UNCERTAINTY MAPS IN MEDICAL IMAGING SEGMENTATION: STATE OF THE ART

WILLIAM BASSOLINO

PAPERS IN THIS REVIEW:

1. *SkiNet: A deep learning framework for skin lesion diagnosis with uncertainty estimation and explainability* – Singh et al. 2022
2. *Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks* – Wang et al. 2019
3. *Exploring uncertainty measures in convolutional neural network for semantic segmentation of oral cancer images* – Song et al. 2022
4. *Inherent Brain Segmentation Quality Control from Fully ConvNet Monte Carlo Sampling* – Roy et al. 2018
5. *Uncertainty Driven Pooling Network for Microvessel Segmentation in Routine Histology Images* – Fraz et al. 2018
6. *MILD-Net: Minimal Information Loss Dilated Network for Gland Instance Segmentation in Colon Histology Images* – Graham et al. 2018
7. *Leveraging Uncertainty Estimates for Predicting Segmentation Quality* – DeVries et Taylor 2018
8. *ADNet++: A few-shot learning framework for multi-class medical image volume segmentation with uncertainty-guided feature refinement* – Hansen et al. 2023
9. *A Quantitative Comparison of Epistemic Uncertainty Maps Applied to Multi-Class Segmentation* – Camarasa et al. 2021

MACRO-TOPICS

- Uncertainty estimation in binary segmentation
  - Papers: 1-7
- Uncertainty estimation in multi-class segmentation
  - Papers: 9
- Uncertainty as method to improve segmentation
  - Papers: 5,6,8

Personal notes:

1. The use of entropy on both test-time data augmentation and Monte Carlo dropout samples to measure uncertainty is something that exists in literature. There is variety of approaches.
2. The question “how to transform the uncertainty map into a single metric?” is an open question, but the first papers have found maybe the best answer (accuracy of the prediction depending on the certainty)
3. Uncertainty estimation in multi-class segmentation is not a topic with extensive literature, but it exists.
4. The main methods to use uncertainty to improve segmentation are to have manual assistance from a domain expert, use it to fine tune model-specific features, or use uncertainty to remove uncertain pixels-voxels to improve performances.
5. Assuring segmentation quality without a test-set is still an open question in literature.

In the next slides, most important aspects of 9 papers are exposed.

SkiNet: A deep learning framework for skin lesion diagnosis with uncertainty estimation and explainability

Rajeev Kumar Singh[1]*, Rohan Gorantla[1,2], Sai Giridhar Rao Allada[1,4], Pratap Narra[1,3]

1 Department of Computer Science, Shiv Nadar University, Delhi NCR, India, **2** School of Informatics, University of Edinburgh, Edinburgh, United Kingdom, **3** Tandon School of Engineering, New York University, New York, New York, United States of America, **4** Luddy School Of Informatics, Computing, And Engineering, Indiana University Bloomington, Bloomington, Indiana, United States of America

* Rajeev.kumar@snu.edu.in

Abstract

Skin cancer is considered to be the most common human malignancy. Around 5 million new cases of skin cancer are recorded in the United States annually. Early identification and evaluation of skin lesions are of great clinical significance, but the disproportionate dermatologist-patient ratio poses a significant problem in most developing nations. Therefore a novel deep architecture, named as SkiNet, is proposed to provide faster screening solution and assistance to newly trained physicians in the process of clinical diagnosis of skin cancer. The main motive behind SkiNet's design and development is to provide a white box solution, addressing a critical problem of trust and interpretability which is crucial for the wider adoption of Computer-aided diagnosis systems by medical practitioners. The proposed SkiNet is a two-stage pipeline wherein the lesion segmentation is followed by the lesion classification. Monte Carlo dropout and test time augmentation techniques have been employed in the proposed method to estimate epistemic and aleatoric uncertainty. A novel segmentation model named Bayesian MultiResUNet is used to estimate the uncertainty on the predicted segmentation map. Saliency-based methods like XRAI, Grad-CAM and Guided Backprop are explored to provide post-hoc explanations of the deep learning models. The ISIC-2018 dataset is used to perform the experimentation and ablation studies. The results establish the robustness of the proposed model on the traditional benchmarks while addressing the black-box nature of such models to alleviate the skepticism of medical practitioners by incorporating transparency and confidence to the model's prediction.

IN-DEPTH ANALYSIS: [1]

ABSTRACT

The proposed SkiNet is a two-stage pipeline wherein the lesion segmentation is followed by the lesion classification.
A novel segmentation model named Bayesian MultiResUNet is used to estimate the uncertainty on the predicted segmentation map. Saliency-based methods like XRAI, Grad-CAM and Guided Backprop are explored to provide post-hoc explanations of the deep learning models.

ALEATORIC VS EPISTEMIC UNCERTAINTY

There are mainly two types of uncertainty visualization, Aleatoric and Epistemic uncertainty:
Aleatoric uncertainty captures noise inherent in the data and cannot be abated by collecting more data.
Epistemic uncertainty, also known as model uncertainty, accounts for variability in the parameters of the model and analyzes what the model is not aware owing to the lack of training data. Epistemic uncertainty is helpful to understand examples that vary from training data especially in situations where we have small and imbalanced datasets, which is common in CAD systems.
Uncertainties are formulated as probability distributions over the model parameters (for epistemic uncertainty) or model inputs (for aleatoric uncertainty).

SkiNet: A deep learning framework for skin lesion diagnosis with uncertainty estimation and explainability – Singh et al. 2022

ALEATORIC UNCERTAINTY

A test sample s* is augmented to form V different versions of the image and is forwarded to the network. The mean μ_a over these iterations is considered the final result of a given test sample. μ_a is computed as shown in the equation:

$$\mu_a \approx \frac{1}{V}\sum_{v=1}^{V} p(y^*|s_v^*, \hat{W})$$

where s^*_v denotes the v^{th} augmented image, W denotes the weights of the network and V is the total number of image augmentations. Among several classes y* , the one with μ_{max} is selected as the outcome for each test sample s*.

EPISTEMIC UNCERTAINTY

Given a test sample s* , we sample the network B times over its parameters and thereby giving an estimate of the predictive posterior distribution. The mean μ_e over these iterations is considered as the final result on a given test sample. μ_e is computed as shown in the equation:

$$\mu_e \approx \frac{1}{B}\sum_{m=1}^{B} p(y^*|s^*, \hat{W}_B)$$

where W_B denotes the weights of the network with dropouts in B^{th} MC iteration and B is the total number of sampled sets of weights. Among several classes y*, the one with μ_{max} is selected as the outcome for each test sample s*.

SkiNet: A deep learning framework for skin lesion diagnosis with uncertainty estimation and explainability – Singh et al. 2022

UNCERTAINTY

The two approaches are then combined to calculate the overall uncertainty where a test sample s* is augmented to form M different versions of the image and is forwarded to the network with the dropout activated during the test time. The mean **μ** over these iterations is considered as the final result on a given test sample. **μ** is computed as shown in the equation:

$$\mu \approx \frac{1}{M}\sum_{m=1}^{M} p(y^*|s_m^*, \hat{W}_m)$$

where s^*_m denotes the augmented image passed and W_m denotes the weights of the network with dropouts during the m^{th} iteration and M is the total number of iterations. Among several classes y*, the one with μ_{max} is selected as the outcome for each test sample s*

In order to estimate the **model uncertainty ϕ**, we calculate the entropy of the averaged probability vector across the N classes using the equation

$$\varphi = -\sum_{n=1}^{N} p_n \log p_n$$

where p_n is the probability of n^{th} class

SkiNet: A deep learning framework for skin lesion diagnosis with uncertainty estimation and explainability – Singh et al. 2022

EVALUATION METRICS: UNCERTAINTY

Normalized uncertainty ϕ_{norm} is calculated:

$$\varphi_{norm} = \frac{\varphi - \varphi_{min}}{\varphi_{max} - \varphi_{min}}$$

To split the predictions into certain and uncertain categories, we set a threshold ϕ_T [0, 1] where a prediction is deemed to be certain if $\phi_{norm} < \phi_T$ and uncertain if ϕ_{norm}

There are 4 kinds of predictions:

- Incorrect-uncertain (iu) → prediction is incorrect and the model is uncertain about it
- correct-uncertain (cu) → prediction is correct but the model is uncertain about it
- correct-certain (cc) → prediction is correct and the model is uncertain about it
- incorrect-certain (ic) → prediction is incorrect and the model is certain about it

Diagnostic Accuracy A is represented in the form:

$$A(\varphi_T) = \frac{L_{cc} + L_{iu}}{L_{cc} + L_{iu} + L_{cu} + L_{ic}}$$

where L represents the count of each possible combination

SkiNet: A deep learning framework for skin lesion diagnosis with uncertainty estimation and explainability – Singh et al. 2022

Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks

Guotai Wang[a,b,c,*], Wenqi Li[a,b], Michael Aertsen[d], Jan Deprest[a,d,e,f], Sébastien Ourselin[b], Tom Vercauteren[a,b,f]

[a] Wellcome / EPSRC Centre for Interventional and Surgical Sciences, University College London, London, UK
[b] School of Biomedical Engineering and Imaging Sciences, King's College London, London, UK
[c] School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China
[d] Department of Radiology, University Hospitals Leuven, Leuven, Belgium
[e] Institute for Women's Health, University College London, London, UK
[f] Department of Obstetrics and Gynaecology, University Hospitals Leuven, Leuven, Belgium

ARTICLE INFO

Article history:
Received 2 August 2018
Revised 26 January 2019
Accepted 27 January 2019
Available online 7 February 2019

Communicated by Pingkun Yan

Keywords:
Uncertainty estimation
Convolutional neural networks
Medical image segmentation
Data augmentation

ABSTRACT

Despite the state-of-the-art performance for medical image segmentation, deep convolutional neural networks (CNNs) have rarely provided uncertainty estimations regarding their segmentation outputs, e.g., model (*epistemic*) and image-based (*aleatoric*) uncertainties. In this work, we analyze these different types of uncertainties for CNN-based 2D and 3D medical image segmentation tasks at both pixel level and structure level. We additionally propose a test-time augmentation-based *aleatoric* uncertainty to analyze the effect of different transformations of the input image on the segmentation output. Test-time augmentation has been previously used to improve segmentation accuracy, yet not been formulated in a consistent mathematical framework. Hence, we also propose a theoretical formulation of test-time augmentation, where a distribution of the prediction is estimated by Monte Carlo simulation with prior distributions of parameters in an image acquisition model that involves image transformations and noise. We compare and combine our proposed *aleatoric* uncertainty with model uncertainty. Experiments with segmentation of fetal brains and brain tumors from 2D and 3D Magnetic Resonance Images (MRI) showed that 1) the test-time augmentation-based *aleatoric* uncertainty provides a better uncertainty estimation than calculating the test-time dropout-based model uncertainty alone and helps to reduce overconfident incorrect predictions, and 2) our test-time augmentation outperforms a single-prediction baseline and dropout-based multiple predictions.

© 2019 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY license. (http://creativecommons.org/licenses/by/4.0/)

IN-DEPTH ANALYSIS: [2]

TEST-TIME AUGMENTATION

We propose a more general estimation of aleatoric uncertainty that is related to not only image noise but also spatial transformations (rotation, scaling, flipping) of the input, considering different possible poses of the object during image acquisition. To obtain the transformation-related uncertainty, we **augment the input image at test time**, and obtain an estimation of the distribution of the prediction based on test-time augmentation.
We propose a mathematical formulation for test-time augmentation, and analyse its performance for the general aleatoric uncertainty estimation in medical image segmentation tasks.

METHODS

Both the variance and entropy of the distribution p(Y|X) can be used to estimate uncertainty. However, variance is not sufficiently representative in the context of multimodal distributions. In this paper we use **[Shannon] entropy** for uncertainty Estimation (both for aleatoric and epistemic uncertainties).

[Note: they use Monte Carlo sampling for Aleatoric Uncertainty estimation because they don't apply inverse transformation to the perturbated images because they don't know the paraemeters]

We use the volume variation coefficient (VVC) to estimate the structure-wise uncertainty

$$VVC = \frac{\sigma_v}{\mu_v}$$

Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks – Wang et al. 2019

EXPERIMENTS AND RESULTS

We compared different types of uncertainties for the segmentation results:

1) the proposed aleatoric uncertainty based on our formulated test-time augmentation (TTA)
2) the epistemic uncertainty based on test-time dropout (TTD)
3) hybrid uncertainty that combines the aleatoric and epistemic uncertainties based on TTA + TTD

We also evaluated the segmentation accuracy of these different prediction methods: TTA, TTD, TTA + TTD and the baseline that uses a single prediction without TTA and TTD

Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks – Wang et al. 2019

PLOTS AND MEASURES

N=20 was chosen →

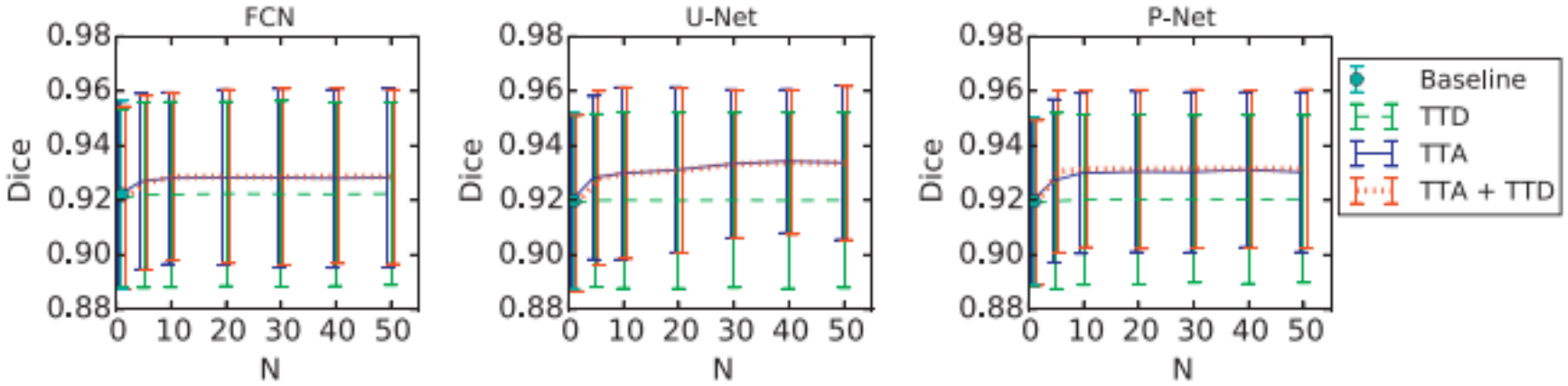


Fig. 2. Dice of 2D fetal brain segmentation with different *N* that is the number of Monte Carlo simulation runs.

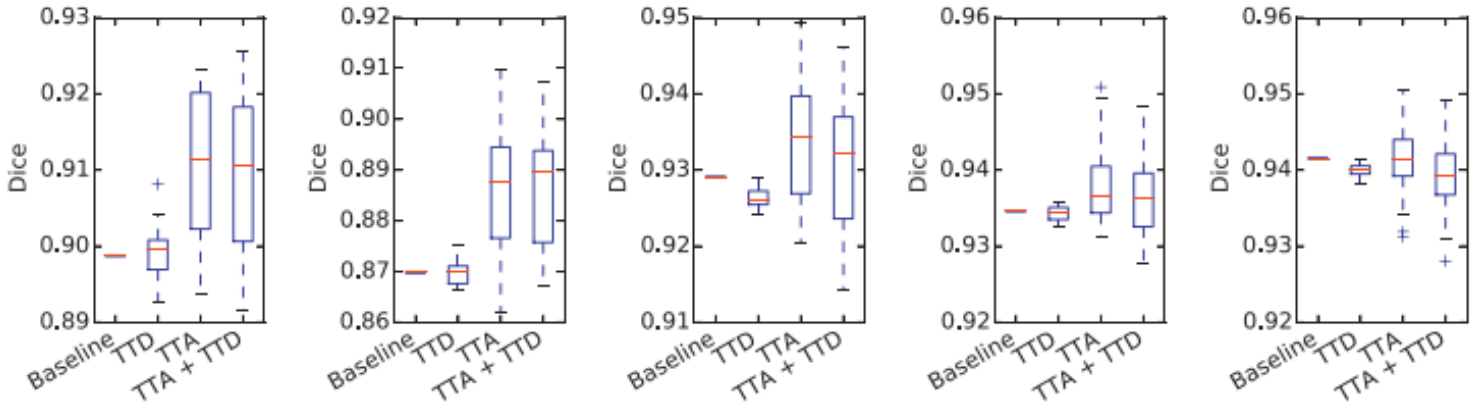


Fig. 3. Dice distributions of segmentation results with different testing methods for five example stacks of 2D slices of fetal brain MRI. Note TTA's higher mean value and variance compared with TTD.

← Plot that shows the dices distribution of the N different inferences on 5 different images

Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks – Wang et al. 2019

Exploring uncertainty measures in convolutional neural network for semantic segmentation of oral cancer images

Bofan Song,[a,*] Shaobai Li,[a] Sumsum Sunny,[b] Keerthi Gurushanth,[c] Pramila Mendonca,[d] Nirza Mukhia,[c] Sanjana Patrick,[e] Tyler Peterson,[a] Shubha Gurudath,[c] Subhashini Raghavan,[c] Imchen Tsusennaro,[f] Shirley T. Leivon,[f] Trupti Kolur,[d] Vivek Shetty,[d] Vidya Bushan,[d] Rohan Ramesh,[f] Vijay Pillai,[d] Petra Wilder-Smith,[g] Amritha Suresh,[b,d] Moni Abraham Kuriakose,[h] Praveen Birur,[e,c] and Rongguang Liang[a,*]

Abstract

Significance: Oral cancer is one of the most prevalent cancers, especially in middle- and low-income countries such as India. Automatic segmentation of oral cancer images can improve the diagnostic workflow, which is a significant task in oral cancer image analysis. Despite the remarkable success of deep-learning networks in medical segmentation, they rarely provide uncertainty quantification for their output.

Aim: We aim to estimate uncertainty in a deep-learning approach to semantic segmentation of oral cancer images and to improve the accuracy and reliability of predictions.

Approach: This work introduced a UNet-based Bayesian deep-learning (BDL) model to segment potentially malignant and malignant lesion areas in the oral cavity. The model can quantify uncertainty in predictions. We also developed an efficient model that increased the inference speed, which is almost six times smaller and two times faster (inference speed) than the original UNet. The dataset in this study was collected using our customized screening platform and was annotated by oral oncology specialists.

Results: The proposed approach achieved good segmentation performance as well as good uncertainty estimation performance. In the experiments, we observed an improvement in pixel accuracy and mean intersection over union by removing uncertain pixels. This result reflects that the model provided less accurate predictions in uncertain areas that may need more attention and further inspection. The experiments also showed that with some performance compromises, the efficient model reduced computation time and model size, which expands the potential for implementation on portable devices used in resource-limited settings.

Conclusions: Our study demonstrates the UNet-based BDL model not only can perform potentially malignant and malignant oral lesion segmentation, but also can provide informative pixel-level uncertainty estimation. With this extra uncertainty information, the accuracy and reliability of the model's prediction can be improved.

IN-DEPTH ANALYSIS: [3]

UNCERTAINTY MEASURE

This work introduced a UNet-based Bayesian deep-learning (BDL) model to segment potentially malignant and malignant lesion areas in the oral cavity. The model can quantify uncertainty in predictions.

The uncertainty is obtained by calculating the variance:

$$v = \frac{1}{\sigma}\sum_{i=1}^{\sigma}(p(y|x^*, W_i) - p(y|x^*, X, Y))^2$$

where $p(y|x^*,W_i)$ represents σ times softmax output with different weights W_i of input x and $p(y|x^*,X,Y)$ is the predictive posterior mean of input x*.

UNCERTAINTY EVALUATION

d two conditional probabilities as uncertainty evaluation metrics, p (accurate | certain) and p (uncertain | inaccurate), and the combination of them, patch accuracy versus patch uncertainty (PAvPU)

$$p(\text{accurate}|\text{certain}) = \frac{n_{ac}}{n_{ac} + n_{ic}},$$

$$p(\text{uncertain}|\text{inaccurate}) = \frac{n_{iu}}{n_{ic} + n_{iu}},$$

$$\text{PAvPU} = \frac{n_{ac} + n_{iu}}{n_{ac} + n_{au} + n_{ic} + n_{iu}}.$$

Exploring uncertainty measures in convolutional neural network for semantic segmentation of oral cancer images – Song et al. 2022

EXPERIMENTS AND RESULTS

The idea is to better the segmentation performances by removing uncertain pixels

We measured the change in pixel accuracy, mean IoU, and weighted IoU **when removing pixels** with uncertainty values higher than a specific level. By adjusting the level of uncertainty thresholding, we plotted the change of these three evaluation metrics in the following figure. We can see a continuous increase in all three evaluation metrics in response to a change in uncertainty thresholding.

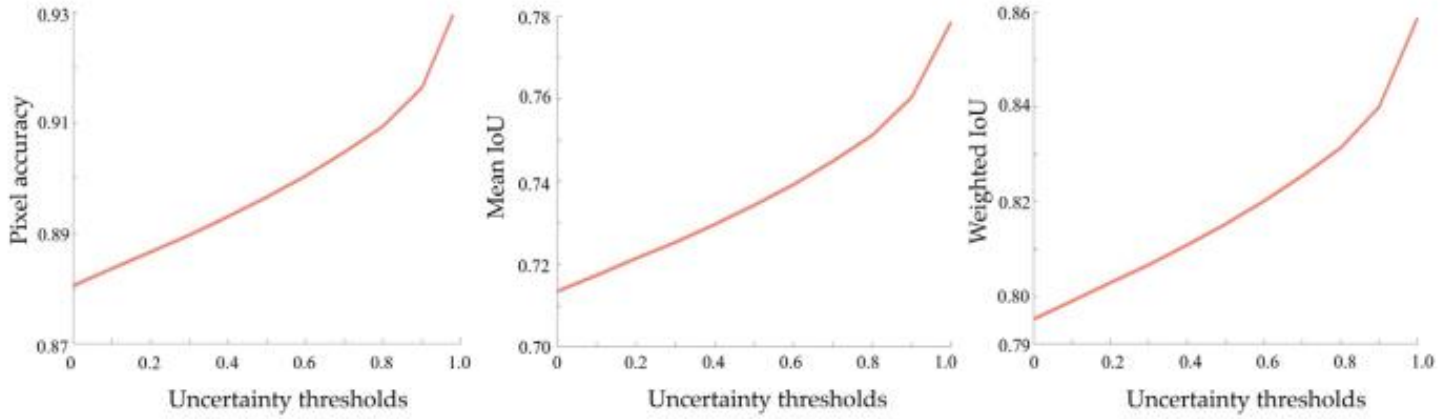
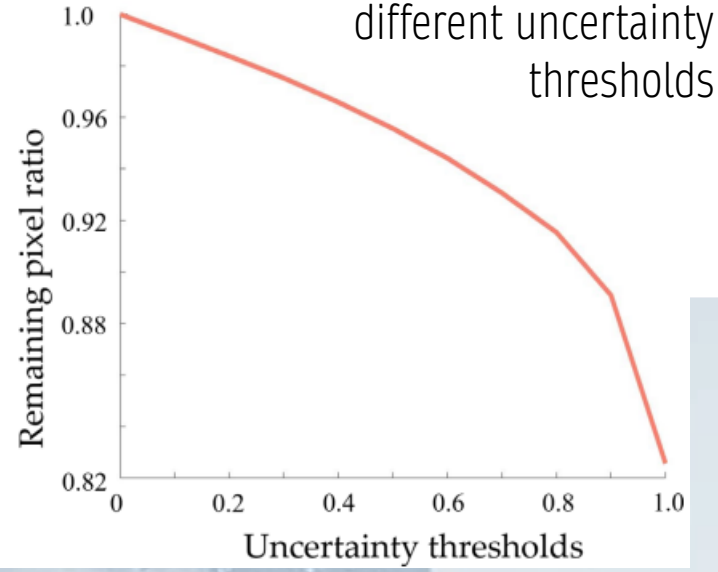


Fig. 4 The change of pixel accuracy, mean IoU, and weighted IoU when removing pixels with uncertainty values higher than a specific level.

The remaining pixel ratios (1—removed pixel ratio) corresponding to different uncertainty thresholds



Exploring uncertainty measures in convolutional neural network for semantic segmentation of oral cancer images – Song et al. 2022

EXPERIMENTS AND RESULTS

To evaluate the uncertainty estimation performance, we calculated and compared the p (accurate | certain), p (uncertain | inaccurate), and PAvPU. We then observed how these metrics varied with a change of uncertainty threshold. A model with A higher value of these metrics is a better performer.

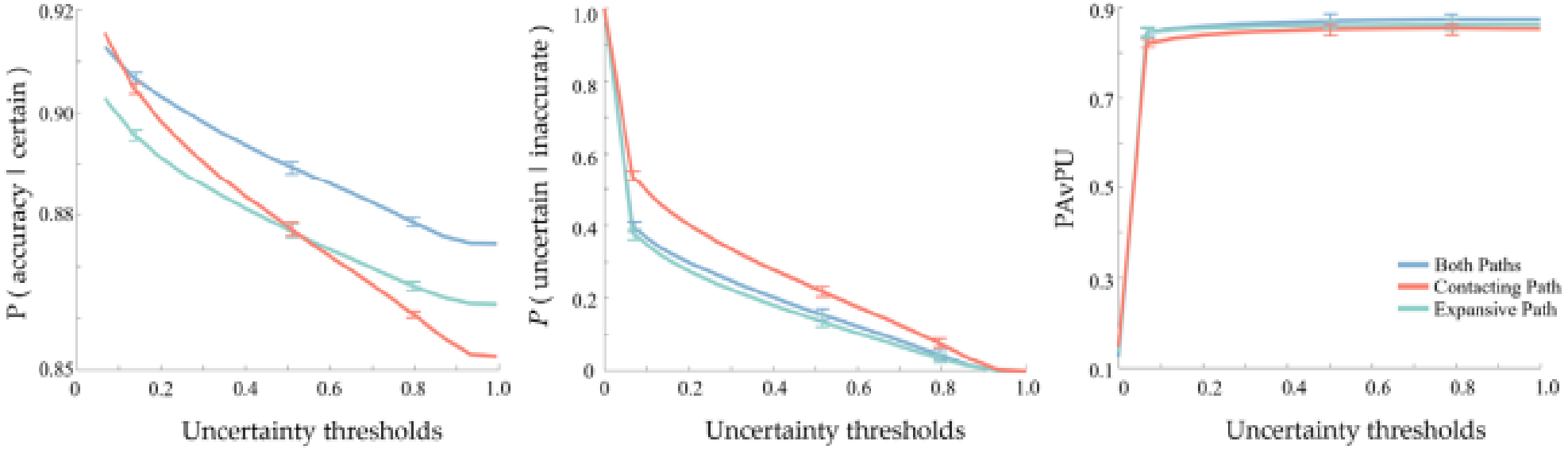


Fig. 6 Uncertainty estimation performance comparison of three models by adding MC dropout layers on contracting or expansive blocks or both, using *p* (accurate | certain), *p* (uncertain | inaccurate), and patch accuracy versus PAvPU.

Exploring uncertainty measures in convolutional neural network for semantic segmentation of oral cancer images – Song et al. 2022

Inherent Brain Segmentation Quality Control from Fully ConvNet Monte Carlo Sampling

Abhijit Guha Roy[1,2(✉)], Sailesh Conjeti[3], Nassir Navab[2,4], and Christian Wachinger[1]

[1] Artificial Intelligence in Medical Imaging (AI-Med), KJP, LMU München, Munich, Germany
abhijit.guha-roy@tum.de
[2] Computer Aided Medical Procedures, Technische Universität München, Munich, Germany
[3] German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany
Computer Aided Medical Procedures, Johns Hopkins University, Baltimore, USA

Abstract. We introduce inherent measures for effective quality control of brain segmentation based on a Bayesian fully convolutional neural network, using model uncertainty. Monte Carlo samples from the posterior distribution are efficiently generated using dropout at test time. Based on these samples, we introduce next to a voxel-wise uncertainty map also three metrics for structure-wise uncertainty. We then incorporate these structure-wise uncertainty in group analyses as a measure of confidence in the observation. Our results show that the metrics are highly correlated to segmentation accuracy and therefore present an inherent measure of segmentation quality. Furthermore, group analysis with uncertainty results in effect sizes closer to that of manual annotations. The introduced uncertainty metrics can not only be very useful in translation to clinical practice but also provide automated quality control and group analyses in processing large data repositories.

IN-DEPTH ANALYSIS: [4]

ABSTRACT

We introduce next to a voxel-wise uncertainty map also three metrics for structure-wise uncertainty. We then incorporate these structure-wise uncertainty in group analyses as a measure of confidence in the observation. The metrics are highly correlated to segmentation accuracy and therefore present an inherent measure of segmentation quality. The introduced uncertainty metrics can not only be very useful in translation to clinical practice but also provide automated quality control and group analyses in processing large data repositories.

UNCERTAINTY MEASURES

Voxel-wise uncertainty → entropy over all N MC probability maps p_S: the voxel-wise uncertainty is the sum over all structures s

$$U_s(\mathbf{x}) = -\sum_{i=1}^{N} p_s^i(\mathbf{x}) \log(p_s^i(\mathbf{x}))$$

Structure-wise uncertainty → three different strategies:

- Variation of the volume across the MC samples → $CV_s = \frac{\sigma_s}{\mu_s}$
- Overlap between samples, computing the average DICE score over all pairs of MC samples → $d_s^{MC} = E\left[\{Dice((S_i == s), (S_j == s))\}_{i \neq j}\right]$
- Mean voxel-wise uncertainty over the voxels which were labelled as s → $\mathcal{U}_s = E\left[\{U(\mathbf{x})\}_{\mathbf{x} \in \{S == s\}}\right]$

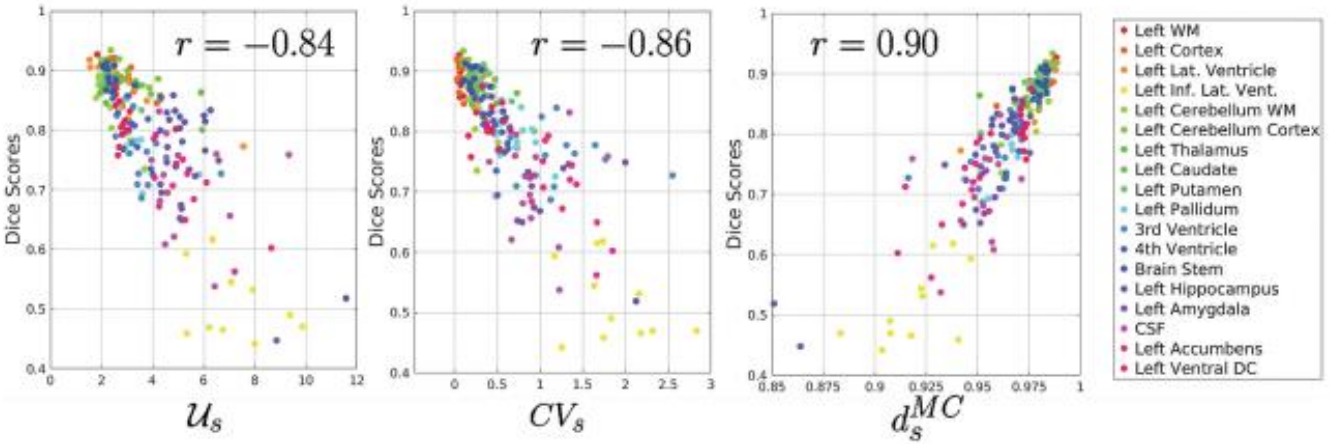
Inherent Brain Segmentation Quality Control from Fully ConvNet Monte Carlo Sampling – Roy et al. 2018

QUANTITATIVE ANALYSIS

We computed the correlation coefficient between the dice scores and the three types of structure-wise uncertainty.

Datasets	Mean dice score (DS)	Mean	Corr(·, DS)		
		CV_s	$\mathcal{U}_s$	CV_s	d_s^{MC}
MALC-15	$\mathbf{0.88} \pm 0.02$	0.38	−0.85	−0.81	**0.86**
ADNI-29	0.83 ± 0.02	0.46	−0.72	−0.71	**0.78**
CANDI-13	0.81 ± 0.03	0.54	−0.84	−0.86	**0.90**
IBSR-18	0.81 ± 0.02	0.57	−0.76	−0.76	**0.80**

The overall high correlation for d_S^{MC} indicates that it is suitable proxy for measuring segmentation accuracy without presence of ground truth annotations.



Scatter plot of three types of uncertainty and Dice scores on CANDI-13 dataset (one dot per scan and structure), with their corresponding correlation coefficient (r). For clarity, structures only on the left hemisphere are shown

Inherent Brain Segmentation Quality Control from Fully ConvNet Monte Carlo Sampling – Roy et al. 2018

Uncertainty Driven Pooling Network for Microvessel Segmentation in Routine Histology Images

Abstract. Lymphovascular invasion (LVI) and tumor angiogenesis are correlated with metastasis, cancer recurrence and poor patient survival. In most of the cases, the LVI quantification and angiogenic analysis is based on microvessel segmentation and density estimation in immunohistochemically (IHC) stained tissues. However, in routine H&E stained images, the microvessels display a high level of heterogeneity in terms of size, shape, morphology and texture which makes microvessel segmentation a non-trivial task. Manual delineation of microvessels for biomarker analysis is labor-intensive, time consuming, irreproducible and can suffer from subjectivity among pathologists. Moreover, it is often beneficial to account for the uncertainty of a prediction when making a diagnosis. To address these challenges, we proposed a framework for microvessel segmentation in H&E stained histology images. The framework extends DeepLabV3+ by using an improved dice coefficient based custom loss function and also incorporating an uncertainty prediction mechanism. The proposed method uses an aligned Xception model, followed by atrous spatial pyramid pooling for feature extraction at multiple scales. This architecture counters the challenge of segmenting blood vessels of varying morphological appearance. To incorporate uncertainty, random transformations are introduced at test time for a superior segmentation result and simultaneous uncertainty map generation, highlighting ambiguous regions. The method is evaluated using 1167 images of size 512 × 512 pixels, extracted from 13 WSIs of oral squamous cell carcinoma (OSCC) tissue at 20x magnification. The proposed net-work achieves state-of-the-art performance compared to current semantic segmentation deep neural networks (FCN-8, U-Net, SegNet and DeepLabV3+).

IN-DEPTH ANALYSIS: [5]

ABSTRACT

We proposed a framework for microvessel segmentation in H&E stained histology images. To incorporate uncertainty, random transformations are introduced at test time for a superior segmentation result and simultaneous uncertainty map generation, highlighting ambiguous regions.

UNCERTAINTY ESTIMATION

we present a framework for precise segmentation of microvessels in H&E stained histology images at multiple scales and resolutions by using an uncertainty aware spatial pyramid pooling deep neural network architecture.

Despite achieving state-of-the art performance in semantic segmentation, the deep networks typically do not inherently model the segmentation uncertainty. For this purpose, we apply random transformations to the images during test time, as a method to generate the approximate predictive distribution. Taking the **average** of these predictions of transformed images yields a superior segmentation and enables us to observe ambiguous areas, where the network is uncertain in a decision.
Transformations of the images comprehend median blur, Gaussian blur, rotation, flipping, Gaussian Noise.

Uncertainty is the variance within the samples:

$$\mu = -\frac{1}{m}\sum_{i}^{m} f(\delta_i(x); W); \quad \sigma = -\frac{1}{m}\sum_{i}^{m} f(\delta_i(x); W-\mu)^2 \tag{2}$$

where, μ is the prediction of microvessel segmentation, σ is the prediction uncertainty and m in the number of applied transformations. δ_i denotes the random transformation applied to the input image x. Taking the average of the prediction of transformed images give better segmentation.

Uncertainty Driven Pooling Network for Microvessel Segmentation in Routine Histology Images – Fraz et al. 2018

MILD-Net: Minimal Information Loss Dilated Network for Gland Instance Segmentation in Colon Histology Images

Simon Graham, Hao Chen, Jevgenij Gamper, Qi Dou,
Pheng-Ann Heng, David Snead, Yee Wah Tsang, Nasir Rajpoot

Abstract

The analysis of glandular morphology within colon histopathology images is an important step in determining the grade of colon cancer. Despite the importance of this task, manual segmentation is laborious, time-consuming and can suffer from subjectivity among pathologists. The rise of computational pathology has led to the development of automated methods for gland segmentation that aim to overcome the challenges of manual segmentation. However, this task is non-trivial due to the large variability in glandular appearance and the difficulty in differentiating between certain glandular and non-glandular histological structures. Furthermore, a measure of uncertainty is essential for diagnostic decision making. To address these challenges, we propose a fully convolutional neural network that counters the loss of information caused by max-pooling by re-introducing the original image at multiple points within the network. We also use *atrous* spatial pyramid pooling with varying dilation rates for preserving the resolution and multi-level aggregation. To incorporate uncertainty, we introduce random transformations during test time for an enhanced segmentation result that simultaneously generates an uncertainty map, highlighting areas of ambiguity. We show that this map can be used to define a

ACCEPTED MANU

IN-DEPTH ANALYSIS: [6]

ABSTRACT

To incorporate uncertainty, we introduce random transformations during test time for an enhanced segmentation result that simultaneously generates an uncertainty map, highlighting areas of ambiguity. This map can be used to define a metric for disregarding predictions with high uncertainty.

INTRODUCTION

Uncertainty of a prediction and the simultaneous segmentation of additional histological components, may give additional diagnostic power.

During uncertainty quantification, we apply random transformations to the input images as a method of generating the predictive distribution. This leads to a superior segmentation result and allows us to observe areas of uncertainty that may be clinically informative. Furthermore, we use this measure of uncertainty to rank images that should be prioritised for pathologist annotation.

MILD-Net: Minimal Information Loss Dilated Network for Gland Instance Segmentation in Colon Histology Images – Graham et al. 2018

RANDOM TRANSFORMATION SAMPLING FOR UNCERTAINTY QUANTIFICATION

To obtain the predictive distribution, we apply a random transformation Φ(x) on a sample of n images, where Φ performs a flip, rotation, Gaussian blur, median blur or adds Gaussian noise on input image x to obtain {Φ1, Φ2, ..., Φn}. Each image within the sample is then processed, where the mean of this processed sample gives the refined prediction and the **variance** gives the uncertainty.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} f(\Phi_i(\mathbf{x}); \boldsymbol{w}); \quad \sigma = \frac{1}{n}\sum_{i=1}^{n} (f(\Phi_i(\mathbf{x}); \boldsymbol{w}) - \mu)^2$$

Where μ defines the segmentation prediction, σ defines the uncertainty and n defines the number of transformations.

We propose a metric to give individual glands a score of uncertainty, based on the uncertainty map generated via random transformation sampling (RTS).

We suggest that it is reasonable to disregard segmented glands that have an uncertainty score above a given threshold. We first remove the boundaries by subtracting the predicted contours that have been output by the network and then calculate the object-level uncertainty score for each predicted instance k as:

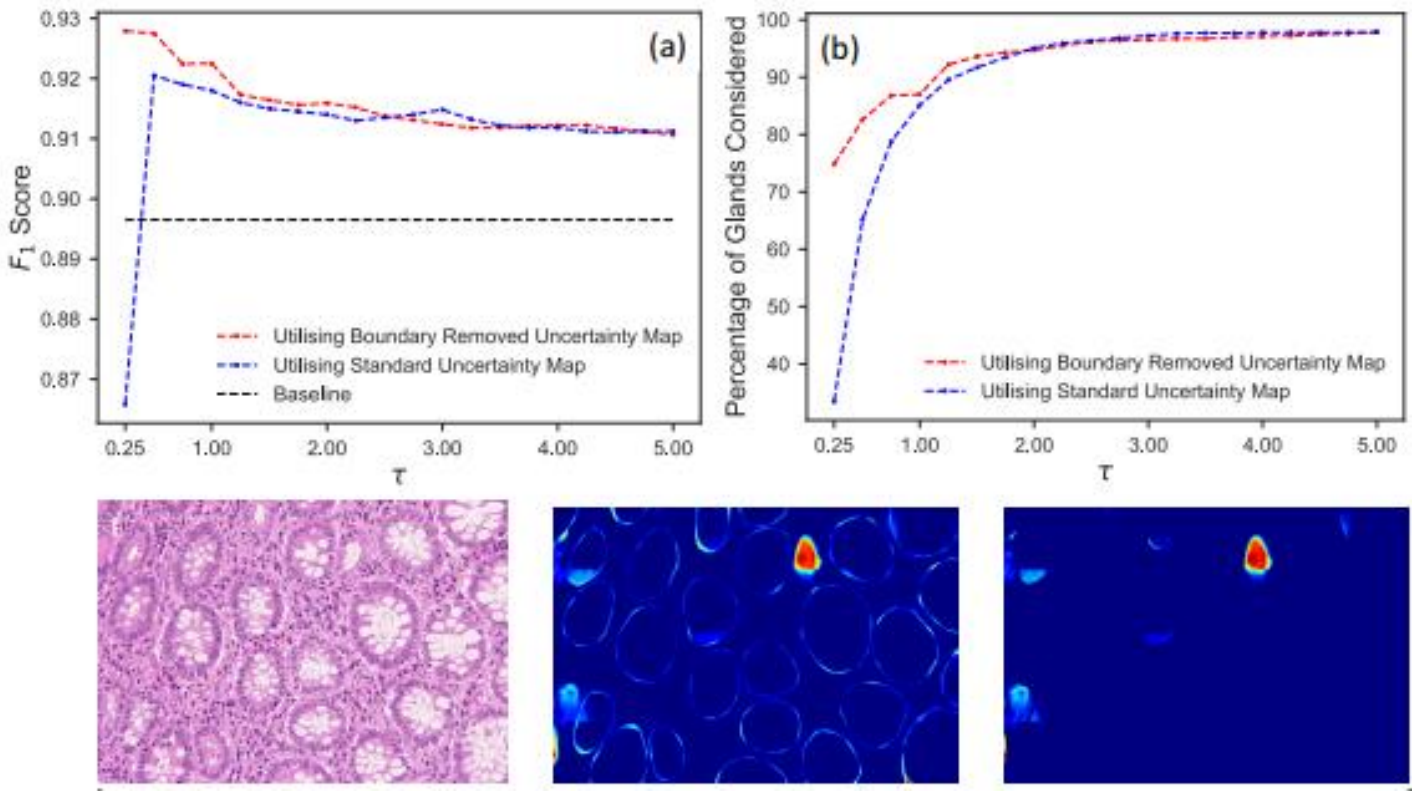
$$\tau_k = \frac{1}{n}\sum_{i=1}^{n} \hat{\sigma}\rho_{k,i}$$

where $\hat{\sigma}$ is the boundary removed uncertainty map and $\rho_{k,i}$ is the predicted binary output of pixel i within instance k. We define n as the number of pixels within predicted instance k.

[They select a **global threshold** for their uncertainty score]
[This study uses RTS, they segment the n images and they use the average as the segmentation of the original image]

MILD-Net: Minimal Information Loss Dilated Network for Gland Instance Segmentation in Colon Histology Images – Graham et al. 2018

RANDOM TRANSFORMATION SAMPLING FOR UNCERTAINTY QUANTIFICATION



Object-level uncertainty quantification.
(a) shows the F1 score as we disregard predictions with an uncertainty score τk greater than a given threshold τ.
(b) The percentage of total instances considered, given a threshold τ. For the red dashed line, we use the boundary removed uncertainty map, whereas for the blue dashed line we use the standard uncertainty map. The black horizontal line shows the F1 score when no glands with a high uncertainty are removed.
(c) from left to right: original image; uncertainty map σ; boundary removed uncertainty map $\hat{\sigma}$. For each instance k within $\hat{\sigma}$, an object-level uncertainty score τ is calculated

Removing the boundary allows us to preserve a larger proportion of the dataset when we are using lower thresholds for the removal of predictions with high uncertainty: utilising the boundary removed uncertainty map is more robust and can be effectively used to select predictions with low uncertainty. The intuition of disregarding glands with high uncertainty means that we should not extract any statistical measures from these disregarded glands. Therefore, when removing predicted instances with high uncertainty, we also remove the corresponding ground truth instance to obtain the above measures

MILD-Net: Minimal Information Loss Dilated Network for Gland Instance Segmentation in Colon Histology Images – Graham et al. 2018

Leveraging Uncertainty Estimates for Predicting Segmentation Quality

Terrance DeVries
University of Guelph and Vector Institute
terrance@uoguelph.ca

Graham W. Taylor
University of Guelph and Vector Institute
Canadian Institute for Advanced Research
gwtaylor@uoguelph.ca

Abstract

The use of deep learning for medical imaging has seen tremendous growth in the research community. One reason for the slow uptake of these systems in the clinical setting is that they are complex, opaque and tend to fail silently. Outside of the medical imaging domain, the machine learning community has recently proposed several techniques for quantifying model uncertainty (i.e. a model knowing when it has failed). This is important in practical settings, as we can refer such cases to manual inspection or correction by humans. In this paper, we aim to bring these recent results on estimating uncertainty to bear on two important outputs in deep learning-based segmentation. The first is producing spatial uncertainty maps, from which a clinician can observe where and why a system thinks it is failing. The second is quantifying an image-level prediction of failure, which is useful for isolating specific cases and removing them from automated pipelines. We also show that reasoning about spatial uncertainty, the first output, is a useful intermediate representation for generating segmentation quality predictions, the second output. We propose a two-stage architecture for producing these measures of uncertainty, which can accommodate any deep learning-based medical segmentation pipeline.

IN-DEPTH ANALYSIS: [7]

ABSTRACT

Estimate uncertainty to bear on two important outputs in deep learning-based segmentation. The first is producing spatial uncertainty maps, from which a clinician can observe where and why a system thinks it is failing. The second is quantifying an image-level prediction of failure, which is useful for isolating specific cases and removing them from automated pipelines. We also show that reasoning about spatial uncertainty, the first output, is a useful intermediate representation for generating segmentation quality predictions, the second output..

ESTIMATING SEGMENTATION QUALITY WITH UNCERTAINTY INFORMATION

The segmentation network takes in some image x, and produces two outputs: class prediction logits ρ and corresponding uncertainty (or confidence) estimates z (z is uncertainty map).
A second network g is then trained to predict the quality of the segmentation v^.
Segmentation quality measurement can be any segmentation-based evaluation metric, or even multiple metrics predicted simultaneously.

In the case that f performs very well on the training set, a holdout set may be necessary, as the lack of examples of poor segmentations will bias g towards always predicting that the segmentation is good.

UNCERTAINTY ESTIMATION

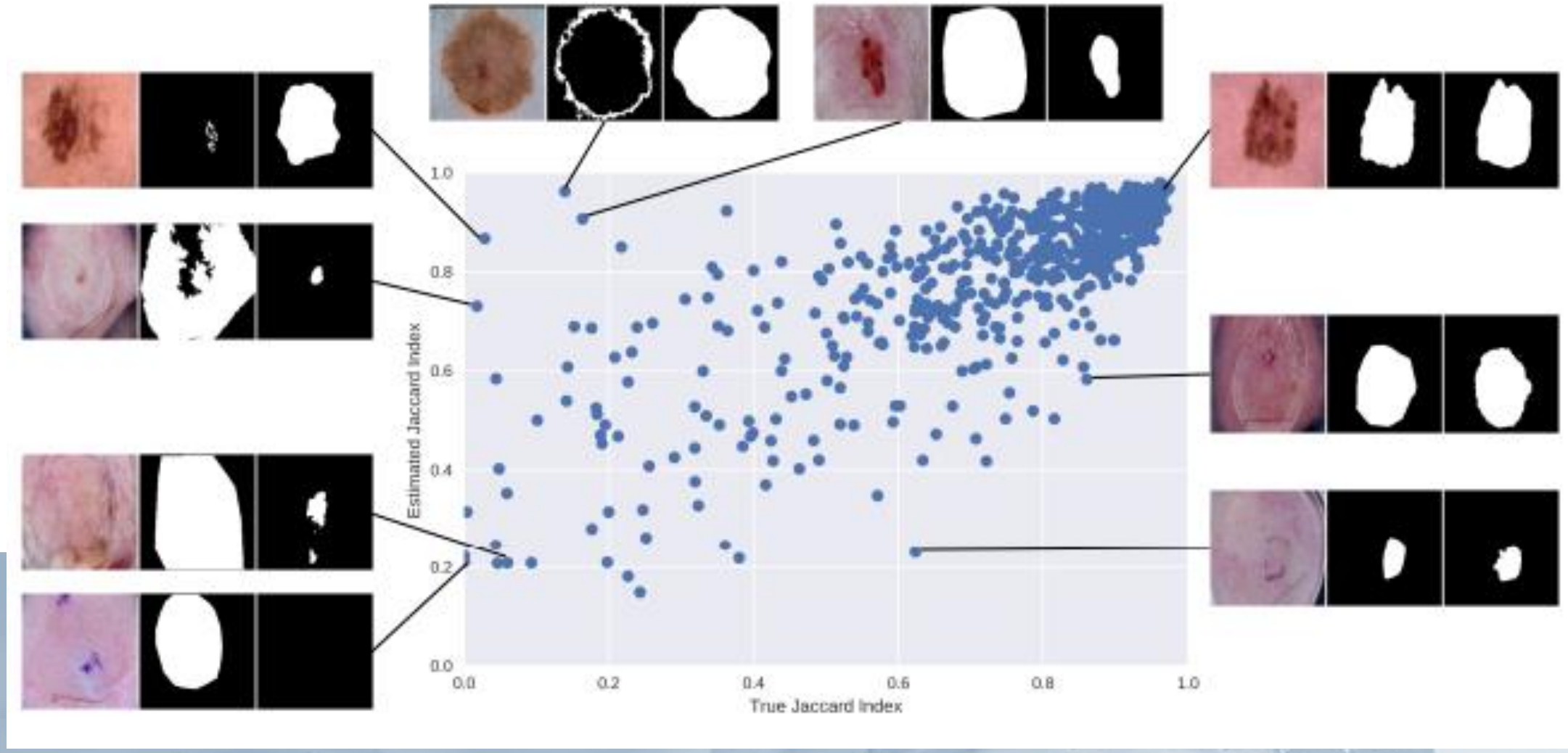
Model uncertainty z is estimated by calculating the **entropy** of the averaged probability vector across the class dimension:

$$z = -\sum_{c=1}^{C} p_c \log p_c$$

Leveraging Uncertainty Estimates for Predicting Segmentation Quality– DeVries et Taylor 2018

SEGMENTATION QUALITY PREDICTION

We use a variety of metrics to evaluate how well our models can predict the quality of segmentations: RMSE, detection error, AUROC, and AUPR.



Leveraging Uncertainty Estimates for Predicting Segmentation Quality– DeVries et Taylor 2018

SEGMENTATION QUALITY PREDICTION

Treating uncertainty explicitly improves performance significantly compared to the case with no uncertainty information.

segmentation quality estimates rarely fall below 0.2 for any method. This is likely caused by the rarity of poor quality segmentations in the dataset used to train the segmentation quality estimation network, since it was trained on the same dataset as the original segmentation network.

Conclusion: making spatial uncertainty explicit aided in predicting a measure of segmentation quality, the Jaccard index.

Leveraging Uncertainty Estimates for Predicting Segmentation Quality – DeVries et Taylor 2018

ADNet++: A few-shot learning framework for multi-class medical image volume segmentation with uncertainty-guided feature refinement

Stine Hansen *, Srishti Gautam, Suaiba Amina Salahuddin, Michael Kampffmeyer, Robert Jenssen

ABSTRACT

A major barrier to applying deep segmentation models in the medical domain is their typical data-hungry nature, requiring experts to collect and label large amounts of data for training. As a reaction, prototypical few-shot segmentation (FSS) models have recently gained traction as data-efficient alternatives. Nevertheless, despite the recent progress of these models, they still have some essential shortcomings that must be addressed. In this work, we focus on three of these shortcomings: (i) the lack of uncertainty estimation, (ii) the lack of a guiding mechanism to help locate edges and encourage spatial consistency in the segmentation maps, and (iii) the models' inability to do one-step multi-class segmentation. Without modifying or requiring a specific backbone architecture, we propose a modified prototype extraction module that facilitates the computation of uncertainty maps in prototypical FSS models, and show that the resulting maps are useful indicators of the model uncertainty. To improve the segmentation around boundaries and to encourage spatial consistency, we propose a novel feature refinement module that leverages structural information in the input space to help guide the segmentation in the feature space. Furthermore, we demonstrate how uncertainty maps can be used to automatically guide this feature refinement. Finally, to avoid ambiguous voxel predictions that occur when images are segmented class-by-class, we propose a procedure to perform one-step multi-class FSS. The efficiency of our proposed methodology is evaluated on two representative datasets for abdominal organ segmentation (CHAOS dataset and BTCV dataset) and one dataset for cardiac segmentation (MS-CMRSeg dataset). The results show that our proposed methodology significantly (one-sided Wilcoxon signed rank test, $p < 0.05$) improves the baseline, increasing the overall dice score with +5.2, +5.1, and +2.8 percentage points for the CHAOS dataset, the BTCV dataset, and the MS-CMRSeg dataset, respectively.

IN-DEPTH ANALYSIS: [8]

INTRODUCTION

l few-shot segmentation (FSS) models have recently shown promise as data efficient alternatives to solving this task by using a small set of labeled examples to extract class-wise prototypes that can be leveraged to segment objects in new images.
existing medical FSS models do not provide any measure of uncertainty for their predictions.
Current medical FSS methods only focus on binary foreground/background segmentation and are forced to segment the images class-by-class.
To facilitate the computation of uncertainty maps in prototypical FSS models we propose a modified prototype extraction module that introduces a Bernoulli distributed variable for each voxel location in the feature representation. Uncertainty maps are then based on the predictive distribution estimated from a set of prototypes extracted by this proposed module: uncertainty maps can be used to automatically guide this feature refinement.

UNCERTAIN ESTIMATION AND REFINEMENT MODULE

Model uncertainty map can be estimated as the predictive [Shannon] entropy.
Dynamic refinement β can be derived from the uncertainty map, so that no labelled data is required to determine that value.

A dynamic β has the potential to increase the flexibility of the feature refinement by allowing different voxels to move with different step lengths, depending on the model's uncertainty: For voxels in regions where the model is unsure about its initial prediction, we will pay more attention to the region information in the input space

ADNet++: A few-shot learning framework for multi-class medical image volume segmentation with uncertainty-guided feature refinement – Hansen et al. 2023

UNCERTAINTY MAPS

to quantify the fidelity of the estimated uncertainty maps, we start by removing all voxels in the predictions and successively add voxels according to their estimated uncertainty, starting with the least uncertain voxels. Segmentation performance decreases for all classes as more uncertain voxels are included: uncertainty maps canbe used to quantify how much a prediction can be trusted.

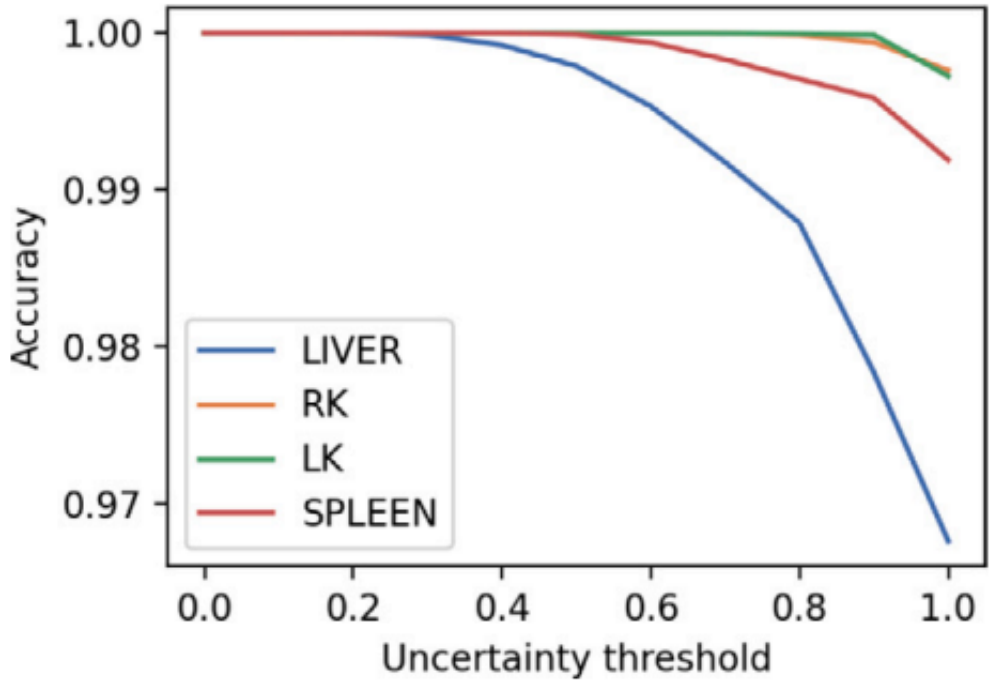


Fig. 5. Relationship between accuracy and estimated uncertainty. By successively including more uncertain voxels, the segmentation accuracy decreases.

ADNet++: A few-shot learning framework for multi-class medical image volume segmentation with uncertainty-guided feature refinement – Hansen et al. 2023

A Quantitative Comparison of Epistemic Uncertainty Maps Applied to Multi-Class Segmentation

Abstract

Uncertainty assessment has gained rapid interest in medical image analysis. A popular technique to compute epistemic uncertainty is the Monte-Carlo (MC) dropout technique. From a network with MC dropout and a single input, multiple outputs can be sampled. Various methods can be used to obtain epistemic uncertainty maps from those multiple outputs. In the case of multi-class segmentation, the number of methods is even larger as epistemic uncertainty can be computed voxelwise per class or voxelwise per image.

This paper highlights a systematic approach to define and quantitatively compare those methods in two different contexts: class-specific epistemic uncertainty maps (one value per image, voxel and class) and combined epistemic uncertainty maps (one value per image and voxel). We applied this quantitative analysis to a multi-class segmentation of the carotid artery lumen and vessel wall, on a multi-center, multi-scanner, multi-sequence dataset of Magnetic Resonance (MR) images. We validated our analysis over 144 sets of hyperparameters of a model.

Our main analysis considers the relationship between the order of the voxels sorted according to their epistemic uncertainty values and the misclassification of the prediction. Under this consideration, the comparison of combined uncertainty maps reveals that the multi-class entropy and the multi-class mutual information statistically out-perform the other combined uncertainty maps under study (the averaged entropy, the averaged variance, the similarity Bhattacharya coefficient and the similarity Kullback-Leibler divergence). In a class-specific scenario, the one-versus-all entropy statistically out-performs the class-wise entropy, the class-wise variance and the one versus all mutual information. The class-wise entropy statistically out-performs the other class-specific uncertainty maps in term of calibration. We made a python package available to reproduce our analysis on different data and tasks.

IN-DEPTH ANALYSIS: [9]

ABSTRACT

Two different contexts: class-specific epistemic uncertainty maps (one value per image, voxel and class) and combined epistemic uncertainty maps (one value per image and voxel).
Multi-class entropy and the multi-class mutual information statistically out-perform the other combined uncertainty maps under study (the averaged entropy, the averaged variance, the similarity Bhattacharya coefficient and the similarity Kullback-Leibler divergence). In a class-specific scenario, the one-versus-all entropy statistically out-performs the class-wise entropy, the class-wise variance and the one versus all mutual information. The classwise entropy statistically out-performs the other class-specific uncertainty maps in term of calibration.

INTRODUCTION

A multi-class segmentation task provide a larger set of epistemic uncertainty maps than a single-class segmentation task as it is possible to compute combined epistemic uncertainty maps (one uncertainty map per voxel, per image) or class-specific epistemic uncertainty maps (one uncertainty map per class, per voxel, per image).. there is limited research focusses on analysing the extra options resulting from the multi-class setting.

1. systematic approach to characterize both class-specific epistemic uncertainty maps, combined epistemic uncertainty maps separating the epistemic uncertainty map into a uncertainty measure and an aggregation method
2. quantitatively and statistically compare the ability of those different epistemic uncertainty maps to assess misclassification on a segmentation of the carotid artery on a multi-center, multi-scanner, multi-sequence dataset of MR images
3. compare our evaluation of
4. class-specific epistemic uncertainty derived from the MC dropout technique to the one proposed in the BRATS challenge (Menze et al., 2014) in a multi-class segmentation setting

A Quantitative Comparison of Epistemic Uncertainty Maps Applied to Multi-Class Segmentation – Camarasa et al. 2021

UNCERTAINTY ANALYSIS

In medical imaging, a straightforward approach is to consider the **inter-observer variability as "ground truth" aleatoric uncertainties**.

QUANTIFYING UNCERTAINTY

To quantify the uncertainty of the prediction distribution per voxel, one can either determine the uncertainty of a class-specific prediction distribution, study the similarity between the different class-specific prediction distributions or directly determine the uncertainty of the prediction distribution.

DESCRIPTION MEASURES → uncertainty of a single class-specific prediction distribution. Examples are variance or entropy
SIMILARITY MEASURES → An approach to quantify the uncertainty is to measure the similarity of two class-specific prediction distributions (where c' and c" are two distinct output classes). The more those two distributions overlap, the more similar they are and the more difficult it is to determine the predicted class, which makes the outcome more uncertain. Examples are Bhattacharya Coefficient (BC) and Kullback-Leibler Divergence (KL)
MULTI-CLASS MEASURES → capture the uncertainty of the prediction distributions in one measure of uncertainty. Examples: entropy of MC dropout distribution (entropy of the mean of the prediction distribution), Mutual Information (MI) between the MC dropout distribution over the model parameters.

A Quantitative Comparison of Epistemic Uncertainty Maps Applied to Multi-Class Segmentation – Camarasa et al. 2021

AGGREGATION METHODS

Now that the different uncertainty measures are defined, it is important to define the different aggregation methods to obtain the epistemic uncertainty maps from those measures of uncertainty. Two families: combined aggregation methods and class-specific aggregation methods.

COMBINED AGGREGATION METHODS

They use the description measure per voxel per class and average those descriptions per voxel over the classes.

AVERAGED AGGREGATION METHOD: $s_D^A(y_j) = \frac{1}{C}\sum_{c=1}^{C} D(y_j^c)$ where D is description measure (entropy or variance)

SIMILARITY AGGREGATION METHOD: $s_S^T(y_j) = S\left(y_j^{c1}, y_j^{c2}\right)$ where c1 and c2 are respectively the most and the second most probable classes of the voxel j and S is a similarity measure (BC or KL)

MULTICLASS DESCRIPTION MEASURE: $s_M^A(y_j) = M(y_j)$ where M is a multi-class measure (entropy or MI)

A Quantitative Comparison of Epistemic Uncertainty Maps Applied to Multi-Class Segmentation – Camarasa et al. 2021

CLASS-SPECIFIC AGGREGATION METHODS

The most direct aggregation method to obtain an uncertainty map per class is to compute a description measure (either distribution variance or distribution entropy) per voxel per class.

This approach is the DESCRIPTION AGGREGATION METHOD: $s_D^{CD}\left(y_j^c\right) = D\left(y_j^c\right)$

Another approach is the ONE VERSUS ALL AGGREGATION METHOD: $s_M^{1vA}\left(y_j^c\right) = M\left(\left(y_j^c, \overline{y_j^c}\right)\right)$

That consists in the application of a multi-class measure to the distribution of the class under study and the sum of the distributions of the other classes.

A Quantitative Comparison of Epistemic Uncertainty Maps Applied to Multi-Class Segmentation – Camarasa et al. 2021

COMBINED EVALUATION

Considering uncertainty as a score that predicts misclassification leads to a redefinition of the notions of true and false positives and negatives in an uncertainty context. A voxel is considered misclassified when its predicted class and its ground truth class mismatch.
Once an uncertainty map is thresholded at a value τ_u, one can define four types of:

- misclassified and uncertain (UTP(τu) in a sense that the uncertainty of the voxel accurately predicts its misclassification)
- misclassified and certain (UFN(τu))
- correctly classified and uncertain (UFP(τu))
- correctly classified and certain (UTN(τu))

For a given value of the uncertainty threshold τ_u, it is possible to compute the precision and the recall of uncertainty as a misclassification predictor following:

$$\mathrm{UPr}(\tau_u) = \frac{\mathrm{UTP}(\tau_u)}{\mathrm{UTP}(\tau_u)+\mathrm{UFP}(\tau_u)} \qquad \mathrm{URc}(\tau_u) = \frac{\mathrm{UTP}(\tau_u)}{\mathrm{UTP}(\tau_u)+\mathrm{UFN}(\tau_u)}$$

The main characteristic of this performance measure is its independence from uncertainty map calibration. Only the order of the voxels sorted according to their epistemic uncertainty values matters as this performance measure is invariant by strictly monotonic increasing transformation.

A Quantitative Comparison of Epistemic Uncertainty Maps Applied to Multi-Class Segmentation – Camarasa et al. 2021

CLASS-SPECIFIC EVALUATION

For each filtering threshold τf , the voxels with uncertainty values in the uncertainty map u above the threshold τf are removed from the prediction. Then, one can derive the ratio of filtered true positives (tpr(τf)), the ratio of filtered true negatives (tnr(τf)) and the filtered Dice score (Dice(τf)). The BraTS uncertainty performance measure (BRATS-UNC) integrates these three measurements over the range of values of the uncertainty map, as follows:

$$\text{BRATS-UNC} = \frac{1}{3}\int_{\min(u)}^{\max(u)} \text{Dice}(\tau_f) + (1 - \text{tnr}(\tau_f)) + (1 - \text{tpr}(\tau_f))\text{d}\tau_\text{f}$$

BRATS-UNC is slightly altered compared to its original formulation by Mehta et al. (2020) to generalise to the case of non-normalised uncertainty maps.

STATISTICAL SIGNIFICANCE →

COMBINED UNCERTAINTY MAPS EVALUATION

Statistically significant best results are observed with the multi-class entropy and the multi-class mutual information. The statistically significant worst result is observed with the similarity KL

Table 2: 95% credible interval of the posterior distribution of $p_{A>B}$ for the combined AUC-PR performance measure and the combined uncertainty maps pairs A and B. The rows correspond to the uncertainty maps A and the columns to the uncertainty map B. The statistically significant differences are reported in bold ($0.5 \notin I_{95\%}$). $p_{A>B} > 0.5$ means that uncertainty map A is better than uncertainty map B and the contrary if $p_{A>B} < 0.5$ (Av = Averaged, Mu = Multi-class, Sim = Similarity)

	Av Variance	Mu Entropy	Mu MI	Sim BC	Sim KL
Av Entropy	[0.43, 0.59]	**[0.20, 0.35]**	**[0.29, 0.45]**	**[0.89, 0.97]**	**[0.90, 0.98]**
Av Variance		**[0.22, 0.37]**	**[0.30, 0.46]**	**[0.93, 0.99]**	**[0.91, 0.98]**
Mu Entropy			[0.49, 0.65]	**[0.93, 0.99]**	**[0.95, 1.00]**
Mu MI				**[0.89, 0.97]**	**[0.91, 0.98]**
Sim BC					**[0.64, 0.79]**

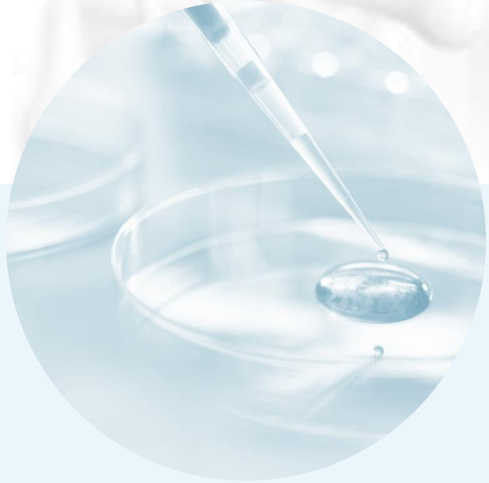
A Quantitative Comparison of Epistemic Uncertainty Maps Applied to Multi-Class Segmentation – Camarasa et al. 2021

DISCUSSION

Multi-class entropy and multi-class mutual information statistically out-perform the averaged variance and the averaged entropy which in turn statistically out-perform the similarity BC which out-performs the similarity KL. This analysis highlights the good performances of the multi-class aggregation method and the averaged aggregation method compared with the similarity aggregation method which is coherent with their extensive use in the literature. The superiority of the multi-class aggregation method over the class-wise aggregation method seems also coherent as the averaged aggregation method consists of a combination of class-specific uncertainty maps where the multi-class aggregation method is designed for multi-class problems.

The entropy uncertainty measure out-performs the other uncertainty measures with a multiclass aggregation method in a combined scenario and with a one versus all aggregation method in a class-specific scenario.

In this article, we proposed a systematic approach to characterize epistemic uncertainty maps that can be used in a multi-class segmentation context. We also proposed a methodology to analyse the quality of multi-class epistemic uncertainty maps.



GRAZIE

WILLIAM BASSOLINO