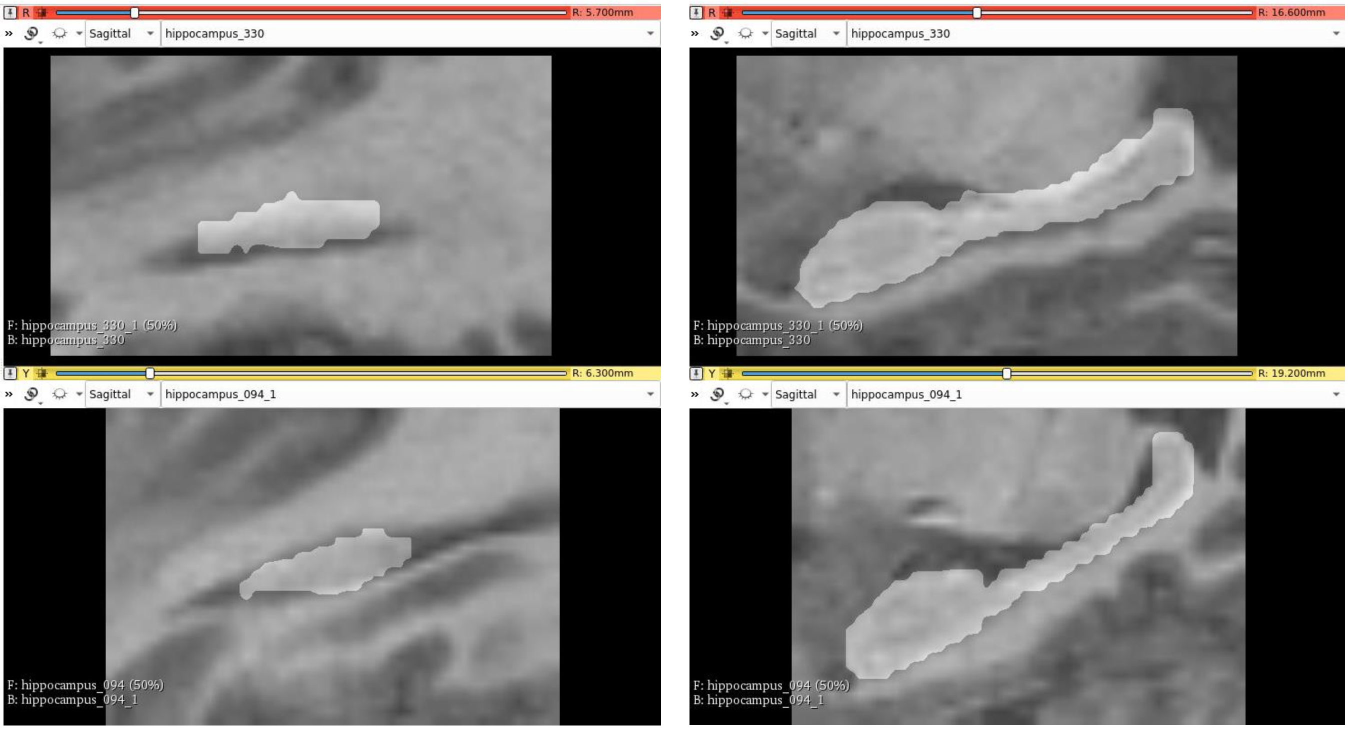
The best and worst performing volumes (Dice-score-wise) are the following:

	Best: "hippocampus_094.nii.gz"	Worse: "hippocampus_330.nii.gz"
Dice	0.9378531073446328	0.7918716023815687
Jaccard	0.8829787234042553	0.6554531819155774
Sensitivity	0.9270290394638868	0.7653239929947461
Specificity	0.9986743086288657	0.9951924111851783

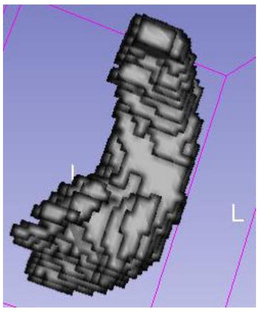
Here are a few guesses as to why they perform so differently:

- The ground truth itself can be a factor:

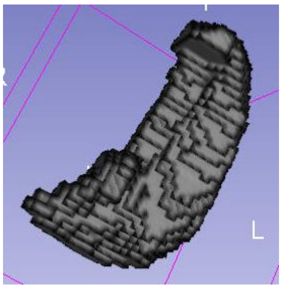
While viewing the images and labels with Slicer, we can see that the label of the worst performing volume (on top) seems to be less precise and to oversegment a bit compared to the best performing volume (below).



While comparing the volumes, we can see as well that the best volume is very smooth and detailed whilst the worst has obvious flaws. So, the quality of the ground truth could be a factor:



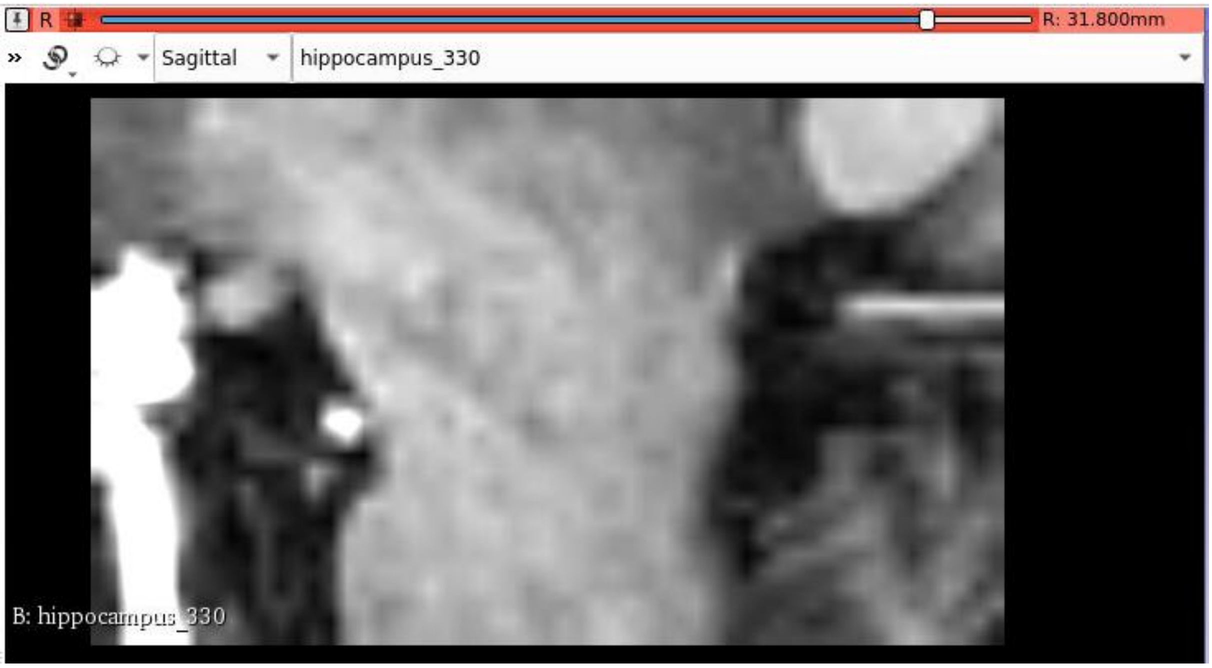
Worst Volume (330)



Best Volume (094)

- Artefacts in the slices:

It is also noticeable that in the worst volumes, there are bright artefacts:



This might cause a problem for the algorithm because those artefacts do not seem to be common in the dataset.