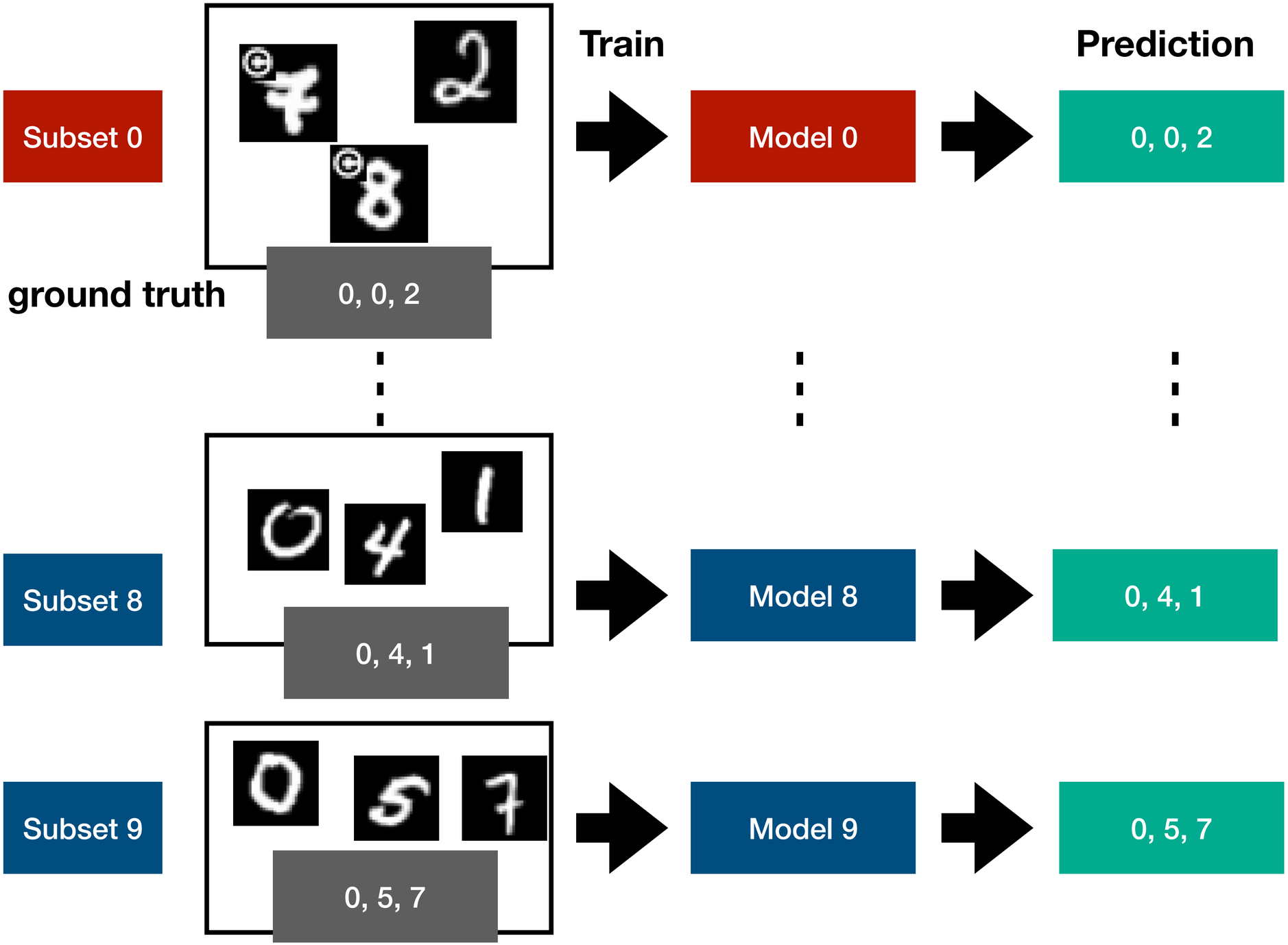
Detection of Backdoor Sources with Cross-Validating Ensembles

CC CHANG

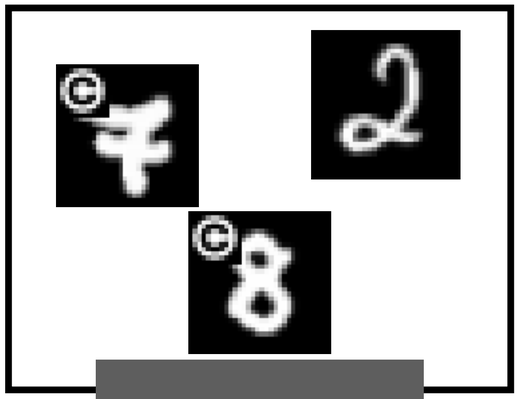
10 Feb 2025

Training weak models



Test weak models

Subset 0



0, 0, 2

Test

Model 1 → 7, 8, 2

Model 2 → 7,8, 2

- - -

Model 9 → 7, 8, 2

Subset 8

0, 4, 1

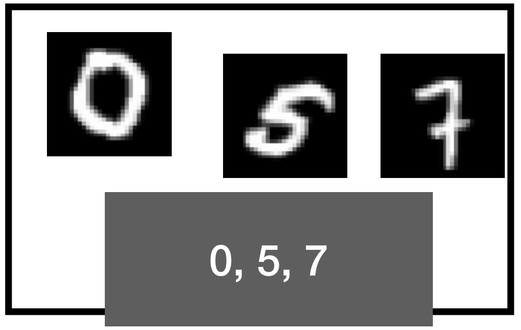
Model 0 → 0, 4, 1

Model 1 → 0, 4, 1

- - -

Model 9 → 0, 4, 1

Subset 9



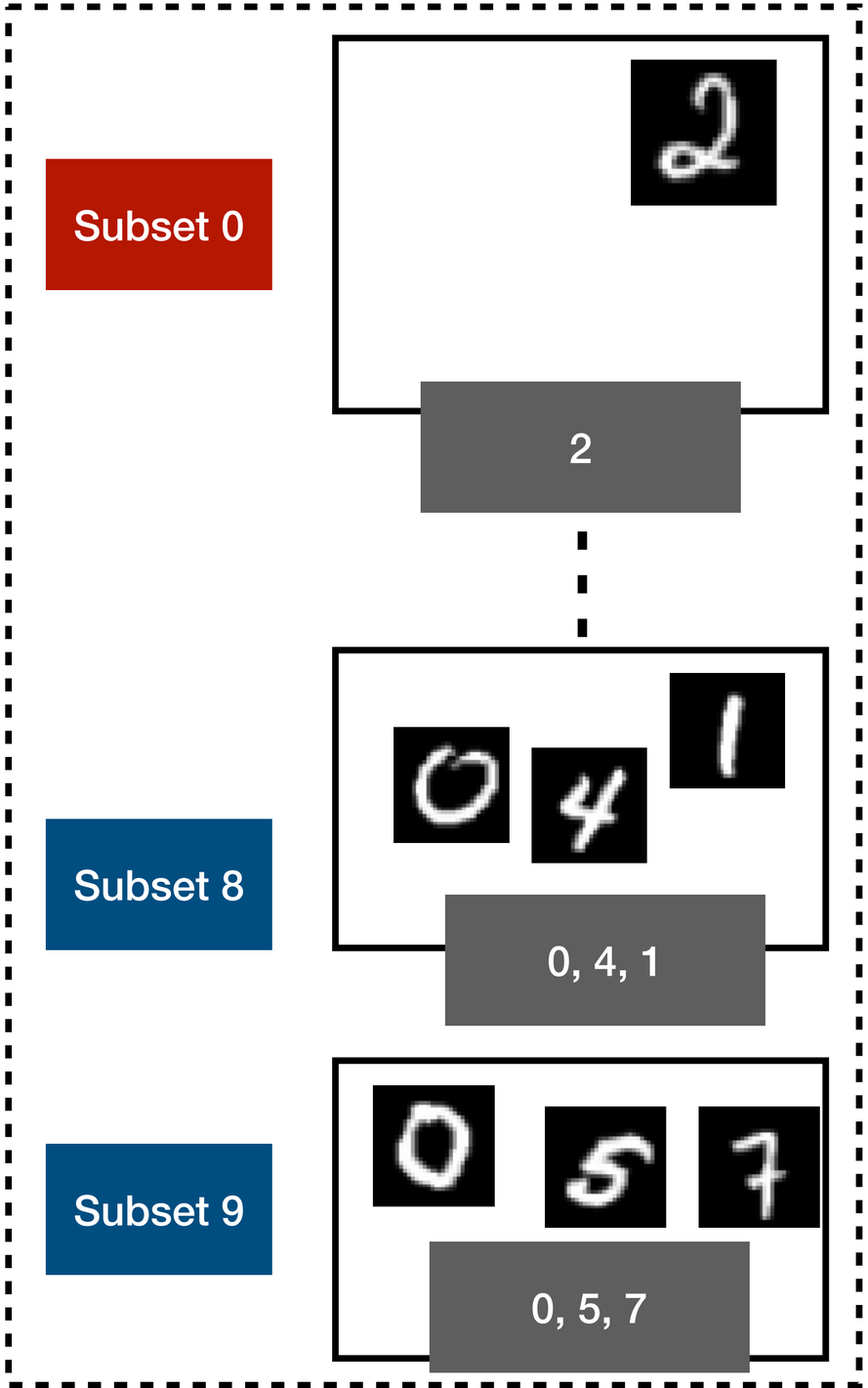
0, 5, 7

Model 0 → 0, 5, 7

Model 1 → 0, 5, 7

- - -

Model 8 → 0, 5, 7

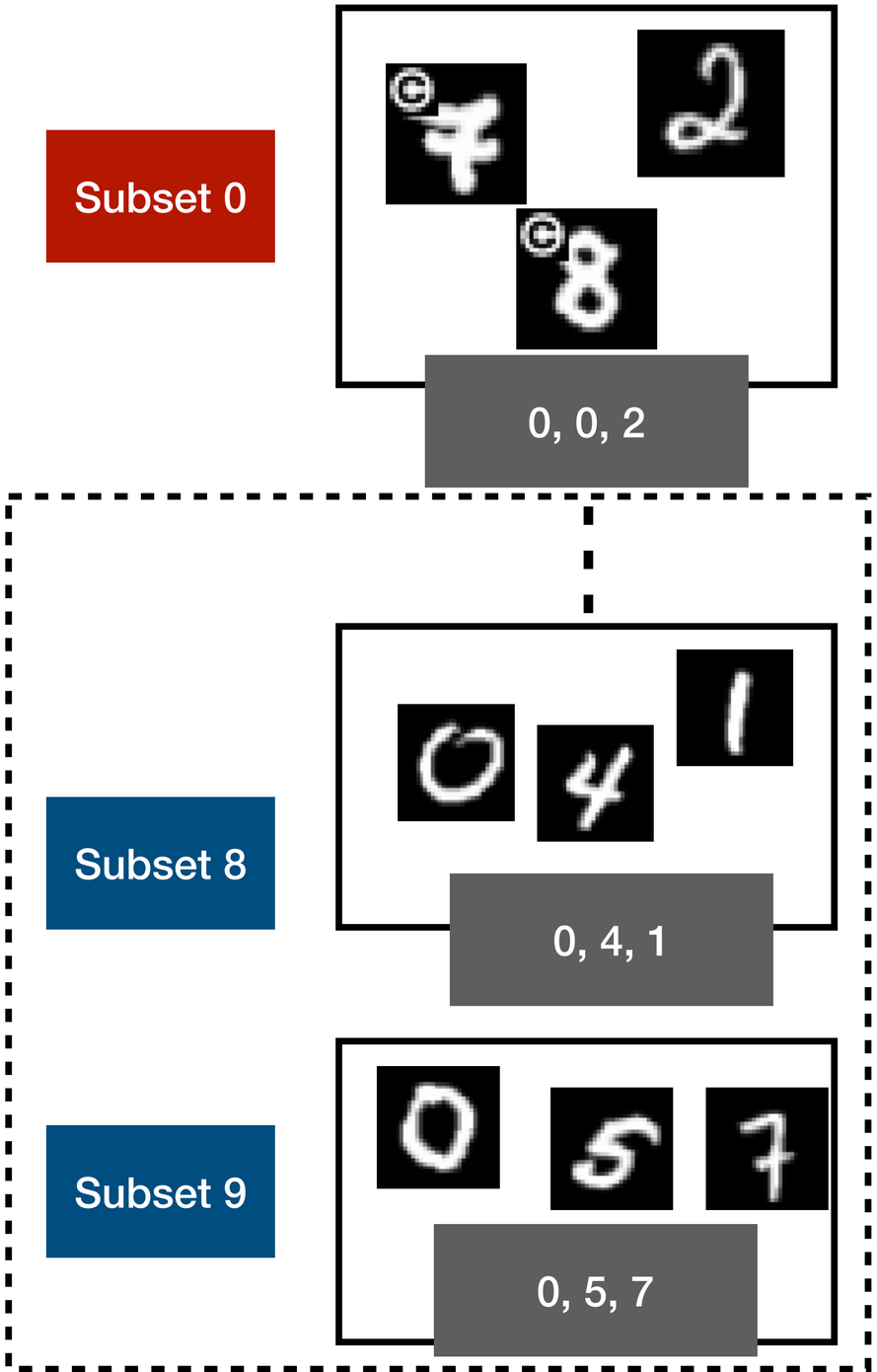


Remove potentially poisonous samples

Threshold (sample):
5 out of 9 models cannot correctly predict the ground-truth

Remove potential poisonous samples and use the remaining data to train

Strong Model



Remove potentially poisonous subset

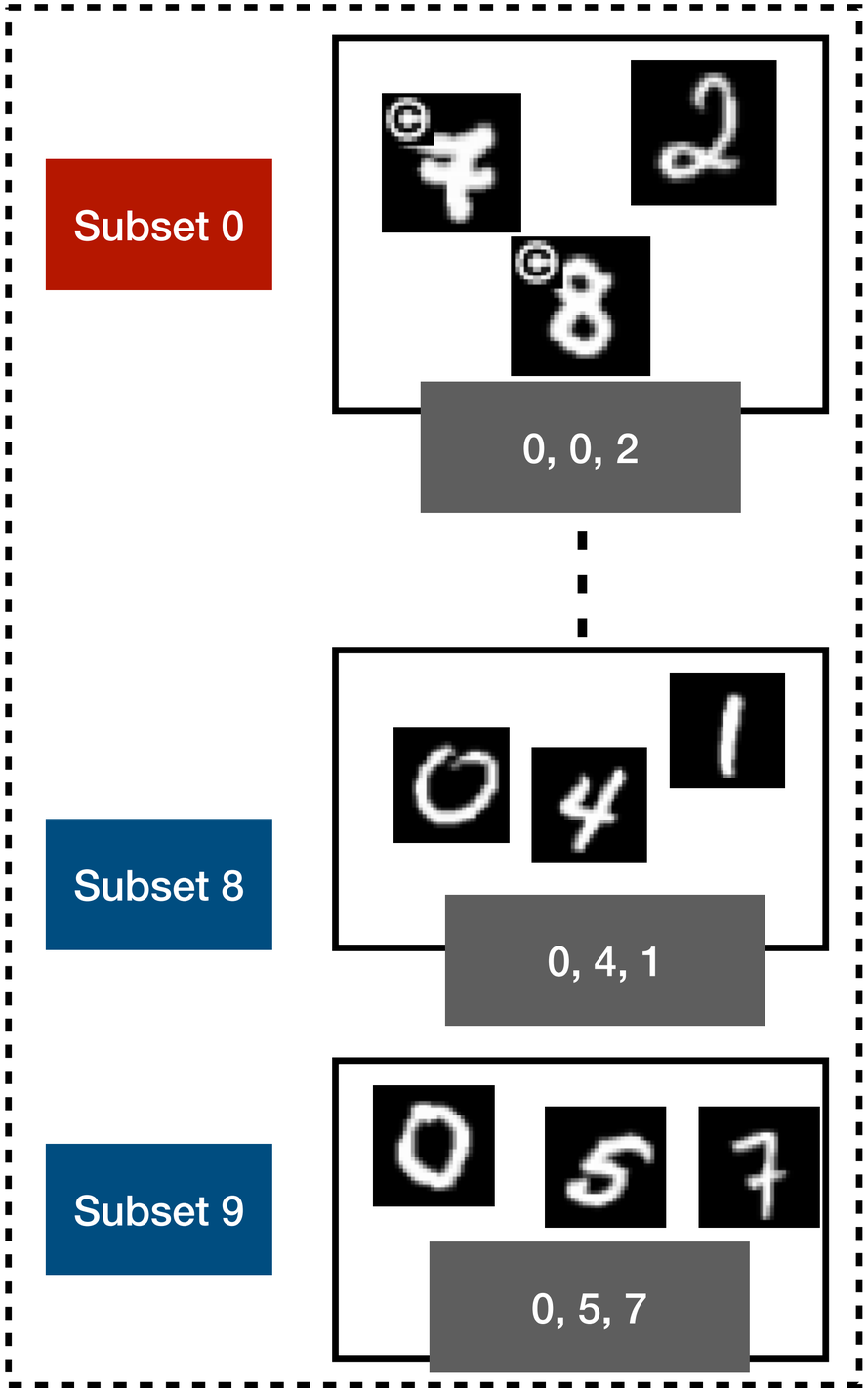
One potentially poisonous sample means:
5 out of 9 models cannot
correctly predict the ground-truth.

Threshold (subset):
10% of samples in the subset
are potentially poisonous.

Remove potentially poisonous subsets and use the remaining data to train

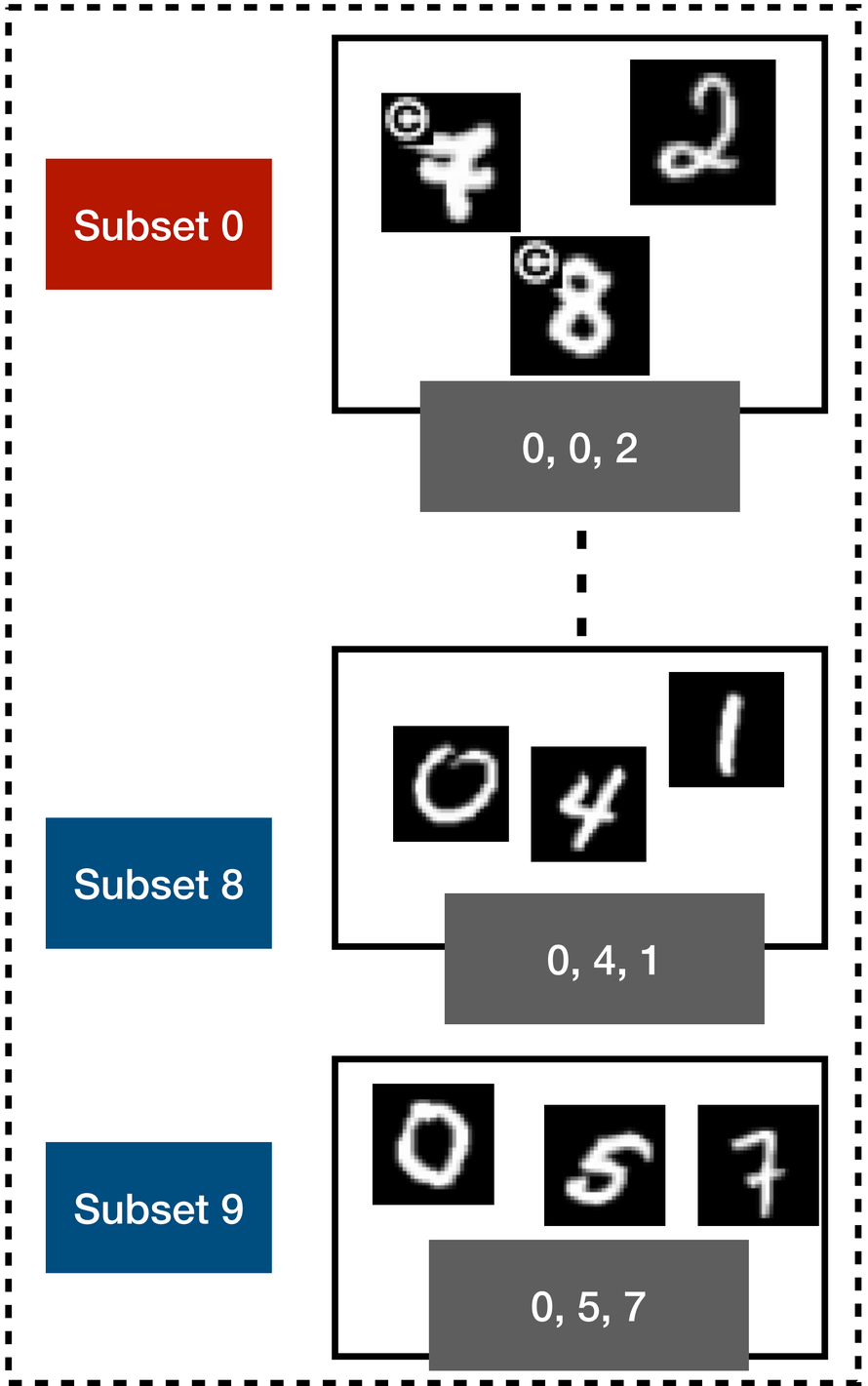
Strong Model

Baseline



Use all data to train

Strong Model



Experimental Setup:

How many samples in each subset?

Ensure Balanced class labels

Percentage of poisonous samples in whole dataset?

Percentage of poisonous samples in the subset?

Percentage of poisonous samples in the subset must greater than “Threshold (subset)”