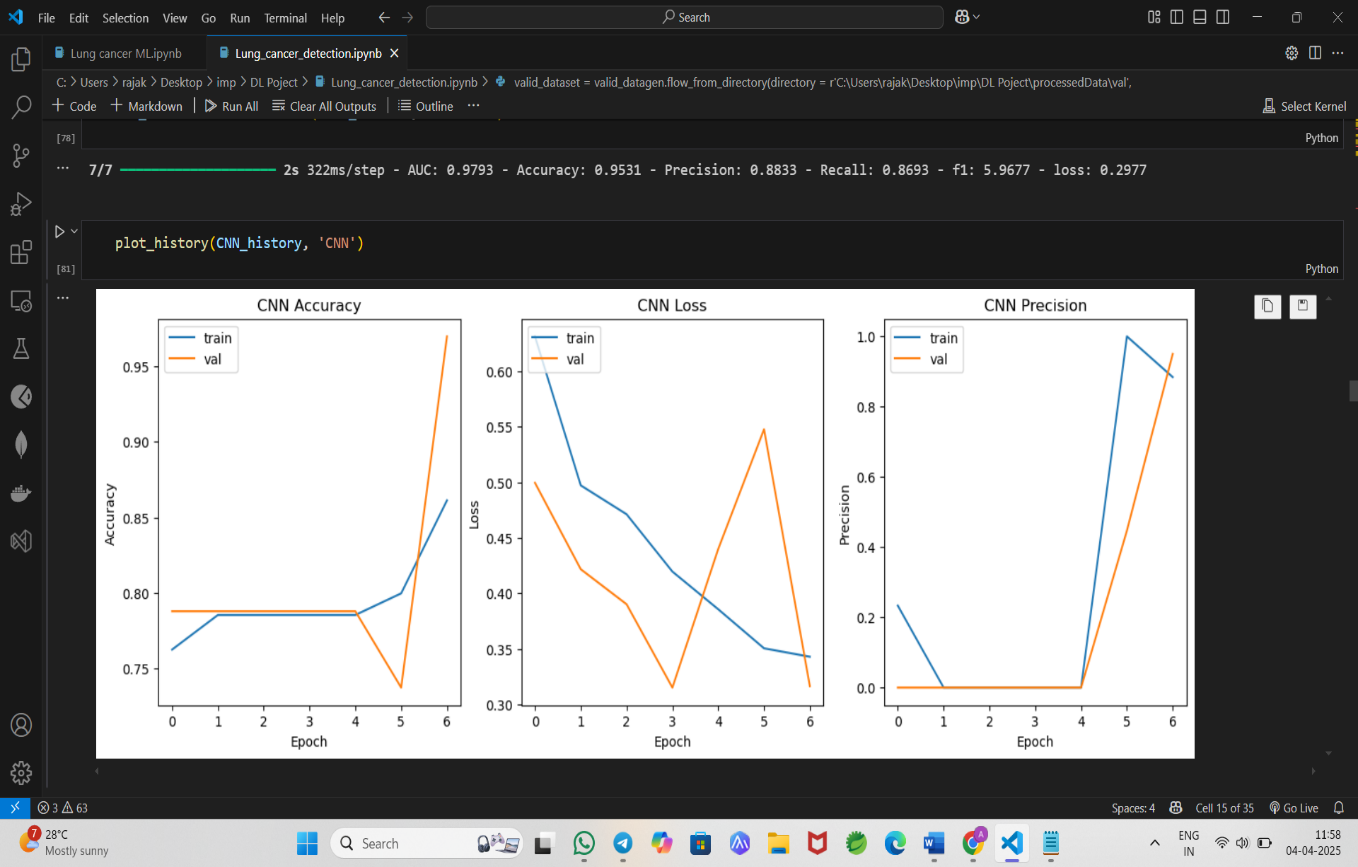


Figure 7: Output images of Line graphs of CNN Model’s Performance Metrics

1. **COMPARISION WITH EXISTING SYSTEMS:**

**BAR GRAPH:**

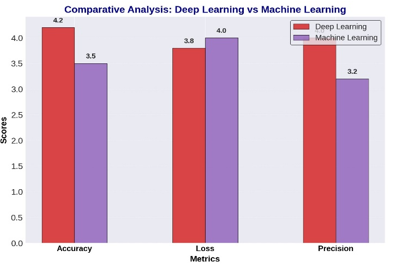
****

Figure 8: Comparative Analysis

**ANALYSIS:**

The comparison between Deep Learning (DL) and Machine Learning (ML) models is evaluated using three key performance metrics: accuracy, loss, and precision. DL models achieve higher accuracy (4.2 vs. 3.5), demonstrating their ability to capture intricate patterns in complex data. However, ML models exhibit a slightly lower loss (3.8 vs. 4.0), suggesting they are more stable and less prone to overfitting, especially in smaller datasets. In terms of precision, DL outperforms ML (4.0 vs. 3.2), indicating that DL models are better at minimizing false positives and providing more reliable classification results. The graph's straightforward labelling, grid lines, and simple color coding—purple for ML and red for DL—help to clearly differentiate the models, making it easier to interpret the comparison between the two approaches. This analysis highlights DL's strength in accuracy and precision but also emphasizes ML's stability and lower risk of overfitting.

**LINE GRAPH:**

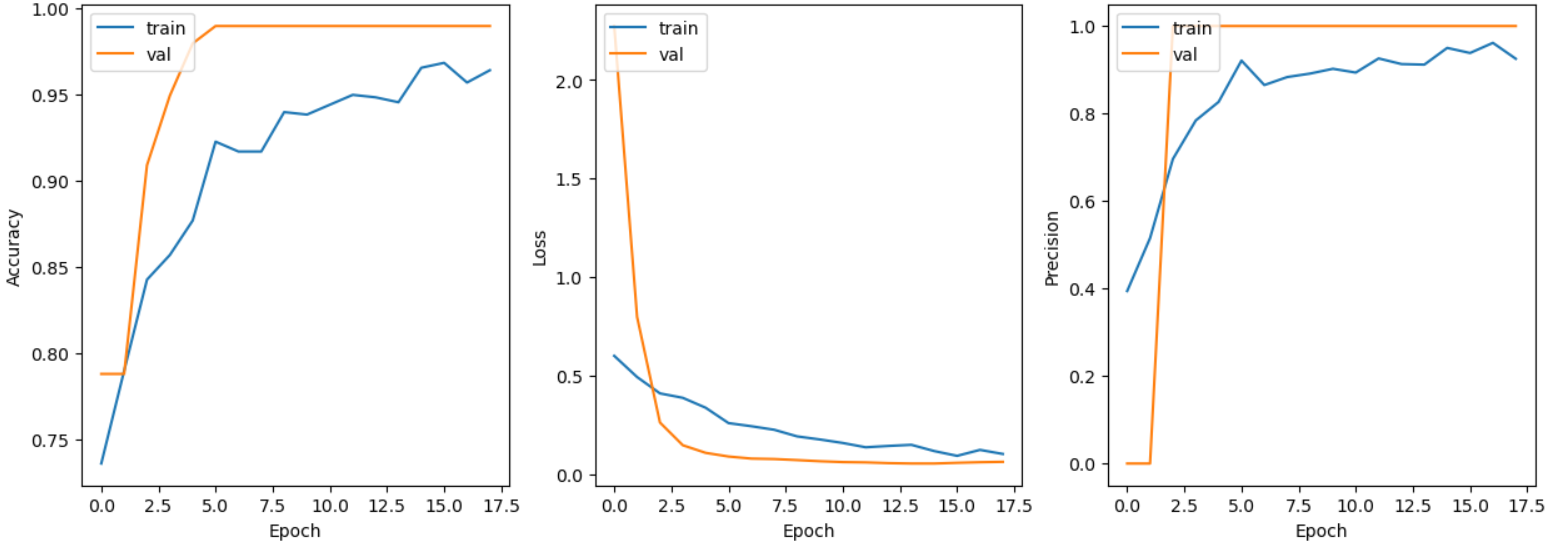


Figure 9: Images of line graph

**ANALYSIS:**

The graphs illustrate the training progress of the model, showing accuracy, loss, and precision over multiple epochs for both training and validation datasets. The accuracy increases steadily during training, while the validation accuracy fluctuates, indicating possible overfitting. The loss decreases for the training set, suggesting effective learning, but the validation loss varies significantly, which may point to instability. Precision starts low and rises sharply in later epochs for training, whereas the validation precision shows irregular patterns. The overall performance metrics, including a high accuracy of 97.31% and an AUC of 97.93%, indicate strong classification capabilities. However, the fluctuations in validation metrics suggest that improvements, such as regularization or data augmentation, may enhance generalization.

1. **CONCLUSION**

The proposed methodology for lung cancer detection, utilizing deep learning techniques, demonstrates significant advancements in identifying malignant nodules in CT scan images. The use of image preprocessing, segmentation, feature extraction, and classification ensures a comprehensive approach to improving detection accuracy. Deep learning models, particularly Convolutional Neural Networks (CNNs), are effective in recognizing complex patterns and variations in medical imaging, which enhances the identification of cancerous nodules. These models, when trained on large datasets, offer superior performance compared to traditional machine learning methods in terms of accuracy, precision, and the ability to generalize across diverse cases. Despite this, challenges such as overfitting and the need for large, annotated datasets remain. The integration of methods such as data augmentation and transfer learning helps address some of these challenges, enabling models to perform well even with limited labeled data. Future improvements can focus on model interpretability, minimizing false positives, and extending the approach to other types of cancer detection or multi-modal medical imaging. Ultimately, the combination of deep learning with traditional medical imaging has the potential to assist healthcare professionals in early diagnosis, thereby improving patient outcomes.

1. **REFERENCES**

 **Liu, X., et al. (2018).** "Lung cancer detection using deep learning with convolutional neural networks." IEEE Access, 6, 4180-4188.

 **Cheng, J. Z., et al. (2019).** "Computer-aided diagnosis with deep learning architecture: Application to lung cancer detection." Scientific Reports, 9(1), 2980.

 **Shin, H. C., et al. (2016).** "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics, and transfer learning." IEEE Transactions on Medical Imaging, 35(5), 1285-1298.

 **He, K., et al. (2015).** "Deep residual learning for image recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770-778.

 **Esteva, A., et al. (2019).** "A guide to deep learning in healthcare." Nature Medicine, 25(1), 24-29.

 **Xie, Y., et al. (2020).** "Deep learning for lung cancer detection: An overview." Journal of Thoracic Disease, 12(1).

 **Tosun, M. S., et al. (2018).** "Lung cancer detection using deep learning methods." Neurocomputing, 275, 66-77.

 **Dou, Q., et al. (2017).** "Multilevel deep learning for lung cancer nodule classification." IEEE Transactions on Biomedical Engineering, 64(6), 1358-1367.

 **Zhu, W., et al. (2017).** "Lung cancer detection using a convolutional neural network model based on deep learning." *Computers in Biology and Medicine*, 87, 106-116.

 **Chen, M., et al. (2019).** "A novel deep learning-based framework for lung cancer detection in CT images." *Computers, Materials & Continua*, 58(2), 335-346.

 **Hassan, S. S., et al. (2020).** "Lung nodule classification using convolutional neural networks." *Pattern Recognition Letters*, 138, 281-288.

 **Wang, J., et al. (2020).** "Lung cancer detection from CT images using deep convolutional neural networks." *Medical Physics*, 47(3), 1240-1250.

 **Roy, A., et al. (2018).** "Deep learning-based lung cancer classification with a support vector machine classifier." *Healthcare Informatics Research*, 24(2), 151-158.