Task 2: Medical Imaging Dataset Exploration

Dataset 1: COVID-19 Radiography Database (Small Medical Imaging Dataset)

Link: [Dataset1]

Dataset Overview

Type of Imaging Data: Chest X-ray images

Medical Use Case: Detection of COVID-19 and related lung infections

Image Format: PNG

Image Size: Mostly standardized (~299×299 pixels)

Dataset Composition:

Class	Approx. Images
COVID-19	360
Normal	400
Viral Pneumonia	400
Total	**1,200**

Classes / Labels:

- **COVID-19** – X-rays of patients infected with SARS-CoV-2
- **Normal** – Healthy chest X-rays
- Viral Pneumonia – Non-COVID viral lung infections

Dataset Imbalance:

- Mild class imbalance exists
- The COVID-19 class has **fewer images** compared to others

- Imbalance is manageable but may affect model bias

Challenges Observed:

1) **Image Quality Variation**

- Different hospitals and machines
- Varying contrast and noise levels

2) **Visual Similarity**

- COVID-19 and viral pneumonia X-rays often look similar

3) **Limited Dataset Size**

- Risk of overfitting in deep learning models

4) **Annotation Reliability**

- Labels based on clinical diagnosis, not always radiologist consensus

Summary

- The COVID-19 Radiography Database is a small-scale chest X-ray dataset designed for rapid exploration and classification of COVID-19 and related lung diseases. It contains approximately 1,200 images across three classes: COVID-19, normal, and viral pneumonia. While the dataset is well-structured and suitable for quick experimentation, challenges such as limited size, mild class imbalance, and visual similarity between disease classes pose difficulties for accurate classification.

Dataset 2: HAM10000 Skin Lesion Dataset (Reduced Size)

Link: Dataset2

Details

- **Type:** Skin lesion (dermoscopic images)
- **Total images:** ~10,000
- **Classes:** 7 skin lesion categories
- **Label type:** Multi-class classification
- **Imbalance:** Highly imbalanced (one class dominates)

Challenges

- Severe class imbalance
- Visual similarity between classes
- Annotation difficulty

Summary

The HAM10000 dataset consists of dermoscopic images of skin lesions categorized into seven classes. While it is suitable for studying multi-class classification, challenges such as class imbalance and visual similarity between lesions make accurate classification difficult.