Quantifying Inter-Operator Variability and its Causes for Medical Semantic Segmentation

Anonymous ICCV submission

Paper ID 9

Abstract

Uncertainty quantification is a vital component of the development of models for medical semantic segmentation, as we are increasingly faced with the question "when is 'good' good enough?". A necessary part of the answer to this will always involve understanding the variability of acceptable segmentation results, which in turn requires the inter-operator variability to be disclosed. This is a stumbling block for many, as quantifying inter-operator variability is expensive and time consuming - especially within medical imaging. In this work we not only highlight how critical understanding this variability is, but also provide a novel framework to better understand and predict regions of variability by considering the texture of the scan. The source code will be released upon publication.

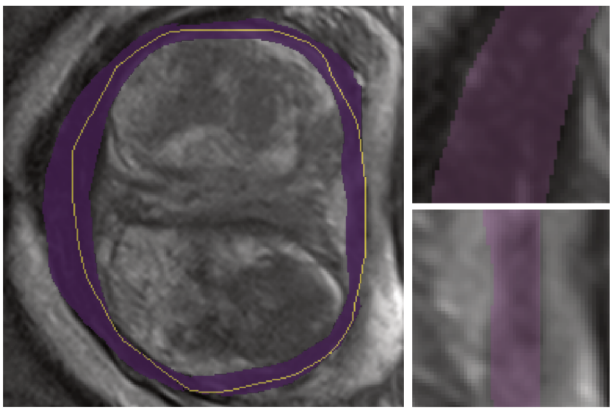

Figure 1. *Left:* An example segmentation (yellow) within the area of other clinician segmentations (purple). *Top Right:* A magnified region with high variability. *Bottom Right:* A magnified region with low variability.

1. Introduction

1.1. Semantic Segmentation

In computer vision, semantic segmentation is the task of determining which pixels in a given image belong to a target class. Within medical imaging the output of semantic segmentation can be used to take measurements [12], detect abnormalities [16], and construct virtual copies of an object [11]. However, ground truth segmentations are often time consuming (and therefore expensive) for a clinician to produce manually [13]. This has necessitated the development of segmentation models to reduce the amount of clinician input needed to generate segmentations [2]. A key difference between how a clinician produces a segmentation versus a standard segmentation model is the use of the object boundary. A clinician will define the object by its boundary; for them it is clear which spaces are in the object and which are not. In contrast, most models will define the object as a collection of pixels, with no regard for which ones form the boundary [17].

Inter-operator variability is an accepted issue across semantic segmentation, whilst also providing a helpful margin of acceptable error for models. Despite the importance of annotator variability, relatively few publicly available datasets include variability statistics in their documentation. When available, variability is often stated as a single measurement (e.g. 0.05 DSC), with no further information on how certain parts of a segmentation may be prone to higher variability than others.

This paper has three aims:

1. To highlight how prevalent inter-operator variability is within semantic segmentation data.
2. To demonstrate how the distribution of this variability is often uneven.
3. To provide a framework for quantifying both the variability and its causes within a single image.

Predicting variability via models is a well-explored area; however, these methods often struggle when trying to validate their results against actual recorded uncertainty. The ideal method to validate such a model would be to compare it against the variability between a group of annotators; unfortunately, this is frequently unrealistic due to the cost of annotation. Instead, if we could understand the causes of

inter-operator variability then we may be able to assess the performance of these models without the need for data fully annotated by multiple clinicians.

1.2. Inter-Operator Variability

Inter-operator variability refers to the variability between different annotators; in contrast to intra-operator variability, which refers to the variability within the segmentations done by the same annotator. Reporting inter-operator variability is particularly important if the dataset contains segmentations drawn by multiple distinct annotators, as 'errors' by the model may actually just be differences in annotator style.

As a basis for this work we considered the 18 semantic segmentation challenges run in the MICCAI 2023 conference, focusing on the number of annotators providing the segmentations and the reported inter-operator variability. This conference was chosen not only because it had a large number of semantic segmentation challenges for medical imaging, but the high-profile nature of the conference meant that the datasets used received a lot of publicity, with the prize money totaling in excess of $40K.

All of the challenges had a single annotation per image, but many had split these between multiple annotators. Of the 18 challenges, only 2 had a single clinician providing all annotations and more than half of the challenges had 5 or more clinicians providing annotations [3]. In the challenges where the exact number of annotators was not given, it was made clear that there were still multiple annotators. This means that we can expect inter-operator variability to be present in at least 16 of the challenges.

Despite this, only 6 of these 16 challenges quantified the annotator uncertainty in some way, and of those the uncertainty was often described in a vague or unhelpful manner. For example, the autoPET 2023 challenge only describes the uncertainty in defining lesion boundaries by saying *"The difference in segmentation volumes can range from 5-30%"* [7].

The datasets used in these challenges will likely go on to be used in many further studies, yet most provide no warning that multiple annotators were used or state the likely variability in the ground truths. As it would be unrealistic to expect each image to be annotated by multiple clinicians, we will instead provide a method to allow single annotations to reflect the uncertainty in the boundary.

2. Existing Work

2.1. Predicting Uncertainty

Previous work in predicting which areas of a ground truth are likely to be the most uncertain often involves methods such as training a model multiple times with different parameters, such as MC Dropout which uses different dropout conditions [5]. The variance in the results on the same image can then be used as a prediction for the uncertainty of that model. Another example is CoraNet [15], in which the model is trained with a conservative setting and a radical setting to determine which regions are the most likely to be uncertain.

Other methods consider that different annotators will have different styles and that some annotators may be more senior and therefore should be considered 'more correct'. An example of this can be seen in the work by Ji et al. [9], which looked to calibrate multiple annotations with expertise levels in order to produce a model that could emulate an expert annotator.

2.2. Annotation Reliability

The standard way to increase the reliability of an annotation is to have the annotation repeated, either by the same clinician (to increase intra-operator reliability) or by multiple clinicians (to increase inter-operator reliability). The annotations are then often combined into a single annotation, by methods such as majority voting [6] or the STAPLE algorithm, which forms a probabilistic estimate of the 'true' ground truth and weights the annotator segmentations accordingly [18].

3. Method

3.1. Data

In order to properly assess the relationship between variability, noise, and segmentation regions, we searched for publicly available semantic segmentation data with at least three annotations per image. The resulting datasets were then reduced to those of universal physiological features (i.e. not lesions, tumors, or other features that have no set location). The reason for this was that we wanted to focus on tasks where the target segmentation had a fixed known location, and therefore have neighboring structures in the same place in each image.

This search resulted in just three datasets: the QUBIQ Prostate dataset, the RIGA fundus dataset, and the Grey Matter (GM) Spinal dataset. Each of these datasets has two tasks per image, resulting in six unique tasks overall. Examples from each dataset can be seen in Figure 2.

3.1.1. QUBIQ Prostate

The QUBIQ Prostate data is a dataset of prostate MRI images, made available for the QUBIQ challenge [10]. It contains 48 2D MRI slices, where each slice has been annotated for two tasks by an expert clinician. The challenge does not describe what the two tasks are segmenting, but it appears that Task 1 is the total prostate volume and Task 2 is the transitional zone volume [19].

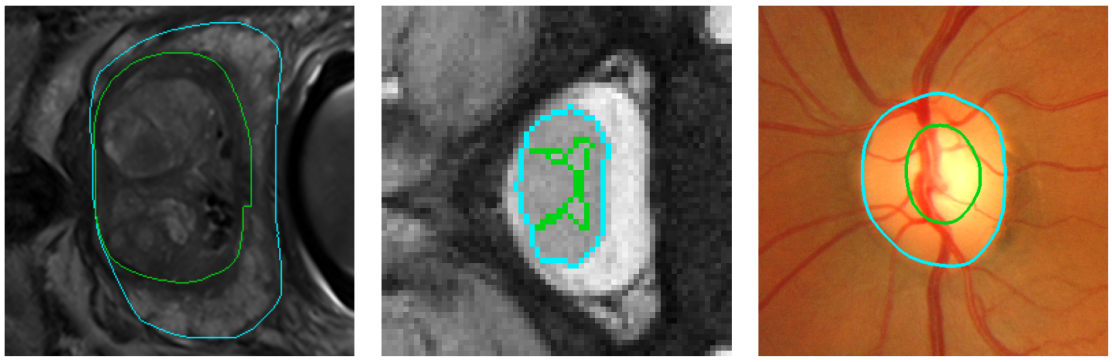

Figure 2. Example segmentations for both tasks for each of: QUBIQ Prostate (Left), GM Spinal (Centre), RIGA (Right)

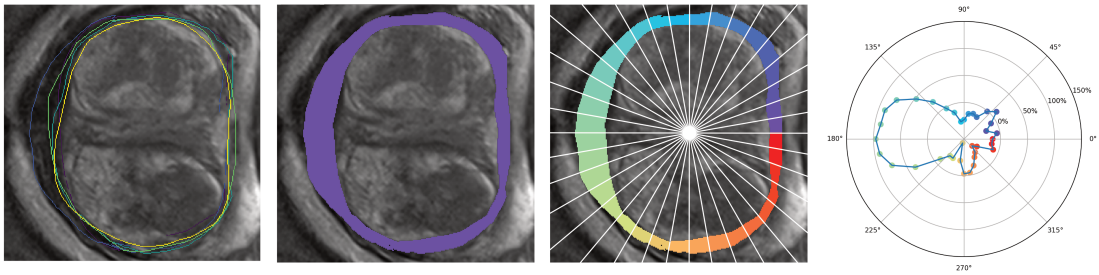

Figure 3. *Far Left:* A scan from the second QUBIQ Prostate task with multiple annotations. *Center Left:* The area disputed between annotators. *Center Right:* The disputed area split into 36 boundary regions. *Center Right:* The corresponding graph showing the change in disputed area width, as described in Section 3.4.1.

3.1.2. RIGA

The RIGA dataset contains 750 fundus images from three different sources [1]. Each fundus image has the optic disc (Task 1) and optic cup (Task 2) annotated by six experienced ophthalmologists. Of the three sources (MESSIDOR, Bin Rushed, and Magrabi), the MESSIDOR subset has 460 images, the Bin Rushed set has 195, and the Magrabi set has 95. The Bin Rushed images are difficult to use due to the ground truths being compressed on top of the fundus images as single JPEG images, so we removed this subset from the analysis.

3.1.3. GM Spinal

The GM Spinal Cord Segmentation Challenge data was released in 2017, containing 3D MRI data for 80 patients from 4 sites, resulting in a total of 407 2D slices [14]. Each slice was annotated by four expert raters, who manually segmented the white matter (Task 1) and gray matter (Task 2). The software used to assist the annotations varied between annotators, and is described in the GM Spinal publication.

3.1.4. Pre Processing

Each image was converted to 256 levels, with values normalized within this range. No other pre-processing was applied.

3.2. Disputed and Agreed Areas

In order to assess how annotator variability changes across a segmentation, we first need to establish which parts of the segmentation are contributing towards the variability.

For a set of annotations A we consider the *agreed segmentation*, $\texttt{AS}(A)$ and the *disputed segmentation* $\texttt{DS}(A)$.

The agreed area is the intersection of all annotations:

$$\texttt{AS}(A) = \bigcap_{a_i \in A} a_i$$

The disputed area is the area that is included by at least one annotator, but not all annotators:

$$\texttt{DS} = \bigcup_{a_i \in A} a_i - \texttt{AS}(A)$$

Finally, we will also refer to the *total segmentation* $\texttt{TS}(A)$, which is the union of the disputed and agreed areas.

An example of the disputed area can be seen in Figure 3.

3.3. Boundary Regions

In order to compare variability between the same areas, we split the disputed area into n non-overlapping regions. The

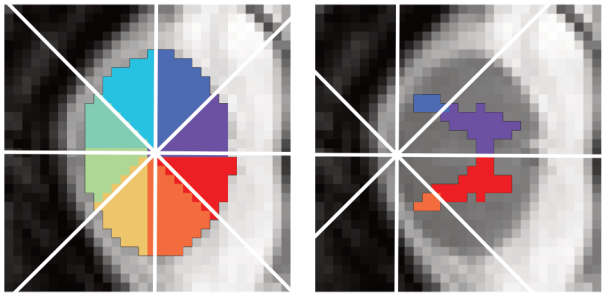

Figure 4. *Left:* An example of 8 regions for the GM Spinal Task 1. *Right:* An example of 8 regions for the GM Spinal Task 2, with centroid aligned to the left to allow more meaningful separation.

exact value of n is determined using the minimum boundary length across the entire dataset, with the constraint that each region must always contain at least one pixel.

The region each pixel p belongs to is determined by the angle between p and the centroid c of the agreed area for all tasks except Task 2 of the GM Spinal data. In this task, the concave shape of the gray matter meant that more meaningful regions were created by instead using a left aligned centroid, as seen in Figure 4.

For given n and c the regions of $\texttt{disp}(A)$ are then defined as:

$$\mathtt{R}_{0\leq i<n} = \{\, p \in \texttt{disp}(A) \;\mid\; i \leq \frac{1}{n} \cdot tan^{-1}\frac{p_y - c_y}{p_x - c_x} < i+1 \,\}$$

An example of this for 36 regions can be seen in Figure 3.

This method has some weaknesses; regions with a longer medial axis (i.e. higher irregularity) are greater represented then those with a shorter medial axis, and if the medial axis is not convex then multiple disconnected segments of the axis could potentially be included in the same region. A more nuanced method would be an improvement for future work.

For an object with genus 0 the disputed area almost always appears as a band around the boundary of the object, as the topology of the object is generally not disputed. An exception to this is seen in the GM Spinal data, where the segmentations are so small that the disputed area is often only 1 pixel wide, and interrupted by sections with no disputed area.

In the simple case where the disputed area is an uninterrupted band, we can measure the width by iterating across the medial axis of the disputed area: the width is calculated as the shorted distance to the boundary of the disputed area. In cases where the medial axis is a set of disconnect segments, we instead join them with the corresponding sections of the boundary of the agreed area (each of these pixels has width 0).

Using these regions, we can investigate whether particular regions of a segmentation have consistently higher inter-operator variability or changes in texture.

3.4. Inter-Operator Variability

3.4.1. Disputed Area Width

To measure the average width of the disputed area for a region we consider the width for each point $p \in \mathtt{RM}_i$, where $\mathtt{RM}_i$ is the intersection of region $\mathtt{R}_i$ and the medial axis. The width at p is the shortest distance to the boundary of the disputed area.

Rather than the absolute width, we instead measure the percentage change between the width w_p at point p and the mean width $\bar{w}$. The mean change in width ΔW for region $\mathtt{R}_i$ is then:

$$\Delta W\,\mathtt{R}_i = \frac{100}{\mid \mathtt{RM}_i \mid} \sum_{p \in \mathtt{RM}_i} w_p - \bar{w}$$

A region with a high change in width will have a higher inter-operator variability than a region in the same image with a lower change in width.

3.4.2. Dice Similarity Coefficient

The second method we will use to measure the inter-operator variability between regions is the change in the Dice Similarity Coefficient (DSC) [4]. DSC is commonly used in medical imaging semantic segmentation as a way to measure the similarity of two segmentations, and as such is also often used to record inter-operator variability.

The value of DSC for a region $\mathtt{R}_i$ measures the similarity between the total segmentation $\mathtt{TS}$ and the agreed segmentation $\mathtt{AS}$:

$$\mathtt{TS}_{R_i} = \mathtt{R}_i \cap \mathtt{TS} \quad (1)$$

$$\mathtt{AS}_{R_i} = \mathtt{R}_i \cap \mathtt{AS} \quad (2)$$

$$DSC(\mathtt{R}_i) = \frac{2 \mid \mathtt{AS}_{R_i} \mid}{\mid \mathtt{AS}_{R_i} \mid + \mid \mathtt{TS}_{R_i} \mid} \quad (3)$$

We then evaluate ΔDSC as the percentage difference against the value of DSC for all regions ($DSC(\cup \mathtt{R})$):

$$\Delta DSC(\mathtt{R}_i) = \frac{100\, DSC(\mathtt{R}_i)}{DSC(\cup \mathtt{R})}$$

3.5. Texture

The aim for this paper is to compare inter-operator variability to a textures measures that could be obtained from a dataset with just one annotation per image. We therefore

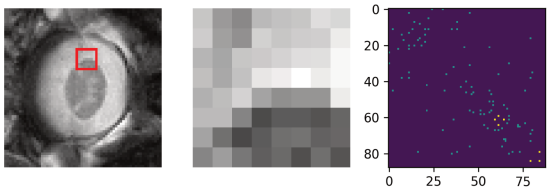

Figure 5. *Left:* A GM Spinal image with patch in red. *Center:* The magnified patch. *Right:* The corresponding patch GLCM.

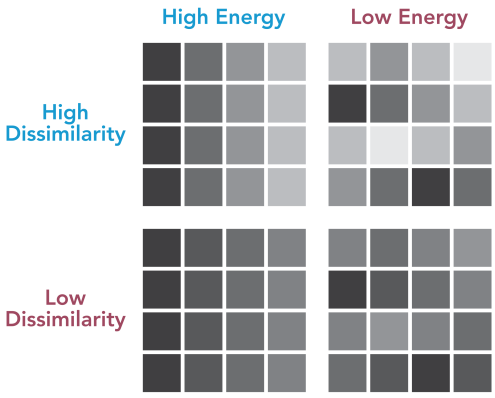

Figure 6. Example patches showing combinations of high and low energy and dissimilarity, for a GLCM with comparison distance of 1 pixel and horizontal direction. Within each patch the difference in intensity between each pixel and its horizontal neighbour is the same.

measure the texture for each point p along the boundary of each individual segmentation for the six tasks.

For each point p on the segmentation boundaries we assess the texture using a Gray Level Co-occurrence Matrix (GLCM) for a patch P of size m centered around p [8]. The value of m is dependent on the smallest segmentation area in the dataset for that task. The GLCM identifies the frequency that each intensity level occurs within the patch with a distance of d and along an axis of θ degrees. This matrix is then used to compute the energy and dissimilarity values.

The value of θ for a point p is decided using the angle of p from the centroid c, as described in Section 3.3. This approximately gives a value of θ that is perpendicular to the boundary, in order to identify any clear boundaries in the original scan. An example patch and corresponding GLCM matrix can be seen in Figure 5.

3.5.1. Energy

The energy of the GLCM measures how orderly the pixel value differences are within the patch [8]. A GLCM with high energy will have pixels of similar intensities close to each other, whereas one with low energy will have pixels of different intensities 'mixed' with each other. For a point p with GLCM matrix P, the energy is then:

$$\mathrm{e}(p) = \sum_{i,j=0}^{levels-1} P_{i,j}(-log(P_{i,j})) \tag{297}$$

For a segmentation boundary region R_i, the mean change in energy is calculated against the mean segmentation energy $\bar{\mathrm{e}}$:

$$\Delta E\,\mathrm{R}_i = \frac{100}{|\,\mathrm{R}_i\,|} \sum_{p \in \mathrm{R}_i} \mathrm{e}_p - \bar{\mathrm{e}} \tag{301}$$

Examples of patches with low versus high energy can be seen in the rows of Figure 6.

3.5.2. Dissimilarity

Dissimilarity is a measure of contrast within the patch [8]. A GLCM with high dissimilarity will have low contrast, regardless of the order of the pixels or the absolute values compared to the rest of the image. For a point p with GLCM matrix P, the dissimilarity is calculated as:

$$\mathrm{d}(p) = \sum_{i,j=0}^{N-1} P_{i,j}\,|\,i-j\,| \tag{310}$$

Similarly, for a segmentation boundary region R_i the mean change in dissimilarity is calculated against the mean segmentation dissimilarity $\bar{\mathrm{d}}$:

$$\Delta D\,\mathrm{R}_i = \frac{100}{|\,\mathrm{R}_i\,|} \sum_{p \in \mathrm{R}_i} \mathrm{d}_p - \bar{\mathrm{d}} \tag{314}$$

Examples of patches with low versus high dissimilarity can be seen in the columns of Figure 6. In these examples we can also see how the combination of energy and dissimilarity affects how easily a boundary can be identified. Patches with a continuous edge that runs perpendicular to θ will have a higher energy, and this edge will be easier to detect if the dissimilarity is higher. In areas with low dissimilarity any boundaries will be much harder to make out – an indicator that the patch is monotone, with no clear boundaries, is high energy with low dissimilarity.

4. Results

4.1. Overall Variability

The mean size and standard deviation for the total segmentation area can be seen in Table 1, which shows the substantial difference in size between the QUBIQ and RIGA tasks versus the GM Spinal tasks. The RIGA tasks also have a much larger standard deviation of the total segmentation area, corroborating the need for variation measures that do not use absolute pixel values.

	QUBIQ Prostate		RIGA		GM Spinal	
	Task 1	Task 2	Task 1	Task 2	Task 1	Task 2
Mean	44075.4	28901.3	52452.7	18523.8	990.5	527.2
S.D.	14339.8	14185.5	66009.5	25454.6	194.7	93.5

Table 1. Total segmentation area in pixels (mean and standard deviation) for all tasks.

	QUBIQ Prostate		RIGA		GM Spinal	
	Task 1	Task 2	Task 1	Task 2	Task 1	Task 2
Mean	20.0%	39.0%	17.3%	53.6%	13.8%	45.7%
S.D.	4.4%	13.8%	5.5%	15.5%	3.7%	6.6%

Table 2. Proportion of disputed area (mean and standard deviation) for all tasks.

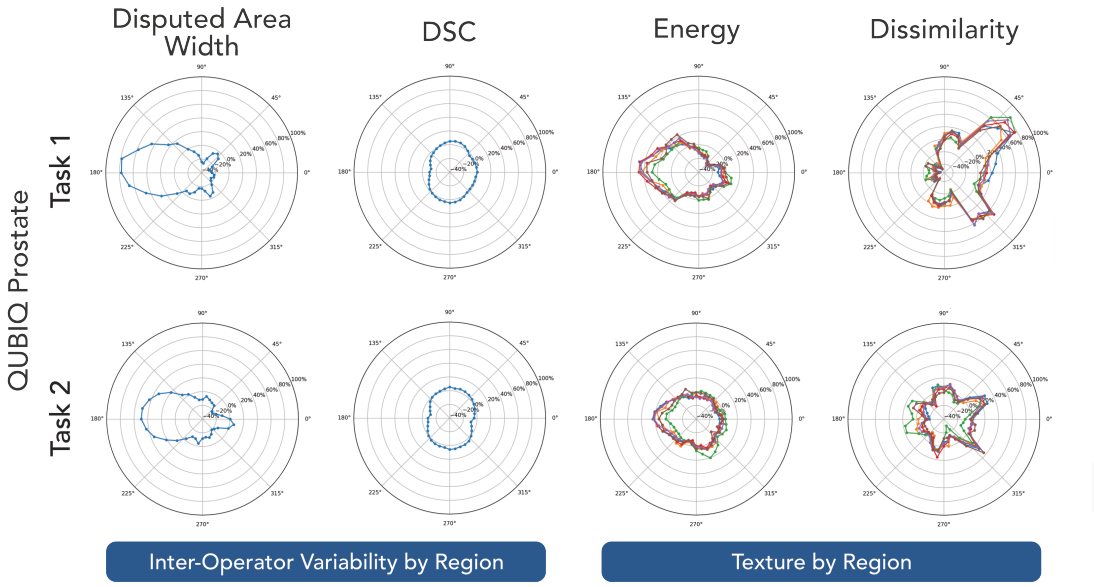

Figure 7. Inter-operator variability and texture measures by region for the QUBIQ Prostate Tasks

Table 2 then shows the overall size of the disputed area as a proportion of the total segmentation for all six tasks. Although this measurement does not show how variability changes across regions of a segmentation, we can see that across all six tasks the disputed area was a significant proportion of the total segmentation.

An interesting observation is that the second task for each of the three datasets had a notable increase in the proportion of the disputed area. The second feature for these datasets was always entirely contained within the first task (and therefore smaller). The cause for this increase in variability could theoretically be because annotators made segmentations for both tasks at the same time, and did not zoom in on the second smaller task.

Finally, Table 3 shows the DSC values between the agreed and total segmentation areas for all tasks. The observations here are similar to that of table 2, with the second tasks having notably lower DSC scores.

4.2. Variability and Texture by Region

For each of the three datasets, we report the disputed area width, DSC, energy, and dissimilarity as a mean value for each region. These results can be seen in Figures 7, 8, and 9 as polar graphs, with the angular coordinate representing the angle of each region from the centroid and the radial coordinate representing the measured value.

4.2.1. QUBIQ Prostate Data

The results for the QUBIQ Prostate data can be seen in Figure 7. For both tasks there is a clear increase in variability in the center left regions, with an average increase in width of up to 80% in Task 1 and 50% in Task 2. This corresponds to a decrease in region DSC of up to -10% in both cases (as Task 1 is of a larger volume).

	QUBIQ Prostate		RIGA		GM Spinal	
	Task 1	Task 2	Task 1	Task 2	Task 1	Task 2
Mean	0.882	0.792	0.899	0.608	0.925	0.705
S.D.	0.029	0.136	0.036	0.154	0.021	0.055

Table 3. DSC between agreed area and total segmentation area (mean and standard deviation) for all tasks.

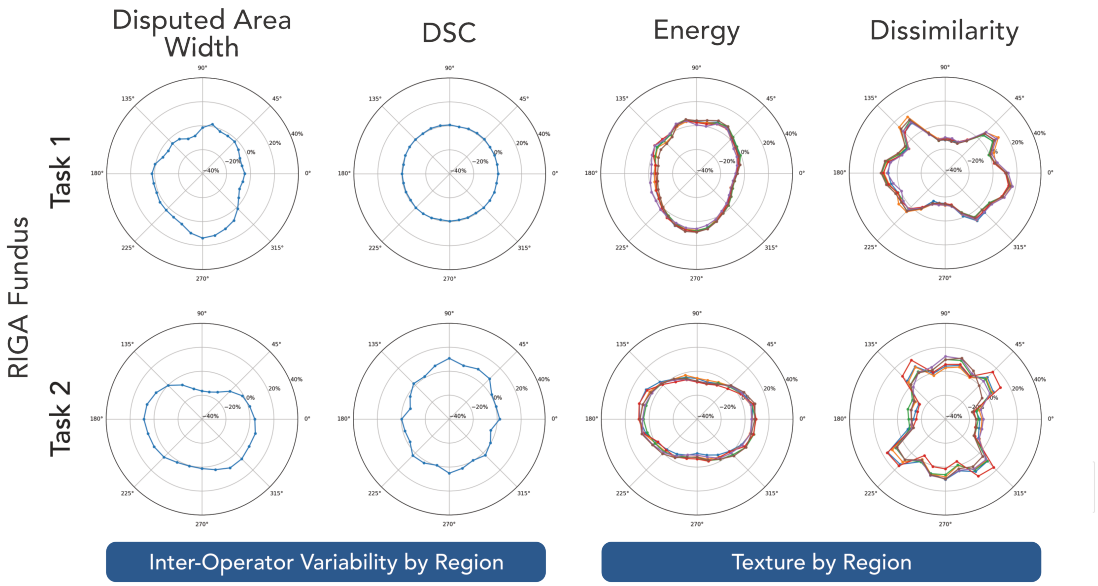

Figure 8. Inter-operator variability and texture measures by region for the RIGA Fundus Tasks

For both tasks, the increase in width in the center right regions corresponds to a similar increase in energy for the same regions, as well as a decrease in dissimilarity. As mentioned in Section 3.5.2 the combination of high energy with low dissimilarity indicates a monotone region, which appears to be the case here.

4.2.2. RIGA Fundus Data

The results for the RIGA fundus data can be seen in Figure 8. As with the QUBIQ data a similar magnitude of difference in width (+15% for Task 1 and -18% in Task 2) results in a much greater effect in region DSC for the task with a smaller overall volume.

A contributing factor to variability that is not considered here is blood vessels obscuring the segmentation boundaries. Though the blood vessels often run perpendicular to the boundary, they may still affect the texture measurements.

Task 1 The optic disc task has a region of high energy and low dissimilarity at the center bottom region, corresponding to a higher disputed area width. Though this area is less monotone than in the QUBIQ data, it once again correlates to an increase in variability.

Task 2 Due to the smaller volume of the optic cup, any blood vessels present will take up a larger proportion of the volume than with the optic disc. Despite the symmetry of the texture measurements the variability is much more asymmetric, with a notable decrease in variability in the top center segment.

4.2.3. GM Spinal Data

The results for the GM Spinal data can be seen in Figure 9.This data was the most difficult to analyze due to the relative size of the segmentations and the complex 'butterfly' shape of the second task. The width of the disputed area was rarely greater than one pixel for both tasks, causing any slight variations to have a disproportionate affect on the results.

Task 1 The first task had relatively low variance per region in both width and DSC, with a slight increase in width for the right hand side and spikes of around +10% at the very top and bottom. The energy of the boundary was consistent for all annotators; however there were notable changes in dissimilarity in the top and bottom quadrants. These changes (ranging from -20% to +10%) occurred where the boundary becomes the most convex, which may have caused a lack in visible features for annotators to follow.

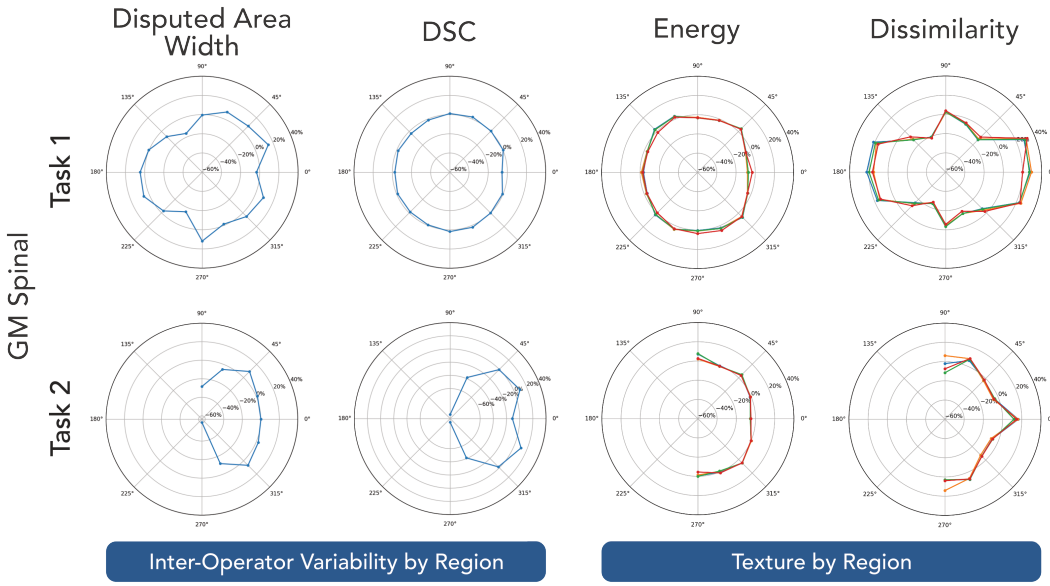

Figure 9. Inter-operator variability and texture measures by region for the GM Spinal Tasks

Task 2 Due to its complex shape, the centroid of this task was instead placed to the left of the actual centroid. The width, DSC, energy, and dissimilarity were all relatively constant for the majority of the segmentation, with the exception of the long and thin dorsal horns. Counter-intuitively both the width and the DSC were significantly lower here, though the DSC much more so. The lack of variance of energy or dissimilarity in these regions suggests that the inter-operator variability here was not caused by the texture of the image; instead, this could potentially be due to the geometric complexity and thinness of the object.

5. Conclusion

When evaluating a model against a dataset it is vital to know the possible variability – with disputed areas being up to 50% of the total segmentation, the choice of annotator could have a significant impact on the measured accuracy of the model. For medical imaging this variability can also help provide the range of potential measurements that can be taken from a segmentation.

This work has compared multiple measures of inter-operator variability for segmentation regions for six tasks across three datasets. From this analysis we can conclude that certain regions may have consistently higher variability, and that for circular segmentations disputed area width is a more sensitive method than DSC for identifying these variations.

We have also compared these measures of inter-operator variability to texture measures of energy and dissimilarity for individual annotations, with the aim of providing a substantiated explanation for the uncertainty.

From this we have found that regions of high energy and low dissimilarity frequently correlate with an increase in inter-operator variability. These regions will be more monotonic than other regions in the image, providing no clear boundary for the annotator to follow. Similarly regions with low energy and high dissimilarity, which appear disordered and noisy, also correlate with regions with higher inter-operator variability.

A limitation of this work is the lack of publicly available datasets to evaluate. While inter-operator availability is described for privately available datasets it is often done so as a single value, with no distinction for regions of the feature. By reporting variability with an approach such as the one described here much more nuance is obtained, without the need to make the dataset publicly available.

An area that would benefit from further study is the relationship between the geometry and the variability of a task. We theorized that this could be a contributor to the uncertainty in the second GM Spinal task, however further analysis would be needed to confirm this.

6. Compliance with Ethical Standards

This study was conducted retrospectively using ethically acquired publicly available human subject data. The authors have no interests to disclose.

References

[1] Ahmed Almazroa, Sami Alodhayb, Essameldin Osman, Eslam Ramadan, Mohammed Hummadi, Mohammed Dlaim, Muhannad Alkatee, Kaamran Raahemifar, and Vasudevan Lakshminarayanan. Retinal fundus images for glaucoma analysis: the riga dataset. In *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*, pages 55–62. SPIE, 2018. 3

[2] Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. Medical image segmentation review: The success of u-net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[3] The Medical Image Computing and Computer Assisted Intervention Society. Miccai registered challenges. https://miccai.org/index.php/special-interest-groups/challenges/miccai-registered-challenges/, 2025. Accessed: 2025-01-20. 2

[4] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945. 4

[5] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. 2

[6] Mudasir A Ganaie, Minghui Hu, Ashwani Kumar Malik, Muhammad Tanveer, and Ponnuthurai N Suganthan. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151, 2022. 2

[7] Sergios Gatidis, Thomas Kustner, Michael Ingrisch, Clemens Cyranand, and Jens Kleesiek. Automated lesion segmentation in whole-body fdg-pet/ct - domain generalization. https://zenodo.org/records/7845727, 2023. Accessed: 2025-02-20. 2

[8] Mryka Hall-Beyer. Glcm texture: A tutorial v. 3.0 march 2017. *University of Calgary*, 2017. 5

[9] Wei Ji, Shuang Yu, Junde Wu, Kai Ma, Cheng Bian, Qi Bi, Jingjing Li, Hanruo Liu, Li Cheng, and Yefeng Zheng. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12341–12351, 2021. 2

[10] Hongwei Bran Li, Fernando Navarro, Ivan Ezhov, Amirhossein Bayat, Dhritiman Das, Florian Kofler, Suprosanna Shit, Diana Waldmannstetter, Johannes C Paetzold, Xiaobin Hu, et al. Qubiq: Uncertainty quantification for biomedical image segmentation challenge. *arXiv preprint arXiv:2405.18435*, 2024. 2

[11] Marco Mandolini, Agnese Brunzini, Giulia Facco, Alida Mazzoli, Archimede Forcellese, and Antonio Gigante. Comparison of three 3d segmentation software tools for hip surgical planning. *Sensors*, 22(14):5242, 2022. 1

[12] Caroline Petitjean, Maria A Zuluaga, Wenjia Bai, Jean-Nicolas Dacher, Damien Grosgeorge, Jérôme Caudron, Su Ruan, Ismail Ben Ayed, M Jorge Cardoso, Hsiang-Chou Chen, et al. Right ventricle segmentation from cardiac mri: a collation study. *Medical image analysis*, 19(1):187–202, 2015. 1

[13] Dzung L Pham, Chenyang Xu, and Jerry L Prince. Current methods in medical image segmentation. *Annual review of biomedical engineering*, 2(1):315–337, 2000. 1

[14] Ferran Prados, John Ashburner, Claudia Blaiotta, Tom Brosch, Julio Carballido-Gamio, Manuel Jorge Cardoso, Benjamin N Conrad, Esha Datta, Gergely Dávid, Benjamin De Leener, et al. Spinal cord grey matter segmentation challenge. *Neuroimage*, 152:312–329, 2017. 3

[15] Yinghuan Shi, Jian Zhang, Tong Ling, Jiwen Lu, Yefeng Zheng, Qian Yu, Lei Qi, and Yang Gao. Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation. *IEEE transactions on medical imaging*, 41 (3):608–620, 2021. 2

[16] Toufique A Soomro, Lihong Zheng, Ahmed J Afifi, Ahmed Ali, Shafiullah Soomro, Ming Yin, and Junbin Gao. Image segmentation for mr brain tumor detection using machine learning: a review. *IEEE Reviews in Biomedical Engineering*, 16:70–90, 2022. 1

[17] Michael J Trimpl, Sergey Primakov, Philippe Lambin, Eleanor PJ Stride, Katherine A Vallis, and Mark J Gooding. Beyond automatic medical image segmentation—the spectrum between fully manual and fully automatic delineation. *Physics in Medicine & Biology*, 67(12):12TR01, 2022. 1

[18] Simon K Warfield, Kelly H Zou, and William M Wells. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging*, 23(7):903–921, 2004. 2

[19] Jianli Yang, Qiaozhi Ma, Jiqiang Liu, Haiping Zu, Siqing Dong, Ying Liu, Gang Guo, Binbin Nie, and Xuetao Mu. Multiparametric magnetic resonance imaging with comprehensive assessment of prostate volume, morphology, and composition better reflects the correlation with international prostate symptom score. *Urology*, 177:134–141, 2023. 2