Title: CliniScan-Lung Abnormality Detection Using AI

Intern: Bhumika Warkhade

Table of Contents:

1. Project Description

CliniScan is a deep learning–based project focused on detecting lung abnormalities from chest X-ray images using machine learning and computer vision techniques. The project involves converting DICOM medical images into PNG format, preprocessing the dataset, extracting bounding box annotations, and preparing the data for training an object detection model (YOLO). The workflow includes:

- Collecting and understanding medical imaging data.
- Converting DICOM images to PNG format.
- Exploring the dataset visually and statistically.
- Preprocessing images (normalization, resizing, augmentation).
- Preparing annotations for object detection.
- Future steps include model training, inference, and evaluation.

2. Dataset Used

- The dataset used is the VinBigData Chest X-ray Abnormalities Detection dataset, available on Kaggle.

Dataset Contents:

- 18,000+ DICOM chest X-ray images
- Bounding box annotations for multiple lung abnormalities
- CSV file containing annotations such as:
- image_id
- class_name
- xmin, ymin, xmax, ymax
- The dataset is ideal for training object detection models for medical imaging tasks

3. Environment Setup

Setting up the development environment involves installing necessary tools and libraries to process medical images and train deep learning models.

Required Libraries:

- Python 3.x
- pydicom
- opencv-python
- numpy
- pandas
- matplotlib
- torch, torchvision
- tqdm

Typical Environment (Kaggle/Colab):

pip install pydicom opencv-python numpy pandas matplotlib torch torchvision tqdm

Hardware Used:

GPU accelerator (Kaggle or Google Colab)

5–10 GB runtime storage

4. Data Exploration

- The purpose of data exploration is to understand the data before preprocessing or model training.

Exploration Steps:

- Load annotation CSV and inspect the structure.
- Read sample DICOM images and visualize them.
- Check the distribution of abnormality classes.
- Identify missing or invalid data.

Example Findings:

- Some images contain no abnormalities (No finding).
- Bounding boxes vary in size depending on abnormality.
- Image resolution varies across samples.
- Visualizing the bounding boxes on images helps verify annotation quality.

5. Data Preprocessing

❑ Data preprocessing prepares the raw data for model training.

Steps Performed:

- DICOM → PNG Conversion
- Normalized pixel values from raw medical format to 0–255.
- Resized all images to a fixed size (512✖512).
- Annotation Extraction
- Loaded bounding boxes from CSV.
- Filtered only images with valid annotations.
- Saved a clean parsed_annotations.csv file.
- Image Normalization & Tensor Conversion
- Converted images to grayscale.
- Used PyTorch transforms for resizing and normalization.
- Prepared tensors for training.
- Selection of Subset for Processing
- Limited processing to the first 1000 images to reduce compute time and storage.
- Preprocessing ensures consistency and prepares the dataset for the next steps such as training YOLO or CNN models.

THANK YOU