Applied Soft Computing Journal 205 (2027) 116818



Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc



Confidence-weighted ensemble strategies based on task-informed data-augmentation to improve segmentation robustness: application to kidney MRI

Elisa Scalco [a,*], Rossella Damiano [a], Anna Caroli [b], Marco L. Della Vedova [c], Andrea Bombarda [d], Ettore Lanzarone [d]

[a] Institute of Biomedical Technologies (ITB), National Research Council of Italy (CNR), Segrate (MI), Italy
[b] Department of Biomedical Engineering, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Ranica (BG), Italy
[c] Department of Mechanical Engineering, Chalmers University of Technology, Gothenburg, Sweden
[d] Department of Management, Information and Production Engineering (DIGIP), University of Bergamo, Dalmine (BG), Italy

HIGHLIGHTS

- Distribution shifts affect segmentation accuracy and uncertainty reliability in kidney MRI.
- Task-informed data augmentation improves robustness of kidney MRI segmentation.
- Confidence-weighted ensembles outperform single models and soft voting.
- Confidence-based aggregation enhances accuracy–uncertainty consistency.
- The framework improves error detection and assists clinician-AI joint decision-making.

ARTICLE INFO

Dataset link: T2-weighted Kidney MRI Segmentation (Reference data)

Code link: Zenodo repository

Keywords:
Deep learning robustness
Data-augmentation
Uncertainty-weighted ensemble
Kidney MRI segmentation
Natural MRI alterations

ABSTRACT

Deep learning–based segmentation models have achieved remarkable performance in medical image analysis, yet their deployment in clinical practice is hindered by limited robustness and unreliable uncertainty estimates under distribution shifts. In renal magnetic resonance imaging (MRI), variability due to geometric transformations, intensity inhomogeneities, and resolution loss can substantially degrade model reliability. In this study, we propose an uncertainty-aware ensemble framework that combines task-informed data augmentation with confidence-weighted prediction aggregation to improve MRI renal segmentation robustness and reliability. Multiple U-Net models equipped with Monte Carlo Dropout are trained using distinct, realistic MRI-specific augmentation strategies, each targeting a specific type of natural variation (affine transformations, bias field artifacts, and blurring). At inference time, predictions are aggregated using ensemble strategies that explicitly leverage predictive uncertainty to weight or select the most reliable outputs. Model performance is evaluated on a renal MRI dataset comprising healthy controls and chronic kidney disease patients, under controlled perturbations reflecting realistic clinical distortions. Robustness is assessed through a unified framework that jointly considers segmentation accuracy and multiple uncertainty-based metrics. Results demonstrate that confidence-weighted ensemble strategies consistently outperform single-model and conventional soft-voting baselines, achieving higher robustness across all perturbation types. The proposed approach enhances error detection and supports more informed human–AI collaboration by identifying unreliable predictions. Overall, these findings indicate that selectively exploiting uncertainty within ensembles of task-specialized models is a more effective strategy than relying on comprehensive augmentation within a single model, offering a robust and clinically relevant solution for renal MRI segmentation.

* Corresponding author.
Email address: elisa.scalco@cnr.it (E. Scalco).

https://doi.org/10.1016/j.asoc.2026.116818
Received 19 May 2026; Received in revised form 28 September 2026; Accepted 1 November 2026
Available online 5 November 2026
1568-4946/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

The widespread development of deep learning (DL) solutions has significantly transformed the medical field, enabling major advancements in applications such as medical image analysis, disease diagnosis, and treatment planning [1]. In particular, DL models have enabled the automated segmentation of organs and tissues from medical images, a crucial step for several clinical and research tasks, including morphological evaluation, volumetric quantification, and disease monitoring [2,3]. These methods allow accurate and efficient identification and delineation of pathophysiological structures, paving the way for more precise and scalable healthcare solutions [4].

However, DL-based segmentation networks often suffer from performance degradation when exposed to data distributions that differ from those seen during training [5]. In the context of medical imaging, these variations include image noise, contrast changes, illumination shifts, resolution loss, and cross-domain acquisition differences [6]. This is the case with Magnetic Resonance Imaging (MRI), which provides high diagnostic quality while being inherently affected by variability arising from acquisition artifacts (e.g., bias field) and geometric differences (e.g., affine transformations).

These variations can compromise the performance of DL models, highlighting the need to assess performance degradation not only in terms of accuracy but also through additional indicators, such as prediction uncertainty. Understanding how a model's confidence levels respond to perturbations can provide deeper insights into its reliability [5]. For instance, a robust model should not only produce accurate outputs under varying conditions, but also exhibit appropriate uncertainty estimates when faced with atypical inputs [7,8].

By quantifying the confidence in a model's predictions, uncertainty estimation enables the identification of potentially unreliable outputs that may require further refinement. In clinical settings, this allows clinicians to make more informed decisions by appropriately complementing automated predictions with human expertise, moving toward effective *human-AI collaboration*, where AI systems work jointly with humans as teammates or partners to solve problems [9]. This approach is particularly desirable in healthcare, where medical professionals are often overworked, thereby increasing the risk of human error.

We previously assessed in [10] the robustness of a DL-based segmentation network against natural MRI variations, defined as its ability to maintain stable performance under increasingly altered inputs, and measured by combining accuracy and uncertainty metrics. We showed that, although accuracy remains acceptable under such perturbations, the model exhibits degraded uncertainty estimates. This effect was more pronounced in patients whose anatomy is inherently more complex and variable and in the presence of bias field alterations.

Data augmentation has been proposed as a regularization- and robustness-enhancing strategy, as it simulates natural input variations during training. Well-designed augmentation consistently improves robustness to geometric variations, intensity and contrast shifts, noise, blur, and cross-domain differences [11]. In this context, task-informed augmentation is defined as a modality- and task-aware strategy that leverages domain physics to synthesize only realistic, label-preserving variations expected at deployment, enabling models to learn invariant properties and remain robust in real-world settings. Task-informed augmentation has proven highly effective, as demonstrated for the nnU-Net, where a standardized geometric and photometric augmentation policy is a key factor in its cross-dataset robustness [12].

In this study, we train multiple U-Net models incorporating Monte Carlo Dropout to provide uncertainty estimates, and combine them using an uncertainty-weighted ensemble. Each model is trained with a specific task-informed data augmentation strategy. The key idea is to specialize each U-Net for a specific type of natural variation and, at inference time, leverage uncertainty estimates to assign higher weights to predictions from the most confident model (and thus best suited) for the input. The robustness of the proposed segmentation architecture against natural MRI variations is then evaluated by jointly considering accuracy and uncertainty metrics, and is compared with models trained using a single data augmentation strategy.

The proposed approach has been validated on renal MRI scans from patients with Chronic Kidney Disease (CKD), where accurate kidney identification is particularly important, as anatomical changes can significantly impact diagnosis and treatment [13].

The specific goals of this study are: *i*) to analyze the impact of different task-informed data augmentation strategies, considering geometric variations (affine transformations), intensity shifts (bias field artifacts), and resolution loss (blurring alterations); *ii*) to combine different task-informed data augmentation strategies using uncertainty-weighted ensemble approaches and compare them with single data augmentation and benchmark strategies; *iii*) to assess model robustness under controlled affine, bias field, and blur transformations applied to test images, by jointly considering accuracy and uncertainty metrics. Moreover, from a clinical perspective, an additional goal of this study is to support decision-making in patients with CKD by providing an effective and reliable renal MRI analysis framework.

2. Literature review

Data augmentation has emerged as a central strategy for improving the robustness of deep learning models in medical imaging, particularly in MRI, where variability induced by acquisition protocols, scanner differences, and physical artifacts is substantial. Conventional augmentation techniques—such as geometric transformations, intensity scaling, noise injection, and blurring—are widely used to expose models to a broader input distribution and mitigate overfitting [11,14]. However, in MRI, generic augmentation policies may introduce unrealistic variations that do not reflect the underlying imaging physics, potentially limiting their effectiveness in real-world deployment. To address this limitation, recent studies have focused on realistic, physics-driven augmentation strategies. Chen et al. [15] proposed an adversarial augmentation method that generates physically plausible signal corruptions modeling intensity inhomogeneities caused by bias field artifacts—a major source of variability in MR imaging—demonstrating improved generalization, including in low-data regimes. This work was later extended to AdvChain [16], a framework that dynamically composes photometric and geometric transformations to generate realistic yet challenging variations, further improving robustness in cardiac and prostate MRI segmentation. More broadly, domain-specific transformations have been shown to better approximate real-world distribution shifts, supporting improved cross-scanner generalization [17]. This has motivated the paradigm of task-informed augmentation, where transformations are selected based on domain knowledge and task-specific invariances, enabling models to learn features that are both discriminative and robust to expected perturbations. More recent work has explored augmentation strategies that improve out-of-distribution robustness without explicitly targeting specific sources of shift. Vaish et al. [18] evaluated MixUp and Auxiliary Fourier Augmentation, a frequency-domain augmentation strategy, on cardiac and prostate MRI, demonstrating improved performance across diverse transformations—including scanner and protocol variability—while promoting more structured feature representations. In a related direction, Panfilov et al. [19] showed that MixUp regularization improves cross-scanner robustness in knee MRI segmentation, providing evidence that interpolation-based augmentation strategies can serve as a computationally efficient alternative to explicit domain adaptation when the target acquisition settings are not available at training time. Importantly, augmentation design also influences uncertainty calibration: models trained on more diverse and realistic input distributions tend to produce better-calibrated confidence estimates, which is critical in clinical settings where identifying unreliable predictions is as important as achieving high accuracy [7]. While existing augmentation strategies improve average generalization, they operate primarily at training time and do not explicitly model the relationship between

specific perturbation types and model reliability at inference time. As a result, these approaches fail to exploit the complementary strengths of models trained under different distributional assumptions, which may exhibit perturbation-dependent robustness.

Assessing robustness is essential, as when models are deployed in real-world settings, they must maintain reliable performance not only on clean data but also under distribution shifts, noise, and other perturbations that could commonly occur in practice. This applies to any models, including those used for classification [20,21], segmentation [10,22], and regression [23]. In the context of image segmentation, robustness is traditionally assessed through accuracy-oriented metrics quantifying the agreement between predicted and ground truth (GT) regions. This agreement is often assessed by considering the Dice score, which evaluates the spatial overlap between areas and is particularly robust to class imbalance, a common characteristic of medical imaging datasets [24,25]. Only a few studies consider pixel-wise accuracy, as it is generally considered inadequate for segmentation tasks because of its sensitivity to dominant background classes, often leading to overly optimistic evaluations [22,26]. However, recent literature highlights that accuracy alone provides an incomplete picture of robustness for medical image segmentation, as models may maintain high overlap scores while producing unreliable or overconfident predictions.

To address this limitation, uncertainty-based metrics have gained increasing attention as complementary indicators of robustness [27]. Deep neural networks, including U-Nets, are known to often be poorly calibrated and produce overconfident outputs even when incorrect [28]. For this reason, uncertainty quantification methods, such as Monte Carlo dropout, Bayesian neural networks, and deep ensembles, have been adopted. They enable the estimation of predictive confidence, typically through measures such as entropy, variance, or mutual information [29]. These uncertainty estimation strategies have been shown to correlate with segmentation errors and can be used to detect out-of-distribution samples or failure cases, thereby enhancing the reliability of the model in safety-critical applications [24,28].

More recently, research has focused on defining frameworks for combining accuracy and uncertainty into unified robustness measures. The motivation is that robust models should not only achieve high segmentation quality but also provide reliable confidence estimates.

In line with this research, in this work, we propose the use of ensembles of U-Net architectures, weighted according to the uncertainty of their outputs, in order to achieve more robust segmentation. Classically, ensemble strategies are used for improving predictive performance by combining multiple models and aggregating their outputs (e.g., by voting or averaging) [30].

A widely used approach is the one referred to as a *deep ensemble* [31], where multiple networks (often with the same architecture) are trained independently with different random initializations or data shuffles, and predictions are averaged [32]. Beyond accuracy gains, deep ensembles have been shown to provide high-quality predictive uncertainty estimates and improved calibration compared with several approximate Bayesian methods, while remaining straightforward to implement and parallelize [33]. The most common aggregation rule in deep ensemble is *soft-voting* of the outputs provided by each ensemble member [32]. However, it implicitly assumes that all components are equally reliable across all inputs and regions. We argue that this assumption is frequently violated in medical images (e.g., because of motion artifacts or low SNR regions). This motivates our uncertainty-weighted approach, where member predictions are combined using uncertainty as a reliability proxy. Similar approaches have been reported in the literature; however, these works typically consider ensemble strategies in which all ensemble members are trained on the same dataset, and they focus primarily on semi-supervised learning settings [34] or on the segmentation of locally unknown regions [3]. For example, uncertainty is used to weight ensemble member predictions in the context of lung cancer segmentation [35]. Unlike our work, where we use multiple U-Nets trained using different data-augmentation strategies, their work adopts an ensemble composed of multiple architectures (K-Net, ResNeSt, Segformer, Twins) and combines predictions from the two most reliable models.

The proposed framework is therefore distinct from several related ensemble and uncertainty-aware paradigms. Standard deep ensembles typically train multiple instances of the same model, often differing only in random initialization, data shuffling, or bootstrap sampling, and then aggregate their predictions to improve accuracy and uncertainty estimation [32,33]. Similarly, conventional soft-voting ensembles combine the predicted probability maps of all members with equal weights, implicitly assuming that each model is equally reliable for every input and spatial region [30,32]. Test-time augmentation, instead, generates multiple transformed versions of the same test image and averages or analyzes the resulting predictions to improve robustness or estimate uncertainty. However, it does not train perturbation-specialized models nor select among models according to perturbation-dependent reliability [36]. Finally, selective prediction and uncertainty-based quality-control methods use confidence or uncertainty to reject, flag, or assess unreliable predictions rather than to dynamically aggregate the outputs of task-specialized segmentation models [28,37,38]. In contrast, our contribution lies in training multiple U-Net models under distinct, task-informed augmentation regimes, with each model targeting a specific class of realistic MRI alterations, and in combining their predictions through voxel-level or case-level uncertainty-informed weighting or selection. Thus, uncertainty serves not merely as a post hoc indicator of prediction quality, but as an active aggregation criterion that enables the framework to exploit the complementary robustness of perturbation-specialized models.

3. Materials and methods

3.1. MRI dataset

A publicly available dataset comprising T2-weighted coronal renal MRI scans from 60 subjects was used in this study [39]. The cohort included 30 patients with CKD and 30 healthy controls (HC). The dataset also included manually delineated kidney masks for each scan, which served as GT annotations.

50 subjects (25 CKD and 25 HC), each imaged once, contributed 649 two-dimensional slices, which were used for model development, with an 80%/20% split at the slice level for training and validation, respectively. The remaining 10 subjects (5 CKD and 5 HC), completely independent of and separated from the training set, underwent 5 repeated scans with repositioning, yielding a total of 650 two-dimensional slices. These data were used as an independent test set to evaluate the performance and robustness of the U-Net-based segmentation models.

Prior to network input, all MRI scans were preprocessed using the same pipeline described in [40], in which the dataset was introduced.

3.2. Training with data-augmentation

Kidney segmentation was performed using a two-dimensional U-Net architecture, following the design originally proposed by Daniel et al. [40] for this dataset, as well as the implementation adopted in previous studies [10,41].

During training, the following on-the-fly data augmentation strategies were applied to increase data variability and simulate realistic MRI-related alterations, using the TorchIO library [42]:

- *Affine*, *Bias Field*, and *Blur* transformations were applied individually;
- combined augmentation strategies were applied by pairing affine transformations with either bias field (*Aff* + *BF*) or blur (*Aff* + *Blur*), as affine augmentation is commonly included in standard training pipelines;
- a benchmark augmentation strategy incorporating all three transformations (*Aff* + *BF* + *Blur*).

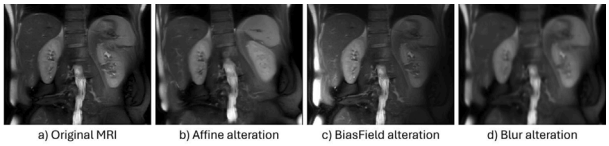

Fig. 1. Illustrative example of MRI digital alterations applied to a CKD test case: a) original MRI; b) *Affine* alteration; c) *Bias Field* alteration; d) *Blur* alteration. Large alterations are reported to visually appreciate the artifacts.

An illustrative example of these transformations is shown in Fig. 1, while all combinations are summarized in Fig. 2 (the DATA AUGMENTATION column on the left).

Model training was performed using the Adam optimizer and the Dice++ loss function ($\mathcal{L}_{DSC++}$), introduced by Yeung et al. [24], which aims to improve model calibration and the reliability of predictive uncertainty estimates, as has already been assessed in previous studies [10,41].

Hyperparameter tuning for the individual models under each data augmentation strategy was performed on the same training partition, by training each configuration once and evaluating it on the same validation set. Specifically, we selected the configuration that achieved the highest Dice score on the validation set using a grid-search strategy. The evaluated hyperparameters included batch size (8 or 16), intensity-normalization method (z-score normalization or min–max normalization), and the number of training epochs (100, 150, or 200). For each model and combination of the other hyperparameters, training began with 100 epochs, after which we assessed whether the training loss had reached a plateau. If convergence had not been achieved, training was extended to 150 epochs and then, if necessary, to 200 epochs. All models reached a plateau within 200 epochs, making this a sufficient upper limit for the search. For each configuration, we identified the epoch with the highest validation performance and retained the corresponding weights. The best-performing configuration was then selected for each model. This procedure ensured that each model was evaluated under near-optimal hyperparameter settings while minimizing unnecessary training, thereby allowing the augmentation strategy itself to remain the primary factor driving performance differences on the altered data.

The selected models were subsequently evaluated on the unaltered test set using the Dice score. Importantly, no hyperparameter selection or model optimization was performed using test-set data. This assessment is reported in Table 1.

3.3. Prediction on test images with MRI alterations

The trained segmentation models were applied while introducing controlled corruptions to the test data through MRI-specific transformations implemented using the TorchIO library [42]. Isolated single-type perturbations were considered to enable a controlled assessment of the robustness of each augmentation strategy and ensemble approach under distinct sources of image degradation. The applied transformations were identical to those used for data augmentation during training, i.e., affine, bias field, and blurring.

For each transformation, the parameter ranges (Table 2) were defined to reflect realistic distribution shifts commonly encountered in clinical practice, as characterized in [10]. Each transformation was applied 20 times, with transformation magnitudes T_r ($r = 1, \dots, 20$) randomly sampled within the specified range. A total of 3000 MRI volumes (25 scans from 5 CKD patients and 25 scans from 5 HC subjects, each altered using 3 transformation types with 20 magnitude levels) were thus used for robustness assessment.

Predictive uncertainty was quantified using the Monte Carlo dropout (MCD) approach, in which dropout layers at the end of each encoder and decoder block remain active during both training and inference [43]. During inference, the network performed $R = 50$ stochastic forward passes, each with a random subset of neurons deactivated according to the specified dropout rate. The resulting set of R predictions can be interpreted as samples from a predictive distribution, enabling a probabilistic characterization of model outputs.

The mean predicted probability $\overline{p}(i,j)$ of each pixel (i,j) was computed by averaging the predicted probabilities $p_r(i,j)$ across the R forward passes. The resulting mean prediction map was then binarized using a threshold of 0.5 on a pixel-wise basis to obtain the mean prediction (MP) mask.

The pixel-wise predictive uncertainty $U(i,j)$ of each pixel (i,j) was quantified using the predictive entropy [44]:

$$U(i,j) = -\overline{p}(i,j)\log\left(\overline{p}(i,j)\right) - \left(1-\overline{p}(i,j)\right)\log\left(1-\overline{p}(i,j)\right) \quad (1)$$

3.4. Ensemble procedures

Predictions obtained from the data augmentation strategies excluding the benchmark – *Affine*, *Bias Field*, *Blur*, *Aff+BF*, and *Aff+Blur* – were combined to produce the final segmentation on the entire MRI volume. They are denoted by $p_\tau(i,j,k)$ with k indicating the MRI slice and $\tau = 1, \dots, 5$.

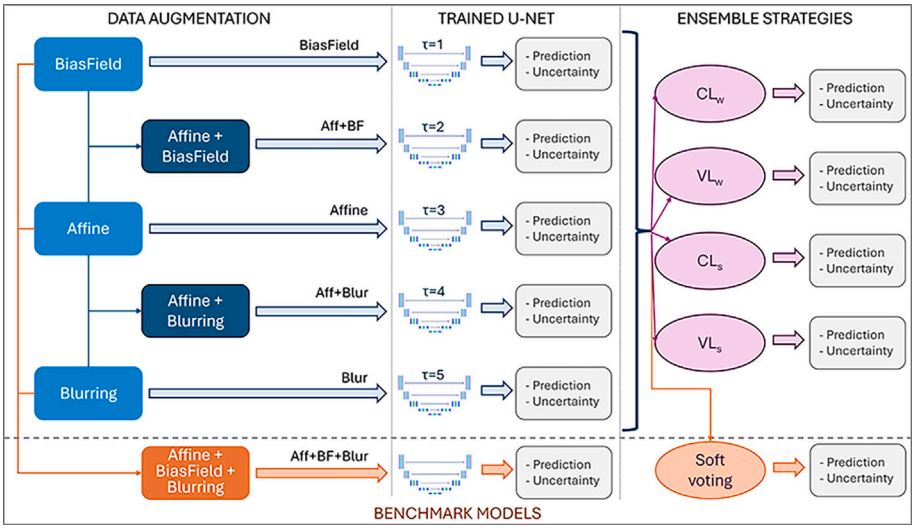

Fig. 2. Framework for data augmentation strategies and ensemble procedures. Benchmark models are also indicated.

Table 1
Hyperparameter tuning of the individual models under each data augmentation strategy. The selected configuration, determined by maximizing the Dice score on the validation set, is reported. Evaluation results on the unaltered independent test set are also provided.

Training type	Data augmentation	Intensity normalization	Batch size	Epochs	*Dice*	
					validation	test
Affine	scale ∈ [0.75, 1.25] degree ∈ [−20, 20] translation up to 0.25 W, 0.25 H	Z-score	16	100	0.88	0.90
Bias Field	BF coeff ∈ [0, 1]	Min-max	8	100	0.88	0.86
Blur	Gaussian kernel std ∈ [0, 3]	Min-max	8	100	0.88	0.83
Aff + BF	scale ∈ [0.75, 1.25] degree ∈ [−20, 20] translation up to 0.25 W, 0.25 H BF coeff ∈ [0, 1]	Z-score	16	150	0.91	0.92
Aff + Blur	Scale [0.75–1.25] degree ∈ [−20, 20] translation up to 0.25 W, 0.25 H Gaussian kernel std ∈ [0, 3]	Min-max	16	150	0.89	0.91
Aff + BF + Blur	Scale [0.75–1.25] degree ∈ [−20, 20] translation up to 0.25 W, 0.25 H BF coeff ∈ [0, 1] Gaussian kernel std ∈ [0, 3]	Min-max	16	200	0.91	0.92

Table 2
Implemented transformations with their respective parameter ranges, according to the library TorchIO. Parameters not listed were set to default values.

Transformation	Parameter	Range
Affine	3D scaling	∈ [0.9, 1.1]
	3D rotation	∈ [−5, 5] degrees
	3D shift	∈ [−3, 3] mm
Bias Field	coefficients	∈ [0, 1]
	order	= 3
Blur	Gaussian kernel std	∈ [0, 3]

The proposed ensemble framework leverages predictive uncertainty to weight the contribution of each augmentation-based prediction, assigning greater influence to more confident outputs. A schematic overview of the ensemble procedure is shown in Fig. 2.

As an initial filtering step, predictions that failed to capture the target anatomy were excluded using a volume-based criterion. For each prediction, the segmented volume V_τ was computed and normalized with respect to the largest predicted volume, yielding

$$v_\tau = \frac{V_\tau}{\max_{\tau=1,\dots,5}\{V_\tau\}}. \tag{2}$$

Predictions with $v_\tau < 0.5$ were excluded from the ensemble. In our data, we empirically observed that severe segmentation failures were often associated with unreliable uncertainty estimates, and that a threshold of 0.5 effectively suppressed such cases.

To define the ensemble weights, the predictive uncertainty $U_\tau(i,j,k)$ was rescaled to the range $[0,1]$ using the maximum uncertainty value within each volume, and the corresponding confidence map was computed voxel-wise as:

$$Conf_\tau(i,j,k) = 1 - \frac{U_\tau(i,j,k)}{\max\{U_\tau(i,j,k)\}}. \tag{3}$$

Four ensemble strategies were investigated: two based on weighted combinations of the predictions and two based on selecting the most confident prediction. For each alternative, one strategy is voxel-wise and the other is at the case level. The mathematical formulations for the voxel-wise probabilities resulting from each strategy are as follows:

- **Ensemble with voxel-level weighting (*VLw*)**: the contribution of each model τ is weighted voxel-by-voxel based on its confidence map $Conf_\tau(i,j,k)$.

$$p^{VLw}(i,j,k) = \frac{\sum_{\tau=1,\dots,5:\,v_\tau>0.5} w_\tau(i,j,k)\,p_\tau(i,j,k)}{\sum_{\tau=1,\dots,5:\,v_\tau>0.5} w_\tau(i,j,k)}$$
$$\text{with } w_\tau(i,j,k) = Conf_\tau(i,j,k) \tag{4}$$

- **Ensemble with case-level weighting (*CLw*)**: the contribution of each model τ is weighted based on its average confidence, computed as the mean value of the confidence map over the entire volume.

$$p^{CLw}(i,j,k) = \frac{\sum_{\tau=1,\dots,5:\,v_\tau>0.5} \overline{w}_\tau\, p_\tau(i,j,k)}{\sum_{\tau=1,\dots,5:\,v_\tau>0.5} \overline{w}_\tau}$$
$$\text{with } \overline{w}_\tau = \frac{\sum_{i,j,k} Conf_\tau(i,j,k)}{\sum_{i,j,k} 1} \tag{5}$$

- **Ensemble with voxel-level selection (*VLs*)**: the most confident prediction across the trained models τ is chosen voxel-by-voxel, thus neglecting the other contributions, based on the confidence map.

$$p^{VLs}(i,j,k) = p_{\tau^*(i,j,k)}(i,j,k)$$
$$\text{with } \tau^*(i,j,k) = \arg\max_{\tau=1,\dots,5:\,v_\tau>0.5} \{Conf_\tau(i,j,k)\} \tag{6}$$

- **Ensemble with case-level selection (*CLs*)**: the most confident prediction across the trained models τ is chosen for the entire MRI volume, thus corresponding entirely to one of the baseline predictions.

$$p^{CLs}(i,j,k) = p_{\tau^*}(i,j,k)$$
$$\text{with } \tau^* = \arg\max_{\tau=1,\dots,5:\,v_\tau>0.5} \left\{ \frac{\sum_{i,j,k} Conf_\tau(i,j,k)}{\sum_{i,j,k} 1} \right\} \tag{7}$$

As for the single transformations, $p^{VLw}(i,j,k)$, $p^{CLw}(i,j,k)$, $p^{VLs}(i,j,k)$, and $p^{CLs}(i,j,k)$ are then binarized using a threshold of 0.5 on a pixel-wise basis.

For each ensemble strategy, the predictive uncertainty map was computed from the final aggregated probability map using the

predictive-entropy formulation defined in Eq. (1), rather than by aggregating the uncertainty maps of the individual models.

The sensitivity of the proposed framework to the volume-ratio threshold and the uncertainty normalization strategy was explicitly assessed. Regarding the volume-ratio threshold, values of 0, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1 were evaluated. A threshold of 0 corresponds to confidence weighting without volume-based rejection, whereas a threshold of 1 effectively corresponds to selecting the prediction(s) with the largest volume. Regarding the normalization strategy, two additional approaches were evaluated alongside the per-volume maximum normalization in (3): *i*) no normalization; and *ii*) per-model maximum normalization in which, for each subject, each uncertainty map is normalized by the maximum uncertainty value observed across the five models. However, the confidence normalization procedure has only a minimal effect on the ensemble strategy, as the maximum uncertainty values across the different baseline models were generally comparable.

The optimal volume-ratio threshold and uncertainty normalization strategy were identified separately based on the baseline metrics of accuracy and uncertainty defined below in Section 3.5. These settings were subsequently used in the robustness analysis. To ensure a fair comparison, the evaluation was restricted to a subset of the altered test set consisting of the first available acquisition for each subject. In addition, we reported how often each individual model was excluded from the ensemble procedure at each volume-ratio threshold, and we evaluated the Spearman's rank correlation coefficient between $Dice$ and confidence. The results of these analyses are reported in the Supplementary Materials.

3.5. Evaluation framework

All models were implemented in PyTorch (version 2.5) using Python (version 3.11.8). Training was performed on a Microsoft Windows workstation equipped with an Intel® Core™ i7-13700 CPU (2.10 GHz), 32 GB of RAM, and an NVIDIA GeForce RTX 4070 Ti GPU. Additional details regarding the training and inference procedures, as well as the software and hardware configurations, are provided in Supplementary Table 11.

3.5.1. Computational cost

The following metrics were calculated for each individual model and for the ensemble approach, to assess computational requirements and execution time: total number of parameters, model size, billions of floating-point operations (GFLOPs), GPU memory usage, training time per epoch, total training time, and inference time per MRI volume.

3.5.2. Accuracy, uncertainty and calibration

Model robustness was assessed using the composite metrics proposed by Essbai et al. [27] and adapted to segmentation tasks in [10], which jointly account for predictive accuracy and uncertainty within a unified evaluation framework.

Predictive accuracy was quantified using $Dice$, computed between the GT and model predictions over the entire three-dimensional volumes.

Model uncertainty was evaluated using three complementary metrics (globally denoted as Unc), all in the range $[0, 1]$:

- **Incorrect–Uncertain Ratio (RIU)**: defined as the ratio between the number of pixels that are both incorrect and uncertain and the total number of incorrect pixels [45,46]. A pixel is defined as correct if the predicted class (after binarization) corresponds to the reference class, while a pixel is considered uncertain if its uncertainty value is higher than 0.1 (threshold chosen empirically). Higher RIU values indicate that incorrect predictions are associated with high uncertainty, suggesting effective identification of unreliable regions.
- **Uncertainty Accuracy (UA)**: defined as the ratio of the sum of incorrect–uncertain pixels and correct–certain pixels to the total number of pixels [45,46]. Maximizing UA favors uncertainty in erroneous predictions and confidence in correct ones.
- **Mean Confidence ($\overline{Conf}$)**: computed as the average of the confidence map $Conf(i, j, k)$ across all pixels. This metric provides a GT–independent measure of model confidence.

Spearman's rank correlation coefficient between $Dice$ and $\overline{Conf}$ was computed to assess the relationship between segmentation accuracy and uncertainty for each model.

Metrics assessing model calibration and error-detection capability were also analyzed. Specifically, error detection was quantified using the Kullback–Leibler (KL) divergence between the uncertainty distributions associated with correctly and incorrectly classified voxels. These distributions were estimated from histograms of voxel-wise uncertainty values, as previously described in [41]. Higher KL-divergence values indicate greater separation between correct and incorrect predictions, suggesting that prediction errors are preferentially associated with higher uncertainty and that the uncertainty estimates are consequently more informative. Model calibration was evaluated using the Brier score [47].

3.5.3. Robustness

The robustness metric rob quantifies the model's ability to maintain stable performance when exposed to altered or perturbed data. A low robustness value indicates that the performance of the model decreases significantly in the presence of such changes, suggesting that the model is sensitive to the transformations and lacks stability.

The robustness against any transformation type T was computed by considering three factors [27]: the probability of each transformation level π_{T_r}, the tolerance of the system to a decrease in performance $tol\left(x^{T_r}, th\right)$, where x denotes any performance or uncertainty metric and th the threshold for the minimum acceptable performance value, and the penalization for unacceptable performance values $dep\left(x^{T_r}, th\right)$:

$$rob^T = \frac{1}{2}\sum_{r=1}^{20}\left[tol\left(x^{T_r}, th\right) - dep\left(x^{T_r}, th\right)\right]\pi_{T_r} + \frac{1}{2}. \tag{8}$$

with

$$tol(x, th) = \frac{\max\{\min\{x, 1\} - th, 0\}}{1 - th} \tag{9}$$

$$dep(x, th) = \frac{\max\{th - \min\{x, 1\}, 0\}}{th} \tag{10}$$

The threshold th was set to 0.7 for $Dice$, 0.6 for RIU and $\overline{Conf}$, and 0.9 for UA. These thresholds were empirically determined based on the distribution of each metric within the population. The probability π_{T_r} was set to $\frac{1}{20}$ for all r, as each of the 20 transformations T_r was randomly sampled from the range of each transformation. The term $\frac{1}{2}$ is a regularization factor to ensure that rob^T falls within the range $[0, 1]$.

3.5.4. Effectiveness

The effectiveness metric (eff) quantifying the network's ability to make effective predictions was defined based on $Dice$ and Unc [27]:

$$eff(Unc) = 2\,Dice\,\frac{Unc}{Unc + 1}. \tag{11}$$

This definition implies that a network that produces exclusively unknown values is considered equally ineffective as a network that produces only incorrect predictions.

3.5.5. Augmented robustness

The augmented robustness rob_{Aug} was adopted in this study to assess robustness by considering both accuracy and uncertainty:

$$rob_{Aug}(Unc)^T = \frac{1}{2}\sum_{j=1}^{20}\left[tol\left(eff(Unc)^{T_j}, \beta\right) - dep\left(eff(Unc)^{T_j}, \beta\right)\right]\pi_{T_j} + \frac{1}{2}. \tag{12}$$

With respect to (8), it considers eff as the quality metric in the definitions of the functions tol and dep. β represents the minimum accepted effectiveness threshold, which in our application was defined following the same criterion adopted in [27]:

$$\beta = \frac{th(Dice)}{th(Unc)+2}\, th(Unc) \tag{13}$$

It quantifies the model's ability not only to provide accurate results, but also to provide reliable confidence estimates associated with them, when encountering altered inputs.

3.5.6. Comparison between models

Accuracy, uncertainty, calibration, error detection, and robustness metrics described in Sections 3.5.3–3.5.5 were computed for the proposed confidence-based ensemble strategies as well as for models trained using individual data augmentation schemes.

In addition, two benchmark approaches were evaluated for comparison:

- a single model trained with a data augmentation that incorporates all three transformations (*Aff + BF + Blur*), which represents a conventional strategy that aggregates all augmentations within a single training pipeline;
- a standard soft voting ensemble without confidence-based weighting (*Soft voting*), where each predicted probability map $p_\tau(i,j)$ contributed equally to the final segmentation, regardless of predictive confidence or segmented volume.

The corresponding uncertainty map was computed using the same procedure adopted for the proposed ensemble methods.

Statistical significance among models was assessed using the Friedman test for paired samples [48]. In the case of significant results, post hoc pairwise comparisons were then performed using two-tailed Wilcoxon signed-rank tests [49]. Comparisons between the confidence-based ensemble approaches, the best-performing single-data-augmentation model, and the two benchmarks (*Aff + BF + Blur* model and *Soft voting*) were directly presented.

To mitigate the dependence introduced by the 5 repeated acquisitions per subject, the values of all metrics were averaged across the repeated acquisitions within each subject. In addition, for metrics not directly related to robustness (namely, $Dice$, $\overline{Conf}$, RIU, UA, and the calibration indices) values were also averaged across all 20 alterations before statistical analysis. This yielded a single value per acquisition and avoided the dependence among observations. Consequently, the statistical analyses were performed at the subject level. For perturbation-specific robustness metrics, each subject contributed one aggregated value for each of the three MRI alteration types, resulting in 30 subject-by-alteration observations (10 subjects × 3 alteration types). For the post hoc analysis, all pairwise comparisons among the 11 considered models were evaluated, resulting in 55 pairwise tests. To account for multiple comparisons, p-values were adjusted using the Benjamini–Hochberg false discovery rate (FDR) correction. Finally, the effect sizes for pairwise Wilcoxon signed-rank tests were quantified using the matched-pairs rank-biserial correlation, with the sign indicating the direction of the effect.

4. Results

4.1. Computational cost

Information on the computational cost and execution time of each model during training and inference is reported in Table 3. All individual models shared the same architecture and therefore had identical model sizes and memory requirements. However, training time increased with the complexity of the adopted data-augmentation strategy, with combinations of multiple transformations requiring longer training times.

Because the ensemble approaches used all five models simultaneously, their computational cost and inference time corresponded to the cumulative contributions of the individual models. In particular, the inference time required for an entire MRI volume was 32.11 s.

4.2. Sensitivity analysis

Sensitivity analysis of the uncertainty-normalization strategy, presented in Supplementary Table 1, revealed a negligible impact on both accuracy- and confidence-based metrics for the CLw, CLs, and VLs ensemble approaches. Minor differences were observed for the VLw approach when no normalization was applied, resulting in a slight increase in $Dice$ and RIU and a slight decrease in UA.

The sensitivity analysis of the volume-ratio threshold showed that the *Blur* model was excluded from the ensemble procedure more often, whilst *Aff + Blur* was the one that mostly contributed to it (Supplementary Table 2). The choice of volume-ratio threshold influenced the final performance, with metric variations ranging from −1.9% to +1.4% for $Dice$ and from −13% to +46% for RIU, with distinct trends observed across the four ensemble strategies (Supplementary Tables 3–6). For the CLw approach, performance generally improved as the threshold increased, reaching optimal values for most metrics at a threshold of 0.8. In contrast, the highest correlation between $Dice$ and confidence was observed at lower thresholds (Supplementary Table 7). Importantly, the CLw strategy consistently outperformed the other ensemble approaches across the entire range of thresholds considered for most metrics.

4.3. Accuracy, uncertainty and calibration

The results of the accuracy, uncertainty, and calibration analyses are reported in Table 4, while the Spearman's rank correlation coefficients between $Dice$ and $\overline{Conf}$ are shown in Fig. 3 for each model.

Median $Dice$ exceeded 0.87 for all methods. The two confidence-weighted ensemble strategies achieved the highest $Dice$ values, followed by the *Aff + BF* model and the two benchmark approaches. The lowest segmentation performance was observed for models trained using a single augmentation strategy, namely *BiasField* and *Blur*.

The highest median $\overline{Conf}$ was obtained by the *VLs* model, as this ensemble strategy explicitly maximizes voxel-wise confidence during prediction aggregation.

The ability of uncertainty to discriminate between correct and incorrect predictions was assessed using the KL divergence between their respective uncertainty distributions. The *CLw* ensemble method

Table 3
Metrics quantifying the computational cost and execution time of each individual model.

Model	Number of parameters	Model size [MB]	GFLOPs	GPU memory [MB]	Training time per epoch [s]	Total training time [min]	Inference time per MRI volume [s]
Affine	$3.11 \cdot 10^7$	364.45	33.798	357.62	15	25	6.4
BiasField	$3.11 \cdot 10^7$	364.45	33.798	358.24	18	30	6.4
Blur	$3.11 \cdot 10^7$	364.45	33.798	357.99	18	30	6.4
Aff + BF	$3.11 \cdot 10^7$	364.45	33.798	356.74	26	65	6.4
Aff + Blur	$3.11 \cdot 10^7$	364.45	33.798	357.99	24	60	6.4
Aff + BF + Blur	$3.11 \cdot 10^7$	364.45	33.798	356.74	36	120	6.4

Table 4
Results of accuracy, uncertainty and calibration for each model. Delta mean/median represents the difference in uncertainty between correct and incorrect voxels. Values are reported as mean ± std in the case of a normal distribution, or median [interquartile range] in the case of a non-normal distribution. The best mean/median value for each metric is reported in bold.

Model	Dice	$\overline{Conf}$	KL Divergence	Delta mean	Delta median	Brier
Affine	0.889 [0.819–0.913]	0.873 [0.860–0.887]	0.351 ± 0.100	0.152 [0.135–0.164]	0.245 [0.221–0.263]	0.0059 [0.0048–0.0090]
BiasField	0.870 [0.826–0.894]	0.884 ± 0.017	0.219 [0.152–0.266]	0.119 [0.097–0.133]	0.196 [0.161–0.217]	0.0069 [0.0054–0.0095]
Blur	0.874 [0.812–0.899]	0.915 ± 0.012	0.154 [0.114–0.197]	0.098 [0.078–0.111]	0.166 [0.134–0.187]	0.0069 [0.0057–0.0092]
Aff + BF	0.906 [0.844–0.924]	0.895 [0.883–0.903]	0.369 [0.311–0.427]	0.150 ± 0.021	0.236 [0.222–0.250]	0.0051 [0.0042–0.0080]
Aff + Blur	0.898 [0.876–0.917]	0.864 [0.848–0.882]	0.389 [0.327–0.461]	0.159 ± 0.019	0.243 ± 0.021	0.0056 [0.0044–0.0072]
VLw	0.906 [0.887–0.922]	0.885 [0.864–0.895]	0.343 [0.258–0.435]	0.141 ± 0.031	0.221 ± 0.045	0.0050 [0.0041–0.0062]
CLw	**0.910 [0.891–0.925]**	0.846 [0.817–0.862]	**0.512 [0.412–0.621]**	**0.174 ± 0.029**	**0.250 [0.223–0.276]**	**0.0048 [0.0039–0.0059]**
CLs	0.884 [0.848–0.907]	0.911 [0.902–0.922]	0.181 [0.139–0.227]	0.106 ± 0.025	0.178 [0.154–0.201]	0.0066 [0.0053–0.0083]
VLs	0.891 [0.868–0.912]	**0.957 [0.951–0.960]**	0.140 [0.100–0.518]	0.068 [0.061–0.079]	0.101 [0.090–0.118]	0.0065 [0.0053–0.0079]
Aff + BF + Blur	0.902 [0.866–0.922]	0.871 [0.859–0.888]	0.405 [0.351–0.465]	0.162 ± 0.017	0.247 ± 0.022	0.0052 [0.0043–0.0068]
Soft Voting	0.903 [0.863–0.924]	0.844 [0.806–0.860]	0.501 [0.384–0.630]	0.172 [0.151–0.191]	0.246 [0.214–0.271]	0.0049 [0.0040–0.0065]

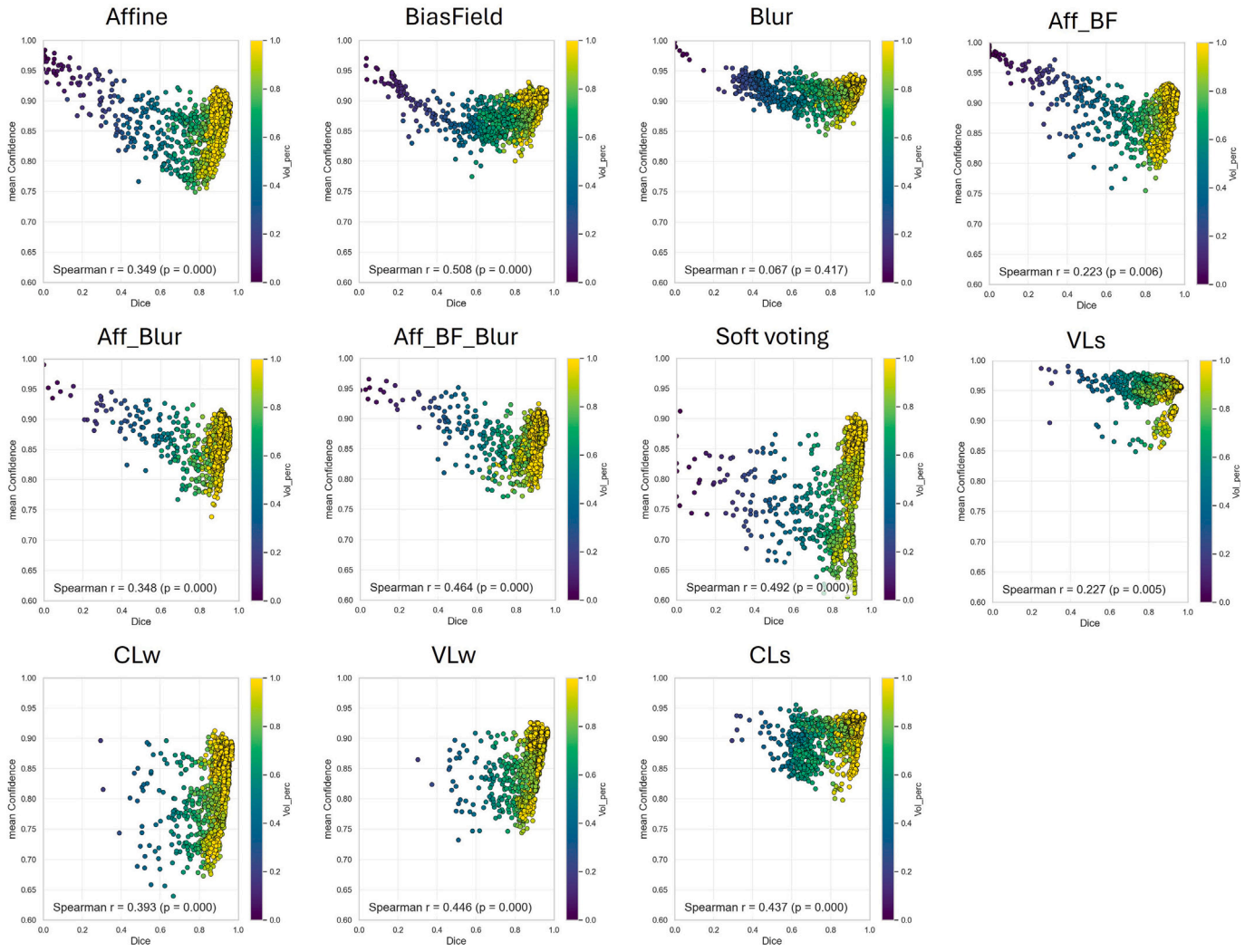

Fig. 3. Scatter plots of *Dice* versus $\overline{Conf}$ for each model. Points are colored according to the ratio between predicted and GT volumes, and Spearman's rank correlation coefficients with corresponding p-values are reported.

achieved the greatest separation between the uncertainty distributions of correct and incorrect voxels, followed by the *Soft voting* approach.

Model calibration, assessed using the *Brier* score, showed that *CLw* again achieved the lowest error, together with *Soft voting*.

Scatter plots of $Dice$ versus $\overline{Conf}$ showed that, among the evaluated approaches, only weighted ensemble-based methods (*CLw*, *VLw* and *Soft voting*) exhibited a monotonic relationship between segmentation accuracy and confidence. In contrast, the remaining models showed either parabolic or weak, near-flat associations, indicating a less consistent relationship between predictive confidence and segmentation accuracy. All models except *Blur* exhibited positive and statistically significant Spearman rank correlation coefficients (p-value < 0.01). However, these coefficients should be interpreted together with the full scatter distributions, as the positive correlations appear to be largely influenced by the dense cluster of observations at higher $Dice$ values. On the contrary, the scatter plots in their entirety also highlighted cases with low $Dice$ values, which may be clinically relevant despite having limited influence on the overall correlation coefficients. Thus, in our data, a significant positive Spearman correlation does not necessarily imply a clear or uniformly monotonic relationship across the entire range of observations.

Furthermore, inspection of the color coding in the scatter points, which represents the ratio between the predicted and GT volumes, revealed that the ensemble strategies tended to produce volume estimates closer to the reference values than the single-data-augmentation models and the two benchmark approaches.

4.4. Robustness

Robustness-related metrics computed on the corrupted test MRI data, considering the three transformations jointly, are reported in Table 5 for all segmentation models.

A detailed analysis of the robustness metric rob_{Aug} is presented in Fig. 4, which illustrates the distributions of the three accuracy–uncertainty metrics (RIU, UA, and $\overline{Conf}$) using violin plots, together with the results of the statistical significance tests comparing ensemble strategies, benchmark methods, and the best-performing single model.

The highest rob_{Aug} values were achieved by the *CLw* strategy for RIU and UA, and by the *VLs* strategy for $\overline{Conf}$. The best-performing ensemble strategy was never outperformed by the benchmark approaches or the strongest single-data-augmentation models, achieving significantly better performance in most cases.

Overall, these results indicate that the proposed ensemble framework provides increased robustness to MRI-specific distortions compared with a single model trained using comprehensive data augmentation. Moreover, confidence-based weighting consistently improved performance over the conventional soft voting strategy, highlighting the benefit of incorporating uncertainty information into the ensemble process. Finally, none of the single models trained with individual or combined data augmentation strategies achieved robust performance across all three transformations.

4.5. Analysis of individual MRI alterations

In this section, the robustness analysis is performed with respect to each specific MRI alteration, i.e., images altered with affine, bias field and blurring transformations.

4.5.1. Images altered with an affine transformation

When considering test images altered exclusively with affine transformations, the highest segmentation accuracy ($Dice = 0.912$) was achieved by the *Aff+BF* and *CLw* models (Supplementary Table 8). Overall, all models trained with augmentation strategies that included affine transformations achieved high accuracy, with $Dice$ values exceeding 0.8 for all MRI volumes. In terms of robustness, *CLw* and *Soft voting* achieved the highest $rob_{Aug}(RIU)$ ($\simeq 0.92$), whereas $rob_{Aug}(UA)$ reached the highest value of 0.938 for *Aff+BF*, followed by *VLw*, *CLw*, and *Soft voting*. The highest $rob_{Aug}(\overline{Conf})$ was obtained by *VLs* (0.924), followed by *Aff+BF*.

A comprehensive comparison is reported in Supplementary Fig. 1. The most significant differences can be observed for $robAug(RIU)$, where *CLw* significantly outperformed all the other models, especially for cases presenting lower robustness. On the contrary, $rob_{Aug}(UA)$ and $rob_{Aug}(Conf)$ were generally comparable or slightly but significantly different among the models.

Scatter plots of $Dice$ versus mean confidence (Supplementary Fig. 2) revealed a weak linear association when affine transformations were included in the training data augmentation. As a general observation, models trained using affine transformations as part of the data-augmentation strategy, including the ensemble approaches, consistently produced highly accurate segmentations. Consequently, segmentation errors were relatively rare, limiting the possibility of fully assessing their ability to identify erroneous predictions through low confidence

Table 5
Metrics for each training type and ensemble strategy, considering all transformed images aggregated. Values are reported as mean ± std in the case of a normal distribution, or median [interquartile range] in the case of a non-normal distribution. The highest median value for each metric is reported in bold.

Training type	$rob(Dice)$	$rob(RIU)$	$rob(UA)$	$rob(\overline{Conf})$	$rob_{Aug}(RIU)$	$rob_{Aug}(UA)$	$rob_{Aug}(\overline{Conf})$
Affine	0.817 [0.727–0.855]	0.743 [0.425–0.848]	0.864 [0.832–0.883]	0.842 [0.825–0.859]	0.858 [0.706–0.906]	0.922 [0.875–0.937]	0.898 [0.847–0.913]
BiasField	0.784 [0.710–0.823]	0.625 [0.363–0.758]	0.882 [0.834–0.895]	0.854 ± 0.021	0.813 [0.696–0.875]	0.910 [0.883–0.924]	0.892 [0.864–0.907]
Blur	0.793 [0.713–0.832]	0.527 [0.289–0.677]	0.906 [0.845–0.919]	0.894 ± 0.015	0.781 [0.664–0.853]	0.910 [0.875–0.930]	0.900 [0.867–0.917]
Aff + BF	0.843 [0.771–0.873]	0.740 [0.356–0.824]	0.879 [0.839–0.902]	0.868 [0.854–0.879]	0.858 [0.681–0.910]	0.932 [0.892–0.944]	0.913 [0.875–0.924]
Aff + Blur	0.831 [0.798–0.861]	0.816 [0.393–0.878]	0.865 [0.837–0.879]	0.830 [0.811–0.853]	0.892 [0.734–0.914]	0.927 [0.912–0.938]	0.900 [0.883–0.913]
VLw	0.844 [0.812–0.871]	0.891 [0.848–0.927]	0.878 [0.859–0.891]	0.861 [0.835–0.871]	0.919 [0.902–0.931]	0.932 [0.920–0.943]	0.912 [0.902–0.922]
CLw	**0.851 [0.818–0.876]**	0.929 [0.888–0.954]	0.849 [0.817–0.867]	0.808 [0.771–0.828]	**0.927 [0.914–0.941]**	**0.933 [0.921–0.942]**	0.898 [0.887–0.911]
CLs	0.808 [0.749–0.844]	0.572 [0.325–0.694]	0.905 [0.845–0.919]	0.889 [0.878–0.903]	0.800 [0.677–0.859]	0.920 [0.896–0.933]	0.905 [0.883–0.919]
VLs	0.819 [0.780–0.854]	0.435 [0.346–0.479]	**0.939 [0.927–0.946]**	**0.946 [0.939–0.951]**	0.758 [0.701–0.794]	0.926 [0.911–0.940]	**0.923 [0.908–0.936]**
Aff + BF + Blur	0.839 [0.787–0.870]	0.775 [0.404–0.872]	0.869 [0.841–0.885]	0.839 [0.823–0.860]	0.878 [0.737–0.914]	0.929 [0.905–0.942]	0.903 [0.882–0.918]
Soft voting	0.842 [0.786–0.874]	**0.932 [0.900–0.960]**	0.849 [0.817–0.867]	0.806 [0.757–0.825]	0.925 [0.899–0.941]	0.930 [0.903–0.942]	0.894 [0.865–0.911]

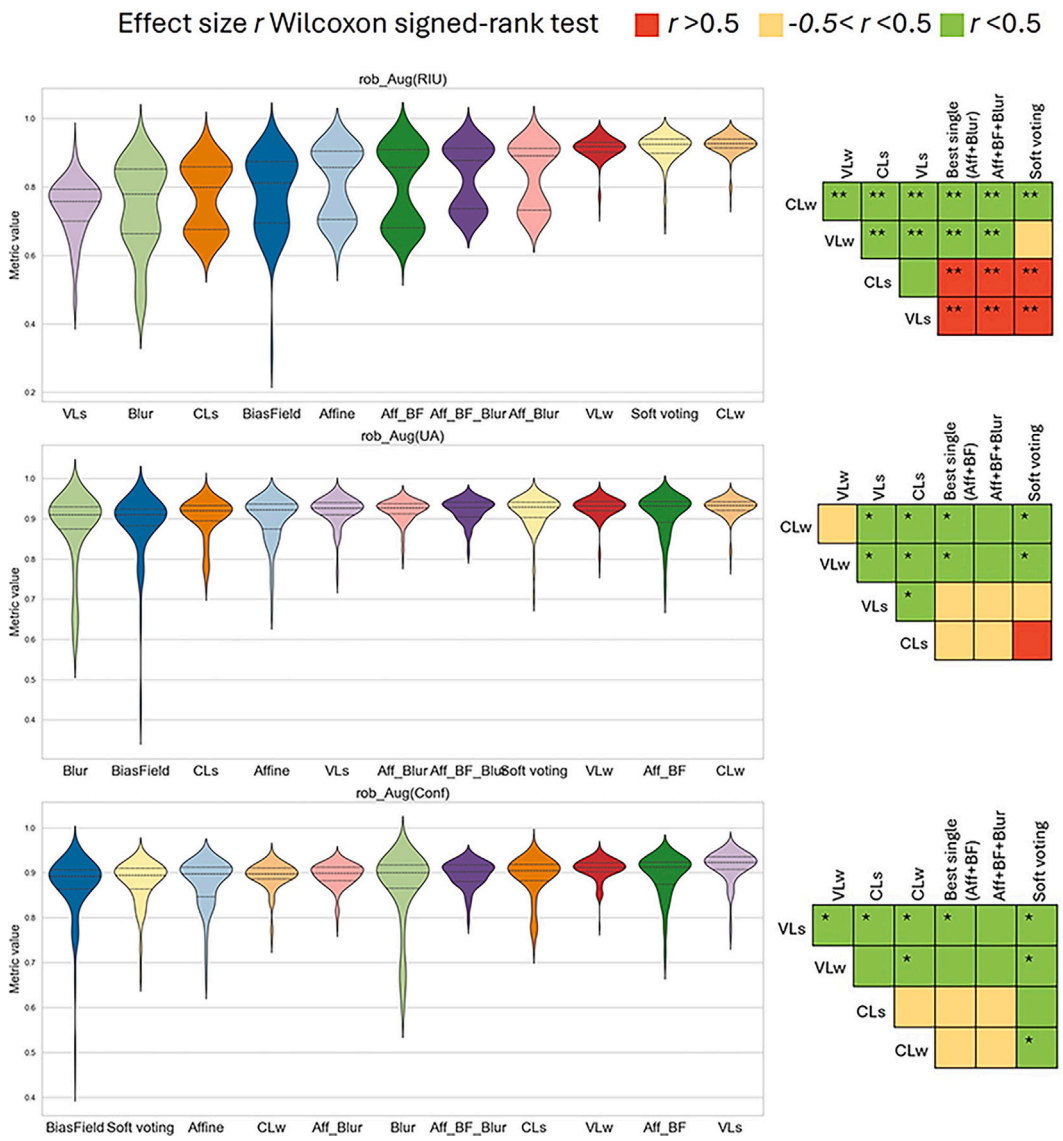

Fig. 4. Violin plots of the Augmented robustness metrics, with all images aggregated. The tables on the right report the effect size (matched-pairs rank-biserial correlation, r) of the post-hoc Wilcoxon signed-rank tests comparing each ensemble approach with the best-performing individual model and the two benchmark models. Colors indicate the direction and magnitude of the effect: green, higher values for the model in the row ($r > 0.5$); red, higher values for the model in the column ($r < -0.5$); yellow, ($-0.5 \leq R \leq 0.5$). Asterisks indicate statistically significant differences after multiple-comparison correction (* adjusted $p<0.05$; ** adjusted $p<0.005$).

values. In contrast, models trained using only bias-field or blur transformations as task-informed augmentations did not consistently achieve accurate segmentations. Importantly, these models often assigned high confidence to erroneous predictions, indicating a limited ability to identify segmentation failures and therefore making their predictions less reliable.

Fig. 5 presents an illustrative example of predictions obtained for a CKD patient with affine-induced distortion. Consistent with the aggregate results, the *BiasField* and *Blur* models failed to correctly segment the right kidney, whereas the remaining models produced satisfactory segmentations. However, for the left kidney, training with affine augmentation alone was insufficient to achieve accurate delineation of the inner region.

Analysis of the uncertainty maps derived from the ensemble strategies revealed qualitative differences not fully captured by quantitative metrics. The weighting-based ensemble approaches also generated maps with very high uncertainty values within correctly segmented kidney regions, an expression of the high variability between models; at the same time, these maps still clearly highlight areas corresponding to segmentation errors. In contrast, the voxel-level selection strategy based on the most confident prediction produced uncertainty maps that more sharply emphasized regions that were consistently uncertain across models.

4.5.2. Images altered with a bias field transformation

The *CLw* ensemble strategy achieved the best performance in terms of $Dice$, $rob_{Aug}(RIU)$, and $rob_{Aug}(UA)$ also when test images were corrupted with bias field distortion. In this setting, *CLw* significantly outperformed both benchmark approaches—*Aff + BF + Blur* and *Soft voting*—as well as all other models (Supplementary Table 9 and Supplementary Fig. 3). Consistent with previous findings, *VLs* achieved the highest performance for $rob_{Aug}(\overline{Conf})$.

In contrast to the affine-distortion scenario, only *CLw*, *VLw* and *Soft voting* exhibited a monotonic relationship between accuracy and confidence under bias field corruption (Supplementary Fig. 4). None of the single–data-augmentation models was able to correctly segment all volumes. Notably, only the uncertainty-informed ensemble strategies were

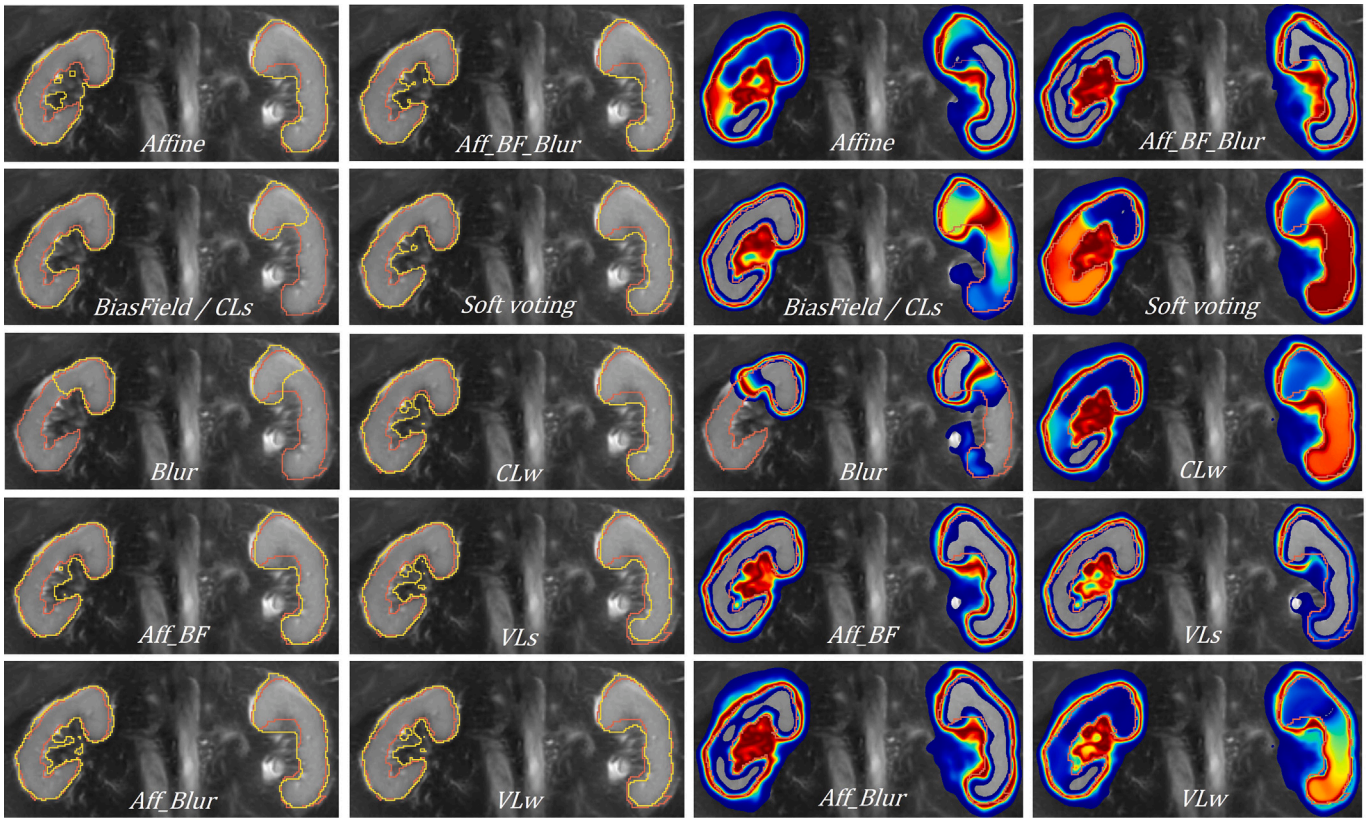

Fig. 5. An illustrative case of a CKD subject altered with an affine transformation. The first two columns: manual contour (red) and predicted contour (yellow) using different approaches are reported; the last two columns: uncertainty maps (from blue: low uncertainty, to red: high uncertainty) are estimated for each approach.

able to improve segmentation accuracy while preserving informative uncertainty estimates.

The illustrative example shown in Fig. 6 further emphasizes that segmentation errors were present in nearly all single–data-augmentation models. The ensemble approaches *CLw*, *VLw*, and *VLs* yielded performance comparable to the benchmark methods (*Aff + BF + Blur* and *Soft voting*). Interestingly, the most confident model selected by *CLs* corresponded to the *Blur* model, which was not the most accurate, as evidenced in the outer region of the right kidney. Uncertainty maps in the weighted ensemble approaches, including *Soft voting*, highlighted the large variability observed in the individual single models, especially in the delineation of the left kidney.

4.5.3. Images altered with a blur transformation

Results obtained on images corrupted with the blur transformation were consistent with those observed for the other distortions. The *CLw* strategy achieved one of the highest robustness values when computed using RIU and UA. However, no statistically significant differences were observed with respect to the *Soft voting* approach for RIU, nor with respect to *Soft voting*, *Aff + BF + Blur*, *Aff + BF*, and *VLw* for UA. As in the previous analyses, $rob_{Aug}(\overline{Conf})$ reached its maximum value with the *VLs* strategy (Supplementary Table 10 and Supplementary Fig. 5).

Scatter plots of $Dice$ versus mean confidence (Supplementary Fig. 6) demonstrated a linear association for all ensemble and benchmark strategies, whereas models trained exclusively with *BiasField* or *Blur* augmentation exhibited non-linear behavior, similar to what was observed for images altered with affine transformations.

The illustrative CKD case shown in Fig. 7 further confirmed that ensemble approaches achieved performance comparable to that of the benchmark models. Additionally, the weighting-based ensemble strategies produced more spatially distributed uncertainty maps, whereas the selection-based approach concentrated uncertainty primarily along kidney boundaries and in regions characterized by consistently low confidence.

5. Discussion and conclusions

In this study, we evaluated the impact of different data augmentation strategies on the robustness of a kidney segmentation model applied to MRI data affected by realistic distortions, jointly considering accuracy and uncertainty. We also introduced an ensemble framework that combines models trained with different augmentation schemes by weighting their contributions based on uncertainty estimates derived from the Monte Carlo dropout approach. Our results indicate that the case-level confidence-weighted ensemble strategy achieved the best overall performance, simultaneously optimizing segmentation accuracy ($Dice$) and uncertainty-related metrics (RIU and UA).

Our findings indicate that the proposed uncertainty-aware ensemble strategy is particularly effective in optimizing uncertainty-related metrics. Specifically, this approach preserves high model calibration, thanks to the lowest Brier score, and better distinguishes between correctly and incorrectly segmented regions, as expressed by the high KL divergence for the *CLw* method: it assigns higher uncertainty to voxels affected by segmentation errors (thereby improving RIU) while simultaneously providing higher confidence for voxels that are segmented correctly, as reflected by increased UA. This dual capability makes the uncertainty-aware ensemble more reliable than the benchmark *Aff + BF + Blur* model, which, despite being trained on all augmentations simultaneously, struggles to manage specific challenging cases. In contrast, the ensemble's use of uncertainty information allows it to modulate the contribution of each individual model and ultimately achieve more consistent performance across heterogeneous input conditions.

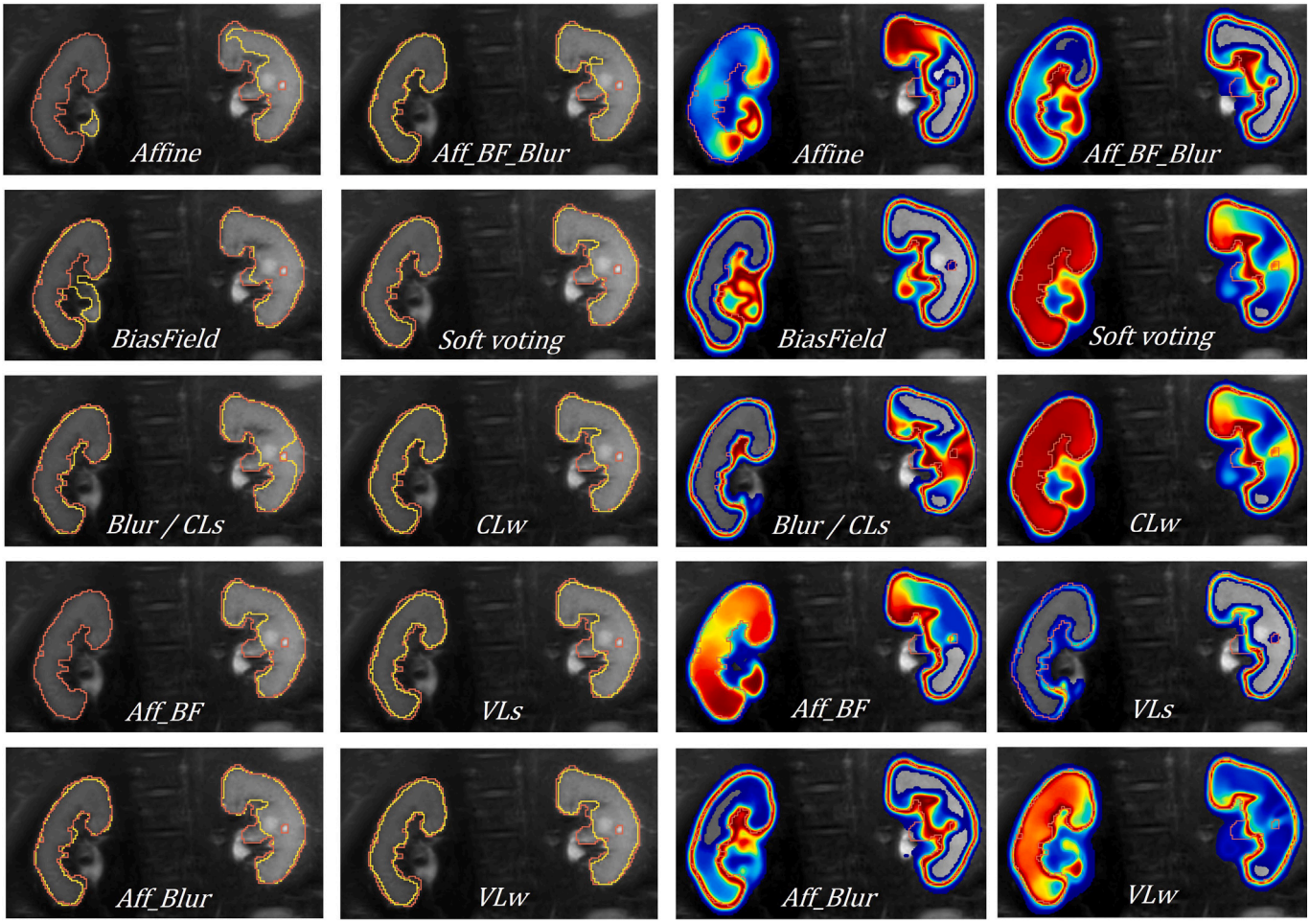

Fig. 6. An illustrative case of a CKD subject altered by bias field transformation. First two columns: manual contour (red) and predicted contour (yellow) using different approaches are reported; last two columns: uncertainty maps (from blue: low uncertainty, to red: high uncertainty) estimated for each approach.

When robustness was evaluated through mean confidence, the voxel-level selection (*VLs*) strategy achieved the best performance. This result is expected, as *VLs* directly selects, voxel by voxel, the prediction associated with the highest confidence among all models. Importantly, this metric does not rely on the GT segmentation and can therefore serve as an independent robustness indicator in real-world clinical applications, where reference labels may be unavailable. Nevertheless, our analysis also confirms that high confidence does not necessarily correspond to high accuracy. This limitation is evident in the association plot relating *Dice* to confidence, which shows a relatively flat trend: segmentations maintain consistently high confidence despite exhibiting a broad spectrum of *Dice* values. This observation reinforces the need for caution when relying solely on confidence-based measures for automated quality assessment.

The comparison with standard *Soft voting* ensembles further highlights the advantages of our method. Soft voting tends to perform similarly to the case-level weighted (*CLw*) strategy when the models being aggregated produce largely concordant predictions, as demonstrated by the non-significant difference between the two approaches for most metrics. However, in instances where one or more constituent models generate incorrect or unstable segmentations, the uncertainty-aware filtering mechanism, implemented through the volume-ratio selection, substantially improves robustness over soft voting. By suppressing unreliable model outputs before aggregation, the ensemble mitigates the negative influence of outlier predictions and maintains higher overall accuracy. This can be appreciated in Fig. 5, where the removal of the *Blur* model from the uncertainty-weighted procedure enhanced the confidence of the *CLw* approach over *Soft voting*.

Our results are consistent with the findings of [35], who showed that an uncertainty-aware ensemble can outperform traditional aggregation strategies such as soft voting and majority voting. Similar to our approach, their method relies on dynamically selecting among multiple baseline models based on estimated uncertainty, thereby adapting the aggregation process to the reliability of individual predictions. However, an important difference emerges when comparing their uncertainty–accuracy relationships with ours. In their study, the models showed a predominantly linear association between uncertainty and *Dice*, enabling straightforward filtering based directly on uncertainty values. In contrast, as illustrated in Fig. 3, each of our baseline models exhibited a parabolic relationship between confidence and accuracy. This nonlinear pattern implies that confidence alone is not a reliable discriminator between good and poor segmentations, as both low-accuracy and high-accuracy predictions may present similarly high confidence values. Consequently, we designed our filtering strategy to rely on the volume ratio of predicted segmentations rather than on confidence itself. This choice provides a more robust mechanism for identifying unreliable model outputs within our dataset. In practical applications, where the GT segmentation is typically unavailable, threshold optimization based on accuracy metrics would not be feasible in a real-world setting. For this reason, we propose a threshold value of 0.5, as it represents a reasonable compromise between segmentation performance and confidence calibration, while also maintaining a moderate correlation between *Dice* and

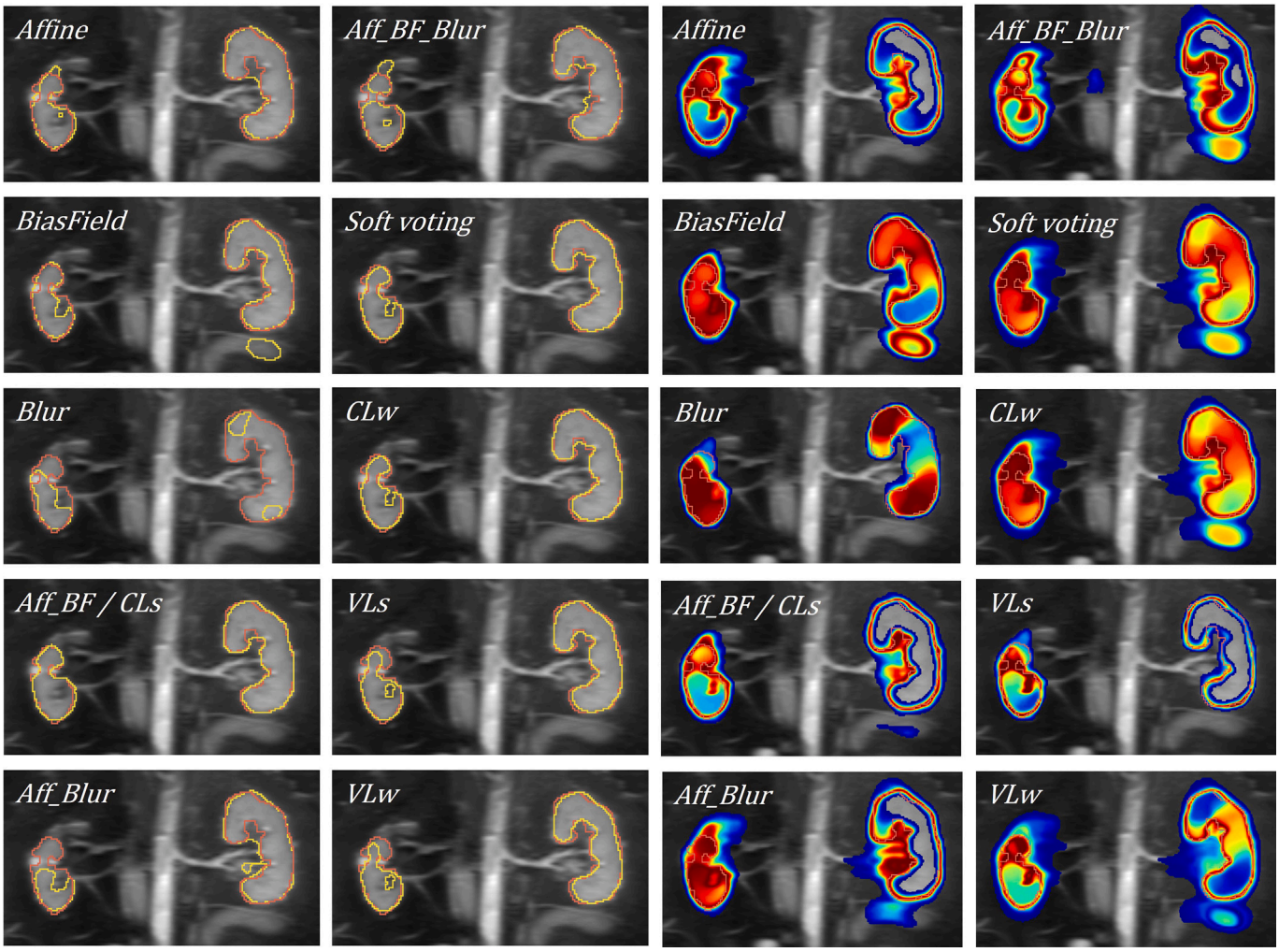

Fig. 7. An illustrative case of a CKD subject altered with a blur transformation. First two columns: manual contour (red) and predicted contour (yellow) using the different approaches are reported; Last two columns: uncertainty maps (from blue: low uncertainty, to red: high uncertainty) estimated for each approach.

confidence. Its generalizability should be tested on different anatomical structures, imaging modalities, or datasets.

The effect of the volume-ratio threshold can be appreciated in the illustrative example presented in Supplementary Fig. 7, where an example of a failure case is shown. Here, the individual models generally performed poorly, particularly for the left kidney, although the magnitude of the errors differed across models. Among the individual models, the *Aff + BF* model achieved the best segmentation performance, whereas the *Aff* model produced the poorest result. The *Blur* model performed slightly better than the *Aff* model and yielded a volume ratio of 51%, thereby meeting the volume-threshold criterion and being retained in the ensemble procedure. Furthermore, the *Blur* model exhibited the highest confidence score and therefore received a substantial weight in the confidence-weighted aggregation process. Nevertheless, the uncertainty maps generated by the CLw and VLw ensemble approaches successfully identified the regions of disagreement among the individual models as highly uncertain. Therefore, despite the segmentation failure, the associated uncertainty estimate was able to flag the problematic region and provide a meaningful warning to the user regarding the reliability of the prediction.

The comparison with single-augmentation models further supports the advantage of diversified augmentation strategies. Training with only one specific alteration consistently resulted in poorer performance. These models achieved acceptable robustness only when tested on images altered with the same transformation used during training. An exception was observed for blur augmentation, where even the blur-trained model did not perform well under blur distortions. Notably, affine augmentation emerged as a necessary component for achieving high performance across nearly all conditions, including bias field and blur distortions. This finding aligns with expectations, as geometric transformations frequently occur in clinical practice due to patient positioning and scanner variability, and is consistent with prior literature emphasizing the importance of geometric robustness [50].

However, incorporating all transformations simultaneously in the training process, while seemingly comprehensive, did not guarantee optimal performance. As shown in Fig. 3, the *Aff + BF + Blur* model produced several poor predictions, including cases with $Dice$ values below 0.2. This model also required the longest training time, making it less practical from a computational standpoint. Additionally, more complex augmentation strategies may require larger datasets and more extensive architectures to fully exploit their benefits. Otherwise, this data diversity could lead to underperforming models that may tend to overfit. In contrast, the proposed ensemble framework provides a modular strategy for incorporating additional types of MRI alterations. Specifically, a new alteration can be addressed by adding dedicated individual models, for example, one trained using the corresponding transformation

alone and another combining it with affine augmentation. This modular design results in an approximately linear increase in the overall training cost as new alteration-specific models are introduced, whereas simultaneously incorporating multiple alterations into a single training strategy results in a larger increase in training time. Moreover, since the individual models are independent, their predictions can be computed in parallel during inference, potentially reducing the ensemble inference time to approximately that of a single model, provided that adequate computational resources are available. The additional memory required to load the individual models was limited to a few hundred megabytes, representing a manageable overhead on commonly available computational hardware. These results suggest that simply maximizing augmentation diversity is not sufficient; instead, selectively exploiting uncertainty within a tailored ensemble framework is a more effective, scalable, and modular strategy, albeit at increased computational cost, for improving both accuracy and uncertainty-aware performance.

5.1. Limitations and future developments

This study presents several limitations that should be considered when interpreting the results. First, the dataset used for training and evaluation was relatively small, homogeneous and acquired in a single center using the same MRI scanner, and involved the segmentation of a single organ with a well-defined morphology. The kidney, although clinically relevant, is a comparatively simple structure to segment; therefore, the generalizability of the proposed framework to more complex anatomical regions remains to be demonstrated. Additionally, the findings of this work cannot be directly extrapolated to other imaging modalities, which would require the use of modality-specific perturbations (e.g., CT-related artifacts in computed tomography). The evaluation of the proposed methodology in these settings is beyond the scope of the present study and will be the subject of future research.

Second, we evaluated only a limited set of image alterations. This choice was intentional, aiming to maintain results that are easy to interpret and to isolate the impact of specific perturbations on model performance. This design facilitated the interpretation of the observed effects and allowed the contribution of each perturbation type to be evaluated independently. However, we acknowledge that real-world MRI examinations may be affected by multiple simultaneous artifacts and image degradations. Therefore, future studies should investigate more complex scenarios involving combinations of perturbations to further assess the generalizability of the proposed framework under realistic clinical conditions.

Third, uncertainty estimation was performed using MCD, which, although widely used, computationally efficient, and able to provide a reasonable approximation of epistemic uncertainty, is not always the most robust or informative method across tasks and architectures. However, in our previous study [10], we systematically compared several uncertainty-quantification techniques and loss functions, finding that the combination adopted here offered the best balance between accuracy and uncertainty-related performance for this application. Nonetheless, future works may assess different uncertainty quantification strategies, especially for the aggregation of individual models, in order to differentiate between disagreement among models and MCD epistemic uncertainty within individual models. Additionally, the ensemble strategy was compared only with two benchmark approaches, namely *Soft voting* and *Aff+BF+Blur*, but alternative and more sophisticated ensemble techniques were not considered.

Finally, the proposed strategies should be validated on real clinical acquisitions, where heterogeneous distortions, scanner-specific variability, and patient-related factors may influence segmentation quality and uncertainty patterns. Testing the framework under these real-world conditions is essential for confirming its practical utility and for refining the filtering and ensemble mechanisms accordingly. This validation step is planned for future work.

CRediT authorship contribution statement

Elisa Scalco: Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Rossella Damiano:** Writing – review & editing, Software, Formal analysis. **Anna Caroli:** Writing – review & editing, Data curation. **Marco L. Della Vedova:** Writing – review & editing, Software, Resources, Methodology, Investigation. **Andrea Bombarda:** Writing – review & editing, Methodology, Investigation, Conceptualization. **Ettore Lanzarone:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Conceptualization.

Ethics approval

This study used publicly available data; thus, no ethics approval was needed.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was partially supported by the Italian Ministry of University and Research, grant protocol number 2022B23JT5, PRIN (Progetti di Rilevante Interesse Nazionale) 2022, funded by the European Union—Next Generation EU (PNRR M4.C2.1.1). A.B., A.C. and E.L. were also supported by the National Plan for NRRP Complementary Investments (PNC, established with the Decree-Law 6 May 2021, n. 59, converted by law n. 101 of 2021) under the call for the funding of research initiatives for technologies and innovative trajectories in the health and care sectors (Directorial Decree n. 931 of 06-06-2022) - project n. PNC0000003 - AdvaNced Technologies for Human-centrEd Medicine (project acronym: ANTHEM). This work reflects only the authors' views and opinions, neither the Ministry of University and Research nor the European Commission can be considered responsible for them.

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi: 10.1016/j.asoc.2026.116818.

Data availability

The MRI dataset used in this study is publicly available at https://doi.org/10.5281/zenodo.5153568.

The source code for the experimental implementation used in this study is publicly available at https://doi.org/10.5281/zenodo.23986081.

References

[1] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A. Van Der Laak, B. Van Ginneken, C.I. Sánchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88, https://doi.org/10.1016/j.media.2017.07.005

[2] Q. Xia, H. Zheng, H. Zou, D. Luo, H. Tang, L. Li, B. Jiang, A comprehensive review of deep learning for medical image segmentation, Neurocomputing 613 (2025) 128740, https://doi.org/10.1016/j.neucom.2024.128740

[3] J. Ma, Y. He, F. Li, L. Han, C. You, B. Wang, Segment anything in medical images, Nat. Commun. 15 (1) (Jan 2024), https://doi.org/10.1038/s41467-024-44824-z

[4] M.E. Rayed, S.S. Islam, S.I. Niha, J.R. Jim, M.M. Kabir, M.F. Mridha, Deep learning for medical image segmentation: state-of-the-art advancements and challenges, Inform. Med. Unlocked 47 (2024) 101504, https://doi.org/10.1016/j.imu.2024.101504

[5] B. Lambert, F. Forbes, S. Doyle, H. Dehaene, M. Dojat, Trustworthy clinical AI solutions: a unified review of uncertainty quantification in deep learning models for medical image analysis, Artif. Intell. Med. 150 (2024) 102830, https://doi.org/10.1016/j.artmed.2024.102830

[6] P. Arcaini, A. Bombarda, S. Bonfanti, A. Gargantini, Dealing with robustness of convolutional neural networks for image classification, in: 2020 IEEE International Conference on Artificial Intelligence Testing, IEEE, 2020, pp. 7–14, https://doi.org/10.1109/AITEST49225.2020.00009

[7] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, J. Snoek, Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift, Adv. Neural Inf. Process. Syst. 32 (2019).

[8] H. Javed, S. El-Sappagh, T. Abuhmed, Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust AI applications, Artif. Intell. Rev. 58 (1) (2024) 12, https://doi.org/10.1007/s10462-024-11005-9

[9] Y. Lai, A. Kankanhalli, D. Ong, Human-AI collaboration in healthcare: a review and research agenda, in: Proceedings of the 54th Hawaii International Conference on System Sciences, 2021, pp. 390–399.

[10] R. Damiano, E. Scalco, M.L. Della Vedova, A. Arrigoni, A. Caroli, A. Bombarda, E. Lanzarone, Integrating uncertainty into U-Net robustness evaluation under natural MRI alterations: application to kidney segmentation, in: International Conference on Artificial Intelligence in Medicine, Springer, 2025, pp. 121–126, https://doi.org/10.1007/978-3-031-95841-0_23

[11] C. Shorten, T.M. Khoshgoftaar, A survey on image data augmentation for deep learning, J. Big Data 6 (1) (2019) 1–48, https://doi.org/10.1186/s40537-019-0197-0

[12] F. Isensee, P.F. Jaeger, S.A. Kohl, J. Petersen, K.H. Maier-Hein, nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, Nat. Methods 18 (2) (2021) 203–211.

[13] S. Francis, C. Buchanan, M. Craig, M. Gilthorpe, P.A. Kalra, I. Mendichovszky, S. Sourbron, M. Taal, N. Selby, # 1885 magnetic resonance imaging assessment of kidney morphometry in chronic kidney disease: baseline data from the afirm study, Nephrol. Dial. Transplant. 40 (Supplement_3) (2025) gfaf116.0447, https://doi.org/10.1093/ndt/gfaf116.0447

[14] P. Chlap, H. Min, N. Vandenberg, J. Dowling, L. Holloway, A. Haworth, A review of medical image data augmentation techniques for deep learning applications, J. Med. Imaging Radiat. Oncol. 65 (5) (2021) 545–563, https://doi.org/10.1111/1754-9485.13261

[15] C. Chen, C. Qin, H. Qiu, C. Ouyang, S. Wang, L. Chen, G. Tarroni, W. Bai, D. Rueckert, Realistic adversarial data augmentation for MR image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2020, pp. 667–677, https://doi.org/10.1007/978-3-030-59710-8_65

[16] C. Chen, C. Qin, C. Ouyang, Z. Li, S. Wang, H. Qiu, L. Chen, G. Tarroni, W. Bai, D. Rueckert, Enhancing MR image segmentation with realistic adversarial data augmentation, Med. Image Anal. 82 (2022) 102597, https://doi.org/10.1016/j.media.2022.102597

[17] M.I. Meyer, E. de la Rosa, N. Pedrosa de Barros, R. Paolella, K. Van Leemput, D.M. Sima, A contrast augmentation approach to improve multi-scanner generalization in MRI, Front. Neurosci. 15 (2021) 708196, https://doi.org/10.3389/fnins.2021.708196

[18] P. Vaish, F. Meister, T. Heimann, C. Brune, J.M. Wolterink, Data-agnostic augmentations for unknown variations: out-of-distribution generalisation in MRI segmentation, arXiv preprint, 2025, https://doi.org/10.48550/arXiv.2505.10223

[19] E. Panfilov, A. Tiulpin, S. Klein, M.T. Nieminen, S. Saarakkala, Improving robustness of deep learning based knee MRI segmentation: mixup and adversarial domain adaptation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019.

[20] C. Liu, Y. Dong, W. Xiang, X. Yang, H. Su, J. Zhu, Y. Chen, Y. He, H. Xue, S. Zheng, A comprehensive study on robustness of image classification models: benchmarking and rethinking, Int. J. Comput. Vis. 133 (2) (2024) 567–589, https://doi.org/10.1007/s11263-024-02196-3

[21] P. Arcaini, A. Bombarda, S. Bonfanti, A. Gargantini, ROBY: a tool for robustness analysis of neural network classifiers, in: 2021 14th IEEE Conference on Software Testing, Verification and Validation (ICST), 2021, pp. 442–447, https://doi.org/10.1109/ICST49551.2021.00057

[22] L. Halmosi, B. Mohos, M. Jelasity, Evaluating the adversarial robustness of semantic segmentation: trying harder pays off, in: European Conference on Computer Vision, Springer, 2024, pp. 1–18, https://doi.org/10.1007/978-3-031-72986-7_1

[23] P. Arcaini, A. Bombarda, S. Bonfanti, A. Gargantini, D. Gamba, R. Pedercini, Robustness assessment and improvement of a neural network for blood oxygen pressure estimation, in: 2022 IEEE Conference on Software Testing, Verification and Validation (ICST), IEEE, 2022, pp. 312–322, https://doi.org/10.1109/ICST53961.2022.00040

[24] M. Yeung, L. Rundo, Y. Nan, E. Sala, C.-B. Schönlieb, G. Yang, Calibrating the dice loss to handle neural network overconfidence for biomedical image segmentation, J. Digit. Imaging 36 (2) (2023) 739–752, https://doi.org/10.1007/s10278-022-00735-3

[25] B. Dai, C. Li, RankSEG: a consistent ranking-based framework for segmentation, J. Mach. Learn. Res. 24 (224) (2023) 1–50, http://jmlr.org/papers/v24/22-0712.html.

[26] K. Li, L. Yu, P.-A. Heng, Towards reliable cardiac image segmentation: assessing image-level and pixel-level segmentation quality via self-reflective references, Med. Image Anal. 78 (2022) 102426, https://doi.org/10.1016/j.media.2022.102426

[27] W. Essbai, A. Bombarda, S. Bonfanti, A. Gargantini, A framework for including uncertainty in robustness evaluation of Bayesian neural network classifiers, in: Proceedings of the 5th IEEE/ACM International Workshop on Deep Learning for Testing and Testing for Deep Learning, 2024, pp. 25–32, https://doi.org/10.1145/3643786.3648026

[28] A. Mehrtash, W.M. Wells, C.M. Tempany, P. Abolmaesumi, T. Kapur, Confidence calibration and predictive uncertainty estimation for deep medical image segmentation, IEEE Trans. Med. Imaging 39 (12) (2020) 3868–3878, https://doi.org/10.1109/TMI.2020.3006437

[29] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U.R. Acharya, et al., A review of uncertainty quantification in deep learning: techniques, applications and challenges, Inf. Fusion 76 (2021) 243–297, https://doi.org/10.1016/j.inffus.2021.05.008

[30] T.G. Dietterich, Ensemble methods in machine learning, in: Multiple Classifier Systems, Springer Berlin Heidelberg, Berlin, Heidelberg, 2000, pp. 1–15, https://doi.org/10.1007/3-540-45014-9_1

[31] N. Rane, S.P. Choudhary, J. Rane, Ensemble deep learning and machine learning: applications, opportunities, challenges, and future directions, Stud. Med. Health Sci. 1 (2) (2024) 18–41, https://doi.org/10.48185/smhs.v1i2.1225

[32] M.A. Ganaie, M. Hu, A.K. Malik, M. Tanveer, P.N. Suganthan, Ensemble deep learning: a review, Eng. Appl. Artif. Intell. 115 (2022) 105151, https://doi.org/10.1016/j.engappai.2022.105151

[33] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc., 2017, https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.

[34] K. Wang, B. Zhan, C. Zu, X. Wu, J. Zhou, L. Zhou, Y. Wang, Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning, Med. Image Anal. 79 (2022) 102447, https://doi.org/10.1016/j.media.2022.102447, https://www.sciencedirect.com/science/article/pii/S1361841522000925.

[35] M. Salvi, A. Mogetta, U. Raghavendra, A. Gudigar, U.R. Acharya, F. Molinari, A dynamic uncertainty-aware ensemble model: application to lung cancer segmentation in digital pathology, Appl. Soft Comput. 165 (2024) 112081, https://doi.org/10.1016/j.asoc.2024.112081

[36] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks, Neurocomputing 338 (C) (2019) 34–45, https://doi.org/10.1016/j.neucom.2019.01.103

[37] Y. Geifman, R. El-Yaniv, Selective classification for deep neural networks, arXiv:1705.08500, 2017.

[38] Y. Geifman, R. El-Yaniv, SelectiveNet: a deep neural network with an integrated reject option, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, Vol. 97, Proceedings of Machine Learning Research, PMLR, 2019, pp. 2151–2159.

[39] A.J. Daniel, C.E. Buchanan, T. Allcock, D. Scerri, E.F. Cox, B.L. Prestwich, S.T. Francis, T2-Weighted Kidney MRI Segmentation, 2021, https://doi.org/10.5281/zenodo.5153568, Version 1.0.0.

[40] A.J. Daniel, C.E. Buchanan, T. Allcock, D. Scerri, E.F. Cox, B.L. Prestwich, S.T. Francis, Automated renal segmentation in healthy and chronic kidney disease subjects using a convolutional neural network, Magn. Reson. Med. 86 (2) (2021) 1125–1136, https://doi.org/10.1002/mrm.28768

[41] R. Damiano, E. Lanzarone, F. Lussana, G. Villa, A.J. Daniel, S. Francis, A. Caroli, E. Scalco, The impact of uncertainty estimation on radiomic segmentation reproducibility and scan–rescan repeatability in kidney MRI, Med. Phys. 52 (7) (2025) e17995, https://doi.org/10.1002/mp.17995

[42] F. Pérez-García, R. Sparks, S. Ourselin, TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning, Comput. Methods Programs Biomed. 208 (2021) 106236, https://doi.org/10.1016/j.cmpb.2021.106236

[43] Y. Gal, Z. Ghahramani, Dropout as a Bayesian approximation: representing model uncertainty in deep learning, in: Proceedings of the 33rd International Conference on Machine Learning, PMLR, 2016, pp. 1050–1059, https://proceedings.mlr.press/v48/gal16.html.

[44] J. Gawlikowski, C.R.N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, et al., A survey of uncertainty in deep neural networks, Artif. Intell. Rev. 56 (Suppl 1) (2023) 1513–1589.

[45] A. Mobiny, P. Yuan, S.K. Moulik, N. Garg, C.C. Wu, H. Van Nguyen, Dropconnect is effective in modeling uncertainty of Bayesian deep networks, Sci. Rep. 11 (1) (2021) 1–14, https://doi.org/10.1038/s41598-021-84854-x

[46] E. Scalco, S. Pozzi, G. Rizzo, E. Lanzarone, Uncertainty quantification in multi-class segmentation: comparison between Bayesian and non-Bayesian approaches in a clinical perspective, Med. Phys. 51 (9) (2024) 6090–6102, https://doi.org/10.1002/mp.17189

[47] W.B. Glenn, et al., Verification of forecasts expressed in terms of probability, Mon. Weather Rev. 78 (1) (1950) 1–3.

[48] T.J. Cleophas, A.H. Zwinderman, Non-parametric tests for three or more samples (friedman and Kruskal-Wallis), in: Clinical Data Analysis on a Pocket Calculator: Understanding the Scientific Methods of Statistical Reasoning and Hypothesis Testing, Springer, 2016, pp. 193–197, https://doi.org/10.1007/978-3-319-27104-0_34

[49] R.F. Woolson, Wilcoxon signed-rank test, in: Wiley Encyclopedia of Clinical Trials, John Wiley & Sons, Ltd., 2008, pp. 1–3, https://doi.org/10.1002/9780471462422.eoct979

[50] E. Goceri, Medical image data augmentation: techniques, comparisons and interpretations, Artif. Intell. Rev. 56 (11) (2023) 12561–12605, https://doi.org/10.1007/s10462-023-10453-z