Available online at www.eccomasproceedia.org
Eccomas Proceedia UNCECOMP (2025) 490-499

ECCOMAS
Proceedia

UNCECOMP 2025
6th ECCOMAS Thematic Conference on
Uncertainty Quantification in Computational Science and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou (eds.)
Rhodes Island, Greece, 15-18 June 2025

IMAGE-WISE UNCERTAINTY QUANTIFICATION FOR MEDICAL IMAGING

Philipp Rosauer[1], Angelina Händeler[1], Felix Terhag[1], Alexander Rüttgers[1], Michael Felderer[1,2]

[1]German Aerospace Center (DLR), Institute of Software Technology
Linder Hoehe, 51147, Cologne, Germany
e-mail: {philipp.rosauer, angelina.haendeler, felix.terhag, alexander.ruettgers, michael.felderer}@dlr.de

[2]University of Cologne, Department of Mathematics and Computer Science
Albertus-Magnus-Platz, 50923, Cologne, Germany

Abstract. *Cine cardiac magnetic resonance imaging (cineCMR) is a well-established imaging modality that is widely used in clinics and research centers to evaluate vital parameters like the exact stroke volume in the left and right ventricle or blood flow in the aortic root. To support medical professionals in the evaluation of cineCMR by machine learning, an uncertainty estimation is essential to replace trust and lack of it with probabilistic (interim) results. We present our workflow that ranks segmented images by image-wise uncertainty and supports manual enhancement by visualizing pixel-wise uncertainty, to improve the efficiency and effectiveness of human-machine collaboration. We present the performance of our nnU-Net configured U-Net with Monte Carlo (MC) dropout to predict uncertainty maps on the open Automated Cardiac Diagnosis Challenge (ACDC) dataset. Furthermore, we present multiple image-wise uncertainty scores derived from these pixel-wise uncertainty maps and compare their performance. Compared to only pixel-wise uncertainty-based workflows, we show how image-wise uncertainty scores can assist in the correction and validation of predicted segmentations through image evaluation. Due to an overwhelming amount of recorded CMR data, the proposed method could enable more reliable and faster evaluation in collaboration between medical professionals and probabilistic models.*

Keywords: Medical Image Segmentation, Cardiac Magnetic Resonance Imaging, Entropy, Monte Carlo Dropout

ISSN:2623-3339
Peer-review under responsibility of the organizing committee of UNCECOMP 2025.
doi: 10.7712/120225.12372.21280

1 INTRODUCTION

Semantic segmentation of CMR plays a crucial role in both diagnosis, monitoring and therapy of cardiovascular diseases. Leveraging deep learning models, like convolutional neural networks (CNNs), has achieved remarkable performance and shows the potential to drastically reduce analysis times [2]. These advances in fully automated segmentation are particularly needed for new imaging techniques like real-time MRI, producing thousands of images. However, despite scoring increasingly well in challenges, these models are inherently uncertain, which poses challenges in both research and clinical applications. Quantifying this uncertainty during application is essential to provide meaningful assistance to medical staff and ensuring reliable decision making.

Uncertainty Quantificatio (UQ) in segmentation is an active area of research, mostly aiming to indicate uncertain regions in one segmentation. While pixel-wise uncertainty measures, such as entropy, provide insights into per-voxel confidence clinicians must review every image, as they do not directly indicate whether the specifi segmentation should be reviewed. Especially when working with hundreds and thousands of individual segmentations on a daily basis, it seems a more efficien approach to somehow aggregate uncertainty information at the image level, allowing for a ranking of segmentations by their reliability [4].

There is a lack of image-wise metrics and their practical use to guide human interaction with model generated segmentations. Similar to the work of Ng et al. [10], in this work, we analyze the approach of using an uncertainty measure for ordering segmentations based on their aggregated pixelwise uncertainty. As the application of volumetric prediction is less concerned with exact geometrical accuracy of class borders, but rather focuses on high overlap, instead of distance-based metrics we use the Dice similarity coefficien (DSC) to measure how much prediction and ground truth coincide. This score can not only assist clinicians and researchers to identify cases that may require human validation or correction, but also serve as a tool to evaluate model performance on specifi data distributions. Combining both image-wise and pixel-wise uncertainty measures could help to reduce the validation and correction effort of medical staff, while increasing understanding and assessability.

To validate the approach, we use a publicly available dataset, as well as state-of-the-art open-source U-Net model configurations We demonstrate high correlation of our implementation in terms of ordering segmentations by performance, as measured by DSC between predictions and ground truth.

2 METHODS

In the following, we present our methodology, detailing the dataset, the segmentation model, and the uncertainty measures at both the pixel and image level.

2.1 Data

The presented methods are tested using the Automated Cardiac Diagnosis Challenge (ACDC) dataset [1]. The ACDC dataset is a widely recognized benchmark dataset for cardiac image analysis. It provides annotated CMR images across a range of pathologies, allowing robust evaluation of segmentation algorithms. The full dataset with 150 patients is divided into 100 patients, with 20 patients for each of the 5 groups that can be used for training, while the test dataset consists of the remaining 50 patients, with 10 for each of the 5 groups. This results in a total of 1902 training and 1076 test images. In all images the left ventricle (LV), the myocardium (MYO) and the right ventricle (RV) are labeled by experts. The dataset contains

images from different slices and physiological timepoints. This results in wide distribution of sizes of all three classes.

2.2 Model Architecture & Training

For CMR segmentation, we employ a U-Net architecture, a widely used convolutional neural CNN designed for biomedical image segmentation [12]. Our implementation follows the parameterization of nnU-Net [5], ensuring a well-optimized configuratio for medical imaging tasks.

The network consists of a five-layer encoder-decoder architecture, where each stage in the contraction and expansion paths contains a double convolutional block. Each block consists of two convolutional layers with kernel size 3, stride 1, and no padding, followed by instance normalization and Leaky ReLU activation. Dropout layers are incorporated into the convolutional blocks in the inner half of the network to estimate uncertainty using Bayesian principles [6]. Unlike standard segmentation pipelines, softmax activation is omitted, as uncertainty estimation requires access to unnormalized logits.

Training is performed using stochastic gradient descent (SGD) with a learning rate of 0.01, momentum of 0.99, and weight decay of 0.001 applied to a regularized cross-entropy loss function. The dropout rate is set to 0.2. The input images are resized to 256 × 224 pixels, and training images are randomly divided into 32 batches, each containing 58 samples.

2.3 Segmentation quality

To assess the quality of the segmentation results, we use two commonly employed and recommended metrics: the Average Symmetric Surface Distance (ASSD), representing distance-based metrics and the DSC, representing overlap-based metrics. However, segmentation quality is difficul to measure in general because it lacks a single definit ve metric, varies across anatomical structures and clinical contexts, and is influence by factors such as image resolution, label uncertainty, and interobserver variability [7].

ASSD ASSD measures the average distance between the surfaces of two segmentations. It is computed by calculating the distance from each point on the surface of one segmentation to the nearest point on the surface of the other, and then averaging these distances symmetrically:

$$\mathrm{ASSD}(A, B) = \frac{\sum_{x \in \partial S_A} d(x, \partial S_B) + \sum_{y \in \partial S_B} d(y, \partial S_A)}{|\partial S_A| + |\partial S_B|} \tag{1}$$

where:

- ∂S_A and ∂S_B are the sets of surface points of segmentations A and B, respectively.
- $d(x, \partial S_B) = \min_{y \in \partial S_B} \|x - y\|$ is the shortest Euclidean distance from point x in ∂S_A to the closest point in ∂S_B.
- Similarly, $d(y, \partial S_A) = \min_{x \in \partial S_A} \|y - x\|$.

A drawback of this metric is how it handles the case of complete segmentation failure. In this case, where a class is present only in one of the two compared segmentations, the denominator is 0. We chose to penalize these cases by setting the ASSD to the highest possible distance in the predicted image, its diagonal, as recommended in [7].

DSC DSC is a measure of overlap between the predicted segmentation and the ground truth. It ranges from 0 (no overlap) to 1 (perfect overlap), and is define as:

$$\mathrm{DSC}(A, B) = \frac{2|A \cap B|}{|A| + |B|}, \tag{2}$$

where $|A|$ and $|B|$ denote the number of elements (e.g., pixels or voxels) in segmentations A and B, and $|A \cap B|$ represents the number of overlapping elements.

2.4 Pixel-wise Uncertainty Quantification

There are various uncertainty measures that can be applied to image segmentation. Here, we compare two methods that provide pixel-wise uncertainty.

Entropy The firs is entropy, which is define as follows [11]:

$$H(x) = -\sum_{i=1}^{C} p_i \log p_i, \tag{3}$$

where C is the number of possible classes to which pixel x can belong, and p_i is the predicted probability that the pixel belongs to the i-th class. As p_i is simply given as an interim result in $\mathrm{softmax}$, the additional cost to compute H_i is minimal.

MC Dropout Variance (MCDV) The second approach we present is to use MC dropout during inference. For this, the dropout layers in the model remain active during the inference phase to obtain probabilistic predictions. The dropout probability is set to $p = 0.2$, the same as during training. For each input image, the model is evaluated five times to obtain five probabilistic, slightly different predictions. This is equivalent to sampling the model weights five times from a Bernoulli distribution with probability p, and making predictions depending on the resulting, slightly different weights. According to the theoretical derivation by Gal et al. [3], the predictive distribution, which is also referred to as the Bayesian model average, is then approximated by taking the sample mean of these probabilistic predictions p_i, $i = 1, \ldots, 5$. It is calculated for each segment class c, $c = 1, \ldots, 4$, and each pixel x separately:

$$\hat{p}^c(x) = \frac{1}{5} \sum_{i=1}^{5} p_i^c(x) \tag{4}$$

The pixelwise uncertainty maps result from this by calculating the pixelwise entropies of the mean outputs $\hat{p}^c$ over the segment classes, $c = 1, \ldots, 4$, as define in (3):

$$u(x) = -\sum_{c=1}^{4} \hat{p}^c(x) \log \hat{p}^c(x).$$

2.5 Image-wise Uncertainty Quantification

Mean To derive an image-wise uncertainty measure from the pixel-wise measure, we can simply compute the mean:

$$\bar{H} = \frac{1}{N} \sum_{i=1}^{N} H(x) \tag{5}$$
$$\bar{u} = \frac{1}{N} \sum_{i=1}^{N} u(x)$$

Surface Distance over Samples Using the pairwise ASSD between samples, we can define

$$\mathrm{ASSD_S} = \frac{2}{N(N-1)} \sum_{0 \le i < j < N} \mathrm{ASSD}(S_i, S_j), \tag{6}$$

where S_i and S_j are pairwise predicted labels of a probabilistic sample each, derived from the softmax outputs by assigning the segment class with the highest probability to a pixel and N the total number of Samples.

DSC over Samples To measure similarity between the softmax probabilities of the predictions we use the pairwise DSC between samples:

$$\mathrm{DSC_S} = \frac{2}{N(N-1)} \sum_{0 \le i < j < N} \mathrm{DSC}(S_i, S_j), \tag{7}$$

similarly to $\mathrm{ASSD_S}$.

2.6 Test setup

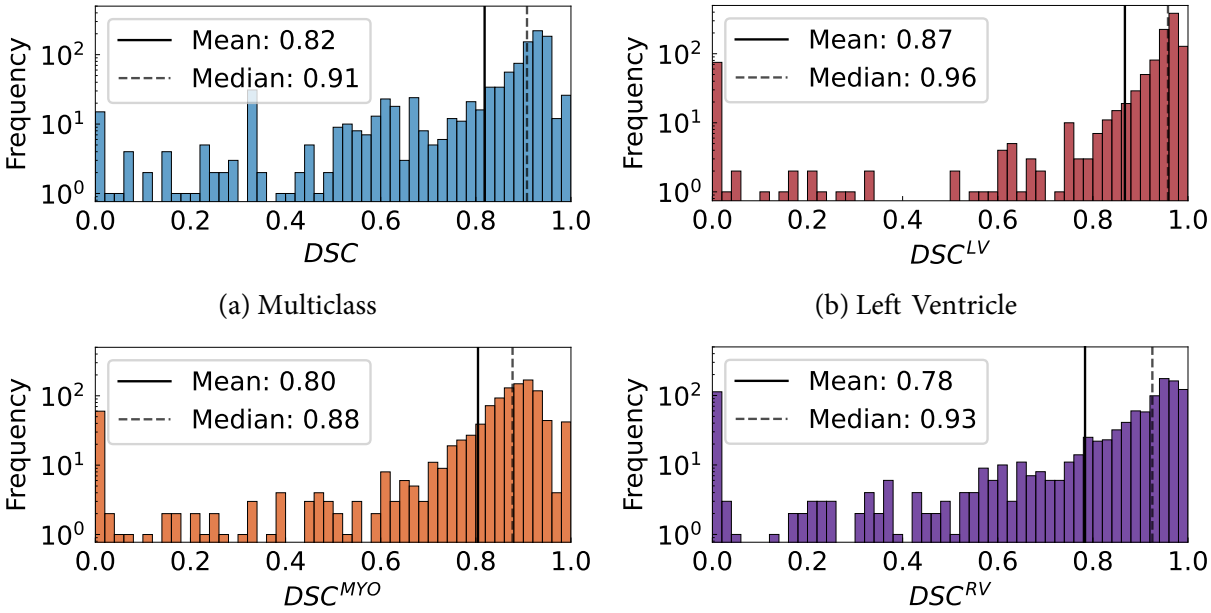
In this work, our primary focus is not on achieving the highest segmentation accuracy but rather on assessing the ability to correctly rank images by their segmentation performance. This ranking is evaluated by comparing the uncertainty quantificatio with the segmentation quality. To quantify the correlation between uncertainty and segmentation quality, we employ Spearman's rank correlation coefficien (ρ). Spearman's ρ is a non-parametricmeasure that evaluates the strength and direction of the monotonic relationship between two variables, making it well-suited for analyzing how uncertainty reflect segmentation performance. According to [13], this is define as:

$$r_S(X, Y) = 1 - \frac{6 \sum_{i=1}^{n} (P_i - Q_i)^2}{n(n^2 - 1)},$$

where $\{(X_i, Y_i)\}_{i=1}^{n}$ are n iid data pairs, P_i the rank of X_i among the $X_1, \ldots, X_n$ and Q_i the rank of Y_i among the $Y_1, \ldots, Y_n$. In this application, X_i represent the image-wise uncertainties and Y_i the DSC of n predictions. It takes values in $[-1, 1]$, where 1 is perfect, -1 is negative, and 0 is no correlation. A strong correlation indicates that a uncertainty quantificatio effectively identifie cases of higher and lower segmentation performance.

3 RESULTS

In this section, we analyze the relationship between uncertainty measures and segmentation quality. We firs evaluate pixel-wise uncertainty using established methods such as entropy and MC dropout. Additionally, we compute four image-wise uncertainty measures: $\bar{H}$, $\bar{u}$ (5), ASSD_S (6), and DSC_S (7). To assess their effectiveness, we compare the Spearman correlations of these metrics with segmentation performance, providing insights into how uncertainty rankings align with actual segmentation performance. The U-Net model achieved a mean DSC of 0.82 and a median DSC of 0.91 on the test dataset.



(a) Multiclass (b) Left Ventricle

(c) Myocardium (d) Right Ventricle

Figure 1: Histograms showing the distribution of binned $Dice$ scores for each segmentation class and all classes combined.

To analyze the segmentation performance across the dataset, we visualize the distribution of DSC using a histogram, see Figure 1. The x-axis represents the DSC values, while the y-axis indicates the frequency of occurrences on a logarithmic scale.

The histograms of the three classes show a bimodal distribution with a high frequency of samples with DSC close to 1, a very low frequency of samples in the mid-range (0.1–0.85) and a notable peak at 0, which occurs when either the prediction or the ground truth is completely missing while the other is present. The mean and median DSC scores for the LV class is highest, with a notably smaller amount of predictions with a score under 0.9. These distributions highlight that while the majority of segmentations are accurate, there are cases of inaccurate segmentations. The range of these inaccuracies spans from complete segmentation failure ($\text{DSC} = 0$) to small deviations only at the edges of classes, which may correspond to difficul or ambiguous samples, underlining the importance of identifying uncertain cases and ranking them for potential human validation.

3.1 Pixel-wise Uncertainty

We computed MCDV and entropy on a pixel basis and provide the five probabilistic inference outputs (Sample 1 - Sample 5), as well as the given GT for three examples in Figure 2. All three images are cropped to fi the ground truth segmentation for visualization, thus the resolution differs between them. Image ★ presents a typical mid-ventricular slice with a high DSC of 0.97. The entropy H_x is only raised at the very borders of the classes, which suggests high agreement between classes within a single prediction. Also, the predictions of all five samples is close to identical, represented in a very high DSC_S of 0.99. As for H_x, u_x is only raised at the borders. In case of Image ■ the DSC is significantl lower, mainly caused by errors in the RV class, but also by disagreement regarding the LV-MYO border. Mislabeled pixels do indeed show an increase in H_x, indicating lower agreement in areas that are predicted incorrectly. However, the five sampled predictions show only slight differences, lowering the DSC_S only by a margin compared to Image ★. Lastly, Image ♦ is an example for segmentation failure with a DSC of only 0.28. The overlap between the five sampled predictions ($\text{DSC}_\text{S} = 0.59$) is very low. Although H_x and also u_x visually indicate high local disagreement, computing the mean value over the entire image $\bar{H}$ and $\bar{u}$ fail to represent this information. This is likely due to the significantl smaller segmentation, compared e.g. to Image ★.

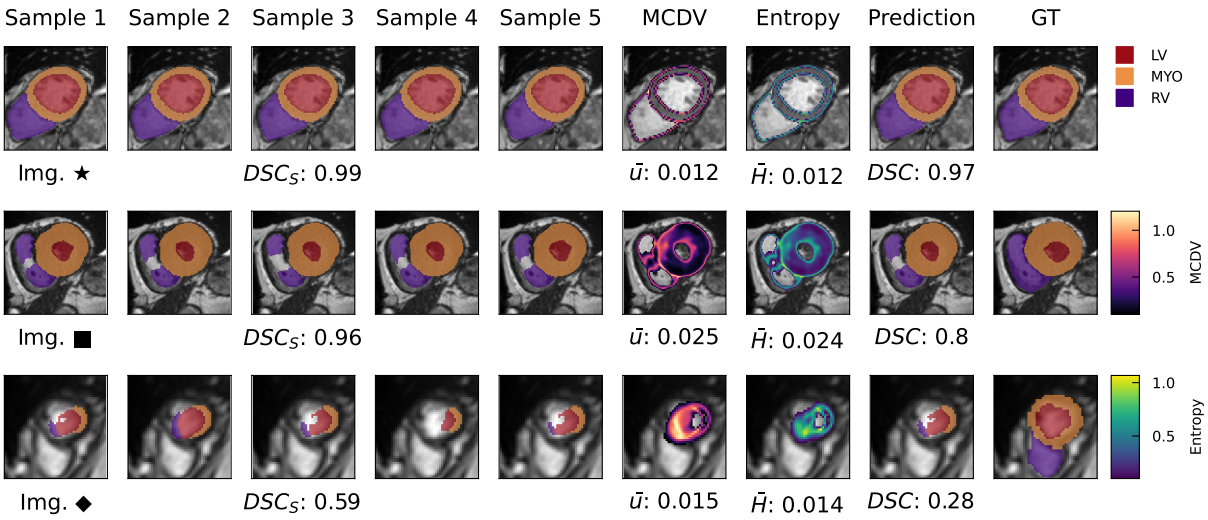


Figure 2: 3 Images from the test dataset with all 5 predictions from the probabilistic dropout inference, MC Dropout Variance (MCDV), deterministic prediction and ground truth (GT)

3.2 Correlation of image-wise uncertainty measures

We investigated the potential of all 4 uncertainty measures for evaluating segmentation performance during or after inference. To measure segmentation performance of the trained model against ground truth on the test data ,we computed both ASSD as well as DSC. As stated in [10] both metrics are well established in medical image analysis.

Table 1 shows the Spearman correlations between uncertainty measures and segmentation performance metrics for all 3 segments. We found little to no correlation between both entropy and MCDV for all classes with both quality metrics. As evident in the examples in Figure 2 this probably caused by sensitivity of the metric in regard of relative segmentation size. The correlations between ASSD_S and ASSD for all classes range between 0.61 for the LV and 0.72

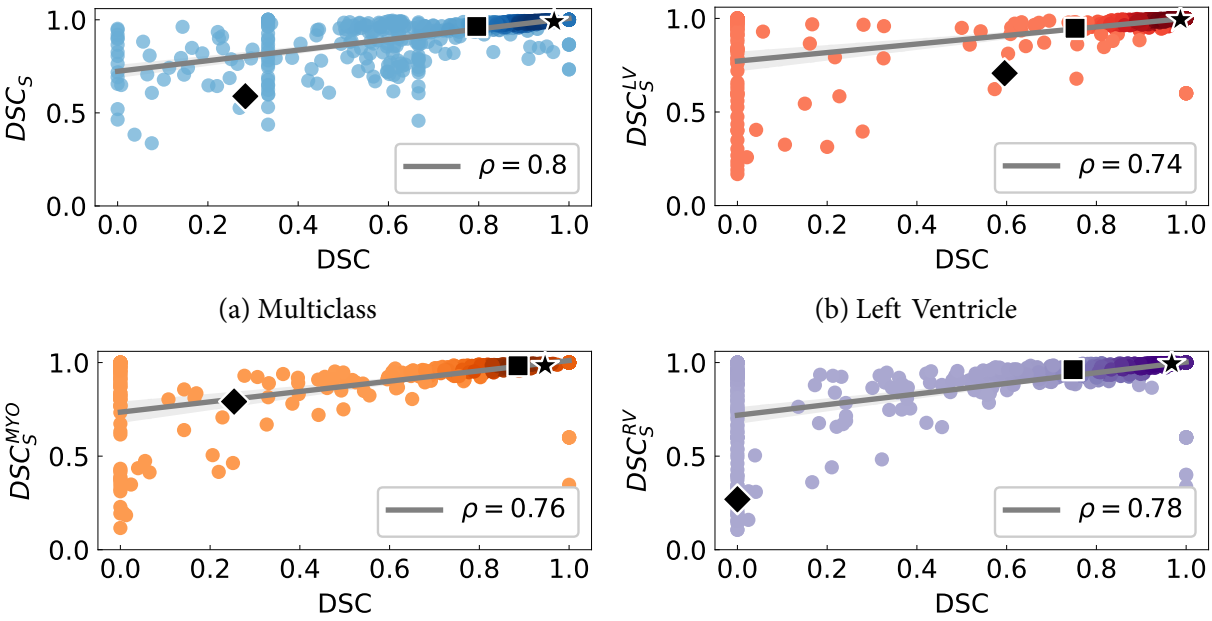
for MYO and are consistent with the finding of Ng et al. [10] and support the feasibility of using the measure for ranking predictions when a distance-based metric is needed. The correlation between DSC_S and DSC is even higher for all segments with lowest correlation for the class LV at 0.74 and highest correlation for RV at 0.78. These finding suggest that DSC_S is an even better predictor for segmentation quality, especially when overlap-based metrics are considered for segmentation quality.

Quality Metric	Left Ventricle				Myocardium				Right Ventricle			
	DSC_S	ASSD_S	$\bar{H}$	$\bar{u}$	DSC_S	ASSD_S	$\bar{H}$	$\bar{u}$	DSC_S	ASSD_S	$\bar{H}$	$\bar{u}$
DSC	**0.74**	0.57	0.12	0.22	**0.76**	0.43	0.01	0.01	**0.78**	0.64	0.12	0.17
$ASSD$	0.58	**0.61**	0.05	0.08	0.51	**0.72**	0.03	0.05	0.68	**0.69**	0.34	0.39

Table 1: Correlations between selected quality and uncertainty metrics

The plots in Figure 1 show the distributions of the computed DSC and DSC_S for every image in the test dataset for each segment as well as the combined score and measure for the multiclass segmentation (blue). It can be observed that for complete segmentation failure the uncertainty measure fails as an indicator for segmentation quality, as we measured a wide range of DSC_S for these cases, but show consistent results for the majority of data points.

Regarding computational cost, in our experiments DSC was nearly twice as fast as our implementation of ASSD, using SimpleITK for findin segmentation contours [9] and signed Maurer distance [8]. When used as a segmentation quality predictor, computation times should be considered, as they scale linearly with the number of compared samples.



(a) Multiclass

(b) Left Ventricle

(c) Myocardium

(d) Right Ventricle

Figure 3: Correlation plots showing the relationship between DSC. The x-axis represents the DSC between predictions and ground truth, while the y-axis represents the DSC between samples. The colormap in each subplot indicates point density estimated using Gaussian kernel density estimation (KDE), and a regression line illustrates the trend.

4 CONCLUSION

We presented our finding on correlations between chosen uncertainty measures and segmentation quality metrics. In particular, when the application requires a high volumetric accuracy, e.g. when determining the stroke volume, the inter-sample DSC offers high potential for ranking segmentations by predicted quality at a lower computational cost than other metrics comparing segmentation samples. Simply computing the entropy of a deterministic model or even computing MC dropout average showed little to no correlation with volumetric as well as distance based quality metrics.

As reasoned in [10], the physical meaning of distance metrices such as ASSD can provide useful insights depending on the medical application. However, we found that dealing with missing classes in one of the two compared segmentations leads to ambiguous results, as the metric itself, as stated in subsection 2.3, is not define in these cases. Aggregating ASSD scores across segments and/or multiple samples introduces potential errors and increases ambiguity. We did not explore further variations of the ASSD that could counteract these shortcomings, but recognize the medical value a distance metric can offer.

If computation time is either not an issue, or is dominated by inference itself, we would consider using both ASSD, or another distance-based metric like the mean average surface distance (MASD), and DSC, following the recommendations in [7]. Furthermore, we do not consider the available results as merely academic, but plan to integrate them into an evaluation tool as a combination of image-wise and pixel-wise UQ measures. The calculated DSC sorts the segmented images and MCDV highlights areas that require special attention.

Acknowledgments The research was partly carried out under the project LBNP ISS4Mars by the German Aerospace Center (DLR).

REFERENCES

[1] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018.

[2] Chen Chen, Chen Qin, Huaqi Qiu, Giacomo Tarroni, Jinming Duan, Wenjia Bai, and Daniel Rueckert. Deep learning for cardiac image segmentation: a review. *Frontiers in cardiovascular medicine*, 7:25, 2020.

[3] 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.

[4] Francesco Galati, Sébastien Ourselin, and Maria A Zuluaga. From accuracy to reliability and robustness in cardiac magnetic resonance image segmentation: a review. *Applied Sciences*, 12(8):3936, 2022.

[5] Fabian Isensee, Paul Jaeger, Simon Kohl, Jens Petersen, and Klaus Maier-Hein. nnU-Net: a self-configurin method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

[6] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015.

[7] Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, et al. Metrics reloaded: recommendations for image analysis validation. *Nature Methods*, 21(2):195–212, 2024.

[8] Calvin Maurer, Rensheng Qi, and Vijay Raghavan. A linear time algorithm for computing exact Euclidean distance transforms of binary images in arbitrary dimensions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):265–270, 2003.

[9] Matthew McCormick, Xiaoxiao Liu, Julien Jomier, Charles Marion, and Luis Ibanez. Itk: enabling reproducible research and open science. *Frontiers in Neuroinformatics*, 8:13, 2014.

[10] Matthew Ng, Fumin Guo, Labonny Biswas, Steffen Petersen, Stefan Piechnik, Stefan Neubauer, and Graham Wright. Estimating uncertainty in neural networks for cardiac MRI segmentation: a benchmark study. *IEEE Transactions on Biomedical Engineering*, 70(6):1955–1966, 2022.

[11] Nikhil Pal and Sankar Pal. Entropy: A new definitio and its applications. *IEEE transactions on systems, man, and cybernetics*, 21(5):1260–1270, 1991.

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[13] Weichao Xu, Yunhe Hou, YS Hung, and Yuexian Zou. A comparative analysis of Spearman's rho and Kendall's tau in normal and contaminated normal models. *Signal Processing*, 93(1):261–276, 2013.