Lung cancer prediction

A project Report

By Vansh Banga

(E23CSEU0727)

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 of Candidate

Vansh banga

Deep Learning Project Report: Lung and Colon Cancer Detection

1. Introduction

Lung and colon cancers are among the most common and deadly forms of cancer worldwide. Early detection is crucial for effective treatment and improved survival rates. Traditional diagnosis methods involve the manual inspection of histopathological images by pathologists, which is time-consuming and susceptible to human error. This project aims to automate the detection of lung and colon cancer using deep learning techniques, particularly Convolutional Neural Networks (CNNs), applied to histopathological image data.

2. Problem Statement

Manual analysis of histopathological slides can be inefficient and inconsistent. This project explores how deep learning can assist in accurately classifying lung and colon tissue into cancerous and non-cancerous categories. The focus is on three types of lung tissues: lung adenocarcinomas, lung squamous cell carcinomas, and normal lung tissue.

3. Dataset Overview

The dataset used for this project was sourced from Kaggle. It contains thousands of high-resolution histopathological images categorized into:

- Lung Adenocarcinomas
- Lung Squamous Cell Carcinomas
- Normal Lung Tissue

The data is structured in separate directories for each category. The images vary in size and detail and are used to train a classification model.

4. 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.

5. Model Selection and Architecture

For this project, we used transfer learning with the VGG16 model. VGG16 is a deep convolutional network pre-trained on ImageNet, a large dataset with millions of labeled images. The convolutional base of VGG16 was used to extract features from the histopathological images, and a custom classification head was added with fully connected layers and a softmax output layer.

The model architecture includes:
- Input layer: Preprocessed image input (typically 224x224 pixels)
- Convolutional layers (from VGG16)
- Flatten layer
- Dense layer with ReLU activation
- Dropout layer to prevent overfitting
- Final dense layer with softmax for classification

6. Training Process

The model was compiled with categorical cross-entropy as the loss function and Adam optimizer. Training was performed over several epochs with batch sizes optimized for GPU usage. During training, loss and accuracy were monitored, and model checkpoints were saved. Validation accuracy was also evaluated to assess generalization. Learning curves showed the model steadily improved and converged without significant overfitting.

7. Results and Evaluation

The model achieved high training and validation accuracy. A confusion matrix was plotted to visualize true vs predicted classifications across categories. The classification report included metrics such as precision, recall, and F1-score for each class. Results demonstrated that the model could accurately distinguish between the three types of tissue with minimal misclassification.

8. Discussion

The use of transfer learning significantly boosted performance while reducing training time. While VGG16 performed well, other architectures like ResNet or Inception could be explored for potentially better accuracy. The main challenge was ensuring consistent image preprocessing and managing large image file sizes.

9. Future Work

Future work may include experimenting with other deep learning architectures, applying ensemble methods, or extending the model to handle multi-class cancer detection across

other organs. Integration with clinical decision support systems could also be explored to aid pathologists in real-time diagnostics.

10. Conclusion

This project showcases the effectiveness of deep learning in medical image analysis. By leveraging VGG16 and histopathological images, a reliable and accurate cancer classification model was built. Such systems have the potential to enhance medical diagnostics and improve patient care.