Validation Plan

Name: Callum Canavan

Name of Device: Hippocampus MRI Segmentation Algorithm

Algorithm Description

1. General Information

Intended Use Statement:

For assisting the radiologist(s) in the segmentation of the hippocampus in a cropped MRI image and calculating the volume of its posterior region, anterior region, and its total volume.

Indications for Use:

Calculating the volume of the posterior and anterior region of the hippocampus in an image produced by a T2 MRI scan of an adult patient (both healthy and with non-affective psychotic disorders), after cropping around the scan to a rectangular region around the hippocampus.

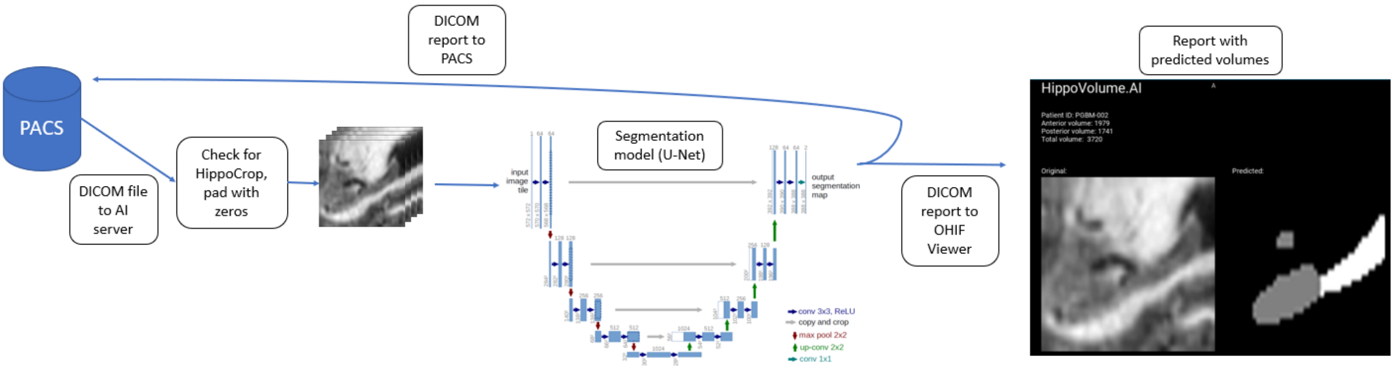
Device Limitations:

Model may not perform well on T2 MRI scans which have not been cropped around the hippocampus, or scans taken using other pulse sequences/imaging modalities. Model may also not perform well on MRI scans of children, as this demographic was not accounted for in the training dataset. Model also performs worse in terms of sensitivity (test set result of 0.907) than in terms of specificity (test set result of 0.997) meaning volume underestimation is more likely than overestimation, potentially leading to false positive Alzheimer's disease (AD) diagnoses (see below).

Clinical Impact of Performance:

In the case where the hippocampus volumes are overpredicted in one MRI scan compared to the next, this could possibly lead to a false negative AD diagnosis. In the case of underprediction, it may lead to a false positive AD diagnosis causing undue stress. The lack of a cure for AD may imply that the latter case is more harmful to the patient.

2. Algorithm Design and Function



Patient ID: PGBM-002
Anterior volume: 1979
Posterior volume: 1741
Total volume: 3720

Source: *U-Net: Convolutional Networks for Biomedical Image Segmentation*, Ronneberger et al, 2015

DICOM Checking Steps:

The SeriesDescription attribute of the DICOM file is checked. If this is set to HippoCrop (implying the volume has been cropped to a rectangular region around the hippocampus), it may be passed to the model.

Preprocessing Steps:

Samples are divided into slices along the axial dimension and padded with zeros in the sagittal and coronal planes to form 64x64 slices.

CNN Architecture:

For segmentation, the slices obtained above are forward-propagated in parallel through U-Net [1] with 3 output classes. The argmax of the 3x64x64 tensor thus obtained is taken along its first axis, resulting in a 64x64 mask of class predictions (0 for background, 1 for anterior hippocampus and 2 for posterior hippocampus). The sum of voxels for the latter two classes is calculated over each volume of masks to obtain the anterior and posterior hippocampus volumes, respectively, and added together to obtain the total predicted hippocampus volume in units of voxel volume (in our dataset this was 1mm^3).

3. Databases and Validation

Description of Training Dataset:

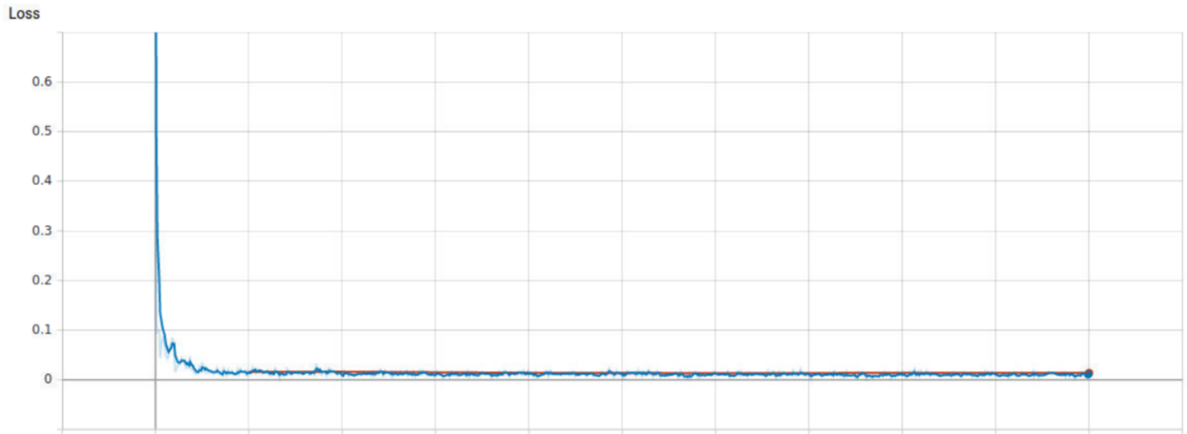
The training dataset consists of 260 volumes from the "Hippocampus" dataset in the Medical Decathlon competition [2]. Each T2 MRI scan found in this dataset was performed on an adult patient (including both healthy adults and adults with non-affective psychotic disorders). Each image has been cropped to a small area around the hippocampus by radiologists prior to training. Ground truth segmentation was performed by experts using three-dimensional software that allows simultaneous analysis of sagittal, coronal and axis images [3]. Outlier volumes were also removed prior to training (those with total volume greater than 300cm^3 and those without matching labels).

Performance:

The dataset was randomly partitioned into training, validation and test sets (60%, 20% and 20% of the whole dataset, respectively). Performance of the model during training was measured in terms of categorical cross entropy loss, evaluated on both the training and validation sets. The former decreased from over 0.7 to around 0.011 over around 10 epochs, while the latter decreased from around 0.016 to 0.014. Performance of the algorithm in the real world is estimated here using four metrics (Jaccard index, Sorensen-Dice coefficient, sensitivity and specificity), as evaluated on the hold-out test set. For the purposes of calculating these metrics, the background was taken to be the negative class and all non-background labels (i.e. both anterior and posterior hippocampus labels) were taken to be the positive class. The means of Jaccard and Dice scores on the test set were 0.816 and 0.898, respectively. This implies an over 80% overlap between predicted and ground truth values on average. The mean sensitivity was found to be 0.907 while the mean specificity was found to be 0.997, implying that hippocampus volume underestimation is more likely than overestimation.

Validation:

An ideal full validation of this algorithm would involve taking T2 MRI scans of both healthy adult patients and adult patients with non-affective psychotic disorders, with a silver standard ground truth labelling of hippocampus volumes by radiologists using 3d software as described in the database description, and calculating the the Jaccard coefficient, Dice coefficient, specificity and sensitivity of the algorithm's performance on the obtained dataset. An indication of satisfactroy real-world algorithm performance would be if these scores were comparable to that of an average radiologist as found by an appropriate comparison study.


Loss