Read Me

This is the project data folder for ‘Predicting HT using machine learning’.

.csv files:

- For each pixel, we extracted its 3x3 area information including DWIb0, DWIb1000, PWI and AIF
- DWI is a single value. PWI and AIF are time-series data sampled at 60 time slots with 1.8s interval. Each pixel has unique PWI. All pixels of one patient have the same AIF.
- In .csv files, there are usually 623 columns:
    - First column indicates patient ID
    - Column 2,3,4 are X,Y,Z locations
    - The following columns are:
        - [9 DWIb1000, 9 DWIb0, 9*60 PWI, 60AIF, Ground Truth](Totally 9+9+540+60+1=619 columns]
- The number of samples in .csv file is shown in the name of file. (eg: train_data_final_50k means 50k samples). The number of bleeding/non-bleeding points is balanced, but the number of samples from each patient is not balanced.
- Keyfactors_*.csv is patients’ clinical information which may be related to HT
- We selected patients from both Chinese and UCLA hospitals. Patient IDs are
    - Chinese patients
        - HT: 9,18,26,34,67,97,115,120,148,154,155,156,168,173,185,189,199,210,212,219,240,283,288,307,356,367,370,371,390,392,406,414,424,448,449,511,525,528,538,628,913
        - NonHT: 7,63,69,77,81,110,133,177,201,206,249,258,312,324,334,353,361,387,477,509,619
    - UCLA patients(HT patients only)
        - HT: 163610,338591,470495,602825,621613,733654,780466,824089,1480101,1617232,2434843,3034158,3271602,3351938,3765588,3857869, 4187978,4233872,4275598,4368582,4379308,4387248,4421302, 4589427,4603399,4647754,4710678,4993051,4993368,5061045, 5067760,5071585,5161120,4696639,4462962,4395400,4332634, 3500907,2368677,1994466,1676642,919386,4428613,4360234,2896041, 2359431,1154474
- 50k data only includes Chinese patients. 80k and 100k data only include HT patients from both China and UCLA.
- ‘test_data’ contains the data used to generate visualizations for single slices. Take ‘115-test-slice34.csv’ as example:

    - It is used to predict slice 34 of patient 115
    - It has 256*256 rows. (each represents one location in the 256*256 image)
    - It has 618 columns. (9+9+540+60)
- For convenience, the ground truth for all test slices is extracted in ‘label’
- Notice: The order you read DCM with MATLAB/Python/… can be different. The coordinates in csv are read by MATLAB function. They are not original coordinates in DCM. Based on the function you use, you should check how coordinates are transformed.
- Attention: For UCLA patients, the DWI b0 and b1000 are reversed!! Be sure to correct it when you sample data