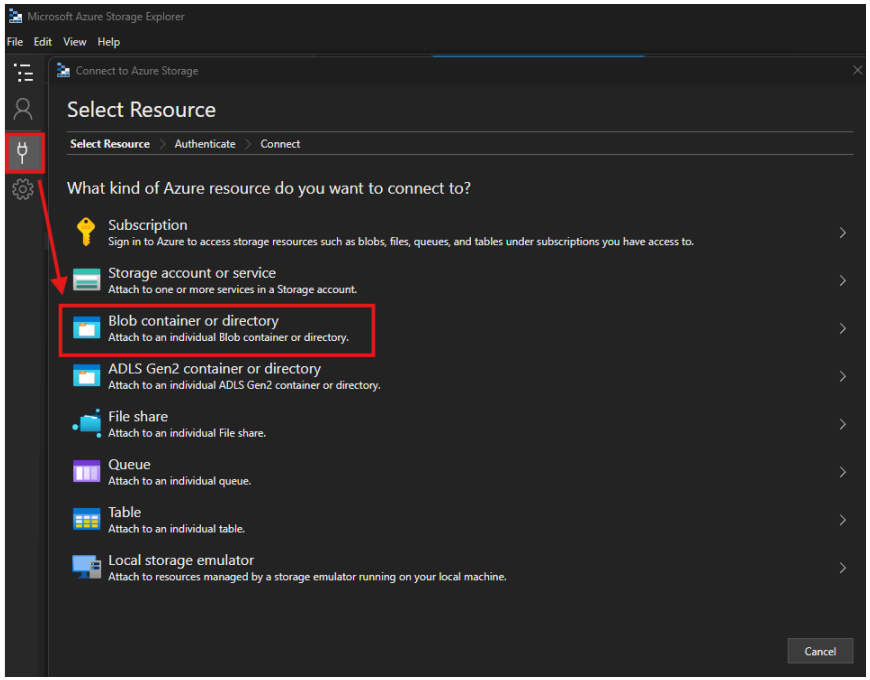
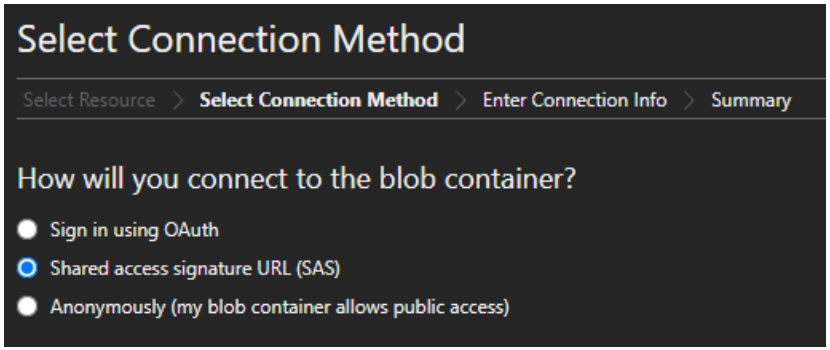
Accessing the Blob container on Microsoft Azure Storage Explorer

To use the CheXpert files, request access using this link:
https://stanfordaimi.azurewebsites.net/datasets/8cbd9ed4-2eb9-4565-affc-111cf4f7ebe2

Then, a link to the blob will be provided which can be accessed using some Microsoft Azure application. We used Microsoft Azure Storage Explorer which provides a useful UI for navigation purposes.

Open the connect dialog box, then select “Blob container or directory.” On the following page, select “Shared access signature URL (SAS)”





On the following page, paste the link to the blob in the “Blob container or directory SAS URL” field, and the “Display name” field will auto populate. Then, click Next. If everything was done correctly, the files should appear in the Storage Explorer UI

Explanation of Files Available on the Blob

Name	Content Type	Size	Last Modified	Parent Directory	Description	Required
CHEXPERT DEMO.xlsx	Excel Spreadsheet	1.94 MB	8/9/2021	-	Contains limited demographic information for each patient in the dataset	**NO**
CheXpert-v1.0 batch 1 (validate & csv).zip	Compressed Zip Folder	486.04 MB	12/30/2023	-	Contains the test set of images and their corresponding labels. This folder also contains valid.csv Despite the name, this is the **test** set, and will be adjusted to reflect this fact during preprocessing	**YES**
valid.csv	Comma Separated Values	31 KB	1/20/2019	CheXpert-v1.0 batch 1 (validate & csv).zip	Contains the **test** labels, and will be adjusted to reflect this fact during preprocessing	**YES**
CheXpert-v1.0 batch 2 (train 1).zip	Compressed Zip Folder	162.39 GB	12/30/2023	-	Contains the first set of training images	**NO**
CheXpert-v1.0 batch 3 (train 2).zip	Compressed Zip Folder	184.82 GB	12/30/2023	-	Contains the second set of training images	**NO**
CheXpert-v1.0 batch 4 (train 3).zip	Compressed Zip Folder	91.09 GB	12/30/2023	-	Contains the third set of training images	**YES**
README.md	Plain-text	3.21 KB	8/9/2021	-	A plain-text file which documents the dataset, the images, and the labeling	**YES**
train_cheXbert.csv	Comma Separated Values	22.06 MB	8/9/2021	-	The training labels produced by the CheXbert labeler which utilizes both a rules-based labeler and a BERT model	**YES**
train_visualCheXbert.csv	Comma Separated Values	28.48 MB	8/9/2021	-	The training labels produced by the VisualCheXbert labeler which combines CheXbert with a CNN computer vision model	**NO**

Starting Schema configuration

After downloading and extracting the training and test data locally, they must be organized into the following directory format in order for the labels to be located properly.

| **[ROOT_DIR]**
|-----| **[MAIN_DIR]**
|----------| **[TRAIN_DIR]**
|----------| **[TEST_DIR]**
|--------------| **valid**
|--------------| valid.csv
|----------| README.md
|----------| train_cheXbert.csv

ROOT_DIR : The root directory path containing the input and output directories
MAIN_DIR: The name of the directory containing the uncompressed files as downloaded from Microsoft Azure Storage Explorer
TRAIN_DIR: The name of the directory containing the extracted training x-ray images (**CheXpert-v1.0 batch 1 (validate & csv)**)
TEST_DIR: The name of the directory containing the extracted test files (**CheXpert-v1.0 batch 1 (validate & csv)**). This folder contains the test labels (**valid.csv**) and a subdirectory (**valid**) containing the test x-ray images.

Source Schema Configuration

|-----| **train**: Contains the individual patient subfolders from batch 3
|----------| Patient ID
|--------------| Study ID
|--------------------| Frontal View / Lateral View
|-----| **test**: Contains the individual patient subfolders from batch 1
|----------| Patient ID
|--------------| Study ID
|--------------------| Frontal View / Lateral View

Data Definitions

Data Transformation

Step	Description	Input Signature	Relevant Prior Step	Output Signature	Tool/Library Required
1	Download, extract, and organize the required files into the required directory structure	-	Reference pages 1-3	-	File Explorer, Microsoft Azure Storage Explorer
	Rename all relevant “validation” names to “test.” This includes file names during step 1 and the image file paths in the training and test datasets.		1		File Explorer, Pandas
	As we are using a subset of the data, filter the training and validation datasets to only contain valid image file paths	Data Frame	-	Data Frame	Pandas
	Insert a column to the training and validation datasets that contain the preprocessed image file paths	Data Frame	3		
	Resize all images to the size (512x512)				

train_visualCheXbert.csv

This contains the file path to each x-ray image along with corresponding features that indicate the presence (or lack thereof) of 14 pathological conditions. It also contains some limited demographic information regarding the patient as well as the configuration parameters of the x-ray itself.

Field Name	Data Type	Field Size	Description	Field Type
Path	TEXT	58	The file path leading to the x-ray image.	ID
Sex	TEXT	7	The gender of the patient.	Demographic
Age	INT64	3	The age of the patient when the study was performed.	Demographic
Frontal/Lateral	TEXT	7	This indicates whether the x-ray view was taken from a frontal position, or a lateral position	X-ray configuration
AP/PA	TEXT	3	X-ray code which indicates the view angle Posteroanterior (PA) refers to when the x-ray beam passes through the patient from the front to the back. This is the most common view, though in this dataset, it is the minority. Anteroposterior (AP) refers to when the x-ray beam passes through the patient from the back to the front. This is typically used when the patient is not well enough to get into the PA position. Lower Lumbar (LL) refers to views of the lumbar spine and sacrum area RL refers to views taken from either the right (R) or left (L) sides of the body	X-ray configuration

Enlarged Cardiomediastinum	FLOAT64	3	This indicates if the cavity containing the heart and other structures is enlarged.	Pathological condition
Cardiomegaly	FLOAT64	3	This indicates the presence of an enlarged heart.	Pathological condition
Lung Opacity	FLOAT64	3	This indicates the presence of hazy, dense areas in the lung that should be darker.	Pathological condition
Lung Lesion	FLOAT64	3	This indicates an abnormal growth in the lung tissue.	Pathological condition
Edema	FLOAT64	3	This indicates the presence of fluid collection in the air sacs	Pathological condition
Consolidation	FLOAT64	3	This indicates when the air within small airways of the lungs is replaced with a fluid, solid, or other material (such as pus, blood, water, etc.).	Pathological condition
Pneumonia	FLOAT64	3	This indicates the presence of a lung infection causing inflammation and fluid buildup in the lungs.	Pathological condition
Atelectasis	FLOAT64	3	This indicates the presence of the collapse of a lung (or part of a lung) which occurs when the alveoli lose air and deflate.	Pathological condition
Pneumothorax	FLOAT64	3	This indicates the presence of air leakage from the lungs into the space between the chest wall and the lungs. This can cause the lungs to collapse.	Pathological condition
Pleural Effusion	FLOAT64	3	This indicates the presence of a collection of fluid in the space between the chest wall and the lungs	Pathological condition
Pleural Other	FLOAT64	3	This indicates the presence of some condition affecting the pleura (The membrane lining the chest wall and lungs)	Pathological condition
Fracture	FLOAT64	3	This indicates the presence of a break in a bone.	Pathological condition
Support Devices	FLOAT64	3	This indicates the presence of support devices in the image.	Pathological condition
No Finding	FLOAT64	3	This indicates the absence of all above pathologies	Pathological condition