greatlearning

Learning for Life

INTERIM REPORT

1. Summary of problem statement, data and findings

**Problem Statement:**

Pneumonia is a serious respiratory condition that requires rapid and accurate diagnosis. Manual analysis of chest X-ray images can be time-consuming and error-prone, especially in resource-limited settings. Our objective is to build a deep learning-based system that can **classify chest X-ray images into three categories**:

* **Normal**
* **Lung Opacity (Pneumonia detected)**
* **No Lung Opacity / Not Normal**

**Dataset:**

The dataset used is derived from the **RSNA Pneumonia Detection Challenge**, which includes:

* Chest X-ray DICOM images
* Bounding box annotations for pneumonia regions
* Label metadata indicating the class (Normal, Lung Opacity, or No Lung Opacity)

**Key Findings:**

* Majority of lung opacity cases have clearly distinguishable patterns in the radiographs.
* Normal and “No Lung Opacity / Not Normal” classes are harder to distinguish without localization.
* There is a class imbalance with fewer "Normal" samples compared to "Lung Opacity".

1. Summary of the Approach to EDA and Pre-processing

**Exploratory Data Analysis (EDA):**

* **Image sample visualization:** Random chest X-rays across each class were visualized with bounding boxes (if present).
* **Class distribution:** A pie chart and bar plot confirmed **imbalance**, with “Lung Opacity” being the most frequent class.
* **Bounding box analysis:** For Lung Opacity images, multiple bounding boxes are often present, indicating severity.
* **Size distribution of bounding boxes:** Heatmaps showed that pneumonia regions are generally centralized in the lungs.

**Pre-processing Steps:**

* **DICOM conversion:** DICOM images were converted to pixel arrays using pydicom.
* **Resizing:** Images resized to 224x224 for uniform input to the CNN model.
* **Normalization:** Applied per-channel mean and standard deviation normalization.
* **Augmentation:** To combat class imbalance and improve generalization, applied transformations like:
  + Random horizontal flips
  + Rotation ±10°
  + Zoom and cropping

**Meaningful Features Identified:**

* Pneumonia presence is visually detectable in dense and irregular lung patterns.
* In “Lung Opacity” images, lesion density and location in lungs (mostly lower lobes) are common.
* Heatmap visualizations (Grad-CAM) show model focuses on actual infected regions when correctly classified.

1. Deciding Models and Model Building

Based on the nature of the problem, decide what algorithms will be suitable and why?

Experiment with different algorithms and get the performance of each algorithm.

1. How to improve your model performance?

What are the approaches you can take to improve your model? Can you do some feature se. ection, data manipulation and model improvements.

Provide your code and as much as visualizations you can share to describe what you have done so far.