

Lung cancer prediction

A project Report

By

Ayush Chauhan (E23CSEU0854)

Sanchit Pandey (E23CSEU0725)

Kshitij (E23CSEU0743)

SCHOOL OF COMPUTER SCIENCE ENGINEERING AND TECHNOLOGY, BENNETT UNIVERSITY

GREATER NOIDA, 201310, UTTAR PRADESH, INDIA

April 2025

**DECLARATION**

I/We hereby declare that the work which is being presented in the report entitled “lung cancer preediction “, is an authentic record of my/our own work carried out during the period from JAN, 2023 to April, 2023 at School of Computer Science and Engineering and Technology, Bennett University Greater Noida.

The matters and the results presented in this report has not been submitted by me/us for the award of any other degree elsewhere.

Signature(s) of Candidate(s)

Ayush Chauhan

Sanchit Pandey

Kshitij

Deep Learning Project Report: Lung Cancer Detection

# Introduction

Lung cancer remains one of the leading causes of cancer-related deaths worldwide. Early and accurate detection is critical to improving treatment outcomes and patient survival rates. Traditional diagnostic methods rely heavily on the manual inspection of CT scan images by radiologists, a process that is not only time-consuming but also prone to human error and subjectivity.

This project focuses on automating the classification of lung cancer into three categories — normal, benign tumor, and malignant tumor — using advanced deep learning techniques. A variety of Convolutional Neural Network (CNN) architectures, including a custom deep CNN as well as popular pretrained models such as ResNet50, VGG16, MobileNetV2, InceptionV3, DenseNet121, and Xception, were employed to analyze CT scan images sourced from a Kaggle dataset. By leveraging the powerful feature extraction capabilities of these models, the project aims to build an efficient and accurate system that can assist in the early diagnosis of lung cancer, ultimately contributing to better clinical decision-making and patient care.

# Problem Statement

Manual analysis of lung CT scans can be inefficient, time-consuming, and subject to inconsistencies. This project explores how deep learning techniques can assist in accurately classifying lung tissue into three categories: normal, benign tumor, and malignant tumor. By automating the analysis of CT scan images, the goal is to improve diagnostic efficiency and support early detection of lung cancer.

# Dataset Overview

The dataset used for this project was sourced from Kaggle. It consists of thousands of high-resolution lung CT scan images, categorized into three classes:

* Normal lung tissue
* Benign lung tumors
* Malignant lung tumors

The dataset is organized into separate directories for each category, making it suitable for supervised learning tasks. The CT images vary in size and resolution, requiring preprocessing steps such as resizing and normalization before being fed into the classification models. This diverse and well-labeled dataset provides a strong foundation for training and evaluating deep learning models for lung cancer detection.

# Data Preparation

The dataset was downloaded and extracted in a Google Colab environment. To ensure consistent input for the model, all images were resized. A sample of images from each category was visualized to verify quality and consistency. Labels were assigned based on folder structure, and the dataset was split into training and testing sets using an 80-20 split. Data augmentation techniques such as rotation, zoom, and flip were applied to enhance model generalization and reduce overfitting.

# Model Selection and Architecture

In this project, multiple deep learning models were explored, including pretrained architectures such as VGG16, ResNet50, MobileNetV2, InceptionV3, DenseNet121, and Xception. However, the highest accuracy was achieved using a custom-built Deep Convolutional Neural Network (CNN) specifically designed for this task.

The custom CNN architecture includes:

* **Input Layer**: Preprocessed CT scan images resized to 256×256×3 dimensions.
* **Convolutional Layers**: Multiple Conv2D layers with 128 and 64 filters, using ReLU activation and 'same' padding to preserve spatial dimensions.
* **Pooling Layers**: Alternating Average Pooling and Max Pooling layers to reduce dimensionality and capture important spatial features.
* **Flatten Layer**: To convert feature maps into a one-dimensional feature vector.
* **Fully Connected (Dense) Layers**:
  + A dense layer with 3000 neurons and ReLU activation.
  + A dense layer with 1500 neurons and ReLU activation.
* **Dropout Layer**: A dropout of 20% was applied to reduce overfitting.
* **Output Layer**: A final dense layer with 3 neurons (for normal, benign, malignant classes) and a softmax activation function for multi-class classification.

The model was compiled using the Adam optimizer and trained with the sparse categorical cross-entropy loss function, which is appropriate for multi-class problems where the labels are integers.

Additionally, **real-time data augmentation** was performed using an ImageDataGenerator with transformations such as:

* Random rotations up to 15 degrees
* Horizontal and vertical shifts up to 10%
* Shearing and zooming transformations
* Horizontal flipping
* Filling empty pixels using nearest neighbor interpolation

This augmentation helped improve the model's generalization ability by synthetically increasing the variability of the training set.

The model was trained for 20 epochs with a batch size of 32, using a validation set for monitoring overfitting and model performance.

# Training Process

The model was compiled using the Adam optimizer and the sparse categorical cross-entropy loss function, which is appropriate for multi-class classification with integer labels. Training was carried out over 20 epochs with a batch size of 32, optimized for efficient GPU utilization.

During training, both the training and validation accuracy and loss were continuously monitored to assess the model's learning progress and generalization performance. Data augmentation was employed to reduce the risk of overfitting and to enhance the model's robustness. The learning curves indicated that the model steadily improved in accuracy over time and successfully converged without significant signs of overfitting, achieving strong generalization on the validation set.

# Results and Evaluation

The model achieved high training and validation accuracy, demonstrating its ability to accurately classify lung tissue into the three categories: normal, benign tumor, and malignant tumor. A confusion matrix was generated to visualize the true vs. predicted classifications for each category, providing a detailed view of the model's performance.

Additionally, a classification report was produced, which included key metrics such as precision, recall, and F1-score for each class. These metrics confirmed that the model performed well across all categories, with minimal misclassification. The results showed that the model was able to accurately distinguish between the different types of lung tissue, offering valuable insight into its generalization capability.

# Discussion

The custom Deep CNN model outperformed other architectures, including VGG16 and InceptionV3, demonstrating the value of a tailored approach for this dataset. While other models like ResNet and DenseNet could offer improvements, the current model achieved high accuracy. A major challenge was managing consistent image preprocessing and large CT scan file sizes, which were critical for model performance.

# Future Work

Future work could involve experimenting with different deep learning architectures, applying ensemble methods, and extending the model to multi-class cancer detection across other organs. Additionally, integrating the model with clinical decision support systems could assist pathologists in real-time diagnostics.

# Conclusion

This project demonstrates the effectiveness of deep learning in medical image analysis. By using a custom Deep CNN model on CT scan images, an accurate and reliable lung cancer classification system was developed. Such systems have the potential to enhance diagnostic accuracy and support clinical decision-making, ultimately improving patient care.