

Uncertainty quantification in multi-class segmentation: Comparison between Bayesian and non-Bayesian approaches in a clinical perspective

Elisa Scalco¹  | Silvia Pozzi² | Giovanna Rizzo³  | Ettore Lanzarone² 

¹Institute of Biomedical Technologies (ITB), National Research Council (CNR), Segrate, Milan, Italy

²Department of Management, Information and Production Engineering, University of Bergamo, Bergamo, Italy

³Institute Of Intelligent Industrial Technologies and Systems (STIIMA), National Research Council (CNR), Milan, Italy

Correspondence

Elisa Scalco, Institute of Biomedical Technologies, National Research Council (ITB-CNR) Via Fratelli Cervi, 93, 20054 Segrate, MI, Italy.
Email: elisa.scalco@itb.cnr.it

Funding information

Ministero dell'Università e della Ricerca, European Union - Next generation EU, Grant/Award Number: 2022B23JT5

Abstract

Background: Automatic segmentation techniques based on Convolutional Neural Networks (CNNs) are widely adopted to automatically identify any structure of interest from a medical image, as they are not time consuming and not subject to high intra- and inter-operator variability. However, the adoption of these approaches in clinical practice is slowed down by some factors, such as the difficulty in providing an accurate quantification of their uncertainty.

Purpose: This work aims to evaluate the uncertainty quantification provided by two Bayesian and two non-Bayesian approaches for a multi-class segmentation problem, and to compare the risk propensity among these approaches, considering CT images of patients affected by renal cancer (RC).

Methods: Four uncertainty quantification approaches were implemented in this work, based on a benchmark CNN currently employed in medical image segmentation: two Bayesian CNNs with different regularizations (Dropout and DropConnect), named BDR and BDC, an ensemble method (Ens) and a test-time augmentation (TTA) method. They were compared in terms of segmentation accuracy, using the Dice score, uncertainty quantification, using the ratio of correct-certain pixels (RCC) and incorrect-uncertain pixels (RIU), and with respect to inter-observer variability in manual segmentation. They were trained with the *Kidney and Kidney Tumor Segmentation Challenge* launched in 2021 (Kits21), for which multi-class segmentations of kidney, RC, and cyst on 300 CT volumes are available. Moreover, they were tested considering this and other two public renal CT datasets.

Results: Accuracy results achieved large differences across the structures of interest for all approaches, with an average Dice score of 0.92, 0.58, and 0.21 for kidney, tumor, and cyst, respectively. In terms of uncertainties, TTA provided the highest uncertainty, followed by Ens and BDC, whereas BDR provided the lowest, and minimized the number of incorrect certain pixels worse than the other approaches. Again, large differences were seen across the three structures in terms of RCC and RIU. These metrics were associated with different risk propensity, as BDR was the most risk-taking approach, able to provide higher accuracy in its prediction, but failing to assign uncertainty on incorrect segmentation in every case. The other three approaches were more conservative, providing large uncertainty regions, with the drawback of giving alert also on correct areas. Finally, the analysis of the inter-observer segmentation variability showed a significant variation among the four approaches on the

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Author(s). *Medical Physics* published by Wiley Periodicals LLC on behalf of American Association of Physicists in Medicine.

external dataset, with BDR reporting the lowest agreement (Dice = 0.82), and TTA obtaining the highest score (Dice = 0.94).

Conclusions: Our outcomes highlight the importance of quantifying the segmentation uncertainty and that decision-makers can choose the approach most in line with the risk propensity degree required by the application and their policy.

KEYWORDS

Bayesian convolutional neural network, Monte Carlo dropout and dropconnect, multi-class segmentation, uncertainty quantification, risk propensity degree

1 | INTRODUCTION

Image segmentation based on Deep Learning (DL) approaches, and on Convolutional Neural Networks (CNNs) in particular, is currently the most widely adopted technique to automatically identify any structure of interest from a medical image, which has shown high accuracy in several applications. The standard U-Net¹ and its variants have been proposed in several cases, with different optimized approaches.^{2–5}

Unfortunately, the adoption of these approaches in clinical practice is slowed down by the difficult trustiness of their results, caused by poor transparency, by the occurrence of overconfident predictions and by the difficulty of providing an accurate quantification of their uncertainty. In addition, CNNs are sensitive to the uncertainty of the manual segmentation taken as a reference during the training phase, which can spill over into inaccurate and uncertain automatic segmentations obtained by the network.

From a clinical application perspective, standard CNNs are unable to provide the reader with a reliable quantification of confidence about the provided segmentation, as standard quality metrics represent only an average assessment of network performance. In this sense, the success of automatic segmentation approaches in clinical practice could be enhanced by tools capable of assessing segmentation uncertainty, that is, of identifying uncertain predictions that require manual human intervention versus those that are more certain, reliable, and accurate.

Different methods have been proposed to quantify the segmentation uncertainties related to data and inadequate knowledge.^{6–8}

A large part of the literature devoted to uncertainty quantification for CNNs exploits Bayesian CNNs (BCNNs) because of their recognized prediction accuracy and uncertainty quantification performance.⁹ They quantify uncertainties by providing predictions in terms of their posterior probability density, via Markov Chain Monte Carlo (MCMC), which samples exactly the posterior, or variational inference, which learns an approximation of the posterior.¹⁰ Within this class of approaches, Monte Carlo Dropout (MCD)¹¹ is largely adopted to effectively implement Bayesian inference into CNNs,⁸ which transforms a deterministic model into a BCNN

by simply using Dropout layers. It requires little additional knowledge or modeling effort, and is faster to train than other variational inference approaches.¹⁰ Recently, Monte Carlo DropConnect (MCDC) has been proposed for the estimation of the posterior, which generalizes MCD. In this case, the Bernoulli distribution is applied directly on each weight of the deep neural network, rather than on their output,¹² with the advantage of improving uncertainty quantification, yielding a precise estimation of model prediction confidence.

Alternatively, non-Bayesian approaches have also been proposed. Among them, ensemble methods (Ens) exploit the fusion of deterministic CNNs by combining predictions from different networks, while test-time augmentation (TTA) methods generate different predictions from augmentations of the actual input and combine them.^{13,14}

However, despite the importance of uncertainty quantification, the use of these methods in medical image segmentation has been addressed only recently.^{7,8} In this context, a few works also considered multi-class segmentation problems.¹⁵ Moreover, the reliability of adopted uncertainty metrics was analyzed based on the relevance of the uncertainty map to predict misclassification.¹⁵ Uncertainty quantification was also considered to effectively correct prediction errors and improve network performance through continuous adversarial learning and alternate training.¹⁶ Finally, the agreement between uncertainty estimated by MCD and assessed by radiation oncologists was evaluated in a recent work for tumor segmentation in lung cancer patients.¹⁷

In this work, we evaluated the MCD and MCDC configurations of BCNNs, an Ens and a TTA method for the multi-class segmentation problem, both in terms of accuracy and uncertainty quantification. We considered the case of Renal Cancer (RC) segmentation as a clinical problem in which the different structures have segmentation difficulties and are differently represented in the dataset, and in which different alternative manual segmentations could be available. This allowed us to evaluate the actual effectiveness of the different approaches in a specific clinical setting, with an eye toward practical applicability.

RC affects more than 400 000 individuals per year,¹⁸ being the eighth more frequent cancer in Europe

according to the European Cancer Information System. RC early diagnosis is important to improve the prognosis and find the best treatment option. Contrast-enhanced computed tomography (CT) is the preferred imaging technique for RC diagnosis, as the detected solid masses are likely to be RC.¹⁹ In addition, RC morphology can provide insight into tumor subtypes, as size, shape, and appearance correlate with aggressiveness and response to treatment.²⁰ Also, the presence of cysts in the kidneys can be detected from CT images and must be carefully identified²¹ because, although benign, they have the potential to develop into malignant masses. Therefore, identification of tumors and surrounding structures is an important task in the clinical decision process.

We applied our analysis to public datasets, to ensure the reproducibility and replicability of our outcomes. The dataset of the *Kidney and Kidney Tumor Segmentation Challenge 2021*, denoted by Kits21,⁴ which includes CT images of kidneys with RC and cysts, presents all of the above-mentioned features. This dataset, when adopted for the challenge, led to variable results across methods and structures, which meets our requirements. Kidney is a fairly easy organ to be segmented, with very high accuracy performance. In contrast, the accuracy of tumor segmentation strongly depends on the type of architecture and training, as can be seen from the challenge leaderboard. Even more challenging is the identification and segmentation of the cysts, which are present in only a subset of patients and are generally characterized by small volumes, often misclassified as a tumor. Moreover, two additional datasets were considered for external tests on completely new data.

The overall aim of the work is to assess whether the uncertainties quantified by a BCNN or a non-Bayesian approach can give effective alerts about the trustiness and correctness of the segmentations provided from CT images, in the presence of a multi-class problem with different accuracy levels. In particular, we want to compare the risk propensity among the approaches, which is a key aspect when they are applied in practice and also relevant for human operators. On the one hand, an approach could associate a high uncertainty to the segmentation, not trusting the predicted mask, even if it is correct for large parts of the area. On the other hand, an approach could make errors even where the uncertainty is not high and the prediction is trusted, still leading to errors.

2 | METHODS

2.1 | Architectures

We considered the *U-Net* proposed by Ronneberger et al.¹ as reference architecture, with an encoder part based on the VGG16 architecture and a decoder part

joined to the encoder by means of skip connections, as illustrated in Figure S1.

Uncertainty quantification was introduced in the *U-Net* by considering both Bayesian and non-Bayesian approaches. More specifically, the following alternatives were taken into account to obtain T segmentation replications from an image:

1. BCNN with MCD regularization, denoted in the following by BDR

BDR exploits neuron dropout as an approximation of a BCNN in which the weights follow Gaussian distributions.¹¹ It works with T deterministic CNNs ($t = 1, \dots, T$), with neurons dropped out, which represent as many Monte Carlo samplings from the space of all possible models associated with a BCNN. The MCD regularization is introduced to every last convolution operation in the convolutional layers, and it is possible to keep MCD active both in the training and in the test phase. MCD activated in the test phase allows to give the same input to the network in the T stochastic perturbations, each time leading to a small difference in the architecture, which could lead to different possible outcomes.

2. BCNN with MCDC regularization, denoted in the following by BDC

BDC differs from BDR in the dropped elements: connections in BDC rather than neurons in BDR.²² It introduces a dynamic sparsity on the network weights, leading to random connections dropping with a given probability at each replication, causing the network behavior to follow a Bernoulli probability distribution. The temporary exclusion of some connections in the training process impacts the model capability to generalize the information learned during training, reducing the excessive customization of the system on data and overfitting.²³ Also the MCDC regularization was introduced to every last convolution operation in the convolutional layers.

3. Ensemble of CNNs, denoted in the following by Ens

It leverages the fusion of T deterministic CNNs (i.e., without Dropout layers) by combining the predictions of different networks.¹³ The ensemble approach was implemented by bringing variety with random initialization and data shuffle, which has been reported to induced sufficient variety.⁷ The deterministic *U-Net* was trained T times, considering a different partition for training/validation sets each time.

4. Single CNN with test-time augmentation method, denoted in the following by TTA

It consists of generating several predictions associated to different augmentations of the actual input data during inference.¹⁴ In particular, one deterministic *U-Net* was trained and, in the prediction phase, T different transformations were applied to each test image, such as rotations, translations and flip in the

same range of the data augmentation used in the training phase.

2.2 | Network training

Data augmentation was adopted to increase variability in the available dataset towards the improvement of CNN robustness and overfitting reduction.²⁴ Specifically, the following rigid modifications were applied, considering physically plausible transformations: image rotation in the range $[0^\circ, 15^\circ]$, width shift in the range $[0\%, 15\%]$, height shift in the range $[0\%, 10\%]$, shear on x-axis in the range $[0^\circ, 10^\circ]$, zoom in the range $[0\%, 20\%]$.

Due to the presence of multiple labels/classes with significant area and volume differences, a multi-class weighted Dice (*mc.wDice*) loss was minimized. It accounts for unbalanced data by weighting the contribution of each class with the inverse of its volume²⁵:

$$mc.wDice = 1 - \frac{2 \sum_{j=1}^C w_j \sum_{i=1}^N p_{ij} g_{ij}}{\sum_{j=1}^C w_j \left(\sum_{i=1}^N p_{ij}^2 + \sum_{i=1}^N g_{ij}^2 \right)}, \quad (1)$$

where w_j is the weight of class j ($j = 1, \dots, C$), while p_{ij} and g_{ij} are the prediction and the Ground Truth (GT) of class j for voxel i , respectively.

The number of training epochs was equal to 100 for each deterministic *U-Net* of Ens and TTA and for BDR, while equal to 125 for BDC, to guarantee the convergence of each network. A batch size of 16, and Adam optimizer with learning rate of 10^{-4} were chosen for each network. Finally, as for BCNNs, the dropout rate was set to 30% and 50% for BDR and BDC, respectively, which provided the best performance.

For each model, we took the weights that provided the best Dice score on cyst segmentation, because of the late learning curve for this class compared to the others.

2.3 | Outcomes and uncertainty quantification

The number of replications T was set to 10 for every approach by identifying the value that provided stable uncertainties estimation (i.e., adding more than T samples does not affect uncertainty computation), as proposed by Roy et al.²⁶

The predicted mask for class c ($c = 1, \dots, C$) in voxel i ($i = 1, \dots, N$) is given in terms of the average value across the estimated probabilities \hat{p}_{cit} for the class in the voxel, obtained in the T samplings:

$$\bar{p}_{ci} = \frac{1}{T} \sum_{t=1}^T \hat{p}_{cit} \quad (2)$$

Binary masks of each class are then obtained by assigning to each pixel the class with the highest probability value.

The predictive uncertainty $U_{ic}^{(pred)}$ for class c in voxel i is determined as follows^{27,28}:

$$U_{ic}^{(pred)} = U_{ic}^{(aleat)} + U_{ic}^{(epist)} = \frac{1}{T} \sum_{t=1}^T \hat{p}_{cit} (1 - \hat{p}_{cit}) + \frac{1}{T} \sum_{t=1}^T (\hat{p}_{cit} - \bar{p}_{ci})^2 \quad (3)$$

The first addend $U_{ic}^{(aleat)}$, called *aleatoric uncertainty*, is the average variance of \hat{p}_{cit} over the T samplings. It derives from data distribution and is intrinsic; therefore, it is irreducible and unavoidable. The second addend $U_{ic}^{(epist)}$, called *epistemic uncertainty*, expresses the model variance. It is caused by the uncertainty in the network parameters. It can be improved by adjusting the architecture, the learning process or the training dataset.

2.4 | CT image datasets

Three datasets of renal CT images were considered:

1. For training and internal test, we used the Kits21 dataset⁴ (<https://kits-challenge.org/kits21>). It includes 300 patients who underwent nephrectomy for suspected malignant RC, and for whom a preoperative contrast-enhanced CT including the entire kidneys was available. Beside RC, some of the patients had one or more cysts. The number of slices per patient was variable. Each image was segmented by three independent operators for each relevant semantic class: kidney, tumor, and cyst. 80% of the patients, corresponding to 240 CT volumes, were randomly extracted and included in the learning process. 2D slices from these patients were randomly split with ratio 80% in the training set (16 151 images) and 20% in the validation set (4038 images). The remaining 20% of the patients, corresponding to 60 CT volumes, were kept aside for the testing phase. To avoid operator-dependent bias, a random selection of one of the manual segmentations was considered as GT for each patient.²⁹
2. For an independent test, the dataset of *Kidney and Kidney Tumor Segmentation Challenge 2023* (<https://kits-challenge.org/kits23>), denoted by Kits23 was used. It made available abdominal CT acquisitions as well as GT segmentation of kidneys, tumor, and cysts, delineated in consensus by different experts. They refer to the same multi-class segmentation problem, but are provided with only one manual segmentation. We considered data added in a second

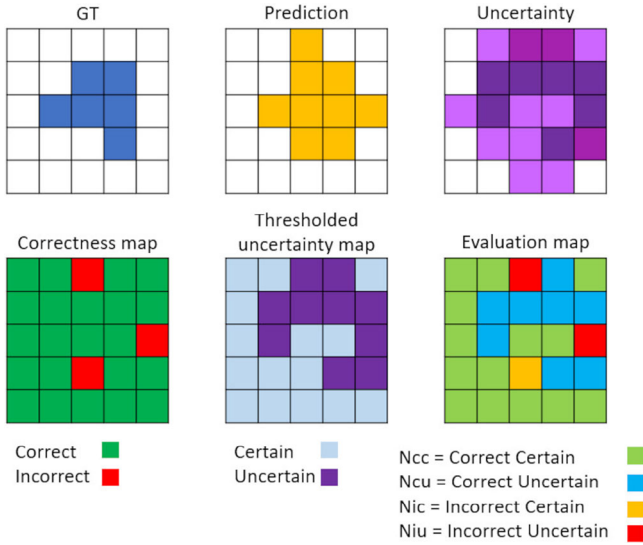


FIGURE 1 Steps for building the uncertainty confusion matrix. From GT and predicted segmentation, the correctness map is obtained by identifying correct pixels. Uncertainty map $U_i^{(pred)}$ is then thresholded using threshold I_T , to obtain a binary map. Finally, the evaluation map is built by considering the four possible combinations: correct certain (N_{CC}), correct uncertain (N_{CU}), incorrect certain (N_{IC}), incorrect uncertain (N_{IU}). GT, ground truth.

moment compared to Kits21, with a slightly different data annotation process. We have randomly selected 60 cases from the additional 200 cases available, in conformity with the test set used for Kits21.

- For an external evaluation with out-of-distribution data, abdominal CT images from the Qubiq challenge (<https://qubiq21.grand-challenge.org>) were used. The dataset consists of 24 cases, with a single 2D CT slice each, with segmentation of one kidney provided by three experts.

All images were resampled to a common size of 128×128 pixels, and CT intensity values were bounded between -79 and 304 , to enhance contrasts in the region of interest, following the procedure described by Heller et al.⁴ Then, all images were normalized by means of the average and standard deviation intensities.

2.5 | Evaluation metrics

The prediction accuracy on the Kits21 and Kits23 test sets was evaluated in terms of the Dice coefficient over the whole 3D volume of each structure and the Average Symmetric Distance (ASD) between predicted and reference contour.³⁰ Metrics have been evaluated for each structure (kidney, tumor, and cyst) and for their combination (total and mass), where mass represent the identification of tumor and cyst together, and total is the union of all three structures. This approach was suggested in the Kits21 challenge to understand the real

performances of the network for the singular class but also to define those cases where the mass is identified correctly but tumor and cyst has been swapped. In this specific situation, the singular structure's Dice indexes would be low, while on the contrary the Dice of the mass would be high.

Uncertainties were quantified using the predictive uncertainties $U_i^{(pred)}$ of each voxel i and class c . Then, two binary maps were created³¹: the *correctness map*, which is equal to 1 if the binary prediction is correct and 0 otherwise, and the *uncertainty map*, which is equal to 1 if $U_i^{(pred)} > I_T$ and 0 otherwise. The Uncertainty Threshold $I_T \in (0, 1)$ separates the uncertainty magnitude that is relevant for the application and determines the discharging of the prediction in the voxel. The combination of these two binary maps leads to the creation of four possible cases in a unique evaluation map (Figure 1), with N_{IU} (number of incorrect and uncertain predictions), N_{CU} (number of correct and uncertain predictions), N_{CC} (number of correct and certain predictions), and N_{IC} (number of incorrect and certain predictions).

These groups allow to derive three metrics upon the goodness of the uncertainty estimation:

- Correct - certain ratio (RCC): conditional probability of having correct pixels, given the certain ones:

$$RCC(I_T) = P_{I_T}(\text{correct}|\text{certain}) = \frac{N_{CC}}{N_{CC} + N_{IC}} \quad (4)$$

A high ratio of correct-certain pixels (RCC) reveals that the number of certain predictions but not correct are low.

- Incorrect - uncertain ratio (RIU): conditional probability of having uncertain pixels, given the incorrect ones:

$$RIU(I_T) = P_{I_T}(\text{uncertain}|\text{incorrect}) = \frac{N_{IU}}{N_{IU} + N_{IC}} \quad (5)$$

A high ratio of incorrect-uncertain pixels (RIU) reveals that incorrect predictions are also uncertain.

- Uncertainty Accuracy (UA): ratio of the previous desirable cases (numerators N_{CC} and N_{IU}) with respect to all cases:

$$UA(I_T) = \frac{N_{CC} + N_{IU}}{N_{CC} + N_{IU} + N_{CU} + N_{IC}} \quad (6)$$

They are deeply affected by the threshold I_T . The desired result maximizes all of the three metrics, favoring uncertainty when the prediction is incorrect and certainty when correct.^{12,28}

Finally, to assess whether the uncertainties provided by the different approaches were comparable to the inter-observer variability in manual delineation, we compared predictions and continuous GT labels (obtained as the average of either multiple annotations

TABLE 1 Prediction accuracy and uncertainty results, expressed as mean and standard deviation, on the Kits21 test set.

Metric	Structure	BDR		BDC		TTA		Ens		p-value
		mean	std	mean	std	mean	std	mean	std	
Dice	kidney	0.92	0.04	0.92	0.05	0.93	0.04	0.93	0.03	0.26
	tumor	0.59	0.26	0.48	0.32	0.59	0.29	0.65	0.26	0.01
	cyst	0.21	0.32	0.18	0.29	0.22	0.31	0.23	0.33	0.90
	total	0.94	0.06	0.94	0.05	0.94	0.04	0.95	0.05	0.55
	mass	0.63	0.25	0.53	0.31	0.62	0.28	0.69	0.25	0.03
ASD (mm)	kidney	1.51	0.95	1.70	1.22	1.54	0.89	1.33	0.87	0.26
	tumor	13.53	22.18	13.39	24.54	9.47	14.62	6.90	8.41	0.17
	cyst	27.40	40.13	20.05	23.89	15.11	16.35	20.44	27.56	0.49
	total	1.31	1.08	1.49	1.13	1.60	2.20	1.07	0.81	0.20
	mass	10.98	15.53	11.92	24.32	8.86	14.52	6.85	12.31	0.39
Uncertainty		$4.62 \cdot 10^{-4}$	$2.32 \cdot 10^{-4}$	$5.31 \cdot 10^{-4}$	$2.64 \cdot 10^{-4}$	$7.37 \cdot 10^{-4}$	$3.75 \cdot 10^{-4}$	$5.97 \cdot 10^{-4}$	$3.59 \cdot 10^{-4}$	< 0.0001
RCC (th = 0.001)		0.9998	0.0005	0.9999	0.0002	1.0000	0.0000	1.0000	0.0001	0.005
RIU (th = 0.001)		0.9523	0.0853	0.9830	0.0369	0.9921	0.0185	0.9923	0.0186	< 0.0001
UA (th = 0.001)		0.9616	0.0086	0.9626	0.0124	0.9260	0.0211	0.9540	0.0154	< 0.0001
RCC (th = 0.01)		0.9997	0.0007	0.9999	0.0003	0.9999	0.0001	0.9999	0.0001	0.004
RIU (th = 0.01)		0.9042	0.1122	0.9546	0.0678	0.9688	0.0396	0.9697	0.0449	< 0.0001
UA (th = 0.01)		0.9922	0.0035	0.9885	0.0050	0.9858	0.0076	0.9888	0.0061	< 0.0001

Abbreviations: ASD, average symmetric distance; BDC, Dropconnect method; BDR, Dropout method; Ens, ensemble method; RCC, ratio of correct-certain pixels; RIU, ratio of incorrect-uncertain pixels; TTA, test-time augmentation; UA, uncertainty accuracy.

by the experts or the T samples) by thresholding the continuous labels at predefined thresholds (i.e., 0.1, 0.2, ..., 0.8, 0.9). The volumetric overlap of the resulting binary volumes was calculated using the Dice score, and Dice scores for all thresholds were finally averaged. This procedure was suggested in the Qubiq challenge. This analysis was performed on the Kits21 test set, for the three segmented structures, and on the Qubiq dataset, by considering kidney segmentation only.

The presence of statistical differences among the networks was evaluated through the Anova test for accuracy metrics, followed by Tukey's post-hoc tests.

3 | RESULTS

3.1 | Accuracy evaluation

The two kidneys and at least one tumor were present in each patient in the Kits21 test set, whilst cyst structure was present in only 32 of them, with a single or multiple presence up to 17. Similarly, the Kits23 test set accounted for the presence of kidneys and tumor in each patient, and for the presence of single or multiple cysts (up to 20) in 35 of them.

Quantitative results about Dice score and ASD for the different structures of interest computed from the four approaches, along with uncertainty quantification metrics are reported in Table 1 for Kits21 data and in Table S1 for Kits23.

Accuracy was very different in the structures of interest, as expected. For Kits21 independent test, kidneys were very well segmented, with Dice scores of about 0.92 and distances between contours lower than 2 mm. Tumor was correctly identified in most cases, but the accuracy of segmentation was far from being ideal, with average Dice in the range 0.48–0.65, but with large dispersion (std of about 0.32). On the contrary, cysts represented a very difficult task, with a large number of false positive and false negative identifications, and poor accuracy (average Dice of 0.20).

Looking at the distribution of the Dice values for the cyst in Figure 2, it can be noted that the overlap between reference and predicted segmentation was almost null in more than 50% of patients. This indicates that the cyst was not even identified but, if detected, Dice was generally greater than 0.5. Instead, tumor segmentation presented a more uniform distribution of Dice values across the accuracy range, indicating very different performance depending on the specific patient. Cysts and tumors were misclassified in some cases; for this reason, accuracy evaluated on the total masses was higher (Dice in range 0.53–0.69).

The accuracy did not present significant differences among the approaches ($p \leq 0.05$ from Anova test), except for the tumor and mass segmentation. In this case Ens provided the highest accuracy and BDC the lowest, as perceivable also from Figure 2, where a lower percentage of low Dice values (Dice less than 0.6)

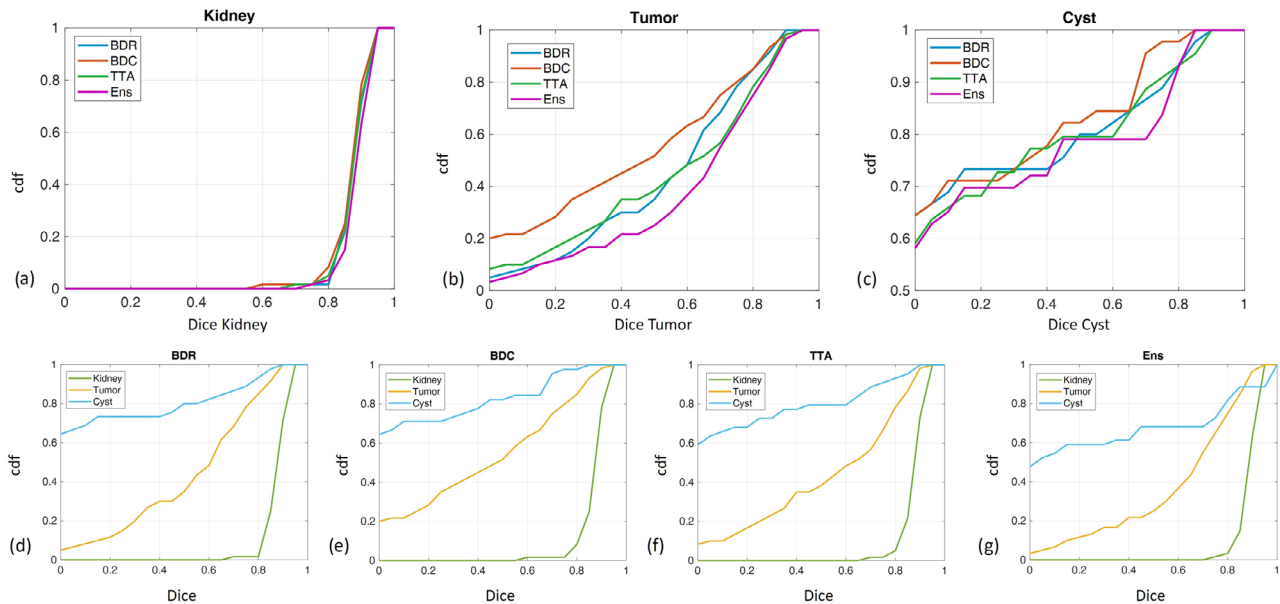


FIGURE 2 The cdf of Dice values over the Kits21 population, on the three different structures and using the four approaches: cdf comparison in kidney (a), tumor (b), and cyst (c) among the four approaches; cdf comparison for BDR (d), BDC (e), TTA (f), and Ens (g) among the three structures. cdf, cumulative density function; Ens, ensemble method; TTA, test-time augmentation.

is present for the former approach, compared to the others.

Similar trends were found for the Kits23 test evaluation, with accuracy values slightly reduced for kidneys and tumor (see Table S1 and Figure S2 for a detailed report). Furthermore, the four approaches did not present significant differences in accuracy for any of the structures.

3.2 | Uncertainty quantification

Looking at the uncertainty quantification reported in Table 1 and in Table S1, significant differences were present between the four approaches for almost every metric. Specifically, TTA uncertainty was the highest, followed by Ens and BDC, whereas BDR was the lowest.

The UA metrics as a function of threshold I_T are shown in Figure 3a for Kits21 test set and in Figure S3a for Kits23. From this graph, a reasonable choice for I_T was 0.01, which was able to provide high values for all the three metrics at the same time for every approach. The cumulative density functions (cdf) of the three metrics over the test populations for $I_T = 0.01$ are shown in Figure 3b–d and in Figure S3b–d. Specifically, looking at the cdf of RIU, BDR minimized the number of incorrect certain pixels worse than the other approaches; on the other hand, it presented also lower uncertainty for correct pixels, as showed by the better curve for UA. Ens and BDC presented a similar intermediate behavior, whilst TTA presented the lowest values for UA.

When the uncertainty metrics are separately considered for the three structures of interest, the differences among the approaches are similar to those with the overall evaluation, with RCC and RIU slightly lower for BDR, UA slightly lower for TTA, and intermediate values for Ens and BDC (Figure 4 and Figure S4). However, a different behavior among the three structures can be observed, with tumor and cyst presenting poor values for RIU. Focusing on the cyst, almost 50% of the patients presented RIU lower than 0.6, which increased to 60% of the patients with BDR. When assessed on the independent Kits23 dataset, the ability of minimizing the number of incorrect certain pixels was lowered, especially for both Bayesian approaches.

An illustrative example of good segmentation is shown in Figure 5, where the segmentation predicted by every network is comparable with GT. In fact, in this subject, the Dice values over the whole 3D structures were: 0.95 for kidneys in every approach, 0.86, 0.80, 0.84, and 0.93 for tumor in BDR, BDC, TTA, and Ens, respectively; 0.92, 0.77, 0.91, and 0.89 for cyst in BDR, BDC, TTA, and Ens, respectively. Looking at the uncertainty maps, it is possible to appreciate that high uncertainties were present only at the tumor boundary, whereas the prediction was certain inside kidneys and tumor.

An example of poor segmentation and misclassification of tumor and cyst is shown in Figure 6, where multiple cysts are present on the same axial slice. In fact, Dice values were: for kidney 0.86 in BDR and TTA, and 0.88 in BDC and Ens; for tumor 0.23, 0.30, 0.21, and 0.28, in BDR, BDC, TTA, and Ens, respectively;

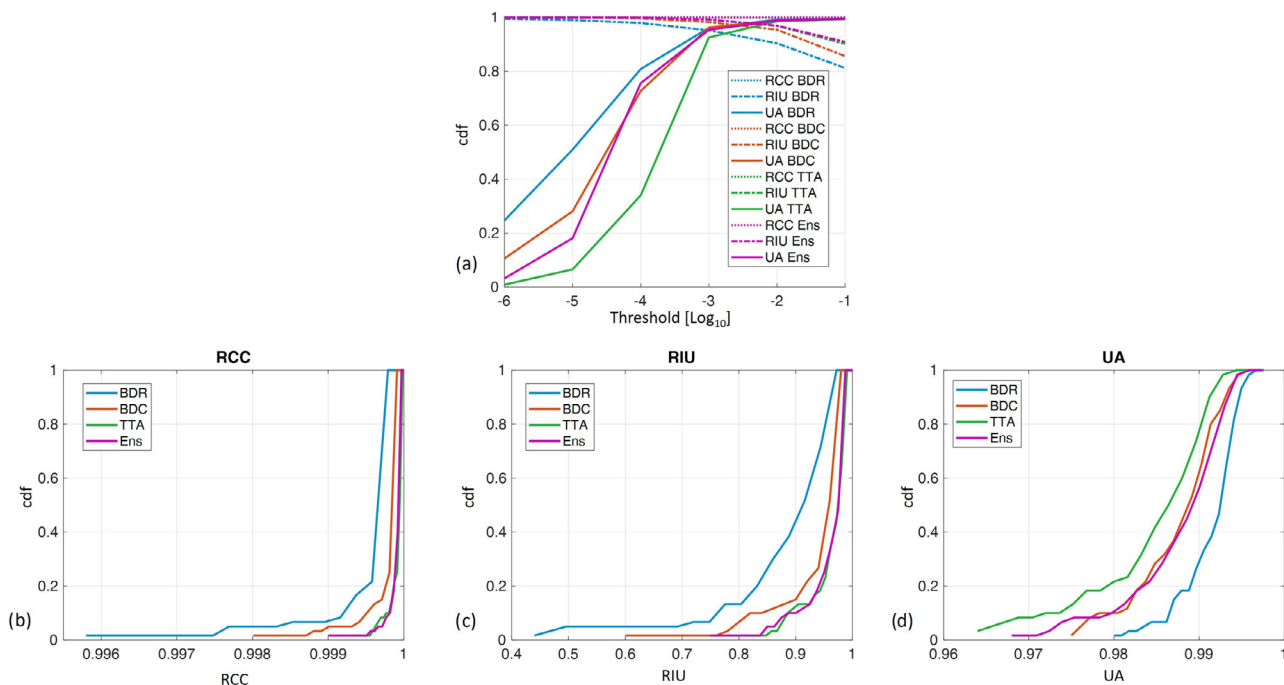


FIGURE 3 RCC, RIU, and UA metrics for Kits21 data as a function of l_T for the four approaches (a). cdf for the three metrics RCC (b), RIU (c), and UA (d) at $l_T = 0.01$. cdf, cumulative density function; RCC, ratio of correct-certain pixels; RIU, ratio of incorrect-uncertain pixels; UA, uncertainty accuracy.

for cyst 0.50, 0.65, 0.50, and 0.46, respectively. In this case, uncertainty maps can help the reader in identifying possible prediction errors, as highlighted by the large uncertainty area. In particular, in the BDR uncertainty map, the tumor and cyst in the right kidney have high uncertainties on the borders only, indicating reliable prediction. On the contrary, the cyst in the left kidney was well segmented but classified as tumor; in this case, the uncertainty map associated to the cyst can provide only some small uncertain areas around the structure. Looking at BDC, the tumor was not well segmented, but the associated uncertainty map provided low confidence within the whole area; the cyst on the right kidney was not identified, but the uncertainty was present on the whole structure, thus helping the reader. Finally, structures on the left kidney presented high uncertainty throughout the whole area, giving low confidence both on the false positive tumor and on the true positive cyst, thus asking to the reader a more in-depth assessment. The presence of higher uncertainty for correct voxels, especially on the kidney, is in agreement with the lower UA score of BDC compared to BDR. The other two non-Bayesian approaches are in line with the previous considerations. TTA provided the highest uncertainty throughout the whole left kidney, in agreement with low UA values, whereas Ens was similar to the BDC prediction.

Other illustrative examples reporting the different uncertainty maps can be found in Figures S5 and S6.

3.3 | Inter-observer variability

The ability of uncertainty quantification of simulating inter-observer contours variability is shown in Table 2. For Kits21 test set, this evaluation was in agreement with the accuracy evaluation: the main differences were related to the structures more than to the quantification methods, with the highest values of Dice score achieved for the kidney segmentation and the lowest for the cysts. On the contrary, the external evaluation on the Qubiq dataset, focused on kidney segmentation only, showed a significant variability among the four approaches, with BDR reporting the lowest agreement (Dice = 0.82), and BDC and TTA obtaining the highest scores (Dice = 0.93 and 0.94).

4 | DISCUSSIONS AND CONCLUSION

The progress for a standardized clinical application of CNNs in RC segmentation, as well as in other medical segmentation problems, mainly concerns the possibility of providing information on the reliability of model segmentation by means of an associated uncertainty map. In this way, the operator can decide whether to trust the segmentation or if the case requires further analyses. Different approaches are available for the quantification of the uncertainties related to image segmentation, spanning from BCNNs to Ensemble

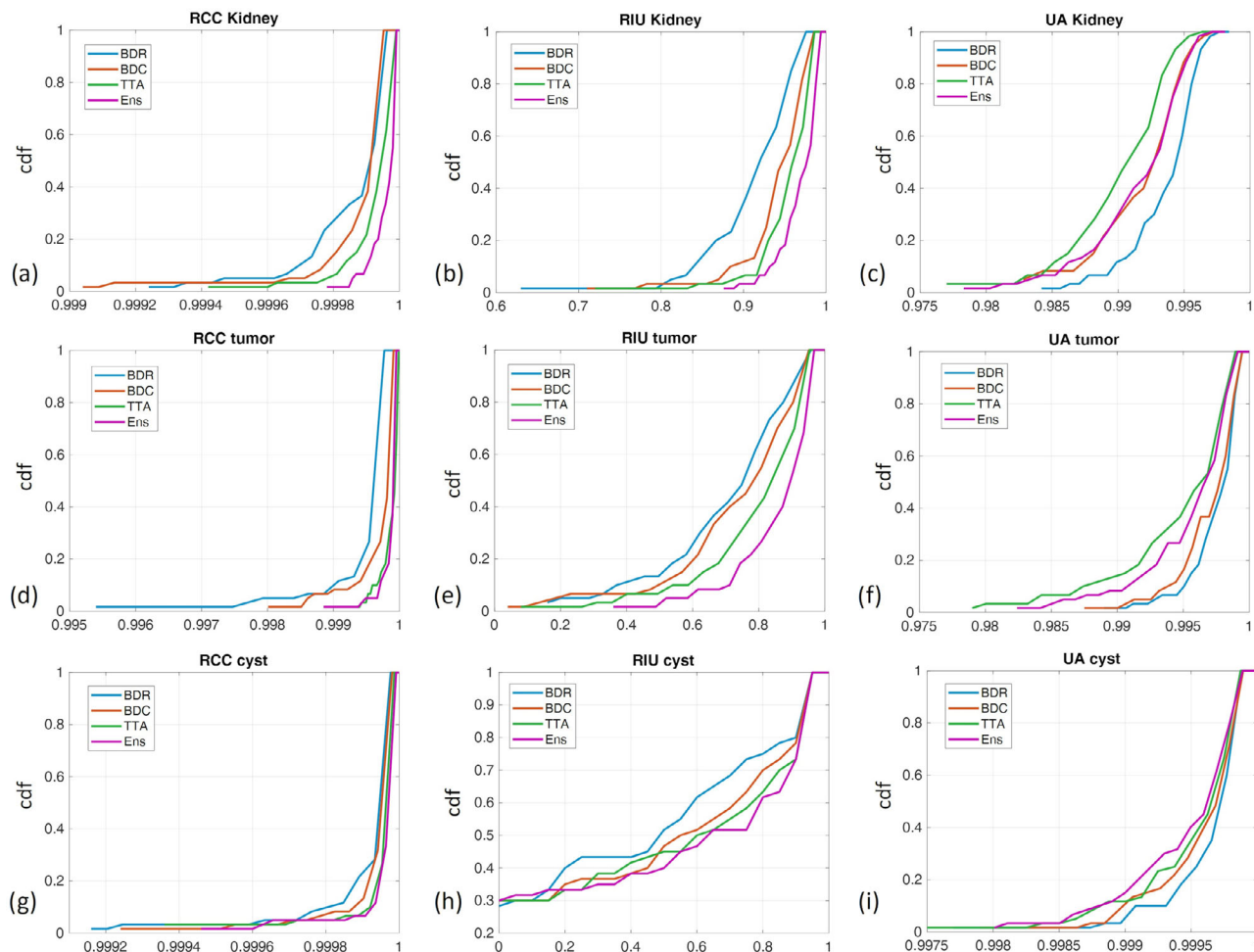


FIGURE 4 The cdf for the three metrics RCC, RIU, and UA and the three regions of interest, at $I_T = 0.01$. Metrics are computed on the Kits21 test set. cdf, cumulative density function; RCC, ratio of correct-certain pixels; RIU, ratio of incorrect-uncertain pixels; UA, uncertainty accuracy.

TABLE 2 Dice for inter-observer variability.

	Structure	BDR		BDC		TTA		Ens		<i>p</i> -value
		mean	std	mean	std	mean	std	mean	std	
Kits21 Test	kidney	0.92	0.04	0.92	0.05	0.92	0.04	0.90	0.04	0.045
	tumor	0.59	0.25	0.47	0.32	0.57	0.29	0.55	0.24	0.072
	cyst	0.23	0.33	0.17	0.27	0.21	0.30	0.19	0.26	0.848
Qubiq	kidney	0.82	0.01	0.93	0.01	0.94	0.01	0.88	0.01	< 0.0001

Abbreviations: BDC, Dropconnect method; BDR, Dropout method; Ens, ensemble method; TTA, test-time augmentation.

and TTA, with MCD as one of the most widely used because of its easy integration in traditional architectures. In this work, we compared two Bayesian and two non-Bayesian approaches to analyze the uncertainty quantification considering three public renal datasets.

Concerning the accuracy of the segmented masks, it was very different in the structures of interest. Kidney, being the easiest structure, was well segmented by all

approaches. Tumor was more or less correctly identified, even if segmentation accuracy was far from being ideal. Cysts presented poor accuracy with a large number of false positive and false negative identifications, as highlighted by the high percentage of subjects presenting Dice values equal to 0. The four approaches did not present significant differences in accuracy, except for tumor segmentation, where BDC achieved the lowest Dice score. These results allowed us to validate the

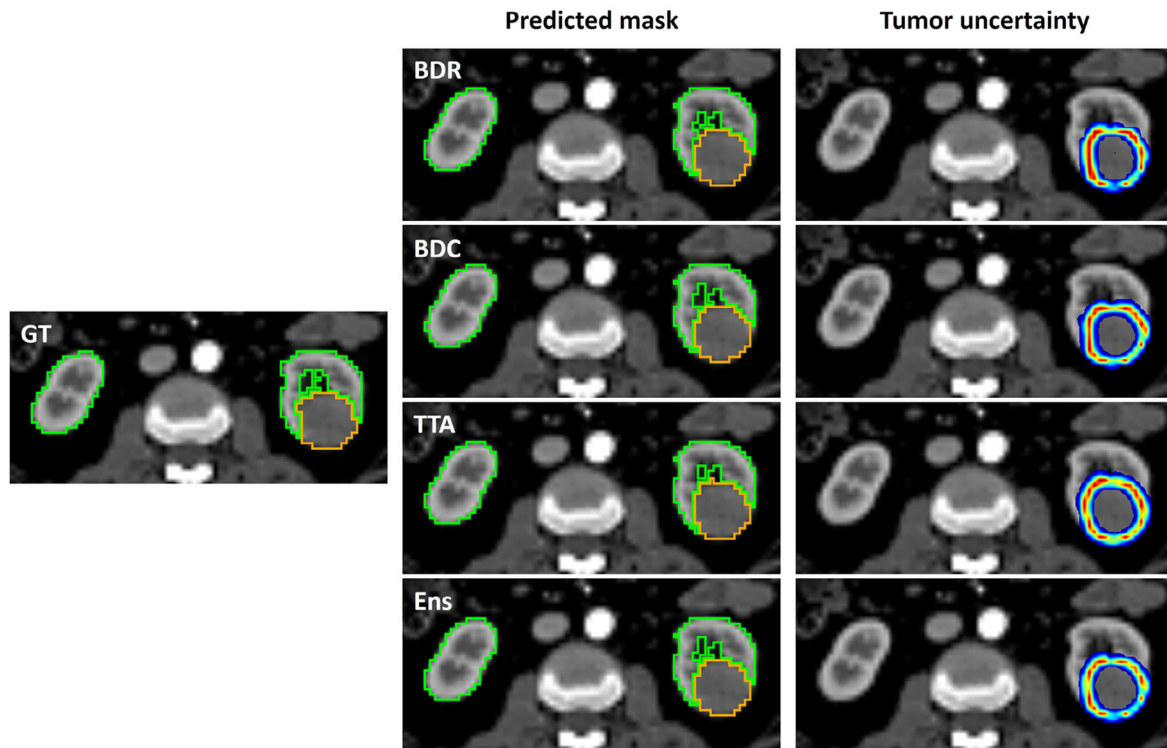


FIGURE 5 Example of segmentation results: kidneys (green) and tumor (orange) predicted by BDR, BDC, TTA, and Ens networks (second column), and the associated tumor uncertainty maps (third column); reference segmentation GT is also provided in the first column. Ens, ensemble method; GT, ground truth; TTA, test-time augmentation.

adopted networks, which performed similarly and poorly only in the difficult structure, confirming our intention to quantify the uncertainties in segmentation tasks of different difficulty.

As Kits21 dataset was used for a segmentation challenge, our accuracy performance can be immediately compared with those achieved by the participants to this challenge. We reached Dice scores lower than the best among those of the competitors, especially for tumor and mass segmentation, while at the same time our results are in line with those of the competitors positioned in the middle of the leaderboard and also comparable with benchmark results.⁴ This can be explained by our choice of evaluating the uncertainty maps with reference to a standard segmentation architecture, that is, the 2D VGG16 *U-Net*, which is still one of the benchmark networks for medical image segmentation.²⁴ A multi-step segmentation process, a cascade *U-Net*, a 3D approach or more advanced architectures, for example including transformers, could increase the accuracy, especially for tumor and cyst segmentation. However, since our aim was to evaluate the uncertainty quantification in regions with different accuracy and in a context that could be easily transferred in clinics, a further improvement in segmentation performance on this specific dataset was out of our scope.

The low performances in terms of accuracy and uncertainty shown for cyst segmentation were due to the limited extension of this class in the training dataset (cysts are present only on 7% of the 2D images used to train the model). Multi-class weighted Dice loss was partly able to handle the high unbalance in the dataset, also combined with the choice of storing the best parameter configuration with respect to the Dice score of cysts. However, when cysts were very small or located in regions not represented in the training set, the predicted masks were often unable to identify the structure and the uncertainty maps did not provide useful information, being overconfident in the wrong classification. This can be quantitatively appreciated in Figure 4h, where for cysts 30% of the population presented null RIU values and 50%–60% had RIU lower than 0.6, while for tumors only 20% of the population had RIU lower than 0.6 and for kidneys RIU was higher than 0.8 in any case.

Looking at the uncertainty maps in Figures 5, it can be seen that uncertainties are focused on the boundary of the object in case of good segmentation, as reported also by Maruccio et al.,¹⁷ whereas the uncertain region can cover the whole structure in case of missed object identification. This could be a practical indication for the user, who would trust the result of the segmentation when the uncertainty is limited to the boundaries, as the uncertainty is mostly due to changes in contrast and

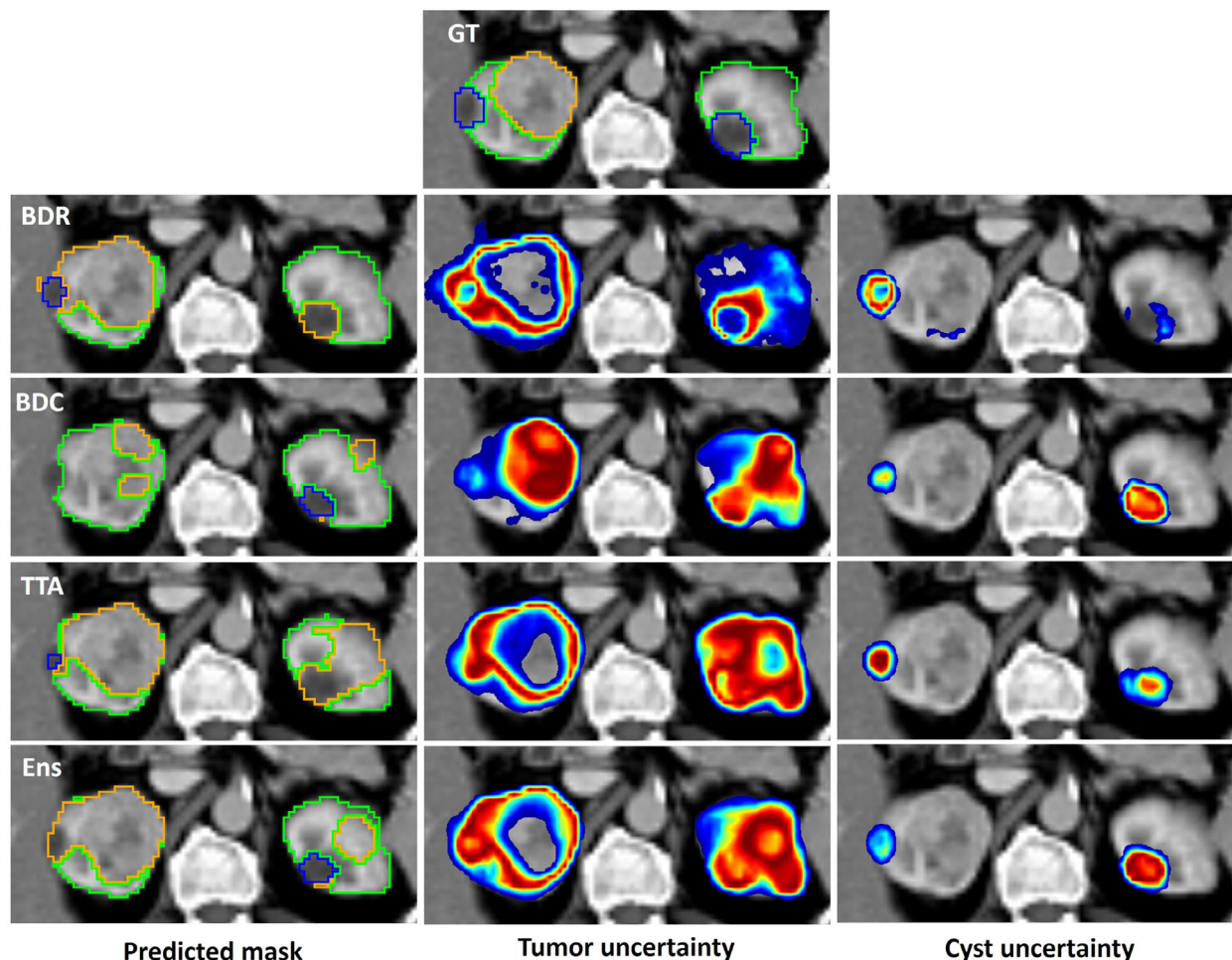


FIGURE 6 Example of segmentation results: kidneys (green), tumor (orange) and cyst (blue) predicted by BDR, BDC, TTA, and Ens networks (second column), and the associated tumor (third column) and cyst (fourth column) uncertainty maps; reference segmentation GT is also provided in the first column. Ens, ensemble method; GT, ground truth; TTA, test-time augmentation.

noise, and would pay more attention in revising the segmentation when the uncertainty covers a whole region, meaning that the model is not sure if the tumor/cyst is actually present. In addition, the possibility of having uncertainty maps separated for tumor and cyst can provide additional information to the reader, especially in case of misclassification.

Interestingly, significant differences were present among the four approaches for every uncertainty quantification metric. Indeed, TTA had always larger uncertainty values, whereas BDR presented the lowest RIU and highest UA, justified by the higher number of correct but uncertain pixels. This difference is qualitatively reflected on the uncertainty regions in Figure 6, being BDR more confident in its prediction, whilst TTA showed the largest uncertainty area. Looking at the differences between the two BCNNs, it can be noted that BDC provided higher uncertainty compared to BDR, especially for misclassified objects and rare class labels, as also reported by Mobiny et al.¹²

These results show a clear different behavior across the four approaches. While BDR is more risk taking, TTA is more conservative because it gives greater uncertainties to larger portions of the masks, thus generating warnings as soon as the segmentation result does not have high confidence. BDC and Ens presented an intermediate behavior, with uncertain regions larger than BDR, as highlighted also by lower UA values, as shown in Figure 4. However, BDC had lower accuracy compared to Ens, as well as lower RIU and higher UA, due to a higher number of incorrect pixels. This trend was confirmed both on the internal text set using Kits21 data and on the independent test on Kits23 data, thus making our findings more robust.

Looking at the inter-observer analysis, results on Kits21 dataset confirmed the trend seen for the accuracy scores, with the main differences observable across structures, rather than across approaches. Specifically, it is clear that none of the approaches was able to mimic inter-observer segmentation variability for the

identification of tumor and cysts. This is explainable by the low accuracy as well as the low RIU values obtained in these structures. As for Qubiq dataset, it can be highlighted that differences were visible across the approaches, where the uncertainty for kidney segmentation obtained from BDR was in lower agreement with the inter-observer variability compared to the others, not visible when assessed on Kits21. This may be attributed to a lower generalization ability for this approach, when it deals with out-of-distribution data, due to the estimation of lower uncertainties, compared to BDC.¹² The inter-observer variability in generating reference labels could also be used to improve model calibration, as proposed by Islam and Glocker.³²

This different behavior is a key aspect for the practical application of uncertainty quantification, because decision-makers are allowed to choose the approach most in line with the risk propensity degree required by the application and by the policy decided by the clinical staff. In this sense, BDR was the most risk-taking approach, able to provide higher accuracy in its prediction, but failing to assign uncertainty on incorrect segmentation in every case. The other three approaches were more conservative, providing large uncertainty regions, especially when they failed in their prediction, with the drawback of giving alert also on correct areas. When kidneys were considered, the approaches performed similarly, with differences only on the external out-of-distribution Qubiq dataset, which highlighted that the uncertainty introduced by BDR was in lower agreement with inter-observer variability compared to the others. In the choice of the approach, one may also consider that BDC and Ens gave similar results in terms of uncertainty metrics, but training Ens required more computational efforts. Finally, it should be noted that neither of the considered approaches was able to satisfactorily address the segmentation of small and under-represented structures, such as the cyst, as illustrated in the Figure S6. Nonetheless, the adoption of one of these approaches can be of practical utility during the segmentation process. When used to facilitate and speed up the manual adjustment process in the semi-automatic segmentation, or to implement a semi-supervised learning,³³ one may be aware of the limitations related to difficult and under-represented structures and of the differences across the possible approaches.

The main limitation of this study lies in the implementation of a 2D architecture, when several 3D networks have been already effectively proposed for this purpose. However, this decision was influenced by our aim to implement a network that could be readily transferred to clinical settings. In fact, despite 3D segmentation advancements, 2D segmentation remains widely used and accepted in several medical imaging applications.³⁴ Moreover, previous research³⁵ has highlighted potential limitations of 3D compared to 2D CNNs, particularly

in cases involving imaging data with large anisotropic voxels, as in our datasets. Finally, we believe that the use of a 3D architecture would not significantly change the main findings on the uncertainty quantification in regions with varying accuracy. Anyway, we acknowledge the potential to extend the evaluation to a 3D framework as well.

In conclusion, our work represents an advancement in bringing state-of-the-art *U-Net* models into clinical application, where the current bottleneck is the trustiness of the segmentations. Moreover, we also paved the way for future analyses, as the approach followed in this work is of general validity and can also be applied to future network configuration that will be the new benchmark from time to time, and also on different datasets and clinical applications.

ACKNOWLEDGMENTS


This work was supported by the Italian Ministry of University and Research, grant protocol number 2022B23JT5, PRIN (PROGETTI DI RICERCA DI RILLEVANTE INTERESSE NAZIONALE) 2022, funded by the European Union – Next generation EU (PNRR M4.C2.1.1).

CONFLICT OF INTEREST STATEMENT


The authors declare no conflicts of interest.

ORCID

Elisa Scalco  <https://orcid.org/0000-0003-2721-5792>

Giovanna Rizzo 

<https://orcid.org/0000-0002-6341-1304>

Ettore Lanzarone 

<https://orcid.org/0000-0001-8816-9086>

REFERENCES

- Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer; 2015:234-241.
- Heller N, Isensee F, Maier-Hein KH, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: results of the KITS19 challenge. *Med Image Anal.* 2021;67:101821.
- Isensee F, Petersen J, Klein A, et al. nnU-net: self-adapting framework for u-net-based medical image segmentation. arXiv preprint arXiv:1809.10486. 2018.
- Heller N, Isensee F, Trofimova D, Tejpaul R, Papanikolopoulos N, Weight C. Kidney and Kidney Tumor Segmentation. In: MICCAI 2021 Challenge, KITS 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings, Vol 13168. Springer Nature; 2022.
- Zhao Z, Chen H, Wang L. A coarse-to-fine framework for the 2021 kidney and kidney tumor segmentation challenge. In: Kidney and Kidney Tumor Segmentation: MICCAI 2021 Challenge, KITS 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings. Springer; 2022: 53-58.

6. Ovadia Y, Fertig E, Ren J, et al. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Adv Neural Inf Process Syst*. 2019;32.
7. Gawlikowski J, Tassi CRN, Ali M, et al. A survey of uncertainty in deep neural networks. *Artif Intell Rev*. 2023;56:1513-1589.
8. Seoni S, Jahmunah V, Salvi M, Barua PD, Molinari F, Acharya UR. Application of uncertainty quantification to artificial intelligence in healthcare: a review of last decade (2013–2023). *Comput Biol Med*. 2023;165:107441.
9. Abdar M, Pourpanah F, Hussain S, et al. A review of uncertainty quantification in deep learning: techniques, applications and challenges. *Inf Fusion*. 2021;76:243-297.
10. Jospin LV, Laga H, Boussaid F, Buntine W, Bennamoun M. Hands-on Bayesian neural networks - a tutorial for deep learning users. *IEEE Comput Intell Mag*. 2022;17:29-48.
11. Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: International Conference on Machine Learning. PMLR; 2016:1050-1059.
12. Mobiny A, Yuan P, Moulik SK, Garg N, Wu CC, Van Nguyen H. Dropconnect is effective in modeling uncertainty of Bayesian deep networks. *Sci Rep*. 2021;11:1-14.
13. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Adv Neural Inf Process Syst*. 2017;30.
14. Wang G, Li W, Aertsen M, Deprest J, Ourselin S, Vercauteren T. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*. 2019;338:34-45.
15. Camarasa R, Bos D, Hendrikse J, et al. Quantitative comparison of monte-carlo dropout uncertainty measures for multi-class segmentation. In: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis: Second International Workshop, UNSURE 2020, and Third International Workshop, GRAIL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings 2. Springer; 2020:32-41.
16. Ruan Y, Li D, Marshall H, et al. Mt-UcGAN: multi-task uncertainty-constrained GAN for joint segmentation, quantification and uncertainty estimation of renal tumors on CT. In: Medical Image Computing and Computer Assisted Intervention—MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23. Springer; 2020:439-449.
17. Maruccio FC, Eppinga W, Laves M-H, et al. Clinical assessment of deep learning-based uncertainty maps in lung cancer segmentation. *Phys Med Biol*. 2024;69:035007.
18. Du Z, Chen W, Xia Q, Shi O, Chen Q. Trends and projections of kidney cancer incidence at the global and national levels, 1990–2030: a Bayesian age-period-cohort modeling study. *Biomark Res*. 2020;8:1-10.
19. Perazella MA, Dreicer R, Rosner MH. Renal cell carcinoma for the nephrologist. *Kidney Int*. 2018;94:471-483.
20. Low G, Huang G, Fu W, Moloo Z, Gzigis S. Review of renal cell carcinoma and its common subtypes in radiology. *World J Radiol*. 2016;8:484-500.
21. Rombolotti M, Sangalli F, Cerullo D, Remuzzi A, Lanzarone E. Automatic cyst and kidney segmentation in autosomal dominant polycystic kidney disease: comparison of U-Net based methods. *Comput Biol Med*. 2022;146:105431.
22. McClure P, Kriegeskorte N. Robustly representing uncertainty in deep neural networks through sampling. *arXiv preprint arXiv:1611.01639*. 2016.
23. Tallón-Ballesteros A, Chen C. A study on the effect of dropconnect to control overfitting in designing neural networks. Machine Learning and Artificial Intelligence: Proceedings of MLIS 2020. 2020;332:178.
24. Wang R, Lei T, Cui R, Zhang B, Meng H, Nandi AK. Medical image segmentation using deep learning: a survey. *IET Image Proc*. 2022;16:1243-1267.
25. Milletari F, Navab N, Ahmadi S-A. V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). IEEE; 2016:565-571.
26. Roy AG, Conjeti S, Navab N, Wachinger C. Bayesian QuickNAT: model uncertainty in deep whole-brain segmentation for structure-wise quality control. *NeuroImage*. 2019;195:11-22.
27. Herzog L, Murina E, Dürr O, Wegener S, Sick B. Integrating uncertainty in deep neural networks for MRI based stroke analysis. *Med Image Anal*. 2020;65:101790.
28. Hasan M, Khosravi A, Hossain I, Rahman A, Nahavandi S. Controlled dropout for uncertainty estimation. In: IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE; 2023: 973-980.
29. Jungo A, Meier R, Ermis E, et al. On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention—MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I. Springer; 2018:682-690.
30. Yeghiazaryan V, Voiculescu I. Family of boundary overlap metrics for the evaluation of medical image segmentation. *J Med Imaging*. 2018;5:015006-015006.
31. Asgharnezhad H, Shamsi A, Alizadehsani R, et al. Objective evaluation of deep uncertainty predictions for covid-19 detection. *Sci Rep*. 2022;12:1-11.
32. Islam M, Glocker B. Spatially varying label smoothing: capturing uncertainty from expert annotations. In: Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings 27. Springer; 2021:677-688.
33. Yu L, Wang S, Li X, Fu C-W, Heng P-A. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In: Medical image computing and computer assisted intervention—MICCAI 2019: 22nd international conference, Shenzhen, China, October 13–17, 2019, proceedings, part II 22. Springer; 2019:605-613.
34. Conze P-H, Andrade-Miranda G, Singh VK, Jaouen V, Visvikis D. Current and emerging trends in medical image segmentation with deep learning. *IEEE Trans Radiat Plasma Med Sci*. 2023;7(6):545-569.
35. Zhu J, Bolsterlee B, Chow BV, et al. Deep learning methods for automatic segmentation of lower leg muscles and bones from MRI scans of children with and without cerebral palsy. *NMR Biomed*. 2021;34:e4609.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Scalco E, Pozzi S, Rizzo G, Lanzarone E. Uncertainty quantification in multi-class segmentation: Comparison between Bayesian and non-Bayesian approaches in a clinical perspective. *Med Phys*. 2024;51:6090–6102.

<https://doi.org/10.1002/mp.17189>