Lung Disease Prediction Using Kaggle Dataset

1. Objective

The goal of this project is to develop a predictive model for identifying lung disease using a dataset sourced from Kaggle. This dataset provides various lung disease images for 3 classes namely Lung Opacity, Normal and Viral Pneumonia, which are leveraged to create a model that can effectively classify and predict potential lung disease cases based on patient's X-ray image.

2. Dataset Overview

- **Source**: Kaggle - Lung Disease Dataset
- **Features**: Various X-ray images categorized in three classes: Lung Opacity, Normal and Viral Pneumonia
- **Target**: Classify these 3 classes for patient's X-ray images

3. Data Pipeline Design

The pipeline processes the data, analyse images, generates processed images, split data in train and test set and applies a Neural Network classification model.

- **Data Loading & Analysis**:
  - Load data downloaded X-ray images from raw data folder.
  - Analyse image dimension and quality for each class images.
- **DataPreprocessing/Feature Engineering**:
  - **Grayscaling and Resizing**: Each image is converted to grayscale and resized to 224x224 pixels to reduce computational load while retraining essential features
  - **Normalizing and Brightening**: Normalization scales pixel values between 0 and 1, improving training stability. Images are also brightened slightly by fixed factor to enhance visibility of certain features in X-rays.
- **Data Splitting**:
  - Split the dataset into an 80% training set and a 20% test set for model evaluation.

4. Modeling Approach

- **First Model Used**: **CNN** (Convolutional Neural Network)
  - CNN is suitable for predicting multi-class data and is commonly used in image data prediction.
  - **Forecasting**: The model was trained on the training set, and predictions were evaluated for the test period with the help of accuracy score.

- **Second Model Used**: ResNet (Residual Neural Network)
    - ResNet is suitable for predicting multi-class image data and is commonly used in complex image classifications like medical imaging data
    - **Forecasting**: The model was trained on the training set, and predictions were evaluated for the test period with the help of accuracy score.

ResNet model performs better then CNN model because of its deep Neural Network Architecture.

6. Next Steps and Improvements

- Project Expansion e.g. using bounding box annotations to highlight specific portion of the X-ray prone to disease
- CNN and ResNet Model tunning
- Implement other Deep Neural Network models.

7. Conclusion

This project demonstrates a structured approach for unstructured lung disease prediction using classical neural network classification modeling. The pipeline serves as a foundation that can be expanded with data and more sophisticated models and feature engineering for enhanced prediction accuracy.