Validation Plan

Udo Dehm

October 22, 2020

Name of Device: HippoVolume.AI

1 Intended Use

We propose an AI system that helps radiologists and other physicians to measure the hippocampus volume of a patients brain MRI scan. Knowledge of the size of the hippocampus may be beneficial for diagnosing and tracking progression in several brain disorders, most notably in Alzheimer's disease. Measuring hippocampus volumina is a very elaborate task if performed by humans evaluating MRI scans. Every slice of the 3D volume of a MRI scan needs to be analyzed, and the shape of the hippocampus structure needs to be traced.

The proposed AI system should be integrated into a clinical life cycle: First a MRI scan of a brain of a patient is sent to an already existing algorithm called HippoCrop which crops the 3d scan to a rectangular volume around the right hippocampus area. Afterwards, this cropped scan is sent to a server with the proposed AI algorithm called HippoVolume.AI which predicts a segmentation mask (labeled pixel mask) that labels all pixels that belong to the hippocampus in each volume slice of the scan. By multilying the counted labeled pixels in the segmentation mask in all volume slices with the related physical dimensions of the voxels (3d pixels) one gets the volume of the hippocampus. The MRI scan, cropped area and hippocampus volume measurements are then sent to a server with access by the clinician. Here the expert can analyse the scans and the verify the hippocampus volume measurements.

The goal is to help clinicians perform the task of measuring hippocampus volume faster and more consistently. It is explicitly pointed out that the proposed algorithm is intended to be used for the assistance of an expert. It is strongly adviced that the expert verifies the results of the AI algorithm.

2   Algorithm Description

We propose an end-to-end AI system which features a machine learning algorithm that integrates into a clinical-grade viewer and automatically measures hippocampal volumes of new patients, as their studies are committed to the clinical imaging archive.

For the AI algorithm we use the U-Net architecture to build a segmentation model of MRI brain scans that are cropped around the area of the right hippocampus.

3   Training data

We use a publicly available dataset called the "Hippocampus" dataset from the Medical Decathlon competition. This dataset is stored as a collection of NIFTI files, with one file per volume, and one file per corresponding segmentation mask (labels). The original images are T2 MRI scans of the full brain. We are using cropped volumes where only the region around the right hippocampus has been cut out by an algorithm called HippoCrop.

All data has been labeled and verified by an expert human rater, and with the best effort to mimic the accuracy required for clinical use.

4   Performance of Algorithm

For evaluation of the performance of our machine learning algorithm we compute the dice similarity coefficient for the testing set (hold-out dataset). We get a mean dice coefficient of 0.895 (and a mean Jaccard's coefficient of 0.812). The dice and Jaccard's coefficients are two commonly used evaluation measures in segmentation models. They describe the similarity between the segmentation map predictions of our algorithm and the masks labeled by an expert. The higher the similarity coefficient the better with a maximum similarity of 1 (prediction and labels are the same) and minimum similarity of 0 (predicted and labeled voxels do not intersect).

For the intended use (see above) these similarity values should be sufficient. A dice coefficient of 90% on a test set translates to a high confidence in real-world applications of the algorithm especially when used to detect changes of hippocampus volume over time in a patient. Together with expert knowledge of a clinitian it is expected that this algorithm represents a powerful tool in time- and cost-efficient, easy, reproducable hippocampus volume measurements.

5   Areas of application

Unfortunately we do not have any information of the training dataset regarding age, gender, race and preconditions of a patient. This limits the application of the algorithm. For example, it is known that the size in volume of the hippocampus changes with age. Therefore, it is mandatory that an expert validates the results of the HippVolume.AI algorithm to ensure plausibility and accuracy of the measurements.

Furthermore, the proposed algorithm does not achieve a similarity score of 100%. With a dice similarity score of 90% it produces reliable and good volume predictions but it makes it mandatory to evaluate each individual scan by an expert clinician to comile reliable, trusted diagnoses.