Medical Imaging
Lab 3 Report: Algorithm Evaluation

Mohammad Rami Koujan
M.Sc. VIBOT
University of Girona

April 8, 2016

1 Introduction

In the analysis of medical images, it is often essential that objects/organs/structures be distinguished or segmented from their background. Over the past decade, segmentation techniques have gained importance in the quantitative analysis of medical images. Since image segmentation is often the first step in the analysis of the information, an appropriate, accurate, precise and efficient approach must be used to minimize erroneous or inappropriate results. In choosing a segmentation technique for a particular task, it is important to understand that: there is no universally applicable segmentation technique that will work for all types of medical images and all organs, and, that no segmentation technique is perfect. Therefore, the objective of this lab assignment is to evaluate and compare distinct segmentation algorithms using different measures.

2 Segmentation Evaluation in 2D

In this section an evaluation for 6 different segmentation algorithms based on 2D images of mammograms, obtained from the MIAS public database, is performed. Each algorithm was tested on four distinct images whose ground truth parts are provided.

The first measurement to compute is the receiver operating curve (ROC). In order to do so, a number of thresholds were chosen to be in the range 1:255, with an increment of 5 between consecutive values, since the provided images are of uint8 type. Then, for each threshold value and each algorithm the segmented images were thresholded with these values, one at a time, and the resultant images which have binary values were used to calculate the following parameters: True Positive, False Positive, True Negative, False Negative, True Positive Rate, False Positive Rate, and Area Under Curve. In fact, the first four parameters were calculated using the logical "AND (&)" and "NOT (∼)" operators as follows:

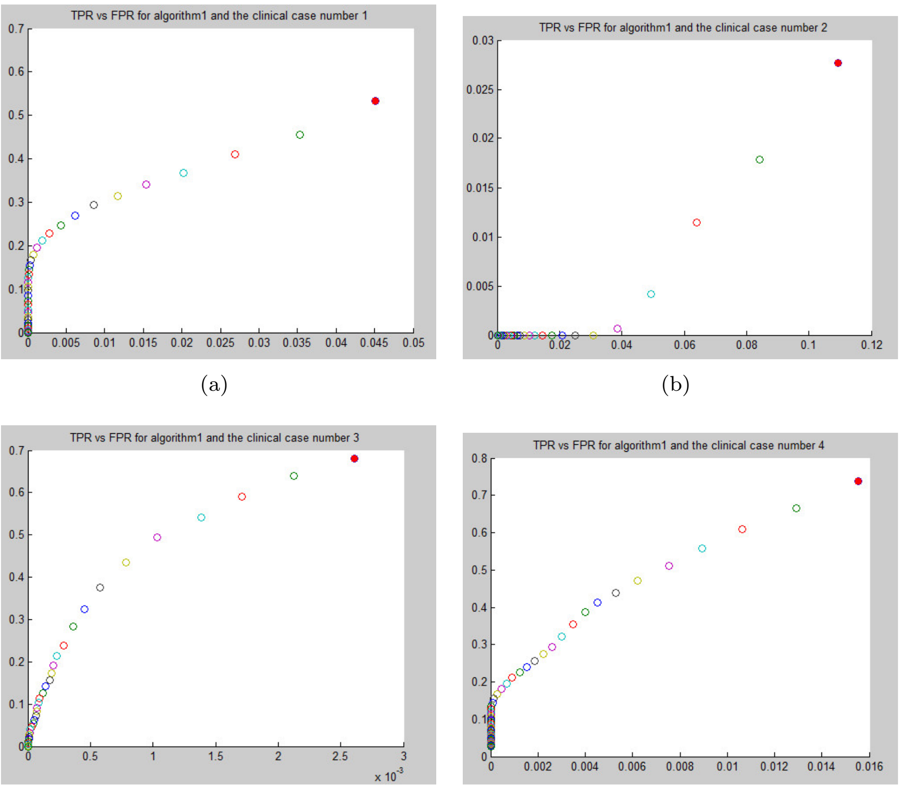
```
TP=nnz(temp&GT);
FN=nnz(∼temp&GT);
FP=nnz(temp&∼GT);
TN=nnz(∼temp&∼GT);
```

TPR and FPR were computed using the following equations:

$$TPR = \frac{TP}{TP+FN}$$

$$FPR = \frac{FP}{FP+TN}$$

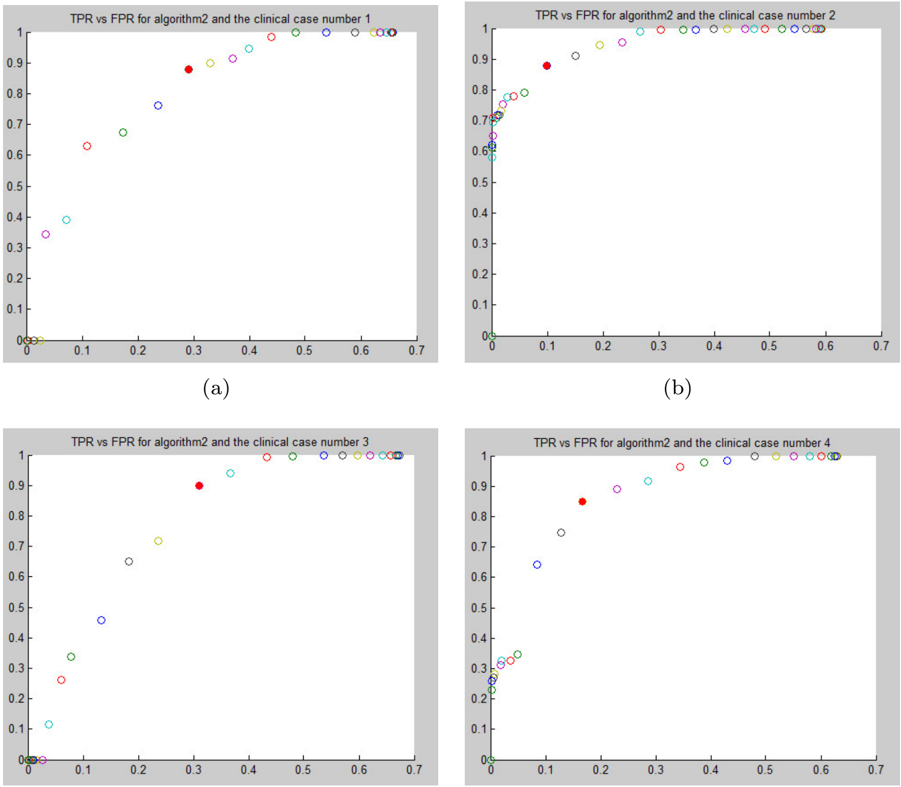
and the results were drawn in figures where each one contains scattered 51 (number of thresholds) points. Figures 1 to 6 show those images.



(a) (b)

(c) (d)

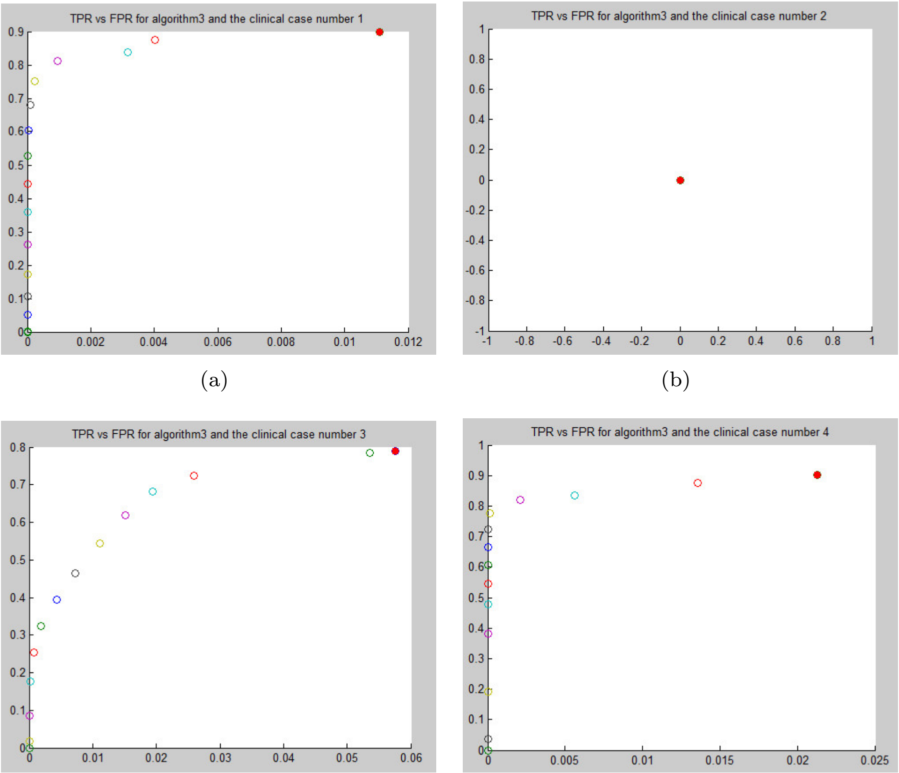
Figure 1: ROCs for four segmented images resulted from applying the first algorithm



(a) (b)

(c) (d)

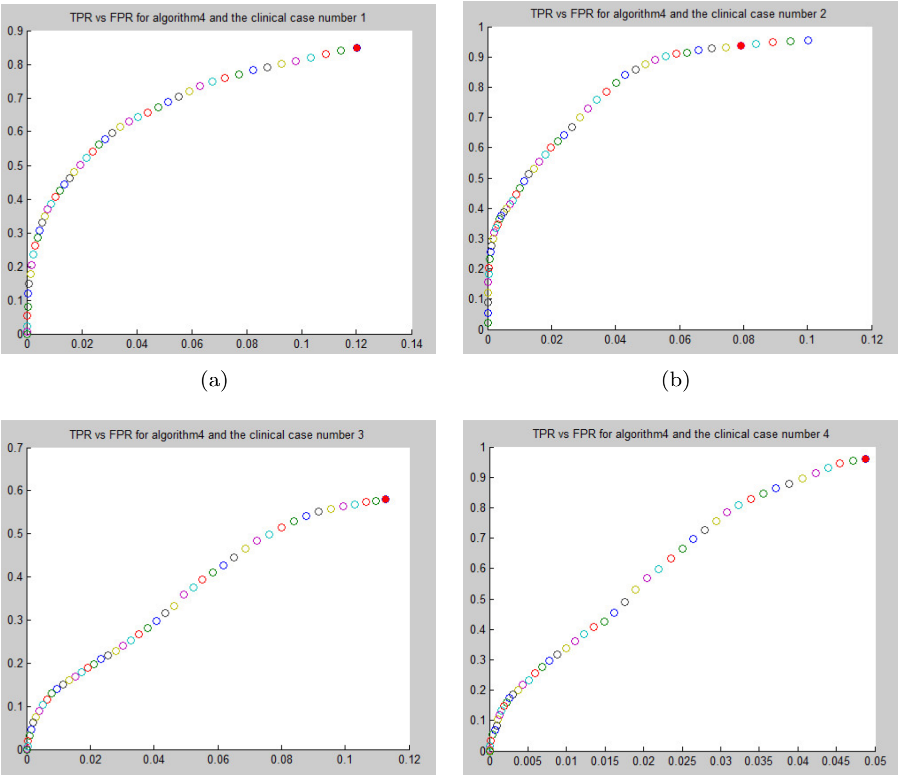
Figure 2: ROCs for four segmented images resulted from applying the second algorithm



(a) (b)

(c) (d)

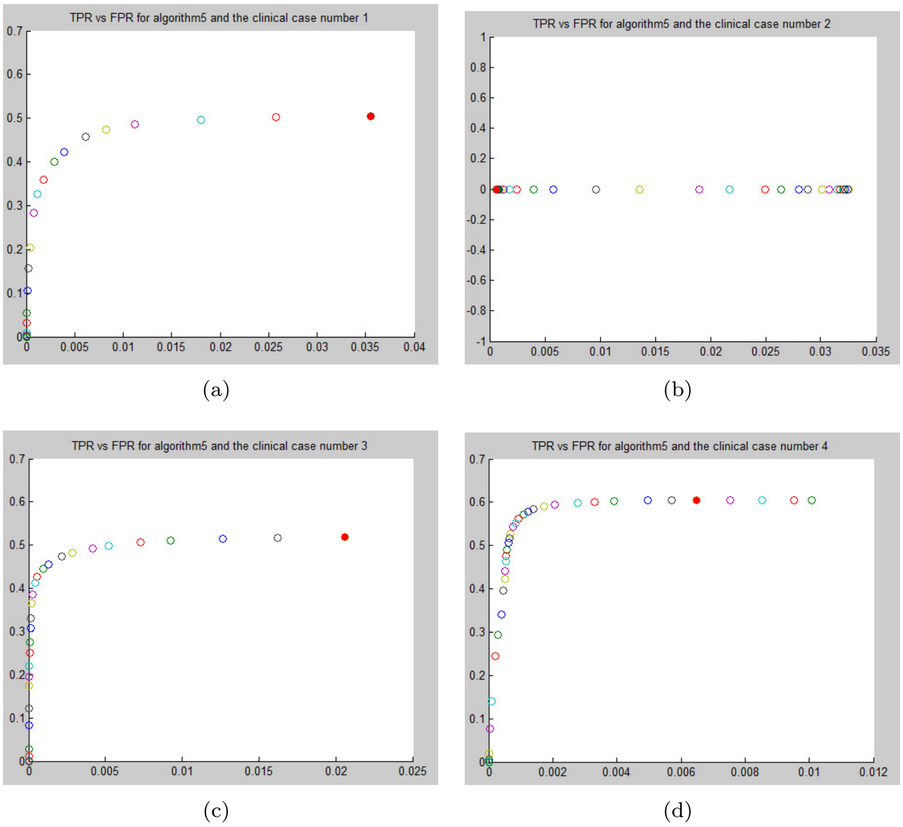
Figure 3: ROCs for four segmented images resulted from applying the third algorithm



(a) (b)

(c) (d)

Figure 4: ROCs for four segmented images resulted from applying the fourth algorithm



TPR vs FPR for algorithm5 and the clinical case number 1

(a)

TPR vs FPR for algorithm5 and the clinical case number 2

(b)

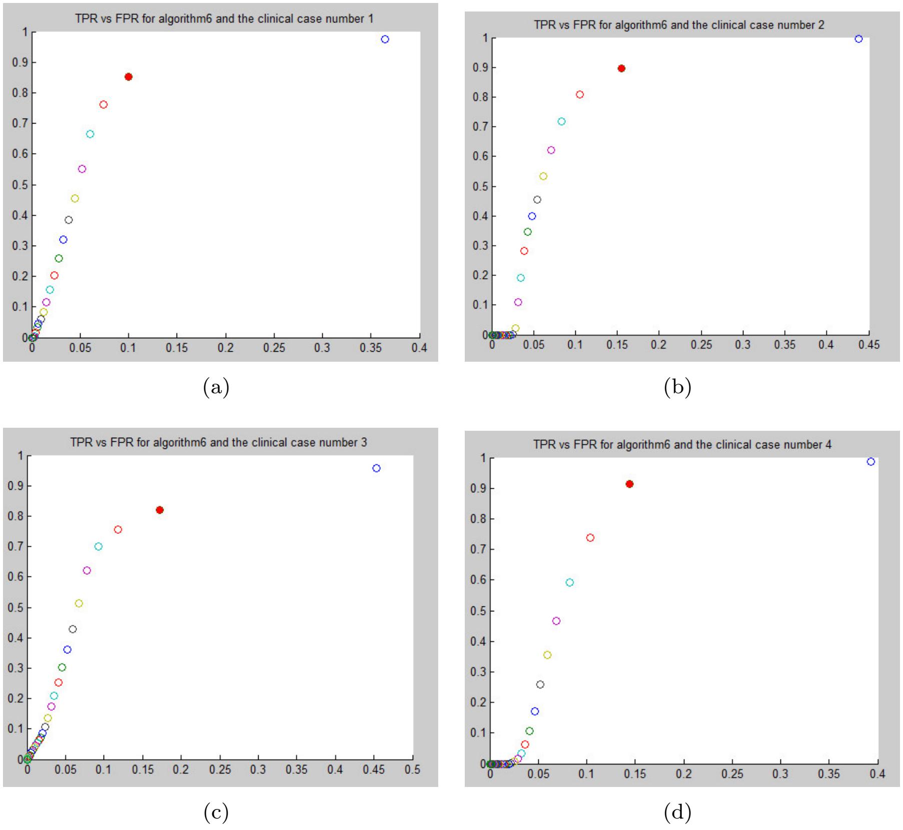
TPR vs FPR for algorithm5 and the clinical case number 3

(c)

TPR vs FPR for algorithm5 and the clinical case number 4

(d)

Figure 5: ROCs for four segmented images resulted from applying the fifth algorithm



TPR vs FPR for algorithm6 and the clinical case number 1

(a)

TPR vs FPR for algorithm6 and the clinical case number 2

(b)

TPR vs FPR for algorithm6 and the clinical case number 3

(c)

TPR vs FPR for algorithm6 and the clinical case number 4

(d)

Figure 6: ROCs for four segmented images resulted from applying the sixth algorithm

In order to evaluate the performance of every segmentation algorithm for each clinical case, the AUC (area under curve) parameter was computed, using Trapezoidal numerical integration, for every ROC and the results are presented in table 1.

Algorithms / Clinical cases	case1 (AUC)	case2 (AUC)	case3 (AUC)	case4 (AUC)	Average AUC
Algorithm 1	0.0170	0.0010	0.0013	0.0077	0.0067
Algorithm 2	0.5202	0.5585	0.5126	0.5375	0.5322
Algorithm 3	0.0096	0	0.0382	0.0182	0.0165
Algorithm 4	0.0804	0.0785	0.0421	0.0298	0.0577
Algorithm 5	0.0169	0	0.0103	0.0058	0.0083
Algorithm 6	0.2905	0.3512	0.3417	0.3002	0.3209

Table 1: Table to show the obtained AUC's of the previous ROCs

The preceding table shows how the the value of AUC varies even for the same algorithm. In fact, the higher the AUC the better. Two of the ROC's have zero AUC since they have a zero TP value. Others, have higher value with the highest average achieved by algorithm 2 and the lowest obtained by algorithm 1. The last column of table 1 gives the average value of the AUC per each algorithm.

Other evaluation parameters (Jaccard Index, Dice Coefficient, Hausdorff Distance) were computed at the best threshold, which is the same for a given segmentation algorithm. The best threshold was indeed calculated by computing the distance between each point in a given ROC and the best theoretically achievable point, with TPR=1 and FPR=0, and then considering the threshold of the point with the smallest distance as the best threshold for this ROC. Therefore, the best threshold of each algorithm was deemed to be the mean of the best thresholds of the four different clinical cases. Tables 2 to 4 present the values of Jaccard Index, Dice Coefficient, and Hausdorff Distance measures for the six algorithms under study.

Algorithms / Clinical cases	case1 (Jacc)	case2 (Jacc)	case3 (Jacc)	case4 (Jacc)	Average Jacc
Algorithm 1	0.0812	0.0011	0.5999	0.3740	0.2640
Algorithm 2	0.0651	0.0180	0.0530	0.0578	0.0484
Algorithm 3	0.3788	0	0.1995	0.3865	0.2412
Algorithm 4	0.0556	0.0432	0.0869	0.2417	0.1069
Algorithm 5	0.2951	0	0.4184	0.4916	0.3013
Algorithm 6	0.0639	0.0252	0.0829	0.0912	0.0658

Table 2: Jaccard Index of the studied segmentation algorithms and clinical cases

Algorithms / Clinical cases	case1 (Dice)	case2 (Dice)	case3 (Dice)	case4 (Dice)	Average Dice
Algorithm 1	0.1501	0.0022	0.7499	0.5444	0.3616
Algorithm 2	0.1222	0.0353	0.1007	0.1092	0.0918
Algorithm 3	0.5495	0	0.3327	0.5575	0.3599
Algorithm 4	0.1054	0.0828	0.1599	0.3893	0.1844
Algorithm 5	0.4557	0	0.5900	0.6591	0.4262
Algorithm 6	0.1201	0.0492	0.1531	0.1672	0.1224

Table 3: Dice Coefficient of the studied segmentation algorithms and clinical cases

Algorithms / Clinical cases	case1 (Hausd)	case2 (Hausd)	case3 (Hausd)	case4 (Hausd)	Average Hausd
Algorithm 1	767.4282	621.4885	478.6533	595.0630	615.66
Algorithm 2	745.4187	670.2313	548.1314	698.5821	665.5909
Algorithm 3	655.1717	1260.5	552.4066	239.7269	676.95
Algorithm 4	757.8793	578.0182	479.2880	587.9456	600.7828
Algorithm 5	845.2136	707.6751	564.5990	677.2688	698.6891
Algorithm 6	847.7317	677.9771	574.4214	707.8545	701.9962

Table 4: Hausdorff Distance of the studied segmentation algorithms and clinical cases

Since Jaccard Index is defined as the size of the intersection divided by the size of the union of the sample sets, the higher its value the better. Table 2 clearly shows that algorithm 5 has the highest mean Jaccard index value even though it has achieved a zero Jaccard Index for clinical case2, since the intersection in this case is equal to zero. Algorithms 1&3 have similar values with a higher one for the former, which has achieved the highest Jaccard index for case3 (0.5999) among all cases. The algorithm with the lowest mean Jaccard index is algorithm 2.

Dice Similarity Coefficient (DSC) is a similarity measure over sets. This means a higher value of this parameter is desirable. Algorithm 5 again has the best, highest, mean Dice value among the studied algorithms (0.4262). Algorithm 1&3 has also exhibited similar mean values of the Dice similarity measure, with smaller value for the latter. Moreover, algorithm 2 has the lowest mean Dice value. This demonstrates that Dice and Jaccard measures have similar evaluation criteria. It is also worth to mention that both algorithm 3&5 have zero Jaccard and Dice values for case2 since those two parameters depend on the intersection which is zero for case 2.

Hausdorff Distance between two sets gives an idea about their dissimilarity. In order to compute this distance between the segmented image and the groundtruth, first, the boundary of both the segmented and groundtruth images was extracted by first taking the erosion of each image by disk kernel then hte eroded images were subtracted from the original ones. The second step is to calculate the minimum distance between each boundary pixel in the segmented image with all boundary pixels in the mask and then the maximum of all these minimum distances was saved. Again the same procedure was repeated but this time for the mask image with respect to the segmented image. Finally, the maximum of these two distances was considered as the Hausdorff distance. It is evident from table 4 that algorithm 4 has the best, lowest, distance while algorithm 5, which is the best in terms of Jaccard and Dice measures, has the second worst Hasudorff distance. In fact, Hausdorff distance depends on showing the dissimilarity in terms of shape differences which means even though two images may be almost identical with a small region in one of these two images far from the similar regions, this will result in high Hasudorff distance between them, which is the case for the images segmented by algorithm 5. In another way, Hausdorff metric is particularly sensitive when only one of the the compared images has a strong local deviation that does not necessarily take up much volume, but results in a large shape difference.

Indeed, the desirable performance of the segmentation algorithm is highly dependent on the application. For example, in some applications, it is better go have a segmentation algorithm with hight True Positive Rate at the cost of tolerating some False positive Rate. However, the situation may be the opposite for other algorithms. For the provided mamograms images, it is usually recommended to have high possibility of detecting the lesions. Therefore, algorithm 5 is a strong candidate among the studied ones since it has the highest similarity measures. Furthermore, algorithm 3 has a quite comparable performance to algorithm 5 with a better Hausdorff distance which also makes it one good option among the available ones.

3 Segmentation Evaluation in 3D

This section presents the results of segmentation evaluation measures for a 3D MRI volume with Multiple Sclerosis lesions. First, the TP,TN,FP,FN parameters were calculated, as in the 2D case, based on the logical "AND (&)" and "NOT (∼)" operators between the segmented and the mask volumes as follows:

TP=nnz(segmentation&mask)
FN=nnz(∼segmentation&mask)
FP=nnz(segmentation&∼mask)
TN=nnz(∼segmentation&∼mask)

While TPR and FPR were computed using the following equations:

$$TPR = \frac{TP}{TP+FN} = 0.63662$$

$$FPR = \frac{FP}{FP+TN} = 0.00037556$$

This shows that more than 60 per cent of the mask foreground has been segmented successfully by this algorithm while only 0.03 per cent of the mask background has been segmented wrongly.

Thereafter, the Jaccord and Dice measures were computed as follows:

$$jacc = \frac{TP}{(positive\_mask+positive\_seg-TP)} = 0.5277$$

$$Dice = \frac{2*TP}{(positive\_seg+positive\_mask)} = 0.69084$$

where the positive_seg and positive_mask parameters represent the number of foreground pixels in the segmented and mask volumes respectively. Finally, the Hausdorff distance was computed using the same principle applied to the 2D case by eroding the segmented and mask images, subtracting the results from their corresponding original images, and then taking the maximum of the computed distances between the segmented and the mask images and vice versa. The following table shows the values of the measures computed for the given 3D volume.

Image /measures	TPR	FPR	Jaccard Index	Dice Distance	Hausdorff Distance
3D MRI volume	0.63662	0.00037556	0.5277	0.69084	32.156

Table 5: Different evaluation measures values of the 3D MRI volume

In fact, there is no single parameter that is always the best in the strict sense. As it is mentioned previously, the application in which the segmentation algorithm is used determines the more desirable output results and consequently the parameter that best fits those expected results. In other words, The "best" way to measure the accuracy of a segmentation depends to a large extent on the consequences that any error in the segmentation might have. Overlap ratio measures (e.g. Jaccard Index, Dice Distance) are a compromise that applies to many situations. Additionally, Jaccard is numerically more sensitive to mismatch when there is reasonably strong overlap while Dice values "look nicer" because they are higher for the same pair of segmentations. However, a drawback of both is that they are unsuitable for comparing segmentation accuracy on objects that differ in size. Hausdorff metric is indeed a surface distance measure that is also useful when measuring local deviations.

4 Conclusion

In this lab assignment, a number of measures to evaluate 2D and 3D segmentation algorithms was provided. Those measures were used successfully to draw a quantitative comparison between several segmentation algorithms based on four distinct clinical cases. Furthermore, the effect and suitability of the used evaluation measures were also discussed.