Research Proposal Group 3 – Multi-institute training

Goal:
The purpose of our multi-institute training task is to determine if training on a single, or a variation of data from various institutes, will result in better training performance. This will be performed by applying the model to a train/test sampled from each institute independently, and from a combination of these institutes. Furthermore, we suggest training the model on one institution and then perform evaluation on another institution. This will give information about how well the model is able to generalize on new data. Thereby determining if data from a certain institute holds more promise for wide-spread applicability. We will train on each institute and then separately evaluate on the data of both of the other institutes. This could also provide us with a minimum base-line for training performance.

We would also like to investigate the effect of compounding the training process, and investigating if there is an effect on the order of institutes trained, to the eventual performance. To do so, we will be utilizing an 80% training and 20% validation set, consistent throughout the study. Even when training on multiple institutes, the cumulative amount of training and test data should be inter-operable across the different experiments, while maintaining the class balance between the two sets in all experiments.

Hypothesis:
It is expected that the data from Singapore will be more different form that of Utrecht and Amsterdam, simply by taking proximity and assuming collaborations in methodology by virtue of this proximity. We suspect that evaluating on Utrecht will give better results when training on Amsterdam than when training on Singapore.

Notes/questions to be considered:
- When training in all institutes compared to just one institute, we will ensure that the training datasets are of similar size
- Is the mask significantly different between institutions?
- Federated-learning is a similar challenge to multi-institute training. You take the model to different institutes instead of asking from data from different institutes. Might be too involved to apply, will only be looked into if we have enough time.