Supplementary: Detecting noisy labels with repeated cross-validations

Jianan Chen, Vishwesh Ramanathan, Tony Xu, and Anne L. Martel

Department of Medical Biophysics, University of Toronto, Toronto, ON, CA
Sunnybrook Research Institute, Toronto, ON, CA

Algorithm 1: Pseudocode of ReCoV in a Python-like style

```
# N_runs: number of runs
# k: number of folds

seeds = GenerateRandomNumbers(N_runs) # generate N_runs seeds
candidates = [] # initialize candidates as an empty list

for seed in seeds: # repeat for N_runs of different seeds
    train_sets, val_sets = FoldSplit(data, k, seed) # k-fold split
    models.train(train_sets) # train k models with k train_sets
    val_metrics = models.test(val_sets) # evaluate trained models
    worst_set = val_sets[Argmin(val_metrics)] # find the worst fold
    candidates.append(worst_set.ids) # add 'worst' ids to candidates

# calculate number of occurrences for each sample
samples, counts = Unique(candidates)
```

Table 1. Hyperparameters used in fastReCoV experiments. Thresholds are either absolute values or percentiles. For CIFAR-10N we choose probability of the image belonging to the given ground true label as the ranking metric. For HECKTOR, we created our own ranking metric inspired from c-index. For a particular sample, we evaluated its concordance with all the other samples both within and across the folds. For PANDA, we used the absolute distance between the ground truth label and predicted label.

Dataset	**CIFAR-10N**	**HECKTOR**	**PANDA**
Sample-level metric	predicted probability	sample-level concordance	regression distance
Threshold T	0.3	4%	10%
N_runs	10	50	15
Temperature τ	0.1	0.5	1.0
Drop rate β	0.8	0.1	0.5
EMA weight α	0.3	0.3	0.3

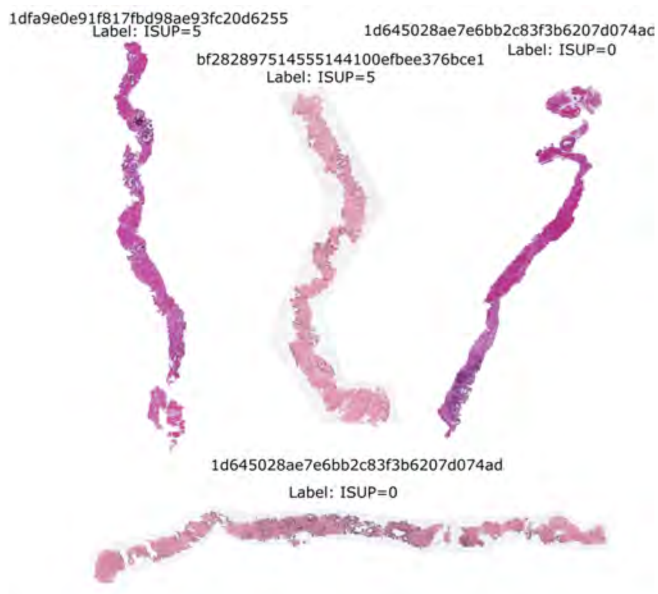

Fig. 1. PANDA samples that are predicted to have highest chance of being noisy in ISUP grade 5 and benign. Sample IDs and original labels are attached above the images.