Team Cancer

https://github.com/CMU-IDS-Spring-2024/final-project-cancer

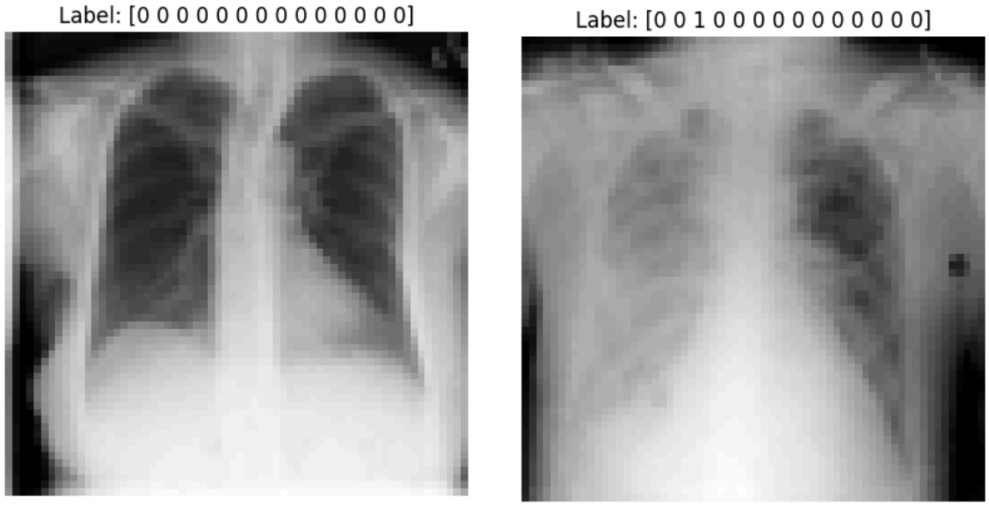
Team members+Problem motivation :

J-how, Parth, Joong Ho (Joshua) and Varun are all first-year MCDS students who are interested in using data science for improving the interpretability of AI in the medical field.

Dataset
We will be exploring ChestMNIST under MedMNIST. ChestMNIST contains 112,120 frontal-view X-Ray images of 30,805 unique patients with the text-mined 14 disease labels, which could be formalized as a multi-label binary-class classification task. The dataset consists of 78,468 training points, 11,219 validation points and 22,433 test points).

We are able to access the dataset by

```
pip install medmnist
import medmnist
from medmnist import ChestMNIST
train_dataset = ChestMNIST(split="train", download=True, size=224)
val_dataset = ChestMNIST(split="val", download=True, size=224)
test_dataset = ChestMNIST(split="test", download=True, size=224)
```



#The images are of PIL.Image.Image data type.

The dataset's distribution is as shown below

labels	('Train', 'obs')	('Train', 'freq')	('Test', 'obs')	('Test', 'freq')
No Finding	50500	0.483392	9861	0.266032
Infiltration	13782	0.131923	6112	0.164891
Effusion	8659	0.082885	4658	0.125664
Atelectasis	8280	0.0792572	3279	0.0884614
Nodule	4708	0.0450656	1623	0.0437856
Mass	4034	0.038614	1748	0.0471578
Consolidation	2852	0.0272997	1815	0.0489654
Pneumothorax	2637	0.0252417	2665	0.0718968
Pleural_Thickening	2242	0.0214607	1143	0.0308361
Cardiomegaly	1707	0.0163396	1069	0.0288397
Emphysema	1423	0.0136211	1093	0.0294871
Edema	1378	0.0131904	925	0.0249548
Fibrosis	1251	0.0119747	435	0.0117355
Pneumonia	876	0.00838518	555	0.0149729
Hernia	141	0.00134967	86	0.00232012

To augment our data, we are considering other potential data sources that could augment our main data. They are
MedMNIST (ChestX-ray 14): https://nihcc.app.box.com/v/ChestXray-NIHCC (huggingface)
CheXpert: https://stanfordmlgroup.github.io/competitions/chexpert/
MIMIC-CXR: https://physionet.org/content/mimic-cxr/2.0.0/
Vindr-CXR: https://github.com/vinbigdata-medical/vindr-cxr
PadCHEST: https://github.com/auriml/Rx-thorax-automatic-captioning

These five potential supplementary datasets cover different imaging modalities. Thus, iIntegrating data from these sources could potentially enhance the diversity and generalization capabilities of our model and our findings later on.

Research Question:

We have a two major data science pursuits in this project :

- First, we try to find the correlation between different types of lung diseases present in the dataset. We do this by exploratory data analysis on the image classes and different clustering algorithms applied on the images.

- Secondly, we will try to show if there are any specific image patterns that can be used to detect the disease the person has been facing.

Here are some steps we plan to perform:

Data Exploration: To uncover correlation patterns in the dataset we will do exploratory data analysis. We will use the multilabel classes as features and try to find which sort of diseases occur together and try to generate visualizations to view the patterns.

Clustering: To find similarities between images of different diseases, we will try to use different clustering techniques to help identify the cluster center for different disease groups and compare how similar they are to figure out if there are any similar features that lead to different diseases.

Model Creation: To try to classify the different types of disease images we will create a model involving CNNs which will output the probabilities of each disease for every image provided.

Interpretability Visualizations: In this section we will try to show the interpretability of the CNN model that we create by trying to visualize what sort of features different filters are focussing on.