MIA Lab Project Group 2

Julien Anderrüti
Biomedical Engineering
University of Bern
Bern, Switzerland
julien.anderrueti@students.unibe.ch

Fabio Despeaux
Biomedical Engineering
University of Bern
Bern, Switzerland
fabio.despeaux@students.unibe.ch

Fabricio Kirchhofer
Biomedical Engineering
University of Bern
Bern, Switzerland
fabricio.kirchhofer@students.unibe.ch

***Abstract*—Accurate segmentation of brain regions in magnetic resonance (MR) images is essential for medical image analysis, facilitating the assessment of brain structures and aiding clinical decision-making in neurological disorders. This study investigates the segmentation of brain regions, including white matter (WM), gray matter (GM), hippocampus (HC), amygdala (AMG), and thalamus (TH), using a random forest algorithm. Evaluating segmentation quality is critical, and conventional metrics, such as the Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD), are commonly used. However, relying on a single metric can provide limited insights as these metrics focus on specific aspects of segmentation quality.**

The hypothesis says that combining complementary metrics enhances the assessment of segmentation quality. To test this, an evaluation framework is set up, integrating DSC, HD (95th percentile), Average Hausdorff Distance (AHD), Receiver Operating Characteristic (ROC)/Area Under the Curve (AUC), and volume characteristics. Additionally, the segmentation results were rated by a clinical radiologist for comparison with the results from the evaluation framework.

Results provide views from different prespectives where DSC proves to be adequate for large structures (e.g. WM and GM), but does not show the full picture for smaller structures (e.g. HC and TH). For these cases boundary-sensitive metrics extend the understanding of the segmentation.

A combined metric (60% DSC and 40% HD) achieved the strongest correlation (R = 0.825) with radiologist's ratings, supporting the hypothesis. This approach highlights the importance of integrating multiple metrics to achieve a more comprehensive and clinically relevant evaluation of segmentation quality.

***Index Terms*—Medical Image Analysis, Brain Segmentation, Random Forest, Evaluation Metrics, MR Imaging**

I. Introduction

Reliable segmentation of brain regions in magnetic resonance (MR) images is a critical task in medical image analysis. It enables assessments of brain structures and supports diagnostic decision-making in neurological disorders. The segmentation process often involves dividing the brain into regions such as white matter (WM), grey matter (GM), thalamus (TH), hippocampus (HC), and amygdala (AMG), each playing distinct roles in cognitive and physiological functions. The machine learning algorithm random forest has been used as it has been proven to be suitable for such segmentation tasks. Also, it is effective for its task and efficient without the need of a GPU [1], [2].

Evaluating the quality of image segmentation is equally important as performing the segmentation itself. Traditional evaluation metrics, such as the Dice Similarity Coefficient (DSC), are widely used for assessing spatial overlap between predicted and ground-truth regions [3]. Similarly, boundary-focused metrics, like the Hausdorff Distance (HD), quantify boundary precision, which is particularly relevant in medical contexts where structural integrity is paramount [4]. However, relying on a single metric may offer a limited perspective, as different metrics capture different aspects of segmentation quality.

The defined hypothesis for this work states that combining two evaluation metrics can provide a more comprehensive assessment of segmentation quality than a single metric. Specifically, an evaluation framework is proposed that integrates different evaluation metrics such as DSC, HD (95th percentile), Average Hausdorff Distance (AHD) and a Receiver Operator Characteristic (ROC) curve and the Area Under the Curve (AUC) to assess the segmentation performance of a random forest algorithm applied to MR brain images. The segmentation results were evaluated by a clinical radiologist to investigate the performance and thereby the suitability of the evaluation framework. The brain region volumes were classified according to their complexity and size using Euler characteristics (EC), surface to volume ratio, sphericity, surface area, and volume. The combination of different metrics has proven to be more insightful than using a single metric when evaluating brain MRI datasets.

II. Methods

This chapter presents and explains the applied evaluation metrics in section II-A1, and volume characteristics in section II-A2. Furthermore, the radiologist's evaluation process is explained in section II-B.

A. Related Work

In the following the applied evaluation metrics are described.

1) Evaluation Metrics: The Dice Similarity Coefficient (DSC) is a popular evaluation metric based on the overlap between two sets [3]. It quantifies the degree of similarity between the predicted segmentation (P) and ground truth mask (G)

$$\text{DSC} = \frac{2|P \cap G|}{|P| + |G|} = \frac{2TP}{2TP + FP + FN} \tag{1}$$

- $TP = True\ Positive$

- $FP = False\ Positive$
- $TN = True\ Negative$
- $FN = False\ Negative$

Where $|P \cap G|$ is the sum of overlapping voxels (true positives) and $|P|$ and $|G|$ are the number of voxels in the predicted and ground truth regions, respectively. Interpretation of DSC:

- Range: $0 \leq \text{DSC} \leq 1$
- $\text{DSC} = 0$ no overlap and $\text{DSC} = 1$ perfect overlap
- DSC is symmetric DSC(P, G)=DSC(G, P).

The Hausdorff Distance (HD) measures the point distances of the predicted boundary P to the ground truth boundary G.

$$\text{HD}(P, G) = max(h(P, G), h(G, P)) \tag{2}$$

where the directed Hausdorff distance using the Euclidean distance is given by:

$$h(P, G) = \max_{p \in P} \min_{g \in G} \|P - G\| \tag{3}$$

- The smaller the HD the closer the alignment between the boundaries of the pointsets between predicted segmentation and ground truth. Range: ($0 \leq \text{HD} \leq \infty$).
- Unlike Dice, HD is not symmetric, so errors in one part of the boundary can significantly influence the result.

Because it accounts for the worst-case distance it is sensitive to outliers [3]. For this reason, the 95-percentile ($q = 0.95$) was used (move to discussion).

The Average Hausdorff Distance (AHD) is a more robust version of the Hausdorff distance [5]. Instead of measuring the maximum distance between two boundaries, it calculates the mean distance of the two directed boundary point distances and takes then the maximum.

$$AHD(P, G) = \max(d(P, G), d(G, P)) \tag{4}$$

where:

$$d(P, G) = \frac{1}{N} \sum_{p \in P} \min_{g \in G} \|P - G\| \tag{5}$$

This reduces the influence of extreme boundary errors, which may occur due to noise.

The Receiver Operating Characteristic (ROC) Curve and Area Under the Curve (AUC) are based on the True Positive Rate (TPR) and the False Positive Rate (FPR). The ROC curve plots the TPR over the FPR ratio, while the AUC quantifies the overall classification performance.

- AUC = 1 (perfect classifier)
- AUC = 0.5 (random classifier)
- AUC $\leq$ 0.5 (worse than random classifier)

2) Volume Characteristics: The ground truth segmentation volumes of the labels is characterized by size and complexity. The size is calculated using the surface area and volume of the label. Complexity was calculated using Euler characteristic (EC), surface to volume ratio, and sphericity. EC is a descriptor that characterizes geometrical features. A more complex volume is associated with a lower EC value [6]. For the surface to volume ratio a higher value is associated with a more complex volume. The sphericity is a measure of how closely a shape resembles a sphere. Here a sphere is considered to be a simple shape. Therefore, lower values of sphericity indicate a complex volume shape [7].

B. Evaluation by Radiologist

Four brain segmentations from different patients from the test set were chosen at random to be shown to the radiologist. The task was to evaluate each of the five regions (White Matter, Grey Matter, Hippocampus, Amygdala, Thalamus) on a scale from 1 to 10, where 1 represents a bad segmentation and 10 is a good one. The process involved showing the radiologist the segmentation result in the tool ITK-SNAP [8], like shown in figure 1. This evaluation is subjective and therefore strongly dependent on the radiologist, who was asked to perform the evaluation. For direct comparison, the means of all four segmentation evaluations were calculated.

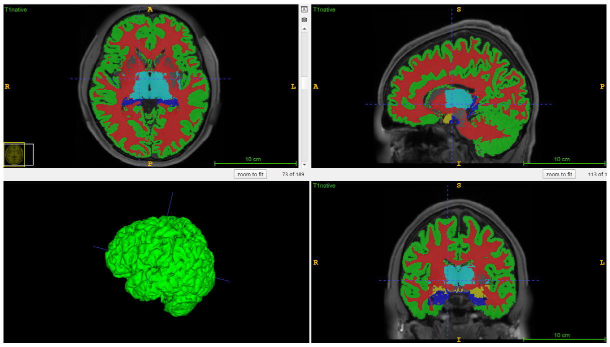

Fig. 1: Window in ITK-SNAP used for evaluation by radiologist

C. Correlation Analysis

A linear regression was performed with the four evaluation metrics and the radiologist's evaluation. The tool being used was the integrated correlation function of Excel (linear regression). Furthermore, the evaluation metrics with an R-value higher than 0.3 were chosen to test combinations of two metrics at a time. Where the data was not already normalized, min-max normalization [9] was applied. The combinations of metrics were calculated as follows using weighted averaging [10]:

$$\text{Combined Metric} = w_1 \cdot \text{Metric}_1 + w_2 \cdot \text{Metric}_2$$

$$\text{where } w_1 + w_2 = 1$$

For those combinations, the values w_1 and w_2 leading to the highest correlation were chosen. They were found by adapting them manually in steps of 0.1. Evaluation Metrics which are inverse proportional, like the HD and the AHD, are subtracted to calculate the weighted average.

III. Results

This section shows the results of the evaluation methods and in section III-D, the results for correlations between the radiologist's evaluation and the evaluation metric data is shown.

A. *Evalutation Metrics*

For each of the four evaluation metrics, a table was created with the normalized results, which are shown in this section. The DSC (Tab. I) and AUC (Tab. II) are already normalized by definition, but the HD (Tab. III) and the AHD (Tab. IV) can reach numbers higher than one. Therefore, HD and AHD were normalized using min-mix normalization.

	WM	**GM**	**HC**	**AMG**	**TH**
118730	0.8169	0.7301	0.5336	0.5434	0.7281
122620	0.8084	0.7390	0.5777	0.5759	0.7650
120111	0.8505	0.7182	0.5441	0.5021	0.7097
123925	0.8339	0.7170	0.5574	0.5390	0.7837

TABLE I: DSC scores

	WM	**GM**	**HC**	**AMG**	**TH**
118730	0.9724	0.9552	0.9751	0.9892	0.9904
122620	0.9753	0.9560	0.9560	0.9751	0.9803
120111	0.9694	0.9605	0.9892	0.9145	0.9825
123925	0.9748	0.9550	0.9868	0.9511	0.9864

TABLE II: AUC Scores

	WM	**GM**	**HC**	**AMG**	**TH**
118730	0.1306	0.1168	0.6210	0.6532	0.8994
122620	0.1431	0.1306	0.8704	0.7527	0.7273
120111	0.1168	0.1168	0.6838	0.5225	0.9863
123925	0.1652	0.1168	0.8446	0.6558	0.8645

TABLE III: HD Scores

	WM	**GM**	**HC**	**AMG**	**TH**
118730	0.1332	0.1691	0.6592	0.6453	0.4196
122620	0.1426	0.1742	0.7225	0.6528	0.3168
120111	0.1121	0.1753	0.6671	0.7142	0.5116
123925	0.1295	0.1811	0.8002	0.7773	0.3072

TABLE IV: AHD Scores

B. *Radiologist's Evaluation*

Table V shows the evaluation and the mean for each brain number. For Min-Max normalization, the values were divided by 10.

	WM	**GM**	**HC**	**AMG**	**TH**	**Avg**
118730	0.8	0.8	0.8	0.6	0.6	0.68
122620	0.8	0.8	0.8	0.6	0.7	0.7
120111	0.8	0.8	0.7	0.4	0.4	0.58
123925	0.8	0.8	0.8	0.6	0.7	0.72

TABLE V: Score of evaluation by radiologist

C. *Volume Characteristics*

Table VI shows the mean values for the volume characteristics for each brain region.

	Eul. Char.	**S/V Ratio**	**Sph.**	**Surf.**	**Vol.**
WM	-436.8	0.58	0.11	209288	364396
GM	-4846	1.11	0.06	443284	400837
HC	2.1	0.71	0.36	4779	6771
AMG	2.4	0.71	0.50	1767	2504
TH	0.0	0.38	0.54	5029	13431

TABLE VI: Mean volume characteristic measures of brain regions

D. *Correlation*

The table (Tab. VII) shows the R-values coming from linear regression between the four evaluation metrics and the evaluation by the radiologist. The table (Tab. VIII), shows the R-values for the linear regression between the combinations of evaluation metrics and the radiologist's evaluation. Only the proportion with the highest R-Value is shown in for each combination. Additionally, figure 2 shows the plot of the linear regression.

Evaluation Metric	**R-value**
DICE	0.632845426
HD	-0.781606441
AHD	-0.73627801
AUC	-0.284132072

TABLE VII: Correlation of radiologist's evaluation and evaluation metrics.

Metric$_1$	**Metric$_2$**	w_1	w_2	**R-value**
DICE	**HD**	0.6	0.4	0.824894584
DICE	**AHD**	0	1	0.736277907
AHD	**HD**	0.5	0.5	-0.811374338

TABLE VIII: Highest correlations of radiologist's evaluation and combined evaluation metrics.

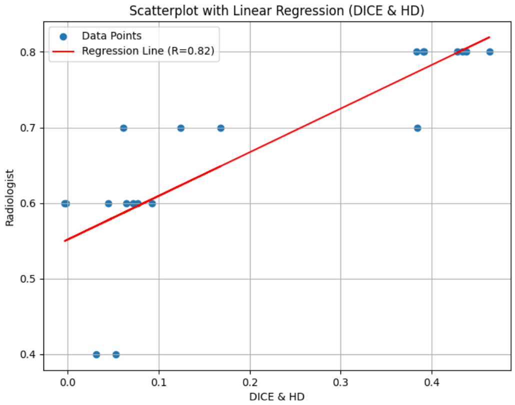

Fig. 2: Linear Regression plot of best combined evaluation metric (60% DICE - 40% HD) and radiologist's evaluation.

IV. Discussion

As recommended by Maier-Hain et al. [4] for semantic segmentation, complementary metrics for overlap and boundary distance measurement were used for brain MR image segmentation. Namely DSC, HD, AHD, and ROC/AUC were applied. While ROC/AUC initially seemed to be promising to obtain an intuition, whether the random forest algorithm is over-/under-segmenting, it turned out that ROC/AUC takes the dominantly large background into consideration. In this case ROC/AUC consistently leads to an overoptimistic and not clinically useful evaluation, as shown in table II. For this reason ROC/AUC is excluded as evaluation metric in further evaluation discussions in this report. While DSC as overlap based metric is known to be more adequate for large structures, HD and AHD are boundary based metrics that are sensitive for small structures. As the results show in figures I, III and IV, DSC indicates better results than HD and AHD for all segmented brain regions. This outcome may be explained by the fact that AHD and, especially HD, strongly penalize incorrectly predicted boundary regions. These errors occur in the visually noticeable, noisy, and rough segmentation boundaries of small structures. Furthermore, AHD proved to be more robust compared to HD for the TH segmentation, which was over-segmented and had dispersed voxels into left and right hemispheres. In figure 1 on the top left and bottom right image, this effect can be identified for the magenta TH segmentation. Using DSC and HD as a combined metric, as described in the methods section III-D, allowed to best reflect the human perception represented by the radiologist. For this combined metric, the R-value of 0.825 (Tab. VIII) suggests a strong relationship between the radiologist's rating and the segmentation scoring with 60% DSC and 40% HD. In other words, in this case, the random forest based segmentation aligns well with the radiologist's assessment. This supports the assumption of obtaining a more comprehensive assessment of brain MR image segmentation quality compared to a single metric. Given time constraints with the radiologist allowed to evaluate only four patient samples. Volume characteristics (shown in table VI) suggests that GM and WM are complex and large volumes, while HC, AMG, and TH are comparably simple and small volumes. Comparing the results from II-A1 with the radiologist's assessment, the random forest segmentation performed worse on smaller structures like HC and AMG. In contrast, it performed better on larger structures such as WM and GM.

V. Conclusion

In summary, it has been shown that combining two evaluation metrics for brain MR images enhances insight and understanding compared to using a single metric. While DSC provides the highest numerical value, it may not be the most suitable method for tubular structures like the HC. Even if the contours are identical, a parallel shift can result in zero overlap. Evaluating with HD and AHD provides better understanding on the smoothness of the segmentation. For further evaluation, applying morphological filters such as opening or closing during post-processing could help reduce noise and smooth the contours. This may improve results, particularly when measuring the TH with HD and AHD. Finally, combining DSC and HD with a 60-to-40 weighting shows a strong correlation with the radiologist's perception. At a next step it might be worth to test for non-linearity e.g. with logistic-regression. Additionally, to have a more robust assessment, it would be beneficial to increase the sample size of the radiologist's evaluation.

References

[1] E. Konukoglu and B. Glocker, "Chapter 19 - Random forests in medical image computing," in *Handbook of Medical Image Computing and Computer Assisted Intervention*, ser. The Elsevier and MICCAI Society Book Series, S. K. Zhou, D. Rueckert, and G. Fichtinger, Eds., Academic Press, Jan. 2020, pp. 457–480, ISBN: 9780128161760. DOI: 10.1016/B978-0-12-816176-0.00024-7. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128161760000247 (visited on 12/11/2024).

[2] D. Hartmann, D. Müller, I. Soto-Rey, and F. Kramer, *Assessing the Role of Random Forests in Medical Image Segmentation*, arXiv:2103.16492, Mar. 2021. DOI: 10.48550/arXiv.2103.16492. [Online]. Available: http://arxiv.org/abs/2103.16492 (visited on 12/11/2024).

[3] A. Aziz Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool," vol. 15, no. 29, Aug. 2015. DOI: DOI10.1186/s12880-015-0068-x. [Online]. Available: https://bmcmedimaging.biomedcentral.com/articles/10.1186/s12880-015-0068-x (visited on 12/05/2024).

[4] L. Maier-Hein, A. Reinke, *et al.*, "Metrics reloaded: Recommendations for image analysis validation," en, *Nature Methods*, vol. 21, no. 2, pp. 195–212, Feb. 2024, ISSN: 1548-7091, 1548-7105. DOI: 10.1038/s41592-023-02151-z. [Online]. Available: https://www.nature.com/articles/s41592-023-02151-z (visited on 12/26/2024).

[5] O. U. Aydin, A. A. Taha, *et al.*, "On the usage of average Hausdorff distance for segmentation performance assessment: Hidden error when used for ranking," *European Radiology Experimental*, vol. 5, no. 1, p. 4, Jan. 2021, ISSN: 2509-9280. DOI: 10.1186/s41747-020-00200-2. [Online]. Available: https://doi.org/10.1186/s41747-020-00200-2 (visited on 12/08/2024).

[6] A. Smith and V. M. Zavala, "The Euler characteristic: A general topological descriptor for complex data," *Computers & Chemical Engineering*, vol. 154, p. 107463, Nov. 2021, ISSN: 0098-1354. DOI: 10.1016/j.compchemeng.2021.107463. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0098135421002416 (visited on 12/28/2024).

[7] S. Knutsson, "Particle Shape Quantities and Measurement Tecniques - A review," en, *Electronic Journal of Geotechnical Engineering*, [Online]. Available: https://www.academia.edu/30653826/Particle_Shape_Quantities_and_Measurement_Tecniques_A_review (visited on 12/28/2024).

[8] *ITK-SNAP Home*. [Online]. Available: http://www.itksnap.org/pmwiki/pmwiki.php (visited on 12/30/2024).

[9] *Normalization*, en. [Online]. Available: https://www.codecademy.com/article/normalization (visited on 12/28/2024).

[10] *Averaging Methods - Testing with Kolena*, en. [Online]. Available: https://docs.kolena.com/metrics/averaging-methods/ (visited on 12/28/2024).

Appendix

A. Reference to GitHub repository

https://github.com/fabriciokirchhofer/mialab.git

B. Linear Regression plots

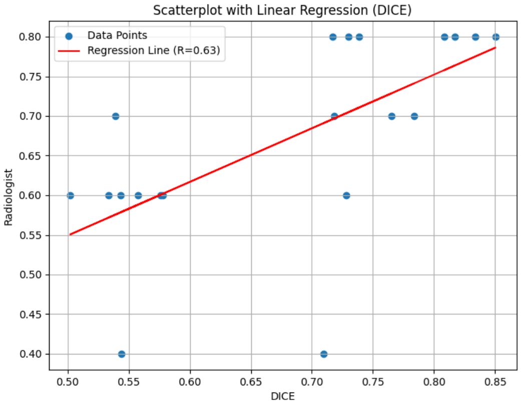

Fig. 3: Linear Regression plot of DICE and radiologist's evaluation. Refer to table VII for exact R-value.

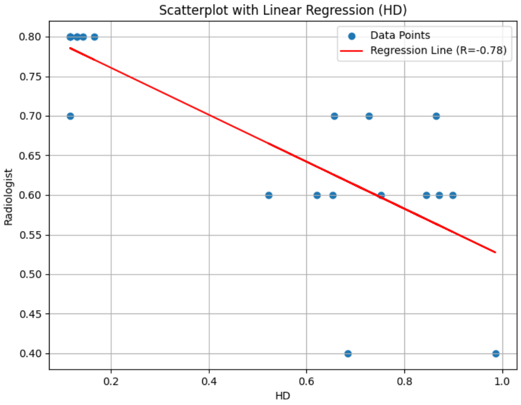

Fig. 4: Linear Regression plot of HD and radiologist's evaluation. Refer to table VII for exact R-value.

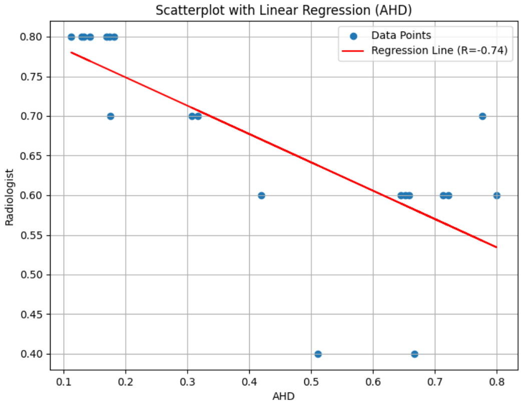

Fig. 5: Linear Regression plot of AHD and radiologist's evaluation. Refer to table VII for exact R-value.

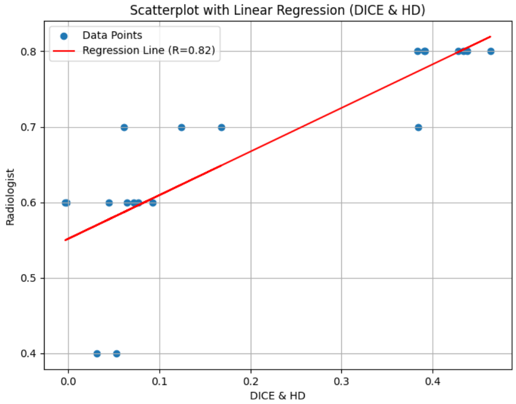

Fig. 6: Linear Regression plot of AUC and radiologist's evaluation. Refer to table VII for exact R-value.

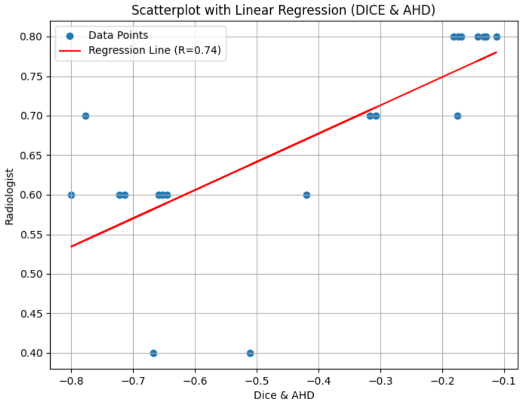

Fig. 7: Linear Regression plot of combined evaluation metric (0% DICE - 100% AHD) and radiologist's evaluation. Refer to table VIII for exact R-value.

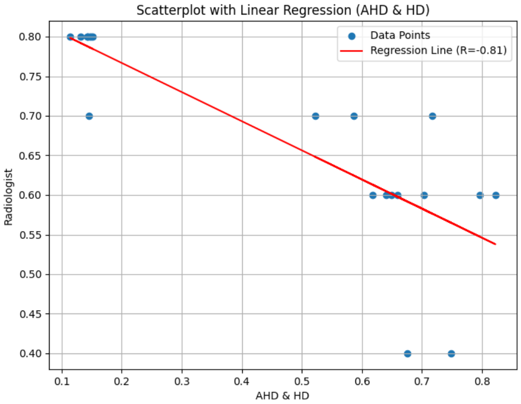

Fig. 8: Linear Regression plot of combined evaluation metric (50% AHD - 50% HD) and radiologist's evaluation. Refer to table VIII for exact R-value.