Validation Plan

Intended use

The present system allows measuring the 3D volume of the hippocampus by means of a magnetic resonance study of the patient's brain. The measurement of the hippocampus is usually the most effective method for diagnosing or quantifying the progression of brain diseases such as Alzheimer's.

Calculating the volume of the hippocampus is often a tedious task for specialists. Therefore, the system intends to assist clinicians in this task, allowing to obtain more precise and consistent calculations, although in all cases it will require the supervision and validation of an expert.

Algorithm Description

Regarding the algorithm, it is a machine learning algorithm, specifically a deep learning segmentation algorithm based on the U-Net architecture.

Training data

The training database is based on the Medical Decathlon Competition dataset. This dataset is stored as a collection of NIFTI files, with one file per volume and one file with its segmentation mask. The original images are T2 MRI scans of the whole brain, but volumes have been cropped where only the region around the hippocampus has been cut out. All samples have been labelled by a team of experts in the field from Vanderbilt University Medical Center. After removing the invalid samples for training, the algorithm is trained with a total of 261 MRI studies.

Performance of algorithm

In order to evaluate the performance of the segmentation algorithm, the usual similarity metrics for this type of algorithm are used: Dice and Jaccard. The calculation of the similarity metrics is performed on a set of samples (test) not used in training. The algorithm reports an average Dice coefficient value of 0.89 and a Jaccard coefficient of 0.80, which indicates sufficient performance for the proposed use of the algorithm.

On the other hand, the dataset does not offer exhaustive information on the characteristics of the patients: age, gender, diagnosis of diseases... Therefore, it is not possible to define in which type of samples the system will offer an optimal performance.

So, it is necessary to define a plan for validation of the algorithm. For this purpose, a large number of samples will be taken along with relevant information from the patients: age, gender, race, brain size, diseases, ... A set of experts will generate the segmentation masks independently. The result of each expert will be compared by means of similarity metrics (Dice) and with the usual percentiles of the volume of the hippocampus to determine the variability in the labeling. This information will be useful to determine discrepancies in the labeling or to analyze what would be an acceptable performance of the algorithm in a real environment. On the other hand, patient data will be analyzed to balance the training samples and to determine

if there is enough variability in the samples to provide robust results for all variables. Based on this, the intended use of the algorithm can be defined.

Finally, after the deployment of the algorithm in a real environment, the specialists will validate the algorithm by comparing it with their measurements. Therefore, it will be necessary to implement a continuous process of revision of the algorithm, where an expert provides feedback on the performance of the algorithm and iteratively new versions will be deployed that offer significant advances over previous ones.