Estimating uncertainty of deep learning-based segmentation for prostate cancer radiotherapy

Maria Leousi

Biomedical Sciences, Utrecht University

***Abstract*—In radiotherapy, structure delineation is crucial to ensure accurate irradiation of the target sparing the adjacent organs. Recently, deep learning-based segmentations achieved encouraging results possibly speeding up the process contrary to manual contouring. Nonetheless, such methods can be overconfident even though their predictions may be incorrect. In this study, we explored four ways to estimate segmentation uncertainty for prostate cancer patients, investigating whether uncertainty can serve as surrogate for the performance of the network. Finally, we inspected the robustness of our approach for out-of-distribution data. A relation between accuracy and uncertainty was observed, suggesting that uncertainty quantification can provide valuable information in clinical settings. However, a discrimination between in- and out-of-distribution data was noticed, implying that a model tested on inputs distant from the training distribution, did not necessarily produce increased uncertainty.**

***Index Terms*—uncertainty estimation, deep learning, radiotherapy, prostate delineation**

I. Introduction

In light of the critical role of medical diagnosis and treatment therapy, structure localization and organ delineation have attracted widespread attention during the past decades [1, 2]. Traditionally, segmentation of the different regions of interest (ROIs) is performed manually [2]. However, this is time-consuming, possibly introducing treatment delays, and presumably comprising large inter-observer variabilities [2, 3]. In an attempt to alleviate the delays caused by manual delineations, make the procedure less error-prone and decrease human intervention, automatic segmentation algorithms are introduced [1, 4, 5, 6]. Akin to manual contouring, deep learning (DL) methods provide precise and robust structure delineations while at the same time they might accelerate the process [2, 5, 7, 8]. As a result, DL has become a broadly recognized and used approach for automatic segmentation over the past years [9, 10, 11, 12, 13].

In radiotherapy (RT), patients undergo a personalized irradiation treatment according to their anatomical and structural composition [2]. During the workflow, magnetic resonance imaging (MRI) assists structure delineations, owing to its great soft-tissue contrast, whereas the dose calculation is performed on computed tomography (CT) scans [1, 5, 14]. Therefore, providing properly delineated contours for both the target and the normal structures, is rather critical for the patients [12, 15]. Automatic segmentation with DL could reduce inter-observer variability and speed up the procedure, with respect to manual contouring [2, 8]. However, the predicted segmentations are revised by experts to ensure their accuracy, since poor predictions may influence the treatment planning and therapy response. Specifically for prostate cancer RT, various methods for automatic segmentation of the target and the surrounding organs-at-risk (OARs) have been published to this day [5, 16].

Generaly, DL-based models are prone to overconfident predictions although their results might be incorrect [11, 12, 15]. Faulty predictions introduce problems and possibly fatal errors in applications for which accuracy is essential, like medical diagnosis and treatment therapy [12, 15, 17, 18]. On that account, quantifying uncertainty of a DL architecture may provide an approach for improving the procedure by alleviating the manual revision and correction [8, 11]. In addition, uncertainty might be used to facilitate quality assurance of the segmentation in clinical settings, if a relation is found with the network's accuracy. In clinic, no information about the performance of the automatic pipeline is available thus, having a surrogate for it is convenient for researchers. Since DL models accomplish good performance on datasets similar to those used during training, their ability to generalize well on discrepant data is another point of interest. Thus, analysis of the ambiguity generated by out-of-distribution data in DL methods might provide additional insights on the network itself. Although pivotal, uncertainty estimation of DL segmentation models and their applications in RT has not been examined to a large extent yet.

In this work, we consider the task of prostate gland segmentation. We adopt a V-net architecture [13] for structure delineation including a pipeline for uncertainty estimation on the predictions. We also investigate the correlation between uncertainty and various performance metrics in order to assess whether uncertainty can be used as a quality measure of the segmentation. Lastly, an exploration of our model's robustness along with the robustness of using uncertainty to signal the quality of the segmentation is investigated by introducing out-of-distribution input data.

II. Preliminaries and Related Work

Uncertainty quantification of DL methods was studied extensively by Kendal and Gal [20] in 2017. According to the authors, uncertainty can be divided into two main types, pursuant to their source of origin: the *aleatoric* (or data uncertainty) and the *epistemic* (or model uncertainty) components. Aleatoric uncertainty reflects the intrinsic noise in the dataset, while the epistemic component represents the uncertainty of the

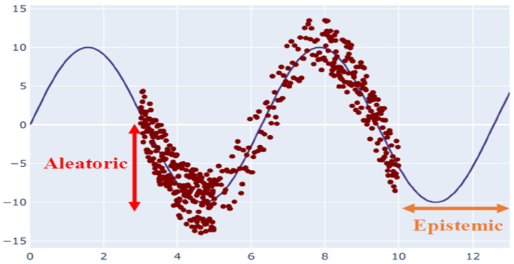

Fig. 1. Schematic representation of the differences between aleatoric and epistemic uncertainty components. Aleatoric uncertainty is an inherent characteristic of the data so it's irreducible even if more data is provided to the network. Epistemic uncertainty is reducible and decreases with an increasing amount of data. Figure encountered in the work of Adbar *et al.* [19].

model's parameters. The epistemic component can be eliminated when giving enough data to the model. Additionally, aleatoric uncertainty can be discriminated even further into homoscedastic and heteroscedastic. Heteroscedastic aleatoric uncertainty considers input-dependent noise for the entire dataset, whereas homoscedastic implies uniform noise across the data. This decomposition on the concept of uncertainty can be very convenient for researches and developers who wish to analyze their results or design their model focusing on the influence of a specific uncertainty type. Nevertheless, it should be noted that this discrimination between aleatoric and epistemic uncertainty is not always clear and becomes strongly dependent on the task and the settings employed [21]. This means that epistemic may transform to aleatoric and vice-versa, and thus, they should not be considered as absolute concepts [21]. A visual representation of the differences between the two uncertainty factors can be seen in Fig. 1.

According to the review of Adbar *et al.* [19], more than 2500 papers have been published on uncertainty quantification in the last decade, irrespective of the field of interest. The main idea behind the implementation of a model that captures uncertainty, leans on the Bayesian theorem, according to which posterior beliefs are affected by prior beliefs. The neural network models (NNs) that are trained with a Bayesian approach are referred to as Bayesian Neural Networks (BNNs) [19, 21, 22]. Such networks place a prior distribution p over their parameters ω (from now on referred to as $p(\omega)$) and encode the training data $D = \{x, y\}$ via a likelihood function $p(D_y|D_x, \omega)$, where D_x and D_y correspond to the training data and labels, respectively. Softmax can be used as the likelihood function for classification tasks, $p(D_y|D_x, \omega) = \mathit{Softmax}(output)$, while the Bayesian posterior distribution $p(\omega|D)$ is captured using the Bayes' theorem, once implying that the parameters and the inputs are independent [22]:

$$p(\omega|D) = \frac{p(D_y|D_x, \omega)p(\omega)}{\int_\omega p(D_y|D_x, \omega')p(\omega')d\omega'} \propto p(D_y|D_x, \omega)p(\omega) \tag{1}$$

Inference or marginalisation is the process of calculating the marginal probability distribution of the output y^*, concerning the posterior distribution $p(\omega|D)$ and given a test input x^* [19, 22]:

$$p(y^*|x^*, D) = \int_\omega p(y^*|x^*, \omega')p(\omega'|D)d\omega' \tag{2}$$

A. *Variational Inference*

The computation of the true posterior in an analytical way is infeasible [22], however, a numerical solution has been suggested to solve this problem. This is achieved through the variational inference (VI) approach, which finds a distribution $q_\theta(\omega)$ that approximates the true posterior distribution $p(\omega|D)$, by being as close as possible to it [17, 19, 20, 21, 22]. This is ensured by measuring the Kullback-Leibler (KL) divergence, a measure of similarity between two distributions, defined as:

$$KL(q_\theta(\omega)|p(\omega|D)) = \int_\omega q_\theta(\omega') \log \frac{q_\theta(\omega')}{p(\omega'|D)} d\omega' \tag{3}$$

Since the aim is to decrease this similarity, the overall goal is the minimization the above-mentioned equation, with respect to θ. Alternatively, the KL divergence minimization can be rearranged into the evidence lower bound (ELBO) score maximization, as referred in [22]:

$$L_{VI} := \int_\omega q_\theta(\omega') \log(p(D_y|D_x, \omega'))\, d\omega' - KL(q_\theta(\omega)|p(\omega)) \tag{4}$$

The solution derived by this process is a distribution $q_\theta(\omega)$ that describes the data well, while also being near the prior distribution [17, 19].

Monte Carlo Dropout

Dropout VI constitutes one of the most widely used techniques to approximate the posterior and estimate model uncertainty [19]. When used during training, dropout is a regularization technique that randomly ignores nodes of the network. Gal and Ghahramani [17] suggested that dropout can also be used to approximate Bayesian inference. In practice, dropout is applied during test time and a finite number of predictions is sampled and averaged. Such an approach is known as Monte Carlo (MC) dropout, due to the stochasticity of the forward passes. In semantic segmentation, where the probabilities of the background and the foreground can be generated, epistemic uncertainty is measured by computing the Shannon's entropy on the probabilities of the latter [20]. Please refer to Supplementary V for more details about the mathematical implementation of the MC dropout technique.

A variety of papers available in literature have used dropout to estimate epistemic uncertainty for segmentation, as summarized in Table I. U-net-like architectures are very popular for this purpose and are mainly employed by studies that tackle medical imaging problems, such as [11], [25], [26] and [28]. DenseNet is encountered in [24], while a fully convolutional version of it is implemented in [20]. Finally, a HighResNet architecture is exploited in [23]. The majority of studies (6 out of 9) incorporate 50% dropout during test time except for [20]

TABLE I
SUMMARY OF THE PAPERS DISCUSSED THAT USED MC DROPOUT, IN CHRONOLOGICAL ORDER. THE ARCHITECTURE, THE DROPOUT PROBABILITY AND THE NUMBER OF MC SAMPLES INVOLVED ARE REPORTED. THE DROPOUT RATE USED DURING TRAINING AGREES WITH THE ONE USED IN TEST TIME FOR EACH STUDY.

Study	Model	Dropout Probability (%)	MC samples
Kendal,2017 [20]	Fully conv. DenseNet	20	50
Bragman,2018 [23]	HighResNet	50	20
DeVries,2018 [11]	U-net	50	20
Jungo,2018 [24]	DenseNet	20	20
Hu,2019 [25]	probabilistic U-net	50	n.a.
Do,2020 [26]	probabilistic U-net	50	1115
Meijenik,2020 [27]	NN: 2 layers/100 nodes	50	100
Jungo,2020 [28]	U-net	5/50	n.a.
Ståhl,2020 [29]	NN: 2 layers/800 nodes	60	n.a.

and [24] where 20% is applied and [29] where 60% was used. In [28] experiments with 5% dropout are included in parallel with the ones of 50%. At the same time, one-third of the papers include 20 MC samples, while this quantity experiences a considerable rise in [27] and [26], where 100 and 1115 MC samples are used respectively. Finally, three papers ([25], [28] and [29]) do not mention this information.

B. Model ensembling

Lakshminarayanan *et al.* [30] suggested that model ensembling could serve as an alterantive approach to BNNs. This pathway introduces the idea that a collection of models would produce more robust predictions contrary to those acquired by a single method [12, 19, 30], while providing proper model uncertainty estimates. The average of the predictions acquired by each model in the ensemble serves as the final predicted segmentation of the network, whereas their variance highlights the epistemic uncertainty [30]. The similarity with the Bayesian approach presented in the previous subsection, is worthy of note [29].

In literature, various combinations of NN architecture types with numerous models combined together have been tested when using the ensemble technique for uncertainty estimation, as showed in Table II. When dealing with medical imaging data, a very common approach is to employ together multiple U-net architectures, as presented in [12] and [28] where 2 and 10 models were applied. In case of computer vision application, however, simple NN architectures with a number of models ranging from 1 to 15 ([30] and [27]) and even 50 ([29]) seem to serve the purpose of capturing epistemic uncertainty. Nevertheless, apart from NNs a combination of 2 DenseNets is also considered effective, as investigated in [12].

TABLE II
SUMMARY OF THE PAPERS DISCUSSED THAT USED MODEL ENSEMBLING, IN CHRONOLOGICAL ORDER. THE NUMBER OF NETWORKS AND THE ARCHITECTURES INVOLVED ARE MENTIONED.

Study	Number of networks	Architectures
Lakshminarayanan,2016 [30]	1-15	NN: 3 layers/200 nodes
Meijerink,2020 [27]	5	NN: 2 layers/100 nodes
Jungo,2020 [28]	10	U-net
JooLee,2020 [12]	2/2	DenseNet/U-net
Ståhl,2020 [29]	50	NN: 2 layers/320 nodes

C. Heteroscedastic aleatoric uncertainty

Epistemic component is considered the most interesting type because it reflects the confidence of the model itself [20]. Nevertheless, quantifying the intrinsic uncertainty of the data is extremely important as well, since it could give a hint about the ambiguity of the input. More specifically, uncertain results would be produced if poor image quality is fed into the network or if noisy inputs are present [7, 8]. This is why is meaningful to assess the aleatoric component along with its epistemic counterpart.

According to Kendal and Gal [20], homoscedastic aleatoric uncertainty is computed by tuning the constant noise σ, which is present in the dataset. Heteroscedastic uncertainty accounts for variations in the value of σ amongst the inputs. Therefore, it requires more adaptations on the original architecture. These adaptations consist of learning the variance throughout training rather than computing it at the end of the inference, like what MC dropout or model averaging approaches do. In practice, information about the variance of the input is lacking, thus the goal is reached in an unsupervised way through the computation of the loss. The model is modified to produce a two-fold result for every input: the logits $\hat{y}$ and the variance σ^2, which corresponds to the aleatoric uncertainty [20]. In classification settings, the variance is corrupted by Gaussian noise $N(0, \sigma^2 I)$ and then added to the logits. The result is processed through a softmax function and the final class probabilities are acquired by integration using multiple stochastic forward passes [20]. The objective function finally used can vary. For instance, negative log-likelihood with a softmax likelihood function scaled by the uncertainty term σ^2 is included in [23], while the cross-entropy loss is referred in [28], [25] and [11]. In [29] the categorical cross-entropy loss is accommodated. A summary of the different studies reported in Sections II-A to II-C, is presented in Table III.

TABLE III

SUMMARY OF THE LITERATURE DISCUSSED ABOUT UNCERTAINTY ESTIMATION, AS PRESENTED IN SECTIONS II-A TO II-C, IN CHRONOLOGICAL ORDER. THE STUDY, ITS PRACTICAL APPLICATION AND THE METHOD FOR UNCERTAINTY DETERMINATION (MC DROPOUT, ENSEMBLE AND/OR HETEROSCEDASTIC ALEATORIC UNCERTAINTY) ARE REPORTED.

Study	**Application**	**Uncertainty approach**		
		MC dropout	Ensembles	Heteroscedastic aleatoric uncertainty
Lakshminarayanan,2016 [30]	computer vision		✓	
Kendal,2017 [20]	computer vision	✓		✓
Bragman,2018 [23]	prostate RT	✓		✓
DeVries,2018 [11]	skin lesion segmentation	✓		✓
Jungo,2018 [24]	brain tumour segmentation	✓		
Hu,2019 [25]	lung & prostate segmentation	✓		✓
Do,2020 [26]	myocardial arterial spin labeling (ASL) segmentation	✓		
Meijenik,2020 [27]	medical tabular data	✓	✓	
Jungo,2020 [28]	brain tumour segmentation	✓	✓	✓
JooLee,2020 [12]	computer vision & endometrium segmentation		✓	
Ståhl,2020 [29]	computer vision	✓	✓	✓

D. Uncertainty as a quality measure

Uncertainty gives a valuable insight on the model's confidence. As presented in the previous subsections, there are diverse ways to perceive uncertainty estimates on DL-based architectures, however, the essential point is discovering how to use this knowledge properly. In clinical settings and most specifically in RT, the information on uncertainty would ideally indicate faulty predictions. Such predictions should be further inspected and corrected by the experts, as there is no room for inaccuracies.

Notwithstanding, only few studies have focused on the assessment of uncertainty information. In particular, most of the papers combine the information about the segmentation performance with the knowledge derived from the uncertainty estimation in one single value. Dice similarity coefficient (DSC) comprises the most prevailing performance metric that has been analyzed, as seen in [24], [28] and [7]. It can be compared to the doubt score (dbt) based on user-defined threshold values, as in [24], or to the Hausdorff distance (HD), as mentioned in [7], in terms of correlation with uncertainty. Moreover, the area under the ROC curve (AUROC) is employed in [28] and [11] and the area under the precision-recall curve (AUPRC) is also mentioned in the latter work. Both these metrics, AUROC and AUPRC, interrogate whether uncertainty can detect segmentation failure. A summary of the aforementioned papers regarding how uncertainty is evaluated, can be found in Table IV.

III. METHODOLOGY

In this work, we implemented a deep learning method to segment the prostate on MRI, employing four approaches to capture uncertainty. We explored whether uncertainty can give profitable insights into the performance of the model. Finally, we investigated how our approach operates when out-of-distribution data is used.

TABLE IV

SUMMARY OF EVALUATION METRICS FOR UNCERTAINTY ENCOUNTERED IN LITERATURE, IN CHRONOLOGICAL ORDER.

Study	**Uncertainty evaluation metrics**
Jungo,2018 [24]	dbt, DSC
DeVries,2018 [11]	AUROC, AUPRC
Pan,2019 [7]	DSC, HD
Jungo,2020 [28]	DSC, AUROC

A. Data collection and preprocessing

The dataset used for this project was the publicly available prostate data collection of the `Medical Segmentation Decathlon`[1]. This collection consists of multi-parametric MR images of 48 patients, 32 for training and 16 for testing, with manual delineations of the prostate gland and the prostate peripheral zone available only in the training set. In this study, due to the lack of labeled images for test set provided, we splitted the Decathlon training set into training (14 patients), validation (9 patients) and test (9 patients) subsets. For each patient, a combination of T2-weighted and apparent diffusion coefficient (ADC) from diffusion-weighted scans was compounded with resolution of 0.6×0.6×4mm and 2×2×4mm correspondingly [31].

Image intensities from both T2-weighted scans and ADC scans were initially normalized according to the 95/5 percentiles of each channel. However, this was taken into consideration during the optimization of the network. Moreover, resampling to unit voxel size (1×1×1 mm^3) was employed, using bilinear and nearest neighbor interpolation for the images and labels, respectively [32]. Cropping or padding for both T2 and ADC scans was applied to acquire volumes of 192×192×64 voxels. For the sake of simplicity, we rejected the label of the peripheral zone and preserved only the label for

[1] http://medicaldecathlon.com

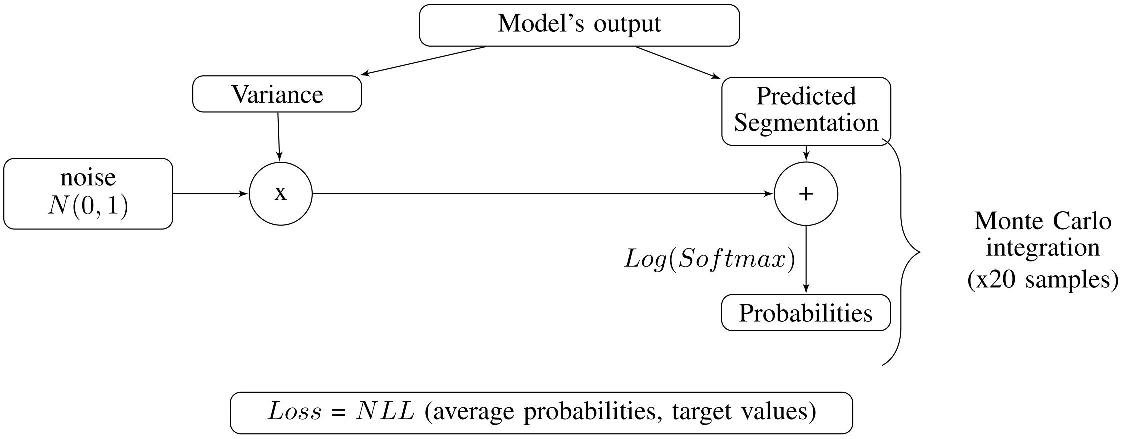
Model's output
Variance
Predicted Segmentation
noise
$N(0,1)$
x
+
$Log(Softmax)$
Probabilities
Monte Carlo integration (x20 samples)
$Loss$ = NLL (average probabilities, target values)

Fig. 2. Representation of the heteroscedastic loss function computation. NLL: the negative log-likelihood loss.

the prostate gland. Because of space constraints, an example pair of images, labels for two different patients from the data collection used, can be seen in Supplementary V.

B. Model architecture and training

For prostate segmentation, a V-net architecture [13] was adopted. V-net uses an encoder-decoder architecture with residual connections while it is composed of 4 convolutional layers for the downsampling and of 4 for the upsampling path. It processes data by performing volumetric convolutions for feature extraction (using kernels of 5×5×5 voxels) and for image resolution reduction (applying kernels of 2×2×2 voxels, with a stride of 2). At the end of the decoder, a softmax activation function is encompassed to produce class probabilities with two outputs: one channel for the foreground and one for the background. The total number of the model's parameters was 45,609,944. We trained the network with dropout $p = 50\%$ and used the ADAM oprimizer with a learning rate of $1e{-}1$. In order to capture the aleatoric uncertainty, we added another channel that computes the variance of the foreground. This channel was learnt during training via the calculation of the loss. Originally, the loss function used for prostate segmentation with the V-net was the Dice loss [13]. However, this could not apprehend the uncertainty information inherent in the input (reflecting the aleatoric component). Therefore, our objective function was adapted to accommodate for that. The modified loss was inspired by Kendal and Gal [20]: samples were drawn from a Gaussian distribution $N(0, 1)$, which corrupted the channel of the variance, and the result was added to the logits. The probabilitites of the classes were computed through the application of the log softmax on the output of this procedure. This process was repeated for $T = 20$ times, finally averaging the obtained probabilities. At the end, the loss for each epoch was calculated as the negative log-likelihood loss (NLL) between the predicted probabilities (p) and the target labels (y) for the C=2 classes:

$$Loss = -\sum_{c=1}^{C=2} y_c(\log \frac{1}{T} \sum_{t=1}^{T=20} \log p_{t,c}) \tag{5}$$

The process of the composition of the loss function is illustrated in Fig. 2.

Data augmentation was used during training applying affine deformations (shift, scale and rotations). The parameters were sampled from uniform distributions and applied to the coronal and sagittal planes only. In particular, the shift parameter was selected from the range [-50, 50] mm, the angle for the rotation was chosen from [-10, 10] degrees and the scaling factor was between [-10%, 10%] + 1. When the training was completed the value of 0.5 was applied to the output, in order to transform the predicted probability maps (for the background and the foreground) into binary ones, by thresholding the voxels [13]. Moreover, small predicted structures that were not part of the largest delineated connected component were removed. The architecture was implemented in `PyTorch` (v1.6.0) using the `MONAI` framework[2] (v0.3.0).

C. Uncertainty estimation

In our approach, we leveraged four different ways to estimate uncertainty:

- **Epistemic uncertainty** was approximated by the MC dropout technique. Since most of the related studies adopted a dropout rate of 0.5 during inference time, we also maintained this value, and we sampled from the network 50 times. Finally, the epistemic uncertainty map was formulated by estimating Shannon's entropy over the mean probabilities of the prostate channel, derived by the multiple MC samples.
- **Aleatoric uncertainty** was captured via the learnt loss function, as described in Section III-B. This uncertainty map was formulated as the average of the variance channel, after the application of the multiple MC samples during inference.
- **Total uncertainty** was considered as the voxel-wise sum of the aleatoric and the epistemic components.
- **1-Max(Softmax) uncertainty** was the inverse of the maximum softmax probability. This uncertainty estimate was calculated by merely finding the maximum of the

[2] https://monai.io/

softmax output across the class dimension and reversing the result. This measure should be the easiest way to capture uncertainty [11, 15], as it is the opposite of the confidence on the predictions.

For all the types mentioned, the calculations were performed per output voxel so that for each type, we procured an uncertainty map of the same size as the input image. We compared all four types for the evaluation of our method.

D. Evaluation metrics

The accuracy of the segmentations were quantitative calculated against the ground truth labels through dice similarity coefficient (DSC), accuracy, precision, 95% percentile hausdorff distance (HD$_{95}$) and average surface distance (ASD). All measures were calculated within a bounding box which enclosed the structure indicated by the labeled image. This bounding box was defined as the minimum box that includes the information from the labeled image through all transverse slices, increased by 5 mm for each side of the coronal and sagittal planes. This was done to ensure that the performance metrics were combined with the uncertainty proxies per transverse slice, so that the slices that generated high uncertainty were pointed out for manual correction. This is discussed in more detail in Section III-E.

Uncertainty assessment constitutes a more complex task, since there was no ground truth available. The evaluation was accomplished by computing the average value of the uncertainty map within the aforementioned bounding box. In addition to that, we used the doubt score (dbt) that was presented by Jungo *et al.* [24], which served as a proxy for uncertainty. This score combines the uncertainty map (h), an Euclidean distance map of every voxel to the outline of the segmentation (w) and a binary mask around the segmenation outline (k) for each voxel $i \in [1, N]$:

$$dbt = \sum_{i=1}^{N} k_i w_i h_i \tag{6}$$

When calculating the dbt, the information from all voxels inside the image volume is included. Hence, contrary to the mean value of the uncertainty map, dbt was computed over the entire transverse slice, rather than inside the bounding box. Under the observation that the equation 6 would generate higher dbt for larger delineated volumes, a slightly adjusted version of it was formulated and applied, normalizing this score over the number of voxels present in each slice:

$$dbt_{modified} = \frac{\sum_{i=1}^{N} k_i w_i h_i}{N_{voxels}} \tag{7}$$

In an attempt to analyze whether uncertainty or its surrogates and various performance metrics were correlated with one another, Spearman's rank correlation coefficient (r_s) was calculated. This coefficient reflects how well the relationship of two variables, a and b, can be described by a monotonic function, while assuming that these variables are not explicitly normally distributed [33]. The resulted value ranges from -1 to 1, with these two bounds indicating an exact negative or positive correlation of the compared quantities within this range, respectively. That is to say if r_s is negative, as the quantity a increases, b decreases, whereas if r_s is positive, b grows as a grows. No correlation is implied in case r_s equals 0.

E. Experiments

Three main experiments have been conducted in this work, namely network optimization, uncertainty assessment and analysis of the model's behaviour on out-of-distribution datasets.

Network optimization

TABLE V
POINT OF REFERENCE FOR DATA NORMALIZATION AND HYPER-PARAMETERS INITALIZATION FOR THE V-NET. THE PARAMETERS WHOSE INFLUENCE WAS EXPLORED DURING THIS EXPERIMENT ARE WRITTEN IN **BOLD**.

Data normalization	95/5 percentiles
Batch size	2
Optimizer	ADAM
Learning rate	$1e-1$
Learning rate scheduler	no
L2 regularization	no
Early stopping	no
Epochs number	500

In order to maximize the segmentation accuracy of the V-net, we investigated data normalization and hyper-parameter optimization in terms of batch size, learning rate scheduler and weight decay. We started by setting the parameters-to-be-explored to their point-of-reference state, and then we modified each of them while monitoring the produced mean DSC for the validation set. We improved our segmentation approach by employing in the final network the parameters which resulted in the highest mean DSC for the above-mentioned set. The baseline for data normalization and parameters initialization, is reported in Table V.

A comparison between input normalization techniques was conducted: we explored the effect of the min–max normalization, applied on percentiles of the image intensities (95/5, 97.2/2.5 and 99/1), and those of the z-score normalization, employing the mean and standard deviation of each volume. The impact of batch size was also questioned, specifically for batches composed of 2, 4 and 6 images. The learning rate was initially set to $1e-1$, and we compared three different schedulers to update it during training, namely $CyclicLR$, $ReduceLROnPlateau$ and $StepLR$ (the names correspond to those found in the documentation of `PyTorch`). $CyclicLR$ makes use of a cyclical learning rate between two boundaries (from $1e-3$ to $1e-1$), where the period of the cycle was set to 2,000 iterations. $ReduceLROnPlateau$ reduces the initial rate by a factor of 0.2 if the DSC obtained for the validation set has stopped improving. $StepLR$ decreases this rate to the half of its previously applied value after every 50 epochs. Finally, L2 regularization of $1e-6$ for the ADAM optimizer was also tested. Early stopping was used to stop training in case no improvement appeared in the last 5 epochs to the mean DSC for the validation set. The architecture was optimized for 500 epochs for each combination of the normalization method and hyper-parameters, reaching a number of 36 combinations.

Uncertainty as a surrogate of performance

The four uncertainty types were analyzed by taking into account the average value of the uncertainty map in the bounding box per transverse slice, and the doubt score and the suggested modified doubt score for the entire image, again per transverse slice. Apart from these uncertainty assessment metrics, the accuracy of the segmentation was computed per slice in the bounding box in terms of accuracy, precision, DSC, HD_{95} and ASD. R_s was employed to examine the relationship between the performance metrics and the uncertainty estimates. Hence, the general pattern of this correlation for the entire dataset was observed based on the average correlation value across the patients. The highest absolute value of the r_s determined the dominant measure which would be used for quality assurance.

In order to obtain further in-depth information on the ambiguous delineations, the cases that need inspection and correction by the experts were highlighted. For this purpose, each transverse slice was represented by a point in a scatter plot, describing the relation between the two correlated quantities. To determine a criterion for evaluation, it was hypothesised that uncertain predictions would be underlined by low segmentation performance, i.e. reduced values for DSC and accuracy and/or elevated ones for HD_{95} and ASD. Once the correlation was determined, thresholds were used to pinpoint profoundly ambivalent slices for each patient. More specifically, these thresholds were pre-defined; the values used were the average values of the each measure. This hypothesis was investigated only for the metric resuted in the highest r_s with uncertainty.

Out-of-distribution robustness

TABLE VI
Transforms used to generate out-of-distribution data. When the factor is of the form $\{a : k : b\}$ it means the value is updated from a to b by a step of k.

Transform	Parameters / Factors
Gaussian noise	mean:0 & std:{0.1, 0.9}
Gaussian smoothing	{0.25 :0.25: 2.0}
Spatial scaling	{0.25 :0.25: 2.25}
Spatial shearing	{-1.0 :0.25: 1.25}
Spatial shifting	{-80 :20: 80} [mm]

For our last experiment, we explored how our approach on uncertainty estimation and assessment act on data that the model has not been trained on. Such data was generated by applying transforms on the test set and keeping track of the results (will be referred to as test-time data augmentation). To establish the robustness of the network, it was assumed that for these out-of-distribution inputs, the model would result in diminished segmentation performance and increased uncertainty [27, 34]. In an attempt to assess this hypothesis, the performance of the model on the perturbed images was explored relative to a network without test-time data augmentation. The evaluation was based on the accuracy, precision and DSC values, averaged over all patients in the test set. Additionally, the aforementioned performance was compared against the mean value of the total predicted uncertainty map. The total uncertainty map was selected amongst the four types since it supports uncertainty information emanating from the data and the model combined. In this way, the efficiency of the model to produce robust segmentations and the effectiveness of the performance metrics to describe the total uncertainty were explored. A list of the transforms used to produce out-of-distribution data are mentioned in Table VI.

TABLE VII
DSC (mean±std) achieved on the validation set while performing network optimization. The results reported refer to the influence of the specific parameter while keeping the rest fixed to their the baseline adaptations (see Table V).

Parameter	DSC (average±std)
Data normalization	
95/5 percentiles	0.75±0.01
97.5/2.5 percentiles	0.73±0.04
99/1 percentiles	0.75±0.04
z-score	0.69±0.06
Batch size	
2	0.74±0.05
4	0.76±0.04
6	0.76±0.02
L2 regularization	
$1e-6$	0.71±0.04
Learning rate scheduler	
StepLR	0.77±0.05
CyclicLR	0.76±0.03
ReduceLROnPlateau	0.71±0.08

IV. Results

Network optimization

Since the outcomes of the 36 different combinations are difficult to be reported, the results on the validation set in terms of DSC (mean±std) for each of the parameters involved separately during network optimization, are mentioned in Table VII. As can be seen, when the $StepLR$ scheduler was used, the average DSC of the method reached its maximum value of 77%, surmounting its two counterparts ($CyclicLR$ and $ReduceLROnPlateau$). Akin to that, a batch size of 6 increased the accuracy of the network by around 2% contrary to the smallest batch applied. However, the image batch of 6 was not selected eventually, since the actual findings from the combination of the various parameters in one model were slightly different. More specifically, the highest DSC accomplished on the validation set, reached a value of 0.77±0.02. The adaptations resulted in this value included min-max image normalization using the 95/5 percentiles for each volume, while selecting a batch size of 2 images. In addition, the $StepLR$ scheduler and L2 regularization of $1e-6$ were

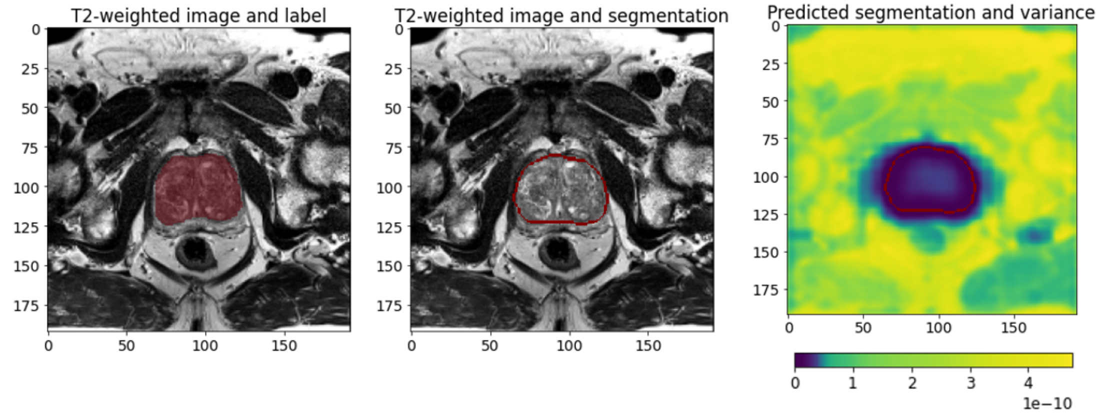
T2-weighted image and label
T2-weighted image and segmentation
Predicted segmentation and variance
0
25
50
75
100
125
150
175
0
50
100
150
0
1
2
3
4
1e−10

Fig. 3. Prediction of our method for patient Pt_1 (slice No21). From left to right: the T2-weighted scan and the labeled image (superimposed in red), the T2-weighted scan and the outline of the predicted segmentation (presented in red), the predicted variance map and the outline of the predicted prostate delineation (displayed by the red contour).

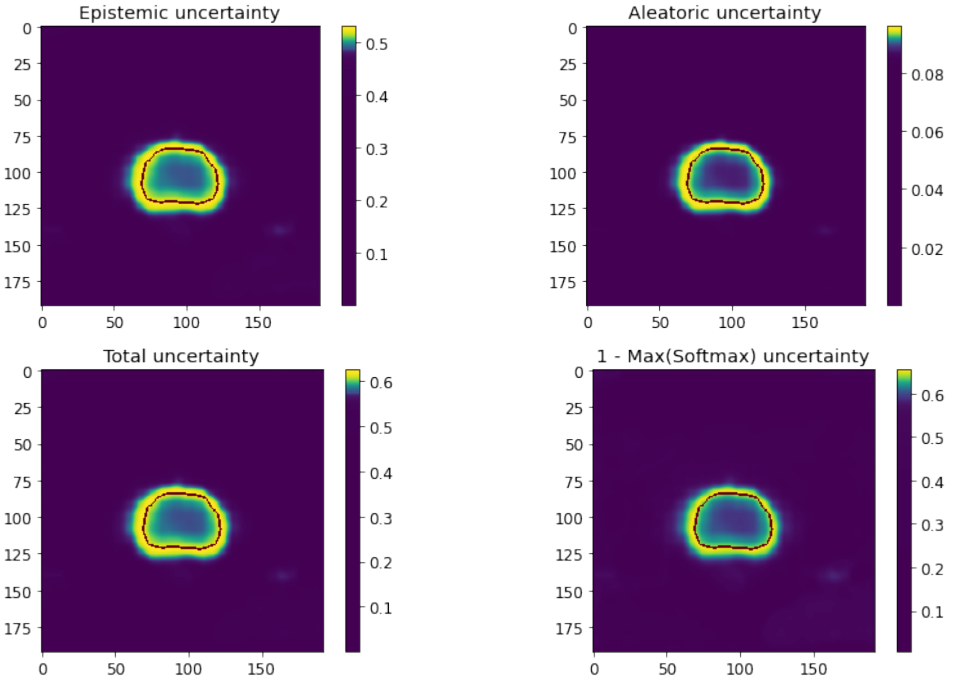

Fig. 4. Example of the predicted prostate segmentation on top of the four different uncertainty types used in this work for patient Pt_1 (slice No21). The predicted prostate structure's outline is presented by the red contour. The colorbars refer to the uncertainty values for each type.

also employed, while finally, early stopping was performed on the epoch number 344. The suggested approach made use of about 13 GB of GPU memory, whereas the training lasted approximately 13 hours. Please refer to Supplementary V to visualize the training curves. Eventually, the network achieved a performance of 0.78±0.06 DSC on the test set. Fig. 3 displays the predicted segmentation and variance as compared to the labeled image for a 2D slice of a single patient from the test set (Pt_1). As observed, the predicted variance presented higher values around the predicted segmentation and lower ones inside the delineated structure.

Uncertainty assessment

Epistemic, aleatoric, total and 1-max(softmax) uncertainty types were compared qualitatively against one another and against the prostate predicted delineation for patient Pt_1, as illustrated in Fig. 4. The visual inspection of these four maps revealed that all techniques operated in a comparable manner, producing a ring of high intensities at the border of the segmentations. However, it is apparent that aleatoric component was composed of lower intensity values compared to its three correspondences. This was also demonstrated in Fig. 5, where an overview of the performance metrics (accuracy,

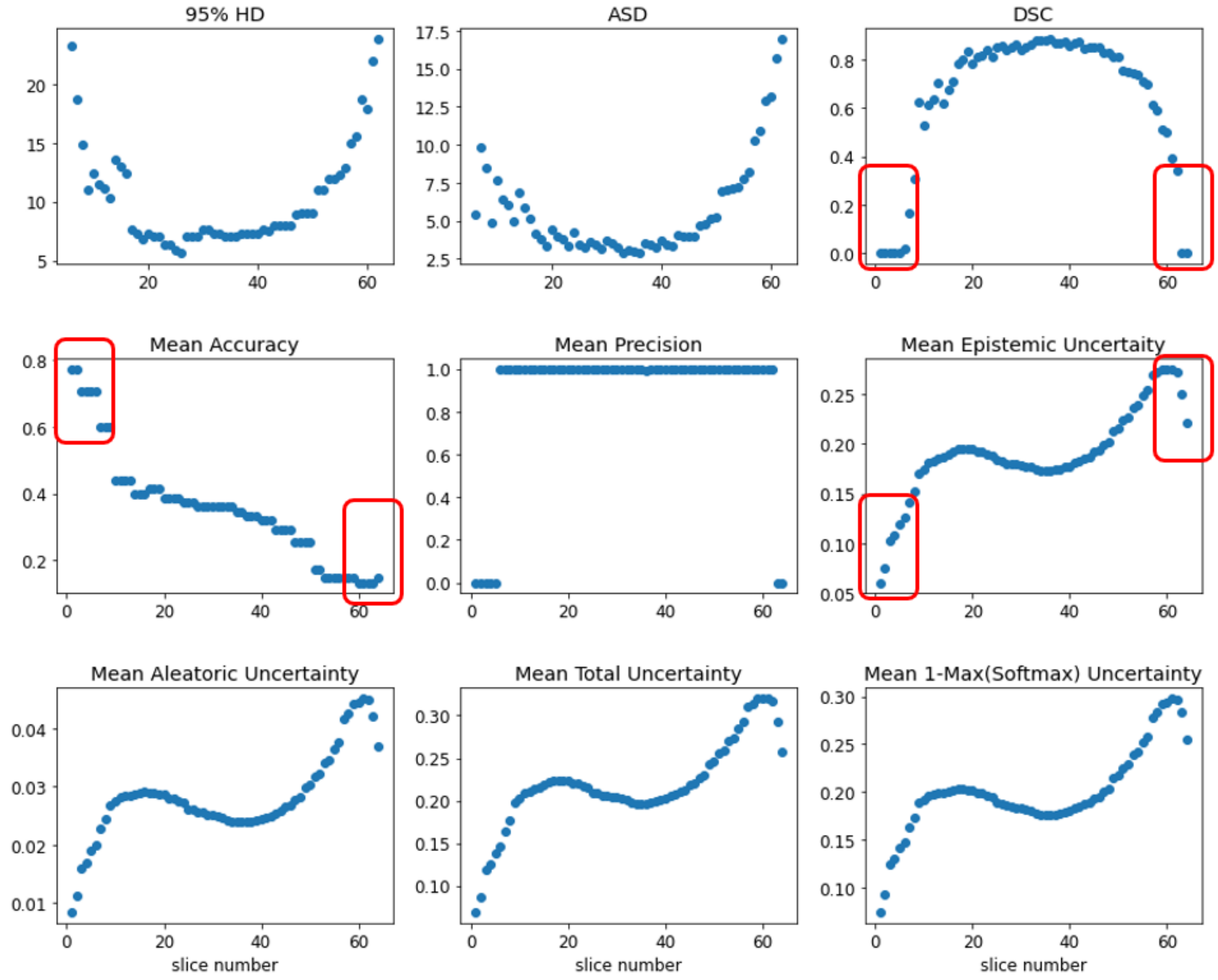

Fig. 5. Scatter plots of the different performance metrics and the four uncertainty types for patient Pt_1. The calculations are obtained within a bounding box per transverse slice. Slice number goes from No1 (closest to caudal direction) to No64 (closest to cranial direction). Accuracy decreases while uncertainty increases, but DSC is almost 0 for low as well as for high uncertain slices (indicated by the red boxes).

precision, DSC, HD_{95} and ASD) and the four uncertainty types is presented for the same patient (Pt_1). As shown, the various uncertainty estimates behaved similarly to one another, while reporting some fluctuations on their exact value range. It was also noticeable that uncertainty reached its maximum value towards the end of the slice range (around slice No60), which refers to the area nearest the cranial direction. From a closer inspection of the scatter plots, it appeared that as the accuracy of the network degrades, the uncertainty rises, suggesting that there might be a correlation between these two quantities. It is also worth noting that in regions near the caudal (around slice No5) and the cranial (close to slice No61) directions, although uncertainty values appeared to be low and high respectively, DSC values were almost zero for both cases. Therefore, DSC may not be an appropriate metric to describe the behavior of the uncretainty. Due to space constraints, more results can be viewed in Supplementary V.

A more quantitative evaluation of the uncertainty was achieved through Spearman's correlation coefficient r_s. The results of this correlational analysis are summarized in Fig. 6. As indicated, the overall pattern was approximately constant amongst the four uncertainty estimates for each metric, irrespective of the uncertainty proxy used. As opposed to the two doubt scores (dbt and $\text{dbt}_{modified}$), the uncertainty maps demonstrated a stronger correlation with the performance metrics, since the obtained average r_s abstained the value of 0 for all case studies. This graph clearly indicates that uncertainty revealed a more robust correlation with accuracy than with the other measures. Particularly for the correlation with accuracy, r_s resulted in an average value around -0.7 for all four uncertainty types, meaning that the correlation can be described as negative. That is to say, for increasing accuracy, decreasing uncertainty was expected. On the contrary, a strong positive correlation between uncertainty and precision was noticed since r_s was approximately 0.6. Even though the dbt achieved a correlation with accuracy of about -0.5, it was not strong enough to surmount the respective value of r_s between the uncertainty maps and the accuracy previously mentioned

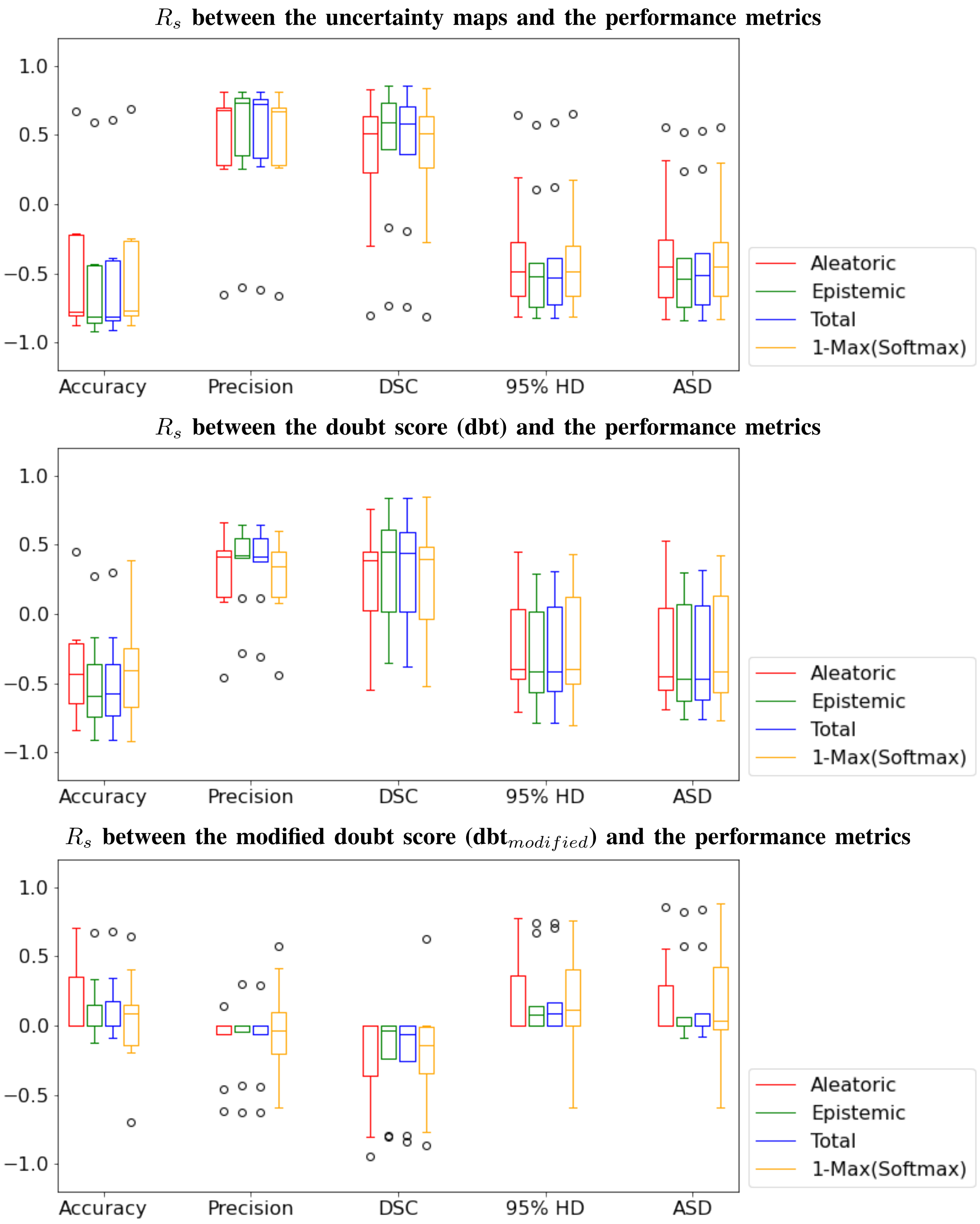

Fig. 6. Boxplots for the Spearman's correlation coefficient R_s between uncertainty proxies (average value of the uncertainty map, doubt score and modified doubt score) and the performance metrics (accuracy, precision, DSC, HD$_{95}$ and ASD) for all patients in the test set.

(-0.7). Contrary to expectations, no correlation between the dbt$_{modified}$ and the performance is generated, implied by the fact that r_s was almost 0 for all metrics explored, as seen in Fig. 6. Hence, it could conceivably be hypothesised that the mean value of the uncertainty map could serve as a proxy for the segmentation accuracy, regardless of the uncertainty type used.

Due to the highest correlation achieved by the accuracy and the average value of the uncertainty map, inspection of the ambiguous slices for each patient can be performed. In Fig. 7, four scatter plots are presented, each one of them describing the relationship between a specific uncertainty type and accuracy for patient Pt_1. Overall, the same slices were depicted by all four uncertainty types. Epistemic uncertainty attained to capture all three of them (No45, No46, No47), whereas total and 1-max(softmax) uncertainties detected two (No46, No47) and one (No47) of them, correspondingly. The labeled image, the predicted segmentation and the epistemic uncertainty obtained for these three slices, are presented in Fig. 8. As seen, the segmentation method underestimated the prostate gland for these ambiguous slices. Moreover, the model's uncertainty (i.e. the epistemic component) expands over a ring around the delineated structure, imposing that ambiguity was roughly present on the boundaries of the predicted segmentation.

Out-of-distribution data analysis

For the dataset without the test-time augmentation applied, the results concerning the accuracy, precision and DSC found to be approximately 89, 90 and 70%, for each one of the metrics respectively, while the average value for the uncertainty map was 0.23. The behaviour of our network on unseen data is assessed from Fig. 9. As observed, accuracy and precision do not experience large variations with increasing image perturbations, yet this is not the case for DSC. For instance, when gaussian noise was applied with 0 mean and a standard deviation increase by 70% (contrary to the unperturbed data), the DSC declined to 0%, suggesting that the noise was so dominant on the image that the network became incapable of identifying any structure at all. However, the segmentation accuracy presented a drop of about 14%, attained an average of 77%, whereas the precision decrease by almost 7.5%, finally reaching the value of 83%. The total uncertainty for this case reduced around 8% contrary to the non-augmented data, approaching the value of 0.02. An example input image for patient Pt_9 and the results obtained when employing gaussian noise with std=0.7, are presented in Fig. 10. When spatial shifting of 80 mm was applied to the images, the structure of interest approached the edges of the FOV. In this case, the DSC fell about 36% from its original value, whereas the accuracy and the precision declined by 9 and 7%, reaching 81 and 84% respectively. Interestingly for this case study, total uncertainty increased by almost 8%, attained the value of 0.25. The most striking feature derived by this experiment on out-of-distribution data was that DSC appeared to be in agreement with the total uncertainty, for 4 out of the 5 types of perturbations applied (except from the spatial shift, as shown in Fig. 9). For this incidents, DSC followed a similar decreasing pattern with uncertainty, as the network became unable to detect robust delineations of the prostate gland.

V. Discussion

In this work, we sought to determine four types of uncertainty on the segmentation of a DL network for prostate delineation. Based on our findings, uncertainty correlates better with accuracy than with the other metrics explored, and out-of-distribution data does not necessarily produce increased uncertainty estimates.

For low image contrast, the prostate delineation becomes challenging even for the clinicians (Fig. 13), and high uncertainty is detected within a broader ring around the structure's outline (Fig. 14). For good image contrast, the segmentation approach works satisfactorily (Fig. 3) and our framework detects uncertainty around the edges of the delineation (Fig. 4). In regions where the segmentation quality is poor, for instance on the borders of the predicted structure, uncertainty maps generate increased values for all four uncertainty types that were investigated (Fig. 4 and Fig. 14). On the one hand, this is extremely relevant especially for the aleatoric uncertainty, as reported in [11, 20]: this component accounts for the uncertainty stemming from the data quality, mostly highlighting the voxels belonging to the structure's outline [11, 18]. On the other hand, this is also well explained by the epistemic uncertainty, which exhibits a rise in regions which have been wrongly segmented [11] or which are comprised of challenging voxels [20]. These results are not surprising even regarding the 1-max(softmax) component, since this uncertainty type is related to the confidence on the predicted class [11, 15]. Thus, this component rises for miss-classified voxels [11, 15]. In addition, epistemic seems to dominate over its aleatoric correspondence. Because of its nature, aleatoric component is confronted by the clinicians as well: poor input image quality would hamper even the manual contouring, leading to variations on the final segmentations [18]. Therefore, data uncertainty is not encountered only in automatic segmentation methods, but is rather inevitable in clinic. Notwithstanding, the total uncertainty might be more useful for further research compared to the other types due to the fact that it combines the uncertainties that originate from the two main sources (epistemic and aleatoric). In reviewing the literature, there has been some work indicating that it might not be a strong and clear distinction between epistemic and aleatoric uncertainties [34]. Therefore what is necessary, useful and relevant for an implementation, is always dependent on the target of the application and the settings used.

Typically, a DL model is expected to produce lower uncertainty values for robust predictions [24]. The results on the Spearman's correlation coefficient (r_s) from Fig. 6, show that uncertainty and doubt score are inversely proportional to the segmentation accuracy (r_s=-0.5), HD$_{95}$ (r_s=-0.4) and ASD (r_s=-0.4) and proportional to the precision (r_s=0.4) and the DSC (r_s=0.5). Even though the r_s was not explored in the study of Jungo *et al.*, the authors demonstrated that increased DSC reflected high doubt scores ($dbt >$ 25,000 average across all patients), which is consistent with what we have obtained in this study ($dbt >$ 4,177 average across all patients). The most interesting finding is that the average value of the uncertainty map together with accuracy reveal the strongest correlation compared to the rest of the combinations (r_s=-0.7), by being numerically close to the upper bound -1. Due to the fact that a relation between the two aforementioned quantities is present, a threshold only for uncertainty might be sufficient to specify the uncertain predictions. Although, a second threshold on the accuracy values could further improve the initial assumptions, by limiting the number of the ambivalent slices. In this case, interference by the experts during the revision of the predictions would be reduced considerably and manual correction of the erroneous image slices would target only the most ambiguous incidents. Therefore, the overall process would

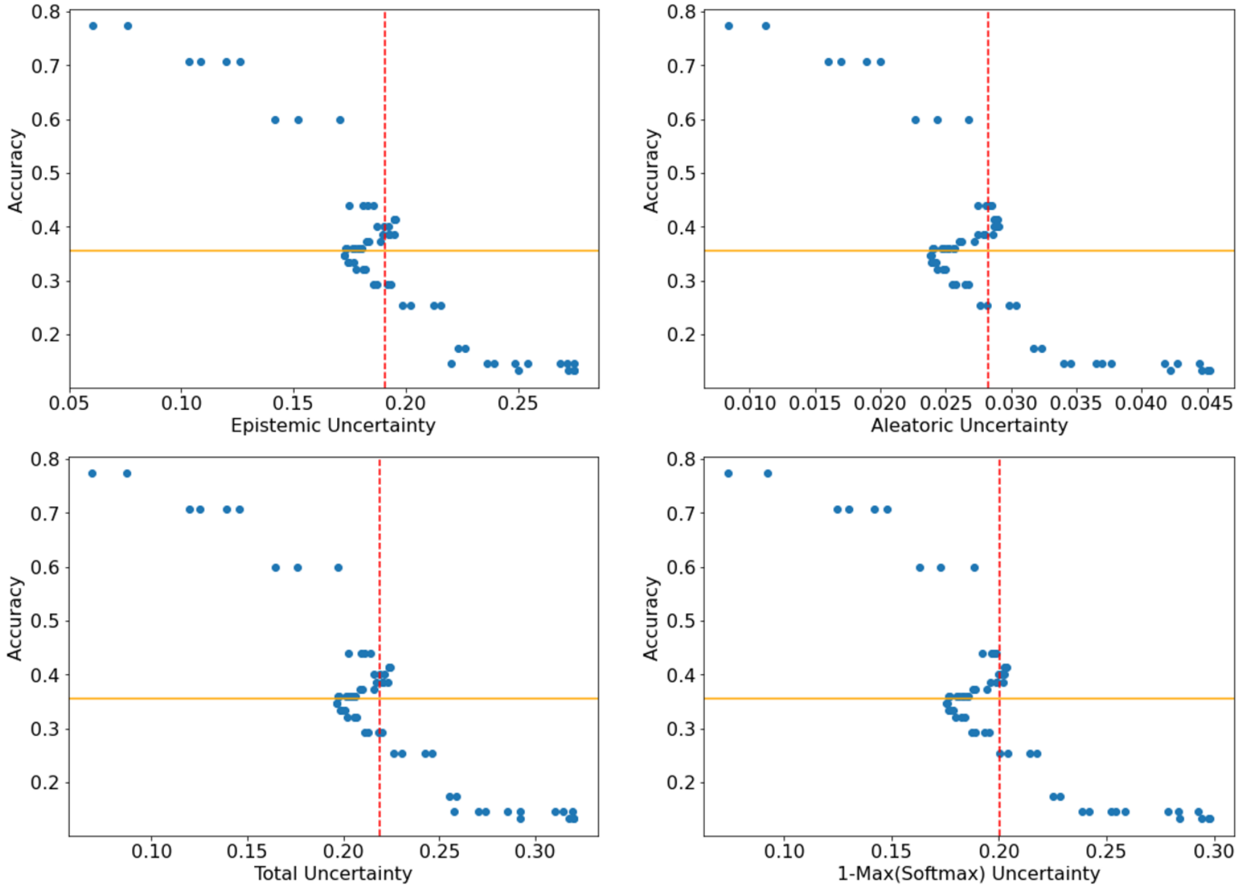

Fig. 7. Scatter plots of the mean accuracy and the various uncertainty types for patient Pt_1. The horizontal orange line represents the threshold value for the accuracy whereas the vertical red dashed line portrays the threshold for the uncertainty. The slices that need human inspection lie on the bottom right quadrant. For these plots, the average value of each quantity served as thresholds.

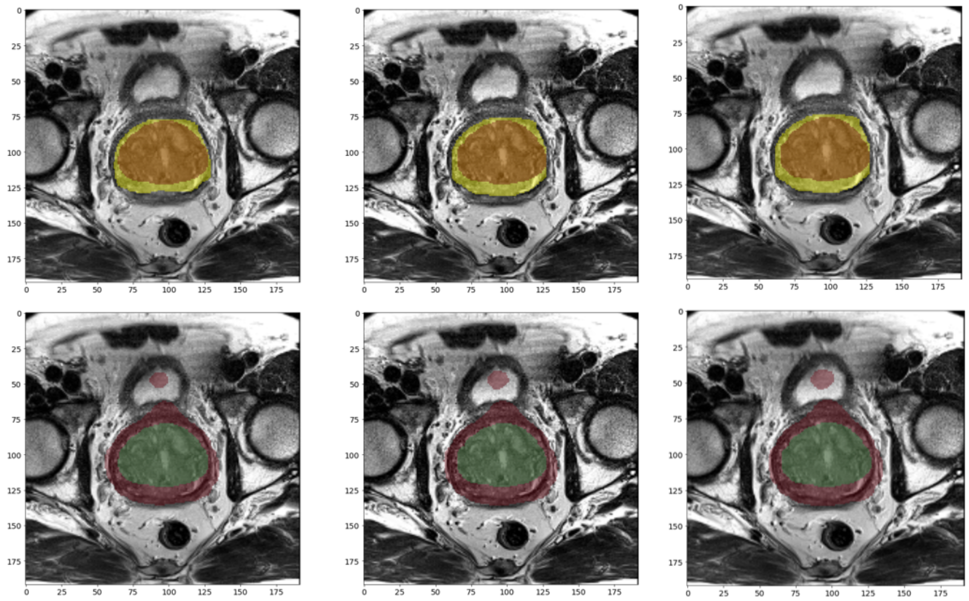

Fig. 8. The three tranverse slices of patient Pt_1 (from left to right: No45, No46, No47) which were highlighted for human revision by the epistemic uncertainty, according to the scatter plot in Fig. 7. Top row: ground truth prostate segmentation in yellow and predicted delineation in orange. Bottom row: prediction in green and epistemic uncertainty in red.

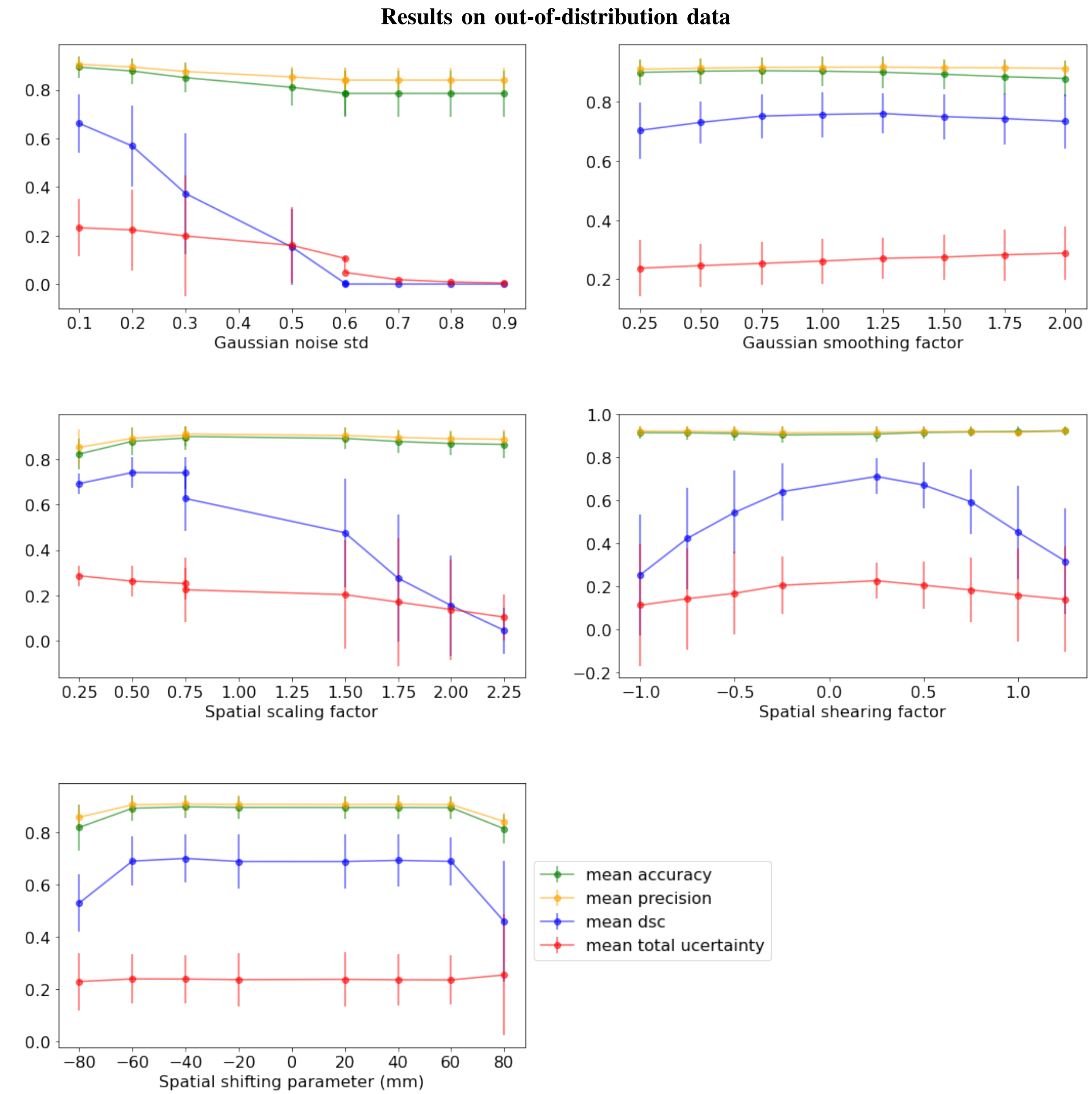

Fig. 9. Errorbars indicating how input image perturbations affect the performance of the network and the quantification of the total uncertainty.

be accelerated providing a feasible solution for an accurate automatic structure segmentation algorithm for the clinic.

Our approach was designed to determine the effect of uncertainty estimation on perturbed datasets. It was assumed that increased perturbations would generate out-of-distribution images and, therefore, the network would not be capable of producing correct delineations of the prostate. This in turn would provoke low accuracy values with raised predicted uncertainty estimates, as also hypothesized by relevant literature [15, 27, 30, 34, 35]. However, the findings do not support this assumption [15, 34]. In most cases, although the segmentation accuracy slightly degrades with increasing perturbations, total uncertainty does not experience a rise, as showed from Fig. 9 and 10. The latter seem to be consistent with the results in [27] and [34]. This outcome may be explained by the absence of a predicted structure. For example, with increasing gaussian noise DSC values approach 0 (Fig. 9), denoting that the prostate structure cannot be identified by the network. Therefore, when the test data differs a lot from the training data, the model might not be able to delineate the prostate at all. Yet, uncertainty may be generated approximately where the location of the prostrate structure should be, although its values would be considerably lower (Fig. 10) than those produced by the in-distribution data (Fig. 4). An unanticipated finding from this experiment is the decreasing pattern that DSC follows for the majority of the case studies, which is consistent with

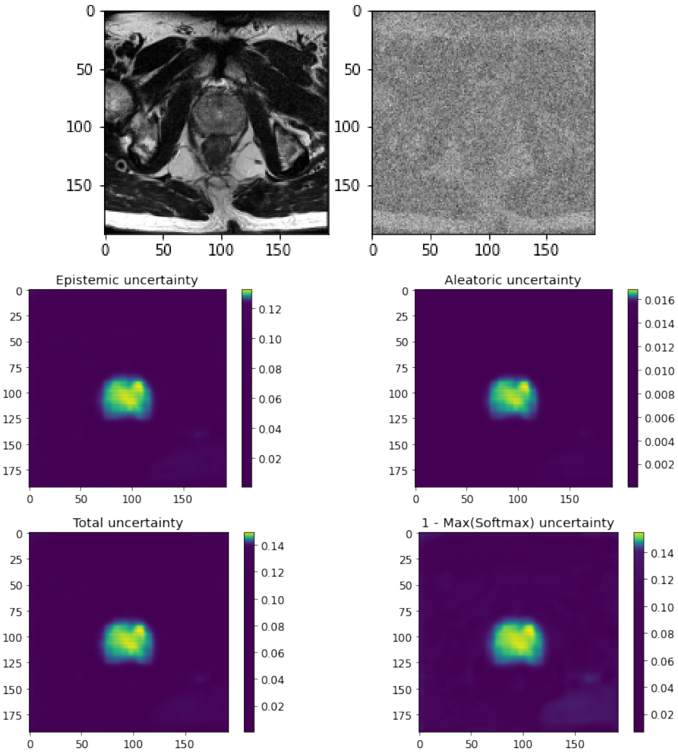

Fig. 10. Example of the network's output when using test-time augmentation with gaussian noise 0±0.7 (mean±std). On the top: T2-weighted image before (left) and after (right) the application of the gaussian noise for patient Pt_9. On the bottom: the four uncertainty maps produced for this patient.

the descending trend of total uncertainty. This observation may support the hypothesis that DSC could describe better than accuracy or precision the segmentation performance and the uncertainty estimation of the model on out-of-distribution data. Uncertainty on the corrupted datasets may be better described by an ensemble model, as reported by Ovadia *et al.* [34] or using a variational autoencoder as suggested by Meijerink *et al.* [27] and Ståhl *et al.* [29], who also argued that out-of-distribution input might not produced high epistemic uncertainty but rather low aleatoric uncertainty.

The suggested approach yields to limitations. For instance, uncertainty estimation is strongly affected by the way epistemic and aleatoric counterparts are quantified. In our approach epistemic uncertainty was gauged through 50 MC samples while applying 50% of dropout in inference time. The variance of the prediction is calculated in the loss function, averaging the 20 MC samples in every epoch. We noticed that the larger this number was, the more expensive the training is concerning GPU capacity and of course total training time. At the end, the aleatoric uncertainty is extracted as the average on the estimated variance channel, after the MC dropout application in inference time. These estimates could be determined differently, employing varied alternatives of the number of multiple samples for each uncertainty component or applying a different dropout rate. In addition, the evaluation of this study was performed in the transverse plane, however coronal and sagittal planes might also be enquired; future studies may investigate this aspect. In general, the dataset included few patients, so even though our results are promising, the lack of sufficient data reduces the validity of the outcomes. We can consider this study as a first feasibility investigation to bring uncertainty estimation into the clinic, but in the future a larger patient cohort should be employed to confirm the conclusions made. Furthermore, accurate error-prone slice identification is task-dependent since the thresholds used are pre-defined, hampering the generalization ability of the approach. Besides, calibration is considered essential when dealing with probability and uncertainty estimates in order to ensure that they are representatives of the true likelihood [34, 36, 37]. However, calibration was not included in this work, possibly impacting the results.

This study, provides a feasible framework for uncertainty estimation and assessment for RT. Our approach could facilitate identifying where the DL model fails and give a valuable insight into the data as well as the model used. This is of great importance when designing an automatic structure segmentation network for irradiation therapy. According to our outcomes, the predicted uncertainty can be used as a surrogate for accuracy in order to evaluate the performance of the segmentation approach. However, there is a clear distinction between in- and out-of-distribution input performance. The suggested method seems encouraging and further work is required to establish its validity, especially using a larger cohort of patients. Future studies might also focus on extending this approach to be used in clinical RT applications, ameliorating the workload. It might also be beneficial for further investigations to broaden the scope of this application to other body districts apart from prostate cancer treatment. This could eventually lead to a more generalized approach for uncertainty estimation in RT.

Acknowledgment

I would like to express my gratitude to my supervisors prof. dr. ir. Nico van den Berg and dr. Matteo Maspero for their valuable advice, guidance and support during the development of this research project. Their willingness to share their visions led to insightful discussions and was very much appreciated.

References

[1] G. Sharp, K. D. Fritscher, V. Pekar, M. Peroni, N. Shusharina, H. Veeraraghavan, and J. Yang, "Vision 20/20: perspectives on automated image segmentation for radiotherapy," *Medical physics*, vol. 41, no. 5, 2014.

[2] C. E. Cardenas, J. Yang, B. M. Anderson, L. E. Court, and K. B. Brock, "Advances in Auto-Segmentation," *Sem Radiat Oncol*, vol. 29, no. 3, pp. 185–197, jul 2019.

[3] K. B. Girum, G. Créhange, R. Hussain, P. M. Walker, and A. Lalande, "Deep generative model-driven multimodal prostate segmentation in radiotherapy," in *Workshop on Artificial Intelligence in Radiation Therapy*. Springer, 2019, pp. 119–127.

[4] K. Kamnitsas, E. Ferrante, S. Parisot, C. Ledig, A. V. Nori, A. Criminisi, D. Rueckert, and B. Glocker, "DeepMedic for Brain Tumor Segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Second International Workshop, BrainLes 2016, with the Challenges on BRATS, ISLES and mTOP 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, Revised Selected Papers*, vol. 10154. Springer, 2017, p. 138.

[5] M. H. Savenije, M. Maspero, G. G. Sikkes, J. R. van der Voort van Zyp, A. N. TJ Kotte, G. H. Bol, and C. A. T. van den Berg, "Clinical implementation of mri-based organs-at-risk auto-segmentation with convolutional networks for prostate radiotherapy," *Radiation Oncology*, vol. 15, pp. 1–12, 2020.

[6] S. Kazemifar, A. Balagopal, D. Nguyen, S. McGuire, R. Hannan, S. Jiang, and A. Owrangi, "Segmentation of the prostate and organs at risk in male pelvic ct images using deep learning," *Biomedical Physics & Engineering Express*, vol. 4, no. 5, p. 055003, 2018.

[7] H. Pan, Y. Feng, Q. Chen, C. Meyer, and X. Feng, "Prostate segmentation from 3d mri using a two-stage model and variable-input based uncertainty measure," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 468–471.

[8] A. Balagopal, D. Nguyen, H. Morgan, Y. Weng, M. Dohopolski, M.-H. Lin, A. S. Barkousaraie, Y. Gonzalez, A. Garant, N. Desai *et al.*, "A deep learning-based framework for segmenting invisible clinical target volumes with estimated uncertainties for post-operative prostate cancer radiotherapy," *arXiv preprint arXiv:2004.13294*, 2020.

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[10] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[11] T. DeVries and G. W. Taylor, "Leveraging uncertainty estimates for predicting segmentation quality," *arXiv preprint arXiv:1807.00502*, 2018.

[12] H. J. Lee, S. T. Kim, N. Navab, and Y. M. Ro, "Efficient ensemble model generation for uncertainty estimation with bayesian approximation in segmentation," *arXiv preprint arXiv:2005.10754*, 2020.

[13] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 565–571.

[14] M. van Herk, A. McWilliam, M. Dubec, C. Faivre-Finn, and A. Choudhury, "Magnetic resonance imaging–guided radiation therapy: A short strengths, weaknesses, opportunities, and threats analysis," *International Journal of Radiation Oncology• Biology• Physics*, vol. 101, no. 5, pp. 1057–1060, 2018.

[15] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," *arXiv preprint arXiv:1802.10501*, 2018.

[16] A. Balagopal, S. Kazemifar, D. Nguyen, M.-H. Lin, R. Hannan, A. Owrangi, and S. Jiang, "Fully automated organ segmentation in male pelvic CT images," *Phys Med Biol*, vol. 63, no. 24, p. 245015, 2018.

[17] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, 2016, pp. 1050–1059.

[18] M. Monteiro, L. L. Folgoc, D. C. de Castro, N. Pawlowski, B. Marques, K. Kamnitsas, M. van der Wilk, and B. Glocker, "Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty," *arXiv preprint arXiv:2006.06015*, 2020.

[19] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, A. Khosravi, U. R. Acharya, V. Makarenkov *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *arXiv preprint arXiv:2011.06225*, 2020.

[20] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Advances in neural information processing systems*, 2017, pp. 5574–5584.

[21] L. Hirschfeld, K. Swanson, K. Yang, R. Barzilay, and C. W. Coley, "Uncertainty quantification using neural networks for molecular property prediction," *Journal of Chemical Information and Modeling*, vol. 60, no. 8, pp. 3770–3780, 2020.

[22] L. V. Jospin, W. Buntine, F. Boussaid, H. Laga, and M. Bennamoun, "Hands-on bayesian neural networks–a tutorial for deep learning users," *arXiv preprint arXiv:2007.06823*, 2020.

[23] F. J. Bragman, R. Tanno, Z. Eaton-Rosen, W. Li, D. J. Hawkes, S. Ourselin, D. C. Alexander, J. R. McClelland, and M. J. Cardoso, "Uncertainty in multitask learning: joint representations for probabilistic mr-only radiotherapy planning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 3–11.

[24] A. Jungo, R. Meier, E. Ermis, E. Herrmann, and M. Reyes, "Uncertainty-driven sanity check: Application to postoperative brain tumor cavity segmentation," *arXiv preprint arXiv:1806.03106*, 2018.

[25] S. Hu, D. Worrall, S. Knegt, B. Veeling, H. Huisman, and M. Welling, "Supervised uncertainty quantification for segmentation with multiple annotations," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 137–145.

[26] H. P. Do, Y. Guo, A. J. Yoon, and K. S. Nayak, "Accuracy, uncertainty, and adaptability of automatic myocardial asl segmentation using deep cnn," *Magnetic resonance in medicine*, vol. 83, no. 5, pp. 1863–1874, 2020.

[27] L. Meijerink, G. Cinà, and M. Tonutti, "Uncertainty es-

timation for classification and risk prediction in medical settings," *arXiv preprint arXiv:2004.05824*, 2020.

[28] A. Jungo, F. Balsiger, and M. Reyes, "Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation," *Frontiers in Neuroscience*, vol. 14, p. 282, 2020.

[29] N. Ståhl, G. Falkman, A. Karlsson, and G. Mathiason, "Evaluation of uncertainty quantification in deep learning," in *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Springer, 2020, pp. 556–568.

[30] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *arXiv preprint arXiv:1612.01474*, 2016.

[31] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. Van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze *et al.*, "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," *arXiv preprint arXiv:1902.09063*, 2019.

[32] M. Perslev, E. B. Dam, A. Pai, and C. Igel, "One network to segment them all: A general, lightweight system for accurate 3d medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 30–38.

[33] H. Akoglu, "User's guide to correlation coefficients," *Turkish journal of emergency medicine*, vol. 18, no. 3, pp. 91–93, 2018.

[34] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," *arXiv preprint arXiv:1906.02530*, 2019.

[35] N. Tagasovska and D. Lopez-Paz, "Frequentist uncertainty estimates for deep learning," *arXiv preprint arXiv:1811.00908*, 2018.

[36] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1321–1330.

[37] M.-H. Laves, S. Ihler, K.-P. Kortmann, and T. Ortmaier, "Well-calibrated model uncertainty with temperature scaling for dropout variational inference," *arXiv preprint arXiv:1909.13550*, 2019.

Supplementary

Monte Carlo Dropout

MC dropout is equivalent to a probabilistic Gaussian process approximation [17], where a distribution that minimizes equation 3 is found [20]. Practically, a predefined number of forward passes is performed during inference and the results are to be averaged, a practice also known as model averaging [17]. Mathematically, the loss function for each point i for given pairs of input data and labels (x,y respectively), is formulated as:

$$L(\theta, d) = -\frac{1}{N}\sum_{i=1}^{N}(y_i|output(x_i)) + \frac{1-d}{2N}||\theta^2|| \tag{8}$$

where N is the number of the voxels, d is the dropout probability that is applied, θ represents the parameters of the optimized distribution q and $\log p(y_i|output(x_i))$ comprises the log-likelihood. According to Kendal and Gal [20], in classification setups a softmax likelihood is used and the results over the multiple samples T are averaged:

$$p(y|x, \omega) \approx \frac{1}{T}\sum_{t=1}^{T} Softmax(output(x_i)) \tag{9}$$

Then the epistemic uncertainty can be seen as the entropy of the predicted probability vector d over the c classes:

$$H(d) = -\sum_{c=1}^{C} d_c \log d_c \tag{10}$$

Medical Segmentation Decathlon dataset

The prostate data collection used for this project is composed of T2-weighted and apparent diffusion coefficient (ADC) from diffusion-weighted MRI, for each patient. An example of two patients from this data collection, can be visualized in Fig. 11.

Results

The training curves obtained by the model with the highest mean DSC on the validation set (77±0.02%) during network optimization, are presented in Figure 12.

The results for 2 patients of the test set (Pt_4 and Pt_5) are presented in Fig. 13 and 14. As shown, the network oversegments the prostate when its boundaries are getting more difficult to be spotted. Especially for patient Pt_5, where the prostate gland is hard to detect even by an expert, a broader ring of high uncertainty values is produced.

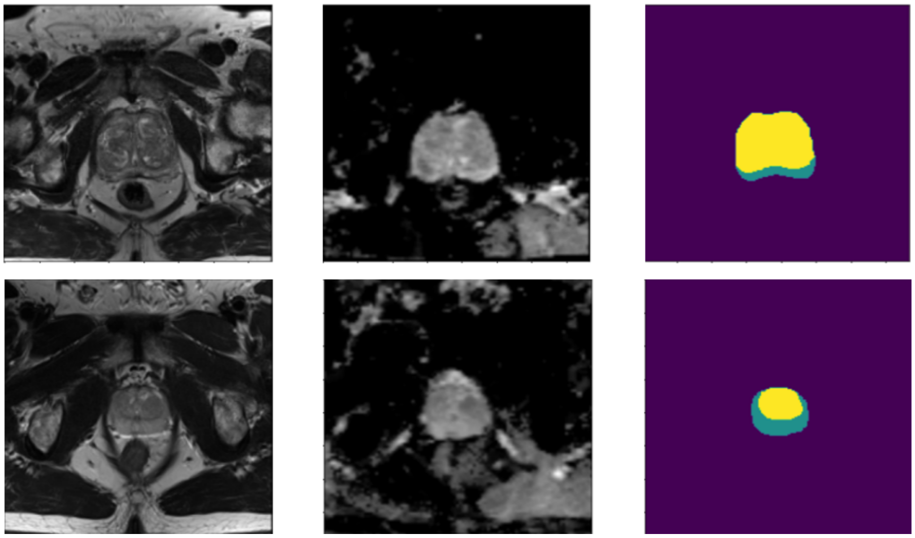

Fig. 11. Example images of two prostate cancer patients from the Medical Segmentation Decathlon dataset (Pt_1 and Pt_4). From left to right: T2 MR scan, ADC image, label image (the prostate gland is illustrated in yellow and the peripheral zone in light blue). For our application we considered only the label of the prostate, discarding the peripheral zone.

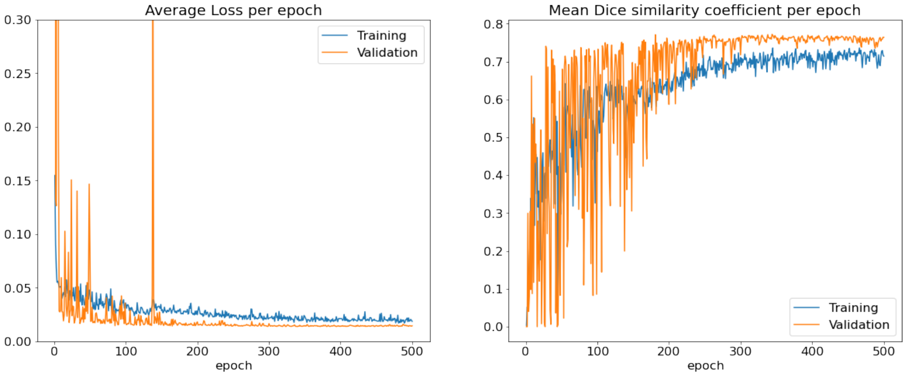

Fig. 12. Training loss and mean DSC curve from the selected model.

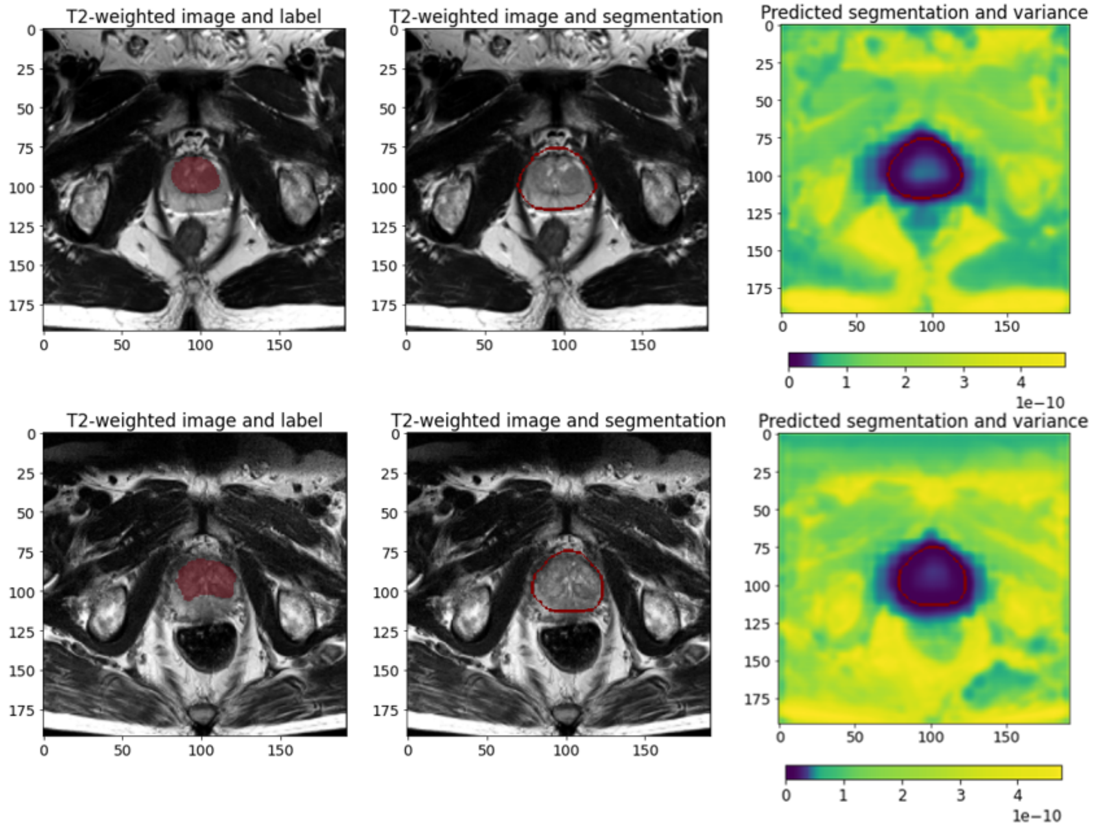
T2-weighted image and label
T2-weighted image and segmentation
Predicted segmentation and variance
0
25
50
75
100
125
150
175
0
50
100
150
0
1
2
3
4
1e−10
T2-weighted image and label
T2-weighted image and segmentation
Predicted segmentation and variance
1e−10

Fig. 13. Prediction of our method for 2 patients Pt_4 (top row), Pt_5 (bottom row). From left to right: the T2-weighted scan and the labeled image (superimposed in red), the T2-weighted scan and the outline of the predicted segmentation (presented in red), the predicted variance map and the outline of the predicted prostate delineation (displayed by the red contour).

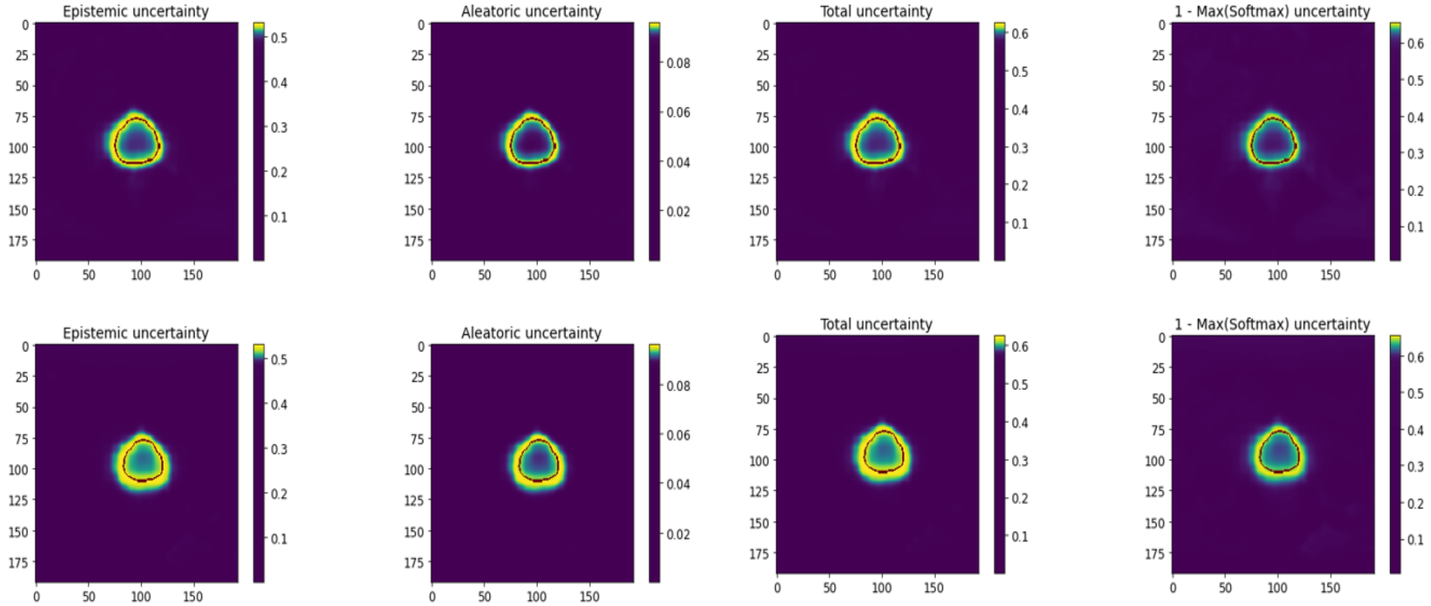

Fig. 14. Example of the predicted prostate segmentations on top of the four different uncertainty types used in this work for 2 patients Pt_4 (top row), Pt_5 (bottom row). The predicted prostate structure's outline is presented by the red contour. The colorbars refer to the uncertainty values for each type. From left to right: epistemic, aleatoric, total and 1-max(softmax) uncertainties.