Data

The majority of the datasets utilized in this study are openly accessible for both training and validation purposes:

1. DeepLesion (https://nihcc.app.box.com/v/DeepLesion/): DeepLesion is a dataset comprising 32,735 lesions from 10,594 studies of 4,427 unique patients collected over two decades from the National Institute of Health Clinical Center PACS server. Various lesions, including kidney, bone, and liver lesions - as well as enlarged lymph nodes and lung nodules, are annotated. The lesions are identified through radiologist-bookmarked RECIST diameters across 32,120 CT slice . In our study we use this dataset both for our pre-training and use-case 1

2. LUNA16 (https://luna16.grand-challenge.org/Data/): LUNA16 is a curated version of the LIDC-IDRI dataset of 888 diagnostic and lung cancer screening thoracic CT scans obtained from seven academic centers and eight medical imaging companies comprising 1,186 nodules. The nodules are accompanied by annotations agreed upon by at least 3 out of 4 radiologists. Alongside nodule location annotations, radiologists also noted various observed attributes like internal composition, calcification, malignancy, suspiciousness, and more. We use this dataset to develop and validate our diagnostic image biomarker

3. LUNG1 (https://wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics): LUNG1 is a cohort of 422 patients with stage I-IIIB NSCLC treated with radiation therapy at MAASTRO Clinic, Maastricht, The Netherlands. FDG PET-CT scans were acquired with or without contrast on the Siemens Biograph Scanner. Radiation oncologists used PET and CT images to delineate the gross tumor volume. Our prognostic image biomarker is validated using this cohort.

4. RADIO (https://wiki.cancerimagingarchive.net/display/Public/NSCLC+Radiogenomics): RADIO (NSCLC-Radiogenomics)
   dataset is a collection of 211 NSCLC stage I-IV patients who were referred for surgical treatment and underwent preoperative CT and PET/CT scans. These patients were recruited from the Stanford University School of Medicine and the Palo Alto Veterans Affairs Healthcare System. Scan scans were obtained using various scanners and protocols depending on the institution and physician. A subset of 144 patients in the cohort has available tumor segmentations independently reviewed by two thoracic radiologists. In addition to imaging data, the dataset includes molecular data from EGFR, KRAS, ALK mutational testing, gene expression microarrays, and RNA sequencing. We use this dataset for validation the performance of our prognostic biomarker and also for our biological analysis.

Note: The training dataset for our prognostic biomarker model, HarvardRT, is internal and unavailable to the public. Nonetheless, our foundational model can be publicly accessed, and the results reproduced using the accessible test datasets.

Along with this codebase, we provide scripts and code to easily download and pre-process this dataset to encourage reproduction of our study. This README contains detailed information on how to do so.

Downloading the datasets

The DeepLesion and LUNA16 dataset can be downloading using download scripts provided in `data/download`. Note that these scripts are provided for a linux environment but can be simply adapted to other environments by downloading equivalent packages. For the LUNG1 and RADIO datasets, we provide an end-to-end way to download and process them using the Imaging Data Commons infrastructure through Google Colab notebooks. This can also be run locally by downloading the google colab jupyter notebook and simply running it on your local machine.

Downloading the DeepLesion and LUNA16 datasets

Note: Conda is required to perform this. Make sure you have enough space in the location you are downloading (~250GB for DeepLesion, ~60GB for LUNA16)

For DeepLesion,

```
cd data/download
bash deeplesion.sh <path_to_download>
```

For LUNA16,

```
cd data/download
bash luna16.sh <path_to_download>
```

Downloading the LUNG1 and RADIO datasets

Refer to Google colab notebook that contains information on how this data can be downloaded and pre-processed.

Preprocessing the datasets

We provide simple linux shell scripts to reproduce the pre-processing pipeline. Incase, you have a different operating system, simply run the python scripts in these shell scripts individually in your environment.

Pre-processing the DeepLesion dataset

Note: Conda is required to perform this

```
cd data/preprocessing/deeplesion
bash run.sh <path_to_download>
```

Once you run this successfully, you should see a file `data/processing/deeplesion/annotations/deeplesion_training_annotations.csv`. At this point you can run the notebook `data/processing/deeplesion/Process_Dataset.ipynb` to get the splits we use in our paper. For reference, we have already provided splits for comparison as generated by us.

NOTE: The pre-processing extracts the image files from zip files. Please delete the zip files from the path `<path_to_download>/DeepLesion/Images_png` using `rm <path_to_download>/DeepLesion/Images_png/*.zip` path after these scripts are successfully run to not inflate your disk space.

Pre-processing the LUNA16 dataset

```
cd data/preprocessing/luna16
bash run.sh <path_to_download>
```

NOTE: The pre-processing extracts the image files from zip files. Please delete the zip files from the path `<path_to_download>/LUNA16` using `rm <path_to_download>/LUNA16/*.zip` after these scripts are successfully run to not inflate your disk space.

Once you run this successfully, you should see a file `data/processing/deeplesion/annotations/luna16_training_annotations.csv`. At this point you can run the notebook `data/processing/deeplesion/Process_Dataset.ipynb` to get the splits we use in our paper. For reference, we have already provided splits for comparison as generated by us.

Pre-processing the LUNG1 and RADIO dataset

Refer to our Google Colab reproducible notebooks that also contains information on how this data can be accessed