Trustworthy clinical AI solutions: a unified review of uncertainty quantification in deep learning models for medical image analysis

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

Trustworthy clinical AI solutions: a unified review of uncertainty quantification in deep learning models for medical image analysis

Benjamin Lambert[1,3], Florence Forbes[2], Alan Tucholka[3], Senan Doyle[3], Harmonie Dehaene[3] and Michel Dojat[1]

[1]Univ. Grenoble Alpes, Inserm, U1216, Grenoble Institut des Neurosciences, Grenoble, 38000, France
[2]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, Grenoble, 38000, France
[3]Pixyl Research and Development Laboratory, Grenoble, 38000, France

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. Particularly, end users are reluctant to rely on the rough predictions of DL models. Uncertainty quantification methods have been proposed in the literature as a potential response to reduce the rough decision provided by the DL black box and thus 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 to DL predictions. We focus on applications to medical image analysis, which present specific challenges due to the high dimensionality of images and their quality variability, as well as constraints associated to real-life clinical routine. We then discuss the evaluation protocols to validate the relevance of uncertainty estimates. Finally, we highlight the open challenges of uncertainty quantification in the medical field.

1 Introduction

These past years, many Deep Learning (DL) medical applications were proposed for the automatic analysis of various imaging modalities, including Magnetic Resonance Imaging (MRI), Computed Tomography (CT), Ultrasound (US) or histopathological images (see Puttagunta and Ravi (2021) for a review). To be accepted and routinely used by clinicians, however, these algorithms must

provide robust and trustable predictions. This is of particular importance in the context of clinical applications, where the automated prediction may have a direct impact on patient care. Yet, DL models are often considered and used as black-boxes, due to the absence of clear decision rules, as well as to the lack of reliable confidence estimates associated with their predictions (Guo et al., 2017). Additionally, DL models proved to be overconfident about their predictions on outliers data (Nguyen et al., 2015), and very sensitive to adversarial attacks (Ma et al., 2021), which suggests a global lack of robustness of this type of models. Due to these limitations, detecting failures or inconsistencies produced by DL models is complex, raising concerns regarding the reliability and safety of using these algorithms in clinical practice (Ford et al., 2016). To tackle this essential aspect, several research directions have emerged in order to mitigate the "black-box issue", including Explainable Artificial Intelligence (XAI) and Uncertainty Quantification (UQ). XAI methods (Arrieta et al., 2020) propose to explain the prediction of the DL model in a way that is understandable to humans. In the context of medical image analysis, an example of XAI approach is the computing of saliency maps showing the image's relevant features identified by the DL model, or example-based explanations consisting in the presentation of cases similar to the one considered, *e.g.* medical images of patients with the same condition, (van der Velden et al., 2022). However, concerns have been raised concerning the fidelity and intelligibility of the explanations provided by XAI methods, which may give the misleading impression of a better understanding of the black-box Adebayo et al. (2018); Rudin (2019). On the other side, UQ methods (Abdar et al., 2021a) were 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, unconfident predictions can indicate an incomplete training dataset. It gives insights regarding 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 property of a deployed AI algorithm (Tonekaboni et al., 2019). As a result, the medical-imaging community is becoming increasingly interested in incorporating UQ to 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. (2021a), authors propose a complete review of UQ methods, as well as their various concrete applications. Hüllermeier and Waegeman (2021) 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. (2021), insights about the various sources of uncertainty are presented. Reviews focusing on Bayesian DL (Jospin et al., 2022; Wang and Yeung, 2020) and prediction intervals Kabir et al. (2018) have been also published. More recently, Zhou et al. (2022) present a review of the latest advances considering epistemic uncertainty quantification in DL from the perspective of generalization error. While these various works propose a complete overview of UQ methods in DL from a general point of view, we have noticed the lack of reviews focusing on medical image processing applications, where being able to correctly identify the confidence of the model is crucial. Kurz et al. (2022) presented a first work in this direction, using a corpus of 22 papers. Their study, however, is restricted to medical image classification. With the present review, we propose to extend the latter by presenting a complete review of 130 peer-reviewed papers implementing UQ applications in supervised DL-based pipelines, for both medical image classification and segmentation. We also aim at providing an in-depth discussion of UQ methods' evaluation procedures, as well as pointing out the challenges of the field and potential future directions. Our review differentiates from other previously published ones by the following contributions:

- A review of UQ methods dedicated to DL medical image processing classification and segmentation.
- A focus on the proposed metrics for uncertainty estimates evaluation.
- 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 report is divided into four sections. Section 2 introduces the key concepts addressed in this study, namely the application of DL models to medical image classification and segmentation (subsection 2.2), as well as the main notions of UQ (subsection 2.3). Section 3 presents the most popular UQ methods applied in the context of medical image analysis. Section 4 then focuses on the evaluation procedures that can be implemented to assess the usefulness of uncertainty estimates. Finally, Section 5 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. With this classical setting, the goal of the DL algorithm is to learn a task T based on a training dataset composed of pairs of input images x, and their associated ground truth y. This target represents a class in the context of classification (*e.g.*, *healthy*, *pathological*), whereas it consists in 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 mapping function $p(y|x)$ from the data.

2.2 Deep Learning for medical image analysis

The common approach for supervised DL medical image processing is the training of a Convolutional Neural Network (CNN) using an annotated dataset (*i.e.* the ground truth). The building block of CNNs is the convolutional layer, which convolves the input data with learnable weighted kernels. This enables the extraction of features within the image, while being insensitive to the position, scale and shape.

For medical image classification, popular convolutional architectures comprises Residual and Dense CNNs (Huang et al., 2017) or EfficientNets (Tan and Le, 2019). These architectures consist of a succession of convolutional layers that extract features from the image at different scales while reducing its size, thus its spatial resolution. For medical image segmentation, popular choices include U-Net (Ronneberger et al., 2015) and its variants, such as Residual U-Net (Kerfoot et al., 2018), V-Net (Milletari et al., 2016), Attention U-Net (Oktay et al., 2018) or Dynamic U-Net (Isensee et al., 2021). These segmentation models are composed of two branches, an encoder and a decoder, forming the U shape. The encoder compresses the dimension of the input image, while the decoder decompresses the signal until it recovers its original size. Between the two modules, skip connections are usually added so that the features learned in the encoder part can be used to generate the segmentation in the decoder part. 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...), the CNNs can be implemented in 2D or 3D.

During the supervised training stage, the CNN uses images from the training set to produces predictions, which are compared to the ground truth targets in order to estimate the error of 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 (Lin et al., 2020). For segmentation tasks, specific loss functions can also be used such as the popular Dice loss (Milletari et al., 2016) and variants (Generalized Dice loss (Fidon et al., 2017) or Tversky loss (Salehi et al., 2017)).

In the context of medical image classification, CNNs 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 CNN predicts one class per pixel or voxel. UQ aims at completing these predictions with uncertainty estimates, allowing a better interpretation of the results with respect to the model's confidence. In the following section, the main concepts of uncertainty are introduced.

2.3 The specific language of uncertainty

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

Epistemic uncertainty describes uncertainty arising from the lack of knowledge about the perfect predictor, considering the current input (Hüllermeier and Waegeman, 2021). In complex scenarios, there is often not a single model, but rather a multitude of models that can explain the observed data (Gal et al., 2016). Thus, uncertainty arises regarding the choice of the model parameters. Epistemic uncertainty is considered to be reducible, meaning that it can be reduced by using additional data. In practice, epistemic uncertainty is expected to be high for images far from the training data distribution (referred to as out-of-distribution (OOD) samples). Such discrepancy between test and training datasets is frequent in medical image analysis, where there may be significant variations between images acquired at different hospitals or using different machines. Additionally, unexpected patterns can be encountered in test images, such as diseases not encountered during training, or artifacts. Popular approaches to improve the generalizability of models to unseen domains include data augmentation (Chen et al., 2020; Ouyang et al., 2021; Zhang et al., 2020) or transfer learning (Ghafoorian et al., 2017).

Aleatoric uncertainty describes intrinsic noise and random effects within the data (Hüllermeier and Waegeman, 2021). It is not intrinsic to the model, but rather a property of the underlying generative distribution of the data. In the context of classification or segmentation, aleatoric uncertainty increases when the number of classes is high and when these classes are fine-grained (Malinin, 2019). Aleatoric uncertainty is considered to be irreducible, meaning that it cannot be reduced with more data. Actually, the only way to diminish aleatoric uncertainty would be to increase the measurement system precision to reduce noise that corrupts the dataset (Gal et al., 2016). Finally, aleatoric uncertainty can be further split into two categories: *homoscedastic uncertainty*, which is identical for each sample of the dataset, and *heteroscedastic uncertainty*, which depends on the query input.

Lastly, closely linked to this notion of data uncertainty, the notion of *label uncertainty* was introduced for segmentation tasks. It has been observed that inter-rater variability in the context of manual delineations of medical images was important (Becker et al., 2019; Joskowicz et al., 2019). 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.

In the next section, we propose an overview of the most employed UQ methods for medical image classification and segmentation, in light of the selected corpus of papers.

3 Review of Uncertainty-Quantification methods for medical image analysis using Deep Learning

We performed a systematic search on June 2022 using Google Scholar and PubMed to identify DL studies implementing UQ methods for medical image classification and segmentation published from 2015 (included) to June 2022. The following combination of keywords was used for the search: "Deep Learning", "Uncertainty", "MRI", "CT", "PET", "X-RAY", "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 were applied 1) non-peer-reviewed studies (exception were made for papers with more than 30 citations); 2) non-English papers; 3) review articles and 4) animal studies. 130 papers were finally selected for analysis. It resulted a total of 199 UQ models, implemented either as principal contributions or as comparison methods (the exhaustive list of methods can be found in A). 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 that is modeled, namely epistemic or aleatoric. Moreover, for real-world deployment of a DL model in a clinical setting, speed is crucial for integration into the routine. This means that the implementation of a UQ protocol should not come with prohibitive inference time or computational cost. Thus, we further distinguish between sampling methods, which require multiple inferences per input image (and thus tend to be slow and/or computationally costly), and single-step methods, which produce uncertainty estimate at the cost of a single inference step (and thus tend to be faster). The resulting taxonomy is presented in Figure 1. In the following of the section, we briefly present each UQ framework.

3.1 Softmax methods

An immediate and simple approach to obtain uncertainty estimates from classification or segmentation by neural network (NN) techniques is to consider the output predicted probabilities $p(y|x)$. Naively, the higher this probability, the more certain is the prediction. Works have been proposed to improve the calibration of the probabilities predicted by a DL model, and ensure that these scores match the true performance of the model (Guo et al., 2017; Kumar et al., 2019; Nixon et al., 2019). More formally, a model is calibrated if, for each prediction with the associated probability p, the model is correct 100 x p of

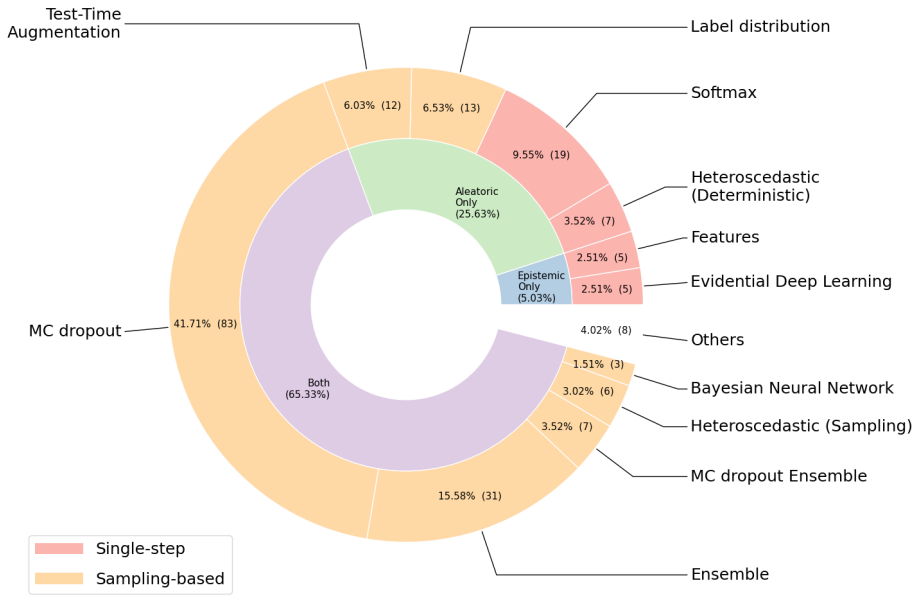

Figure 1: Implemented UQ methods in the 130 selected papers. The percentage (and the number) of the selected papers for each class of methods is indicated in the outer ring and the corresponding percentage below the class name. Orange identifies sampling-based methods, and red single-step methods. The inner ring classifies methods according to the type of uncertainty modeled: aleatoric, epistemic or both.

the time. However, UQ based on softmax probabilities only consider the distribution over the model's outputs, and not on the model's weights. Thus, this type of deterministic uncertainty estimates only consider aleatoric uncertainty Hüllermeier and Waegeman (2021); Kendall and Gal (2017).

3.2 Bayesian Neural Network methods

In Bayesian Deep Learning (BDL), each weight of the NN is replaced by a distribution, rather than having a single fixed value (Blundell et al., 2015). 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 account for the less and more 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. During inference,

the distribution is marginalized by repeatedly sampling weights from the shared distribution and averaging the predictions. Uncertainty estimates such as the entropy of the predictive distribution, its variance or its mutual information can be computed. BNN places a distribution on the model's outputs as well as on the model's weights, hence is able to model both aleatoric and epistemic uncertainties. While being theoretically founded, BNN requires extensive changes to both the model architecture and training paradigm, and also significantly increases the computational cost of training and inference.

3.3 Monte Carlo dropout methods

In Gal and Ghahramani (2016), authors demonstrated that a NN trained with dropout is able to efficiently approximate Bayesian inference without the associated prohibitive computational cost. Based on this principle, the Monte Carlo Dropout (MC dropout) technique proposes to train a model with dropout and keeps it activated during inference. For a given query input, multiple forward passes are then performed. Each time, a different dropout mask is randomly sampled, producing different predictions. Following this process, a predictive distribution is obtained, similarly to BNN. MC dropout allowing to approximate a BNN in any network trained with dropout, it thus rapidly gained popularity. As a counterpart, finding the optimal dropout strategy (rate and position within the NN) is not straight-forward (Jungo et al., 2020).

3.4 Ensemble methods

Deep Ensemble (DE) (Lakshminarayanan et al., 2017) proposes to sequentially train a series of NN. As the weights of the neural network are initialized randomly, the models reach different optimums during training. As a result, they produce diverse predictions for the same query input. As for BDL and MC dropout, uncertainty estimates can then be extracted from the ensemble's predictive distribution. A DE does not require any changes to model architecture or training paradigm. Yet, it requires to repeat the training several times, and the aggregation of each individual prediction at inference, which increases the computational cost of this approach. Finally, it is worth noticing that some works propose to associate ensemble and MC dropout in order to get the best of both methods. This allows the combination of individual models uncertainty (though Monte Carlo dropout) as well as the overall ensemble uncertainty (though Deep Ensemble).

3.5 Heteroscedastic model-based methods

As its denomination indicates, an heteroscedastic model aims at evaluating the heteroscedastic part of aleatoric uncertainty. Within this framework, uncertainty is directly learned during training from the data itself, without the need for ground truth labels for uncertainty. Heteroscedastic models can be categorized into two subtypes: sampling methods, which extend MC dropout models,

and deterministic approaches, which are a recent improvement and produce the prediction together with its related uncertainty in a single-step.

3.5.1 Sampling Heteroscedastic models

In Kendall and Gal (2017), authors hypothesize that the network output logits are corrupted by Gaussian noise with mean equal to 0 and variance z. The higher the variance, the higher is the aleatoric uncertainty. A model can then be trained to predict the mean logits ρ, as well as the noise variance z. To do so, authors duplicate the outputs of a MC dropout network: one for the logits, and one for the variance. At each training step, the loss of the model is evaluated by integrating over multiple samples of noise. Paired with the MC dropout framework, this sampling formulation of heteroscedastic models enables the modeling of both epistemic and aleatoric uncertainties.

3.5.2 Deterministic Heteroscedastic models

Recently, deterministic variants of the heteroscedastic model have been proposed. In this approach, uncertainty is still learned during training from the data itself, but do not require the integration of the loss over multiple samples of noise. As in the sampling approach, the NN is modified by adding a dedicated output for uncertainty. Then, an uncertainty-augmented loss function is used to learn the predictive task (classification or segmentation), while also learning to predict high uncertainty scores for samples that are likely to be incorrect. In this context, aleatoric uncertainty is learned without any additional cost, as it simply exploits the ground truth label (class or segmentation).

3.6 Label-distribution model-based methods

A branch of the UQ literature focuses on modeling label uncertainty in the context of image segmentation. These approaches focus on datasets for which multiple expert manual segmentations are provided for each image, interpreting the inter-rater variability as a form of ground truth uncertainty. In this setting, it becomes possible to approximate the expert label distribution using generative segmentation neural networks (Kohl et al., 2018). At inference, sampling from the learned distribution produces diverse segmentation masks, which reproduce the inter-rater variability. We refer to this family of methods as label-distribution models. Despite being intuitive, it is not clear whether or not inter-rater variability can be used as ground truth for uncertainty. In the context of medical image segmentation, there are many cases where a unique segmentation cannot be obtained, for instance due to partial volume effect observed in MRI at the boundaries between healthy tissue and lesions. In that context, experts segmentations exhibit somewhat random variations around the boundaries of the target object. Moreover, experts can over-segment or alternatively under-segment the same object of interest, based on their annotation style. This inter-rater variability is thus rather linked to contextual biases (e.g,

radiologist experience or annotation habits) rather that on the true uncertainty of the label (Mehta et al., 2022).

3.7 Test-Time Augmentation

Test-Time Augmentation (TTA) (Ayhan and Berens, 2018) 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). This process aims at exploring the impact of input-image transformations on the prediction. Using TTA, the model generates a set of predictions for the same initial input image. From this distribution of predictions, uncertainty metrics can be extracted such as the mediane or the variance.

3.8 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 were recently proposed to detect OOD from the feature map signature of a trained NN (Postels et al., 2021). This builds on the hypothesis that feature maps contain information regarding the correctness of a prediction. Despite being efficient for OOD detection, it has been observed that feature-based uncertainty estimates are generally poorly calibrated (Postels et al., 2021). These methods are computationally efficient, however their application to medical images remains rare.

3.9 Evidential Deep Learning

The Dempster–Shafer Theory of Evidence (DST) is a framework for dealing with epistemic uncertainty (Dempster, 1968). In a K-class classification (respectively, segmentation) problem, DST proposes to assign belief masses to each possible class, as well as an overall uncertainty mass. 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. In practice, DST can be applied to Deep Learning models by fitting a Dirichlet distribution on the model's outputs, in place of the standard categorical distribution. Additionally, the Bayes-Risk loss function (Sensoy et al., 2018) is used to train the model in replacement of the standard cross-entropy loss.

3.10 Other UQ methods

Finally, we found a few methods not conforming to any of the frameworks previously introduced. We list these applications in B, with a short description of each UQ approach proposed.

4 How to evaluate uncertainty quantification approaches

In the previous section, 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 straight forward, as we typically do not dispose of ground-truth uncertainty values. Proxy metrics are thus developed to estimate the performances of uncertainty quantification methods. We have identified 7 types of evaluation protocols (see Figure 2). In the following, we present each protocol and identify their use cases. Table C lists use cases of each metrics in the reviewed corpus of papers.

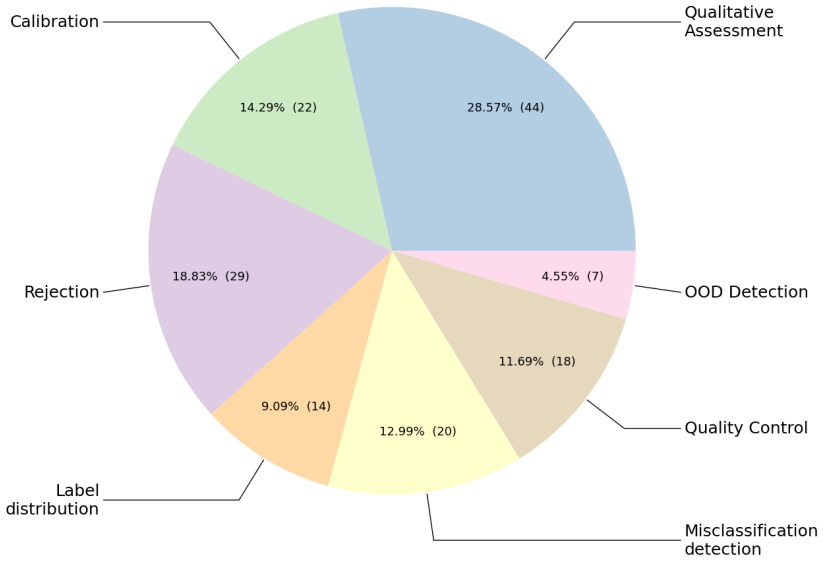

Figure 2: Implemented UQ evaluation protocols in the reviewed papers. The percentage (and the number) of the reviewed papers per class is mentioned in the Pie chart.

4.1 Qualitative assessment protocol

As computing quantitative metrics for uncertainty is not direct, several works focused on a qualitative assessment of the computed 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. Alternatively, the pertinence 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 (*e.g.* training-image selection in a semi-supervised learning, or improvement of the predictive performance).

4.2 Calibration protocol

As presented in Section 3.1, the output softmax probabilities of a NN can directly be used as a marker of (un)certainty. 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 error rates. Usual choices consist of the Expected Calibration Error (ECE) (Guo et al., 2017), the Brier Score, or the Negative Log-Likelihood (NLL) score.

4.3 Misclassification detection protocol

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 dramatic impact, especially in real-world medical image applications. In that sense, the uncertainty estimates can be turned into a binary classifier that aims at distinguishing between correct and incorrect predictions (*i.e.*, sample for which the predicted label y and the ground truth label z differ). As in the binary classification setting, an uncertainty threshold is applied to distinguish between positive (*i.e.* certain) and negative (*i.e.* uncertain) samples. The result of this classification is then compared to the true label 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 possibles cases, as shown in Figure 3. 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 (FP): 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 Negative (FN): the classification is uncertain but the prediction and the expected label are identical.

4.4 Rejection protocol

Another way of exploiting uncertainty estimates in an automated pipeline is the rejection mechanism. 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, and the performance of the model is computed on the remaining predictions. If uncertainty estimates efficiently identify uncertain cases that are more likely to be incorrect, then the performance and the remaining prediction should improve. Multiple fractions can be used, producing a curve showing the performance of the model with respect to the fraction of rejected data. The area under the resulting curve is used as a qualitative score.

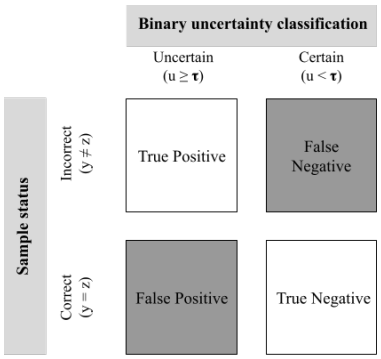

Figure 3: Confusion matrix for uncertainty-based misclassification detection. Desired cases are represented in white, while undesired cases are presented in gray.

This rejection-based evaluation protocol essentially highlights the same trends as the previous misclassification detection setting.

4.5 OOD detection protocol

A desired property of uncertainty is to be high for abnormal images that are different from the images seen during training. Similarly to the misclassification detection setting, the uncertainty estimates can be translated into a binary classifier that aim at distinguishing between in-distribution (ID) and OOD images. Standard classification metrics can further be computed from the confusion matrix, as presented above in Section 4.3.

4.6 Quality-control protocol

For segmentation tasks, uncertainty estimates are obtained for every pixel (or voxel) in the medical image. These scores can then be aggregated into image-wise uncertainty, for instance by taking their mean. The correlation between this image-wise uncertainty, and image-wise metrics quantifying the quality of the segmentation, such as the Dice score, can be computed. 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.

4.7 Label-distribution protocol

Finally, label-distribution protocol consists in comparing the predicted distribution of labels P_{out} with the ground truth distribution of the experts P_{gt}. A

popular choice of metric is the Generalized Energy Distance (Kohl et al., 2018) between both distributions.

5 Discussion

We reviewed the most popular UQ methods for DL-based medical image analysis and the associated evaluation protocols. In this section, we list the keys 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 the human-machine collaboration and break the black-box effect.

Bayesian methodology, although providing a strong theoretical background for uncertainty, is scarcely implemented for medical image analysis. This can be explained by the complex implementation that requires (i) the modification of the NN (weights are replaced by distributions, thus doubling the number of parameters to estimate) and (ii) the modification of the training paradigm. Additionally, convergence tends to be slow for complex scenarios (Osawa et al., 2019) and gradient descent more noisy and unstable using a Bayesian NN than with a standard non-deterministic NN (Jospin et al., 2022). Finally, it has also been observed that Bayesian NNs tend to underfit (Dusenberry et al., 2020) and that their predictive performances are lower than standard NNs (Wenzel et al., 2020). 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 nearly half of the implemented methods (44.73%, considering both the standard MC dropout methods (41.71%) as well as sampling Heteroscedastic models (3.02%), 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 preventing over-fitting during training, which is a common problem in medical domain, where training-dataset size is limited. However, the performance of MC dropout is highly dependent on the applied dropout rate (Osawa et al., 2019), which can makes it impractical to tune. Moreover, it requires multiple inferences for the same input image, considerably extending the inference time, which may not be compatible with AI applications in clinical environments.

Ensembling approaches are also commonly employed for UQ, although less common than MC dropout models. Aggregating the predictions of multiple models is a popular trick to improve the predictive performance, while also providing quality uncertainty estimates. The drawback is an increased computational cost and time as it requires multiple training and their predictions

aggregation at testing.

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 predicted one. However, most of the time, the corresponding uncertainty values are not provided. Thus, evaluation of UQ usually relies on proxy tasks, such as the detection of misclassification, Out-of-distribution, or Quality Control. These methods are inspired from concrete applications of uncertainty in a real-world scenario. Yet, although commonly used, UQ evaluation based on misclassification detection should not directly be used for ranking methods. Indeed, the set of correct and incorrect predictions is specific to each model that produces its own binary misclassification. It is then inappropriate to compare them directly (Ashukha et al., 2020).

In this review, we have distinguished between sampling and single-step UQ methods. The latter approaches offer to compute uncertainty in a quick and efficient manner, which is generally required for medical applications. Although they currently represent a niche, their easy practical considerations may promote their rapid adoption. These methods, however, only partially model uncertainty, by computing either aleatoric or epistemic uncertainties, but not both.

Finally, it must be acknowledged that the effort of the community is promoted by challenges, such as the 2020 edition of the BraTS challenge (Mehta et al., 2022; Menze et al., 2014) that included an UQ task, the MICCAI QUBIQ challenge[1] that focused on label uncertainty, and the SHIFT 2022 challenge[2] that will contain a task of uncertainty quantification for Multiple Sclerosis lesions segmentation (Malinin et al., 2022).

5.1 Future directions

Based on the gaps identified above, we suggest several future research directions for UQ in DL-based medical image analysis.

As shown, the vast majority (81.15%) of the implemented UQ methods are based on a sampling protocol, aiming at generating multiple predictions for the same query input. Yet, this significantly increases the computational burden of UQ, which may prevent its adoption in an automated pipeline in medical domain. Deterministic UQ methods requiring a single-step to compute uncertainty, are very promising and should be more intensively explored.

Overall, the detection of Out-of-distribution (OOD) predictions using uncertainty concerns few studies, despite of being crucial in a real-world medical scenarios. In an automated medical image pipeline, input samples can exhibit various anomalies that may disturb the 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 contrast setting due to a particular

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

acquisition protocol. In such situations, uncertainty associated to the computed predictions are expected to be high and should represent a warning for 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 (Snoek et al., 2019; Ulmer and Cinà, 2021). This motivates the development of feature-based methods, specially tailored for OOD detection. Note that OOD detection is a very active research topic not specific to the UQ field (Bulusu et al., 2020), but currently rarely exploited for medical image analysis. We hypothesize that the lack of OOD-based evaluation protocol may be due to the difficulty of gathering relevant data. A simple solution may be to use two distinct datasets, one for training the neural network and its evaluation on in-distribution data, and one during testing for OOD detection (Karimi and Gholipour, 2020). Another approach would be to use data augmentation to corrupt images with synthetic artifacts, helping to achieve a more realistic setting (Shaw et al., 2021).

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, helping the user to understand the sources of uncertainty in an intelligible way and possibly contribute to its improvement.

6 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 phenomenons can cause predictive uncertainty, such as noisy images, imperfect ground truth labels, lack 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 topics.

7 Declaration of competing interest

BL, AT, SD, HD are employees of the Pixyl company. MD and FF serve on Pixyl scientific advisory board.

8 Acknowledgments

Benjamin Lambert is supported by a CIFRE convention (ANRT 2020/1555). The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

A Classification of the reviewed papers with respect to the UQ methods implemented.

Study	Year	Modality	Applications
Softmax			
Wang et al. (2018)	2018	fetal MRI brain MRI	organ segmentation tumor segmentation
Jungo et al. (2020)	2020	brain MRI	tumor segmentation
Mojabi et al. (2020)	2020	breast US	tissue segmentation
DeVries and Taylor (2018)	2018	dermatoscopy	segmentation
Filos et al. (2019)	2019	retinal images	classification
Rousseau et al. (2021)	2020	brain MRI brain CT	tumor segmentation stroke segmentation
Mehta et al. (2020)	2020	brain MRI	tumor segmentation
Diao et al. (2022)	2022	brain MRI	tumor segmentation
Hoebel et al. (2020)	2020	lung CT	nodule segmentation
Lourenço-Silva and Oliveira (2022)	2021	MRI / CT	segmentation tasks
Carneiro et al. (2020)	2021	colonoscopy	polyps classification
Calderon-Ramirez et al. (2021b)	2021	chest X-Ray	COVID detection
Calderon-Ramirez et al. (2021a)	2021	X-Ray	mammogram classification
Berger et al. (2021)	2021	chest X-Ray	disease classification
Liang et al. (2020)	2020	CT / histology	classification tasks
Ayhan et al. (2020)	2020	retinal images	disease classification
Belharbi et al. (2021)	2020	histology	cancer cell segmentation
Lin et al. (2021)	2021	CT	multi-organ segmentation
Judge et al. (2022)	2022	US lung X-Ray	cardiac segmentation lung segmentation
Bayesian Neural Network			
Dhakal and Joshi (2021)	2021	knee MRI	classification
Filos et al. (2019)	2019	retinal images	classification
Li et al. (2021)	2021	lung CT nasal endoscopy	lesion segmentation
Monte Carlo dropout			
Jungo et al. (2018a)	2018	brain MRI	cavity segmentation
Zhang et al. (2022)	2022	lung CT abdominal CT	nodule segmentation tumor segmentation
Ghosal et al. (2021)	2021	biopsy	tumor segmentation
Ghoshal and Tucker (2020)	2020	lung CT	COVID detection
Yu et al. (2019)	2019	cardiac MRI	cardiac segmentation
Wickstrøm et al. (2020)	2018	colonoscopy	polyps segmentation

Ghoshal et al. (2021)	2021	microscopy brain MRI	nuclei segmentation tumor classification
Ozdemir et al. (2019)	2019	lung CT	nodule segmentation
Sander et al. (2019)	2018	cardiac MRI	segmentation
Eaton-Rosen et al. (2018)	2018	brain MRI	tumor segmentation
Abdar et al. (2021c)	2021	dermatoscopy	classification
Tousignant et al. (2019)	2019	brain MRI	classification
Jungo et al. (2020)	2020	brain MRI	tumor segmentation
Balagopal et al. (2021)	2021	prostate CT	target segmentation
Hu et al. (2020)	2020	brain CT/PET	tumor segmentation
Xia et al. (2020)	2020	CT	segmentation
Sedai et al. (2019)	2019	retinal image	layer segmentation
Wang et al. (2019a)	2019	brain MRI	brain segmentation tumor segmentation
Mehrtash et al. (2020)	2020	MRI	segmentation
Roy et al. (2019)	2018	brain MRI	brain segmentation
DeVries and Taylor (2018)	2018	dermatoscopy	segmentation
Liu et al. (2020)	2020	prostate MRI	segmentation
Dhakal and Joshi (2021)	2021	knee MRI	classification
Jungo et al. (2017)	2017	brain MRI	tumor segmentation
Rączkowski et al. (2019)	2019	microscopy	classification
Filos et al. (2019)	2019	retinal image	classification
Abideen et al. (2020)	2020	chest CT	tuberculosis detection
Thagaard et al. (2020)	2020	histology	metastasis detection
Jungo et al. (2018b)	2018	brain MRI	cavity segmentation
Hiasa et al. (2019)	2019	muscle CT	muscle segmentation
Orlando et al. (2019)	2019	retinal images	layer segmentation
Leibig et al. (2017)	2017	retinal images	classification
Kwon et al. (2020)	2020	retinal images brain MRI	vessel segmentation stroke segmentation
Rousseau et al. (2021)	2020	brain MRI brain CT	tumor segmentation stroke segmentation
Mehta et al. (2020)	2020	brain MRI	tumor segmentation
Asgharnezhad et al. (2022)	2022	lung CT	COVID detection
Yang et al. (2021)	2020	lung CT	nodule detection
Ozdemir et al. (2017)	2017	lung CT	nodule detection
Soberanis-Mukul et al. (2020)	2019	CT	segmentation
Pan et al. (2019)	2019	prostate MRI	segmentation
Lee et al. (2022)	2022	brain MRI	tumor segmentation
Bhat et al. (2021)	2021	liver MRI	metastase segmentation
Huang et al. (2020)	2020	cardiac OCT	tissue segmentation
Iwamoto et al. (2021)	2021	microscopy	segmentation
McClure et al. (2019)	2019	brain MRI	atlas segmentation
Natekar et al. (2020)	2019	brain MRI	tumor segmentation
Herzog et al. (2020)	2020	brain MRI	stroke classification
Mehta et al. (2019)	2019	brain MRI	MS lesion segmentation tumor segmentation
Molle et al. (2019)	2019	dermatoscopy	classification
Zou et al. (2022)	2022	brain MRI	tumor segmentation
Diao et al. (2022)	2022	brain MRI	tumor segmentation
Hoebel et al. (2020)	2020	lung CT	nodule segmentation
Karimi and Gholipour (2020)	2020	MRI / CT	segmentation
Redekop and Chernyavskiy (2021)	2020	CT dermatoscopy	segmentation

Bhat and Kuijf (2022)	2021	MRI/CT	segmentation tasks
Song et al. (2021)	2021	tongue images	oral cancer classification
Gou and He (2021)	2021	head CT	detection of subarachnoid hemorrhages
Carneiro et al. (2020)	2021	colonoscopy	polyps classification
Mahapatra et al. (2021)	2021	chest X-Ray histology	disease classification gland segmentation
Hasan and Linte (2021)	2022	cardiac MRI	segmentation
Mojiri Forooshani et al. (2022)	2022	brain MRI	White Matter Hyperintensities segmentation
Laves et al. (2019)	2019	retina images	disease classification
Mobiny et al. (2019)	2019	skin images	disease classification
Calderon-Ramirez et al. (2021b)	2021	chest X-Ray	COVID detection
Calderon-Ramirez et al. (2021a)	2021	X-Ray	mammogram classification
Linmans et al. (2020)	2020	microscopy	segmentation of breast cancer metastasis
Berger et al. (2021)	2021	chest X-Ray	disease classification
Ayhan et al. (2020)	2020	retina images	disease classification
Ju et al. (2022)	2021	dermatoscopy	disease classification
Hasan and Linte (2022)	2022	cardiac MRI	segmentation
Jiménez-Sánchez et al. (2022)	2020	femur X-Ray	fracture classification
Cao et al. (2021)	2019	breast ultrasound	breast mass segmentation
Pocevičiūtė et al. (2022)	2022	microscopy	cancer classification
Rajaraman et al. (2022)	2022	chest X-Ray	tuberculosis consistent region segmentation
Senousy et al. (2021)	2021	histology	breast cancer classification
Ahsan et al. (2022)	2022	retina images	disease classification
Javadi et al. (2022)	2022	prostate ultrasound	cancer detection
Tardy et al. (2019)	2019	X-Ray	mammogram classification
Yang and Fevens (2021)	2021	CT / histology	disease classification
Jensen et al. (2019)	2019	dermatoscopy	disease classification
Lambert et al. (2022)	2022	brain MRI	MS lesions segmentation
Judge et al. (2022)	2022	US lung X-Ray	cardiac segmentation lung segmentation
Zhao et al. (2022)	2022	MRI	cardiac segmentation
Ensemble			
Yang et al. (2017)	2017	histology US	gland / lymph nodes segmentation
Abdar et al. (2021c)	2021	dermatoscopy	cancer classification
Jungo et al. (2020)	2020	brain MRI	tumor segmentation
Shamsi et al. (2021)	2021	lung CT	COVID detection
Mehrtash et al. (2020)	2020	MRI	segmentation tasks
Filos et al. (2019)	2019	retinal images	classification
Thagaard et al. (2020)	2020	histology	metastasis detection
Mehta et al. (2020)	2020	brain MRI	tumor segmentation
Asgharnezhad et al. (2022)	2022	lung CT	COVID detection
Vu et al. (2020)	2020	brain MRI	tumor segmentation
Zhou et al. (2022)	2022	brain MRI	tumor segmentation
Yang et al. (2022)	2022	CT / MRI	segmentation tasks
Hoebel et al. (2020)	2020	lung CT	nodule segmentation
Mehrtash et al. (2021)	2021	prostate MRI	cancer classification
Redekop and Chernyavskiy (2021)	2021	liver CT dermatoscopy	segmentation

Cetindag et al. (2022)	2021	MRI/CT	segmentation tasks
Pal (2022)	2021	MRI/CT	segmentation tasks
Wang et al. (2020)	2020	retinal images	disease classification
Linmans et al. (2020)	2020	microscopy	segmentation of breast cancer metastasis
Berger et al. (2021)	2021	chest X-Ray	disease classification
Ghesu et al. (2019)	2019	Chest X-Ray	classification
Ghesu et al. (2021)	2019	Chest X-Ray brain MRI abdominal US	disease classification detection of metastases view classification
Ayhan et al. (2020)	2020	retinal images	disease classification
Guo et al. (2022)	2022	cardiac MRI	segmentation
Rosas-Gonzalez et al. (2021)	2021	brain MRI	tumor segmentation
Pocevičiūtė et al. (2022)	2022	microscopy	cancer classification
Yang and Fevens (2021)	2021	CT / histology	disease classification
Jensen et al. (2019)	2019	dermatoscopy	disease classification
Xiang et al. (2022)	2022	CT MRI	pancreas segmentation cardiac segmentation
Kushibar et al. (2022)	2022	mammogram	mass segmentation
Zhao et al. (2022)	2022	MRI	cardiac segmentation
Monte Carlo dropout Ensemble			
Ghoshal et al. (2021)	2021	microscopy brain MRI	nuclei segmentation tumor classification
Abdar et al. (2021c)	2021	dermatoscopy	cancer classification
Filos et al. (2019)	2019	retinal images	classification
Mehta et al. (2020)	2020	brain MRI	tumor segmentation
Asgharnezhad et al. (2022)	2022	lung CT	COVID detection
Abdar et al. (2021b)	2022	chest X-RAY / CT	COVID detection
Yang and Fevens (2021)	2021	CT / histology	disease classification
Heteroscedastic (Sampling)			
Nair et al. (2020)	2020	brain MRI	MS lesion segmentation
Eaton-Rosen et al. (2018)	2018	brain MRI	tumor segmentation
Jungo et al. (2020)	2020	brain MRI	tumor segmentation
Sedai et al. (2018)	2018	retinal images	retinal layer segmentation
DeVries and Taylor (2018)	2018	dermatoscopy	segmentation
Shaw et al. (2021)	2021	brain MRI	brain segmentation
Heteroscedastic (Deterministic)			
McKinley et al. (2020a)	2020	brain MRI	tumor segmentation
McKinley et al. (2020b)	2020	brain MRI	new MS lesion segmentation
McKinley et al. (2018)	2018	brain MRI	tumor segmentation
McKinley et al. (2019)	2019	brain MRI	tumor segmentation
DeVries and Taylor (2018)	2018	dermatoscopy	segmentation
Diao et al. (2022)	2022	brain MRI	tumor segmentation
Judge et al. (2022)	2022	US lung X-Ray	cardiac segmentation lung segmentation
Label Distribution models			
Kohl et al. (2018)	2018	lung CT	nodule segmentation
Kohl et al. (2019)	2019	lung CT microscopy	nodule segmentation neocortex segmentation
Baumgartner et al. (2019)	2019	lung CT prostate MRI	nodule segmentation prostate segmentation

Hu et al. (2019)	2019	lung CT prostate MRI	nodule segmentation prostate segmentation
Li and Luo (2020)	2020	prostate MRI	prostate segmentation
Gantenbein et al. (2020)	2020	lung CT prostate MRI	nodule segmentation prostate segmentation
Monteiro et al. (2020)	2020	lung CT brain MRI	nodule segmentation tumor segmentation
Zou et al. (2022)	2022	brain MRI	tumor segmentation
Diao et al. (2022)	2022	brain MRI	tumor segmentation
Selvan et al. (2020)	2020	lung CT retinal image	nodule segmentation vessel segmentation
Cetindag et al. (2022)	2021	MRI/CT	segmentation tasks
Ji et al. (2020)	2020	MRI/CT	segmentation tasks
Bhat and Kuijf (2022)	2021	MRI/CT	segmentation tasks
Test-Time Augmentation			
Wang et al. (2019a)	2019	brain MRI	fetal brain segmentation tumor segmentation
Norouzi et al. (2019)	2019	cardiac MRI	cardiac segmentation
Ayhan and Berens (2018)	2018	retinal images	classification
Pan et al. (2019)	2019	prostate MRI	prostate segmentation
Diao et al. (2022)	2022	brain MRI	tumor segmentation
Redekop and Chernyavskiy (2021)	2021	liver CT dermatoscopy	segmentation
Combalia et al. (2020)	2020	dermatoscopy	lesion classification
Ayhan et al. (2020)	2020	retinal images	disease classification
Wang et al. (2019b)	2019	brain MRI	tumor segmentation
Pocevičiūtė et al. (2022)	2022	microscopy	cancer classification
Javadi et al. (2022)	2022	prostate US	cancer detection
Jensen et al. (2019)	2019	dermatoscopy	disease classification
Feature-based methods			
Karimi and Gholipour (2020)	2020	CT, MRI	segmentation tasks
Diao et al. (2022)	2022	brain MRI	tumor segmentation
Calderon-Ramirez et al. (2021b)	2021	chest X-Ray	COVID detection
Berger et al. (2021)	2021	chest X-Ray	disease classification
Tardy et al. (2019)	2019	X-Ray	mammogram classification
Evidential Deep Learning			
Ghesu et al. (2019)	2019	chest X-Ray	classification
Ghesu et al. (2021)	2019	chest X-Ray brain MRI Abdominal US	disease classification detection of metastases view classification
Tardy et al. (2019)	2019	X-Ray	mammogram classification
Huang et al. (2021)	2021	PET / CT	lymphomas segmentation
Zou et al. (2022)	2020	brain MRI	tumor segmentation

B Uncategorized approaches for UQ in medical image processing applications.

Study	Year	Modality	Application
Jungo et al. (2020)	2020	brain MRI	Training of an auxiliary net to predict the voxel-wise errors of a brain tumor segmentation model.
Mishra et al. (2021)	2021	retina images	Modeling task-dependent (homoscedastic) uncertainty in a multi-tasking vessel segmentation setting.
Föllmer et al. (2022)	2021	heart CT	Modeling task-dependent (homoscedastic) uncertainty in a multi-tasking heart-segmentation setting.
Laves et al. (2019)	2019	retina images	Approximating the output posterior distribution of the classification network by a normal distribution $\mathcal{N}(\mu, \sigma^2)$ and learning both parameters using a variational network.
Toledo-Cortés et al. (2020)	2020	retina images	Addition of a Gaussian Process at the end of a DL model to quantify classification uncertainty.
Jensen et al. (2019)	2019	dermatoscopy	Use of Monte Carlo Batch Normalization (MCBN) to quantify uncertainty by relying on the stochasticity of Batch Normalization layers.
Judge et al. (2022)	2022	US lung X-Ray	Use of Contrastive Learning to model the joint distribution of valid segmentations and associated images.
Lu et al. (2022)	2022	MRI	Use Conformal Predictions to provide a set of plausible classes for a given image with coverage guarantees.

C Evaluation protocols proposed in the corpus of papers and classified according to their framework.

Evaluation protocol	Papers
Qualitative Assessment	Wang et al. (2018), Yu et al. (2019), Wickstrøm et al. (2020), Eaton-Rosen et al. (2018), Mojabi et al. (2020), Shamsi et al. (2021), Sedai et al. (2018), Norouzi et al. (2019), Hu et al. (2020), Xia et al. (2020), Sedai et al. (2019), Li and Luo (2020), Dhakal and Joshi (2021), Jungo et al. (2017), Kwon et al. (2020), Mishra et al. (2021), Soberanis-Mukul et al. (2020), Lee et al. (2022), Bhat et al. (2021), Huang et al. (2020), McKinley et al. (2018), McKinley et al. (2019), McKinley et al. (2020b), Natekar et al. (2020), Mehta et al. (2019), Li et al. (2021), Mehrtash et al. (2021), Redekop and Chernyavskiy (2021), Wang et al. (2020), Mahapatra et al. (2021), Hasan and Linte (2021), Mojiri Forooshani et al. (2022), Laves et al. (2019), Toledo-Cortés et al. (2020), Lin et al. (2021),Guo et al. (2022), Ju et al. (2022), Belharbi et al. (2021),Hasan and Linte (2022), Jiménez-Sánchez et al. (2022), Cao et al. (2021),Senousy et al. (2021), Huang et al. (2021),Xiang et al. (2022)
Calibration	Ghoshal et al. (2021), Ozdemir et al. (2019), Sander et al. (2019), Jungo et al. (2020), Mehrtash et al. (2020), Thagaard et al. (2020), Rousseau et al. (2021), Asgharnezhad et al. (2022), Ozdemir et al. (2017),Karimi and Gholipour (2020), Zou et al. (2022), Herzog et al. (2020),Gou and He (2021), Carneiro et al. (2020), Berger et al. (2021),Liang et al. (2020),Ayhan et al. (2020), Javadi et al. (2022), Jensen et al. (2019),Judge et al. (2022),Kushibar et al. (2022),Zhao et al. (2022)
Misclassification detection	Ghoshal and Tucker (2020), Yang et al. (2017), Ghoshal et al. (2021), Abdar et al. (2021c), Jungo et al. (2020), Wang et al. (2019a), Rączkowski et al. (2019), Thagaard et al. (2020), Asgharnezhad et al. (2022), Iwamoto et al. (2021), McClure et al. (2019),Molle et al. (2019),Zou et al. (2022), Mobiny et al. (2019), Calderon-Ramirez et al. (2021b), Calderon-Ramirez et al. (2021a), Pocevičiūtė et al. (2022), Ahsan et al. (2022),Abdar et al. (2021b),Judge et al. (2022)
Rejection	Nair et al. (2020), Zhang et al. (2022), Ghoshal et al. (2021), Ozdemir et al. (2019), Sander et al. (2019),Abdar et al. (2021c),Tousignant et al. (2019), Filos et al. (2019),Abideen et al. (2020), Ayhan and Berens (2018),Leibig et al. (2017),Mehta et al. (2020),McKinley et al. (2020a),Yang et al. (2021),Vu et al. (2020), Herzog et al. (2020),Diao et al. (2022), Song et al. (2021),Carneiro et al. (2020),Mobiny et al. (2019), Ghesu et al. (2019),Ghesu et al. (2021), Combalia et al. (2020), Ayhan et al. (2020),Rosas-Gonzalez et al. (2021), Rajaraman et al. (2022), Tardy et al. (2019),Yang and Fevens (2021),Lambert et al. (2022)

OOD detection	Karimi and Gholipour (2020),Diao et al. (2022),Thagaard et al. (2020), Linmans et al. (2020), Berger et al. (2021), Combalia et al. (2020), Tardy et al. (2019)
Quality Control	Ghosal et al. (2021), Balagopal et al. (2021), Wang et al. (2019a),Mehrtash et al. (2020), Roy et al. (2019), DeVries and Taylor (2018), Jungo et al. (2018b), Jungo et al. (2020), Hiasa et al. (2019),Orlando et al. (2019), Pan et al. (2019), McClure et al. (2019), Hoebel et al. (2020), Shaw et al. (2021), Rosas-Gonzalez et al. (2021), Wang et al. (2019b),Judge et al. (2022),Kushibar et al. (2022)
Label distribution	Jungo et al. (2018a), Baumgartner et al. (2019), Hu et al. (2019), Li and Luo (2020), Kohl et al. (2018), Kohl et al. (2018),Gantenbein et al. (2020), Monteiro et al. (2020), Selvan et al. (2020), Yang et al. (2022),Cetindag et al. (2022), Bhat and Kuijf (2022), Pal (2022), Lourenço-Silva and Oliveira (2022)

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., et al., 2021a. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion 76, 243–297.

Abdar, M., Salari, S., Qahremani, S., Lam, H.K., Karray, F., Hussain, S., Khosravi, A., Acharya, U.R., Nahavandi, S., 2021b. Uncertaintyfusenet: Robust uncertainty-aware hierarchical feature fusion with ensemble monte carlo dropout for covid-19 detection. arXiv e-prints .

Abdar, M., Samami, M., Mahmoodabad, S.D., Doan, T., Mazoure, B., Hashemifesharaki, R., Liu, L., Khosravi, A., Acharya, U.R., Makarenkov, V., et al., 2021c. Uncertainty quantification in skin cancer classification using three-way decision-based bayesian deep learning. Computers in biology and medicine 135, 104418.

Abideen, Z.U., Ghafoor, M., Munir, K., Saqib, M., Ullah, A., Zia, T., Tariq, S.A., Ahmed, G., Zahra, A., 2020. Uncertainty assisted robust tuberculosis identification with bayesian convolutional neural networks. IEEE Access 8, 22812–22825.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I.J., Hardt, M., Kim, B., 2018. Sanity checks for saliency maps. Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018 , 9525–9536.

Ahsan, M.A., Qayyum, A., Razi, A., Qadir, J., 2022. An active learning method for diabetic retinopathy classification with uncertainty quantification. Medical & Biological Engineering & Computing , 1–15.

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al., 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information fusion 58, 82–115.

Asgharnezhad, H., Shamsi, A., Alizadehsani, R., Khosravi, A., Nahavandi, S., Sani, Z.A., Srinivasan, D., Islam, S.M.S., 2022. Objective evaluation of deep uncertainty predictions for covid-19 detection. Scientific Reports 12, 1–11.

Ashukha, A., Lyzhov, A., Molchanov, D., Vetrov, D.P., 2020. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. 8th International Conference on Learning Representations, ICLR 2020 .

Ayhan, M.S., Berens, P., 2018. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. International conference on Medical Imaging with Deep Learning .

Ayhan, M.S., Kühlewein, L., Aliyeva, G., Inhoffen, W., Ziemssen, F., Berens, P., 2020. Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection. Medical Image Analysis 64, 101724.

Balagopal, A., Nguyen, D., Morgan, H., Weng, Y., Dohopolski, M., Lin, M.H., Barkousaraie, A.S., Gonzalez, Y., Garant, A., Desai, N., et al., 2021. A deep learning-based framework for segmenting invisible clinical target volumes with estimated uncertainties for post-operative prostate cancer radiotherapy. Medical image analysis 72, 102101.

Baumgartner, C.F., Tezcan, K.C., Chaitanya, K., Hötker, A.M., Muehlematter, U.J., Schawkat, K., Becker, A.S., Donati, O., Konukoglu, E., 2019. Phiseg: Capturing uncertainty in medical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention , 119–127.

Becker, A.S., Chaitanya, K., Schawkat, K., Muehlematter, U.J., Hötker, A.M., Konukoglu, E., Donati, O.F., 2019. Variability of manual segmentation of the prostate in axial t2-weighted mri: A multi-reader study. European journal of radiology 121, 108716.

Belharbi, S., Rony, J., Dolz, J., Ayed, I.B., McCaffrey, L., Granger, E., 2021. Deep interpretable classification and weakly-supervised segmentation of histology images via max-min uncertainty. IEEE Transactions on Medical Imaging 41, 702–714.

Berger, C., Paschali, M., Glocker, B., Kamnitsas, K., 2021. Confidence-based out-of-distribution detection: a comparative study and analysis. Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis , 122–132.

Bhat, I., Kuijf, H.J., 2022. Extending probabilistic u-net using mc-dropout to quantify data and model uncertainty. International MICCAI Brainlesion Workshop , 555–559.

Bhat, I., Kuijf, H.J., Cheplygina, V., Pluim, J.P., 2021. Using uncertainty estimation to reduce false positives in liver lesion detection. 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) , 663–667.

Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D., 2015. Weight uncertainty in neural network. International Conference on Machine Learning , 1613–1622.

Bulusu, S., Kailkhura, B., Li, B., Varshney, P.K., Song, D., 2020. Anomalous example detection in deep learning: A survey. IEEE Access 8, 132330–132347.

Calderon-Ramirez, S., Murillo-Hernandez, D., Rojas-Salazar, K., Calvo-Valverd, L.A., Yang, S., Moemeni, A., Elizondo, D., Lopez-Rubio, E., Molina-Cabello, M.A., 2021a. Improving uncertainty estimations for mammogram classification using semi-supervised learning. 2021 International Joint Conference on Neural Networks (IJCNN) , 1–8.

Calderon-Ramirez, S., Yang, S., Moemeni, A., Colreavy-Donnelly, S., Elizondo, D.A., Oala, L., Rodríguez-Capitán, J., Jiménez-Navarro, M., López-Rubio, E., Molina-Cabello, M.A., 2021b. Improving uncertainty estimation with semi-supervised deep learning for covid-19 detection using chest x-ray images. IEEE Access 9, 85442–85454.

Cao, X., Chen, H., Li, Y., Peng, Y., Wang, S., Cheng, L., 2021. Dilated densely connected u-net with uncertainty focus loss for 3d abus mass segmentation. Computer Methods and Programs in Biomedicine 209, 106313.

Carneiro, G., Pu, L.Z.C.T., Singh, R., Burt, A., 2020. Deep learning uncertainty and confidence calibration for the five-class polyp classification from colonoscopy. Medical Image Analysis 62, 101653.

Cetindag, S.C., Yergin, M., Alis, D., Oksuz, I., 2022. Meta-learning for medical image segmentation uncertainty quantification. International MICCAI Brainlesion Workshop , 578–584.

Chen, C., Qin, C., Qiu, H., Ouyang, C., Wang, S., Chen, L., Tarroni, G., Bai, W., Rueckert, D., 2020. Realistic adversarial data augmentation for mr image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention , 667–677.

Combalia, M., Hueto, F., Puig, S., Malvehy, J., Vilaplana, V., 2020. Uncertainty estimation in deep neural networks for dermoscopic image classification. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops , 744–745.

Dempster, A.P., 1968. A generalization of bayesian inference. Journal of the Royal Statistical Society: Series B (Methodological) 30, 205–232.

DeVries, T., Taylor, G.W., 2018. Leveraging uncertainty estimates for predicting segmentation quality. arXiv e-prints .

Dhakal, P., Joshi, S.R., 2021. Uncertainty estimation in detecting knee abnormalities on mri using bayesian deep learning. Proceedings of 10th IOE Graduate Conference 10.

Diao, Z., Jiang, H., Shi, T., 2022. A unified uncertainty network for tumor segmentation using uncertainty cross entropy loss and prototype similarity. Knowledge-Based Systems 246, 108739.

Dusenberry, M., Jerfel, G., Wen, Y., Ma, Y., Snoek, J., Heller, K., Lakshminarayanan, B., Tran, D., 2020. Efficient and scalable bayesian neural nets with rank-1 factors. International Conference on Machine Learning , 2782–2792.

Eaton-Rosen, Z., Bragman, F., Bisdas, S., Ourselin, S., Cardoso, M.J., 2018. Towards safe deep learning: accurately quantifying biomarker uncertainty in neural network predictions. International Conference on Medical Image Computing and Computer-Assisted Intervention , 691–699.

Fidon, L., Li, W., Garcia-Peraza-Herrera, L.C., Ekanayake, J., Kitchen, N., Ourselin, S., Vercauteren, T., 2017. Generalised wasserstein dice score for imbalanced multi-class segmentation using holistic convolutional networks. International MICCAI Brainlesion workshop , 64–76.

Filos, A., Farquhar, S., Gomez, A.N., Rudner, T.G., Kenton, Z., Smith, L., Alizadeh, M., De Kroon, A., Gal, Y., 2019. A systematic comparison of bayesian deep learning robustness in diabetic retinopathy tasks. arXiv e-prints .

Föllmer, B., Biavati, F., Wald, C., Stober, S., Ma, J., Dewey, M., Samek, W., 2022. Active multi-task learning with uncertainty weighted loss for coronary calcium scoring. Medical Physics .

Ford, R.A., Price, W., Nicholson, I., 2016. Privacy and accountability in black-box medicine. Mich. Telecomm. & Tech. L. Rev. 23, 1.

Gal, Y., Ghahramani, Z., 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. International Conference on Machine Learning , 1050–1059.

Gal, Y., et al., 2016. Uncertainty in deep learning. Ph.D. thesis. University of Cambridge.

Gantenbein, M., Erdil, E., Konukoglu, E., 2020. Revphiseg: A memory-efficient neural network for uncertainty quantification in medical image segmentation. Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis , 13–22.

Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al., 2021. A survey of uncertainty in deep neural networks. arXiv e-prints .

Ghafoorian, M., Mehrtash, A., Kapur, T., Karssemeijer, N., Marchiori, E., Pesteie, M., Guttmann, C.R., Leeuw, F.E.d., Tempany, C.M., Ginneken, B.v., et al., 2017. Transfer learning for domain adaptation in mri: Application in brain lesion segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention , 516–524.

Ghesu, F.C., Georgescu, B., Gibson, E., Guendel, S., Kalra, M.K., Singh, R., Digumarthy, S.R., Grbic, S., Comaniciu, D., 2019. Quantifying and leveraging classification uncertainty for chest radiograph assessment. International Conference on Medical Image Computing and Computer-Assisted Intervention , 676–684.

Ghesu, F.C., Georgescu, B., Mansoor, A., Yoo, Y., Gibson, E., Vishwanath, R., Balachandran, A., Balter, J.M., Cao, Y., Singh, R., et al., 2021. Quantifying and leveraging predictive uncertainty for medical image assessment. Medical Image Analysis 68, 101855.

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

Ghoshal, B., Tucker, A., 2020. Estimating uncertainty and interpretability in deep learning for coronavirus (covid-19) detection. arXiv e-prints .

Ghoshal, B., Tucker, A., Sanghera, B., Lup Wong, W., 2021. Estimating uncertainty in deep learning for reporting confidence to clinicians in medical image segmentation and diseases detection. Computational Intelligence 37, 701–734.

Gou, X., He, X., 2021. Deep learning-based detection and diagnosis of subarachnoid hemorrhage. Journal of Healthcare Engineering 2021.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks. International Conference on Machine Learning , 1321–1330.

Guo, F., Ng, M., Kuling, G., Wright, G., 2022. Cardiac mri segmentation with sparse annotations: Ensembling deep learning uncertainty and shape priors. Medical Image Analysis , 102532.

Hasan, S.K., Linte, C.A., 2021. A multi-task cross-task learning architecture for ad hoc uncertainty estimation in 3d cardiac mri image segmentation. 2021 Computing in Cardiology (CinC) 48, 1–4.

Hasan, S.K., Linte, C.A., 2022. Calibration of cine mri segmentation probability for uncertainty estimation using a multi-task cross-task learning architecture. Medical Imaging 2022: Image-Guided Procedures, Robotic Interventions, and Modeling 12034, 174–179.

Herzog, L., Murina, E., Dürr, O., Wegener, S., Sick, B., 2020. Integrating uncertainty in deep neural networks for mri based stroke analysis. Medical Image Analysis 65, 101790.

Hiasa, Y., Otake, Y., Takao, M., Ogawa, T., Sugano, N., Sato, Y., 2019. Automated muscle segmentation from clinical ct using bayesian u-net for personalized musculoskeletal modeling. IEEE Transactions on Medical Imaging 39, 1030–1040.

Hoebel, K., Andrearczyk, V., Beers, A., Patel, J., Chang, K., Depeursinge, A., Müller, H., Kalpathy-Cramer, J., 2020. An exploration of uncertainty information for segmentation quality assessment. Medical Imaging 2020: Image Processing 11313, 381–390.

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

Hu, X., Guo, R., Chen, J., Li, H., Waldmannstetter, D., Zhao, Y., Li, B., Shi, K., Menze, B., 2020. Coarse-to-fine adversarial networks and zone-based uncertainty analysis for nk/t-cell lymphoma segmentation in ct/pet images. IEEE journal of biomedical and health informatics 24, 2599–2608.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition , 4700–4708.

Huang, L., Ruan, S., Decazes, P., Denoeux, T., 2021. Evidential segmentation of 3d pet/ct images. International Conference on Belief Functions , 159–167.

Huang, Z., Gan, Y., Lye, T., Zhang, H., Laine, A., Angelini, E.D., Hendon, C., 2020. Heterogeneity measurement of cardiac tissues leveraging uncertainty information from image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention , 782–791.

Hüllermeier, E., Waegeman, W., 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Machine Learning 110, 457–506.

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods 18, 203–211.

Iwamoto, S., Raytchev, B., Tamaki, T., Kaneda, K., 2021. Improving the reliability of semantic segmentation of medical images by uncertainty modeling with bayesian deep networks and curriculum learning. Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis , 34–43.

Javadi, G., Bayat, S., Kazemi Esfeh, M.M., Samadi, S., Sedghi, A., Sojoudi, S., Hurtado, A., Chang, S., Black, P., Mousavi, P., et al., 2022. Towards targeted ultrasound-guided prostate biopsy by incorporating model and label uncertainty in cancer detection. International Journal of Computer Assisted Radiology and Surgery 17, 121–128.

Jensen, M.H., Jørgensen, D.R., Jalaboi, R., Hansen, M.E., Olsen, M.A., 2019. Improving uncertainty estimation in convolutional neural networks using inter-rater agreement. International Conference on Medical Image Computing and Computer-Assisted Intervention , 540–548.

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

Jiménez-Sánchez, A., Mateus, D., Kirchhoff, S., Kirchhoff, C., Biberthaler, P., Navab, N., Ballester, M.A.G., Piella, G., 2022. Curriculum learning for improved femur fracture classification: Scheduling data with prior knowledge and uncertainty. Medical Image Analysis 75, 102273.

Joskowicz, L., Cohen, D., Caplan, N., Sosna, J., 2019. Inter-observer variability of manual contour delineation of structures in ct. European radiology 29, 1391–1399.

Jospin, L.V., Laga, H., Boussaid, F., Buntine, W., Bennamoun, M., 2022. Hands-on bayesian neural networks—a tutorial for deep learning users. IEEE Computational Intelligence Magazine 17, 29–48.

Ju, L., Wang, X., Wang, L., Mahapatra, D., Zhao, X., Zhou, Q., Liu, T., Ge, Z., 2022. Improving medical images classification with label noise using dual-uncertainty estimation. IEEE Trans. Medical Imaging 41, 1533–1546.

Judge, T., Bernard, O., Porumb, M., Chartsias, A., Beqiri, A., Jodoin, P.M., 2022. Crisp - reliable uncertainty estimation for medical image segmentation , 492–502.

Jungo, A., Balsiger, F., Reyes, M., 2020. Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. Frontiers in neuroscience , 282.

Jungo, A., McKinley, R., Meier, R., Knecht, U., Vera, L., Pérez-Beteta, J., Molina-García, D., Pérez-García, V.M., Wiest, R., Reyes, M., 2017. Towards uncertainty-assisted brain tumor segmentation and survival prediction. International MICCAI Brainlesion Workshop , 474–485.

Jungo, A., Meier, R., Ermis, E., Blatti-Moreno, M., Herrmann, E., Wiest, R., Reyes, M., 2018a. On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention , 682–690.

Jungo, A., Meier, R., Ermis, E., Herrmann, E., Reyes, M., 2018b. Uncertainty-driven sanity check: Application to postoperative brain tumor cavity segmentation. arXiv e-prints .

Kabir, H.D., Khosravi, A., Hosen, M.A., Nahavandi, S., 2018. Neural network-based uncertainty quantification: A survey of methodologies and applications. IEEE Access 6, 36218–36234.

Karimi, D., Gholipour, A., 2020. Improving calibration and out-of-distribution detection in medical image segmentation with convolutional neural networks. arXiv e-prints .

Kendall, A., Gal, Y., 2017. What uncertainties do we need in bayesian deep learning for computer vision? Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017 , 5574–5584.

Kerfoot, E., Clough, J., Oksuz, I., Lee, J., King, A.P., Schnabel, J.A., 2018. Left-ventricle quantification using residual u-net. International Workshop on Statistical Atlases and Computational Models of the Heart , 371–380.

Kohl, S., Romera-Paredes, B., Meyer, C., Fauw, J.D., Ledsam, J.R., Maier-Hein, K.H., Eslami, S.M.A., Rezende, D.J., Ronneberger, O., 2018. A probabilistic u-net for segmentation of ambiguous images , 6965–6975.

Kohl, S.A., Romera-Paredes, B., Maier-Hein, K.H., Rezende, D.J., Eslami, S., Kohli, P., Zisserman, A., Ronneberger, O., 2019. A hierarchical probabilistic u-net for modeling multi-scale ambiguities. arXiv e-prints .

Kumar, A., Liang, P., Ma, T., 2019. Verified uncertainty calibration. Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019 , 3787–3798.

Kurz, A., Hauser, K., Mehrtens, H.A., Krieghoff-Henning, E., Hekler, A., Kather, J.N., Fröhling, S., von Kalle, C., Brinker, T.J., et al., 2022. Uncertainty estimation in medical image classification: Systematic review. JMIR Medical Informatics 10, e36427.

Kushibar, K., Campello, V., Garrucho, L., Linardos, A., Radeva, P., Lekadir, K., 2022. Layer ensembles: A single-pass uncertainty estimation in deep learning for segmentation , 514–524.

Kwon, Y., Won, J.H., Kim, B.J., Paik, M.C., 2020. Uncertainty quantification using bayesian neural networks in classification: Application to biomedical image segmentation. Computational Statistics & Data Analysis 142, 106816.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017 , 6402–6413.

Lambert, B., Forbes, F., Doyle, S., Tucholka, A., Dojat, M., 2022. Beyond voxel prediction uncertainty: Identifying brain lesions you can trust 13611.

Laves, M.H., Ihler, S., Ortmaier, T., 2019. Uncertainty quantification in computer-aided diagnosis: Make your model say" i don't know" for ambiguous cases. arXiv e-prints .

Lee, J., Shin, D., Oh, S.H., Kim, H., 2022. Method to minimize the errors of ai: Quantifying and exploiting uncertainty of deep learning in brain tumor segmentation. Sensors 22, 2406.

Leibig, C., Allken, V., Ayhan, M.S., Berens, P., Wahl, S., 2017. Leveraging uncertainty information from deep neural networks for disease detection. Scientific reports 7, 1–14.

Li, H., Luo, H., 2020. Uncertainty quantification in medical image segmentation. 2020 IEEE 6th International Conference on Computer and Communications (ICCC) , 1936–1940.

Li, Y., Chen, X., Quan, L., Zhang, N., 2021. Uncertainty-guided robust training for medical image segmentation. 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) , 1471–1475.

Liang, G., Zhang, Y., Jacobs, N., 2020. Neural network calibration for medical imaging classification using dca regularization. International Conference on Machine Learning, Workshop on Uncertainty & Robustness in Deep Learning .

Lin, H., Li, Z., Yang, Z., Wang, Y., 2021. Variance-aware attention u-net for multi-organ segmentation. Medical Physics 48, 7864–7876.

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2020. Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 42, 318–327.

Linmans, J., van der Laak, J., Litjens, G., 2020. Efficient out-of-distribution detection in digital pathology using multi-head convolutional neural networks. Medical Imaging with Deep Learning , 465–478.

Liu, Y., Yang, G., Hosseiny, M., Azadikhah, A., Mirak, S.A., Miao, Q., Raman, S.S., Sung, K., 2020. Exploring uncertainty measures in bayesian deep attentive neural networks for prostate zonal segmentation. IEEE Access 8, 151817–151828.

Lourenço-Silva, J., Oliveira, A.L., 2022. Using soft labels to model uncertainty in medical image segmentation. International MICCAI Brainlesion Workshop , 585–596.

Lu, C., Angelopoulos, A.N., Pomerantz, S., 2022. Improving trustworthiness of ai disease severity rating in medical imaging with ordinal conformal prediction sets , 545–554.

Ma, X., Niu, Y., Gu, L., Wang, Y., Zhao, Y., Bailey, J., Lu, F., 2021. Understanding adversarial attacks on deep learning based medical image analysis systems. Pattern Recognition 110, 107332.

Mahapatra, D., Poellinger, A., Shao, L., Reyes, M., 2021. Interpretability-driven sample selection using self supervised learning for disease classification and segmentation. IEEE Transactions on Medical Imaging 40, 2548–2562.

Malinin, A., 2019. Uncertainty estimation in deep learning with application to spoken language assessment. Ph.D. thesis. University of Cambridge.

Malinin, A., Athanasopoulos, A., Barakovic, M., Cuadra, M.B., Gales, M.J., Granziera, C., Graziani, M., Kartashev, N., Kyriakopoulos, K., Lu, P.J., et al., 2022. Shifts 2.0: Extending the dataset of real distributional shifts. arXiv preprint arXiv:2206.15407 .

McClure, P., Rho, N., Lee, J.A., Kaczmarzyk, J.R., Zheng, C.Y., Ghosh, S.S., Nielson, D.M., Thomas, A.G., Bandettini, P., Pereira, F., 2019. Knowing what you know in brain segmentation using bayesian deep neural networks. Frontiers in neuroinformatics 13, 67.

McKinley, R., Meier, R., Wiest, R., 2018. Ensembles of densely-connected cnns with label-uncertainty for brain tumor segmentation. International MICCAI Brainlesion Workshop , 456–465.

McKinley, R., Rebsamen, M., Daetwyler, K., Meier, R., Radojewski, P., Wiest, R., 2020a. Uncertainty-driven refinement of tumor-core segmentation using 3d-to-2d networks with label uncertainty. International MICCAI Brainlesion Workshop , 401–411.

McKinley, R., Rebsamen, M., Meier, R., Wiest, R., 2019. Triplanar ensemble of 3d-to-2d cnns with label-uncertainty for brain tumor segmentation. International MICCAI Brainlesion workshop , 379–387.

McKinley, R., Wepfer, R., Grunder, L., Aschwanden, F., Fischer, T., Friedli, C., Muri, R., Rummel, C., Verma, R., Weisstanner, C., et al., 2020b. Automatic detection of lesion load change in multiple sclerosis using convolutional neural networks with segmentation confidence. NeuroImage: Clinical 25, 102104.

Mehrtash, A., Kapur, T., Tempany, C.M., Abolmaesumi, P., Wells, W.M., 2021. Prostate cancer diagnosis with sparse biopsy data and in presence of location uncertainty. 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) , 443–447.

Mehrtash, A., Wells, W.M., Tempany, C.M., Abolmaesumi, P., Kapur, T., 2020. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE Transactions on Medical Imaging 39, 3868–3878.

Mehta, R., Christinck, T., Nair, T., Lemaitre, P., Arnold, D., Arbel, T., 2019. Propagating uncertainty across cascaded medical imaging tasks for improved deep learning inference. Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Clinical Image-Based Procedures , 23–32.

Mehta, R., Filos, A., Baid, U., Sako, C., McKinley, R., Rebsamen, M., Dätwyler, K., Meier, R., Radojewski, P., Murugesan, G.K., et al., 2022. Qu-brats: Miccai brats 2020 challenge on quantifyinguncertainty in brain tumor segmentation-analysis of ranking scores and benchmarking results. Journal of Machine Learning for Biomedical Imaging 1.

Mehta, R., Filos, A., Gal, Y., Arbel, T., 2020. Uncertainty evaluation metric for brain tumour segmentation. arXiv e-prints .

Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al., 2014. The multimodal brain tumor image segmentation benchmark (brats). IEEE Transactions on Medical Imaging 34, 1993–2024.

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

Mishra, S., Chen, D.Z., Hu, X.S., 2021. Objective-dependent uncertainty driven retinal vessel segmentation. 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) , 453–457.

Mobiny, A., Singh, A., Van Nguyen, H., 2019. Risk-aware machine learning classifier for skin lesion diagnosis. Journal of clinical medicine 8, 1241.

Mojabi, P., Khoshdel, V., Lovetri, J., 2020. Tissue-type classification with uncertainty quantification of microwave and ultrasound breast imaging: A deep learning approach. IEEE Access 8, 182092–182104.

Mojiri Forooshani, P., Biparva, M., Ntiri, E.E., Ramirez, J., Boone, L., Holmes, M.F., Adamo, S., Gao, F., Ozzoude, M., Scott, C.J., et al., 2022. Deep Bayesian networks for uncertainty estimation and adversarial resistance of white matter hyperintensity segmentation. Technical Report. Wiley Online Library.

Molle, P.V., Verbelen, T., Boom, C.D., Vankeirsbilck, B., Vylder, J.D., Diricx, B., Kimpe, T., Simoens, P., Dhoedt, B., 2019. Quantifying uncertainty of deep neural networks in skin lesion classification. Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Clinical Image-Based Procedures , 52–61.

Monteiro, M., Le Folgoc, L., Coelho de Castro, D., Pawlowski, N., Marques, B., Kamnitsas, K., van der Wilk, M., Glocker, B., 2020. Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. Advances in Neural Information Processing Systems 33, 12756–12767.

Nair, T., Precup, D., Arnold, D.L., Arbel, T., 2020. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. Medical image analysis 59, 101557.

Natekar, P., Kori, A., Krishnamurthi, G., 2020. Demystifying brain tumor segmentation networks: interpretability and uncertainty analysis. Frontiers in computational neuroscience 14, 6.

Nguyen, A., Yosinski, J., Clune, J., 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition , 427–436.

Nixon, J., Dusenberry, M.W., Zhang, L., Jerfel, G., Tran, D., 2019. Measuring calibration in deep learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops , 38–41.

Norouzi, A., Emami, A., Najarian, K., Karimi, N., Soroushmehr, S.R., et al., 2019. Exploiting uncertainty of deep neural networks for improving segmentation accuracy in mri images. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) , 2322–2326.

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al., 2018. Attention u-net: Learning where to look for the pancreas. Medical Imaging with Deep Learning .

Orlando, J.I., Seeböck, P., Bogunović, H., Klimscha, S., Grechenig, C., Waldstein, S., Gerendas, B.S., Schmidt-Erfurth, U., 2019. U2-net: A bayesian u-net model with epistemic uncertainty feedback for photoreceptor layer segmentation in pathological oct scans. 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019) , 1441–1445.

Osawa, K., Swaroop, S., Khan, M.E., Jain, A., Eschenhagen, R., Turner, R.E., Yokota, R., 2019. Practical deep learning with bayesian principles. Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019 , 4289–4301.

Ouyang, C., Chen, C., Li, S., Li, Z., Qin, C., Bai, W., Rueckert, D., 2021. Causality-inspired single-source domain generalization for medical image segmentation. arXiv e-prints .

Ozdemir, O., Russell, R.L., Berlin, A.A., 2019. A 3d probabilistic deep learning system for detection and diagnosis of lung cancer using low-dose ct scans. IEEE Transactions on Medical Imaging 39, 1419–1429.

Ozdemir, O., Woodward, B., Berlin, A.A., 2017. Propagating uncertainty in multi-stage bayesian convolutional neural networks with application to pulmonary nodule detection. arXiv e-prints .

Pal, J.B., 2022. Holistic network for quantifying uncertainties in medical images. International MICCAI Brainlesion Workshop , 560–569.

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

Pocevičiūtė, M., Eilertsen, G., Jarkman, S., Lundström, C., 2022. Generalisation effects of predictive uncertainty estimation in deep learning for digital pathology. Scientific Reports 12, 1–15.

Postels, J., Segu, M., Sun, T., Van Gool, L., Yu, F., Tombari, F., 2021. On the practicality of deterministic epistemic uncertainty. International Conference on Machine Learning .

Puttagunta, M., Ravi, S., 2021. Medical image analysis based on deep learning approach. Multimedia Tools and Applications 80, 24365–24398.

Rączkowski, Ł., Możejko, M., Zambonelli, J., Szczurek, E., 2019. Ara: accurate, reliable and active histopathological image classification framework with bayesian deep learning. Scientific reports 9, 1–12.

Rajaraman, S., Zamzmi, G., Yang, F., Xue, Z., Jaeger, S., Antani, S.K., 2022. Uncertainty quantification in segmenting tuberculosis-consistent findings in frontal chest x-rays. Biomedicines 10, 1323.

Redekop, E., Chernyavskiy, A., 2021. Uncertainty-based method for improving poorly labeled segmentation datasets. 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) , 1831–1835.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention , 234–241.

Rosas-Gonzalez, S., Birgui-Sekou, T., Hidane, M., Zemmoura, I., Tauber, C., 2021. Asymmetric ensemble of asymmetric u-net models for brain tumor segmentation with uncertainty estimation. Frontiers in Neurology , 1421.

Rousseau, A.J., Becker, T., Bertels, J., Blaschko, M.B., Valkenborg, D., 2021. Post training uncertainty calibration of deep networks for medical image segmentation. 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) , 1052–1056.

Roy, A.G., Conjeti, S., Navab, N., Wachinger, C., Initiative, A.D.N., et al., 2019. Bayesian quicknat: Model uncertainty in deep whole-brain segmentation for structure-wise quality control. NeuroImage 195, 11–22.

Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 1, 206–215.

Salehi, S.S.M., Erdogmus, D., Gholipour, A., 2017. Tversky loss function for image segmentation using 3d fully convolutional deep networks. International Workshop on Machine Learning in Medical Imaging , 379–387.

Sander, J., de Vos, B.D., Wolterink, J.M., Isgum, I., 2019. Towards increased trustworthiness of deep learning segmentation methods on cardiac MRI. Medical Imaging 2019: Image Processing 10949, 1094919.

Sedai, S., Antony, B., Mahapatra, D., Garnavi, R., 2018. Joint segmentation and uncertainty visualization of retinal layers in optical coherence tomography images using bayesian deep learning. Computational Pathology and Ophthalmic Medical Image Analysis , 219–227.

Sedai, S., Antony, B., Rai, R., Jones, K., Ishikawa, H., Schuman, J., Gadi, W., Garnavi, R., 2019. Uncertainty guided semi-supervised segmentation of retinal layers in oct images. International Conference on Medical Image Computing and Computer-Assisted Intervention , 282–290.

Selvan, R., Faye, F., Middleton, J., Pai, A., 2020. Uncertainty quantification in medical image segmentation with normalizing flows. International Workshop on Machine Learning in Medical Imaging , 80–90.

Senousy, Z., Abdelsamea, M.M., Gaber, M.M., Abdar, M., Acharya, U.R., Khosravi, A., Nahavandi, S., 2021. Mcua: Multi-level context and uncertainty aware dynamic deep ensemble for breast cancer histology image classification. IEEE Transactions on Biomedical Engineering 69, 818–829.

Sensoy, M., Kaplan, L.M., Kandemir, M., 2018. Evidential deep learning to quantify classification uncertainty. Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018 , 3183–3193.

Shamsi, A., Asgharnezhad, H., Jokandan, S.S., Khosravi, A., Kebria, P.M., Nahavandi, D., Nahavandi, S., Srinivasan, D., 2021. An uncertainty-aware transfer learning-based framework for covid-19 diagnosis. IEEE Transactions on neural networks and learning systems 32, 1408–1417.

Shaw, R., Sudre, C.H., Ourselin, S., Cardoso, M.J., Pemberton, H.G., 2021. A decoupled uncertainty model for mri segmentation quality estimation. arXiv e-prints .

Snoek, J., Ovadia, Y., Fertig, E., Lakshminarayanan, B., Nowozin, S., Sculley, D., Dillon, J.V., Ren, J., Nado, Z., 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019 , 13969–13980.

Soberanis-Mukul, R.D., Navab, N., Albarqouni, S., 2020. Uncertainty-based graph convolutional networks for organ segmentation refinement. Medical Imaging with Deep Learning , 755–769.

Song, B., Sunny, S., Li, S., Gurushanth, K., Mendonca, P., Mukhia, N., Patrick, S., Gurudath, S., Raghavan, S., Tsusennaro, I., et al., 2021. Bayesian deep learning for reliable oral cancer image classification. Biomedical Optics Express 12, 6422–6430.

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. International Conference on Machine Learning , 6105–6114.

Tardy, M., Scheffer, B., Mateus, D., 2019. Uncertainty measurements for the reliable classification of mammograms. International Conference on Medical Image Computing and Computer-Assisted Intervention , 495–503.

Thagaard, J., Hauberg, S., Vegt, B.v.d., Ebstrup, T., Hansen, J.D., Dahl, A.B., 2020. Can you trust predictive uncertainty under real dataset shifts in digital pathology? International Conference on Medical Image Computing and Computer-Assisted Intervention , 824–833.

Toledo-Cortés, S., De La Pava, M., Perdómo, O., González, F.A., 2020. Hybrid deep learning gaussian process for diabetic retinopathy diagnosis and uncertainty quantification. International Workshop on Ophthalmic Medical Image Analysis , 206–215.

Tonekaboni, S., Joshi, S., McCradden, M.D., Goldenberg, A., 2019. What clinicians want: contextualizing explainable machine learning for clinical end use. Machine learning for healthcare conference , 359–380.

Tousignant, A., Lemaître, P., Precup, D., Arnold, D.L., Arbel, T., 2019. Prediction of disease progression in multiple sclerosis patients using deep learning analysis of mri data. International conference on medical imaging with deep learning , 483–492.

Ulmer, D., Cinà, G., 2021. Know your limits: Uncertainty estimation with relu classifiers fails at reliable ood detection. Uncertainty in Artificial Intelligence , 1766–1776.

van der Velden, B.H., Kuijf, H.J., Gilhuijs, K.G., Viergever, M.A., 2022. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. Medical Image Analysis , 102470.

Vu, M.H., Nyholm, T., Löfstedt, T., 2020. Multi-decoder networks with multi-denoising inputs for tumor segmentation. International MICCAI Brainlesion Workshop , 412–423.

Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., 2019a. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing 338, 34–45.

Wang, G., Li, W., Ourselin, S., Vercauteren, T., 2019b. Automatic brain tumor segmentation based on cascaded convolutional neural networks with uncertainty estimation. Frontiers in computational neuroscience 13, 56.

Wang, G., Li, W., Zuluaga, M.A., Pratt, R., Patel, P.A., Aertsen, M., Doel, T., David, A.L., Deprest, J., Ourselin, S., et al., 2018. Interactive medical image segmentation using deep learning with image-specific fine tuning. IEEE Transactions on Medical Imaging 37, 1562–1573.

Wang, H., Yeung, D.Y., 2020. A survey on bayesian deep learning. ACM Computing Surveys (CSUR) 53, 1–37.

Wang, X., Tang, F., Chen, H., Luo, L., Tang, Z., Ran, A.R., Cheung, C.Y., Heng, P.A., 2020. Ud-mil: uncertainty-driven deep multiple instance learning for oct image classification. IEEE journal of biomedical and health informatics 24, 3431–3442.

Wenzel, F., Roth, K., Veeling, B.S., Swiatkowski, J., Tran, L., Mandt, S., Snoek, J., Salimans, T., Jenatton, R., Nowozin, S., 2020. How good is the bayes posterior in deep neural networks really? International Conference on Machine Learning 119, 10248–10259.

Wickstrøm, K., Kampffmeyer, M., Jenssen, R., 2020. Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. Medical image analysis 60, 101619.

Xia, Y., Yang, D., Yu, Z., Liu, F., Cai, J., Yu, L., Zhu, Z., Xu, D., Yuille, A., Roth, H., 2020. Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. Medical Image Analysis 65, 101766.

Xiang, J., Qiu, P., Yang, Y., 2022. Fussnet: Fusing two sources of uncertainty for semi-supervised medical image segmentation , 481–491.

Yang, J., Liang, Y., Zhang, Y., Song, W., Wang, K., He, L., 2021. Exploring instance-level uncertainty for medical detection. 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) , 448–452.

Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z., 2017. Suggestive annotation: A deep active learning framework for biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention , 399–407.

Yang, S., Fevens, T., 2021. Uncertainty quantification and estimation in medical image classification. International Conference on Artificial Neural Networks , 671–683.

Yang, Y., Guo, X., Pan, Y., Shi, P., Lv, H., Ma, T., 2022. Uncertainty quantification in medical image segmentation with multi-decoder u-net. International MICCAI Brainlesion Workshop , 570–577.

Yu, L., Wang, S., Li, X., Fu, C.W., Heng, P.A., 2019. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention , 605–613.

Zhang, G., Dang, H., Xu, Y., 2022. Epistemic and aleatoric uncertainties reduction with rotation variation for medical image segmentation with convnets. SN Applied Sciences 4, 1–11.

Zhang, L., Wang, X., Yang, D., Sanford, T., Harmon, S., Turkbey, B., Wood, B.J., Roth, H., Myronenko, A., Xu, D., et al., 2020. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. IEEE Transactions on Medical Imaging 39, 2531–2540.

Zhao, Y., Yang, C., Schweidtmann, A., Tao, Q., 2022. Efficient bayesian uncertainty estimation for nnu-net , 535–544.

Zhou, X., Liu, H., Pourpanah, F., Zeng, T., Wang, X., 2022. A survey on epistemic (model) uncertainty in supervised learning: Recent advances and applications. Neurocomputing 489, 449–465.

Zou, K., Yuan, X., Shen, X., Wang, M., Fu, H., 2022. Tbrats: Trusted brain tumor segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention 13438, 503–513.