Trustworthy clinical AI solutions: A unified review of uncertainty quantification in Deep Learning models for medical image analysis

Benjamin Lambert, Florence Forbes, Senan Doyle, Harmonie Dehaene, Michel Dojat

Artificial Intelligence In Medicine 150 (2024) 102830

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

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

Trustworthy clinical AI solutions: A unified review of uncertainty quantification in Deep Learning models for medical image analysis

Benjamin Lambert [a,c], Florence Forbes [b], Senan Doyle [c], Harmonie Dehaene [c], Michel Dojat [a,*]

[a] Univ. Grenoble Alpes, Inserm, U1216, Grenoble Institut des Neurosciences, Grenoble, 38000, France
[b] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, Grenoble, 38000, France
[c] Pixyl Research and Development Laboratory, Grenoble, 38000, France

ARTICLE INFO

Keywords:
Medical imaging
Machine learning
Classification
Segmentation
Quality control
Artificial intelligence

ABSTRACT

The full acceptance of Deep Learning (DL) models in the clinical field is rather low with respect to the quantity of high-performing solutions reported in the literature. End users are particularly reluctant to rely on the opaque predictions of DL models. Uncertainty quantification methods have been proposed in the literature as a potential solution, to reduce the black-box effect of DL models and increase the interpretability and the acceptability of the result by the final user. In this review, we propose an overview of the existing methods to quantify uncertainty associated with DL predictions. We focus on applications to medical image analysis, which present specific challenges due to the high dimensionality of images and their variable quality, as well as constraints associated with real-world clinical routine. Moreover, we discuss the concept of structural uncertainty, a corpus of methods to facilitate the alignment of segmentation uncertainty estimates with clinical attention. We then discuss the evaluation protocols to validate the relevance of uncertainty estimates. Finally, we highlight the open challenges for uncertainty quantification in the medical field.

1. Introduction

In recent years, many Deep Learning (DL) medical applications have been proposed for the automatic analysis of various imaging modalities, including Magnetic Resonance Imaging (MRI), Computed Tomography (CT), Ultrasound (US) or histopathological images (see [1] for a review). To be accepted and routinely used by clinicians, however, these algorithms must provide robust and trustworthy predictions. This is of particular importance in the context of clinical applications, where the automated prediction may have a direct impact on patient care. However, Neural Network (NN) models are often considered and used as black-boxes, due to the absence of clear decision rules, as well as the lack of reliable confidence estimates associated with their predictions [2]. Additionally, DL models prove to be overconfident about their predictions on outlier data [3], and very sensitive to adversarial attacks [4], which suggests a global lack of robustness of this type of model. Due to these limitations, detecting failures or inconsistencies produced by DL models is complex, raising concerns as to the reliability and safety of these algorithms in clinical routine use [5]. To tackle this important issue, Uncertainty Quantification (UQ) methods [6] have been developed to quantify the predictive uncertainty of a given DL model. Enhancing an automated prediction with an estimation of its confidence has numerous benefits. First, it allows the identification of uncertain samples that need human reviewing. In a medical setting, this is particularly crucial to prevent silent errors that may lead to inaccurate diagnosis or treatment. Second, it enables the identification of the model's pitfalls. For example, uncertain predictions can indicate an incomplete training dataset. It provides insight into the knowledge captured by the model and can be used to extend the training set with supplementary data if needed. High uncertainty can also reveal anomalies within the input data, which is critical for Quality Control (QC). Overall, UQ increases trust in the algorithm and facilitates the interaction between the algorithm and the user. Moreover, UQ benefits from strong theoretical foundations and has emerged, from the clinical point of view, as one of the expected properties of any deployed AI algorithm [7]. As a result, the medical imaging community is becoming increasingly interested in incorporating UQ into image processing pipelines in order to highlight model failures or weaknesses. In this work, we propose a comprehensive overview of such an UQ integration in medical image processing pipelines.

1.1. Research outline

Several review articles focusing on uncertainty in DL can be found in the literature. In Abdar et al. [6], authors propose a complete review of

* Corresponding author.
E-mail address: Michel.Dojat@inserm.fr (M. Dojat).

https://doi.org/10.1016/j.artmed.2024.102830
Received 21 June 2023; Received in revised form 28 February 2024; Accepted 1 March 2024
Available online 4 March 2024

UQ methods, as well as their various concrete applications. Hullermeier et al. [8] focus their article on the definition of the two main categories of uncertainty, namely aleatoric and epistemic uncertainties, in the context of machine learning applications. In Gawlikowski et al. [9], insights about the various sources of uncertainty are presented. Reviews focusing on Bayesian DL [10,11] and prediction intervals [12] have been also published. More recently, Zhou et al. [13] presented a review of the latest advances considering epistemic uncertainty quantification in DL from the perspective of generalization error. These different works propose an overview of UQ methods in DL from a general point of view, but they failed to consider the specific aspects of UQ for medical image processing applications, a domain where the correct identification of the confidence of the model is crucial. Recently, Kurz et al. [14] and Loftus et al. [15] presented reviews in this direction, using a corpus of 22 and 30 papers, respectively. Here, we propose to extend the latter by presenting a complete review of 218 peer-reviewed papers implementing UQ applications in supervised DL-based pipelines, for both medical image classification and segmentation. We also explore the emerging concept of structural segmentation uncertainty, allowing the conversion of pixel uncertainty to instance uncertainty (*e.g.* lesion or organ) and case uncertainty (*e.g.* input image, output segmentation), providing a readable uncertainty estimation from a clinician's point of view. Finally, we aim to provide an in-depth discussion of UQ methods' evaluation procedures, as well as to point out the challenges in the field and potential future research directions. Our review differentiates from those previously published, due to the following contributions:

- A review of UQ methods dedicated to DL medical image processing classification and segmentation.
- A presentation of structural uncertainty frameworks for medical image segmentation tasks.
- A focus on the proposed metrics for uncertainty estimates evaluation.
- A discussion on the current challenges and limitations of UQ for medical image analysis, and suggestion of future work directions.

1.2. Organization of this review

This review is divided into five sections. Section 2 introduces the key concepts addressed in this study, namely the application of DL models to medical image classification and segmentation (Section 2.1), as well as the main notions of UQ (Section 2.2). Section 3 presents the most popular UQ methods applied in the context of medical image analysis. Section 4 introduces the notion of structural uncertainty and the various methods proposed for its quantification. Section 5 then focuses on the evaluation procedures that can be implemented to assess the usefulness of uncertainty estimates. Finally, Section 6 proposes a discussion of the current challenges and gaps in the literature in the field of UQ for DL medical image processing.

2. Framework

2.1. Problem setting

In this work, we focus on supervised learning approaches, where a model learns a task T based on a training dataset composed of pairs of input images X, and their associated ground truths Y. The targets represent a class in the context of classification (*e.g.*, healthy, pathological), and represent a mask for segmentation tasks (*e.g.*, the manual delineation of tumors). By observing multiple examples of pairs of images and their corresponding labels during training, the learning agent estimates the probability distribution $p(y|x)$ from the data.

For medical image classification, popular architectures include Residual and Dense Convolutional Neural Networks (CNNs) [16], EfficientNets [17] or Transformers [18]. For medical image segmentation, popular choices include U-Net [19] and its variants, such as Residual U-Net [20], V-Net [21], Attention U-Net [22] or Dynamic U-Net [23]. Recently, transformer-based architectures gained interest, including the U-NETR [24], Swin U-NETR [25] or Trans U-Net [26]. Similarly to medical images that can be either 2-dimensional (*e.g.* 2D CT, Optical coherence tomography (OCT), microscopy, or colonoscopy) or 3D (*e.g.* MRI, 3D CT, PET...), these models can be implemented in 2D or 3D.

During the supervised training stage, the NN uses images from the training set to produce predictions, which are compared to the ground truth targets in order to estimate the error in the model. To do so, a loss function is introduced to estimate the discrepancy between predicted and true labels. Standard choices for both image classification and segmentation include the cross-entropy loss or focal loss [27]. For segmentation tasks, specific loss functions can also be used such as the popular Dice loss [21] and variants: Generalized Dice loss [28] or Tversky loss [29].

In the context of medical image classification, NNs provide a categorical probability distribution over the different observable classes, by applying a Softmax function on the model's output. The final assigned class corresponds to the one having the highest probability. The same process is applied for medical image segmentation, except that the NN predicts one class per pixel or voxel. UQ aims to complement these predictions with uncertainty estimates, allowing for a better interpretation of the results with respect to the model's confidence. In the following section, the main concepts of uncertainty are introduced.

2.2. The specific language of uncertainty

Predictive uncertainty, meaning the uncertainty associated with the prediction of a DL model, is generally divided into two parts: model (or epistemic) and data (or aleatoric) uncertainty.

Epistemic uncertainty describes the lack of knowledge of the learning model [8]. It is considered to be reducible, meaning that it can be reduced by using additional data. In practice, epistemic uncertainty is expected to be high when the model is confronted with images that are different from those observed during the training stage [30]. Such discrepancy between test and training datasets is frequent in medical image analysis, where there may be significant variation between images acquired at different hospitals or using different machines [31]. Additionally, unexpected patterns can be encountered in test images, such as diseases not encountered during training, and artifacts.

Aleatoric uncertainty describes intrinsic noise and random effects within the data [8]. It is not intrinsic to the model, but rather a property of the underlying generative distribution of the data. In the context of medical image analysis, noise can be observed in both the input data (low signal-to-noise ratio, artifacts, partial volume effect, ...), but also in the ground-truth. It has been observed that inter-rater variability in the context of ground truth annotations of medical images was important [32,33]. This has a direct impact on the model's overall uncertainty as the same object of interest (*e.g.* a brain tumor) may have significantly different ground truth delineations depending on the rater. An illustration of such aleatoric uncertainty intrinsic to ground truth labels is presented in Fig. 1.

3. Review of uncertainty-quantification methods for medical image analysis using deep learning

We performed a systematic search on October 2023 using Google Scholar and PubMed to identify DL studies implementing UQ methods for medical image classification and segmentation published from 2015 (included) to October 2023. The following combination of keywords was used for the search: "Deep Learning", "Uncertainty", "MRI", "CT", "PET", "X-RAY", "Ultrasound", "Medical image". Studies were included if they (1) implemented supervised DL models for medical image classification or segmentation; and (2) proposed a quantification of the uncertainty of their algorithms. The following exclusion criteria

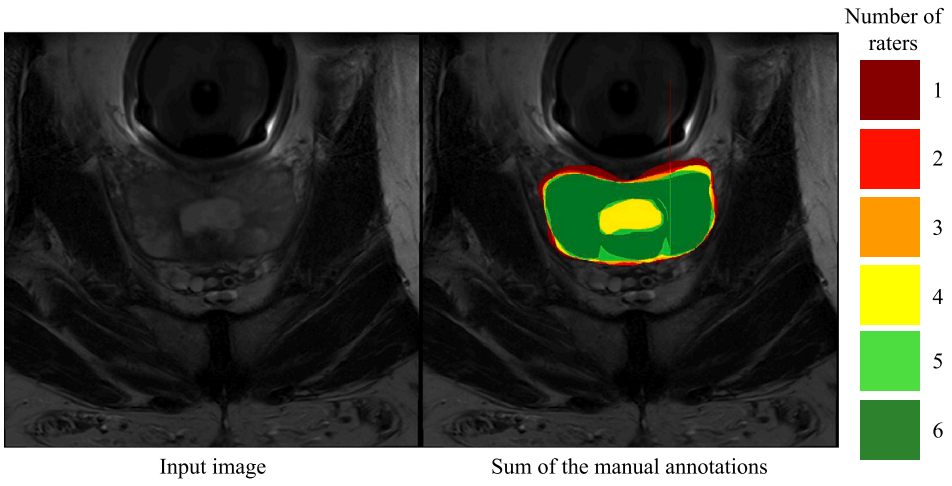

Fig. 1. Illustration of aleatoric uncertainty in prostate MRI segmentation. We present an example image (left) from the QUBIQ 2021 dataset superimposed with the sum of the annotations by six distinct annotators (right). As expected, rater uncertainty is observed at the prostate border.

were applied (1) non-peer-reviewed studies (exceptions were made for papers with more than 30 citations); (2) non-English papers; (3) review articles and (4) animal studies. 218 papers were finally selected for analysis. It resulted in a total of 338 UQ models, implemented either as principal contributions or as comparison methods (the exhaustive list of methods can be found in Appendix, Table 1). We first clustered them according to the method used for uncertainty estimation. We further proposed a categorization of these methods according to the type of uncertainty being modeled, namely epistemic, aleatoric, or both. The resulting taxonomy is presented in Fig. 2. In the following of the section, we briefly present each UQ framework.

3.1. Softmax probabilities

An immediate and intuitive approach to obtain uncertainty estimates from classification or segmentation models is to consider the output predicted probabilities $p(y|x)$ of the NN (Fig. 3). This is based on the assumption that the predicted probability of the NN reflects the true probability of an event, *e.g* when considering all predictions made with a probability of 0.80, the model should be correct 80% of times. This desired property occurs when the NN is well *calibrated*. More formally, writing Y and $\hat{Y}$ the ground truth and predicted label classes and $\hat{P}$ the associated probability, a perfectly calibrated model respects:

$$P(\hat{Y} = Y|\hat{P} = p) = p, \forall p \in [0, 1] \quad (1)$$

In practice, calibrated probabilities do represent meaningful uncertainty estimates, with higher probabilities associated with more confident model predictions. However, modern NNs tend to be highly miscalibrated, meaning that the produced probabilities are unreliable, and usually over-confident [2]. To transform the raw probabilities into real certainty estimates, various calibration methods have been proposed in the literature. Pioneering work proposed the Temperature Scaling approach [2], which consists of rescaling the logits of the NN by a single scalar value, the temperature, which proves to empirically reduce the calibration error without altering the classification result. However, Temperature Scaling also reduces the confidence of correct predictions. More sophisticated approaches were proposed afterward, based on binning approaches [34] or Dirichlet distributions [35]. Finally, while these methods imply a post-hoc calibration once training is completed, another field of work enforces calibration through the learning objective, in order to obtain ad-hoc calibrated models [36].

Due to its simplicity, the utilization of Softmax probabilities as uncertainty estimates was naturally explored for medical image processing applications, often serving as a simple baseline for comparison to more sophisticated approaches. As an illustrative example, Diao et al. [37] and Jungo et al. [38] leveraged the entropy of the (uncalibrated) Softmax probability vectors for brain tumor segmentation in MRI. Alternatively, DeVries et al. [39] used the Maximum Softmax Probability (MSP) uncertainty estimator, corresponding to the highest probability class for each voxel, for skin lesion segmentation in RGB images. A similar score is used for out-of-distribution (OOD) detection experiments in the context of chest X-ray pathology classification [40] and COVID-19 lesions segmentation in CT scans [41], respectively. Calibration was explored in Carneiro et al. [42], where authors employ Temperature Scaling to recalibrate the predicted probabilities of a polyp classification model. Finally, Murugesan et al. [36] and Liang et al. [43] proposed incorporating calibration terms in the training objective of their NN in the context of segmentation and classification of medical images, respectively, in order to obtain well-calibrated predicted probabilities.

It is important to note that UQ based on Softmax probabilities only considers the distribution over the model's outputs and not the model's weights. Thus, this type of deterministic uncertainty estimate only considers aleatoric uncertainty [8,30].

3.2. Conformal Prediction

Conformal Prediction (CP, Fig. 4) is a statistical approach for uncertainty quantification that has been attracting a lot of attention lately in the ML community. While its fundamental concepts are not new [44], CP has been extensively revisited in DL pipelines as it has several appealing properties: it makes no assumption about the black-box predictor nor the distribution of the data, and it provides provable statistical guarantees. The core concept of CP is to transform the point-wise prediction of a model into a predictive set. In the classification setting, these predictive sets correspond to a list of probable class labels while for regression tasks, they correspond to predictive intervals (PIs) associated to the regressed value [45]. These sets are constructed so that the ground truth label is guaranteed to be included with a user-defined confidence level, such as 90% or 95%. This corresponds to the desired coverage level. To achieve this result, CP performs a post-processing of the raw predictions of the model (class probabilities for classification, or predicted scores for regression) and is usually fit using a set-aside labeled calibration dataset that comes from the same distribution as the test dataset. This procedure is called split CP. It is important to note that in contrast to other UQ methods that aim at complementing a prediction with an uncertainty estimate, CP instead starts by defining a target level of uncertainty, and then adapts the prediction accordingly.

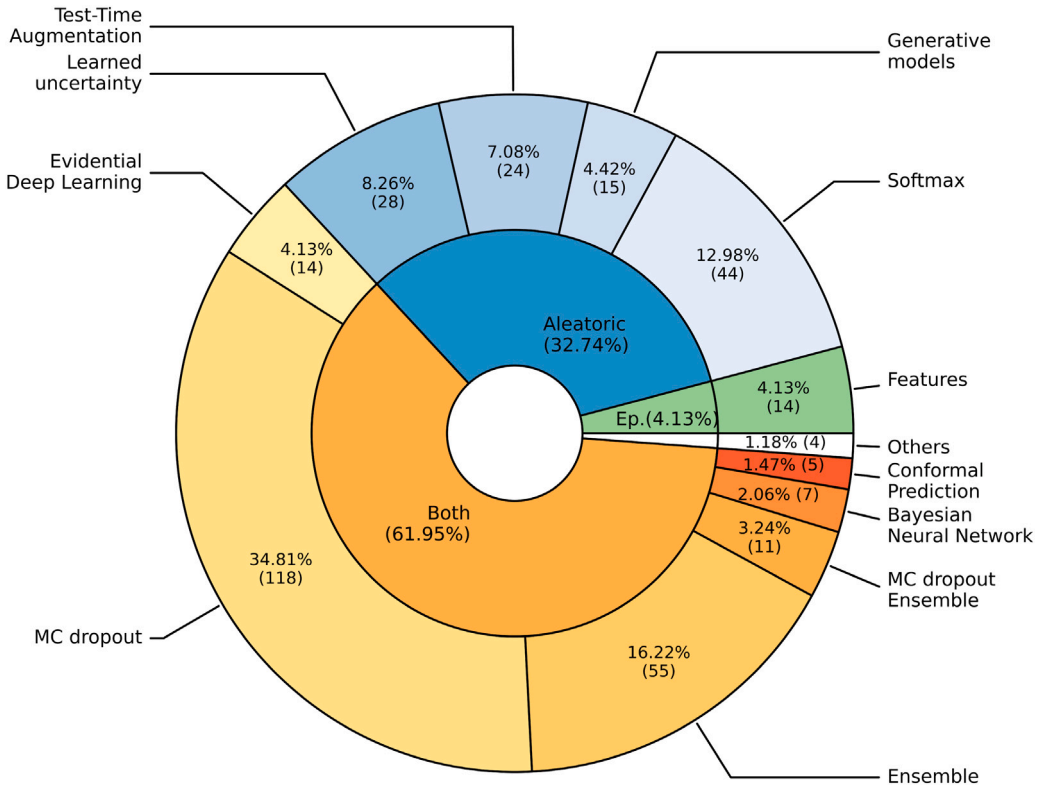

Fig. 2. Implemented UQ methods in the 218 selected papers. The percentage (and the number) of the selected papers for each class of methods is indicated in the outer ring and the corresponding percentage is below the class name. The inner ring classifies methods according to the type of uncertainty modeled: aleatoric, epistemic or both.

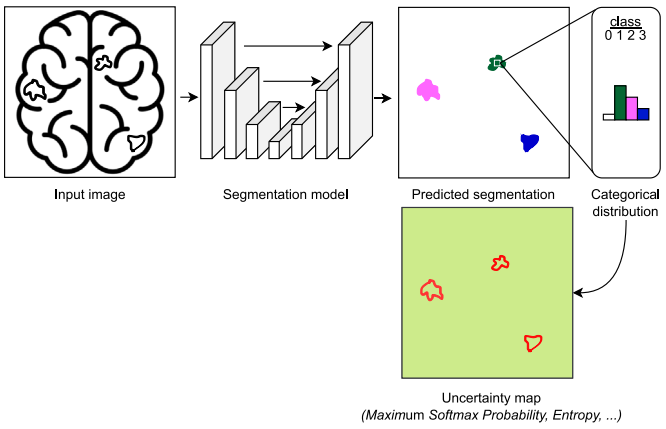

Fig. 3. Illustration of the Softmax uncertainty paradigm.

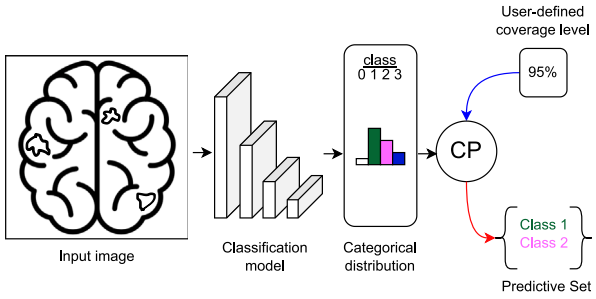

Fig. 4. Illustration of the Conformal Prediction paradigm.

CP has found many applications in natural image classification [46], regression tasks [47], or drug discovery [48]. Applications to medical images are emerging: CP is employed in the AmnioML framework [49] to provide PIs associated with Amniotic Fluid volume prediction. A similar objective is pursued in the TriadNet model [50], which enhances a multi-head segmentation model with CP in order to provide 90% PIs associated with tumor volume estimation. As opposed to volume prediction, Eaton et al. [51] focuses on computing PIs for counting tasks, applied to cell and brain lesions counting. Finally, CP is also investigated in two recent studies focusing in medical image classification [52,53].

While extremely promising for medical applications as it provides statistical guarantees to the user concerning the error of the deployed model, CP suffers from 2 major limitations that may hinder its usage in the field. First, it is based on the assumption that calibration and test data are exchangeable, meaning that they come from the same distribution. However, it is well known that domain shifts are extremely common in medical imaging applications, which hinder the effectiveness of the conformal procedure [54]. Second, the calibration dataset must be large enough to perform the split CP procedure. The current guideline is to use 1000 calibration samples [45]. In medical applications, data is often scarce, hence obtaining high-confidence PIs using split CP may not always be feasible.

3.3. Bayesian Neural Network methods

In Bayesian Deep Learning (BDL, Fig. 5), each weight of the NN is replaced by a distribution, rather than having a single fixed value

[55,56]. To achieve this, a prior distribution $p(w)$ (usually Gaussian) is first initialized over the NN weights. It follows that each weight is represented by a mean and a variance (thus doubling the number of parameters of the model). Then, during training, the model learns the posterior distribution $p(w|D)$ given the training dataset D and the prior distribution, which accounts for the least- and most likely parameters given the observed data. The trained Bayesian Neural Network (BNN) is akin to a virtually infinite ensemble of NNs, where each instance has its weights drawn from the learned posterior distribution. At inference, to obtain a prediction, the distribution should be marginalized over all model parameters. Although it is possible for very simple neural networks, modern neural networks are over-parameterized, which makes the exact computation of the posterior intractable [57]. To deal with this issue, a branch of work has focused on approximating the true posterior using Variational Inference (VI). VI proposes to approximate the posterior using a variational distribution $q(w|\theta)$ [10]. The parameters θ of the variational distribution are learned during training to be as close as possible to the exact posterior. This is achieved by minimizing the variational free-energy cost function, usually referred to as the expected lower bound (ELBO) [55]. Minimization of this loss is achieved using Stochastic Gradient Descent (SGD), as in standard neural networks. This training paradigm is called Bayes by Backprop (BBB) [55]. VI thus allows to address Bayesian Inference as a classical optimization problem. Once training is completed, various uncertainty estimates can be obtained, such as the entropy of the predictive distribution, its variance, or its mutual information. BDL places a distribution on the model's weights, hence it is rooted in epistemic uncertainty quantification. However, BDL applied to classification and segmentation tasks also produces a categorical probability distribution, so it can be easily coupled with the Softmax probabilities framework previously introduced (Section 3.1) to also quantify aleatoric uncertainty.

Applications of BDL to medical image processing are scarce. Studies initially focused on applying Bayesian convolutions associated with VI approaches. We found applications for 2D medical image classification [58,59], knee abnormality detection [60], lung and nasal endoscopy CT segmentation [61] and brain tumor segmentation [62]. However, this approach requires extensive changes in the model architecture and training paradigm [55,56], associated with an increase in the computational cost of both training and inference. This has motivated recent studies on scalable BDL solutions. For example, in Adams et al. [63], authors evaluate Rank-1 Bayesian networks [64] as well as latent posterior BNN on a task of organ segmentation in 3D CT. These two approaches were recently proposed as scalable alternatives to the standard BDL framework.

3.4. Monte Carlo dropout methods

In Gal et al. [65], authors demonstrated that a NN trained with dropout [66], a regularization technique based on the random dropping of activation, is able to efficiently approximate Bayesian inference without the associated prohibitive computational cost. Based on this principle, the Monte Carlo Dropout (MC dropout, Fig. 6) technique proposes to train a model with dropout and keep it activated during inference. For a given query input, multiple forward passes are then performed. Each time, a different dropout mask is randomly sampled (generally following a Bernoulli distribution), producing different predictions. Following this process, a predictive distribution is obtained, similar to BNN. As for BDL, MC dropout was initially proposed to tackle epistemic uncertainty, although it still produces a categorical probability distribution from which aleatoric uncertainty estimates can be computed [67]. MC dropout allows the approximation of a BNN in any network trained with dropout, it thus rapidly gained popularity, and applications in the medical imaging field are numerous. Implementations of this framework vary little. Studies that stand out have studied in further detail the importance of the dropout layers' position and type, rate, and the number of drawn MC dropout samples. Jungo et al.

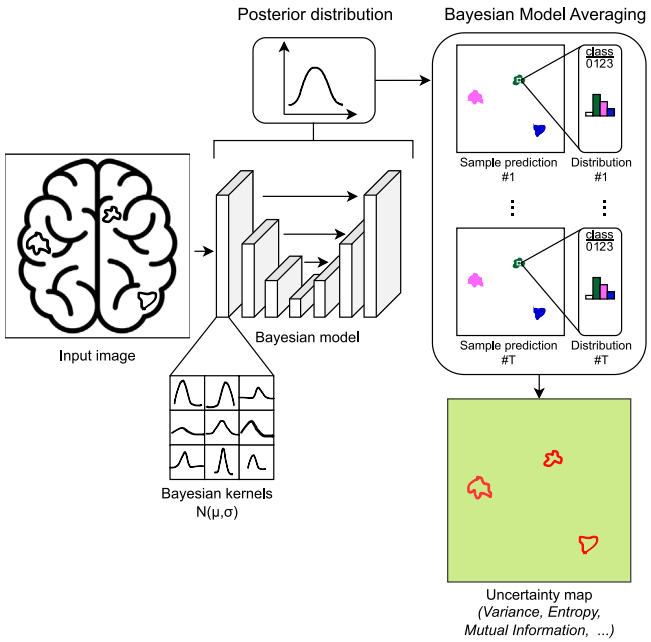

Fig. 5. Illustration of a Bayesian Neural Network.

[38,68] studied the importance of the positioning of the dropout layer within a convolutional network for brain tumor segmentation in MRI. Orlando et al. [69] evaluated the impact of the number of MC samples at inference on the segmentation accuracy of the photoreceptor layer in OCT scans, and found no improvement after 20 samples. Similar work was carried out in Camarasa et al. [70], where the impact of the dropout rate and type (Bernoulli or Gaussian dropout) was assessed on a task of cardiac MRI segmentation. While the dropout type had little impact on the segmentation performance, they found that the choice of the dropout probability was critical, as the performance of their model significantly decreased for a dropout probability superior to $p = 0.50$. In Ghoshal et al. [71], authors propose to quantify both aleatoric and epistemic uncertainty using MC dropout in a task of nuclei segmentation in microscopy images and found that increasing the number of MC samples led to a decrease of the measured aleatoric uncertainty.

Other studies have proposed improvements to the standard MC dropout technique. In Jungo et al. [38], authors propose to use concrete dropout, a variant where the dropout rate at each layer is learned as part of the optimization process [72], but found no improvement as compared to the standard Bernoulli dropout. In a similar vein, [73] proposed a novel Spike-and-Slab dropout strategy, allowing to learn during training the dropout probability for each convolutional filter independently, for brain parcellation in T1-weighted MRI. Alternatively, Mobiny et al. [74] applied a variant of dropout called DropConnect [75], following which weights are randomly set to zero instead of activation, and demonstrated its advantages on various segmentation tasks, including organ segmentation from CT scans.

3.5. Ensembling methods

Deep Ensemble [76] (DE, Fig. 7) proposes to quantify uncertainty from a series of sequentially trained NN. As the weights of the neural networks are initialized randomly, the models reach different optimum during training. As a result, they produce diverse predictions for the same query input. As for BDL and MC dropout, uncertainty estimates or both aleatoric and epistemic uncertainties can then be extracted from the ensemble's predictive distribution [77]. A DE does not require any changes to model architecture or training paradigm and is known for

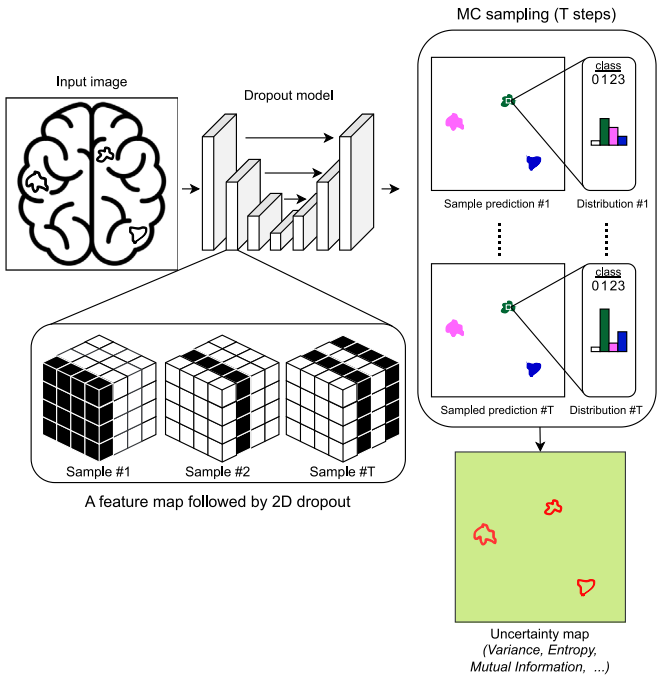

Fig. 6. Illustration of the Monte Carlo dropout paradigm.

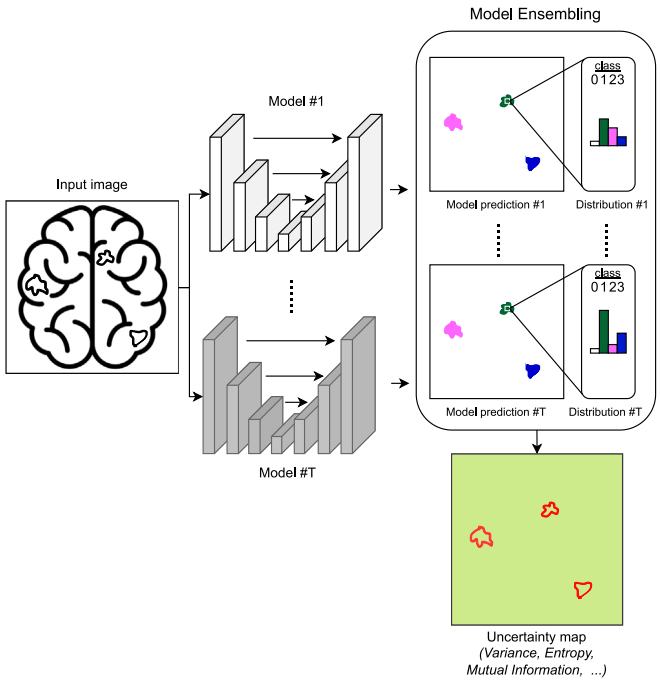

Fig. 7. Illustration of the deep Ensemble paradigm.

boosting predictive performance. Yet, it requires to repeat the training several times, which is cumbersome when the model is complex. Moreover, the aggregation of each individual prediction at inference increases the computational cost of this approach.

Ensembling techniques have been widely studied for UQ in medical imaging, with various studies demonstrating superior predictive performance and uncertainty quantification quality, as compared to the MC dropout approach [78–80]. Noticeably, efforts are carried out to develop more efficient ensembling strategies that preserve the performance and uncertainty gain of the standard DE approach, without the prohibitive computational cost. Typical examples include multi-output architectures, that are able to produce a diverse set of predictions in a single pass using different branches or heads. This concept has been applied to both medical image classification [81] and segmentation [62]. In a similar vein, Layer Ensemble [82] proposes to add a dedicated output to each intermediate layer of a segmentation model to form an ensemble within a single network. The same concept is adopted in the Early-Exit Ensemble [83] and applied to medical image classification tasks. Other illustrative examples include Checkpoint Ensembling [84] which builds an ensemble from different checkpoints saved during the course of a single NN training. Alternatively, the Stochastic Weight Averaging Gaussian (SWAG) framework [85] can be viewed as an efficient way of ensembling. It aims at estimating a Gaussian approximate posterior over the weights of a NN by sampling its weight configurations during training, using a constant learning rate. At inference, from this distribution, it is possible to sample an ensemble of diverse models. Two applications of SWAG can be found in the medical image processing literature, for retinal artery-venous segmentation [86] and chest X-ray classification [87], respectively. Finally, another research lead corresponds to developing techniques to improve diversity within ensembles, which is known to be a key factor of this technique [88,89]. Two efforts in this direction are Orthogonal Ensembles [90], which optimize the orthogonality between ensemble members' weights to promote variety among predictions, and diversity-promoting ensembles, composed of varied NN architectures explicitly chosen to minimize the correlation between their predictions [91].

Finally, it is worth noticing that some works propose to associate ensemble and MC dropout, forming the so-called Ensemble Monte Carlo (EMC) [58,92,93]. This allows to investigate two different types of uncertainty, namely (i) uncertainty resulting from the random seed used to perform SGD training, yielding to different optima when sequentially training NNs and (ii) the weight uncertainty within each unique ensemble member assessed using the MC dropout approach.

3.6. Learned uncertainty frameworks

Learned uncertainty frameworks (LU, Fig. 8) are built on the idea that aleatoric uncertainty can be learned during training directly from the data itself. The most immediate approach, for segmentation tasks, is to treat the inter-rater variability as a ground truth for uncertainty. Supervised learning strategies can then be adopted to reproduce the distribution of the raters annotations [94–97]. However, this approach is limited to datasets where multiple ground truth segmentations are available, which is usually not the case. Most LU approaches have thus developed strategies to learn the segmentation and uncertainty conjointly without the need for explicit ground truth labels for uncertainty. In this direction, the initial proposal is to suppose that the network output logits are corrupted by Gaussian noise with mean equal to 0 and variance z. The higher the variance, the higher the aleatoric uncertainty. A model can then be trained to predict the mean logits ρ, as well as the noise variance z [30]. To do so, the model is equipped with two outputs: one for the logits, and one for the variance. During training, a modified version of the cross-entropy loss is adopted to learn the variance. Illustrative examples of this approach can be found for Multiple Sclerosis lesions [98], tumor [99], and atlas segmentation [100] in brain MRI. This framework was recently extended to skewed-Gaussian distributions in order to quantify asymmetric contour uncertainty [101]. Another lead consists of learning uncertainty estimates directly correlated with the errors of the model, usually for segmentation applications. This is performed using an uncertainty-augmented loss function. The initial idea is that the errors are available at each training iteration by computing the differences between the ground truth labels and the predicted labels. Then, it is possible to use the computed error maps as ground truth indicators for uncertainty. In McKinley et al. [102], authors present a modified segmentation

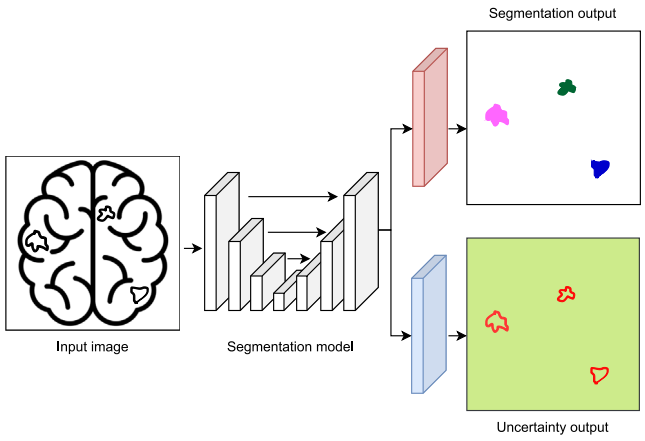

Fig. 8. Illustration of the learned uncertainty paradigm.

architecture with 2 output convolutions: one for the segmentation probabilities, and one for an uncertainty output called labelflip probability. They also propose the labelflip loss in order to correlate this predicted uncertainty score with the incorrectly segmented voxels. An alternative approach is the Learned Confidence Estimates loss proposed in Devries et al. [39]. In this framework, the segmentation model is also equipped with two separate outputs for the segmentation and for a confidence estimate, respectively. This estimate is used to interpolate between the predicted probability distribution and the target distribution so that low-confidence voxels are pushed toward the correct output. Finally, Diao et al. [37] proposed the uncertainty cross-entropy loss, which is an extension of the standard cross-entropy with an extra class corresponding to the uncertain case, when no other classes can be predicted with confidence. The loss can be minimized in two fashions, either by (i) predicting the correct class label or (ii) predicting the uncertainty class. The same motivation is at the core of the Deep Gambler model [103], recently applied to medical image classification [104], that converts a $m-$class classification problem into a $m+1$ problem where the extra class represents abstention from answering.

3.7. Test Time Augmentation

Test-Time Augmentation [105] (TTA, Fig. 9) was proposed as an UQ method to evaluate aleatoric uncertainty. At test time, multiple variants of the input image are generated using Data Augmentation. This can include spatial transformations (*e.g.* flipping, rotation) as well as intensity augmentations (*e.g.* contrast modification, noise injection, or artifacts). The model generates a prediction for each augmented variant of the input image. From this distribution, uncertainty metrics can be extracted such as the entropy or the variance. The TTA process aims to explore the impact of input-image transformations on the prediction. TTA is particularly interesting as it is completely model-agnostic: it does not require any particular architecture or training design, and can thus be used with any pretrained or open-source model. Various TTA strategies were proposed for medical image applications. In its simplest setting, only flipping is applied [41]. Noise and intensity shifts are also commonly added to the augmentation pipeline [37,106,107]. Ayhan et al. [105] give an example of a more elaborate setting, where a TTA scheme based on 128 variants is applied per input image using both extensive intensity and geometry transformations. In a more original manner, [108] proposed to use in-painting as TTA for uncertainty estimation.

3.8. Generative models

Image-conditional generative models have been explored for UQ in medical image segmentation. The main objective is to generate

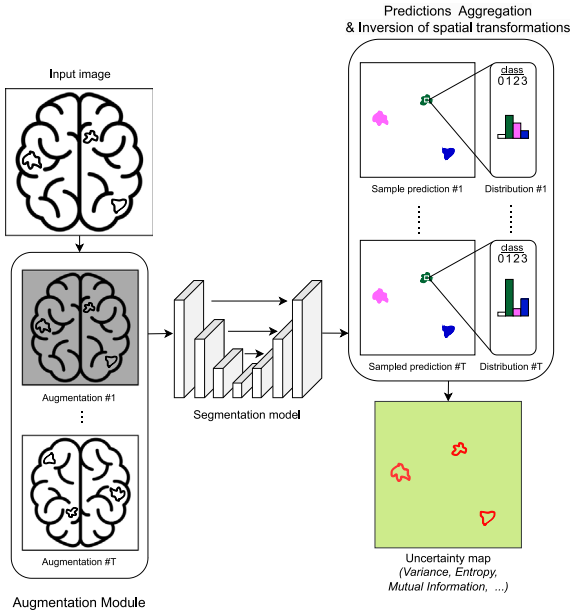

Fig. 9. Illustration of the Test-Time Augmentation paradigm.

various plausible, spatially correlated segmentation masks for a given input image. The first attempt in this direction was achieved with the Probabilistic U-Net [109] which proposed a segmentation architecture based on a Variational Autoencoder (VAE). This allows the encoding of the input image into several multivariate normal latent variables which are then decoded into diverse variations of the same region of interest. This framework is illustrated in Fig. 10. Several improvements were then proposed to extend the expressivity of this generative model, such as with the Hierarchical Probabilistic U-Net [110], the PHISeg [111], the RevPHISeg models [112], or by the insertion of Normalizing Flows [113–115]. Another interesting variant was proposed in the Stochastic Segmentation Network [116], which uses a low-rank multivariate normal distribution on the logit space instead of a VAE, allowing the generation of a set of spatially coherent segmentation for each input image. More recently, diffusion models were applied to this problem [117]. Note however, that these different approaches are based on the sampling (either sampling several plausible masks at test time for Probabilistic U-Net and variants, or an iterative generative process for diffusion models), hence their computation cost is higher than that of a standard segmentation model.

3.9. Feature-based methods

From a practical point-of-view, epistemic uncertainty is expected to be high for Out-of-distribution (OOD) images, *e.g.* images that are far from the training image distribution. Based on this concrete application, efficient epistemic uncertainty techniques have been recently proposed to detect OOD from the intermediate features of a trained NN [118]. This builds on the hypothesis that the feature maps computed when processing an input image contain information regarding its conformity (Fig. 11). These methods are computationally efficient and are increasingly experimented in medical image processing applications. For instance, the Mahalanobis Distance was investigated to detect outliers in the context of COVID-19 lesions segmentation [41], X-ray classification [40,119,120] mammography classification [121] and more recently liver segmentation in T1-w MRI [122]. As alternatives, Karimi et al. [123] proposed to study the spectral signature of the intermediate feature map by computing its Singular Value Decomposition in order to detect OOD images, while Diao et al. [37] computed class-wise prototypes from the feature representations in the context

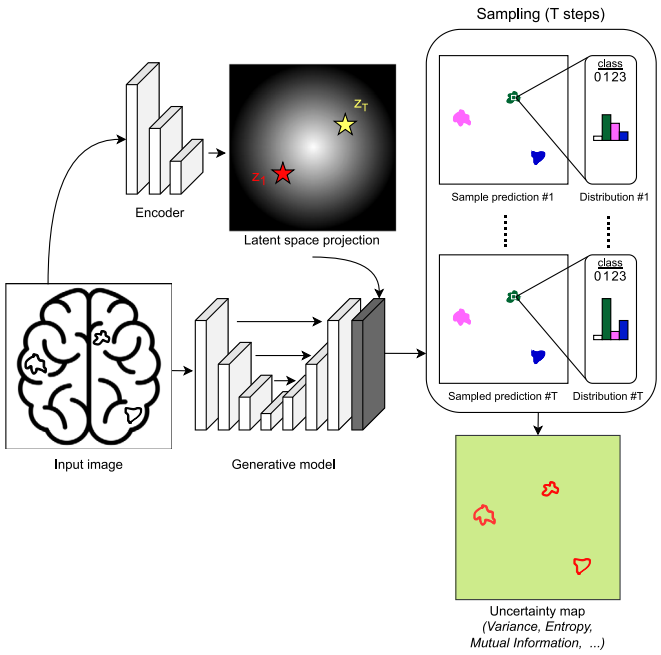

Fig. 10. Illustration of the Probabilistic U-Net framework.

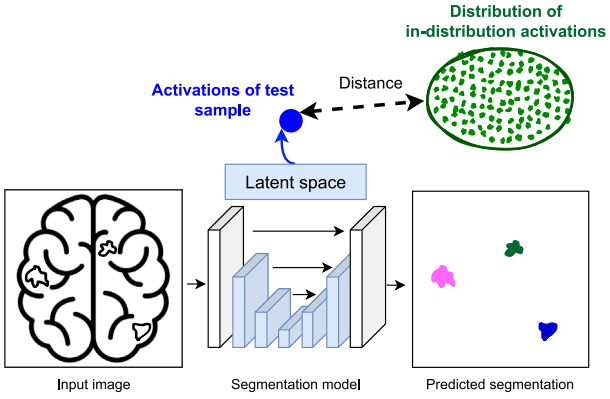

Fig. 11. Illustration of feature-based Out-of-distribution detection paradigm.

of brain tumor segmentation in MRI, allowing to detect train–test mismatches. Finally, Lambert et al. [124] proposed a benchmark of the aforementioned solutions on a task of tumor segmentation in 3D brain MRI, and analyzed the importance of the number of feature maps used to compute the OOD scores.

3.10. Evidential Deep Learning

The Dempster–Shafer Theory of Evidence (DST) is a framework for dealing with both epistemic and aleatoric uncertainty [125]. In a K-class classification problem, DST proposes to assign belief masses b^k to each possible class for a given input image as well as an overall uncertainty mass u such that:

$$1 = \sum_{k=1}^{K} b^k + u \tag{2}$$

where $b^k > 0$ and $u > 0$. Beliefs are computed from the evidences e^k, which are typically obtained by applying a Softplus operator to the raw logits predicted by the NN [126], such that:

$$b^k = \frac{e^k}{S} \text{ and } u = \frac{K}{S} \tag{3}$$

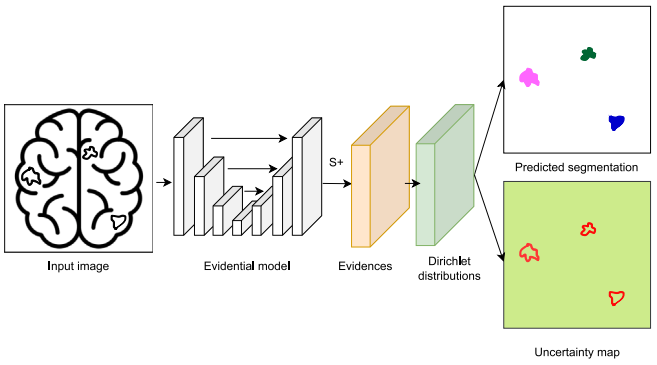

Fig. 12. Illustration of the Evidential Deep Learning paradigm. S+: Softplus function.

$$\text{with } S = \sum_{k=1}^{K}(e^k + 1) = \sum_{k=1}^{K} \alpha^k$$

where S is called the Dirichlet strength. When there is no evidence collected guiding to any of the K classes, the beliefs reach their minimal values 0, while the overall uncertainty reaches its maximal value 1. Finally, DST proposed to parametrize a Dirichlet distribution on the model's outputs in place of the categorical distribution using the parameters $\{\alpha^1, \alpha^2, \ldots, \alpha^K\}$. Interestingly, the realization of a Dirichlet distribution is still a distribution. It can be intelligently used to replace the standard categorical probability distribution of a classification NN by a distribution over possible Softmax outputs, thus modeling second-order probabilities and uncertainty [127]. It is thus more expressive in terms of UQ than the standard Softmax probability framework (Section 3.1). In EDL, the class probabilities are obtained as $p^k = \alpha^k / S$ [127]. Aleatoric uncertainty estimates can be obtained from the estimated probability, similarly to the Softmax uncertainty framework, while epistemic uncertainty can be evaluated using the estimated u, which encompass the accumulated evidences. DSL applications can be found for both medical images segmentation [126,128,129] and classification [130–132]. This framework is illustrated in Fig. 12.

3.11. Other UQ methods

Finally, we found a few methods not conforming to any of the frameworks previously introduced. In Jensen et al. [133], the authors explore the Monte Carlo Batch Normalization (MCBN) framework, a variant of MC dropout making use of the stochasticity of batch normalization layers. In Jungo et al. [38], an auxiliary net is proposed to detect the errors of a segmentation model, and the voxel-wise error probability is used as an uncertainty metric, allowing the decoupling of the uncertainty and segmentation tasks. Toledo et al. [134] plugged a Gaussian Process at the end of a DL feature extractor for diabetic retinopathy classification. Finally, Wang et al. [135] addressed contour uncertainty by replacing the binary segmentation masks with a soft alpha matte mask.

In conclusion, Fig. 2 summarizes the set of UQ methods used in the current literature for medical imaging classification and segmentation. Note the importance of MC dropout and Deep Ensemble that represent altogether 50% of the set of UQ methods and 80% of those addressing both aleatoric and epistemic uncertainty in medical applications.

4. From voxel uncertainty to lesion-level and case-level uncertainties

In the previous section, the most popular approaches for UQ in medical image analysis have been introduced. In a segmentation setting, all methods except the feature-based approach produce voxel-wise uncertainty estimates when applied to 3D medical image segmentation.

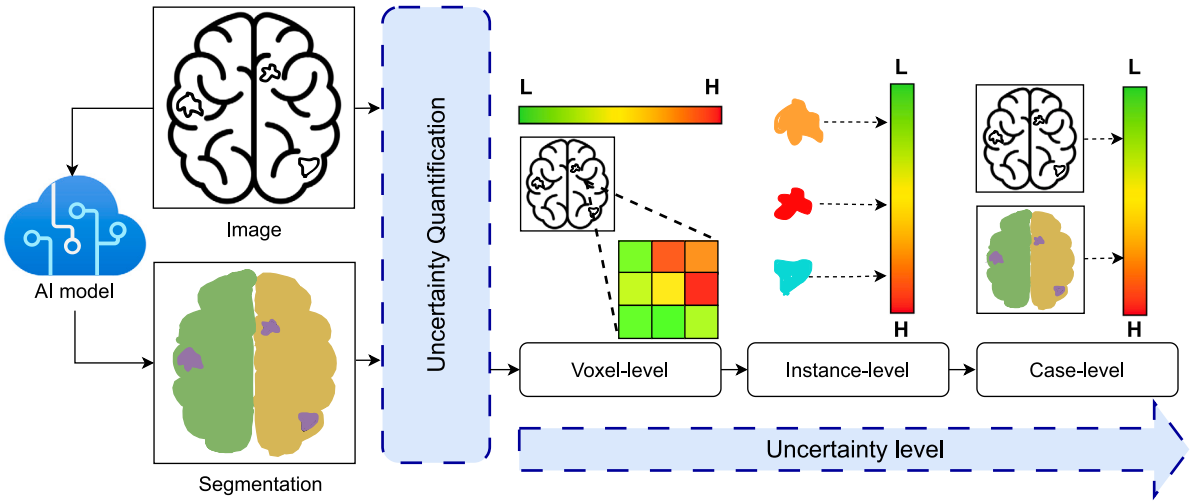

Fig. 13. Illustration of the different levels of uncertainty that can be calculated for medical image segmentation. In this example, the segmentation of the model contains 2 anatomical classes (green and yellow) as well as one lesion class (purple). Voxel uncertainty estimates can be used to derive instance-level uncertainty estimates for each lesion. The analysis can be complemented with case-level scores to estimate the overall quality of the input image and the output segmentation, respectively. L: Low uncertainty. H: High uncertainty. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

While this is convenient for visualization purposes, this is not totally aligned with medical attention, which is usually located at the structure level (*e.g.* lesion or anatomical region). Moreover, when processing 3D medical images, the visual inspection of the entire volume to monitor uncertain areas can be time-consuming. A branch of the medical imaging UQ literature has thus focused on scaling uncertainty estimates at higher levels. These structural uncertainty scores have been mainly explored in 2 settings: (1) the binary segmentation of lesions, for which an uncertainty score is assigned to each identified lesion instance, and (2) case-level uncertainty, where the goal is to identify non-conform images (input QC) or poor-quality predictions (output QC). These different scales are illustrated in Fig. 13, and the corresponding literature is further presented in the following.

4.1. Lesion-level uncertainty estimates

Lesion-level confidence estimation is an emerging UQ application that consists in attributing one uncertainty score for each identified lesion in a medical image. This is relevant for applications that rely on the detection of multiple lesions, and for which precise counting is required. A pioneering study in this direction is the work of Nair et al. [98] which proposes to fuse voxel-level certainty scores to lesion-level scores, for Multiple Sclerosis lesions segmentation. The proposed strategy consists of using the logsum operator on the uncertainty of the voxels composing each lesion. One downside is that this metric systematically attributes higher uncertainties to smaller lesions. Other aggregation operators have further been proposed for Multiple Sclerosis, such as the average of the voxel uncertainties [136]. In a different direction, several plausible masks can be obtained for each unique brain lesion using a DE, and the agreement between the ensemble members has been proposed as a lesion-level uncertainty sore, although not using directly the voxel uncertainty maps [136].

Lesion-level estimates were also explored for liver tumor lesions [137,138]. In these two studies, the authors start by computing voxel-uncertainty maps, for example using MC dropout, DE or TTA. Then, they compute radiomics [139] for each lesion using the uncertainty maps. These features are used to train a classifier (SVM) to predict the status of the lesion: True Positive (TP) or False Positive (FP). While their primary focus is the reduction of FP liver lesions, it seems that the FP probability of the lesion constitutes a viable lesion-level uncertainty score, where higher values are associated with more ambiguous lesions. A similar strategy was proposed by Ozdemir et al. [140] for FP reduction in nodules detected in lung CT. The CT is first processed by a segmentation model, providing a segmentation and uncertainty map. Bounding boxes are further extracted from the segmentation, centered at each identified nodule. An auxiliary CNN classifier then predicts the probability that the lesion is a FP.

4.2. Case-level uncertainty estimates

A frequent question when dealing with uncertainty is to wonder if the overall prediction can be trusted. To answer this, case-level uncertainty scores can be furnished to the user to provide a general impression regarding the model confidence for a given case. In the context of medical image segmentation, such scores can actually be computed to detect non-conform inputs (input-level QC) or poor-quality segmentations (output-level QC).

4.2.1. Input-level QC

Input-level QC, akin to OOD detection, aims at detecting samples nonsuitable for analysis. For these images, epistemic (i.e. model) uncertainty is expected to be high [30]. Thus, monitoring the output uncertainty of predictive models could theoretically be used to detect poor-quality input images. This idea was explored in McClure et al. [73] where authors used MC dropout to estimate the voxel uncertainty of a brain tumor segmentation model. From these uncertainty maps, image-level scores are derived by averaging voxel scores across the volume. The scores are then used to detect poor-quality scans. A similar process is adopted by Gonzalez et al. [41] for non-conform input detection in chest CT segmentation, using MC dropout and TTA as voxel uncertainty estimators. In line with these works, the previously presented feature-based OOD detection methods (Section 3.9) produce case-level scores that can be used to perform input-level QC.

4.2.2. Output-level QC

Output-level uncertainty estimates aim at evaluating the overall quality of an automated segmentation. An intuitive solution is to fuse the voxel uncertainty estimates computed by a standard UQ methodology (MC dropout, DE, TTA...) to a case-level score, for example, using the mean [141]. Following the observation that voxel uncertainty is often concentrated at the boundaries between classes, two studies have proposed to reduce the weight of contour voxels to get more accurate structural uncertainties. This prior knowledge-based aggregation was investigated in Jungo et al. [38] and Graham et al. [142].

Instead of relying on voxel uncertainties, a series of studies have proposed to focus on the set of plausible segmentation masks generated by standard UQ methodologies. Pionnering work in this direction is the study carried out by Roy et al. for whole brain segmentation [141]. They use MC dropout to generate a set of segmentation masks for each input image and use the disagreement between the samples as structural uncertainty estimates. This follows the intuition that disagreement between the predictions should be higher for poor predictions. In this direction, they propose different proxies: the Coefficient of Variations among the volumes, and the Dice & IoU agreements between the MC samples. They further show that these proxy metrics correlate strongly with the true Dice, unknown during inference. This concept of using a set of plausible segmentation masks to compute structural uncertainty metrics was further explored via MC dropout sampling [143,144], Deep ensemble [145] and TTA [106,146]. Other proxies were proposed, including the Predictive Dice Coefficient [143], the Contour Quality metric [147] or the Doubt score [148], which have all demonstrated a strong correlation with the true Dice coefficient.

To further improve the output QC procedure, several studies have explored the use of these uncertainty metrics as features to train a ML model to directly infer the prediction quality, in a regression setting. Ghosal et al. [149] and Hann et al. [150] trained linear regression models to predict the Dice directly from uncertainty estimates of MC dropout models, for digital histopathology image segmentation and cardiac MRI segmentation, respectively. Alternatively, Arega et al. [151] used a Random Forest either in a binary classification approach (accept/reject poor segmentation) or regression (predict the Dice score) from the outputs of a MC dropout model. These approaches require building a training dataset comprising automated predictions and associated quality to allow the training of the auxiliary ML model.

5. How to evaluate uncertainty quantification approaches

In the previous sections, we have presented the main UQ approaches that are applied to DL-based medical image classification and segmentation. In this section, we now propose to introduce the different protocols that are implemented in these papers to evaluate the relevance of the UQ approaches. Evaluating UQ approaches is not straightforward, as we generally do not dispose of ground-truth uncertainty values. Proxy metrics are thus developed to estimate the performances of uncertainty quantification methods. We have identified 6 types of evaluation protocols (see Fig. 14, the exhaustive list of protocols and the corresponding cited references can be found in Appendix, Table 2). In the following, we present each protocol and identify their use cases.

5.1. Qualitative assessment protocol

As computing quantitative metrics for uncertainty is not direct, several works focused on a qualitative assessment of the predicted uncertainty estimates. In this context, a visual inspection of the cases considered as certain/uncertain is usually performed to verify whether they correspond to cases that a human would consider as uncertain [152–154]. Alternatively, the relevance of the incorporation of UQ in a medical image processing pipeline can be assessed via the monitoring of its beneficial impact on a downstream task such as active learning [155], curriculum learning [156–158], weakly-supervised learning [159], semi-supervised learning [160–164], cascaded inference tasks [165,166], segmentation refinement [167], federated learning [168], cross-domain generalization [169], or predictive performance [170].

5.2. Calibration metrics

As presented in Section 3.1, the output Softmax probabilities of a NN can directly be used as a marker of uncertainty. A popular way of estimating the accuracy of such uncertainty estimates is the use of calibration metrics, that verify the correspondence between predicted probabilities and actual error rates. Usual choices consist of the Expected Calibration Error (ECE) [36,38,79,132,140], or the Negative Log-Likelihood (NLL) score [78]. It is also common to represent calibration visually using calibration plots, also sometimes called reliability diagrams [2]. Such a diagram is illustrated in Fig. 14, **Calibration**. The main idea is to plot the model's accuracy as a function of its confidence. This is performed by binning predictions according to the associated predicted probability, and computing the accuracy for each bin. Following Eq. (1), the calibration plot of a perfectly calibrated model should correspond to the identity function. Any deviation to this identity function corresponds to a calibration gap (either over or under confidence). Additionally, the ECE can be directly extracted from the plot by computing the difference between the average accuracy and confidence, for each bin.

5.3. Coverage error

As presented in Section 3.2, Conformal Prediction is traditionally defined around the notion of coverage, following which a fraction of the ground truth labels should be included in the predictive sets (*e.g.* 95% or 99%). A natural way of evaluating uncertainty under this framework is to compute the distance between the empirical coverage on test data and the user-defined target coverage [45]. If implemented properly, CP is statistically guaranteed to approximate the target coverage. However, this can be achieved with unnecessarily large intervals. Let us take the example of predicting the volume of a brain lesion, that we want to equip with a predictive interval using CP. A perfect coverage of 100% could be achieved by predicting a lower bound of 0 mL, and an upper bound corresponding to the overall intracranial volume. However, these intervals would be useless in practice. Thus, CP methods are also generally evaluated with respect to the average interval width, where narrower values are preferred [50,51].

5.4. Error detection and referral

A direct downstream application of uncertainty in an automated pipeline is the detection of samples for which the prediction is likely to be incorrect. This is crucial to prevent silent errors that could have a dramatic impact, especially in real-world medical image applications. Error detection is thus commonly used to estimate the quality of uncertainty estimates. In this scenario, the model's predictions are classified into two groups, certain and uncertain samples, by thresholding their associated uncertainty. The result of this classification is then compared to the correctness of each sample, namely correct or incorrect. In that context, a confusion matrix from the uncertainty point of view can be constructed, by distinguishing 4 possible cases, as shown in Fig. 14, **Error Detection**. Usual classification metrics can then be computed based on the counts of True Positive (TP): the classification is uncertain and the expected label and the prediction differ, False Negative (FN): the classification is certain but the expected label and the prediction differ, True Negative (TN): the classification is certain and the expected label and the prediction are identical, and False Positive (FP): the classification is uncertain but the prediction and the expected label are identical. Illustrative applications of this metric can be found for COVID-19 detection [80], cell colony segmentation [171] and skin disease assessment [172]. In a similar vein but specifically for image segmentation, the uncertainty-error overlap was also proposed [38] and further extended using mutual information [173]. Other variants include the use of distance metrics such as the Wasserstein distance [174] or Jensen–Shannon [175] to measure how much the predicted uncertainty correlates with the distribution of errors of the model.

Another variant of this framework is the referral mechanism (also sometimes referred to as rejection or filtering) [176]. In this context, predictions of the model are ordered from the most certain to the most uncertain. A fraction of the most uncertain predictions are then rejected (i.e. referred to the expert), and the performance of

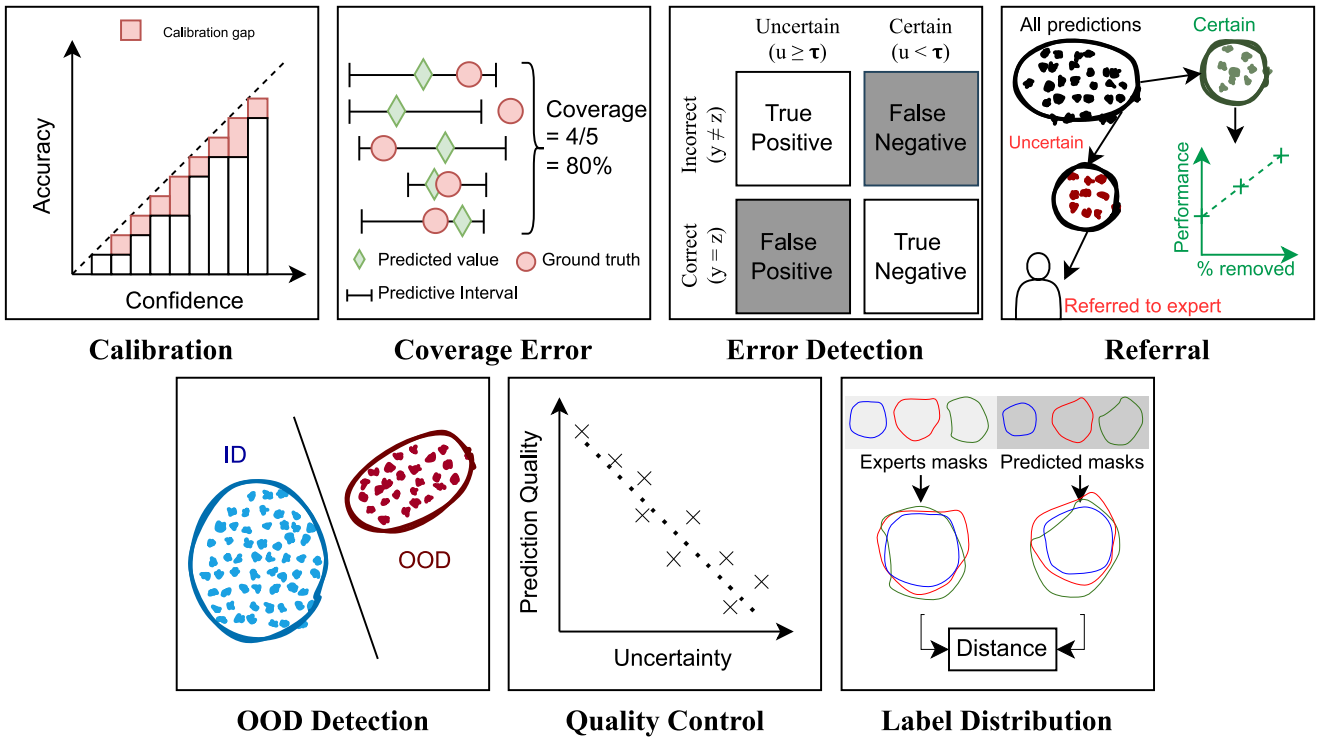

Fig. 14. Illustration of the different approaches used to estimate **quantitatively** the quality of uncertainty estimates. See text for details.

the model is computed on the remaining predictions. If uncertainty estimates efficiently identify uncertain cases that are more likely to be incorrect, then the error rate on the remaining prediction should decrease. Multiple fractions can be used, producing a curve showing the error rate of the model with respect to the fraction of rejected data. The area under the resulting curve is used as a qualitative score. This referral-based evaluation protocol aims at mimicking a human-in-the-loop process where the model abstains from uncertain predictions which are redirected to an expert for correction. It essentially highlights the same trends as the previous misclassification detection setting. Such metric was for instance used for MS lesions segmentation in brain MRI [98], cardiac MRI segmentation [177], brain stroke detection [178] or diabetic retinopathy detection [179]. Interestingly, such referral metrics were also used in the context of the QU-BraTS 2020 challenge focusing on UQ for brain tumor segmentation in MRI [180], as well as in the more recent SHIFT 2023 challenge focusing on MS lesions segmentation [181].

5.5. OOD detection protocol

A desired property of uncertainty is to be high in the context of a train–test mismatch, occurring when input images are significantly different from the images seen during training. Similarly to the misclassification detection setting, the uncertainty estimates can be translated into a binary classifier that aims at distinguishing between in-distribution (ID) and OOD images. Standard classification metrics can further be computed.

Protocols for OOD detection can be characterized based on the type of OOD data used. The most obvious setting corresponds to extreme OOD data, corresponding to samples that share little to no similarity with the training data. For instance, [41] proposes to train a model to segment COVID lesions in chest CT, and further use colon and spleen CT as OOD data. A more realistic shift in the context of medical image processing is diagnostic shifts, where a disease unobserved during training is included in the test images. This setting has been explored in Berger et al. [40] where authors train a binary classifier to distinguish between cardiomegaly and pneumothorax in chest X-ray, then use images with fracture as OOD. Similarly, in the context of digital pathology detection, [79,81] trained breast metastasis detection models and included images with new unseen abnormalities at test-time as OOD data. [182] trained an 8-class classifier on the ISIC 2019 dataset to detect skin disease, which also contains a test set of OOD images belonging to none of the 8 classes for OOD evaluation. Modality shifts can also be encountered in medical image processing tasks, where the imaging acquisition protocol is altered between training and testing images. Tardy et al. [121] train a 2D mammography classification model and use images acquired from a different manufacturer at test-time, simulating a common data shift encountered in clinical routine. [119] considered posteroanterior chest X-rays as training images and tested OOD detection on anteroposterior images. Finally, transformation shifts were also investigated, where transformations are applied to the input images to push them away from the training images distribution. For instance, [41] generated OOD data by applying affine transformations and synthetic artifacts to the input images.

5.6. Quality control

For segmentation tasks uncertainty is expected to be higher for poorly-segmented images than for well-segmented ones. Based on this desired property, several works studied the correlation between image-wise uncertainty scores (in contrast with the usual pixel-wise estimates) and the segmentation quality, such as the Dice score. In an automated medical image segmentation pipeline, this process can be used to detect images for which the produced segmentation does not meet quality standards. We refer to this mode of evaluation, specific to segmentation tasks, as QC-based evaluation protocols. In this setting, the goal is generally to maximize Pearson's correlation between the true segmentation quality and the proxy score [82,106,141,143,145,147,183]. However, we found that such QC-based protocols usually only focus on the correlation with the Dice score, which has been shown to be correlated to the size of the segmented object [184,185]. As a result, it is possible that these QC protocols demonstrate a correlation between the size of

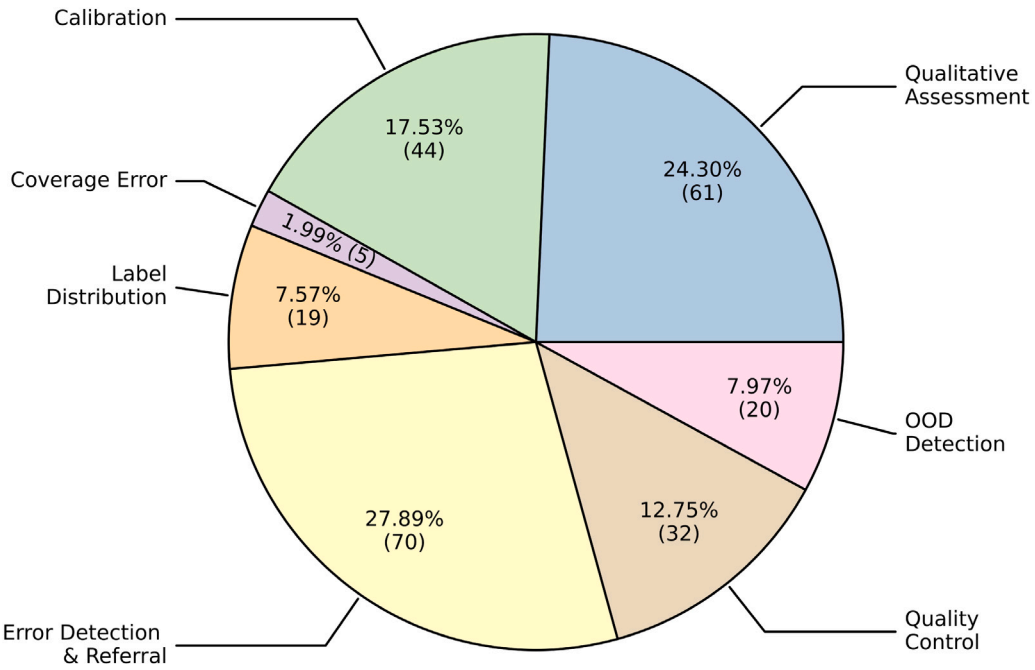

Fig. 15. Implemented UQ evaluation protocols in the reviewed papers. The percentage (and the number) of the reviewed papers per class is mentioned in the Pie chart.

the segmented region and the uncertainty level (with smaller regions being more uncertain), rather than the true quality of the segmentation. Studying the correlation with other segmentation metrics is only rarely envisaged, with an exception in Kushibar et al. [82] where authors demonstrate a correlation for both the Dice and Hausdorff metrics.

5.7. Label-distribution protocol

Finally, for segmentation tasks where several expert delineations are available per image, label-distribution metrics can be employed. This consists of comparing the predicted distribution of labels P_{out} with the ground truth distribution of the experts P_{gt} and is thus commonly used paired with Generative UQ models presented in Section 3.8. A popular choice of metric is the Generalized Energy Distance [109–111] between both distributions. Other proposed metrics include the normalized cross-entropy [97], the normalized cross-correlation [111] or the weighted mean of the predictive entropy between the true and predicted uncertainty maps [186].

The distribution of the UQ evaluation protocols in the studied corpus is shown in Fig. 15.

5.8. Distinguishing aleatoric and epistemic uncertainties during evaluation

While the distinction between aleatoric and epistemic uncertainties is possible when defining UQ approaches, this distinction is less clear when dealing with evaluation. Aleatoric and epistemic uncertainty estimates are often compared using the same metrics [37,38,126,173,187]. Some evaluation scenarios clearly focus on one particular side of uncertainty, for instance, OOD detection is clearly linked to epistemic uncertainty, while label distribution is akin to aleatoric uncertainty. Other metrics including Coverage Error, Error Detection, Referral, and Quality Control do not make any assumption about the source of error, which could come from a noisy data point (aleatoric uncertainty) or the lack of knowledge of the model (epistemic uncertainty), and hence cannot be clearly associated to one or the other. Calibration metrics, that consider the Softmax probabilities of the model, could be cast as a way to evaluate aleatoric uncertainty. However, there is an increasing literature on the calibration under domain-shift, which bridges the gap with epistemic uncertainty [188–191].

6. Discussion

In the first section, we reviewed the most popular UQ methods for DL-based medical image analysis. Then, we presented how structural uncertainty could be assessed to align with clinical interests. Finally, the various UQ evaluation protocols were investigated. In this section, we list the key insights of this review and identify potential future research directions.

First, the large number of studies incorporating UQ in their medical analysis pipeline proves that the need for UQ is well taken into account by the DL community. This shows that efforts are being made to develop AI tools that are not only powerful, but also useful in a real clinical setting. In this context, the predictive performance of the model only is not enough to reach a good acceptance. UQ is key to facilitate human–machine collaboration and break the 'black-box effect'.

UQ methods. Bayesian methodology, although providing a strong theoretical background for uncertainty, is scarcely implemented for medical image analysis (only 2.06% of the papers, see Fig. 2). This can be explained by the complex implementation that requires (i) the modification of the NN and (ii) the modification of the training paradigm. Less formal approximations of the Bayesian framework, such as dropout-based methods, are thus generally preferred. Overall, MC dropout method seems to be the most popular approach for UQ in medical image analysis, representing more than a third of the implemented methods (38.05%), considering both the standard MC dropout methods (34.81%) as well as MC dropout Ensemble models (3.24%), which are an MC-dropout extension. This popularity can be explained by its easy implementation in any NN trained with dropout, indeed a large majority of NNs. Additionally, dropout helps prevent overfitting during training, which is a common problem in the medical domain, where the training dataset size is limited. However, the performance of MC dropout is highly dependent on the applied dropout rate [70,192], which can make it impractical to tune. Moreover, it requires multiple inferences for the same input image, increasing the inference time, which may not be compatible with AI applications in clinics. Ensembling approaches are also commonly employed for UQ (16.22% of the implemented UQ methods), although less common than MC dropout models. Aggregating the predictions of multiple models is a well-known trick to improve predictive performance, while also

providing quality uncertainty estimates. The drawback is an increased computational cost and time, as it requires multiple training and the aggregation of their predictions at testing. Other popular UQ metrics is Softmax probability (12.98%), which provides intuitive and easy-to-use uncertainty estimates, and learned uncertainty methods (8.26%) that propose to learn aleatoric uncertainty from the data. Finally, we note (see Fig. 2) that most implemented methods simultaneously estimate aleatoric and epistemic uncertainties (61.95%), followed by aleatoric-only (32.74%) and epistemic-only (4.13%).

For segmentation tasks, structural uncertainty has been proposed to align uncertainty quantification with medical attention. In clinical routine, clinicians may be more interested in knowing which automatically detected region (e.g. individual lesion) can be trusted, rather than knowing which pixels/voxels may be inaccurate. Several leads have been proposed in this direction, either resorting to the aggregation of pixel-wise estimates or using an auxiliary model (radiomics-based ML model or DL model) to produce region-wise uncertainty scores.

UQ evaluation protocols. In the literature, a large variety of evaluation protocols are reported, aiming at assessing the quality of uncertainty estimates. In the context of medical image segmentation, if multiple manual expert delineations are available for a given input image, the inter-rater variability can be used as ground truth uncertainty, to be compared with the one predicted (representing 7.57% of the implemented evaluation protocols, see Fig. 15). However, most of the time, such an uncertainty gold standard is not provided. Thus, the evaluation of UQ usually relies on proxy tasks, such as the detection of errors (27.89%), Quality Control (12.75%), or Out-of-distribution (7.97%). These methods are inspired by concrete applications of uncertainty in a real-world scenario. Yet, although commonly used, UQ evaluation based on error detection is not ideal for ranking methods. Indeed, the set of correct and incorrect predictions is specific to each predictive model. It is then inappropriate to compare them directly [193]. Calibration evaluation metrics are also commonly used (17.53%). The use of such metrics seems particularly interesting because many popular uncertainty estimates, such as variance, entropy, or mutual information, can be directly extracted from probability distributions. Thus, guaranteeing that the probability estimates are well-calibrated seems to be essential to obtain meaningful uncertainty estimates.

Finally, it must be acknowledged that the effort of the community is promoted by the organization of uncertainty-oriented challenges. The 2020 edition of the BraTS challenge included an uncertainty task (QU-BraTS) asking participants to provide brain tumor segmentation models that are able to provide voxel-wise uncertainty estimates correlating with segmentation errors [180,194]. The MICCAI QUBIQ challenge,[1] hosted in 2020 and 2021, focused on label uncertainty. Participants were provided with images annotated by several experts, on a variety of medical image segmentation tasks, and were asked to develop methods able to reproduce the rater's annotation distribution. Finally, the SHIFT 2023 challenge[2] contained a task of uncertainty quantification for Multiple Sclerosis lesions segmentation [181]. The challenge focused on the development of models robust to train–test mismatches and uncertainty was evaluated through an error detection setting.

6.1. Future directions

Based on our review, we suggest several future research directions for UQ in DL-based medical image analysis.

First, the majority (67.83%) of the implemented UQ methods are based on a sampling protocol (MC dropout, Deep Ensemble, BNN, MC dropout Ensemble, TTA, and Generative models), aiming at generating multiple predictions for the same query input. Yet, this process may significantly increase the computational burden of UQ, especially when processing large 3D volumes, which may prevent its adoption in an automated pipeline in the medical domain. Single-step UQ methods such as Softmax (calibrated) probabilities, Evidential Deep Learning, or Features-based methods are thus promising especially for time-critical applications.

Overall, the detection of Out-of-distribution (OOD) predictions using uncertainty concerns few studies (7.97% of UQ evaluation protocols), despite being crucial in real-world medical scenarios. In an automated medical image pipeline, input samples can exhibit various anomalies or be inappropriate for the correct functioning of the NN, thus resulting in very poor predictions. Real clinical cases may include artifacts, present a pathology unseen during training, or an unusual imaging contrast due to a particular acquisition protocol. In such situations, uncertainty associated with the computed predictions is expected to be high and should represent a warning to the user (*e.g.*, the medical practitioner). In practice, this is usually not the case with standard approaches such as Softmax uncertainty, MC dropout, or Deep Ensemble, which have limited performance in terms of OOD detection [195]. The rarity of OOD-based evaluation protocol in the reviewed corpus may be explained by the difficulty of gathering relevant OOD data representative of real-world scenarios. A potential lead is to use Data Augmentation, which allows to generate OOD samples from ID samples by introducing well-controlled intensity noise or artefacts in the image [41,62]. Alternatively, OOD detection is not only studied from an uncertainty point-of-view. For example, Unsupervised Anomaly Detection [196,197] uses the reconstruction error of Auto-Encoder models for OOD detection, and did not fall under the scope of this review.

Experimental results outside the medical imaging field suggest that the quality of uncertainty estimates degrades in the presence of domain shift [190,198,199]. Thus, high-quality uncertainty estimates on in-domain data may not generalize well in the presence of train–test mismatches. In the context of medical image processing, this major limitation of UQ is only rarely studied. In Murugesan et al. [36], authors explore the deterioration of probability calibration with increasing levels of corruption on the input images. Similarly, Thagaard et al. [79] demonstrated that the ability to detect errors from uncertainty estimates is reduced in the presence of train–test mismatches. Finally, the SHIFT 2023 challenge evaluates the robustness of uncertainty estimates on a hidden test dataset not used for training. We argue that the robustness of uncertainty estimates under domain shift should be more commonly included in UQ evaluation protocols, especially in the context of medical image processing where domain shifts are frequent.

Finally, while the need for UQ in medical applications is unquestionable, we argue that being able to understand the prediction process of the DL model is also crucial to promote a trustable usage of AI in medicine. Then, the link between explainability and uncertainty should be studied, which would allow to understand both how the prediction is made and whether or not it should be trusted. An interesting research direction would be to complement uncertainty estimates with explanations [200], helping the user to understand the sources of uncertainty in an intelligible way and possibly contribute to its improvement.

7. Conclusion

In this review, we have proposed an overview of the most popular UQ methods implemented in DL-based medical image applications, a specific domain with inherent uncertainty. Numerous phenomena can cause predictive uncertainty, such as noisy images, imperfect ground truth labels, lack of or incomplete data, and inter-site image variability. The literature proposes various methods to quantify this uncertainty which are applied to a very large range of medical image applications. As demonstrated in this review, developing trustable AI solutions integrating uncertainty quantification of the computed predictions is an active research topic.

[1] https://qubiq21.grand-challenge.org/.
[2] https://shifts.grand-challenge.org/.

Table 1
List of reviewed papers according to the uncertainty framework. The same study can be present in different rows, if several uncertainty approaches have been compared.

Uncertainty Frameworks	Count	Studies
Monte Carlo dropout	118	[37,38,58,79,106,140,141,143,147,183,201–206] [40,42,78,81,144,154,156,173,174,207–213] [39,73,123,133,152,157,171,175,178,179,204,214–218] [80,99,108,126,165,172,177,186,219–226] [60,67–69,92,121,149,153,161,182,227–232] [41,62,70,71,74,83,107,137,148,162,163,167,233–235] [50,51,63,101,104,138,176,236–244] [59,150,151,187,245–247]
Deep ensemble	55	[38,40,58,78,79,81,145,160,183,210,216,248–250] [80,82,92,96,126,130,131,133,136,228,229,251–255] [51,63,83,86,87,90,138,176,233,239,242,256–258] [59,150,187,243,245–247,259–262]
Softmax	44	[36–38,40,42,58,159,173,175,183,210,217,250,263,264] [39,41,43,51,138,221,228,238,239,265–270] [50,104,242,243,246,271–279]
Learned uncertainty	28	[37–39,95,97,98,132,155,173,213,280–284] [61,94,99–102,104,166,230,285–288]
TTA	24	[37,41,105–108,133,146,170,172,182,216,224,233,250] [50,59,101,138,142,187,240,244,289]
Generative models	15	[37,109–112,116,215] [62,113–115,117,126,173,290]
Features	14	[37,40,41,119,121,123,175] [120,122,124,238,239,242,291]
Evidential deep learning	14	[121,126,128,130–132] [129,158,168,237,247,292–294]
Dropout ensemble	11	[58,80,92,212,228,229] [59,93,150,176,295]
Bayesian neural networks	7	[58,60,62,296] [59,63,297]
Conformal prediction	5	[49,51,298] [50,299]
Other	4	[38,134] [133,135]
Total	**338**	

Table 2
List of reviewed papers according to the uncertainty evaluation paradigm. The same study can be present in different rows, if several evaluation strategies have been used.

Evaluation frameworks	Count	Studies
Error detection - Referral	70	[37,38,42,58,79,106,140,145,173,174,202–205,211,212] [73,80,98,134,136,171,175,178,179,204,213,216,217,222,250,252] [92,105,121,126,130,131,172,177,182,225,226,228–231,254] [70,74,87,93,100–102,107,138,142,176,234,256,286,287,297] [59,63,243,247,277]
Qualitative assessment	61	[128,144,154–156,159,170,201,206,207,209,248,249,281,282] [68,99,152,157,160,165,214,218–220,253,264,265,283,284] [60,67,94,137,153,161,162,167,232,233,255,267,285,296] [71,129,158,163,168,235,236,241,271,272] [166,257,273,278,292,294,295]
Quality control	32	[38,39,69,73,78,82,106,141,143,145–147,173,183,224,280] [101,148–151,240,244,245,259–262,288–291]
Calibration	44	[36,38,40,42,78,79,123,132,133,140,173,208,210,212,250,263] [43,62,80,82,83,86,90,108,126,177,178,221,223,237,268,269] [59,101,243,245,247,258,270,274–277,279]
OOD detection	20	[37,40,41,79,81,83,119,121,123,182] [104,120,122,124,238,239,242,246,291,293]
Label distribution	19	[95,97,109–112,116,186,215,251] [61,96,113–115,117,135,187,266]
Coverage error	5	[49,51,298] [50,53]

CRediT authorship contribution statement

Benjamin Lambert: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Florence Forbes:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Conceptualization. **Senan Doyle:** Writing – review & editing, Writing – original draft, Funding acquisition. **Harmonie Dehaene:** Writing – review & editing, Writing – original draft. **Michel Dojat:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Benjamin Lambert reports financial support was provided by Pixyl. Senan Doyle reports financial support was provided by Pixyl. Harmonie Dehaene reports financial support was provided by Pixyl. Michel Dojat reports a relationship with Pixyl that includes: board membership. Florence Forbes reports a relationship with Pixyl that includes: board membership.

Acknowledgments

Benjamin Lambert is supported by a CIFRE convention (ANRT 2020/1555).

Appendix. Classification of the reviewed papers

This appendix presents the classification of the reviewed papers according to the implemented uncertainty framework (Table 1) and the uncertainty evaluation paradigm (Table 2).

References

[1] Puttagunta M, Ravi S. Medical image analysis based on deep learning approach. Multimedia Tools Appl 2021;80(16):24365–98.
[2] Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: International conference on machine learning. 2017, p. 1321–30.
[3] Nguyen A, Yosinski J, Clune J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2015, p. 427–36.
[4] Ma X, Niu Y, Gu L, Wang Y, Zhao Y, Bailey J, Lu F. Understanding adversarial attacks on deep learning based medical image analysis systems. Pattern Recognit 2021;110:107332.
[5] Ford RA, Price W, Nicholson I. Privacy and accountability in black-box medicine. Mich Telecomm Technol Law Rev 2016;23:1.
[6] Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, Fieguth P, Cao X, Khosravi A, Acharya UR, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Inf Fusion 2021;76:243–97.
[7] Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What clinicians want: contextualizing explainable machine learning for clinical end use. In: Machine learning for healthcare conference. 2019, p. 359–80.
[8] Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Mach Learn 2021;110(3):457–506.
[9] Gawlikowski J, Tassi CRN, Ali M, Lee J, Humt M, Feng J, Kruspe A, Triebel R, Jung P, Roscher R, et al. A survey of uncertainty in deep neural networks. Artif Intell Rev 2023;56(Suppl 1):1513–89.

[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(2):29–48.

[11] Wang H, Yeung D-Y. A survey on Bayesian deep learning. ACM Comput Surv 2020;53(5):1–37.

[12] Kabir HD, Khosravi A, Hosen MA, Nahavandi S. Neural network-based uncertainty quantification: A survey of methodologies and applications. IEEE Access 2018;6:36218–34.

[13] Zhou X, Liu H, Pourpanah F, Zeng T, Wang X. A survey on epistemic (model) uncertainty in supervised learning: Recent advances and applications. Neurocomputing 2022;489:449–65.

[14] Kurz A, Hauser K, Mehrtens HA, Krieghoff-Henning E, Hekler A, Kather JN, Fröhling S, von Kalle C, Brinker TJ, et al. Uncertainty estimation in medical image classification: Systematic review. JMIR Med Inform 2022;10(8):e36427.

[15] Loftus TJ, Shickel B, Ruppert MM, Balch JA, Ozrazgat-Baslanti T, Tighe PJ, Efron PA, Hogan WR, Rashidi P, Upchurch Jr GR, et al. Uncertainty-aware deep learning in healthcare: a scoping review. PLoS Digit Health 2022;1(8):e0000085.

[16] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2017, p. 4700–8.

[17] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. 2019, p. 6105–14.

[18] Dai Y, Gao Y, Liu F. Transmed: Transformers advance multi-modal medical image classification. Diagnostics 2021;11(8):1384.

[19] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. 2015, p. 234–41.

[20] Kerfoot E, Clough J, Oksuz I, Lee J, King AP, Schnabel JA. Left-ventricle quantification using residual u-net. In: International workshop on statistical atlases and computational models of the heart. 2018, p. 371–80.

[21] 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, 2016, p. 565–71.

[22] Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla NY, Kainz B, et al. Attention u-net: Learning where to look for the pancreas. Med Imaging Deep Learn 2018.

[23] Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. Nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021;18(2):203–11.

[24] Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman B, Roth HR, Xu D. Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2022, p. 574–84.

[25] Hatamizadeh A, Nath V, Tang Y, Yang D, Roth HR, Xu D. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries: 7th international workshop, brainLes 2021, held in conjunction with MICCAI 2021, virtual event, September 27 2021, revised selected papers, part i. 2022, p. 272–84.

[26] Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, et al. Transunet: Transformers make strong encoders for medical image segmentation. 2021, arXiv preprint arXiv:2102.04306.

[27] Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell 2020;42(2):318–27.

[28] Fidon L, Li W, Garcia-Peraza-Herrera LC, Ekanayake J, Kitchen N, Ourselin S, Vercauteren T. Generalised wasserstein dice score for imbalanced multi-class segmentation using holistic convolutional networks. In: International MICCAI brainlesion workshop. 2017, p. 64–76.

[29] Salehi SSM, Erdogmus D, Gholipour A. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In: International workshop on machine learning in medical imaging. 2017, p. 379–87.

[30] Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision? In: Advances in neural information processing systems 30: annual conference on neural information processing systems 2017. 2017, p. 5574–84.

[31] Xue Z, Yang F, Rajaraman S, Zamzmi G, Antani S. Cross dataset analysis of domain shift in cxr lung region detection. Diagnostics 2023;13(6):1068.

[32] Becker AS, Chaitanya K, Schawkat K, Muehlematter UJ, Hötker AM, Konukoglu E, Donati OF. Variability of manual segmentation of the prostate in axial t2-weighted mri: A multi-reader study. Eur J Radiol 2019;121:108716.

[33] Joskowicz L, Cohen D, Caplan N, Sosna J. Inter-observer variability of manual contour delineation of structures in ct. Eur Radiol 2019;29(3):1391–9.

[34] Kumar A, Liang P, Ma T. Verified uncertainty calibration. In: Advances in neural information processing systems 32: annual conference on neural information processing systems 2019. 2019, p. 3787–98.

[35] Kull M, Perello Nieto M, Kängsepp M, Silva Filho T, Song H, Flach P. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. Adv Neural Inf Process Syst 2019;32.

[36] Murugesan B, Liu B, Galdran A, Ayed IB, Dolz J. Calibrating segmentation networks with margin-based label smoothing. Med Image Anal 2023;102826.

[37] Diao Z, Jiang H, Shi T. A unified uncertainty network for tumor segmentation using uncertainty cross entropy loss and prototype similarity. Knowl-Based Syst 2022;246:108739.

[38] Jungo A, Balsiger F, Reyes M. Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. Front Neurosci 2020;282.

[39] DeVries T, Taylor GW. Leveraging uncertainty estimates for predicting segmentation quality. 2018, arXiv e-prints.

[40] Berger C, Paschali M, Glocker B, Kamnitsas K. Confidence-based out-of-distribution detection: a comparative study and analysis. In: Uncertainty for safe utilization of machine learning in medical imaging, and perinatal imaging, placental and preterm image analysis. 2021, p. 122–32.

[41] González C, Gotkowski K, Fuchs M, Bucher A, Dadras A, Fischbach R, Kaltenborn IJ, Mukhopadhyay A. Distance-based detection of out-of-distribution silent failures for covid-19 lung lesion segmentation. Med Image Anal 2022;82:102596.

[42] Carneiro G, Pu LZCT, Singh R, Burt A. Deep learning uncertainty and confidence calibration for the five-class polyp classification from colonoscopy. Med Image Anal 2020;62:101653.

[43] Liang G, Zhang Y, Jacobs N. Neural network calibration for medical imaging classification using dca regularization. In: International conference on machine learning, workshop on uncertainty and robustness in deep learning. 2020.

[44] Vovk V, Gammerman A, Shafer G. Algorithmic learning in a random world, Vol. 29. Springer; 2005.

[45] Angelopoulos AN, Bates S. Conformal prediction: A gentle introduction. Found Trends® Mach Learn 2023;16(4):494–591.

[46] Angelopoulos A, Bates S, Malik J, Jordan MI. Uncertainty sets for image classifiers using conformal prediction. 2020, arXiv preprint arXiv:2009.14193.

[47] Romano Y, Patterson E, Candes E. Conformalized quantile regression. Adv Neural Inf Process Syst 2019;32.

[48] Alvarsson J, McShane SA, Norinder U, Spjuth O. Predicting with confidence: using conformal prediction in drug discovery. J Pharm Sci 2021;110(1):42–9.

[49] Csillag D, Paes LM, Ramos T, Romano JV, Schuller R, Seixas RB, Oliveira RI, Orenstein P. Amnioml: amniotic fluid segmentation and volume prediction with uncertainty quantification. Proc AAAI Conf Artif Intell 2023;37(13):15494–502.

[50] Lambert B, Forbes F, Doyle S, Dojat M. Triadnet: sampling-free predictive intervals for lesional volume in 3d brain MR images. In: Uncertainty for safe utilization of machine learning in medical imaging - 5th international workshop, UNSURE 2023, held in conjunction with MICCAI 2023, Vancouver, BC, Canada, October 12 2023, proceedings. Vol. 14291, 2023, p. 32–41.

[51] Eaton-Rosen Z, Varsavsky T, Ourselin S, Cardoso MJ. As easy as 1 2. 4? uncertainty in counting tasks for medical imaging. In: Medical image computing and computer assisted intervention–MICCAI 2019: 22nd international conference, Shenzhen, China, October (2019) 13–17, proceedings, part IV 22. 2019, p. 356–64.

[52] Zhang Y, Wang S, Zhang Y, Chen DZ. Rr-cp: Reliable-region-based conformal prediction for trustworthy medical image classification. In: International workshop on uncertainty for safe utilization of machine learning in medical imaging. 2023, p. 12–21.

[53] Mehrtens H, Bucher T, Brinker TJ. Pitfalls of conformal predictions for medical image classification. In: International workshop on uncertainty for safe utilization of machine learning in medical imaging. 2023, p. 198–207.

[54] Barber RF, Candes EJ, Ramdas A, Tibshirani RJ. Conformal prediction beyond exchangeability. Ann Statist 2023;51(2):816–45.

[55] Blundell C, Cornebise J, Kavukcuoglu K, Wierstra D. Weight uncertainty in neural network. In: International conference on machine learning. 2015, p. 1613–22.

[56] Shridhar K, Laumann F, Liwicki M. A comprehensive guide to Bayesian convolutional neural network with variational inference. 2019, arXiv preprint arXiv:1901.02731.

[57] Gal Y, et al. Uncertainty in deep learning (Ph.D. thesis), University of Cambridge; 2016.

[58] Filos A, Farquhar S, Gomez AN, Rudner TG, Kenton Z, Smith L, Alizadeh M, De Kroon A, Gal Y. A systematic comparison of Bayesian deep learning robustness in diabetic retinopathy tasks. In: 4th workshop on Bayesian deep learning (NeurIPS 2019), Vancouver, Canada. 2019.

[59] Mehrtens HA, Kurz A, Bucher T-C, Brinker TJ. Benchmarking common uncertainty estimation methods with histopathological images under domain shift and label noise. Med Image Anal 2023;89:102914.

[60] Dhakal P, Joshi SR. Uncertainty estimation in detecting knee abnormalities on mri using Bayesian deep learning. In: Proceedings of 10th IOE graduate conference, Vol. 10. 2021.

[61] Li H, Luo H. Uncertainty quantification in medical image segmentation. In: 2020 IEEE 6th international conference on computer and communications. ICCC, 2020, p. 1936–40.

[62] Fuchs M, Gonzalez C, Mukhopadhyay A. Practical uncertainty quantification for brain tumor segmentation. Med Imaging Deep Learn 2021.

[63] Adams J, Elhabian SY. Benchmarking scalable epistemic uncertainty quantification in organ segmentation. In: International workshop on uncertainty for safe utilization of machine learning in medical imaging. 2023, p. 53–63.

[64] Dusenberry M, Jerfel G, Wen Y, Ma Y, Snoek J, Heller K, Lakshminarayanan B, Tran D. Efficient and scalable Bayesian neural nets with rank-1 factors. In: International conference on machine learning. 2020, p. 2782–92.

[65] Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: International conference on machine learning. 2016, p. 1050–9.

[66] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15(1):1929–58.

[67] Kwon Y, Won J-H, Kim BJ, Paik MC. Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation. Comput Statist Data Anal 2020;142:106816.

[68] Jungo A, McKinley R, Meier R, Knecht U, Vera L, Pérez-Beteta J, Molina-García D, Pérez-García VM, Wiest R, Reyes M. Towards uncertainty-assisted brain tumor segmentation and survival prediction. In: International MICCAI brainlesion workshop. 2017, p. 474–85.

[69] Orlando JI, Seeböck P, Bogunović H, Klimscha S, Grechenig C, Waldstein S, Gerendas BS, Schmidt-Erfurth U. U2-net: A Bayesian u-net model with epistemic uncertainty feedback for photoreceptor layer segmentation in pathological oct scans. In: 2019 IEEE 16th international symposium on biomedical imaging. ISBI 2019, 2019, p. 1441–5.

[70] Camarasa R, Bos D, Hendrikse J, Nederkoorn P, Kooi E, van der Lugt A, de Bruijne M. 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. 2020, p. 32–41.

[71] Ghoshal B, Tucker A, Sanghera B, Wong WL. Estimating uncertainty in deep learning for reporting confidence to clinicians when segmenting nuclei image data. In: 2019 IEEE 32nd international symposium on computer-based medical systems. CBMS, 2019, p. 318–24.

[72] Gal Y, Hron J, Kendall A. Concrete dropout. Adv Neural Inf Process Syst 2017;30.

[73] McClure P, Rho N, Lee JA, Kaczmarzyk JR, Zheng CY, Ghosh SS, Nielson DM, Thomas AG, Bandettini P, Pereira F. Knowing what you know in brain segmentation using Bayesian deep neural networks. Front Neuroinform 2019;13:67.

[74] 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):1–14.

[75] Wan L, Zeiler M, Zhang S, Le Cun Y, Fergus R. Regularization of neural networks using dropconnect. In: International conference on machine learning. 2013, p. 1058–66.

[76] Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Annual conference on neural information processing systems 2017. 2017, p. 6402–13.

[77] Malinin A, Gales M. Uncertainty estimation in autoregressive structured prediction. In: International conference on learning representations. 2020.

[78] Mehrtash A, Wells WM, Tempany CM, Abolmaesumi P, Kapur T. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE Trans Med Imaging 2020;39(12):3868–78.

[79] Thagaard J, Hauberg S, Vegt Bvd, Ebstrup T, Hansen JD, Dahl AB. Can you trust predictive uncertainty under real dataset shifts in digital pathology? In: International conference on medical image computing and computer-assisted intervention. 2020, p. 824–33.

[80] Asgharnezhad H, Shamsi A, Alizadehsani R, Khosravi A, Nahavandi S, Sani ZA, Srinivasan D, Islam SMS. Objective evaluation of deep uncertainty predictions for covid-19 detection. Sci Rep 2022;12(1):1–11.

[81] Linmans J, van der Laak J, Litjens G. Efficient out-of-distribution detection in digital pathology using multi-head convolutional neural networks. Med Imaging Deep Learn 2020;465–78.

[82] Kushibar K, Campello V, Garrucho L, Linardos A, Radeva P, Lekadir K. Layer ensembles: A single-pass uncertainty estimation in deep learning for segmentation. In: Medical image computing and computer assisted intervention–MICCAI 2022: 25th international conference, Singapore, September 18–22, 2022, proceedings, part VIII. 2022, p. 514–24.

[83] Qendro L, Campbell A, Lio P, Mascolo C. Early exit ensembles for uncertainty quantification. Mach Learn Health 2021;181–95.

[84] Zhao G, Liu F, Oler JA, Meyerand ME, Kalin NH, Birn RM. Bayesian convolutional neural network based mri brain extraction on nonhuman primates. Neuroimage 2018;175:32–44.

[85] Maddox WJ, Izmailov P, Garipov T, Vetrov DP, Wilson AG. A simple baseline for Bayesian uncertainty in deep learning. Adv Neural Inf Process Syst 2019;32.

[86] Lindén M, Garifullin A, Lensu L. Weight averaging impact on the uncertainty of retinal artery-venous 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. 2020, p. 52–60.

[87] Liu Y, Zhao C, Rubin J. Uncertainty quantification in chest x-ray image classification using Bayesian deep neural networks. In: KDH@ ECAI. 2020, p. 19–26.

[88] Granitto PM, Verdes PF, Ceccatto HA. Neural network ensembles: evaluation of aggregation algorithms. Artificial Intelligence 2005;163(2):139–62.

[89] Kuncheva LI, Whitaker CJ. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Mach Learn 2003;51(2):181.

[90] Larrazabal AJ, Martínez C, Dolz J, Ferrante E. Orthogonal ensemble networks for biomedical image segmentation. In: Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1 2021, proceedings, part III 24. 2021, p. 594–603.

[91] Georgescu M-I, Ionescu RT, Miron AI. Diversity-promoting ensemble for medical image segmentation. In: The 38th ACM/SIGAPP symposium on applied computing. 2022.

[92] Abdar M, Samami M, Mahmoodabad SD, Doan T, Mazoure B, Hashemifesharaki R, Liu L, Khosravi A, Acharya UR, Makarenkov V, et al. Uncertainty quantification in skin cancer classification using three-way decision-based Bayesian deep learning. Comput Biol Med 2021;135:104418.

[93] Abdar M, Salari S, Qahremani S, Lam H-K, Karray F, Hussain S, Khosravi A, Acharya UR, Makarenkov V, Nahavandi S. Uncertaintyfusenet: Robust uncertainty-aware hierarchical feature fusion model with ensemble monte carlo dropout for covid-19 detection. Inf Fusion 2023;90:364–81.

[94] Ji W, Chen W, Yu S, Ma K, Cheng L, Shen L, Zheng Y. Uncertainty quantification for medical image segmentation using dynamic label factor allocation among multiple raters. In: MICCAI on QUBIQ workshop. 2020.

[95] Cetindag SC, Yergin M, Alis D, Oksuz I. Meta-learning for medical image segmentation uncertainty quantification. In: International MICCAI brainlesion workshop. 2022, p. 578–84.

[96] Yang Y, Guo X, Pan Y, Shi P, Lv H, Ma T. Uncertainty quantification in medical image segmentation with multi-decoder u-net. In: International MICCAI brainlesion workshop. 2022, p. 570–7.

[97] Hu S, Worrall D, Knegt S, Veeling B, Huisman H, Welling M. Supervised uncertainty quantification for segmentation with multiple annotations. In: International conference on medical image computing and computer-assisted intervention. 2019, p. 137–45.

[98] Nair T, Precup D, Arnold DL, Arbel T. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. Med Image Anal 2020;59:101557.

[99] Eaton-Rosen Z, Bragman F, Bisdas S, Ourselin S, Cardoso MJ. Towards safe deep learning: accurately quantifying biomarker uncertainty in neural network predictions. In: International conference on medical image computing and computer-assisted intervention. 2018, p. 691–9.

[100] Graham MS, Sudre CH, Varsavsky T, Tudosiu P-D, Nachev P, Ourselin S, Cardoso MJ. Hierarchical brain parcellation with uncertainty. 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. 2020, p. 23–31.

[101] Judge T, Bernard O, Cho Kim W-J, Gomez A, Chartsias A, Jodoin P-M. Asymmetric contour uncertainty estimation for medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. 2023, p. 210–20.

[102] McKinley R, Rebsamen M, Daetwyler K, Meier R, Radojewski P, Wiest R. Uncertainty-driven refinement of tumor-core segmentation using 3d-to-2d networks with label uncertainty. In: International MICCAI brainlesion workshop. 2020, p. 401–11.

[103] Liu Z, Wang Z, Liang PP, Salakhutdinov RR, Morency L-P, Ueda M. Deep gamblers: Learning to abstain with portfolio theory. Adv Neural Inf Process Syst 2019;32.

[104] Bungert TJ, Kobelke L, Jäger PF. Understanding silent failures in medical image classification. In: International conference on medical image computing and computer-assisted intervention. 2023, p. 400–10.

[105] Ayhan MS, Berens P. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In: International conference on medical imaging with deep learning. 2018.

[106] 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.

[107] Ballestar LM, Vilaplana V. Mri brain tumor segmentation and uncertainty estimation using 3d-unet architectures. In: Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries: 6th international workshop, brainLes 2020, held in conjunction with MICCAI 2020, Lima, Peru, October 4 2020, revised selected papers, part i 6. 2021, p. 376–90.

[108] Javadi G, Bayat S, Kazemi Esfeh MM, Samadi S, Sedghi A, Sojoudi S, Hurtado A, Chang S, Black P, Mousavi P, et al. Towards targeted ultrasound-guided prostate biopsy by incorporating model and label uncertainty in cancer detection. Int J Comput Assist Radiol Surg 2022;17(1):121–8.

[109] Kohl S, et al. A probabilistic u-net for segmentation of ambiguous images. In: Advances in neural information processing systems 31: annual conference on neural information processing systems 2018. 2018, p. 6965–75.

[110] Kohl SA, Romera-Paredes B, Maier-Hein KH, Rezende DJ, Eslami S, Kohli P, Zisserman A, Ronneberger O. A hierarchical probabilistic u-net for modeling multi-scale ambiguities. 2019, arXiv e-prints.

[111] Baumgartner CF, Tezcan KC, Chaitanya K, Hötker AM, Muehlematter UJ, Schawkat K, Becker AS, Donati O, Konukoglu E. Phiseg: Capturing uncertainty in medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. 2019, p. 119–27.

[112] Gantenbein M, Erdil E, Konukoglu E. Revphiseg: A memory-efficient neural network for uncertainty quantification in medical image segmentation. In: Uncertainty for safe utilization of machine learning in medical imaging, and graphs in biomedical image analysis. 2020, p. 13–22.

[113] Valiuddin MA, Viviers CG, van Sloun RJ, de With PH, van der Sommen F. Improving aleatoric uncertainty quantification in multi-annotated medical image segmentation with normalizing flows. In: Uncertainty for safe utilization of machine learning in medical imaging, and perinatal imaging, placental and preterm image analysis: 3rd international workshop, UNSURE 2021, and 6th international workshop, PIPPI 2021, held in conjunction with MICCAI 2021, Strasbourg, France, October 1 2021, proceedings 3. 2021, p. 75–88.

[114] Viviers CG, Valiuddin AM, de With PH, van der Sommen F. Probabilistic 3d segmentation for aleatoric uncertainty quantification in full 3d medical data. In: Medical imaging 2023: computer-aided diagnosis, Vol. 12465. 2023, p. 343–53.

[115] Selvan R, Faye F, Middleton J, Pai A. Uncertainty quantification in medical image segmentation with normalizing flows. In: International workshop on machine learning in medical imaging. 2020, p. 80–90.

[116] Monteiro M, Le Folgoc L, Coelho de Castro D, Pawlowski N, Marques B, Kamnitsas K, van der Wilk M, Glocker B. Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. Adv Neural Inf Process Syst 2020;33:12756–67.

[117] Amit T, Shichrur S, Shaharabany T, Wolf L. Annotator consensus prediction for medical image segmentation with diffusion models. In: Medical image computing and computer assisted intervention – MICCAI 2023. 2023, p. 544–54.

[118] Postels J, Segu M, Sun T, Van Gool L, Yu F, Tombari F. On the practicality of deterministic epistemic uncertainty. In: International conference on machine learning. 2021.

[119] Calli E, Van Ginneken B, Sogancioglu E, Murphy K. Frodo: An in-depth analysis of a system to reject outlier samples from a trained neural network. IEEE Trans Med Imaging 2022;42(4):971–81.

[120] Anthony H, Kamnitsas K. On the use of mahalanobis distance for out-of-distribution detection with neural networks for medical imaging. In: International workshop on uncertainty for safe utilization of machine learning in medical imaging. 2023, p. 136–46.

[121] Tardy M, Scheffer B, Mateus D. Uncertainty measurements for the reliable classification of mammograms. In: International conference on medical image computing and computer-assisted intervention. 2019, p. 495–503.

[122] Woodland M, Patel N, Al Taie M, Yung JP, Netherton TJ, Patel AB, Brock KK. Dimensionality reduction for improving out-of-distribution detection in medical image segmentation. In: International workshop on uncertainty for safe utilization of machine learning in medical imaging. 2023, p. 147–56.

[123] Karimi D, Gholipour A. Improving calibration and out-of-distribution detection in deep models for medical image segmentation. IEEE Trans Artif Intell 2022.

[124] Lambert B, Forbes F, Doyle S, Dojat M. Multi-layer aggregation as a key to feature-based ood detection. In: International workshop on uncertainty for safe utilization of machine learning in medical imaging. 2023, p. 104–14.

[125] Dempster AP. A generalization of Bayesian inference. J R Stat Soc Ser B Stat Methodol 1968;30(2):205–32.

[126] Zou K, Yuan X, Shen X, Wang M, Fu H. Tbrats: Trusted brain tumor segmentation. In: International conference on medical image computing and computer-assisted intervention, Vol. 13438. 2022, p. 503–13.

[127] Sensoy M, Kaplan LM, Kandemir M. Evidential deep learning to quantify classification uncertainty. In: Advances in neural information processing systems 31: annual conference on neural information processing systems 2018. 2018, p. 3183–93.

[128] Huang L, Ruan S, Decazes P, Denoeux T. Evidential segmentation of 3d pet/ct images. In: International conference on belief functions. 2021, p. 159–67.

[129] Huang L, Ruan S, Denoeux T. Belief function-based semi-supervised learning for brain tumor segmentation. In: 2021 IEEE 18th international symposium on biomedical imaging. ISBI, 2021, p. 160–4.

[130] Ghesu FC, Georgescu B, Gibson E, Guendel S, Kalra MK, Singh R, Digumarthy SR, Grbic S, Comaniciu D. Quantifying and leveraging classification uncertainty for chest radiograph assessment. In: International conference on medical image computing and computer-assisted intervention. 2019, p. 676–84.

[131] Ghesu FC, Georgescu B, Mansoor A, Yoo Y, Gibson E, Vishwanath R, Balachandran A, Balter JM, Cao Y, Singh R, et al. Quantifying and leveraging predictive uncertainty for medical image assessment. Med Image Anal 2021;68:101855.

[132] Dawood T, Chan E, Razavi R, King AP, Puyol-Anton E. Addressing deep learning model calibration using evidential neural networks and uncertainty-aware training. In: 2023 IEEE 20th international symposium on biomedical imaging. ISBI, 2023.

[133] Jensen MH, Jørgensen DR, Jalaboi R, Hansen ME, Olsen MA. Improving uncertainty estimation in convolutional neural networks using inter-rater agreement. In: Medical image computing and computer assisted intervention–MICCAI 2019: 22nd international conference, Shenzhen, China, October (2019) 13–17, proceedings, part IV 22. 2019, p. 540–8.

[134] Toledo-Cortés S, De La Pava M, Perdómo O, González FA. Hybrid deep learning gaussian process for diabetic retinopathy diagnosis and uncertainty quantification. In: International workshop on ophthalmic medical image analysis. 2020, p. 206–15.

[135] Wang L, Ju L, Zhang D, Wang X, He W, Huang Y, Yang Z, Yao X, Zhao X, Ye X, et al. Medical matting: a new perspective on medical segmentation with uncertainty. In: Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1 2021, proceedings, part III 24. 2021, p. 573–83.

[136] Molchanova N, Raina V, Malinin A, La Rosa F, Muller H, Gales M, Granziera C, Graziani M, Cuadra MB. Novel structural-scale uncertainty measures and error retention curves: application to multiple sclerosis. In: 2023 IEEE 20th international symposium on biomedical imaging. ISBI, 2022.

[137] Bhat I, Kuijf HJ, Cheplygina V, Pluim JP. Using uncertainty estimation to reduce false positives in liver lesion detection. In: 2021 IEEE 18th international symposium on biomedical imaging. ISBI, 2021, p. 663–7.

[138] Bhat I, Pluim JP, Viergerver MA, Kuijf HJ. Influence of uncertainty estimation techniques on false-positive reduction in liver lesion detection. Mach Learn Biomed Imaging 2022;1:1–33.

[139] Van Griethuysen JJ, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, Beets-Tan RG, Fillion-Robin J-C, Pieper S, Aerts HJ. Computational radiomics system to decode the radiographic phenotype. Cancer Res 2017;77(21):e104–7.

[140] Ozdemir O, Russell RL, Berlin AA. A 3d probabilistic deep learning system for detection and diagnosis of lung cancer using low-dose ct scans. IEEE Trans Med Imaging 2019;39(5):1419–29.

[141] Roy AG, Conjeti S, Navab N, Wachinger C, Initiative ADN, et al. Bayesian quicknat: Model uncertainty in deep whole-brain segmentation for structure-wise quality control. NeuroImage 2019;195:11–22.

[142] Graham S, Chen H, Gamper J, Dou Q, Heng P-A, Snead D, Tsang YW, Rajpoot N. Mild-net: Minimal information loss dilated network for gland instance segmentation in colon histology images. Med Image Anal 2019;52:199–211.

[143] Hiasa Y, Otake Y, Takao M, Ogawa T, Sugano N, Sato Y. Automated muscle segmentation from clinical ct using Bayesian u-net for personalized musculoskeletal modeling. IEEE Trans Med Imaging 2019;39(4):1030–40.

[144] Hu X, Guo R, Chen J, Li H, Waldmannstetter D, Zhao Y, Li B, Shi K, Menze B. Coarse-to-fine adversarial networks and zone-based uncertainty analysis for nk/t-cell lymphoma segmentation in ct/pet images. IEEE J Biomed Health Inform 2020;24(9):2599–608.

[145] Rosas-Gonzalez S, Birgui-Sekou T, Hidane M, Zemmoura I, Tauber C. Asymmetric ensemble of asymmetric u-net models for brain tumor segmentation with uncertainty estimation. Front Neurol 2021;1421.

[146] Wang G, Li W, Ourselin S, Vercauteren T. Automatic brain tumor segmentation based on cascaded convolutional neural networks with uncertainty estimation. Front Comput Neurosci 2019;13:56.

[147] Balagopal A, Nguyen D, Morgan H, Weng Y, Dohopolski M, Lin M-H, Barkousaraie AS, Gonzalez Y, Garant A, Desai N, et al. A deep learning-based framework for segmenting invisible clinical target volumes with estimated uncertainties for post-operative prostate cancer radiotherapy. Med Image Anal 2021;72:102101.

[148] Jungo A, Meier R, Ermis E, Herrmann E, Reyes M. Uncertainty-driven sanity check: Application to postoperative brain tumor cavity segmentation. Med Imaging Deep Learn 2018.

[149] Ghosal S, Xie A, Shah P. Uncertainty quantified deep learning for predicting dice coefficient of digital histopathology image segmentation. 2021, arXiv e-prints.

[150] Hann E, Gonzales RA, Popescu IA, Zhang Q, Ferreira VM, Piechnik SK. Ensemble of deep convolutional neural networks with monte carlo dropout sampling for automated image segmentation quality control and robust deep learning using small datasets. In: Annual conference on medical image understanding and analysis. 2021, p. 280–93.

[151] Arega TW, Bricq S, Legrand F, Jacquier A, Lalande A, Meriaudeau F. Automatic uncertainty-based quality controlled t1 mapping and ecv analysis from native and post-contrast cardiac t1 mapping images using bayesian vision transformer. Med Image Anal 2023;86:102773.

[152] Huang Z, Gan Y, Lye T, Zhang H, Laine A, Angelini ED, Hendon C. Heterogeneity measurement of cardiac tissues leveraging uncertainty information from image segmentation. In: International conference on medical image computing and computer-assisted intervention. 2020, p. 782–91.

[153] Wickstrøm K, Kampffmeyer M, Jenssen R. Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. Med Image Anal 2020;60:101619.

[154] Natekar P, Kori A, Krishnamurthi G. Demystifying brain tumor segmentation networks: interpretability and uncertainty analysis. Front Computat Neurosci 2020;14:6.

[155] Föllmer B, Biavati F, Wald C, Stober S, Ma J, Dewey M, Samek W. Active multitask learning with uncertainty weighted loss for coronary calcium scoring. Med Phys 2022.

[156] Jiménez-Sánchez A, Mateus D, Kirchhoff S, Kirchhoff C, Biberthaler P, Navab N, Ballester MAG, Piella G. Curriculum learning for improved femur fracture classification: Scheduling data with prior knowledge and uncertainty. Med Image Anal 2022;75:102273.
[157] Ju L, Wang X, Wang L, Mahapatra D, Zhao X, Zhou Q, Liu T, Ge Z. Improving medical images classification with label noise using dual-uncertainty estimation. IEEE Trans Med Imaging 2022;41(6):1533–46.
[158] Li C, Li M, Peng C, Lovell BC. Dynamic curriculum learning via in-domain uncertainty for medical image classification. In: International conference on medical image computing and computer-assisted intervention. 2023, p. 747–57.
[159] Belharbi S, Rony J, Dolz J, Ayed IB, McCaffrey L, Granger E. Deep interpretable classification and weakly-supervised segmentation of histology images via max–min uncertainty. IEEE Trans Med Imaging 2021;41(3):702–14.
[160] Xiang J, Qiu P, Yang Y. Fussnet: Fusing two sources of uncertainty for semi-supervised medical image segmentation. In: Medical image computing and computer assisted intervention–MICCAI 2022: 25th international conference, Singapore, September 18–22, 2022, proceedings, part VIII. 2022, p. 481–91.
[161] Sedai S, Antony B, Rai R, Jones K, Ishikawa H, Schuman J, Gadi W, Garnavi R. Uncertainty guided semi-supervised segmentation of retinal layers in oct images. In: International conference on medical image computing and computer-assisted intervention. 2019, p. 282–90.
[162] 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: International conference on medical image computing and computer-assisted intervention. 2019, p. 605–13.
[163] Cao X, Chen H, Li Y, Peng Y, Wang S, Cheng L. Uncertainty aware temporal-ensembling model for semi-supervised abus mass segmentation. IEEE Trans Med Imaging 2020;40(1):431–43.
[164] Lu W, Lei J, Qiu P, Sheng R, Zhou J, Lu X, Yang Y. Upcol: Uncertainty-informed prototype consistency learning for semi-supervised medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. 2023, p. 662–72.
[165] Mehta R, Christinck T, Nair T, Lemaitre P, Arnold D, Arbel T. Propagating uncertainty across cascaded medical imaging tasks for improved deep learning inference. In: Uncertainty for safe utilization of machine learning in medical imaging and clinical image-based procedures. 2019, p. 23–32.
[166] Feiner LF, Menten MJ, Hammernik K, Hager P, Huang W, Rueckert D, Braren RF, Kaissis G. Propagation and attribution of uncertainty in medical imaging pipelines. In: International workshop on uncertainty for safe utilization of machine learning in medical imaging. 2023, p. 1–11.
[167] Soberanis-Mukul RD, Navab N, Albarqouni S. Uncertainty-based graph convolutional networks for organ segmentation refinement. In: Medical imaging with deep learning. 2020, p. 755–69.
[168] Wang M, Wang L, Xu X, Zou K, Qian Y, Goh RSM, Liu Y, Fu H. Federated uncertainty-aware aggregation for fundus diabetic retinopathy staging. In: Medical image computing and computer assisted intervention – MICCAI 2023. 2023, p. 222–32.
[169] Zhu J, Bolsterlee B, Chow BV, Song Y, Meijering E. Uncertainty and shape-aware continual test-time adaptation for cross-domain segmentation of medical images. In: International conference on medical image computing and computer-assisted intervention. 2023, p. 659–69.
[170] Norouzi A, Emami A, Najarian K, Karimi N, Soroushmehr SR, et al. Exploiting uncertainty of deep neural networks for improving segmentation accuracy in mri images. In: ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing. ICASSP, 2019, p. 2322–6.
[171] Iwamoto S, Raytchev B, Tamaki T, Kaneda K. Improving the reliability of semantic segmentation of medical images by uncertainty modeling with bayesian deep networks and curriculum learning. In: Uncertainty for safe utilization of machine learning in medical imaging, and perinatal imaging, placental and preterm image analysis. 2021, p. 34–43.
[172] Singh RK, Gorantla R, Allada SGR, Narra P. Skinet: A deep learning framework for skin lesion diagnosis with uncertainty estimation and explainability. Plos One 2022;17(10):e0276836.
[173] Judge T, Bernard O, Porumb M, Chartsias A, Beqiri A, Jodoin P-M. Crisp-reliable uncertainty estimation for medical image segmentation. In: Medical image computing and computer assisted intervention–MICCAI 2022: 25th international conference, Singapore, September 18–22, 2022, proceedings, part VIII. 2022, p. 492–502.
[174] Ghoshal B, Tucker A. Estimating uncertainty and interpretability in deep learning for coronavirus (covid-19) detection. 2020, arXiv e-prints.
[175] Calderon-Ramirez S, Yang S, Moemeni A, Colreavy-Donnelly S, Elizondo DA, Oala L, Rodríguez-Capitán J, Jiménez-Navarro M, López-Rubio E, Molina-Cabello MA. Improving uncertainty estimation with semi-supervised deep learning for covid-19 detection using chest x-ray images. IEEE Access 2021;9:85442–54.
[176] Zhang R, Gatsonis C, Steingrimsson JA. Role of calibration in uncertainty-based referral for deep learning. Stat Methods Med Res 2023;09622802231158811.
[177] Sander J, de Vos BD, Wolterink JM, Isgum I. Towards increased trustworthiness of deep learning segmentation methods on cardiac MRI. In: Medical imaging 2019: image processing. vol. 10949, 2019, 1094919.
[178] 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.
[179] Leibig C, Allken V, Ayhan MS, Berens P, Wahl S. Leveraging uncertainty information from deep neural networks for disease detection. Sci Rep 2017;7(1):1–14.
[180] Mehta R, Filos A, Baid U, Sako C, McKinley R, Rebsamen M, Dätwyler K, Meier R, Radojewski P, Murugesan GK, et al. Qu-brats: Miccai brats 2020 challenge on quantifyinguncertainty in brain tumor segmentation-analysis of ranking scores and benchmarking results. J Mach Learn Biomed Imaging 2022;1.
[181] Malinin A, Athanasopoulos A, Barakovic M, Cuadra MB, Gales MJ, Granziera C, Graziani M, Kartashev N, Kyriakopoulos K, Lu P-J, et al. Shifts 2.0: Extending the dataset of real distributional shifts. 2022, arXiv preprint arXiv:2206.15407.
[182] Combalia M, Hueto F, Puig S, Malvehy J, Vilaplana V. Uncertainty estimation in deep neural networks for dermoscopic image classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020, p. 744–5.
[183] Hoebel K, Andrearczyk V, Beers A, Patel J, Chang K, Depeursinge A, Müller H, Kalpathy-Cramer J. An exploration of uncertainty information for segmentation quality assessment. In: Medical imaging 2020: image processing. vol. 11313, 2020, p. 381–90.
[184] Maier-Hein L, Menze B, et al. Metrics reloaded: Pitfalls and recommendations for image analysis validation. 2022, arXiv. org, no. 2206.01653.
[185] Raina V, Molchanova N, Graziani M, Malinin A, Muller H, Cuadra MB, Gales M. Tackling bias in the dice similarity coefficient: Introducing ndsc for white matter lesion segmentation. In: 2023 IEEE 20th international symposium on biomedical imaging. ISBI, 2022.
[186] Jungo A, Meier R, Ermis E, Blatti-Moreno M, Herrmann E, Wiest R, Reyes M. On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. 2018, p. 682–90.
[187] Roshanzamir P, Rivaz H, Ahn J, Mirza H, Naghdi N, Anstruther M, Battié MC, Fortin M, Xiao Y. How inter-rater variability relates to aleatoric and epistemic uncertainty: a case study with deep learning-based paraspinal muscle segmentation. In: International workshop on uncertainty for safe utilization of machine learning in medical imaging. 2023, p. 74–83.
[188] Wald Y, Feder A, Greenfeld D, Shalit U. On calibration and out-of-domain generalization. Adv Neural Inf Process Syst 2021;34:2215–27.
[189] Munir MA, Khan MH, Sarfraz M, Ali M. Towards improving calibration in object detection under domain shift. Adv Neural Inf Process Syst 2022;35:38706–18.
[190] Tomani C, Gruber S, Erdem ME, Cremers D, Buettner F. Post-hoc uncertainty calibration for domain drift scenarios. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021, p. 10124–32.
[191] Gong Y, Lin X, Yao Y, Dietterich TG, Divakaran A, Gervasio M. Confidence calibration for domain generalization under covariate shift. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021, p. 8958–67.
[192] Osawa K, Swaroop S, Khan ME, Jain A, Eschenhagen R, Turner RE, Yokota R. Practical deep learning with bayesian principles. In: Advances in neural information processing systems 32: annual conference on neural information processing systems 2019. 2019, p. 4289–301.
[193] Ashukha A, Lyzhov A, Molchanov D, Vetrov DP. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. In: 8th international conference on learning representations. ICLR 2020, 2020.
[194] Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, Burren Y, Porz N, Slotboom J, Wiest R, et al. The multimodal brain tumor image segmentation benchmark (brats). IEEE Trans Med Imaging 2014;34(10):1993–2024.
[195] Ulmer D, Cinà G. Know your limits: Uncertainty estimation with relu classifiers fails at reliable ood detection. In: Uncertainty in artificial intelligence. 2021, p. 1766–76.
[196] Baur C, Wiestler B, Albarqouni S, Navab N. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. In: Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries: 4th international workshop, brainLes 2018, held in conjunction with MICCAI 2018, granada, Spain, September 16, 2018, revised selected papers, part i 4. 2019, p. 161–9.
[197] Zimmerer D, Isensee F, Petersen J, Kohl S, Maier-Hein K. Unsupervised anomaly localization using variational auto-encoders. In: Medical image computing and computer assisted intervention–MICCAI 2019: 22nd international conference, shenzhen, China, October (2019) 13–17, proceedings, part IV 22. 2019, p. 289–97.
[198] Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, Dillon J, Lakshminarayanan B, Snoek J. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. Adv Neural Inf Process Syst 2019;32:13969–80.
[199] Minderer M, Djolonga J, Romijnders R, Hubis F, Zhai X, Houlsby N, Tran D, Lucic M. Revisiting the calibration of modern neural networks. Adv Neural Inf Process Syst 2021;34:15682–94.
[200] Combi C, Amico B, Bellazzi R, Holzinger A, Moore JH, Zitnik M, Holmes JH. A manifesto on explainability for artificial intelligence in medicine. Artif Intell Med 2022;133:102423.

[201] Hasan SK, Linte CA. A multi-task cross-task learning architecture for ad hoc uncertainty estimation in 3d cardiac mri image segmentation. In: 2021 computing in cardiology. CinC, vol. 48, 2021, p. 1–4.
[202] Ahsan MA, Qayyum A, Razi A, Qadir J. An active learning method for diabetic retinopathy classification with uncertainty quantification. Med Biol Eng Comput 2022;1–15.
[203] Rączkowski Ł, Możejko M, Zambonelli J, Szczurek E. Ara: accurate, reliable and active histopathological image classification framework with bayesian deep learning. Sci Rep 2019;9(1):1–12.
[204] Song B, Sunny S, Li S, Gurushanth K, Mendonca P, Mukhia N, Patrick S, Gurudath S, Raghavan S, Tsusennaro I, et al. Bayesian deep learning for reliable oral cancer image classification. Biomed Opt Express 2021;12(10):6422–30.
[205] Lambert B, Forbes F, Doyle S, Tucholka A, Dojat M. Beyond voxel prediction uncertainty: Identifying brain lesions you can trust. In: Interpretability of machine intelligence in medical image computing: 5th international workshop, IMIMIC 2022, held in conjunction with MICCAI 2022, Singapore, Singapore, September 22 2022, proceedings. 2022, p. 61–70.
[206] Hasan SK, Linte CA. Calibration of cine mri segmentation probability for uncertainty estimation using a multi-task cross-task learning architecture. In: Medical imaging 2022: image-guided procedures, robotic interventions, and modeling. vol. 12034, 2022, p. 174–9.
[207] Mojiri Forooshani P, Biparva M, Ntiri EE, Ramirez J, Boone L, Holmes MF, Adamo S, Gao F, Ozzoude M, Scott CJ, et al. Deep Bayesian networks for uncertainty estimation and adversarial resistance of white matter hyperintensity segmentation. Tech. rep., Wiley Online Library; 2022.
[208] Gou X, He X, et al. Deep learning-based detection and diagnosis of subarachnoid hemorrhage. J Healthc Eng 2021;2021.
[209] Cao X, Chen H, Li Y, Peng Y, Wang S, Cheng L. Dilated densely connected u-net with uncertainty focus loss for 3d abus mass segmentation. Comput Methods Programs Biomed 2021;209:106313.
[210] Zhao Y, Yang C, Schweidtmann A, Tao Q. Efficient bayesian uncertainty estimation for nnu-net. In: Medical image computing and computer assisted intervention–MICCAI 2022: 25th international conference, Singapore, September 18–22, 2022, proceedings, part VIII. 2022, p. 535–44.
[211] Zhang G, Dang H, Xu Y. Epistemic and aleatoric uncertainties reduction with rotation variation for medical image segmentation with convnets. SN Appl Sci 2022;4(2):1–11.
[212] Ghoshal B, Tucker A, Sanghera B, Lup Wong W. Estimating uncertainty in deep learning for reporting confidence to clinicians in medical image segmentation and diseases detection. Comput Intell 2021;37(2):701–34.
[213] Yang J, Liang Y, Zhang Y, Song W, Wang K, He L. Exploring instance-level uncertainty for medical detection. In: 2021 IEEE 18th international symposium on biomedical imaging. ISBI, 2021, p. 448–52.
[214] Liu Y, Yang G, Hosseiny M, Azadikhah A, Mirak SA, Miao Q, Raman SS, Sung K. Exploring uncertainty measures in bayesian deep attentive neural networks for prostate zonal segmentation. IEEE Access 2020;8:151817–28.
[215] Bhat I, Kuijf HJ. Extending probabilistic u-net using mc-dropout to quantify data and model uncertainty. In: International MICCAI brainlesion workshop. 2022, p. 555–9.
[216] Pocevičiūtė M, Eilertsen G, Jarkman S, Lundström C. Generalisation effects of predictive uncertainty estimation in deep learning for digital pathology. Sci Rep 2022;12(1):1–15.
[217] Calderon-Ramirez S, Murillo-Hernandez D, Rojas-Salazar K, Calvo-Valverd L-A, Yang S, Moemeni A, Elizondo D, Lopez-Rubio E, Molina-Cabello MA. Improving uncertainty estimations for mammogram classification using semi-supervised learning. In: 2021 international joint conference on neural networks. IJCNN, 2021, p. 1–8.
[218] Mahapatra D, Poellinger A, Shao L, Reyes M. Interpretability-driven sample selection using self supervised learning for disease classification and segmentation. IEEE Trans Med Imaging 2021;40(10):2548–62.
[219] Senousy Z, Abdelsamea MM, Gaber MM, Abdar M, Acharya UR, Khosravi A, Nahavandi S. Mcua: Multi-level context and uncertainty aware dynamic deep ensemble for breast cancer histology image classification. IEEE Trans Biomed Eng 2021;69(2):818–29.
[220] Lee J, Shin D, Oh S-H, Kim H. Method to minimize the errors of ai: Quantifying and exploiting uncertainty of deep learning in brain tumor segmentation. Sensors 2022;22(6):2406.
[221] Rousseau A-J, Becker T, Bertels J, Blaschko MB, Valkenborg D. Post training uncertainty calibration of deep networks for medical image segmentation. In: 2021 IEEE 18th international symposium on biomedical imaging. ISBI, 2021, p. 1052–6.
[222] Tousignant A, Lemaître P, Precup D, Arnold DL, Arbel T. Prediction of disease progression in multiple sclerosis patients using deep learning analysis of mri data. In: International conference on medical imaging with deep learning. 2019, p. 483–92.
[223] Ozdemir O, Woodward B, Berlin AA. Propagating uncertainty in multi-stage bayesian convolutional neural networks with application to pulmonary nodule detection. In: 2nd workshop on Bayesian deep learning. NeurIPS 2017, Vancouver, Canada; 2017.
[224] Pan H, Feng Y, Chen Q, Meyer C, Feng X. Prostate segmentation from 3d mri using a two-stage model and variable-input based uncertainty measure. In: 2019 IEEE 16th international symposium on biomedical imaging. ISBI 2019, 2019, p. 468–71.
[225] Molle PV, Verbelen T, Boom CD, Vankeirsbilck B, Vylder JD, Diricx B, Kimpe T, Simoens P, Dhoedt B. Quantifying uncertainty of deep neural networks in skin lesion classification. In: Uncertainty for safe utilization of machine learning in medical imaging and clinical image-based procedures. 2019, p. 52–61.
[226] Mobiny A, Singh A, Van Nguyen H. Risk-aware machine learning classifier for skin lesion diagnosis. J Clin Med 2019;8(8):1241.
[227] Abideen ZU, Ghafoor M, Munir K, Saqib M, Ullah A, Zia T, Tariq SA, Ahmed G, Zahra A. Uncertainty assisted robust tuberculosis identification with bayesian convolutional neural networks. IEEE Access 2020;8:22812–25.
[228] Mehta R, Filos A, Gal Y, Arbel T. Uncertainty evaluation metric for brain tumour segmentation. In: Medical imaging with deep learning. 2020.
[229] Yang S, Fevens T. Uncertainty quantification and estimation in medical image classification. In: International conference on artificial neural networks. 2021, p. 671–83.
[230] Laves M-H, Ihler S, Ortmaier T. Uncertainty quantification in computer-aided diagnosis: Make your model say i don't know for ambiguous cases. In: Medical imaging with deep learning. 2019.
[231] Rajaraman S, Zamzmi G, Yang F, Xue Z, Jaeger S, Antani SK. Uncertainty quantification in segmenting tuberculosis-consistent findings in frontal chest x-rays. Biomedicines 2022;10(6):1323.
[232] Xia Y, Yang D, Yu Z, Liu F, Cai J, Yu L, Zhu Z, Xu D, Yuille A, Roth H. Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. Med Image Anal 2020;65:101766.
[233] Redekop E, Chernyavskiy A. Uncertainty-based method for improving poorly labeled segmentation datasets. In: 2021 IEEE 18th international symposium on biomedical imaging. ISBI, 2021, p. 1831–5.
[234] Dolezal JM, Srisuwananukorn A, Karpeyev D, Ramesh S, Kochanny S, Cody B, Mansfield AS, Rakshit S, Bansal R, Bois MC, et al. Uncertainty-informed deep learning models enable high-confidence predictions for digital histopathology. Nature Commun 2022;13(1):6572.
[235] Ruan Y, Li D, Marshall H, Miao T, Cossetto T, Chan I, Daher O, Accorsi F, Goela A, Li S. 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. 2020, p. 439–49.
[236] Hu L, Li J, Peng X, Xiao J, Zhan B, Zu C, Wu X, Zhou J, Wang Y. Semi-supervised npc segmentation with uncertainty and attention guided consistency. Knowl-Based Syst 2022;239:108021.
[237] Huang L, Ruan S, Decazes P, Denœux T. Lymphoma segmentation from 3d pet-ct images using a deep evidential network. Internat J Approx Reason 2022;149:39–60.
[238] Gonzalez C, Gotkowski K, Bucher A, Fischbach R, Kaltenborn I, Mukhopadhyay A. Detecting when pre-trained nnu-net models fail silently for covid-19 lung lesion segmentation. In: Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, strasbourg, France, September 27–October 1 2021, proceedings, part VII 24. 2021, p. 304–14.
[239] Vasiliuk A, Frolova D, Belyaev M, Shirokikh B. Limitations of out-of-distribution detection in 3d medical image segmentation. J Imaging 2023;9(9).
[240] Lin Q, Chen X, Chen C, Garibaldi JM. A novel quality control algorithm for medical image segmentation based on fuzzy uncertainty. IEEE Trans Fuzzy Syst 2023;31(8):2532–44.
[241] Zhang Y, Xi R, Fu H, Towey D, Bai R, Higashita R, Liu J. Elongated physiological structure segmentation via spatial andscale uncertainty-aware network. In: Medical image computing and computer assisted intervention – MICCAI 2023. 2023, p. 323–32.
[242] Vasiliuk A, Frolova D, Belyaev M, Shirokikh B. Redesigning out-of-distribution detection on 3d medical images. In: International workshop on uncertainty for safe utilization of machine learning in medical imaging. 2023, p. 126–35.
[243] Wang S, Nuyts J, Filipovic M. Uncertainty estimation in liver tumor segmentation using the posterior bootstrap. In: International workshop on uncertainty for safe utilization of machine learning in medical imaging. 2023, p. 188–97.
[244] Thibeau-Sutre E, Alblas D, Buurman S, Brune C, Wolterink JM. Uncertainty-based quality assurance of carotid artery wall segmentation in black-blood mri. In: International workshop on uncertainty for safe utilization of machine learning in medical imaging. 2023, p. 95–103.
[245] Buddenkotte T, Sanchez LE, Crispin-Ortuzar M, Woitek R, McCague C, Brenton JD, Öktem O, Sala E, Rundo L. Calibrating ensembles for scalable uncertainty quantification in deep learning-based medical image segmentation. Comput Biol Med 2023;107096.
[246] Linmans J, Elfwing S, van der Laak J, Litjens G. Predictive uncertainty estimation for out-of-distribution detection in digital pathology. Med Image Anal 2023;83:102655.
[247] Li H, Nan Y, Del Ser J, Yang G. Region-based evidential deep learning to quantify uncertainty and improve robustness of brain tumor segmentation. Neural Comput Appl 2023;35(30):22071–85.

[248] Shamsi A, Asgharnezhad H, Jokandan SS, Khosravi A, Kebria PM, Nahavandi D, Nahavandi S, Srinivasan D. An uncertainty-aware transfer learning-based framework for covid-19 diagnosis. IEEE Trans Neural Netw Learn Syst 2021;32(4):1408–17.
[249] Guo F, Ng M, Kuling G, Wright G. Cardiac mri segmentation with sparse annotations: Ensembling deep learning uncertainty and shape priors. Med Image Anal 2022;102532.
[250] Ayhan MS, Kühlewein L, Aliyeva G, Inhoffen W, Ziemssen F, Berens P. Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection. Med Image Anal 2020;64:101724.
[251] Pal JB. Holistic network for quantifying uncertainties in medical images. In: International MICCAI brainlesion workshop. 2022, p. 560–9.
[252] Vu MH, Nyholm T, Löfstedt T. Multi-decoder networks with multi-denoising inputs for tumor segmentation. In: International MICCAI brainlesion workshop. 2020, p. 412–23.
[253] Mehrtash A, Kapur T, Tempany CM, Abolmaesumi P, Wells WM. Prostate cancer diagnosis with sparse biopsy data and in presence of location uncertainty. In: 2021 IEEE 18th international symposium on biomedical imaging. ISBI, 2021, p. 443–7.
[254] Yang L, Zhang Y, Chen J, Zhang S, Chen DZ. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. 2017, p. 399–407.
[255] Wang X, Tang F, Chen H, Luo L, Tang Z, Ran A-R, Cheung CY, Heng P-A. Ud-mil: uncertainty-driven deep multiple instance learning for oct image classification. IEEE J Biomed Health Inform 2020;24(12):3431–42.
[256] Mei H, Lei W, Gu R, Ye S, Sun Z, Zhang S, Wang G. Automatic segmentation of gross target volume of nasopharynx cancer using ensemble of multiscale deep neural networks with spatial attention. Neurocomputing 2021;438:211–22.
[257] Lu W, Lei J, Qiu P, Sheng R, Zhou J, Lu X, Yang Y. Upcol: Uncertainty-informed prototype consistency learning for semi-supervised medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. 2023, p. 662–72.
[258] Galdran A, Verjans JW, Carneiro G, González Ballester MA. Multi-head multi-loss model calibration. In: International conference on medical image computing and computer-assisted intervention. 2023, p. 108–17.
[259] Hann E, Popescu IA, Zhang Q, Gonzales RA, Barutçu A, Neubauer S, Ferreira VM, Piechnik SK. Deep neural network ensemble for on-the-fly quality control-driven segmentation of cardiac mri t1 mapping. Med Image Anal 2021;71:102029.
[260] Alves N, Bosma JS, Venkadesh KV, Jacobs C, Saghir Z, de Rooij M, Hermans J, Huisman H. Prediction variability to identify reduced ai performance in cancer diagnosis at mri and ct. Radiology 2023;308(3):e230275.
[261] Hann E, Biasiolli L, Zhang Q, Popescu IA, Werys K, Lukaschuk E, Carapella V, Paiva JM, Aung N, Rayner JJ, et al. Quality control-driven image segmentation towards reliable automatic image analysis in large-scale cardiovascular magnetic resonance aortic cine imaging. In: Medical image computing and computer assisted intervention–MICCAI 2019: 22nd international conference, shenzhen, China, October 13–17, 2019, proceedings, part II 22. 2019, p. 750–8.
[262] Zhang X, Cerna AEU, Stough JV, Chen Y, Carry BJ, Alsaid A, Raghunath S, VanMaanen DP, Fornwalt BK, Haggerty CM. Generalizability and quality control of deep learning-based 2d echocardiography segmentation models in a large clinical dataset. Int J Cardiovasc Imaging 2022;38(8):1685–97.
[263] Lu C, Angelopoulos AN, Pomerantz S. Improving trustworthiness of ai disease severity rating in medical imaging with ordinal conformal prediction sets. In: Medical image computing and computer assisted intervention–MICCAI 2022: 25th international conference, Singapore, September (2022) 18–22, proceedings, part VIII. 2022, p. 545–54.
[264] Wang G, Li W, Zuluaga MA, Pratt R, Patel PA, Aertsen M, Doel T, David AL, Deprest J, Ourselin S, et al. Interactive medical image segmentation using deep learning with image-specific fine tuning. IEEE Trans Med Imaging 2018;37(7):1562–73.
[265] Mojabi P, Khoshdel V, Lovetri J. Tissue-type classification with uncertainty quantification of microwave and ultrasound breast imaging: A deep learning approach. IEEE Access 2020;8:182092–104.
[266] Lourenço-Silva J, Oliveira AL. Using soft labels to model uncertainty in medical image segmentation. In: International MICCAI brainlesion workshop. 2022, p. 585–96.
[267] Lin H, Li Z, Yang Z, Wang Y. Variance-aware attention u-net for multi-organ segmentation. Med Phys 2021;48(12):7864–76.
[268] Wieslander H, Harrison PJ, Skogberg G, Jackson S, Fridén M, Karlsson J, Spjuth O, Wählby C. Deep learning with conformal prediction for hierarchical analysis of large-scale whole-slide tissue images. IEEE J Biomed Health Inform 2020;25(2):371–80.
[269] Thiagarajan JJ, Venkatesh B, Rajan D, Sattigeri P. Improving reliability of clinical models using prediction calibration. 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. 2020, p. 71–80.
[270] Larrazabal AJ, Martínez C, Dolz J, Ferrante E. Maximum entropy on erroneous predictions: Improving model calibration for medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. 2023, p. 273–83.
[271] Xiao X, Hu QV, Wang G. Edge-aware multi-task network for integrating quantification segmentation and uncertainty prediction of liver tumor on multi-modality non-contrast mri. In: International conference on medical image computing and computer-assisted intervention. 2023, p. 652–61.
[272] Zhao X, Shen Z, Chen D, Wang S, Zhuang Z, Wang Q, Zhang L. One-shot traumatic brain segmentation with adversarial training and uncertainty rectification. In: International conference on medical image computing and computer-assisted intervention. 2023, p. 120–9.
[273] Zhu J, Bolsterlee B, Chow BV, Song Y, Meijering E. Uncertainty and shape-aware continual test-time adaptation for cross-domain segmentation of medical images. In: International conference on medical image computing and computer-assisted intervention. 2023, p. 659–69.
[274] Murugesan B, Adiga Vasudeva S, Liu B, Lombaert H, Ben Ayed I, Dolz J. Trust your neighbours: Penalty-based constraints for model calibration. In: International conference on medical image computing and computer-assisted intervention. 2023, p. 572–81.
[275] Karani N, Dey N, Golland P. Boundary-weighted logit consistency improves calibration of segmentation networks. In: International conference on medical image computing and computer-assisted intervention. 2023, p. 367–77.
[276] Shui C, Szeto J, Mehta R, Arnold DL, Arbel T. Mitigating calibration bias without fixed attribute grouping for improved fairness in medical imaging analysis. In: International conference on medical image computing and computer-assisted intervention. 2023, p. 189–98.
[277] Philps B, Valdes Hernandez MdC, Bernabeu Llinares M. Proper scoring loss functions are simple and effective for uncertainty quantification of white matter hyperintensities. In: International workshop on uncertainty for safe utilization of machine learning in medical imaging. 2023, p. 208–18.
[278] Lambert B, Forbes F, Doyle S, Dojat M. Anisotropic hybrid networks for liver tumor segmentation with uncertainty quantification. In: Resource-efficient medical image analysis - 2nd international workshop, REMIA 2023, held in conjunction with MICCAI 2023, vancouver, BC, Canada, October 12, 2023, proceedings. vol. 14394, 2023.
[279] Yeung M, Rundo L, Nan Y, Sala E, Schönlieb C-B, Yang G. Calibrating the dice loss to handle neural network overconfidence for biomedical image segmentation. J Digit Imaging 2023;36(2):739–52.
[280] Shaw R, Sudre CH, Ourselin S, Cardoso MJ, Pemberton HG. A heteroscedastic uncertainty model for decoupling sources of mri image quality. Mach Learn Biomed Imaging 2021;1:1–23.
[281] McKinley R, Wepfer R, Grunder L, Aschwanden F, Fischer T, Friedli C, Muri R, Rummel C, Verma R, Weisstanner C, et al. Automatic detection of lesion load change in multiple sclerosis using convolutional neural networks with segmentation confidence. NeuroImage: Clin 2020;25:102104.
[282] McKinley R, Meier R, Wiest R. Ensembles of densely-connected cnns with label-uncertainty for brain tumor segmentation. In: International MICCAI brainlesion workshop. 2018, p. 456–65.
[283] Sedai S, Antony B, Mahapatra D, Garnavi R. Joint segmentation and uncertainty visualization of retinal layers in optical coherence tomography images using bayesian deep learning. In: Computational pathology and ophthalmic medical image analysis. 2018, p. 219–27.
[284] Mishra S, Chen DZ, Hu XS. Objective-dependent uncertainty driven retinal vessel segmentation. In: 2021 IEEE 18th international symposium on biomedical imaging. ISBI, 2021, p. 453–7.
[285] McKinley R, Rebsamen M, Meier R, Wiest R. Triplanar ensemble of 3d-to-2d cnns with label-uncertainty for brain tumor segmentation. In: International MICCAI brainlesion workshop. 2019, p. 379–87.
[286] Araújo T, Aresta G, Mendonça L, Penas S, Maia C, Carneiro Â, Mendonça AM, Campilho A. Dr| graduate: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images. Med Image Anal 2020;63:101715.
[287] Vaseli H, Gu AN, Ahmadi Amiri SN, Tsang MY, Fung A, Kondori N, Saadat A, Abolmaesumi P, Tsang TS. Protoasnet: Dynamic prototypes for inherently interpretable and uncertainty-aware aortic stenosis classification in echocardiography. In: International conference on medical image computing and computer-assisted intervention. 2023, p. 368–78.
[288] Xue W, Guo T, Ni D. Left ventricle quantification with sample-level confidence estimation via bayesian neural network. Comput Med Imaging Graph 2020;84:101753.
[289] Lin Q, Chen X, Chen C, Garibaldi JM. Quality quantification in deep convolutional neural networks for skin lesion segmentation using fuzzy uncertainty measurement. In: 2022 IEEE international conference on fuzzy systems. FUZZ-IEEE, 2022, p. 1–8.
[290] Puyol-Antón E, Ruijsink B, Baumgartner CF, Masci P-G, Sinclair M, Konukoglu E, Razavi R, King AP. Automated quantification of myocardial tissue characteristics from native t 1 mapping using neural networks with uncertainty-based quality-control. J Cardiovasc Magn Reson 2020;22:1–15.

[291] Lennartz J, Schultz T. Segmentation distortion: Quantifying segmentation uncertainty under domain shift via the effects of anomalous activations. In: International conference on medical image computing and computer-assisted intervention. 2023, p. 316–25.
[292] Ren K, Zou K, Liu X, Chen Y, Yuan X, Shen X, Wang M, Fu H. Uncertainty-informed mutual learning for joint medical image classification and segmentation. In: Medical image computing and computer assisted intervention – MICCAI 2023. 2023, p. 35–45.
[293] Fu W, Chen Y, Liu W, Yue X, Ma C. Evidence reconciled neural network for out-of-distribution detection in medical images. In: International conference on medical image computing and computer-assisted intervention. 2023, p. 305–15.
[294] Jones CK, Wang G, Yedavalli V, Sair H. Direct quantification of epistemic and aleatoric uncertainty in 3d u-net segmentation. J Med Imaging 2022;9(3):034002.
[295] Arco JE, Ortiz A, Ramirez J, Martinez-Murcia FJ, Zhang Y-D, Gorriz JM. Uncertainty-driven ensembles of multi-scale deep architectures for image classification. Inf Fusion 2023;89:53–65.
[296] Li Y, Chen X, Quan L, Zhang N. Uncertainty-guided robust training for medical image segmentation. In: 2021 IEEE 18th international symposium on biomedical imaging. ISBI, 2021, p. 1471–5.
[297] Prince EW, Ghosh D, Görg C, Hankinson TC. Uncertainty-aware deep learning classification of adamantinomatous craniopharyngioma from preoperative mri. Diagnostics 2023;13(6):1132.
[298] Zhang Y, Wang S, Zhang Y, Chen DZ. Rr-cp: Reliable-region-based conformal prediction for trustworthy medical image classification. In: International workshop on uncertainty for safe utilization of machine learning in medical imaging. 2023, p. 12–21.
[299] Mehrtens H, Bucher T, Brinker TJ. Pitfalls of conformal predictions for medical image classification. In: International workshop on uncertainty for safe utilization of machine learning in medical imaging. 2023, p. 198–207.