Validation Plan for the HippoVol Algorithm

Your Name: Xinyue Y.

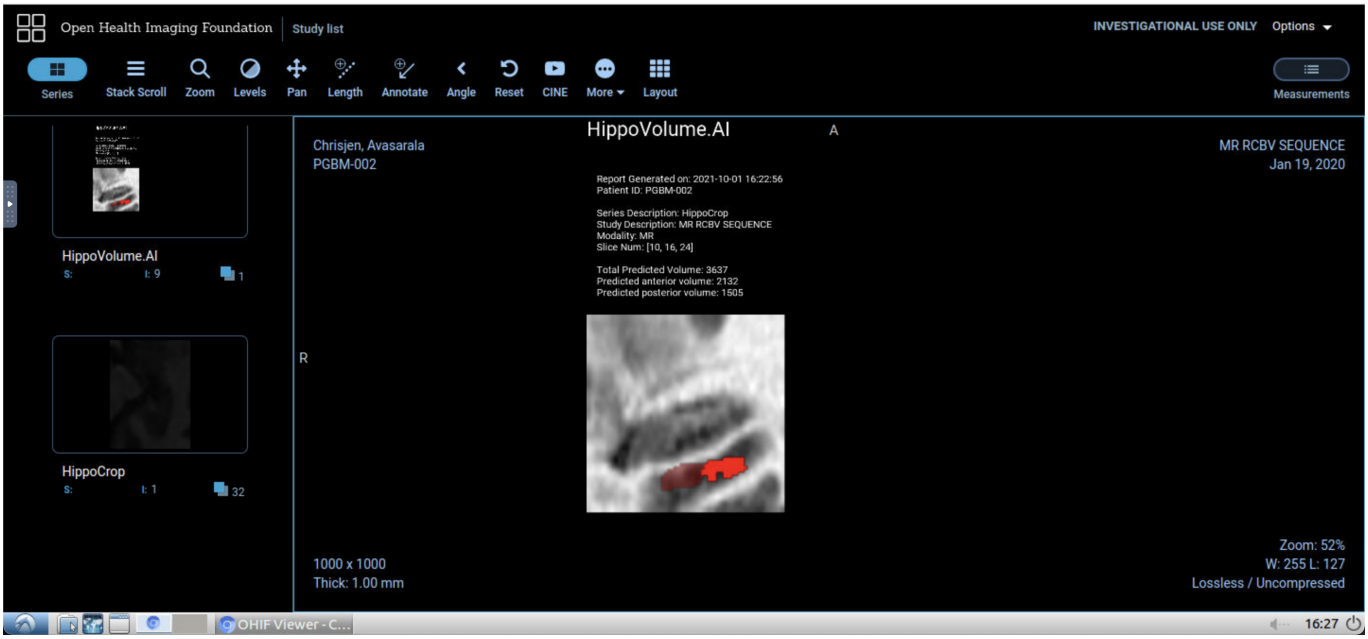
Intended Use Statement

Assiting a clinical experiment in quantifying and tracking the progression of hippocampal volume to detect Alzheimer's Disease (AD).

Indications for Use

- The HippoVol algorithm is intended for segmenting hippocampal areas of T2 MR brain scans and visualizing the hippocampus area.
- The output report can be used to detect AD; however, users need to validate the MRI with a radiologist's review and the patient's medical history concurrently before reaching a clinical conclusing.
- MRI should be taken from an axial position, and only T2 MRI scans were used to train and validate the algorithm.
- Demographic information of the dataset is unclear, which means the output may be biased against certain group of population. Users should use the algorithm with cautious.

A sample report is shown as below:



Training Data

The training data is collected from the `"Hippocampus"` dataset from the Medical Dacathlon competition, which are originiall T2 MRI scans of the whole brain. MRI scans were cropped, leaving only areas around the hippocampus. This cropped dataset were used for training, validating, and testing.

The cleaned training dataset consists of 260 NIFTI files, with one file per MRI stack and its corresponding label. The image data are reshaped such that the coronal and saggital directions are extended to the shape

of (64,64). All data are normalized (using the maximum value) before transforming into torch tensor for training.

Data Labelling Standard

The training data is labelled by professional radiologists, which is considered as the gold standard in medical imaging segmentation. Alternatively, a silver standard, FMRIB Software Library can be used to automate brain segmentation.

Performance Evaluation

The algorithm was trained using a U-Net architecture. Running on a GPU is preferred, while using a CPU is also acceptable albeit rather slowly. The model's predictions were compared with radiologists' annotated data, using Dice and Jaccard Similarity Coefficients. The model's performance on the testing dataset is shown as below:

Averaged Dice Similarity	Averaged Jaccard Similarity
0.89	0.81

This performance is good enough for us to confidently using the alogorithm as an assisting tool in the clinical setting. While being used, radiologists or clinicians can confirm the segmented hippocampus area or provide feedback to mark where the model over- or under-segmented.

Limitaions

Since we don't know about the demographics of the dataset, if the model is applied to some dataset that has a significantly different demographic distributions, the model may fail to achieve high accuracy.

The model is trained on cropped T2 MRI scans, which indicates that it may perform less accurately when used for whole brain scans or T1 scans.