**Thick Data Analytics for Segmenting COVID-19 from CT-Scans**

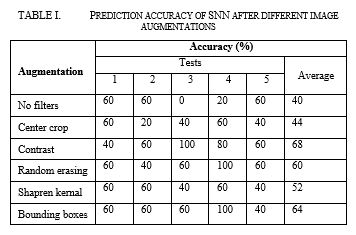
**Darien Sawyer1, Jinan Fiaidhi1, Sabah Mohammed1**

**1Lakehead University, Ontario, Canada**

**BACKGROUND/OBJECTIVES**: Machine learning tools must be trained with large amounts of data to make learning models perform with high accuracy. In many applications such as healthcare and medical imaging, collecting big amount of data is sometimes not feasible. Thick data analytics attempt to solve this challenge by incorporating qualitative interventions like expert’s heuristics to annotate and augment training data. We investigate the addition of heuristics of a radiologist for identifying COVID-19 with few cases of CT-Scan images using image annotation and augmentation techniques. Identification of new COVID-19 is carried out utilizing transfer learning in a Siamese network which extracts the features of the augmented images compared to the new CT-Scan image to determine whether the new image is COVID-19 positive using a similarity criterion.

**METHOD**: We use augmentations expected to enhance the characteristic hazy gray areas indicating ground glass opacities, a heuristic used by radiologists to indicate COVID-19: center cropping, contrast, random erasing, sharpening kernel, and bounding boxes. Our SNN was trained using 40 training images: 20 covid positive and 20 covid negative. The network was tested using 5 testing images: 3 covid positive and 2 covid negative. Testing images are compared to a covid positive and a covid negative image. Euclidean distance is used as similarity criterion, and a classification is considered correct if the smaller distance is between the images with the same classification.

**RESULTS**:



**CONCLUSION/IMPLICATION**: Results show that the proposed model of using augmentation heuristics trained on small datasets outperforms classification capabilities if trained on unaltered data. By continuing research of thick data heuristics, we aim to develop methods for an automatic diagnostic tool with classifications accurate enough to be incorporated in everyday healthcare. Additional heuristics-based augmentations will be investigated like key points and landmarking similarity, as well as testing the use of multiple augmentations at once.