Project Proposal: Classification of Lung Cancer Using Machine Learning

Team Members:

- Abdullah Nadeem
- Usman Akram

1. Introduction / Background

Lung cancer is one of the most common and deadly forms of cancer worldwide. Early and accurate diagnosis is essential for effective treatment, yet manual interpretation of medical images can be time■consuming, subjective, and prone to error.

With advances in machine learning and deep learning, automated image classification systems have emerged as valuable tools to assist radiologists and pathologists. This project proposes a machine learning-based approach to classify lung cancer types using CT and histopathological images, aiming to support clinical decision-making and improve diagnostic efficiency.

2. Objectives

- Develop a machine learning-based system for classifying lung cancer images.
- Preprocess and augment image data to enhance model robustness.
- Implement and train Convolutional Neural Network (CNN) architectures, optionally with transfer learning.
- Evaluate model performance using accuracy, precision, recall, F1-score, and confusion matrix.
- Visualize classification results and highlight key image regions using Grad-CAM.
- Deploy the model through a simple interface for real-time image prediction.

3. Problem Statement

Lung cancer remains a leading cause of cancer-related deaths globally. The manual classification of CT and histopathological images is labor-intensive and subject to human variability, which can delay diagnosis.

The problem is to design and develop an automated classification model that can accurately identify different lung cancer types, improving diagnostic speed and reducing error rates.

4. Dataset

The dataset will be sourced from Kaggle:
https://www.kaggle.com/datasets/programmer3/lung-ct-and-histopathological-images-dataset

It includes labeled CT and histopathological images representing different lung tissue categories and cancer types. The dataset is suitable for training deep learning models for multi-class image classification.

5. Methodology

- **Data Preprocessing**: Resize and normalize images, split into training/validation/testing sets, and apply augmentation (rotation, flipping, zooming).
- **Feature Extraction**: Use CNNs for automatic feature learning. Optionally apply transfer learning with pre-trained models like VGG16, ResNet50, or EfficientNet.
- **Model Training**: Train CNN using categorical cross-entropy loss and Adam optimizer while monitoring validation accuracy.
- **Model Evaluation**: Assess accuracy, precision, recall, F1-score, and confusion matrix. Use Grad-CAM to visualize key regions.
- **Deployment**: Develop a GUI or web interface using Streamlit or Flask to enable image uploads and real-time predictions.

6. Expected Outcomes

We expect to develop a robust, accurate lung cancer classification model capable of distinguishing between various tissue and cancer types. The system will assist medical professionals by improving diagnostic accuracy and reducing interpretation time, ultimately contributing to better patient outcomes.

7. Tools and Technologies

- Programming Language: Python
- Libraries: TensorFlow, Keras, OpenCV, NumPy, Pandas, Matplotlib
- Platform: Jupyter Notebook / Google Colab
- Dataset: Kaggle Lung CT and Histopathological Images Dataset
- GUI Framework: Streamlit or Flask

8. Timeline

- Week 1: Data collection, cleaning, and preprocessing
- Week 2: Model development and transfer learning implementation
- Week 3: Model evaluation, visualization, and performance tuning
- Week 4: GUI development, documentation, and final presentation

9. References

- Kaggle Lung CT Dataset – https://www.kaggle.com/datasets/programmer3/lung-ct-and-histopathological-images-dataset
- TensorFlow Documentation – https://www.tensorflow.org/
- Keras Documentation – https://keras.io/
- OpenCV Documentation – https://docs.opencv.org/