# Week 1: Input Data Documentation

## Dataset I'm Using

\*\*Dataset Name:\*\* COVID-19 Chest X-ray Dataset

\*\*Where to get it:\*\* https://github.com/ieee8023/covid-chestxray-dataset

#### What it contains:

* About 930 chest X-ray images
* From multiple medical institutions globally
* Both PA and AP view chest X-rays
* All images are in PNG/JPG format
* Comes with a CSV metadata file that has labels and patient info

## About the Dataset

The images are chest X-rays from patients with COVID-19, pneumonia, and other respiratory conditions. Each image has metadata including the diagnosis, patient age, sex, and other clinical information.

#### What makes this good for my project:

* Small size - I can actually download and work with this locally
* Has COVID-19 labels which is perfect for my research paper
* Other researchers have used it so I know it works
* Free to download from GitHub
* Good variety of cases for training

#### Potential issues:

* Smaller dataset means less training data
* Images from different hospitals might have different quality/equipment
* Limited to certain disease types

## How I'll Use It

I'm planning to focus on binary classification: COVID-19 vs Pneumonia. This gives me a manageable problem to work on with the Vision Transformer architecture.

#### My Plan

I'm planning to use a subset of this data since 112k images is probably too much for me to handle with my computer. I'll probably take around 10,000 normal images and 10,000 with diseases to make it balanced.

### Getting the Data

\*\*How to Get It:\*\* Free to download - no sign up needed

#### Download Info:

* \*\*Link:\*\* https://nihcc.app.box.com/v/ChestXray-NIHCC
* \*\*Files:\*\* .tar.gz files (12 parts) + CSV files with labels
* \*\*Info Page:\*\* https://www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available-chest-x-ray-datasets-scientific-community

### How I'll Split the Data

#### What the Original Paper Did:

* Training: 70% - 78,468 images
* Validation: 15% - 16,818 images
* Testing: 15% - 16,834 images

#### What I Plan to Do:

* Training: 70% - for teaching the model (78,484 images)
* Test: 20% - for checking how good it is (22,424 images)
* Validation: 10% - final check (11,212 images)

### Getting the Data Ready

#### What I Need to Do:

1. \*\*Resize images:\*\* Change from 1024x1024 to 224x224 (what ViT needs)

2. \*\*Normalize:\*\* Make pixel values between 0 and 1, use ImageNet standards

3. \*\*Data tricks:\*\* Random flips, small rotations, brightness changes

4. \*\*Fix labels:\*\* Change from multiple labels to just Normal vs. Disease

5. \*\*Split carefully:\*\* Make sure same patient isn't in training and test sets

#### What I'll Use:

* \*\*pandas\*\* for working with the CSV files
* \*\*PIL/torchvision\*\* for loading and changing images
* \*\*PyTorch\*\* for loading data
* \*\*scikit-learn\*\* for splitting data properly
* \*\*numpy\*\* for math stuff

### Other Datasets I Might Use

#### Extra Dataset 1:

* \*\*Name:\*\* COVID-19 Chest X-ray Dataset
* \*\*Why:\*\* More COVID-19 cases (NIH dataset doesn't have many)
* \*\*Where:\*\* https://github.com/ieee8023/covid-chestxray-dataset
* \*\*Cost:\*\* Free on GitHub

#### Extra Dataset 2:

* \*\*Name:\*\* RSNA Pneumonia Detection Challenge Dataset
* \*\*Why:\*\* More pneumonia cases for better training
* \*\*Where:\*\* https://www.kaggle.com/c/rsna-pneumonia-detection-challenge
* \*\*Cost:\*\* Free download from Kaggle

### Rules and Ethics

\*\*License:\*\* NIH dataset is public domain (anyone can use it)

#### Rules:

* Can't sell it or make money from it directly
* Have to say it came from NIH Clinical Center
* Can't try to figure out who the patients are (already anonymous)

#### Citation:

I have to cite: Xiaosong Wang et al. "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks for Weakly-Supervised Classification and Localization of Common Thorax Diseases." IEEE CVPR 2017

#### Privacy:

* All patient names/IDs already removed
* People over 89 are just listed as 90 to keep them anonymous
* No privacy problems since this is already public medical research data

### My Timeline

#### Getting Data:

* Week 1: Download NIH dataset (45GB - might take 6-12 hours)
* Week 1-2: Check that everything downloaded right and look at the data
* Week 2: Write code to process the images
* Week 2: Split into train/test/validation sets (making sure same patient isn't in multiple sets)

#### Checking Everything Works:

1. Dataset downloads correctly (NIH website works)

2. Data actually downloaded (45GB downloaded successfully)

3. Format is right (PNG images + CSV with labels)

4. Right number of images (112,120 images)

5. Numbers match what the paper says

6. My processing code works

4. \*\*Email authors:\*\* Ask paper authors for help with dataset if needed

---

## Links and Files

### What I Have:

1. \*\*Dataset Link:\*\* https://nihcc.app.box.com/v/ChestXray-NIHCC

2. \*\*More Info:\*\* https://www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available-chest-x-ray-datasets-scientific-community

3. \*\*Kaggle Version:\*\* https://www.kaggle.com/nih-chest-xrays/data

4. \*\*Original Paper:\*\* https://arxiv.org/abs/1705.02315

### Status:

* Found main dataset and I can download it
* Found backup datasets (Kaggle, CheXpert)
* Know how to download (direct download, no sign up needed)
* Know what I need to do to the images (resize, normalize, augment)
* Know what tools to use (PyTorch, torchvision, pandas)
* Know how to split the data (70/20/10 with patients separated properly)