A review of uncertainty quantification in medical image analysis: probabilistic and non-probabilistic methods

Ling Huang, Su Ruan, Yucheng Xing, Mengling Feng

A review of uncertainty quantification in medical image analysis: probabilistic and non-probabilistic methods

Ling Huang[a], Su Ruan[b], Yucheng Xing[a], Mengling Feng[a,c]

[a]*Saw Swee Hock School of Public Health, National University of Singapore, Singapore*
[b]*Quantif, LITIS, University of Rouen Normandy, France*
[c]*Institute of Data Science, National University of Singapore, Singapore*

Abstract

The comprehensive integration of machine learning healthcare models within clinical practice remains suboptimal, notwithstanding the proliferation of high-performing solutions reported in the literature. A predominant factor hindering widespread adoption pertains to an insufficiency of evidence affirming the reliability of the aforementioned models. Recently, uncertainty quantification methods have been proposed as a potential solution to quantify the reliability of machine learning models and thus increase the interpretability and acceptability of the result. In this review, we offer a comprehensive overview of prevailing methods proposed to quantify uncertainty inherent in machine learning models developed for various medical image tasks. Contrary to earlier reviews that exclusively focused on probabilistic methods, this review also explores non-probabilistic approaches, thereby furnishing a more holistic survey of research pertaining to uncertainty quantification for machine learning models. Analysis of medical images with the summary and discussion on medical applications and the corresponding uncertainty evaluation protocols are presented, which focus on the specific challenges of uncertainty in medical image analysis. We also highlight some potential future research work at the end. Generally, this review aims to allow researchers from both clinical and technical backgrounds to gain a quick and yet in-depth understanding of the research in uncertainty quantification for medical image analysis machine learning models.

Keywords: Uncertainty quantification, Probabilistic methods, Non-probabilistic methods, Epistemic uncertainty, Aleatory uncertainty, Uncertainty evaluation, Medical image analysis

1. Introduction

With the augmented investment of financial and human resources into artificial intelligence (AI), society has experienced notable transformations. Healthcare is definitely one of the areas where we see great potential for AI to introduce revolutionizing improvements. In particular, for medical image analysis (MIA), many deep neural network-based machine learning models with powerful learning and feature representation abilities have been developed. Despite the excellent performance of recent MIA methods, doubts about the reliability of their results still remain (Thagaard et al., 2020; Hüllermeier and Waegeman, 2021; Czolbe

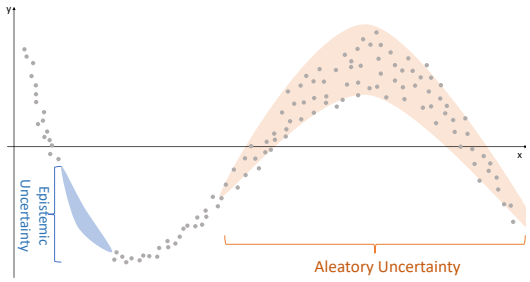


Figure 1: An example explanation of aleatory uncertainty inherently random and epistemic uncertainty caused by a lack of knowledge about the best analysis model (reproduced based on (Yang and Li, 2023))

et al., 2021), which explains why their application to therapeutic decision-making for complex oncological cases is still limited. Learning, in the sense of generalizing beyond observed data so far, relies inherently on induction, i.e., replacing specific observations with general models of the data-generating process. Such models, however, are inherently speculative and lack definitive correctness; they remain hypothetical and, consequently, uncertain. The uncertainty extends to the predictions generated by these models as well. In addition to the inductive inference uncertainty, other sources of uncertainty, such as incorrect model assumptions and noisy or imprecise data, exist, too.

In general, there are two sources of uncertainty: aleatory and epistemic uncertainty (Hora, 1996; Der Kiureghian and Ditlevsen, 2009). Aleatory uncertainty refers to the notion of randomness, i.e., the variability in an experimental outcome due to inherently random effects, which can not be reduced. In contrast, epistemic uncertainty refers to uncertainty caused by a lack of knowledge (ignorance) about the best analysis model, i.e., the ignorance of the learning algorithm or decision-maker. As opposed to uncertainty caused by randomness, uncertainty caused by ignorance can be reduced based on additional information or the design of a suitable learning algorithm. Figure 1 provides an example explanation of aleatory and epistemic uncertainty.

To fully harness the potential benefits of ML in MIA systems, a trustworthy representation of uncertainty is desirable and should be considered a key feature of developing state-of-the-art (SOTA) MIA methods. Traditionally, in fields like statistics and machine learning, probabilistic methods that rely on probability theory to represent, propagate, and analyze uncertainty have been perceived as the ultimate tool for uncertainty handling. They are commonly used with Bayesian inference to model uncertainty in various parameters or variables (Hinton and Van Camp, 1993; MacKay, 1992). Moreover, the recent popularity of deep models has revived research on model uncertainty and has given rise to specific methods such as Monte Carlo dropout (Gal and Ghahramani, 2016; Tran et al., 2019) and model ensembles (Lakshminarayanan et al., 2017; Rupprecht et al., 2017). However, probabilistic models make strong assumptions about the real distribution, which can potentially bring erroneous uncertainty estimation and fail to predict uncertainty correctly when the actual

distribution is different. Moreover, probabilistic models are essentially based on capturing knowledge in terms of probability distribution and fail to distinguish between aleatory and epistemic uncertainty, limiting the exploitation of the results.

Non-probabilistic methods, which handle uncertainty without relying on explicit probabilistic models, are usually used to characterize and analyze uncertainty when probabilistic information, such as precise probabilities or distributions, is unavailable or difficult to determine. Instead of building strong assumptions about the real distribution, these methods use alternative mathematical frameworks or representations such as intervals (Rao and Berke, 1997), fuzzy sets (Zadeh, 1965), Credal partition (Denœux and Masson, 2004), or distance-based evidence reasoning mechanisms (Denoeux, 1995) to quantify uncertainty.

The recent study on uncertainty quantification significantly improved the performance of ML models and increased researchers' interest in analyzing those studies systematically. In 2016, Guney Gusel examined and explained fuzzy logic-based uncertainty methods in healthcare decision-making (Gürsel, 2016). In 2018, Kabir et al. reviewed neural network-based uncertainty quantification methods with a particular focus on probabilistic forecasting and prediction intervals (Kabir et al., 2018). In 2021, there are booming analyses about uncertainty. Alizadehsani et al. reviewed the research handling uncertainty in medical data using machine learning and probability theory techniques in the last 30 years (Alizadehsani et al., 2021). Hüllermeier and Waegeman provided a comprehensive introduction to concepts and methods about aleatory and epistemic uncertainty in ML (Hüllermeier and Waegeman, 2021). Abdar et al. reviewed uncertainty quantification in deep learning with discussions on techniques, applications and potential challenges with a particular focus on Bayesian statistics and ensemble learning (Abdar et al., 2021b). Gillmann et al. studied uncertainty-aware visualization methods, showing readers which approaches can be combined to form uncertainty-aware medical imaging pipelines (Gillmann et al., 2021). However, the above-mentioned review work can not provide a global overview of uncertainty quantification methods in MIA with recent ML methods, limiting the development of uncertainty analysis studies.

Contributions. Unlike previous uncertainty review papers that provide a general picture of uncertainty quantification in ML applications (Abdar et al., 2021b; Hüllermeier and Waegeman, 2021), or focus on discussing several specific uncertainty quantification methods (Alizadehsani et al., 2021), this study reviews both probabilistic and non-probabilistic uncertainty methods in MIA under the ML framework in the last ten years, in which the later one is still ignored. It is worth mentioning that the primary purpose of this study is not to introduce the performance of various existing uncertainty quantification methods. Instead, we focus on outlining the most common uncertainty quantification and evaluation methods, the important application areas, as well as potential research work. We hope this review paper can provide guidance to researchers in the fields of machine learning and clinical practice and pave the way for future research in order to generate reliable and explainable decisions based on quantified uncertainty or improve the fairness of the overall healthcare systems by combining multiple source information with uncertainty. The main contributions of this study are as follows:

- To our best knowledge, this is the first comprehensive review paper regarding the study of both probabilistic and non-probabilistic uncertainty quantification methods in MIA tasks.
- Existing uncertainty evaluation criteria applied for MIA are studied and discussed.
- The main categories of important clinical applications of uncertainty quantification methods are presented and discussed.
- The major advantages and limitations of existing uncertainty quantification research are pointed out, as well as the potential future work.

Organization. The rest of this paper is organized as follows: Section 2 explained the search criteria. Section 3 presents the commonly used probabilistic and non-probabilistic uncertainty quantification methods in MIA. Section 4 introduces the uncertainty evaluation criteria. Section 5 summarizes MIA applications with the mentioned uncertainty quantification methods. Finally, Section 6 provides a discussion of the advantages and limitations of the literature, and Section 7 gives the overall conclusion of this review.

2. Search criteria

To perform this review, we performed a search on Web of Science for the papers published between 1 January 2013 and 15 July 2023. The search keywords used for this study are 'Uncertainty quantification' OR 'fuzzy systems' OR 'Monte Carlo simulation' OR 'rough classification' OR 'Dempster–Shafer theory' OR 'Imprecise probability' AND 'Medical image analysis.' We note that, from 2013 to 2023, more than 5,000 papers studying uncertainty in MIA tasks were published. To ensure the criticality of the study, we only include papers published in related journals and conferences and screen their title and abstracts. Then, about 700 papers with full access and good citation records were selected, and those lacking adequate connection with the topic of our review were removed from the list. Then, we read the full-text paper with the inclusion criteria illustrated in Figure 2. In the end, 301 papers are investigated in this review.

Figure 3 shows the number of papers focused on uncertainty analysis in the last ten years, where we can see that handling uncertainty in machine learning has received increasing attention, especially when machine learning methods have been able to achieve promising accuracy performance with the popularization of deep learning after 2015. Researchers' interest in the study of uncertainty in MIA models remained at a steady state until 2018, i.e., around 400 published papers each year. There are two main reasons: 1) uncertainties in medical image reconstruction or registration tasks are easily observable and awarded; 2) the study of ML models for MIA is lagging behind ML research itself. Once the ML research for MIA has reached an accuracy saturation situation, people then turn to study uncertainty. Therefore, after 2018, increasing efforts have been involved in studying the MIA uncertainty.

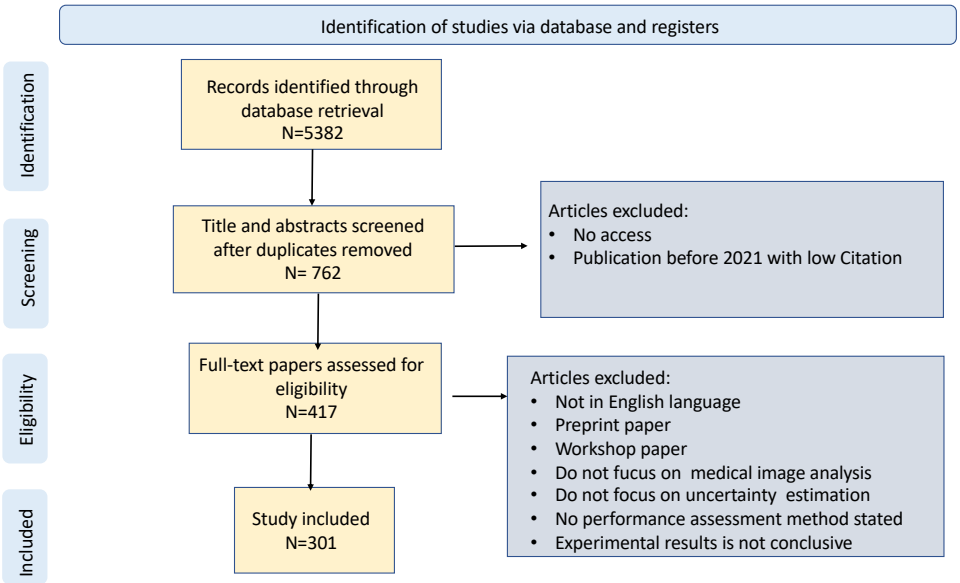
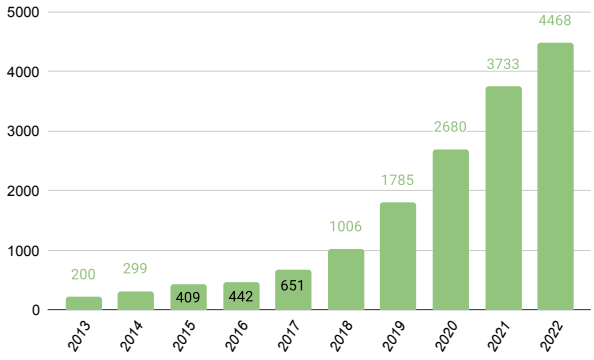
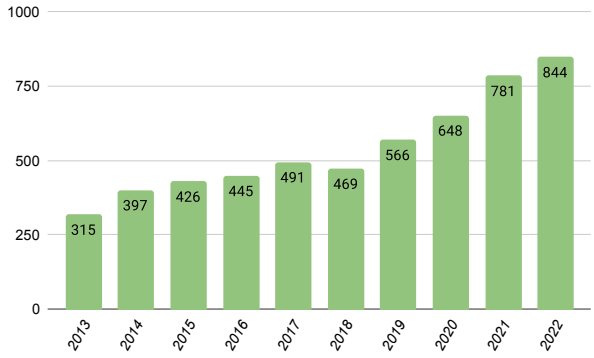


Figure 2: Illustration of selecting eligible publications for inclusion.



(a) Number of papers focused on uncertainty in ML



(b) Number of papers focused on uncertainty in MIA

Figure 3: Number of papers focused on uncertainty in the last 10 years.

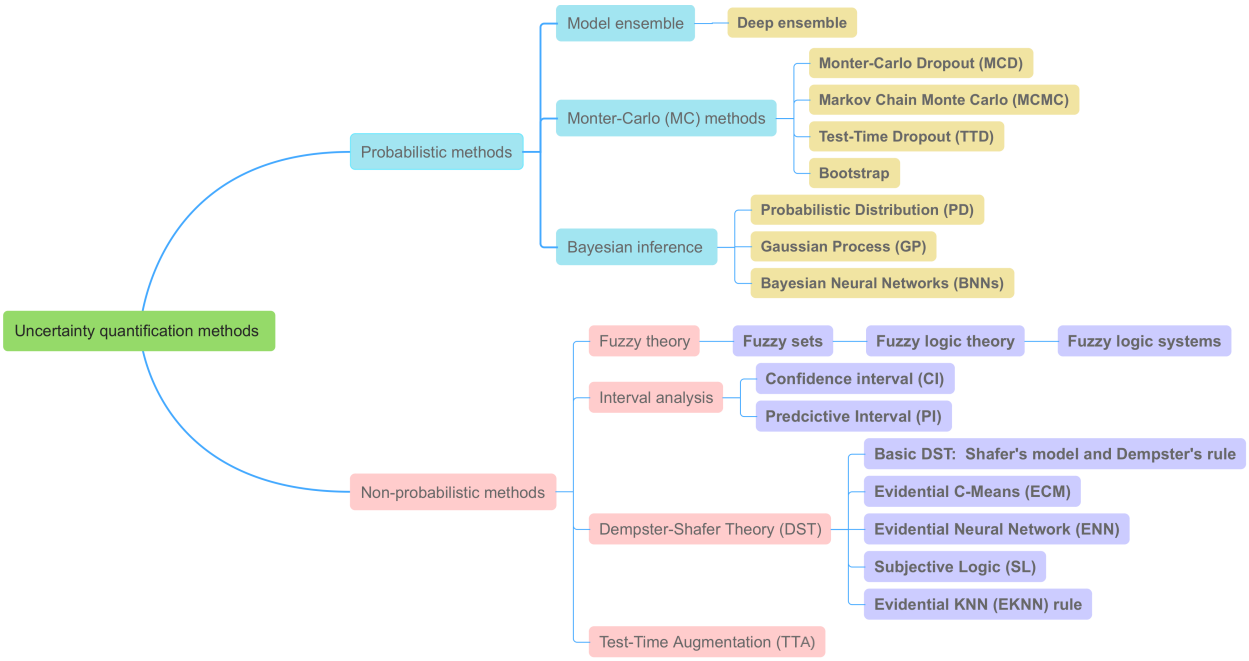


Figure 4: Overview of uncertainty quantification methods: Probabilistic and Non-probabilistic methods

3. Methods to uncertainty quantification

Different from existing literature reviews that focus on analyzing uncertainty methods in a specific field or with a specific methodology, in this paper, we provide a comprehensive overview of uncertainty analysis methods in medical images, including analysis from both probabilistic and non-probabilistic sides of the application to different medical image tasks. Figure 4 shows an overview of uncertainty quantification methods. It should be noted that both probabilistic and non-probabilistic methods represent, propagate, and reason uncertainty in a systematic manner and the choice of method depends on the nature of uncertainty, available information, and the specific problem domain.

3.1. Probabilistic uncertainty quantification methods

Probabilistic uncertainty quantification methods leverage probability theory to represent uncertainty using probability distributions, allowing for calculating probabilities, quantiles, and other statistical measures. To capture uncertainty in ML model predictions, predictive entropy or variance are typically used to estimate distributions over the outputs. Predictive entropy (Shannon, 1948) measures the diversity or spread of a single probability distribution and quantifies how uncertain or ambiguous a model's prediction is by considering the distribution's entropy, which reflects the amount of information or randomness in the distribution, i.e., combining both aleatory and epistemic uncertainties into a single measure that reflects the overall uncertainty in a probabilistic prediction. Higher predictive entropy indicates greater uncertainty and ambiguity in the model's prediction. Predictive variance (Fisher, 1919) is a measure of uncertainty related to the inherent randomness or noise in the data,

i.e., aleatory uncertainty. Higher predictive variance suggests that individual predictions are more scattered around the mean prediction, indicating that the model's predictions are more sensitive to changes in the input. According to our literature review, there are three main probabilistic methods: *Bayesian Inference, Monte-Carlo method, and Model ensemble*, which can generate predictive entropy or variance.

Bayesian inference is a statistical approach that inherently involves probabilistic modeling and updates prior beliefs or knowledge to estimate uncertain quantities using observed data. It combines prior distributions with likelihood functions to obtain posterior distributions that reflect updated beliefs. Among Bayesian inference methods, Probabilistic Distribution (PD) (Wallman et al., 2014), Gaussian Process (GP) (Wachinger et al., 2014) and Bayesian Neural Networks (BNNs) (Blundell et al., 2015) are three main popular methods used for uncertainty quantification. Readers can refer to Supplementary Material A for detailed analysis.

Monte Carlo (MC) methods (Kroese et al., 2014) involves generating random samples from probability distributions. As the number of samples increases, the simulated outcomes converge toward the true distribution of possible outcomes, allowing us to obtain accurate estimates of uncertainty. Aleatory uncertainty is then captured through the variability introduced by dropout during forward passes, reflecting data randomness. Epistemic uncertainty is captured through the diversity of predictions across passes, indicating the model's uncertainty about its own parameters and structure. Among the MC methods, Monte Carlo sampling, Monte Carlo dropout (MCD) (Gal and Ghahramani, 2016), Markov Chain Monte Carlo (MCMC) (Gilks et al., 1995; Brooks, 1998), and Bootstrap are the most common algorithms. Since Test-Time Dropout (TTD) (Srivastava et al., 2014) also involves repeated sampling from the data, here we classify it into the MC methods as well (details can be found in Supplementary Material A). It should be noted that while the MC methods are not inherently a Bayesian inference method, it is often employed in Bayesian inference to estimate posterior distributions and perform various Bayesian analyses.

Model ensemble (Dietterich, 2000; Zhou, 2012) typically focuses on capturing different sources of variability or uncertainty in model assumptions rather than explicitly quantifying uncertainty using probabilistic measures. Each model in the ensemble framework may be trained using different initializations, subsets of the training data, or variations in the model architecture. Standard ensemble methods, such as bagging, boosting, or random forests, generate an ensemble of models that collectively represent uncertainty with the variability among the predictions. Recently, deep ensemble (Lakshminarayanan et al., 2017; Ganaie et al., 2022) models have become a popular uncertainty quantification method integrated with deep neural networks. Details about deep ensemble models can be found in Supplementary Material A.

To sum up, probabilistic uncertainty quantification methods provide a rigorous and quantitative approach to characterizing and analyzing uncertainty.

3.2. Non-probabilistic uncertainty quantification methods

Non-probabilistic uncertainty methods, free from the strong assumption of the prior distribution of the data, are more flexible and applicable for most applications, especially when

precise probabilistic information is not available. Methods such as *interval analysis* (Rao and Berke, 1997), *fuzzy sets and fuzzy logic theory* (Yager and Zadeh, 2012), *Dempster-Shafer theory* (Dempster, 1967; Shafer, 1976), and *Test-Time Augmentation* do not directly involve probability distributions to represent uncertainty. Instead, they introduce conceptions such as predictive intervals (Eaton-Rosen et al., 2018), fuzzy membership functions (Wang et al., 2023a) or plausible sets (Adiga Vasudeva et al., 2022), Credal partition (Denœux and Masson, 2004), evidence-based reasoning mechanisms (Huang et al., 2021a) or test-time data augmentation to model uncertainty.

Dempster-Shafer theory (DST) (Dempster, 1967; Shafer, 1976), also known as Belief function theory or Evidence theory, was first originated by Dempster (Dempster, 1967) in the context of statistic inference in 1968 and was formalized by Shafer (Shafer, 1976) as a theory of evidence in 1976. It is a theoretical framework for modeling, reasoning with, and combining imperfect (imprecise, uncertain, and partial) information. With DST, we can quantify uncertainty in a single forward pass and further explore the possibility of improving the model reliability based on the quantified uncertainty. Based on DST, there are some commonly used uncertainty quantification methods, i.e., Evidential KNN (EKNN) rule Denoeux (1995), Evidential C-Means (ECM) (Masson and Denoeux, 2008), Evidential Neural Network (ENN) (Denœux, 2000), and Subjective Logic (SL) (Josang et al., 2006; Jøsang, 2016), readers can refer to Supplementary Material or paper (Huang et al., 2023a) for more information.

Fuzzy sets (Zadeh, 1965) define the linguistic terms and their membership functions; fuzzy rules capture the relationships between inputs and outputs using if-then statements; and the fuzzy inference mechanism combines the rules and performs fuzzy reasoning to compute the system's output (Dubois, 1980). Fuzzy logic (Hájek, 2013) employs fuzzy sets to capture the degree of membership of elements to a particular linguistic variable such as "high likelihood," "medium uncertainty," or "low confidence." *Fuzzy logic systems* (Mendel, 1995; Yager and Zadeh, 2012) are the implementations of fuzzy logic principles to solve specific problems by utilizing fuzzy sets, fuzzy rules, and fuzzy inference mechanisms.

Interval analysis (Rao and Berke, 1997) represents uncertainty by bounding the possible range of values for variables or parameters using intervals, thus offering a systematic and robust approach to uncertainty quantification, especially in cases where rigorous bounds on uncertainty are essential for decision-making or risk assessment. In interval analysis, uncertainty is characterized by assigning intervals to model parameters, inputs, or outputs rather than specifying precise probabilities. These intervals represent ranges of possible values rather than probabilities of occurrence. Thus, it can be defined based on available information, expert opinion, or experimental data. Confidence intervals (CI) (Hosmer and Lemeshow, 1992; Smithson, 2003) and prediction intervals (PI) (Hwang and Ding, 1997) are two common interval algorithms used to quantify the uncertainty associated with a given estimate. However, it can also lead to wide intervals if the input uncertainties are too large or if the model's behavior is nonlinear and complex.

Test-Time Augmentation (TTA) (Ayhan and Berens, 2018; Wang et al., 2019a) is a technique used in machine learning to improve model performance and enhance the robustness of predictions. At test time, multiple variants of the input image are generated using data

augmentation such as spatial transformations (e.g., flipping, rotation), intensity augmentations (e.g., contrast modification, noise injection, or artifacts), etc. 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 median or variance.

It should also be noted that there are some researches that hybrid more than one uncertainty quantification method, e.g., integrating MCD in deep ensemble models or integrating fuzzy set with DST, for uncertainty analysis and yield more promising performance. The detailed applications of the above-mentioned methods will be introduced in Section 5.

4. Methods to uncertainty evaluation

The previous section presented the main uncertainty estimation approaches applied to MIA tasks. In this section, we introduce the protocols implemented in these papers to evaluate the performance of the uncertainty estimation approaches.

Direct uncertainty evaluation methods such as mean square error validate the correctness of uncertainty quantification techniques with given uncertainty ground truths. While in real-world medical scenarios, ground-truth uncertainty is unavailable or difficult to obtain.

Indirect uncertainty evaluation methods, e.g., calibration metrics, coverage metrics, scoring rules, and prediction entropy, on the other hand, focus on a qualitative assessment of the computed uncertainty estimates by evaluating how well their predicted uncertainties correlate with the actual outcomes or data variability when uncertainty ground truth is unavailable. Misclassification or Out-of-Distribution detection are downstream applications of uncertainty in an automated pipeline of prediction, thus also used quite often in assessing the quality of the uncertainty quantification model. According to the literature review, we grouped five common uncertainty evaluation protocols (see Table 1).

4.1. Coverage metrics

Coverage metrics measure the proportion of cases where predicted uncertainty intervals (e.g., confidence intervals) contain the true value or the average width of prediction intervals. It can be estimated by sample variance or coverage probability. Sample variance computes the output variance across all samples collected using Bayesian inference, MC methods, or model ensembles with the definition:

$$variance = \sqrt{\frac{\sum_{n=1}^{N}(y_n - \bar{y})^2}{N-1}}, \tag{1}$$

where y_n is the value of the observation corresponding to pixel/voxel n, $\bar{y}$ is the mean value of all observations, and N is the number of observations. Coverage probability measures the proportion of true outcomes within the predicted uncertainty intervals (Dodge et al., 2003). The construction of the confidence interval ensures that the probability of finding the true vector θ in the sample dependent interval $[T_u, T_v]$ is (at least) γ:

$$P(T_u < \theta < T_v) = \gamma \tag{2}$$

Table 1: Evaluation criteria

Evaluation criteria	Papers
Coverage metrics	Judge et al. (2022); Mehta et al. (2023); Nair et al. (2020); Valen et al. (2022 Qian et al. (2020); Mehta et al. (2021); Eaton-Rosen et al. (2018); Bian et al Yang et al. (2021); Wickstrøm et al. (2020); Wallman et al. (2014); Ebadi et Herzog et al. (2020); Le Folgoc et al. (2016); Gour and Jain (2022); Corrado Jafari et al. (2021); Balagopal et al. (2021); Awate et al. (2019); Risholm et a
Predictive Entropy	Hamedani et al. (2023); Jungo et al. (2018b); Mehta et al. (2023, 2021); War Del Amor et al. (2023); Gour and Jain (2022); Ghoshal and Tucker (2020); N Ebadi et al. (2023); Kushibar et al. (2022); Camarasa et al. (2021); Rajaram Herzog et al. (2020); Dai and Tian (2013); Arega et al. (2023)
Calibration metrics	Hamedani et al. (2023); Jungo and Reyes (2019); Sambyal et al. (2022); Lave Judge et al. (2022); Carneiro et al. (2020); Liao et al. (2019); Ayhan et al. (2 Dawood et al. (2023); Thagaard et al. (2020); Pandey et al. (2022); Javadi et Ghoshal and Tucker (2022); Laves et al. (2021); Ghoshal and Tucker (2021); Dawood et al. (2023); Buddenkotte et al. (2023); Mehrtash et al. (2020); Rou Li et al. (2022a); Jena and Awate (2019); Arega et al. (2023)
Misclassification detection & OoD	Hamedani et al. (2023); Jungo and Reyes (2019); DeVries and Taylor (2018) Ghoshal et al. (2019); Iwamoto et al. (2021); Belharbi et al. (2021); Asgharn Linmans et al. (2023); Fuchs et al. (2021); Thagaard et al. (2020)
Scoring functions	Sambyal et al. (2022); Mehrtash et al. (2020); Arega et al. (2023); Thagaard Mehrtash et al. (2020); Tanno et al. (2017); Thagaard et al. (2020); Lemay e

4.2. Predictive entropy

The predictive entropy measures the informativeness of the model's predictive density function for each model output y_i with the definition

$$Entropy = -\sum_{i=1}^{C} p(i) \log p(i), \tag{3}$$

where $p(i)$ denotes the probability density function (PDF) (Parzen, 1962) or probability mass function (PMF) (Stewart, 2009) of the predicted variable i, and C is the set of possible values for the predicted variable.

4.3. Calibration metrics

Calibration metrics measure the agreement between the predicted uncertainty and the observed frequency of correct predictions. A well-calibrated uncertainty estimation method should provide uncertainty estimates that align with the true error level or uncertainty in the predictions. Calibration can be assessed using calibration plots, reliability diagrams, or calibration metrics such as Calibration Error (CE), Maximal Calibration Error (MCE), and Expected Calibration Error (ECE). Here, we introduce ECE as an example. It measures the correspondence between predicted probabilities and ground truth (Guo et al., 2017). The output normalized plausibility of the model is first binned into equally spaced bin E_h,

$h \in [1, H]$. The accuracy of bin E_h is defined as

$$\mathsf{acc}(E_h) = \frac{1}{\mid E_h \mid} \sum_{n \in E_h} \mathbf{1}(S_n = G_n), \tag{4}$$

where S_n and G_n are, respectively, the predicted and true class labels for pixel/voxel n. The average confidence of bin E_h is defined as

$$\mathsf{conf}(E_h) = \frac{1}{\mid E_h \mid} \sum_{n \in E_h} P_n, \tag{5}$$

where P_n is the predicted probability for pixel/voxel n. The ECE is the weighted average of the difference in accuracy and confidence of the bins:

$$ECE = \sum_{h=1}^{H} \frac{\mid E_h \mid}{N} \mid \mathsf{acc}(E_h) - \mathsf{conf}(E_h) \mid, \tag{6}$$

where N is the total number of pixels/voxels in all bins here, $\mid E_h \mid$ is the number of elements in bin E_h. A model is perfectly calibrated when $\mathsf{acc}(E_h) = \mathsf{conf}(E_h)$ for all $h \in \{1, ..., H\}$.

4.4. Scoring functions

Brier score (Brier, 1950) and Negative log-likelihood (NLL) are two commonly used scoring functions for evaluating the performance of uncertainty estimation methods. Brier score (Brier, 1950) measures the mean squared difference between predicted probabilities and actual outcomes with:

$$BS = \frac{1}{N} \sum_{n=1}^{N} (P_n - G_n)^2, \tag{7}$$

where G_n is the ground truth of pixel/voxel n and P_n is the predicted probability of pixel/voxel n, N is the number of pixels/voxels here. The lower the Brier score, the better the calibration and accuracy of the uncertainty estimates. NLL is usually used for evaluating probabilistic models and assessing their calibration and accuracy in capturing the uncertainty in predictions. It measures the average log probability assigned by a model to the observed outcomes by:

$$NLL = -\sum_{n=1}^{N} G_n log(P_n) + (1 - G_n) log(1 - P_n), \tag{8}$$

where G_n is the ground truth of pixel/voxel$n$ and P_n is the predicted probability of pixel/voxel n, N is the number of pixels/voxels here. A lower NLL value indicates better calibration and accuracy of the uncertainty estimates.

4.5. Misclassification & Out-of-Distribution 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 or Out-of-Distribution (OoD). This is crucial to prevent silent errors that could have a dramatic impact, especially in real-world medical image applications. In that sense, the uncertainty estimates can be turned into a binary classifier that aims to distinguish between correct and incorrect predictions. 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 (Stehman, 1997) can be constructed from the uncertainty point of view by distinguishing four possible cases with the following counts:

- True Positive (TP): The prediction is uncertain, and the expected label and the prediction differ,
- False Negative (FP): The prediction is certain, but the expected label and the prediction differ,
- True Negative (TN): The prediction is certain, and the expected label and the prediction are identical,
- False Negative (FN): The prediction is uncertain, but the prediction and the expected label are identical.

4.6. Discussion

It's important to note that different evaluation metrics and approaches may be suitable depending on the context and specific application. To evaluate the performance of uncertainty estimation methods, it is necessary to employ additional quantitative evaluation measures, such as calibration, coverage probability, mean squared error, discrimination metrics, or task-specific performance metrics. These metrics assess the uncertainty estimates' accuracy, calibration, and discriminative ability. Apart from the numerical uncertainty evaluation, a visual inspection of uncertainty (Sedai et al., 2018; Gillmann et al., 2020) is also a valuable tool that is usually performed to verify whether they correspond to cases that a human would consider uncertain and usually be used for exploring, interpreting, and communicating uncertainty.

The most controversial point of the current research method is the lack of uncertainty ground truth. With ground truth uncertainty, measuring the exact agreement between estimated and actual uncertainties becomes possible, while the lack of ground truth makes it challenging to assess the accuracy and calibration of uncertainty estimates quantitatively, reducing the development of uncertainty evaluation methods.

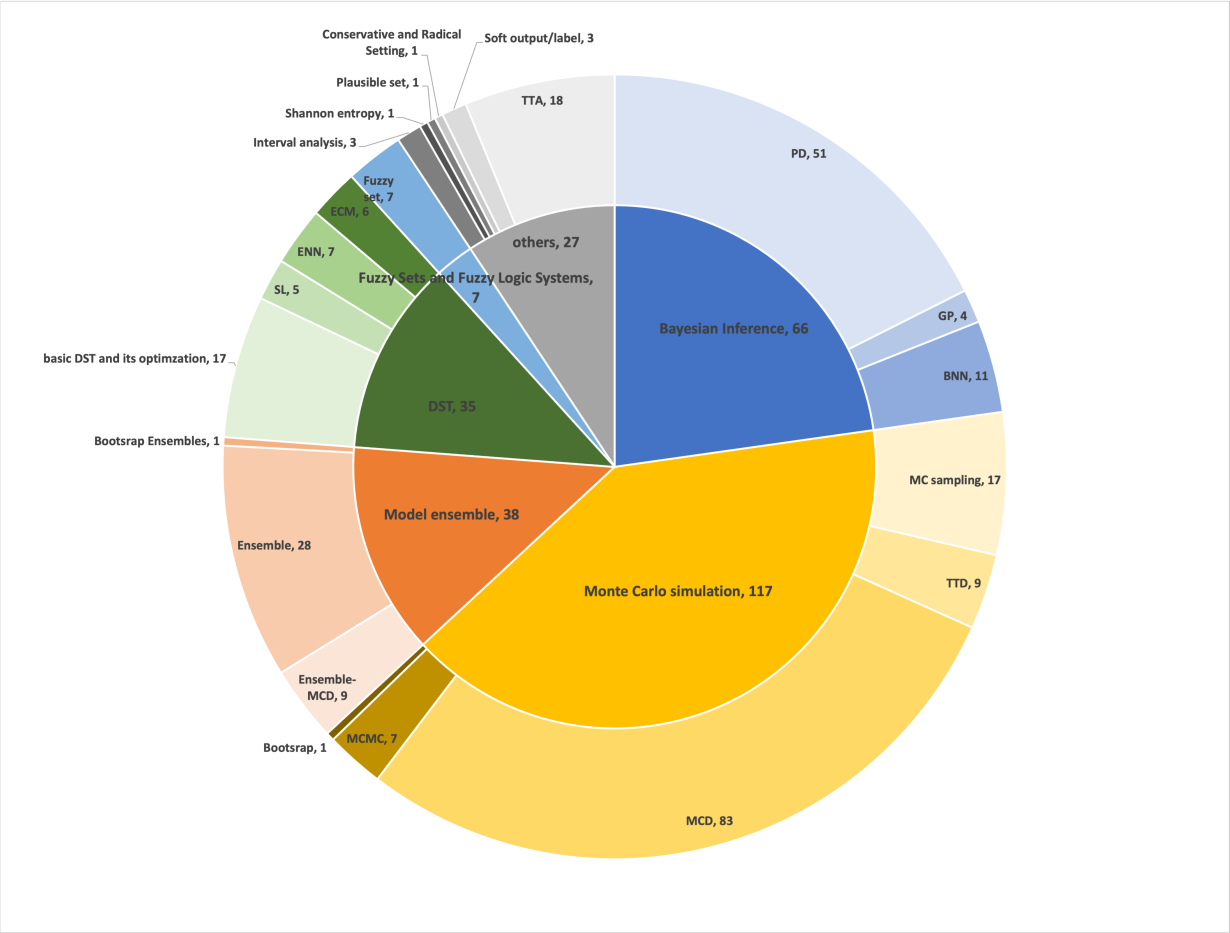


Figure 5: Statistics of uncertainty methods used in medical image analysis

5. Applications of uncertainty quantification in MIA

The application of uncertainty quantification can help increase the accuracy of different MIA tasks. In MIA tasks, the uncertainty can be decomposed into three levels (Lakshminarayanan et al., 2017): pixel/voxel-level, instance-level and subject-level. *Pixel/voxel-level* uncertainty quantification is useful for interaction with physicians by providing additional guidance for correcting reconstruction/registration/segmentation results. *Instance-level* uncertainty is the uncertainty aggregated by a set of pixel/voxel-level uncertainty. Its quantification can be used to reduce the false discovery rate for detection/prediction/classification tasks. *Subject-level* uncertainty offers information on whether or not the model is about a subject. Therefore, quantifying uncertainty in MIA tasks is a critical step in advancing the field of medical imaging, allowing for better decision-making, fostering continual improvement of algorithms and risk assessment, and promoting transparency and trust between experts and patients, as well as ensuring safe and effective healthcare practices. Figure 5 shows the overall statistics of probabilistic and non-probabilistic uncertainty methods used in MIA tasks.

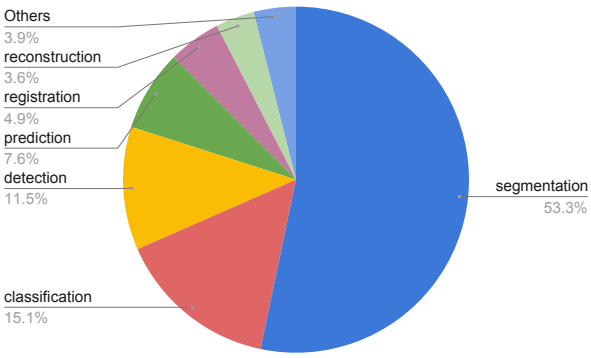


Figure 6: Distribution of different applications with uncertainty estimation in reviewed papers

In this section, we mainly focus on the introduction of recent research that studies the uncertainty in medical image reconstruction, registration, detection, prediction, classification, and segmentation. Figure 6 shows the application types.

Apart from the main applications mentioned above, uncertainties in other medical image tasks such as microstructure estimation (Ye et al., 2020; Adler et al., 2019), image quality estimation (Shaw et al., 2020), survival analysis (Feng et al., 2020; Gomes et al., 2021), risk analysis (Qian et al., 2020), image denoising (Laves et al., 2020b,a; Cui et al., 2022), cellularity assessment (Li et al., 2022b), lesion localization Wu et al. (2021b); Duchateau et al. (2016); Schobs et al. (2022), etc, are also mentioned and studied. Since the uncertainty is similar to the methods we introduced before, we will not go into details about those applications.

5.1. Medical image reconstruction

Medical image reconstruction plays a critical role in modern healthcare and medical imaging. It involves creating high-quality and accurate images of the internal structures of the human body from acquired raw data. The real-world factor is that medical imaging is subject to various sources of variability, including patient motion, imaging artifacts, and variations in imaging protocols. Therefore, developing reconstruction algorithms that can handle such variability and generalize well across different imaging scenarios is an ongoing challenge. Advanced ML reconstruction algorithms, despite providing good reconstruction performance, often lack reliability and explainability (e.g., understanding why a specific reconstruction was produced or tracing back when the results become unreliable), limiting the adoption and acceptance of these methods in clinical application. Therefore, studying reconstruction uncertainty is of great importance to ensure reconstruction reliability and provide explainable results. Table 2 shows the related medical image reconstruction methods considering reconstruction uncertainty.

5.1.1. Bayesian inference

Bayesian inference is the most common approach to quantifying reconstruction uncertainty. In 2014, Wallman et al. developed an electrical propagation model based on Bayesian

Table 2: Uncertainty quantification methods in medical image reconstruction

	Publications	Uncertainty methods	Number of Dataset	Clinical applications
MC methods	Neumann et al. (2014)	MCMC	1	Electromechanical heart M
	Zhou et al. (2020)	MCMC	1	PET reconstruction
	Luo et al. (2023)	MCMC	1	MRI reconstruction
	Edupuganti et al. (2020)	MCD	1	knee MRI reconstruction
Bayesian inference	Wallman et al. (2014)	PD	1	CT-derived ventricular mo
	Zhang et al. (2019)	PD	1	Knee MRI reconstruction
	Narnhofer et al. (2021)	PD	1	Undersampled MRI reconst
	Vlašić et al. (2023)	PD	1	Low/standard-dose PET re
	Barbano et al. (2021)	BNN	1	Sparse view CT reconstruc
	Wang et al. (2023a)	Fuzzy sets	1	COVID-19 CT reconstructi

inference with probabilistic distribution for tissue conduction properties inferred from electroanatomical data and designed strategies to optimize the location and number of measurements required to maximize information and reduce uncertainty (Wallman et al., 2014). The proposed method provides a simultaneous description of clinically relevant electrophysiological conduction properties and their associated uncertainty for various levels of noise.

In 2019, Zhang et al. proposed an uncertainty reduction model in undersampled MRI reconstruction with an active acquisition that, at inference time, dynamically selects the measurements to take and iteratively refines the prediction to reduce the reconstruction error and uncertainty (Zhang et al., 2019). The authors modeled pixel-level uncertainty as a Gaussian distribution centered at reconstruction mean and with variance similar to the method proposed by (Kendall and Gal, 2017).

In 2021, Narnhofer et al. introduced a Bayesian variational framework to quantify the epistemic reconstruction uncertainty (Narnhofer et al., 2021). They first solved the linear inverse problem of undersampled MRI reconstruction in a variational setting and then obtained epistemic uncertainty from a multivariate Gaussian distribution, whose mean and covariance matrix are learned in a stochastic optimal control problem. In the same year, Barbano et al. developed a scalable, data-driven, knowledge-aided computational framework to quantify reconstruction uncertainty via Bayesian neural networks (Barbano et al., 2021). This framework extended to a developed greedy iterative training scheme, deep gradient descent, and recast it within a probabilistic framework. The last layer of each block is Bayesian, with the rest of the layers remaining deterministic to achieve scalability. The framework is showcased on computed tomography with either sparse or limited view data and exhibits competitive performance with respect to SOTA benchmarks, e.g., total variation, deep gradient descent, and learned primal-dual.

In 2023, Vlavsic et al. proposed a DL-based posterior sampling method for uncertainty quantification in PET image reconstruction (Vlašić et al., 2023). The method is based on training a conditional generative adversarial network whose generator approximates sam-

pling from the posterior in Bayesian inversion. The generator is conditioned to a reconstruction from a low-dose PET scan obtained by a conventional reconstruction method. It can, therefore, generate corresponding standard dose PET images.

5.1.2. Monte Carlo methods

Monte Carlo methods are also popular for reconstruction uncertainty estimation. In 2014, Neumann et al. presented a stochastic method to estimate the parameters of an image-based electromechanical heart model and the corresponding uncertainty due to measurement noise (Neumann et al., 2014). First, Bayesian inference was applied to fully estimate the posterior probability density function (PDF) of the model. Second, MCMC sampling was used with computationally tractable designing that employed a fast Polynomial Chaos Expansion-based surrogate model instead of the true forward model. Then, the mean-shift algorithm was used to automatically find the modes of the PDF and select the most likely one while being robust to noise.

In 2020, Zhou et al. provided a framework for performing infinite-dimensional Bayesian inference and uncertainty quantification for image reconstruction with Poisson data (Zhou et al., 2020). They first introduced a positivity-preserving reparametrization and a dimension-independent MCMC algorithm based on the preconditioned Crank Nicolson Langevin method, in which a primal-dual scheme is used to compute the offset direction. Then, a fusion method that combines the model discrepancy and maximum likelihood estimation was proposed to determine the regularization parameter in the hybrid prior. In the same year, Edupuganti et al. quantified the image recovery uncertainty within DL models (Edupuganti et al., 2020). First, variational autoencoders (VAEs) were first leveraged to develop a probabilistic reconstruction scheme that maps out (low-quality) short scans with aliasing artifacts to the diagnostic-quality ones and then encoded the acquisition uncertainty in a latent code and naturally offers a posterior of the image from which one can generate pixel variance maps using MCD.

In 2023, Luo et al. introduced a framework that enables efficient sampling from learned probability distributions for MRI reconstruction where the samples were drawn from the posterior distribution given the measured k-space using MCMC (Luo et al., 2023). Therefore, in addition to the maximum posterior estimate for the image using the log-likelihood, the minimum mean square error estimate and uncertainty maps can also be computed from those drawn samples.

5.1.3. Non-probabilistic methods

Fuzzy theory can also be applied to quantify image reconstruction uncertainty. In 2023, Wang et al. proposed a new fuzzy metric to characterize image reconstruction uncertainty. It first designed a fuzzy hierarchical fusion attention neural network based on multiscale guided learning (Wang et al., 2023a) to convert input images into a fuzzy domain using fuzzy membership functions. The uncertainty of the pixels was processed using the proposed fuzzy rules, and then the output of the fuzzy rule layer was fused with the result of the convolution in the neural network. Simultaneously, a multiscale guided-learning dense residual block and pyramidal hierarchical attention module were designed to extract hierarchical image

Table 3: Uncertainty quantification methods in medical image registration

	Publications	Uncertainty methods	Number of dataset	Clinical applications
MC methods	Risholm et al. (2013)	MCMC	1	Neurosurgery for resection c
	Le Folgoc et al. (2016)	MCMC	1	Medical image registration
	Xu et al. (2022c)	MCD	2	Abdominal CT-MRI registra
Bayesian inference	Oreshkin and Arbel (2013)	PD	1	Medical image registration
	Parisot et al. (2014)	PD	1	Atlas to diseased patient reg
	Yang and Niethammer (2015)	PD	2	Heart&brain image registrat
	Wang et al. (2018b)	PD	2	Synthetic and brain MRI re
	Khawaled and Freiman (2022)	PD	4	Brain MRI registration
	Wachinger et al. (2014)	GP	2	MRI image registration
	Peter et al. (2021)	GP	4	Histology inter-modal, Opti Microscopy image and chest
Hybrid methods	Gong et al. (2022)	MCD, Bootstrap Ensembles	2	Deformable medical image r

information. Finally, a recurrent memory module with a residual structure was used to process the output features of the hierarchical attention modules and a recursive sub-pixel reconstruction module was used at the tail of the network to reconstruct the images.

5.2. Medical image registration

Medical image registration, a fundamental technique in medical image preprocessing, aligns and overlays multiple images of the same patient or anatomical region acquired at different times, from different modalities, or from different imaging devices. By aligning the images, it becomes easier to compare and analyze changes in anatomy or pathology over time. However, given the current SOTA registration technology and the difficulty of the problem, an uncertainty measure that highlights locations where the algorithm had difficulty finding a suitable alignment can be beneficial. According to our literature review, the predominant way to quantify the registration uncertainty is by using summary statistics of the transformation distribution. Table 3 listed the related papers. For medical image registration tasks, two probabilistic methods, Bayesian inference and MC methods, are mainly developed.

5.2.1. Bayesian inference

In 2013, Oreshkin et al. proposed a voxel selection strategy for medical image registration with the uncertainty of the transformation parameters (Oreshkin and Arbel, 2013). First, a Bayesian framework was used to build a voxel sampling probability field (VSPF) based on the variance of this optimal Bayesian estimator, different voxel subsets were then sampled based on the obtained VSPF.

In 2014, Parisot et al. presented a graph-based concurrent brain tumor segmentation and atlas to disease patient registration framework based on a unified pairwise discrete Markov Random Field (MRF) model with non-uniform sampling (Parisot et al., 2014),

following the sampling method proposed in (Oreshkin and Arbel, 2013). First, to get an appropriate sampling solution and reduce memory requirements, content-driven samplings of the discrete displacement set and the sparse grid were considered based on the local segmentation and registration uncertainties recovered by the min marginal energies (Kohli and Torr, 2008). Then, both segmentation and registration problems were modeled using a unified pairwise discrete MRF model on a sparse grid superimposed on the image domain. The registration uncertainty is then calculated by normalizing the min-marginals over all the possible displacements associated with the same segmentation label, while the segmentation uncertainty is evaluated by measuring the energy variation when the segmentation label changes.

In 2015, Yang et al. followed the idea that consisted of mapping displacement into uncertainty by energy information and approximating the covariance matrix by the inverse of the Hessian of the registration energy to quantify registration uncertainty for large deformation diffeomorphic metric mapping (Yang and Niethammer, 2015). The covariance matrix of the Gaussian process posterior distribution was also applied in (Wachinger et al., 2014) to estimate registration uncertainty.

In 2018, Wang et al. presented a large deformation diffeomorphic metric mapping approach similar to (Yang and Niethammer, 2015) that models posterior distribution with a Laplace approximation of Bayesian registration models (Wang et al., 2018b).

In 2021, Peter et al. introduced a principled strategy for the construction of a gold standard for deformable registration by building on the true transformation into a Gaussian process model and then annotating the most informative location in an active learning fashion to minimize the uncertainty of the true transformation (Peter et al., 2021). It should be noted that, in addition to a landmark correspondence for each queried location, this framework supports the specification of an annotation uncertainty, either directly estimated by the annotator or obtained by merging annotations from multiple users.

In 2022, khawaled et al. developed a non-parametric Bayesian method to assess the uncertainty in diffeomorphic deformable MRI registration (Khawaled and Freiman, 2022). It sampled the true posterior distribution of the network weights by noise injection in the training loss gradients with the Adam optimizer and estimated the registration uncertainty according to the voxel-wise diagonal variance.

5.2.2. Monte Carlo methods

In 2013, Risholm et al. proposed a non-rigid registration framework where conventional dissimilarity and regularization energies were included in the likelihood and the prior distribution on deformations, respectively, through Boltzmann's distribution (Risholm et al., 2013). MCMC was used to characterize the posterior distribution with Boltzmann temperature hyper-parameters marginalized under broad uninformative hyper-prior distributions, permitting the estimation of the most likely deformation as well as the associated uncertainty.

In 2016, Le Folgoc et al. investigated uncertainty quantification under a sparse Bayesian model of medical image registration with a focus on the theoretical and empirical quality of uncertainty estimates derived under the approximate scheme and under the exact model

(Le Folgoc et al., 2016). In this paper, the authors implemented an (asymptotically) exact inference scheme based on reversible jump MCMC sampling to characterize the posterior distribution of the transformation.

In 2022, Xu et al. proposed a mean-teacher registration framework, which incorporates an additional temporal consistency regularization term by encouraging the teacher model's prediction to be consistent with that of the student model (Xu et al., 2022c). Instead of searching for a fixed weight, the teacher model enables automatically adjusting the weights of the spatial regularization and the temporal consistency regularization by taking advantage of the transformation uncertainty and appearance uncertainty calculated based on MCD.

5.2.3. Hybrid methods

Compared with other probabilistic uncertainty quantification methods, model ensemble is less popular and is usually used with other methods to construct a hybrid model. In 2023, Gong et al. (Gong et al., 2022) proposed a predictive module to learn the registration and uncertainty in correspondence simultaneously by inducing three empirical randomness and registration error-based uncertainty prediction methods: MCD, deep ensembles, and Bootstrap.

In general, the majority of existing research focuses on trying out different summary statistics as well as means to exploit registration uncertainty. Those researches do have promising contributions, e.g., risk assessment based on the trustworthiness of the registered image data.

5.3. Medical image detection

Medical image detection, aiming at detecting small or subtle abnormalities, anatomical structures, lesions, tumors, or other pathologies, plays a vital role in early diagnosis, treatment planning, and medical conditions monitoring. Images used for detection may have low contrast, low signal-to-noise ratio, or be overshadowed by surrounding structures. These factors can make it difficult for detection methods to identify and localize small objects accurately, leading to false negatives or reduced sensitivity. Therefore, it is necessary to estimate the detection uncertainty. Table 3 listed the related papers.

5.3.1. Bayesian inference

In 2020, Araujo et al. proposed a deep learning-based Diabetic Retinopathy grading computer-aided diagnosis system that supports its decision by providing a medically interpretable explanation and estimation of prediction uncertainty with a novel Gaussian-sampling approach and a multiple-instance learning framework, allowing the ophthalmologist to measure how much that decision should be trusted (Araujo et al., 2020). In the same year, Mao et al. proposed an abnormality detection approach based on an autoencoder that outputs not only the reconstructed normal version of the input image but also a pixel-wise uncertainty prediction with probabilistic distribution (Mao et al., 2020).

In 2021, Sudarshan et al. proposed a sinogram-based uncertainty-aware deep BNN framework to estimate a standard-dose PET image (Sudarshan et al., 2021). Here, the detection uncertainty is modeled through the per-voxel heteroscedasticity of the residuals between the

Table 4: Uncertainty quantification methods in medical image detection

	Publications	Uncertainty methods	Number of dataset	Clinical applications
Bayesian inference	Kwon et al. (2020)	PD	2	Retinal blood vessels detecti
	Araujo et al. (2020)	PD	2	Diabetic retinopathy grading
	Mao et al. (2020)	PD	2	Lung abnormal detection
	Akrami et al. (2021)	PD	2	Brain lesions detection
	Jafari et al. (2021)	PD	1	Video keyframes detection
	Sudarshan et al. (2021)	PD	1	PET-MRI OoD detection
	Wang et al. (2022b)	PD	1	Abnormal lymph nodes dete
	Huang et al. (2022a)	BNN	5	Anomaly detection
MC methods	Leibig et al. (2017)	MCD	1	Diabetic retinopathy detecti
	Gill et al. (2019)	MCD	1	Focal cortical dysplasia dete
	Nair et al. (2020)	MCD	1	Sclerosis lesion detection
	Ghoshal and Tucker (2020)	MCD	1	COVID-19 detection
	Yang et al. (2021)	MCD	1	Lung nodule detection
	Dong et al. (2021)	MCD	1	COVID-19 detection
	Calderon et al. (2021)	MCD	1	Breast cancer detection
	Tang et al. (2022)	MCD	4	Retinal vessel detection
	Ghoshal and Tucker (2021)	MC sampling	1	COVID-19 detection
	Bhat et al. (2021)	TTD	1	Liver lesions detection
TTA	Ayhan et al. (2020)	TTA	2	Diagnosing diabetic retinopa
	Ayhan and Berens (2022)	TTA	1	Diabetic retinopathy detecti
DST	Ben Atitallah et al. (2022)	Basic DST	1	Pneumonia diagnosis
	Rahman et al. (2023b)	Basic DST	1	Fetal plane detection
Model ensemble	Kabir et al. (2022)	Ensemble	1	COVID-19 detection
Interval analysis	Mazoure et al. (2022)	Confidence interval	1	Skin cancer detection
Hybrid methods	Tabarisaadi et al. (2022)	MCD, Ensemble, Spectral GP	1	Skin cancer detection
	Asgharnezhad et al. (2022)	MCD, Ensemble, Ensemble-MCD	1	COVID-19 detection
	Javadi et al. (2022)	TTD, TTA	1	Prostate cancer detection
	Abdar et al. (2023)	Ensemble-MCD	2	COVID-19 detection
	Linmans et al. (2023)	MCD, Ensembles	5	Lymph node tissue, prostate cancer/biopsies, foreign tissu

predicted and the high-quality reference images. Jafari et al. presented a video keyframe landmark detection framework by leveraging the uncertainty of landmark prediction obtained from a deep Bayesian network (Jafari et al., 2021). Akrami et al. described a quantile regression VAE model to avoid variance shrinkage problems by estimating conditional quantiles for the given input image (Akrami et al., 2021). Using the estimated quantiles, the conditional mean and variance for input images were computed under the Gaussian model to estimate detection uncertainty.

In 2022, Huang et al. presented an uncertainty-aware prototypical transformer model, considering both the anomaly diversity and uncertainty to achieve accurate pixel-level visual anomaly detection (Huang et al., 2022a). First, a memory-guided prototype learning transformer encoder was designed to learn the diversity of prototypical representations of anomalies. Second, an anomaly detection uncertainty quantizer was designed by a Bayesian Neural Network with Gaussian distribution to learn the distributions of anomaly detection. Then, an uncertainty-aware transformer decoder was proposed to leverage the detection uncertainties to guide the model to focus on the uncertain areas. In the same year, Wang et al. proposed an improved Mask RCNN framework with a global-local channel attention mechanism and multi-task Gaussian inference-based uncertainty loss for the detection of abnormal lymph nodes in MR images (Wang et al., 2022b).

5.3.2. Monte Carlo methods

In 2017, Leibig et al. evaluated the impact of MCD-based Deep Bayesian uncertainty measures in diagnosing diabetic retinopathy and showed that uncertainty-informed decision referrals could improve diagnostic performance (Leibig et al., 2017). Similar research has been investigated in (Gill et al., 2019), (Ghoshal and Tucker, 2020) and (Nair et al., 2020) for the detection of COVID-19, focal cortical dysplasia detection and lesion, respectively.

In 2021, Yang et al. improved performance of a detection CNN performance with two different bounding-box-level (or instance-level) uncertainty estimates with predictive variance and MC sampling variance, respectively (Yang et al., 2021); Dong et al. proposed a novel deep network for robust COVID-19 detection that employs Deformable Mutual Information Maximization (DeIM), Mixed High-order Moment Feature (MHMF), and Multiexpert Uncertainty-aware Learning (MUL) (Dong et al., 2021). With DeIM, the mutual information between input data and the corresponding latent representations can be estimated and maximized to capture compact and disentangled representational characteristics. MHMF is used to extract discriminative features of complex distributions, and MUL creates multiple parallel MCD networks for each image to evaluate uncertainty and thus prevent performance degradation caused by the noise in the data.

The same year, Ghoshal et al. proposed a Bayesian inference model with MC sampling (Ghoshal and Tucker, 2021) for uncertainty quantification and measured bias-corrected uncertainty using the Jackknife resampling technique (Sahinler and Topuz, 2007); Bhat et al. proposed to use TTD to reduce false positive detections made by a neural network using an SVM classifier trained with features derived from the uncertainty map of the neural network prediction (Bhat et al., 2021).

Later in this year, Calderón-Ramírez et al. explored the impact of using unlabeled data

through the implementation of a recent successful semi-supervised approach, MixMatch (Berthelot et al., 2019), for breast cancer detection on mammogram images (Calderon et al., 2021). They improved uncertainty estimations, i.e., Normalized entropy of Softmax, Maximum value of Softmax and MCD, by using unlabeled data under regimes with a very limited number of labeled observations for training. Moreover, following (Asgharnezhad et al., 2022), the authors used the proposed "uncertainty confusion matrix" that groups uncertainty estimations for each of a model's predictions according to their "correctness" and "confidence." Based on this, the authors proposed an uncertainty-balanced accuracy to ease the comparison of uncertainty estimation approaches in real-world usage scenarios.

5.3.3. Model ensemble

In 2022, Kabir et al. proposed an aleatory-aware deep uncertainty quantification method for COVID-19 detection with an application for transfer learning and deep ensembles that converted the outputted K-nearest posteriors of each DNN into opacity scores to represent aleatory uncertainty (Kabir et al., 2022).

5.3.4. Non-probabilistic methods

In 2020, Ayhan et al. studied an intuitive framework based on TTA to quantify the diagnostic uncertainty of a state-of-the-art DNN for diagnosing diabetic retinopathy (Ayhan et al., 2020). Based on the first work, Ayhan et al. proposed a simple but effective method using traditional data augmentation methods such as geometric and color transformations at test time, allowing us to examine how much the network output varies in the vicinity of examples in the input spaces (Ayhan and Berens, 2022).

In 2022, Ben et al. proposed a disease detection approach based on a DST-based evidence fusion theory, allowing the combination of a set of deep learning classifiers to provide more accurate disease detection results (Ben Atitallah et al., 2022). The main contribution of this work is the application of Dempster's rule for the fusion of five pre-trained convolutional neural networks (CNNs), including VGG16, Xception, InceptionV3, ResNet50, and DenseNet201 for the diagnosis of pneumonia from chest X-ray images. In the same year, Mazoure et al. released a web server, Deep Uncertainty Estimation for Skin Cancer (DUESC) (Mazoure et al., 2022), that performs an intuitive, in-depth analysis of uncertainty in commonly used skin cancer classification models based on CNNs and confidence intervals.

5.3.5. Hybrid methods

In 2021, Javadi et al. proposed a UNet-based deep network for prostate cancer detection in systematic biopsy considering both the label and model uncertainty using TTA and TTD, respectively (Javadi et al., 2022). Uncertainty metrics were then used to report the cancer probability for regions with high confidence to help the clinical decision-making during the biopsy procedure.

In 2022, Tabarisaadi et al. studied the automatic diagnosis of skin cancer using dermatologist spot images (Tabarisaadi et al., 2022). Three different uncertainty-aware training algorithms (MCD, Model ensembling, and Spectral Normalized Neural Gaussian Process (Liu et al., 2020a)) were utilized to detect skin cancer. In the same year, Asgharnezhad et

Table 5: Uncertainty quantification methods in medical image prediction

	Publications	Uncertainty methods	Number of Dataset	Clinical application
Bayesian inference	Bliesener et al. (2019)	PD	1	Brain tumor longitudinal monitorin
	Corrado et al. (2020)	PD	1	Left atrium electro-physiology simu
	Wu et al. (2021b)	Sparse GP	2	Bone age prediction and lesion loca
MC methods	Rafael-Palou et al. (2022)	MC sampling	1	Lung tumour growth prediction
	Corrado et al. (2023)	MC sampling	1	Left atrium anatomy prediction
	Huang and Chung (2020)	MCD	4	Autism spectrum disorder, Alzheim and ocular diseases prediction
	Hemsley et al. (2020)	MCD	1	Brain metastases/glioblastoma radi
	Kannan et al. (2021)	MCD	1	Assessment of paediatric dysplasia
	Dolezal et al. (2022)	MCD	2	lung adenocarcinoma and squamous carcinoma prediction
DST	Lian et al. (2016b)	ECM	2	Lung and esophageal cancer treatm
	Lian et al. (2016a)	ECM	3	Lung, lymph and esophageal cancer treatment outcomes prediction
	Wu et al. (2018)	Basic DST	4	Cancer treatment outcome predictio
	Liu et al. (2023a)	Basic DST	1	Knee replacement prediction
	Ahmad et al. (2023)	Basic DST	2	COVID-19 progression and prognos
Hybrid methods	Jensen et al. (2019)	Ensemble, TTA MC sampling, MCD	1	Skin conditions prediction

al. applied and evaluated three uncertainty quantification techniques, MCD, Ensembles and Ensembles-MCD, for COVID-19 detection (Asgharnezhad et al., 2022). Moreover, a novel concept of uncertainty confusion matrix was proposed and new performance metrics for the objective evaluation of uncertainty estimates were introduced.

In 2023, Abdar et al. presented a simple but efficient deep learning feature fusion model, UncertaintyfuseNet, for COVID-19 detection by using the Ensemble-MCD technique to model detention uncertainty and the obtained results prove the efficiency of the model with robustness to noise and unseen data (Abdar et al., 2023). In the same year, Linmans et al. provided a benchmark for evaluating prevalent uncertainty methods by comparing the uncertainty estimates on both ID and realistic near and far OoD data on a whole-slide level using MCD and model ensembles (Linmans et al., 2023).

5.4. Medical image prediction

Radiomics aim to predict future outcomes or conditions from medical images. Although it has been widely studied recently, it also has certain limitations. For example, disease progression in many medical conditions is complex and multifactorial. Predicting the progression or response to treatment from medical images alone may oversimplify the underlying dynamics. Moreover, radiomic methods often encounter uncertainty and variability in image-based measurements. Quantifying and addressing these uncertainties is crucial for reliable predictions and their subsequent use in clinical decision-making. Table 5 lists the related work.

5.4.1. Bayesian inference

In 2020, Corrado et al. used a Bayesian probabilistic approach to detect the left atrium derived from cardiac MRI and to quantify the uncertainty about the shape (Corrado et al., 2020). In 2021, Wu et al. proposed an uncertainty-aware deep kernel learning model that permits the estimation of the uncertainty in the prediction by a pipeline of a CNN and a sparse Gaussian Process (Wu et al., 2021b). In 2022, Rafael et al. proposed a deep hierarchical generative and Bayesian probabilistic network that, given an initial image of the nodule, predicts whether it will grow, quantifies its future size and provides its expected semantic appearance at a future time and estimates the uncertainty in the predictions from the intrinsic noise in medical images and the inter-observer variability in the annotations (Rafael-Palou et al., 2022). In 2023, Corrado et al. described the left atrium anatomy using a Bayesian shape model that captures anatomical uncertainty in medical images and validated the model on independent clinical images (Corrado et al., 2023).

5.4.2. Monte Carlo methods

In 2018, Jungo et al. proposed an MCD-based full-resolution residual CNN for brain tumor segmentation and survival prediction(Jungo et al., 2018a). In 2019, Bliesener et al. used a neural network to estimate the approximate joint posterior distribution of tracer-kinetic parameters, where uncertainties are estimated for each voxel and are specific to the patient, exam, and lesion (Bliesener et al., 2019). The predicted parameter ranges correlate well with tracer-kinetic parameter ranges observed across different noise realizations and regression algorithms.

In 2020, Huang et al. proposed a concept of MC edge dropout to estimate the predictive uncertainty related to the graph topology (Huang and Chung, 2020). After that, Hemsley et al. proposed an MCD-based conditional generative adversarial model for brain metastases or glioblastoma radiation treatment prediction (Hemsley et al., 2020) and Dolezal et al. trained Bayesian Neural models with MCD to identify lung adenocarcinoma and squamous cell carcinoma (Dolezal et al., 2022).

5.4.3. Non-probabilistic methods

For medical image prediction tasks, DST is the most commonly used non-probabilistic uncertainty quantification method. In 2016, Lian et al. proposed a radiomics feature-based radiotherapy treatment outcomes prediction system using a feature selection method based on DST for modeling and reasoning with uncertain and/or imprecise information (Lian et al., 2016b). The proposed method aimed to reduce the imprecision and overlaps between different classes in the selected feature subspace, thus finally improving the prediction accuracy. Based on the proposed feature selection model, Lian et al. proposed a radiotherapy treatment outcomes prediction system that uses EKNN for radiomic features selection with the consideration of a data balancing procedure and specified prior knowledge (Lian et al., 2016a). After that, Wu et al. proposed a similar method for cancer treatment outcome prediction with a feature selection module and an EKNN classifier (Wu et al., 2018).

In 2023, Ahmad et al. presented a complete COVID-19 progression and prognosis prediction framework using a two-stage reasoning process based on the DST (Ghesu et al., 2021).

In the same year, Liu et al. proposed an evidence-aware multi-modal data fusion framework based on DST that considers the reliability of each source data and the prediction output when making a final decision (Liu et al., 2023a). The backbone models contain an image, a non-image branch and a fusion branch. For each branch, there is an evidence network that takes the extracted features as input and outputs an evidence score, which is designed to represent the reliability of the output from the current branch. The output probabilities, along with the evidence scores from multiple branches are combined with Dempster's combination rule to make a final prediction.

5.4.4. Hybrid methods

In 2019, Jensen et al. experimentally showed that models trained to predict skin conditions become overconfident and then proposed to train models with a label sampling scheme that takes advantage of inter-rater variability to achieve a better-calibrated model (Jensen et al., 2019). Thus, Model Ensemble, TTA, MC Batch Normalization (Teye et al., 2018) and MCD were used to quantify prediction uncertainty.

5.5. Medical image classification

Similar to previous MIA tasks, the performance of medical image classification methods depends on the quality of the image itself and the corresponding annotations. Quantifying instance-level uncertainty helps to classify images where the classification model might be uncertain or incorrect, allowing for manual correction or expert review and improving diagnosis quality and treatment planning. Considering that we have already introduced the main uncertainty quantification methods in sections 3.1 and 3.2, and also the research focused on image classification is similar to the medical image analysis tasks mentioned earlier, here we only briefly describe their corresponding methods, datasets and clinical applications in Table 6.

5.6. Medical image segmentation

Medical image segmentation is more challenging than classification tasks due to the inherent variations in the appearance of anatomical structures, leading to potential errors or inaccuracies in defining boundaries or segment structures. Therefore, quantifying pixel/voxel-level uncertainty helps identify regions where the model might be uncertain or incorrect, allowing for manual correction or expert review and improving radiotherapy treatment performance. Table 8, 7 and 9 list three main probabilistic uncertainty quantification methods used in medical image segmentation tasks. Table 10 shows the most frequent non-probabilistic uncertainty methods DST and Table 11 shows the rest of the non-probabilistic uncertainty methods. Table 12 shows the hybrid uncertainty quantification methods. Among the retrieved methods, MCD and ENN is the most commonly used probabilistic and non-probabilistic uncertainty quantification method for medical image segmentation, respectively.

In MIA tasks, fully supervised learning has gained huge success based on the satisfying condition that large-scale annotated training datasets are available (Ronneberger et al., 2015; Myronenko, 2018; Isensee et al., 2018). However, region labeling in medical image

Table 6: Uncertainty quantification methods in medical image classification

	Publications	Uncertainty methods	Number of Dataset	Clinical applications
DST	Tardy et al. (2019)	SL	2	Mammograms classification
	Ghesu et al. (2019)	SL	2	Chest radiograph assessment
	Yuan et al. (2020)	ENN	2	Breast infiltrating ductal carc radiograph pneumonia classif
	Huang et al. (2021c)	ENN	1	COVID-19 classification
	Ghesu et al. (2021)	SL	2	Chest radiographs abnormalit
	Xu et al. (2022a)	SL	2	Pancreatic tumor subtype an
	Liu et al. (2023b)	DST with new basic probability assignment	1	Grading of breast cancer
MC methods	Abdar et al. (2021a)	MCD	4	COVID-19, chest, optical coh and skin cancer classification
	Ju et al. (2022)	MCD	3	Skin lesions, prostate cancer and retinal diseases classificat
	Valen et al. (2022)	MCD	2	Chest and skin cancer classifi
	Feng et al. (2022)	MCD	3	Optical coherence tomograph and chest classification
	Ahsan et al. (2022)	MCD	1	Diabetic retinopathy classific
	Abdar et al. (2022)	MCD	3	Retinal OCT, lung and chest
	Aljuhani et al. (2022)	MCD	1	Tumor region classification
	Ghoshal et al. (2022)	MC sampling	2	Pancreatic adenocarcinoma g
Bayesian inference	Peressutti et al. (2013)	PD	4	Cardiac interventions
	Thiagarajan et al. (2021)	BNN	1	Breast histopathology images
	Belharbi et al. (2021)	PD	2	Histology images classificatio
	Liu and Zheng (2022)	PD	2	Skin lesion and thorax diseas
	Jiménez-Sánchez et al. (2022)	PD	1	Femur fracture classification
Ensemble	Senousy et al. (2021)	Ensemble	1	Breast cancer classification
	Qendro et al. (2021)	Early exit ensemble	3	Heart attack, epileptic seizur and skin melanoma classificat
	Arco et al. (2023)	Ensemble	1	Bacterial/viral pneumonia cla
Fuzzy sets	Pham (2014)	Fuzzy sets	1	Hernia mesh classification
	Rahman et al. (2023a)	Fuzzy sets	1	Brain tumour classification
Others	Galdran et al. (2019)	Soft label	5	Retinal images classification
	Del Amor et al. (2023)	Soft label	1	Histology image classification
Hybrid methods	Carneiro et al. (2020)	PD, TTA	1	Polyp classification
	Abdar et al. (2021c)	MCD, Ensemble Ensemble-MCD	2	Skin cancer classification
	Gour and Jain (2022)	MCD, CI	3	Breast histopathology images
	Yang and Fevens (2021)	MCD, Ensemble, and Ensemble-MCD	2	COVID-19 and breast tumor
	Dawood et al. (2023)	PD, TTA	2	Cardiac classification
	Cifci (2023)	TTD	1	Lung cancer diagnosis and tr
	Mehta et al. (2023)	Ensemble-MCD	3	Skin lesion classification
	Hamedani et al. (2023)	MCD, Ensemble, and Ensemble-MCD	1	Breast cancer classification

segmentation tasks requires skilled expertise with domain knowledge and careful delineation of boundaries. The contradiction between the increasing demand for segmentation accuracy on the one hand, and the shortage of perfect (precise and reliable) annotations on the other hand has so far limited the performance of learning-based medical image segmentation methods. Therefore, in this section, we focus our uncertainty analysis on semi-supervised medical image segmentation.

Techniques for semi-supervised medical image segmentation can be divided into three groups: graph-constrained methods (Xu et al., 2016; Reiß et al., 2022), self-learning methods (Li et al., 2019; Min et al., 2019), and generative adversarial learning methods (Mondal et al., 2018; Sun et al., 2019). Though these methods can break the dependence of machine learning models on training labels and the experimental results are promising, the uncertainty caused by the low quality of the images and the lack of annotations still need to be further studied for a more accurate and reliable medical image segmentation model.

According to our literature review, current uncertainty-based semi-supervised learning methods (the methods be marked in blue color in Tables 8, 7, 9, 10, 11 and 12) can be classified into two main groups: consistency learning (Yu et al., 2019; Shi et al., 2021) and uncertainty-aware learning (Sedai et al., 2019; Meyer et al., 2021). Consistent learning regularizes the model's predictions to be consistent across different perturbations of the same input and imposes feature-level, data-level, model-level, or task-level consistency on unlabeled data. The common applications are to optimize a teacher-student or multi-view framework with consistent learning, where the teacher/main model provides consistent predictions for guiding the student/ auxiliary model. Uncertainty-aware learning integrates estimated uncertainty into the training process directly when dealing with a mix of labeled and unlabeled data. It leverages the unlabeled data to enhance the model's predictions while providing uncertainty estimates reflecting the model's confidence in those predictions.

In the rest of the section, we will introduce the semi-supervised medical image segmentation methods with uncertainty quantification in detail.

5.6.1. Bayesian inference

Consistent learning. In 2021, Shi et al. presented a conservative radical network with probabilistic uncertainty estimation for medical image segmentation (Shi et al., 2021). The general idea is that if the segmentation result of a pixel becomes inconsistent, this pixel shows a relative uncertainty with probabilistic distribution.

In 2023, Shi et al. proposed an uncertainty-weighted prediction consistency training strategy and a relation-driven consistency training strategy in a semi-supervised fashion for nasopharyngeal carcinoma segmentation (Shi et al., 2023). The architecture was composed of a shared encoder, a main decoder, and several auxiliary decoders. Various perturbations were applied to the shared encoder's output to leverage the unlabeled data and enforce consistency between the predictions of the main and auxiliary decoders and uncertainty estimation was applied to avoid being misled by unreliable outputs during training due to annotation scarcity.

Table 7: Bayesian inference-based uncertainty quantification for medical image segmentation. The semi-supervised methods are highlighted in blue.

Publications	Uncertainty methods	Number of Dataset	Clinical applications
Parisot et al. (2014)	PD	2	Low-grade glioma and brain tumor segmenta
Lê et al. (2016)	PD	1	Brain tumor segmentation
Ghoshal et al. (2019)	PD	1	Nuclei images segmentation
Wang et al. (2018a)	PD	2	Organs and brain tumor core segmentation
Behnami et al. (2019)	PD	1	Infants born MRI tumor segmentation
Ouyang et al. (2019)	PD	1	Pneumothorax segmentation
Baumgartner et al. (2019)	Hierarchical PD	2	thoracic and prostate segmentation
Camarasa et al. (2021)	PD	1	Carotid artery segmentation
Luo et al. (2021)	PD	1	Nasopharyngeal carcinoma segmentation
Zhang et al. (2021)	PD	1	Liver tumor segmentation
Zhao et al. (2021)	PD	1	Carotid artery segmentation
Li et al. (2022c)	PD	2	Subcortical structures segmentation
Li et al. (2021a)	Multi-head PD	1	Intracranial hemorrhage segmentation
Luo et al. (2021)	PD	1	Nasopharyngeal carcinoma segmentation
Mahani et al. (2022)	PD	1	Skin lesions segmentation
Wang et al. (2022c)	PD	2	Cardiac and skin lesion segmentation
Liu et al. (2022)	PD	2	Atrial, brain tumor, liver tumor segmentatio
Xie et al. (2022)	PD with confidence map	3	Ultrasound Image segmentation
Li et al. (2022a)	PD	1	Brain tumor segmentation
Diao et al. (2022)	PD	4	Soft tissue, lymphoma and liver tumor segme
Jones et al. (2022)	PD	1	Brain tumor segmentation, tissue class predi
Luo et al. (2022)	PD	3	Nasopharyngeal carcinoma, brain tumor and pancreas segmentation
Shi et al. (2023)	PD	2	Neck tumor segmentation
Zhang et al. (2023a)	PD	2	Atrial segmentation, brain tumor segmentati
Islam et al. (2023)	PD	2	Breast segmentation
Sedai et al. (2019)	BNN	1	Optical coherence tomography segmentation
Xia et al. (2020a)	BNN	2	Pancreas and liver tumor segmentation
Bian et al. (2020)	BNN	2	Retinal OCT images segmentation
Kwon et al. (2020)	BNN	2	Ischemic stroke lesion segmentation, blood v
Senapati et al. (2020)	BNN	1	Liver segmentation and disease classification
Krygier et al. (2021)	BNN	2	Spine and aorta segmentation
Li et al. (2021b)	BNN	2	Lung and nasal endoscopy segmentation

Table 8: MC methods-based uncertainty quantification for medical image segmentation. The semi-supervised methods are highlighted in blue.

Publication	Uncertainty methods	Number of dataset	Clinical applications
Jungo et al. (2018b)	MCD	2	Synthetic and brain tumor segmentation
Jungo et al. (2018a)	MCD	1	Brain tumor Segmentation, Survival Prediction
Seeböck et al. (2019)	MCD	6	Retinal OCT anatomy segmentation
Hu et al. (2019)	MCD	2	Lung nodule and prostate segmentation
Yu et al. (2019)	MCD	1	Left atrium segmentation
Soberanis-Mukul et al. (2020)	MCD	2	Pancreas and spleen segmentation
Wang et al. (2020b)	MCD	2	Left atrium and kidney segmentation
Xia et al. (2020b)	MCD	4	Pancreas segmentation
Monteiro et al. (2020)	MCD	2	Thorax and brain tumor segmentation
Ruan et al. (2020)	MCD	1	Renal tumors segmentation
Liu et al. (2020b)	MCD	1	Prostate zonal segmentation
Hu et al. (2020)	MCD	1	Natural killer T cell and lymphoma segmentatio
Nair et al. (2020)	MCD	1	Sclerosis lesion detection and segmentation
Wickstrøm et al. (2020)	MCD	1	Polyp segmentation
Hasan and Linte (2021)	MCD	1	Cardiac segmentation
Meyer et al. (2021)	MCD	3	Prostate zones segmentation
Wu et al. (2021a)	MCD	2	Mitochondria segmentation
Cao et al. (2021)	MCD	1	Breast segmentation
Wang et al. (2021a)	MCD	2	Cardiac and prostate segmentation
Ghoshal et al. (2021)	MCD	2	Cell and brain tumor detection
Rousseau et al. (2021)	MCD	2	Ischemic stroke and brain tumor segmentation
Balagopal et al. (2021)	MCD	1	post-operative prostate cancer radiotherapy
Wang et al. (2021b)	MCD	3	Thoracic, white matter and skin lesion segmenta
Silva and Oliveira (2021)	MCD	4	Brain growth, brain tumor, kidney and prostate
Wang et al. (2023b)	MCD	3	Thoracic skin lesion and brain's white matter tissue myelination process
Hu et al. (2022)	MCD	2	Nasopharyngeal carcinoma segmentation
Qiao et al. (2022)	MCD	3	Chest segmentation
Wang et al. (2022c)	MCD	1	Cardiac segmentation
Mojiri Forooshani et al. (2022)	MCD	2	white matter hyperintensity segmentation
Kuang et al. (2022)	MCD	1	Perihematomal edema segmentation
Tang et al. (2022)	MCD	4	Nasopharyngeal carcinoma, lung, optic disc segr
Judge et al. (2022)	MCD	3	Cardiac ultrasound, myocardial infarction and l
Wang et al. (2022a)	MCD	2	Cardiac and prostate segmentation
Xiao et al. (2022)	MCD	1	Cardiac segmentation
Zheng et al. (2022)	MCD	3	Cardiac, spinal cord gray matter and spleen seg
Xiang et al. (2022)	MCD	2	Left atrium and pancreas segmentation
Sambyal et al. (2022)	MCD	1	Brain tumor segmentation
Lu et al. (2023)	MCD	1	Atrial Segmentation
Farooq et al. (2023)	MCD	2	Breast masses segmentation
Zimmer et al. (2023)	MCD	1	Placenta segmentation
Norouzi et al. (2019)	MC sampling	1	Cardiac segmentation
Eaton-Rosen et al. (2019)	MC sampling	2	white-matter hyperintensity segmentation
Huang et al. (2020)	MC sampling	1	Atria and ventricles segmentation
Alonso-Caneiro et al. (2021)	MC sampling	1	Retinal OCT images segmentation
Zhao et al. (2022)	MC sampling	2	Cardiac segmentation
Chlebus et al. (2022)	MC sampling	5	Liver segmentation
Chen et al. (2022)	MC sampling	3	Cardiac, spinal cord gray matter and spleen seg
Wang et al. (2023c)	MC sampling	1	Dental panoramic caries segmentation
Arega et al. (2023)	MC sampling	2	Cardiac pathologies
Natekar et al. (2020)	TTD	1	Brain tumor segmentation
Redekop and Chernyavskiy (2021)	TTD	2	Skin lesion and liver segmentation

Table 9: Model Ensemble uncertainty quantification for medical image segmentation. The semi-supervised methods are highlighted in blue.

Publications	Uncertainty method	Number of dataset	Clinical applications
Nath et al. (2020)	Ensemble	2	Pancreas and tumor segmentation
Fuchs et al. (2021)	Ensemble	1	Brain tumor segmentation
Cao et al. (2020)	Ensemble	1	Breast mass segmentation
Li et al. (2021c)	Ensemble	1	COVID-19 lesion segmentation
Kushibar et al. (2022)	Ensemble	2	Breast cancer and cardiac segmentation
Guo et al. (2022)	Ensemble	4	Cardiac segmentation
Buddenkotte et al. (2023)	Ensemble	2	Cancer and kidney tumor segmentation
Zhang et al. (2023b)	Ensemble	2	Tumor segmentation

Table 10: DST-based uncertainty quantification for medical image segmentation. The semi-supervised methods are highlighted in blue.

Publications	Uncertainty methods	Number of dataset	Clinical applications
Ghasemi et al. (2013)	Basic DST	2	Brain MRI segmentation
Lelandais et al. (2014)	ECM	1	Tumor estimation and dos
Makni et al. (2014)	ECM	1	Prostate multi-parametric
Liu et al. (2015)	DST with fuzzy c-means	1	Brain MRI segmentation
Derraz et al. (2015)	DST optimization	1	Non-small cell lung cancer
Xiao et al. (2017)	GD with Dempster's rule	1	Vascular segmentation
Lian et al. (2017c)	ECM	1	Tumor delineation
Lian et al. (2017a)	ECM	1	Tumor Segmentation
Lian et al. (2017b)	ECM	1	Lung cancer Segmentation
Lian et al. (2018)	ECM	1	Lung cancer Segmentation
Tavakoli and Ghasemi (2018)	DST with fuzzy c-means	1	Brain MRI segmentation
Lima and Islam (2019)	DST with fuzzy c-means	1	Brain MRI segmentation
Huang et al. (2021b)	ENN	1	Brain tumor segmentation
(Huang et al., 2021a)	ENN	1	Lymphoma segmentation
Huang et al. (2022c)	ENN	1	Brain tumor Segmentation
Huang et al. (2022d)	ENN, DST with Radial basis function	1	Lymphoma segmentation
Fidon et al. (2022)	DST with new basic probability assignment	1	fetal brain MRI segmentat
Hu et al. (2023)	SL	1	Liver tumor segmentation
Zou et al. (2023)	SL	3	Skin lesion, liver and brain
Zhang et al. (2023c)	DST with deep hyperspherical clustering	4	Brain MRI segmentation
Huang et al. (2023b)	ENN	1	Brain tumor segmentation

Table 11: Other non-probabilistic uncertainty quantification methods for medical image segmentation. The semi-supervised methods are highlighted in blue.

Publications	Uncertainty methods	Number of dataset	Clinical applications
Alberts et al. (2016)	TTA	15	Brain tumor segmentation
Wang et al. (2019b)	TTA	1	Brain tumor segmentation
Xu et al. (2022b)	TTA	1	Prostate ultrasound segmentation
Wu et al. (2023)	TTA	1	Fetal brain segmentation
Zheng et al. (2020a)	Fuzzy sets	2	Pancreas segmentation
Bertels et al. (2021)	Soft label	4	Lower-left third molar and brain
Shi et al. (2021)	Conservative and Radical Setting	3	Cancreas and endocardium segme
Adiga Vasudeva et al. (2022)	Plausible sets	1	Left atrium segmentation
Huang et al. (2022b)	Fuzzy logic theory	3	Breast segmentation

Uncertainty-aware learning. In 2021, Meyer et al. proposed an uncertainty-aware temporal self-learning (UATS) model to combine the techniques of temporal ensembling and uncertainty-guided self-learning to benefit from unlabeled images (Meyer et al., 2021). In the same year, Luo et al. proposed a semi-supervised medical image segmentation framework with uncertainty rectified pyramid consistency regularization in (Luo et al., 2021, 2022), where uncertainty is estimated via the KL-divergence among multi-scale predictions, which only need a single forward pass compared with MCD.

In 2022, Qiao et al. used a complementary uncertainty pairing rule to dilute the unreliability in semi-supervised learning by assembling reliable annotated data into unreliable unannotated data (Qiao et al., 2022), where a mixed sample data augmentation method was proposed to integrate annotated data into unannotated data for training the model in a low-unreliability manner. In the same year, Wang et al. proposed an uncertainty-guided pixel contrastive learning method (Wang et al., 2022c), where an uncertainty map for unlabeled data was constructed based on the entropy of the average probability distribution by a well-designed consistency learning mechanism, which generates comprehensive predictions for unlabeled data by encouraging consistent network outputs from two different decoders.

5.6.2. Monte Carlo methods

Consistent learning. In 2019, Yu et al. presented a teacher-student-based uncertainty-aware semi-supervised framework for left atrium segmentation (Yu et al., 2019) with an uncertainty-aware scheme that enables the student model to gradually learn from meaningful and reliable targets by exploiting the uncertainty information using MCD. Following the idea that explores uncertainty caused by lack of annotation, researchers optimized or extended semi-supervised or un-supervised MIA models that use the teacher-student framework with the MC methods. For example, Sedai et al. proposed an uncertainty-guided semi-supervised learning network based on a student-teacher framework for medical image segmentation with MCD (Sedai et al., 2019).

In 2022, Chen et al. proposed an MC Sampling-based uncertainty teacher-student framework with dense focal loss and deep co-training (Chen et al., 2022). In the same year, Xiao

et al. designed a teacher-student segmentation method through synchronous training and consistent regular constraints by screening uncertainty assessment with MCD during the training process (Xiao et al., 2022); Hu et al. proposed a two-stage teacher-student semi-supervised segmentation framework where an MCD-based uncertainty estimation was introduced to guide the student model to gradually learn reliable predictions from the teacher model (Hu et al., 2022). In 2023, Farooq et al. proposed a residual-attention-based MCD uncertainty-guided mean teacher framework that incorporates the residual and attention blocks (Farooq et al., 2023).

In addition to using the MC methods in the teacher-student framework, the MC methods are also popular in multi-view frameworks for uncertainty quantification. In 2020, Xia et al. proposed an uncertainty-aware multi-view co-training framework by exploiting the multi-viewpoint consistency of 3D medical images (Xia et al., 2020a,b). They applied co-training by enforcing multi-view consistency generated from MCD on unlabeled data, where an uncertainty estimation of each view is utilized to achieve accurate labeling. A similar approach can be found in (Wang et al., 2023c). In the same year, Zhang et al. proposed an MCD uncertainty-guided mutual consistency learning framework to effectively exploit unlabeled data by integrating intra-task consistency learning from up-to-date predictions for self-ensembling and cross-task consistency learning from task-level regularization to exploit geometric shape information (Zhang et al., 2023a).

Uncertainty-aware learning. In 2019, Sedai et al. proposed an uncertainty-guided semi-supervised learning network based on a student-teacher framework for medical image segmentation (Sedai et al., 2019). First, a teacher segmentation model was trained from the labeled samples using deep learning with MCD to generate soft segmentation labels and uncertainty maps for the unlabeled set. The student model was then updated using the softly segmented samples and the corresponding pixel-wise confidence of the segmentation quality estimated from the uncertainty of the teacher model using a newly designed uncertainty-based loss function. A similar method with an additional learnable uncertainty consistency loss was proposed in (Wang et al., 2020b).

In 2020, Soberanis-Mukul et al. proposed a segmentation refinement method based on MCD uncertainty analysis and graph convolutional networks (Soberanis-Mukul et al., 2020).

In 2022, Zheng et al. proposed an uncertainty-aware scheme to make models learn segmentation regions purposefully (Zheng et al., 2022). The model employed MCD as an estimation method to attain uncertainty maps, which serve as a weight for losses to force the models to focus on the valuable region according to the characteristics of supervised learning and unsupervised learning.

5.6.3. Model ensemble

Consistent learning. In 2021, Li et al. proposed a semi-supervised uncertainty-guided dual-consistency learning segmentation network (UDC-Net) that imposes image transformation equivalence and feature perturbation invariance to effectively harness the knowledge from unlabeled data (Li et al., 2021c). The segmentation uncertainty was then quantified in two forms: confidence uncertainty calculated by the entropy of the mean prediction of multiple

Table 12: Hybrids uncertainty quantification for medical image segmentation. The semi-supervised methods are highlighted in blue.

Publications	Uncertainty methods	Number of dataset	Clinical applications
Eaton-Rosen et al. (2018)	MC sampling, TTD	1	Brain tumour segmentation
Dhamala et al. (2018)	MCMC, PDF	2	Cardiac electrophysiology segm
Wang et al. (2019a)	MCD, TTA	1	Fetal brain and brain tumor se
Jungo and Reyes (2019)	MCD, Ensemble	2	Brain tumor and skin lesion se
Jungo et al. (2020)	MCD, Ensemble	1	Brain tumor segmentation
Venturini et al. (2020)	TTA, TTD	2	Hippocampal and fetal brain s
Zheng et al. (2020b)	Bootstrap, Ensemble	1	Cartilage segmentation
Wang et al. (2020a)	MCD, Ensemble, BNN	1	Fetal brain segmentation
Mehrtash et al. (2020)	MCD, Ensemble	5	Brain tumor, ventricular and p
Czolbe et al. (2021)	MCD, Ensemble, TTA	2	Skin lesion& lung cancer segm
Mehta et al. (2021)	MCD, Deep Ensemble, Ensemble-MCD	2	Lesion detection and brain tun
Zheng et al. (2021)	MC sampling, PD	3	Skin lesion segmentation
Lin et al. (2022b)	Fuzzy set, TTA	1	Skin lesion segmentation
Lin et al. (2022a)	MCD, fuzzy set, TTA	5	Skin lesion, nuclei, lung, breast
Pandey et al. (2022)	MCD, Ensemble, TTA	1	Ultrasound bone segmentation
Rajaraman et al. (2022)	MCD, Interval analysis	1	Tuberculosis segmentation
Ng et al. (2022)	MCD, Ensemble	2	Cardiac Segmentation
Sagar (2022)	MCD, Ensemble, Ensemble-MCD	1	Brain tumor segmentation
Ammari et al. (2023)	MCD, TTA, Shannon entropy	2	Right ventricular segmentation

perturbated inputs, and consensus uncertainty quantified by the standard deviation over the multi-decoders' predictions.

Uncertainty-aware learning. In 2020, Cao et al. presented an uncertainty-aware temporal ensembling model for semi-supervised breast mass segmentation (Cao et al., 2020). A temporal ensembling segmentation model was designed to segment breast mass using a few labeled and a large number of unlabeled images and an uncertainty map was estimated from MCD for each image; an adaptive ensembling momentum map and an uncertainty-aware unsupervised loss was designed and integrated with the temporal ensembling model.

5.6.4. Non-probabilistic methods

Compared to the probabilistic-based method to quantify uncertainty due to lack of annotation, there are only a few non-probabilistic researches that study uncertainty in semi-supervised medical image segmentation frameworks.

In 2022, Venturini et al. proposed an uncertainty-based method to improve the performance of segmentation networks when limited manual labels and estimated segmentation uncertainty on unlabeled data using TTA and TTD (Venturini et al., 2020). In the same year, Xiang et al. proposed a medical image segmentation framework that combines epistemic uncertainty-guided unsupervised learning and aleatory uncertainty-guided supervised learning with the ensemble of decoders (Xiang et al., 2022) Adiga et al. estimated the

pixel-level uncertainty by leveraging the labeling representation of segmentation into a set of plausible masks (Adiga Vasudeva et al., 2022).

In 2023, Huang et al. addressed the uncertainty caused by the low quality of the images and the lack of annotations using DST and deep learning (Huang et al., 2023b) with a semi-supervised learning algorithm proposed based on an image transformation strategy, a probabilistic deep neural network and an evidential neural network used in parallel to provide two sources of segmentation evidence, and Dempster's rule used to combine the two pieces of evidence and reach a final segmentation result.

In the same year, Xu et al. proposed a dual uncertainty-guided mixing consistency network with a contrastive training module that improves the quality of augmented images by retaining the invariance of data augmentation between original data and their augmentations (Xu et al., 2023). The dual uncertainty strategy calculates dual uncertainty obtained from N stochastic forward passes with random dropout between two models to select a more confident area for subsequent segmentation. The mixing volume consistency module guides the consistency between the volume before and after segmentation using dual uncertainty.

6. Discussion

In this section, we first list the key insights of applying uncertainty quantification in MIA and discuss the limitations. We then identify some potential future research points for readers' convenience.

6.1. Uncertainty quantification methods

First, the large number of studies incorporating uncertainty quantification in their medical analysis pipeline proves that the need to quantify uncertainty is well taken into account by the AI research community, showing that efforts are being made to bridge the gap between scientific research and clinical applications.

Bayesian inference, although providing a strong theoretical background for uncertainty, is scarcely implemented for medical image analysis because of the requirement for the modification of the NN weights and the training paradigm, as well as the slow convergence tends (Osawa et al., 2019) and noisy gradient descent (Jospin et al., 2022) in complex scenarios.

MC methods tended to be the most popular approach for uncertainty quantification in MIA, representing around half of the implemented methods. This popularity can be explained by its easy implementation in a large majority of neural networks trained with dropout. However, MC sampling requires multiple inferences for the same input image, considerably extending the inference time, which may not be compatible with high requirements in clinical efficiency.

Model ensemble is a popular trick to improve predictive performance while also providing quality uncertainty estimates. Similar to MC methods, it also has drawbacks in computational cost and efficiency.

Though the above probabilistic methods have gained enough attention in MIA and have achieved promising performance in estimating Out-of-Distribution (OoD) uncertainty when the model faces inputs that fall outside the range or distribution of the training data,

their limitations still remain when addressing or representing complex scenarios, e.g., In-Distribution (ID) uncertainty that arises from the inherent variability and noise within the dataset. For example, in the case of a multiclass problem (a three-class classification task ($\Omega = \{a, b, c\}$) as an example here), a good uncertainty model should be able to model the possible intermediate classes between the totally certain and totally uncertain about a class (i.e., any subset of Ω, e.g., $\{a, b\}$, $\{b, c\}$.), depending on the informativeness of the training data with respect to the class membership of the pattern under consideration (Denœux, 2000). Take three disease diagnoses as an example: an expert confirms that the patient does not have disease a but may have disease b or c; a good uncertainty model should then have the ability to model such ID uncertainty in an informative way, i.e, the degree of belief or plausibility that the patient be classified in to subset $\{b, c\}$. In practical scenarios, standard probabilistic uncertainty approaches, such as MCD or Ensemble, often fall short of effectively quantifying ID uncertainty. These approaches attempt to capture ID uncertainty by generating a set of predictions and calculating statistical indicators such as variance, offering only a singular uncertainty value without further context. Consequently, this limitation hampers the effectiveness of probabilistic methods in modeling ID uncertainty (Snoek et al., 2019; Ulmer and Cinà, 2021).

Non-probabilistic methods attract people's attention in modeling fuzzy, noisy, or uncertain information and motivate the development of methods tailored for uncertain both ID and OoD. Compared with the probabilistic uncertainty methods, non-probabilistic uncertainty quantification methods release the requirement of strong assumptions about the real distribution and modeling uncertainty based on fuzzy or soft conception. DST, the most popular non-probabilistic uncertainty method, can model OoD uncertainty with full ignorance about prediction and model ID uncertainty by providing comprehensive belief and plausibility context about any subset of Ω. Besides uncertainty quantification, DST also offers a way to combine multiple unreliable information, which is particularly useful in fusing multi-modality or cross-modality medical image data (Huang et al., 2022c). Moreover, the introduction of DST with neural networks, i.e., EKNN (Denoeux, 1995) and ENN (Denœux, 2000), makes it possible to integrate DST with SOTA deep learning models and, therefore, popularized its application in MIA. Other non-probabilistic uncertainty methods, such as fuzzy sets and fuzzy logic theory, interval analysis, and test time augmentation, although less frequently mentioned as DST, are also good choices for uncertainty quantification and can be further studied to integrate them with SOTA deep learning models.

6.2. Evaluation criteria

According to our literature review, a large variety of evaluation protocols are reported to assess the quality of uncertainty estimation. In the context of MIA, if multiple manual expert delineations are available for a given input image, the inter-rater variability is usually used as ground truth uncertainty to be compared with the predicted one. The related research has gained promising achievement and contributed to the development of uncertainty estimation in MIA. However, most of the time, the corresponding uncertainty values are not provided. Thus, evaluating uncertainty results relies on proxy tasks, such as detecting sample variance, predictive entropy, misclassification, OoD, or calibration performance.

One possible evaluation method is to determine whether performing a task that takes uncertainty into account improves the performance calculated on the criteria dedicated to this task, for example, the Dice coefficient for a segmentation task. These methods are inspired by concrete applications of uncertainty in a real-world scenario.

However, while several metrics exist to evaluate uncertainty estimation methods, none capture the complete picture. Metrics like calibration and coverage probability provide insights into specific aspects of uncertainty estimation but may not fully capture other important characteristics, such as the ability to capture epistemic and aleatory uncertainty separately. Therefore, we suggest researchers take task-depended clinical expectations/requirements into consideration when choosing uncertainty quantification evaluation criteria and ensure the fairness and pertinence of the evaluation criteria.

6.3. Applications

Analyzing uncertain information in image reconstruction and registration can improve the quality of medical images. Uncertainty quantification assesses the impact of radiation dose or contrast agent usage on reconstructed images and can help find the most optimizing imaging condition. Medical image registration involves aligning and transforming multiple images to enable comparison or fusion. Uncertainty estimates help understand the confidence level of the registration process. This is important when the alignment is challenging due to image noise, artifacts, or deformations, especially for multi-modal medical image registration tasks.

In medical diagnosis, using a detection, classification, or segmentation model developed from an imbalanced dataset (which is a common situation in the medical domain) is risky because the model might be overconfident or overconfident. Uncertainty estimation can thus be used to identify where pixel/voxel or object-level predictions are less certain, therefore helping clinicians understand the reliability of the prediction results and identify areas where automatic prediction may fail and manual intervention might be necessary by providing insights into regions of high ambiguity or uncertainty. This can be particularly useful in minimizing false positives and false negatives and detecting Out-of-Distribution or ambiguous In-Distribution samples that might need specialized handling. Apart from disease diagnosis, prediction of treatment outcomes or disease development is also important to improve the cure rate. Uncertainty estimates provide insights into the range of possible outcomes, supporting personalized treatment strategies and allowing researchers to set realistic expectations for model performance.

To conclude, uncertainty quantification provides critical information about the reliability and confidence of the analysis. This information is particularly valuable in medical applications due to the critical nature of the decisions made based on these predictions, impacting patient care and treatment outcomes. By incorporating uncertainty estimation, MIA becomes more transparent, trustworthy, and aligned with the clinical workflow, which helps bridge the gap between artificial intelligence algorithms and clinical practice, enhancing the acceptance and trustworthiness of AI-assisted medical decisions. Furthermore, building public trust will also help to improve the general fairness of AI healthcare systems.

Apart from the methods mentioned above that focus on studying the uncertainty of the medical image analysis results, a branch of literature also focuses on modeling or analyzing the uncertainty of image labels itself. Medical experts may have varied interpretations of the same image, leading to intra-observer variability (Vinod et al., 2016; Jungo et al., 2018b). Additionally, the same expert may interpret an image differently on different occasions, causing inter-observer variability (Sampat et al., 2006; Schmidt et al., 2023). Such discrepancies in annotations introduce uncertainty and complexity in medical image analysis. Therefore, the label uncertainty modeling approaches focus on such datasets, and studying effective methods for modeling and reducing the inter-observer and intra-observer variability is necessary and important. There are some researches that take into account medical image labeling uncertainty, which can be classified according to the focus on inner uncertainty or inter-observer uncertainty modeling, i.e., image label uncertainty modeling and fusion of uncertain image labels. Moreover, there are some researchers who contribute to open-source new datasets with uncertain ground truth. Readers can refer to Supplementary Material B for related analysis.

6.4. Perspectives

Based on the discussion of the advantages and limitations of existing uncertainty quantification methods, we suggest several future research points to further improve the implications of uncertainty quantification in MIA.

Effectiveness. The most critical limitation of present uncertainty quantification research is the lack of ground truth uncertainty, leading to the lack of standardized evaluation metrics for uncertainty quantification methods. The uncertainty associated with ground truth labels can propagate and affect model uncertainty estimates. However, ground truth labels are not always definitive due to inherent inter-observer variability, ambiguous cases, or inherent limitations of manual annotations. Moreover, the lack of the uncertainty ground truth limits the understanding of sources and reasons behind uncertainty and the explanation of uncertainty to clinicians or users. Though some researchers use inter-rater variability as uncertainty ground truth, it is still unclear whether it is theoretically guaranteed. For example, for a segmentation task, experts can somewhat give random variations around the boundaries of the target object, over-segment, or alternatively under-segment the same object of interest based on their annotation style. This inter-rater variability is thus instead linked to contextual biases (e.g., radiologist experience or annotation habits) rather than to the true uncertainty of the label (Mehta et al., 2022). Therefore, we encourage researchers to put efforts into constructing MIA datasets with both accuracy and uncertain ground truth and set up standardized evaluation metrics for uncertainty quantification methods. A simple solution can be providing diagnosis/detection/prediction/segmentation/classification ground truths as well as providing a corresponding confidence index.

Explainability. SOTA uncertainty quantification methods, such as deep learning ensembles or MCD, may lack interpretability, making it challenging to explain the uncertainty estimation process to clinicians or patients. Therefore, the link between explainability and

uncertainty would be interesting to study. Studying the relationship allows us to understand both how the prediction is made and whether or not it should be trusted, in other words, whether or not the results are reliable. An interesting research point would be to complement uncertainty estimates with explanations, helping the user understand the uncertainty of each source and how the uncertain sources are summarized and summed to reach a final decision. For example, in (Huang et al., 2023a), Huang et al. proposed a deep evidential fusion framework with uncertainty quantification and contextual discounting for multimodal medical image segmentation. This approach is the first attempt to explain the decision-making process by quantifying subject-level uncertainty with contextual discounting to the fusion of deep neural networks and applying it to multimodal medical image segmentation tasks. Another potential research work is studying the relationship between uncertainty and reliability. Conventional research typically treats uncertainty as an opposite indicator of reliability, (Modarres et al., 2016; Ovadia et al., 2019), i.e., the lower the uncertainty, the higher the reliability, which is just an approximation and has limitations in explaining more complex situations such as uncertain but reliable models. Therefore, integrating uncertainty with reliability, i.e., studying the relationship between uncertainty and reliability, could also be an exciting and significant subject.

Efficiency. As shown, the vast majority of the implemented uncertainty quantification methods are based on a sampling protocol, such as MCD and Bayesian inference, aiming at generating multiple predictions. However, they can be computationally expensive and time-consuming, which, therefore, limits their practical application in real-time or clinical settings, where quick and efficient analysis is crucial. The recently popular deep ensemble models, their superior uncertainty measure, along with the high computational cost. Non-probabilistic methods, such as DST, compute the uncertainty in a quick and efficient manner that requires only a single forward step, which is generally required for medical applications, indicating a promising direction to be further explored.

Clinical applications. Integrating uncertainty quantification into clinical workflows and decision-making processes can be challenging due to the limited trust in existing ML models and the limited clinical validation. Therefore, careful consideration and adaptation of uncertainty quantification are required to align research with clinical guidelines and to fit it within the clinical context. We thus suggest researchers integrate clinical validation and take ethical and legal problems into consideration when developing their MIA models to 1) enable more reliable, interpretable, and applicable uncertainty quantification models; 2) ensure their clinical utility, interpretability, and impact on patient outcomes; 3) ensure their fairness to the public.

7. Conclusion

This review provides an overview of the uncertainty quantification methods commonly implemented in machine learning-based medical image applications. Numerous phenomena can cause predictive uncertainty, such as noisy images, imperfect ground truth labels,

incomplete data, and inter-site image variability. The literature proposes various methods to quantify uncertainty applied to an extensive range of medical image applications. As demonstrated in this review, developing trustable AI solutions integrating uncertainty quantification of the computed predictions is an active search topic that has many potential future directions.

Acknowledgments

This research is supported by A*STAR, CISCO Systems (USA) Pte. Ltd, and National University of Singapore under its Cisco-NUS Accelerated Digital Economy Corporate Laboratory (Award I21001E0002) and the NMRC Health Service Research Grant (MOH-000030-00).

References

Abdar, M., Fahami, M.A., Chakrabarti, S., Khosravi, A., Pławiak, P., Acharya, U.R., Tadeusiewicz, R., Nahavandi, S., 2021a. Barf: A new direct and cross-based binary residual feature fusion with uncertainty-aware module for medical image classification. Information Sciences 577, 353–378.

Abdar, M., Fahami, M.A., Rundo, L., Radeva, P., Frangi, A.F., Acharya, U.R., Khosravi, A., Lam, H.K., Jung, A., Nahavandi, S., 2022. Hercules: Deep hierarchical attentive multilevel fusion model with uncertainty quantification for medical image classification. IEEE Transactions on Industrial Informatics 19, 274–285.

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., et al., 2021b. 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., Makarenkov, V., Nahavandi, S., 2023. Uncertaintyfusenet: robust uncertainty-aware hierarchical feature fusion model with ensemble monte carlo dropout for covid-19 detection. Information Fusion 90, 364–381.

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.

Adiga Vasudeva, S., Dolz, J., Lombaert, H., 2022. Leveraging labeling representations in uncertainty-based semi-supervised segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 265–275.

Adler, T.J., Ardizzone, L., Vemuri, A., Ayala, L., Gröhl, J., Kirchner, T., Wirkert, S., Kruse, J., Rother, C., Köthe, U., et al., 2019. Uncertainty-aware performance assessment of optical imaging modalities with invertible neural networks. International journal of computer assisted radiology and surgery 14, 997–1007.

Ahmad, J., Saudagar, A.K.J., Malik, K.M., Khan, M.B., AlTameem, A., Alkhathami, M., Hasanat, M.H.A., 2023. Prognosis prediction in covid-19 patients through deep feature space reasoning. Diagnostics 13, 1387.

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 60, 2797–2811.

Akrami, H., Joshi, A., Aydore, S., Leahy, R., 2021. Quantile regression for uncertainty estimation in vaes with applications to brain lesion detection, in: Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings 27, Springer. pp. 689–700.

Alberts, E., Rempfler, M., Alber, G., Huber, T., Kirschke, J., Zimmer, C., Menze, B.H., 2016. Uncertainty quantification in brain tumor segmentation using crfs and random perturbation models, in: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 428–431.

Alizadehsani, R., Roshanzamir, M., Hussain, S., Khosravi, A., Koohestani, A., Zangooei, M.H., Abdar, M., Beykikhoshk, A., Shoeibi, A., Zare, A., et al., 2021. Handling of uncertainty in medical data using machine learning and probability theory techniques: A review of 30 years (1991–2020). Annals of Operations Research , 1–42.

Aljuhani, A., Casukhela, I., Chan, J., Liebner, D., Machiraju, R., 2022. Uncertainty aware sampling framework of weak-label learning for histology image classification, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 366–376.

Alonso-Caneiro, D., Kugelman, J., Tong, J., Kalloniatis, M., Chen, F.K., Read, S.A., Collins, M.J., 2021. Use of uncertainty quantification as a surrogate for layer segmentation error in stargardt disease retinal oct images, in: 2021 Digital Image Computing: Techniques and Applications (DICTA), IEEE. pp. 1–8.

Ammari, A., Mahmoudi, R., Hmida, B., Saouli, R., Bedoui, M.H., 2023. Deep-active-learning approach towards accurate right ventricular segmentation using a two-level uncertainty estimation. Computerized Medical Imaging and Graphics 104, 102168.

Araujo, T., Aresta, G., Mendonça, L., Penas, S., Maia, C., Carneiro, Â., Mendonça, A.M., Campilho, A., 2020. Dr— graduate: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images. Medical Image Analysis 63, 101715.

Arco, J.E., Ortiz, A., Ramirez, J., Martinez-Murcia, F.J., Zhang, Y.D., Gorriz, J.M., 2023. Uncertainty-driven ensembles of multi-scale deep architectures for image classification. Information Fusion 89, 53–65.

Arega, T.W., Bricq, S., Legrand, F., Jacquier, A., Lalande, A., Meriaudeau, F., 2023. Automatic uncertainty-based quality controlled t1 mapping and ecv analysis from native and post-contrast cardiac t1 mapping images using bayesian vision transformer. Medical image analysis 86, 102773.

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, 815.

Awate, S.P., Garg, S., Jena, R., 2019. Estimating uncertainty in mrf-based image segmentation: A perfect-mcmc approach. Medical image analysis 55, 181–196.

Ayhan, M.S., Berens, P., 2018. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks.(2018). URL https://openreview. net/pdf .

Ayhan, M.S., Berens, P., 2022. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks, in: 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.

Barbano, R., Zhang, C., Arridge, S., Jin, B., 2021. Quantifying model uncertainty in inverse problems via bayesian deep gradient descent, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE. pp. 1392–1399.

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, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22, Springer. pp. 119–127.

Behnami, D., Liao, Z., Girgis, H., Luong, C., Rohling, R., Gin, K., Tsang, T., Abolmaesumi, P., 2019. Dual-view joint estimation of left ventricular ejection fraction with uncertainty modelling in echocardiograms, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22, Springer. pp. 696–704.

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.

Ben Atitallah, S., Driss, M., Boulila, W., Koubaa, A., Ben Ghezala, H., 2022. Fusion of convolutional neural networks based on dempster–shafer theory for automatic pneumonia detection from chest x-ray images. International Journal of Imaging Systems and Technology 32, 658–672.

Bertels, J., Robben, D., Vandermeulen, D., Suetens, P., 2021. Theoretical analysis and experimental validation of volume bias of soft dice optimized segmentation maps in the context of inherent uncertainty. Medical Image Analysis 67, 101833.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A., 2019. Mixmatch: A holistic approach to semi-supervised learning. Advances in neural information processing systems 32.

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

Bian, C., Yuan, C., Wang, J., Li, M., Yang, X., Yu, S., Ma, K., Yuan, J., Zheng, Y., 2020. Uncertainty-aware domain alignment for anatomical structure segmentation. Medical Image Analysis 64, 101732.

Bliesener, Y., Acharya, J., Nayak, K.S., 2019. Efficient dce-mri parameter and uncertainty estimation using a neural network. IEEE transactions on medical imaging 39, 1712–1723.

Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D., 2015. Weight uncertainty in neural network, in: International conference on machine learning, PMLR. pp. 1613–1622.

Brier, G.W., 1950. Verification of forecasts expressed in terms of probability. Monthly weather review 78, 1–3.

Brooks, S., 1998. Markov chain monte carlo method and its application. Journal of the royal statistical society: series D (the Statistician) 47, 69–100.

Buddenkotte, T., Sanchez, L.E., Crispin-Ortuzar, M., Woitek, R., McCague, C., Brenton, J.D., Öktem, O., Sala, E., Rundo, L., 2023. Calibrating ensembles for scalable uncertainty quantification in deep learning-based medical image segmentation. Computers in Biology and Medicine , 107096.

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

Camarasa, R., Bos, D., Hendrikse, J., Nederkoorn, P.J., Kooi, E., van der Lugt, A., de Bruijne, M., 2021. A quantitative comparison of epistemic uncertainty maps applied to multi-class segmentation. The Journal of Machine Learning for Biomedical Imaging 13, 1–39.

Cao, X., Chen, H., Li, Y., Peng, Y., Wang, S., Cheng, L., 2020. Uncertainty aware temporal-ensembling model for semi-supervised abus mass segmentation. IEEE transactions on medical imaging 40, 431–443.

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.

Chen, J., Fu, C., Xie, H., Zheng, X., Geng, R., Sham, C.W., 2022. Uncertainty teacher with dense focal loss for semi-supervised medical image segmentation. Computers in Biology and Medicine 149, 106034.

Chib, S., Greenberg, E., 1995. Understanding the metropolis-hastings algorithm. The american statistician 49, 327–335.

Chlebus, G., Schenk, A., Hahn, H.K., Van Ginneken, B., Meine, H., 2022. Robust segmentation models using an uncertainty slice sampling-based annotation workflow. IEEE Access 10, 4728–4738.

Christophe, A., Freitas, N.D., Doucet, A., , Jordan., M.I., 2023. An introduction to mcmc for machine learning. Machine learning 50, 5–43. doi:https://doi.org/10.1016/j.inffus.2022.11.008.

Cifci, M.A., 2023. A deep learning-based framework for uncertainty quantification in medical imaging using the dropweak technique: An empirical study with baresnet. Diagnostics 13, 800.

Corrado, C., Razeghi, O., Roney, C., Coveney, S., Williams, S., Sim, I., O'Neill, M., Wilkinson, R., Oakley, J., Clayton, R.H., et al., 2020. Quantifying atrial anatomy uncertainty from clinical data and its impact on electro-physiology simulation predictions. Medical Image Analysis 61, 101626.

Corrado, C., Roney, C.H., Razeghi, O., Lemus, J.A.S., Coveney, S., Sim, I., Williams, S.E., O'Neill, M.D., Wilkinson, R.D., Clayton, R.H., et al., 2023. Quantifying the impact of shape uncertainty on predicted arrhythmias. Computers in Biology and Medicine 153, 106528.

Cui, J., Xie, Y., Joshi, A.A., Gong, K., Kim, K., Son, Y.D., Kim, J.H., Leahy, R., Liu, H., Li, Q., 2022. Pet denoising and uncertainty estimation based on nvae model using quantile regression loss, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 173–183.

Czolbe, S., Arnavaz, K., Krause, O., Feragen, A., 2021. Is segmentation uncertainty useful?, in: Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings 27, Springer. pp. 715–726.

Dai, J., Tian, H., 2013. Entropy measures and granularity measures for set-valued information systems. Information Sciences 240, 72–82.

Davison, A.C., Hinkley, D.V., 1997. Bootstrap methods and their application. 1, Cambridge university press.

Dawood, T., Chen, C., Sidhu, B.S., Ruijsink, B., Gould, J., Porter, B., Elliott, M.K., Mehta, V., Rinaldi, C.A., Puyol-Antón, E., et al., 2023. Uncertainty aware training to improve deep learning model calibration for classification of cardiac mr images. Medical Image Analysis , 102861.

Del Amor, R., Silva-Rodríguez, J., Naranjo, V., 2023. Labeling confidence for uncertainty-aware histology image classification. Computerized Medical Imaging and Graphics 107, 102231.

Dempster, A.P., 1967. Upper and lower probability inferences based on a sample from a finite univariate population. Biometrika 54, 515–528.

Denoeux, T., 1995. A k-nearest neighbor classification rule based on dempster-shafer theory. IEEE transactions on systems, man, and cybernetics 25, 804–813.

Denœux, T., 2000. A neural network classifier based on Dempster-Shafer theory. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 30, 131–150.

Denœux, T., Masson, M.H., 2004. Evclus: evidential clustering of proximity data. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 34, 95–109.

Der Kiureghian, A., Ditlevsen, O., 2009. Aleatory or epistemic? does it matter? Structural safety 31, 105–112.

Derraz, F., Pinti, A., Peyrodie, L., Bousahla, M., Toumi, H., 2015. Joint variational segmentation of ct/pet data using non-local active contours and belief functions. Pattern Recognition and Image Analysis 25, 407–412.

DeVries, T., Taylor, G.W., 2018. Leveraging uncertainty estimates for predicting segmentation quality. arXiv preprint arXiv:1807.00502 .

Dhamala, J., Arevalo, H.J., Sapp, J., Horácek, B.M., Wu, K.C., Trayanova, N.A., Wang, L., 2018. Quantifying the uncertainty in model parameters using gaussian process-based markov chain monte carlo in cardiac electrophysiology. Medical image analysis 48, 43–57.

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.

Dietterich, T.G., 2000. Ensemble methods in machine learning, in: International workshop on multiple classifier systems, Springer. pp. 1–15.

Dodge, Y., Cox, D., Commenges, D., 2003. The Oxford dictionary of statistical terms. Oxford University Press on Demand.

Dolezal, J.M., Srisuwananukorn, A., Karpeyev, D., Ramesh, S., Kochanny, S., Cody, B., Mansfield, A.S., Rakshit, S., Bansal, R., Bois, M.C., et al., 2022. Uncertainty-informed deep learning models enable high-confidence predictions for digital histopathology. Nature communications 13, 6572.

Dong, S., Yang, Q., Fu, Y., Tian, M., Zhuo, C., 2021. Rconet: Deformable mutual information maximization and high-order uncertainty-aware learning for robust covid-19 detection. IEEE Transactions on Neural Networks and Learning Systems 32, 3401–3411.

Dubois, D.J., 1980. Fuzzy sets and systems: theory and applications. volume 144. Academic press.

Duchateau, N., De Craene, M., Allain, P., Saloux, E., Sermesant, M., 2016. Infarct localization from myocardial deformation: prediction and uncertainty quantification by regression from a low-dimensional space. IEEE transactions on medical imaging 35, 2340–2352.

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, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I, Springer. pp. 691–699.

Eaton-Rosen, Z., Varsavsky, T., Ourselin, S., Cardoso, M.J., 2019. As easy as 1, 2... 4? uncertainty in counting tasks for medical imaging, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22, Springer. pp. 356–364.

Ebadi, N., Li, R., Das, A., Roy, A., Nikos, P., Najafirad, P., 2023. Cbct-guided adaptive radiotherapy using self-supervised sequential domain adaptation with uncertainty estimation. Medical Image Analysis 86, 102800.

Edupuganti, V., Mardani, M., Vasanawala, S., Pauly, J., 2020. Uncertainty quantification in deep mri reconstruction. IEEE Transactions on Medical Imaging 40, 239–250.

Efron, B., 1992. Bootstrap methods: another look at the jackknife, in: Breakthroughs in statistics: Methodology and distribution. Springer, pp. 569–593.

Farooq, M.U., Ullah, Z., Gwak, J., 2023. Residual attention based uncertainty-guided mean teacher model for semi-supervised breast masses segmentation in 2d ultrasonography. Computerized Medical Imaging and Graphics 104, 102173.

Feng, D., Chen, X., Wang, X., Lv, J., Bai, L., Zhang, S., Zhou, Z., 2022. Penalized entropy: a novel loss function for uncertainty estimation and optimization in medical image classification, in: 2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS), IEEE. pp. 306–310.

Feng, X., Dou, Q., Tustison, N., Meyer, C., 2020. Brain tumor segmentation with uncertainty estimation and overall survival prediction, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part I 5, Springer. pp. 304–314.

Fidon, L., Aertsen, M., Kofler, F., Bink, A., David, A.L., Deprest, T., Emam, D., Guffens, F., Jakab, A., Kasprian, G., et al., 2022. A dempster-shafer approach to trustworthy ai with application to fetal brain mri segmentation. arXiv preprint arXiv:2204.02779 .

Fisher, R.A., 1919. Xv.—the correlation between relatives on the supposition of mendelian inheritance. Earth and Environmental Science Transactions of the Royal Society of Edinburgh 52, 399–433.

Fuchs, M., Gonzalez, C., Mukhopadhyay, A., 2021. Practical uncertainty quantification for brain tumor segmentation, in: Medical Imaging with Deep Learning.

Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: international conference on machine learning, PMLR. pp. 1050–1059.

Galdran, A., Meyer, M., Costa, P., Campilho, A., et al., 2019. Uncertainty-aware artery/vein classification on retinal images, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE. pp. 556–560.

Ganaie, M.A., Hu, M., Malik, A., Tanveer, M., Suganthan, P., 2022. Ensemble deep learning: A review. Engineering Applications of Artificial Intelligence 115, 105151.

Ghasemi, J., Ghaderi, R., Mollaei, M.K., Hojjatoleslami, S., 2013. A novel fuzzy dempster–shafer inference system for brain mri segmentation. Information Sciences 223, 205–220.

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, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22, Springer. pp. 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.

Ghoshal, B., Ghoshal, B., Tucker, A., 2022. Leveraging uncertainty in deep learning for pancreatic adenocarcinoma grading, in: Annual Conference on Medical Image Understanding and Analysis, Springer. pp. 565–577.

Ghoshal, B., Tucker, A., 2020. Estimating uncertainty and interpretability in deep learning for coronavirus (covid-19) detection. arXiv preprint arXiv:2003.10769 .

Ghoshal, B., Tucker, A., 2021. On cost-sensitive calibrated uncertainty in deep learning: An application on covid-19 detection, in: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS), IEEE. pp. 503–509.

Ghoshal, B., Tucker, A., 2022. On calibrated model uncertainty in deep learning. arXiv preprint arXiv:2206.07795 .

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.

Ghoshal, B., Tucker, A., Sanghera, B., Wong, W.L., 2019. Estimating uncertainty in deep learning for reporting confidence to clinicians when segmenting nuclei image data, in: 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), IEEE. pp. 318–324.

Gilks, W.R., Richardson, S., Spiegelhalter, D., 1995. Markov chain Monte Carlo in practice. CRC press.

Gill, R.S., Caldairou, B., Bernasconi, N., Bernasconi, A., 2019. Uncertainty-informed detection of epileptogenic brain malformations using bayesian neural networks, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22, Springer. pp. 225–233.

Gillmann, C., Saur, D., Wischgoll, T., Hoffman, K.T., Hagen, H., Maciejewski, R., Scheuermann, G., 2020. Uncertainty-aware brain lesion visualization, in: Eurographics Workshop on Visual Computing for Biology and Medicine, p. 97.

Gillmann, C., Saur, D., Wischgoll, T., Scheuermann, G., 2021. Uncertainty-aware visualization in medical imaging-a survey, in: Computer Graphics Forum, Wiley Online Library. pp. 665–689.

Gomes, J., Kong, J., Kurc, T., Melo, A.C., Ferreira, R., Saltz, J.H., Teodoro, G., 2021. Building robust pathology image analyses with uncertainty quantification. Computer Methods and Programs in Biomedicine 208, 106291.

Gong, X., Khaidem, L., Zhu, W., Zhang, B., Doermann, D., 2022. Uncertainty learning towards unsupervised deformable medical image registration, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2484–2493.

Gour, M., Jain, S., 2022. Uncertainty-aware convolutional neural network for covid-19 x-ray images classification. Computers in biology and medicine 140, 105047.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks, in: International conference on machine learning, PMLR. pp. 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 81, 102532.

Gürsel, G., 2016. Healthcare, uncertainty, and fuzzy logic. Digital Medicine 2, 101–112.

Hájek, P., 2013. Metamathematics of fuzzy logic. volume 4. Springer Science & Business Media.

Hamedani, K.F., Tavakkoli-Moghaddam, R., Tajally, A.R., Aria, S.S., 2023. Breast cancer classification by a new approach to assessing deep neural network-based uncertainty quantification methods. Biomedical Signal Processing and Control 79, 104057.

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, in: 2021 Computing in Cardiology (CinC), IEEE. pp. 1–4.

Hemsley, M., Chugh, B., Ruschin, M., Lee, Y., Tseng, C.L., Stanisz, G., Lau, A., 2020. Deep generative model for synthetic-ct generation with uncertainty predictions, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8,

2020, Proceedings, Part I 23, Springer. pp. 834–844.
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.
Hinton, G.E., Van Camp, D., 1993. Keeping the neural networks simple by minimizing the description length of the weights, in: Proceedings of the sixth annual conference on Computational learning theory, pp. 5–13.
Hora, S.C., 1996. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. Reliability Engineering & System Safety 54, 217–223.
Hosmer, D.W., Lemeshow, S., 1992. Confidence interval estimation of interaction. Epidemiology , 452–456.
Hu, C., Xia, T., Cui, Y., Zou, Q., Wang, Y., Xiao, W., Ju, S., Li, X., 2023. Trustworthy multi-phase liver tumor segmentation via evidence-based uncertainty. arXiv preprint arXiv:2305.05344 .
Hu, L., Li, J., Peng, X., Xiao, J., Zhan, B., Zu, C., Wu, X., Zhou, J., Wang, Y., 2022. Semi-supervised npc segmentation with uncertainty and attention guided consistency. Knowledge-Based Systems 239, 108021.
Hu, S., Worrall, D., Knegt, S., Veeling, B., Huisman, H., Welling, M., 2019. Supervised uncertainty quantification for segmentation with multiple annotations, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22, Springer. pp. 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, C., Liu, C., Zhang, Z., Wu, Z., Wen, J., Jiang, Q., Xu, Y., 2022a. Pixel-level anomaly detection via uncertainty-aware prototypical transformer, in: Proceedings of the 30th ACM International Conference on Multimedia, pp. 521–530.
Huang, K., Zhang, Y., Cheng, H.D., Xing, P., 2022b. Trustworthy breast ultrasound image semantic segmentation based on fuzzy uncertainty reduction, in: Healthcare, MDPI. p. 2480.
Huang, L., Denoeux, T., Vera, P., Ruan, S., 2022c. Evidence fusion with contextual discounting for multi-modality medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 401–411.
Huang, L., Ruan, S., Decazes, P., Denoeux, T., 2021a. Evidential segmentation of 3d pet/ct images, in: International Conference on Belief Functions, Springer. pp. 159–167.
Huang, L., Ruan, S., Decazes, P., Denœux, T., 2022d. Lymphoma segmentation from 3d pet-ct images using a deep evidential network. International Journal of Approximate Reasoning 149, 39–60.
Huang, L., Ruan, S., Denoeux, T., 2021b. Belief function-based semi-supervised learning for brain tumor segmentation, in: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 160–164.
Huang, L., Ruan, S., Denoeux, T., 2021c. Covid-19 classification with deep neural network and belief functions, in: The Fifth International Conference on Biological Information and Biomedical Engineering, pp. 1–4.
Huang, L., Ruan, S., Denœux, T., 2023a. Application of belief functions to medical image segmentation: A review. Information Fusion 91, 737–756. URL: https://www.sciencedirect.com/science/article/pii/S1566253522002184, doi:https://doi.org/10.1016/j.inffus.2022.11.008.
Huang, L., Ruan, S., Denœux, T., 2023b. Semi-supervised multiple evidence fusion for brain tumor segmentation. Neurocomputing 535, 40–52.
Huang, Y., Chung, A.C., 2020. Edge-variational graph convolutional networks for uncertainty-aware disease prediction, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VII 23, Springer. pp. 562–572.
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, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23, Springer. pp. 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.

Hwang, J.G., Ding, A.A., 1997. Prediction intervals for artificial neural networks. Journal of the American Statistical Association 92, 748–757.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al., 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, in: Proceedings of the AAAI conference on artificial intelligence, pp. 590–597.

Isensee, F., Petersen, J., Klein, ., Zimmerer, D., 2018. nnu-net: Self-adapting framework for u-net-based medical image segmentation. arXiv preprint arXiv:1809.10486 .

Islam, M., Glocker, B., 2021. Spatially varying label smoothing: Capturing uncertainty from expert annotations, in: Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings 27, Springer. pp. 677–688.

Islam, M., Seenivasan, L., Sharan, S., Viekash, V., Gupta, B., Glocker, B., Ren, H., 2023. Paced-curriculum distillation with prediction and label uncertainty for image segmentation. International Journal of Computer Assisted Radiology and Surgery , 1–9.

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, in: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis: 3rd International Workshop, UNSURE 2021, and 6th International Workshop, PIPPI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 3, Springer. pp. 34–43.

Jafari, M.H., Luong, C., Tsang, M., Gu, A.N., Van Woudenberg, N., Rohling, R., Tsang, T., Abolmaesumi, P., 2021. U-land: uncertainty-driven video landmark detection. IEEE Transactions on Medical Imaging 41, 793–804.

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.

Jena, R., Awate, S.P., 2019. A bayesian neural net to segment images with uncertainty estimates and good calibration, in: International Conference on Information Processing in Medical Imaging, Springer. pp. 3–15.

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, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22, Springer. pp. 540–548.

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.

Jones, C.K., Wang, G., Yedavalli, V., Sair, H., 2022. Direct quantification of epistemic and aleatoric uncertainty in 3d u-net segmentation. Journal of Medical Imaging 9, 034002–034002.

Jøsang, A., 2016. Subjective logic. volume 3. Springer.

Josang, A., Hayward, R., Pope, S., 2006. Trust network analysis with subjective logic, in: Conference Proceedings of the Twenty-Ninth Australasian Computer Science Conference (ACSW 2006), Australian Computer Society. pp. 85–94.

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 transactions on 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, in: International Conference on Medical Image Computing

and Computer-Assisted Intervention, Springer. pp. 492–502.
Jungo, A., Balsiger, F., Reyes, M., 2020. Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. Frontiers in neuroscience 14, 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., 2018a. Towards uncertainty-assisted brain tumor segmentation and survival prediction, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers 3, Springer. pp. 474–485.
Jungo, A., Meier, R., Ermis, E., Blatti-Moreno, M., Herrmann, E., Wiest, R., Reyes, M., 2018b. On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I, Springer. pp. 682–690.
Jungo, A., Reyes, M., 2019. Assessing reliability and challenges of uncertainty estimations for medical image segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22, Springer. pp. 48–56.
Kabir, H.D., Khanam, S., Khozeimeh, F., Khosravi, A., Mondal, S.K., Nahavandi, S., Acharya, U.R., 2022. Aleatory-aware deep uncertainty quantification for transfer learning. Computers in Biology and Medicine 143, 105246.
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.
Kannan, A., Hodgson, A., Mulpuri, K., Garbi, R., 2021. Leveraging voxel-wise segmentation uncertainty to improve reliability in assessment of paediatric dysplasia of the hip. International Journal of Computer Assisted Radiology and Surgery 16, 1121–1129.
Kendall, A., Gal, Y., 2017. What uncertainties do we need in bayesian deep learning for computer vision? Advances in neural information processing systems 30.
Khawaled, S., Freiman, M., 2022. Npbdreg: Uncertainty assessment in diffeomorphic brain mri registration using a non-parametric bayesian deep-learning based approach. Computerized Medical Imaging and Graphics 99, 102087.
Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J.R., Maier-Hein, K., Eslami, S., Jimenez Rezende, D., Ronneberger, O., 2018. A probabilistic u-net for segmentation of ambiguous images. Advances in neural information processing systems 31.
Kohli, P., Torr, P.H., 2008. Measuring uncertainty in graph cut solutions. Computer Vision and Image Understanding 112, 30–38.
Kotz, S., Balakrishnan, N., Johnson, N.L., 2004. Continuous multivariate distributions, Volume 1: Models and applications. volume 1. John Wiley & Sons.
Kozumi, H., Kobayashi, G., 2011. Gibbs sampling methods for bayesian quantile regression. Journal of statistical computation and simulation 81, 1565–1578.
Kroese, D.P., Brereton, T., Taimre, T., Botev, Z.I., 2014. Why the monte carlo method is so important today. Wiley Interdisciplinary Reviews: Computational Statistics 6, 386–392.
Krygier, M.C., LaBonte, T., Martinez, C., Norris, C., Sharma, K., Collins, L.N., Mukherjee, P.P., Roberts, S.A., 2021. Quantifying the unknown impact of segmentation uncertainty on image-based simulations. Nature communications 12, 5414.
Kuang, Z., Yan, Z., Yu, L., Deng, X., Hua, Y., Li, S., 2022. Uncertainty-aware deep learning with cross-task supervision for phe segmentation on ct images. IEEE Journal of Biomedical and Health Informatics 26, 2615–2626.
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, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 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.
Laves, M.H., Ihler, S., Fast, J.F., Kahrs, L.A., Ortmaier, T., 2020a. Well-calibrated regression uncertainty in medical imaging with deep learning, in: Medical Imaging with Deep Learning, PMLR. pp. 393–412.
Laves, M.H., Ihler, S., Fast, J.F., Kahrs, L.A., Ortmaier, T., 2021. Recalibration of aleatoric and epistemic regression uncertainty in medical imaging. arXiv preprint arXiv:2104.12376 .
Laves, M.H., Tölle, M., Ortmaier, T., 2020b. Uncertainty estimation in medical image denoising with bayesian deep image prior, in: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis: Second International Workshop, UNSURE 2020, and Third International Workshop, GRAIL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings 2, Springer. pp. 81–96.
Lê, M., Unkelbach, J., Ayache, N., Delingette, H., 2016. Sampling image segmentations for uncertainty quantification. Medical image analysis 34, 42–51.
Le Folgoc, L., Delingette, H., Criminisi, A., Ayache, N., 2016. Quantifying registration uncertainty with sparse bayesian modelling. IEEE transactions on medical imaging 36, 607–617.
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, 17816.
Lelandais, B., Ruan, S., Denœux, T., Vera, P., Gardin, I., 2014. Fusion of multi-tracer pet images for dose painting. Medical image analysis 18, 1247–1259.
Lemay, A., Gros, C., Karthik, E.N., Cohen-Adad, J., 2022. Label fusion and training methods for reliable representation of inter-rater uncertainty. arXiv preprint arXiv:2202.07550 .
Li, H., Nan, Y., Del Ser, J., Yang, G., 2022a. Region-based evidential deep learning to quantify uncertainty and improve robustness of brain tumor segmentation. Neural Computing and Applications , 1–15.
Li, X., Liang, X., Luo, G., Wang, W., Wang, K., Li, S., 2022b. Ultra: Uncertainty-aware label distribution learning for breast tumor cellularity assessment, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 303–312.
Li, X., Luo, G., Wang, W., Wang, K., Gao, Y., Li, S., 2021a. Hematoma expansion context guided intracranial hemorrhage segmentation and uncertainty estimation. IEEE Journal of Biomedical and Health Informatics 26, 1140–1151.
Li, X., Wei, Y., Hu, Q., Wang, C., Yang, J., 2022c. Learning to segment subcortical structures from noisy annotations with a novel uncertainty-reliability aware learning framework. Computers in Biology and Medicine 151, 106326.
Li, X., Yu, L., Chen, H., Fu, C.W., Heng, P.A., 2019. Transformation consistent self-ensembling model for semi-supervised medical image segmentation. arXiv preprint arXiv:1903.00348 .
Li, Y., Chen, X., Quan, L., Zhang, N., 2021b. Uncertainty-guided robust training for medical image segmentation, in: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 1471–1475.
Li, Y., Luo, L., Lin, H., Chen, H., Heng, P.A., 2021c. Dual-consistency semi-supervised learning with uncertainty quantification for covid-19 lesion segmentation from ct images, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24, Springer. pp. 199–209.
Lian, C., Ruan, S., Denœux, T., Guo, Y., Vera, P., 2017a. Accurate tumor segmentation in fdg-pet images with guidance of complementary ct images, in: 2017 IEEE International Conference on Image Processing (ICIP), IEEE. pp. 4447–4451.
Lian, C., Ruan, S., Denœux, T., Jardin, F., Vera, P., 2016a. Selecting radiomic features from fdg-pet images for cancer treatment outcome prediction. Medical image analysis 32, 257–268.
Lian, C., Ruan, S., Denœux, T., Li, H., Vera, P., 2016b. Robust cancer treatment outcome prediction dealing with small-sized and imbalanced data from fdg-pet images, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19, Springer. pp. 61–69.

Lian, C., Ruan, S., Denoeux, T., Li, H., Vera, P., 2017b. Spatial evidential clustering with adaptive distance metric for tumor segmentation in fdg-pet images. IEEE Transactions on Biomedical Engineering 65, 21–30.

Lian, C., Ruan, S., Denoeux, T., Li, H., Vera, P., 2017c. Tumor delineation in fdg-pet images using a new evidential clustering algorithm with spatial regularization and adaptive distance metric, in: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), IEEE. pp. 1177–1180.

Lian, C., Ruan, S., Denœux, T., Li, H., Vera, P., 2018. Joint tumor segmentation in pet-ct images using co-clustering and fusion based on belief functions. IEEE Transactions on Image Processing 28, 755–766.

Liao, Z., Girgis, H., Abdi, A., Vaseli, H., Hetherington, J., Rohling, R., Gin, K., Tsang, T., Abolmaesumi, P., 2019. On modelling label uncertainty in deep neural networks: Automatic estimation of intra-observer variability in 2d echocardiography quality assessment. IEEE transactions on medical imaging 39, 1868–1883.

Lima, S.A., Islam, M.R., 2019. A modified method for brain mri segmentation using dempster-shafer theory, in: 2019 22nd International Conference on Computer and Information Technology (ICCIT), IEEE. pp. 1–6.

Lin, Q., Chen, X., Chen, C., Garibaldi, J.M., 2022a. A novel quality control algorithm for medical image segmentation based on fuzzy uncertainty. IEEE Transactions on Fuzzy Systems .

Lin, Q., Chen, X., Chen, C., Garibaldi, J.M., 2022b. Quality quantification in deep convolutional neural networks for skin lesion segmentation using fuzzy uncertainty measurement, in: 2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE. pp. 1–8.

Linmans, J., Elfwing, S., van der Laak, J., Litjens, G., 2023. Predictive uncertainty estimation for out-of-distribution detection in digital pathology. Medical Image Analysis 83, 102655.

Liu, J., Lin, Z., Padhy, S., Tran, D., Bedrax Weiss, T., Lakshminarayanan, B., 2020a. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. Advances in Neural Information Processing Systems 33, 7498–7512.

Liu, J., Lu, X., Li, Y., Chen, X., Deng, Y., 2015. A new method based on dempster–shafer theory and fuzzy c-means for brain mri segmentation. Measurement Science and Technology 26, 105402.

Liu, P., Zheng, G., 2022. Handling imbalanced data: Uncertainty-guided virtual adversarial training with batch nuclear-norm optimization for semi-supervised medical image classification. IEEE Journal of Biomedical and Health Informatics 26, 2983–2994.

Liu, S., Wang, H., Li, Y., Li, X., Cao, G., Cao, W., 2022. Ahu-multinet: Adaptive loss balancing based on homoscedastic uncertainty in multi-task medical image segmentation network. Computers in Biology and Medicine 150, 106157.

Liu, X., Wang, J., Zhou, S.K., Engstrom, C., Chandra, S.S., 2023a. Evidence-aware multi-modal data fusion and its application to total knee replacement prediction. arXiv preprint arXiv:2303.13810 .

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

Liu, Z., Lin, F., Huang, J., Wu, X., Wen, J., Wang, M., Ren, Y., Wei, X., Song, X., Qin, J., et al., 2023b. A classifier-combined method for grading breast cancer based on dempster-shafer evidence theory. Quantitative Imaging in Medicine and Surgery 13, 3288.

Lu, L., Yin, M., Fu, L., Yang, F., 2023. Uncertainty-aware pseudo-label and consistency for semi-supervised medical image segmentation. Biomedical Signal Processing and Control 79, 104203.

Luo, G., Blumenthal, M., Heide, M., Uecker, M., 2023. Bayesian mri reconstruction with joint uncertainty estimation using diffusion models. Magnetic Resonance in Medicine 90, 295–311.

Luo, X., Liao, W., Chen, J., Song, T., Chen, Y., Zhang, S., Chen, N., Wang, G., Zhang, S., 2021. Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24, Springer. pp. 318–329.

Luo, X., Wang, G., Liao, W., Chen, J., Song, T., Chen, Y., Zhang, S., Metaxas, D.N., Zhang, S., 2022.

Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. Medical Image Analysis 80, 102517.

MacKay, D.J., 1992. A practical bayesian framework for backpropagation networks. Neural computation 4, 448–472.

Mahani, G.K., Li, R., Evangelou, N., Sotiropolous, S., Morgan, P.S., French, A.P., Chen, X., 2022. Bounding box based weakly supervised deep convolutional neural network for medical image segmentation using an uncertainty guided and spatially constrained loss, in: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 1–5.

Makni, N., Betrouni, N., Colot, O., 2014. Introducing spatial neighbourhood in evidential c-means for segmentation of multi-source images: Application to prostate multi-parametric mri. Information Fusion 19, 61–72.

Mao, Y., Xue, F.F., Wang, R., Zhang, J., Zheng, W.S., Liu, H., 2020. Abnormality detection in chest x-ray images using uncertainty prediction autoencoders, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23, Springer. pp. 529–538.

Masson, M.H., Denoeux, T., 2008. Ecm: An evidential version of the fuzzy c-means algorithm. Pattern Recognition 41, 1384–1397.

Mazoure, B., Mazoure, A., Bédard, J., Makarenkov, V., 2022. Dunescan: a web server for uncertainty estimation in skin cancer detection with deep neural networks. Scientific Reports 12, 179.

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., Bussy, A., Premasiri, S., Costantino, M., Chakravarthy, M.M., Arnold, D.L., Gal, Y., Arbel, T., 2021. Propagating uncertainty across cascaded medical imaging tasks for improved deep learning inference. IEEE Transactions on Medical Imaging 41, 360–373.

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 quantifying uncertainty in brain tumor segmentation-analysis of ranking scores and benchmarking results. The journal of machine learning for biomedical imaging 2022.

Mehta, R., Shui, C., Arbel, T., 2023. Evaluating the fairness of deep learning uncertainty estimates in medical image analysis. arXiv preprint arXiv:2303.03242 .

Mendel, J.M., 1995. Fuzzy logic systems for engineering: a tutorial. Proceedings of the IEEE 83, 345–377.

Meyer, A., Ghosh, S., Schindele, D., Schostak, M., Stober, S., Hansen, C., Rak, M., 2021. Uncertainty-aware temporal self-learning (uats): Semi-supervised learning for segmentation of prostate zones and beyond. Artificial Intelligence in Medicine 116, 102073.

Min, S., Chen, X., Zha, Z.J., Wu, F., Zhang, Y., 2019. A two-stream mutual attention network for semi-supervised biomedical segmentation with noisy labels, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 4578–4585.

Modarres, M., Kaminskiy, M.P., Krivtsov, V., 2016. Reliability engineering and risk analysis: a practical guide. CRC press.

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.

Mondal, A., Dolz, J., Desrosiers, C., 2018. Few-shot 3D multi-modal medical image segmentation using generative adversarial learning. arXiv preprint arXiv:1810.12241 .

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.

Myronenko, A., 2018. 3d mri brain tumor segmentation using autoencoder regularization, in: International MICCAI Brainlesion Workshop, Springer. pp. 311–320.

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.
Narnhofer, D., Effland, A., Kobler, E., Hammernik, K., Knoll, F., Pock, T., 2021. Bayesian uncertainty estimation of learned variational mri reconstruction. IEEE Transactions on Medical Imaging 41, 279–291.
Natekar, P., Kori, A., Krishnamurthi, G., 2020. Demystifying brain tumor segmentation networks: interpretability and uncertainty analysis. Frontiers in computational neuroscience 14, 6.
Nath, V., Yang, D., Landman, B.A., Xu, D., Roth, H.R., 2020. Diminishing uncertainty within the training pool: Active learning for medical image segmentation. IEEE Transactions on Medical Imaging 40, 2534–2547.
Neal, R.M., 2003. Slice sampling. The annals of statistics 31, 705–767.
Neumann, D., Mansi, T., Georgescu, B., Kamen, A., Kayvanpour, E., Amr, A., Sedaghat-Hamedani, F., Haas, J., Katus, H., Meder, B., et al., 2014. Robust image-based estimation of cardiac tissue parameters and their uncertainty from noisy data, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14-18, 2014, Proceedings, Part II 17, Springer. pp. 9–16.
Ng, M., Guo, F., Biswas, L., Petersen, S.E., Piechnik, S.K., Neubauer, S., Wright, G., 2022. Estimating uncertainty in neural networks for cardiac mri segmentation: A benchmark study. IEEE Transactions on Biomedical Engineering .
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, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 2322–2326.
Olkin, I., Rubin, H., 1964. Multivariate beta distributions and independence properties of the wishart distribution. The Annals of Mathematical Statistics , 261–269.
Oreshkin, B.N., Arbel, T., 2013. Uncertainty driven probabilistic voxel selection for image registration. IEEE transactions on medical imaging 32, 1777–1790.
Osawa, K., Swaroop, S., Khan, M.E.E., Jain, A., Eschenhagen, R., Turner, R.E., Yokota, R., 2019. Practical deep learning with bayesian principles. Advances in neural information processing systems 32.
Ouyang, X., Xue, Z., Zhan, Y., Zhou, X.S., Wang, Q., Zhou, Y., Wang, Q., Cheng, J.Z., 2019. Weakly supervised segmentation framework with uncertainty: A study on pneumothorax segmentation in chest x-ray, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22, Springer. pp. 613–621.
Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., Snoek, J., 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. Advances in neural information processing systems 32.
Pandey, P.U., Guy, P., Hodgson, A.J., 2022. Can uncertainty estimation predict segmentation performance in ultrasound bone imaging? International Journal of Computer Assisted Radiology and Surgery 17, 825–832.
Parisot, S., Wells III, W., Chemouny, S., Duffau, H., Paragios, N., 2014. Concurrent tumor segmentation and registration with uncertainty-based sparse non-uniform graphs. Medical image analysis 18, 647–659.
Parzen, E., 1962. On estimation of a probability density function and mode. The annals of mathematical statistics 33, 1065–1076.
Peressutti, D., Penney, G.P., Housden, R.J., Kolbitsch, C., Gomez, A., Rijkhorst, E.J., Barratt, D.C., Rhode, K.S., King, A.P., 2013. A novel bayesian respiratory motion model to estimate and resolve uncertainty in image-guided cardiac interventions. Medical image analysis 17, 488–502.
Peter, L., Alexander, D.C., Magnain, C., Iglesias, J.E., 2021. Uncertainty-aware annotation protocol to evaluate deformable registration algorithms. IEEE Transactions on Medical Imaging 40, 2053–2065.
Pham, H.H., Le, T.T., Tran, D.Q., Ngo, D.T., Nguyen, H.Q., 2021. Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels. Neurocomputing 437, 186–194.
Pham, T.D., 2014. Nonstationary mapping of spatial uncertainty for medical image classification, in: 2014 International Conference on Medical Biometrics, IEEE. pp. 164–168.
Qendro, L., Campbell, A., Lio, P., Mascolo, C., 2021. Early exit ensembles for uncertainty quantification, in: Machine Learning for Health, PMLR. pp. 181–195.

Qian, L., Chen, J., Urakov, T., Gu, W., Liang, L., 2020. Cq-vae: Coordinate quantized vae for uncertainty estimation with application to disk shape analysis from lumbar spine mri images, in: 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE. pp. 580–585.

Qiao, P., Li, H., Song, G., Han, H., Gao, Z., Tian, Y., Liang, Y., Li, X., Zhou, S.K., Chen, J., 2022. Semi-supervised ct lesion segmentation using uncertainty-based data pairing and swapmix. IEEE Transactions on Medical Imaging .

Rafael-Palou, X., Aubanell, A., Ceresa, M., Ribas, V., Piella, G., Ballester, M.A.G., 2022. Prediction of lung nodule progression with an uncertainty-aware hierarchical probabilistic network. Diagnostics 12, 2639.

Rahman, A.U., Saeed, M., Saeed, M.H., Zebari, D.A., Albahar, M., Abdulkareem, K.H., Al-Waisy, A.S., Mohammed, M.A., 2023a. A framework for susceptibility analysis of brain tumours based on uncertain analytical cum algorithmic modeling. Bioengineering 10, 147.

Rahman, R., Alam, M.G.R., Reza, M.T., Huq, A., Jeon, G., Uddin, M.Z., Hassan, M.M., 2023b. Demystifying evidential dempster shafer-based cnn architecture for fetal plane detection from 2d ultrasound images leveraging fuzzy-contrast enhancement and explainable ai. Ultrasonics 132, 107017.

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.

Rao, S.S., Berke, L., 1997. Analysis of uncertain structural systems using interval analysis. AIAA journal 35, 727–735.

Redekop, E., Chernyavskiy, A., 2021. Uncertainty-based method for improving poorly labeled segmentation datasets, in: 2021 IEEE 18th international symposium on biomedical imaging (ISBI), IEEE. pp. 1831–1835.

Reiß, S., Seibold, C., Freytag, A., Rodner, E., Stiefelhagen, R., 2022. Graph-constrained contrastive regularization for semi-weakly volumetric segmentation, in: European Conference on Computer Vision, Springer. pp. 401–419.

Risholm, P., Janoos, F., Norton, I., Golby, A.J., Wells III, W.M., 2013. Bayesian characterization of uncertainty in intra-subject non-rigid registration. Medical image analysis 17, 538–555.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer. pp. 234–241.

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

Ruan, Y., Li, D., Marshall, H., Miao, T., Cossetto, T., Chan, I., Daher, O., Accorsi, F., Goela, A., Li, S., 2020. Mt-ucgan: Multi-task uncertainty-constrained gan for joint segmentation, quantification and uncertainty estimation of renal tumors on ct, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23, Springer. pp. 439–449.

Rupprecht, C., Laina, I., DiPietro, R., Baust, M., Tombari, F., Navab, N., Hager, G.D., 2017. Learning in an uncertain world: Representing ambiguity through multiple hypotheses, in: Proceedings of the IEEE international conference on computer vision, pp. 3591–3600.

Sagar, A., 2022. Uncertainty quantification using variational inference for biomedical image segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 44–51.

Sahinler, S., Topuz, D., 2007. Bootstrap and jackknife resampling algorithms for estimation of regression parameters. Journal of Applied Quantitative Methods 2, 188–199.

Sambyal, A.S., Krishnan, N.C., Bathula, D.R., 2022. Towards reducing aleatoric uncertainty for medical imaging tasks, in: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 1–4.

Sampat, M.P., Wang, Z., Markey, M.K., Whitman, G.J., Stephens, T.W., Bovik, A.C., 2006. Measuring intra-and inter-observer agreement in identifying and localizing structures in medical images, in: 2006 International Conference on Image Processing, IEEE. pp. 81–84.

Schmidt, A., Morales-Álvarez, P., Molina, R., 2023. Probabilistic modeling of inter-and intra-observer variability in medical image segmentation. arXiv preprint arXiv:2307.11397 .

Schobs, L.A., Swift, A.J., Lu, H., 2022. Uncertainty estimation for heatmap-based landmark localization. IEEE Transactions on Medical Imaging 42, 1021–1034.

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, in: Computational Pathology and Ophthalmic Medical Image Analysis: First International Workshop, COMPAY 2018, and 5th International Workshop, OMIA 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16-20, 2018, Proceedings 5, Springer. pp. 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, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22, Springer. pp. 282–290.

Seeböck, P., Orlando, J.I., Schlegl, T., Waldstein, S.M., Bogunović, H., Klimscha, S., Langs, G., Schmidt-Erfurth, U., 2019. Exploiting epistemic uncertainty of anatomy segmentation for anomaly detection in retinal oct. IEEE transactions on medical imaging 39, 87–98.

Senapati, J., Roy, A.G., Pölsterl, S., Gutmann, D., Gatidis, S., Schlett, C., Peters, A., Bamberg, F., Wachinger, C., 2020. Bayesian neural networks for uncertainty estimation of imaging biomarkers, in: Machine Learning in Medical Imaging: 11th International Workshop, MLMI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Proceedings 11, Springer. pp. 270–280.

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.

Shafer, G., 1976. A mathematical theory of evidence. volume 42. Princeton university press.

Shannon, C.E., 1948. A mathematical theory of communication. The Bell system technical journal 27, 379–423.

Shaw, R., Sudre, C.H., Ourselin, S., Cardoso, M.J., 2020. A heteroscedastic uncertainty model for decoupling sources of mri image quality, in: Medical Imaging with Deep Learning, PMLR. pp. 733–742.

Shi, Y., Zhang, J., Ling, T., Lu, J., Zheng, Y., Yu, Q., Qi, L., Gao, Y., 2021. Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation. IEEE transactions on medical imaging 41, 608–620.

Shi, Y., Zu, C., Yang, P., Tan, S., Ren, H., Wu, X., Zhou, J., Wang, Y., 2023. Uncertainty-weighted and relation-driven consistency training for semi-supervised head-and-neck tumor segmentation. Knowledge-Based Systems 272, 110598.

Silva, J.L., Oliveira, A.L., 2021. Using soft labels to model uncertainty in medical image segmentation. arXiv preprint arXiv:2109.12622 .

Smithson, M., 2003. Confidence intervals. 140, Sage.

Snoek, L., Miletić, S., Scholte, H.S., 2019. How to control for confounds in decoding analyses of neuroimaging data. Neuroimage 184, 741–760.

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

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research 15, 1929–1958.

Stehman, S.V., 1997. Selecting and interpreting measures of thematic classification accuracy. Remote sensing of Environment 62, 77–89.

Stewart, W.J., 2009. Probability, Markov chains, queues, and simulation: the mathematical basis of performance modeling. Princeton university press.

Sudarshan, V.P., Upadhyay, U., Egan, G.F., Chen, Z., Awate, S.P., 2021. Towards lower-dose pet using physics-based uncertainty-aware multimodal learning with robustness to out-of-distribution data. Medical Image Analysis 73, 102187.

Sun, Y., Zhou, C., Fu, Y., Xue, X., 2019. Parasitic gan for semi-supervised brain tumor segmentation, in:

2019 IEEE International Conference on Image Processing (ICIP), IEEE, Taipei, Taiwan. pp. 1535–1539.
Tabarisaadi, P., Khosravi, A., Nahavandi, S., 2022. Uncertainty-aware skin cancer detection: The element of doubt. Computers in Biology and Medicine 144, 105357.
Tang, P., Yang, P., Nie, D., Wu, X., Zhou, J., Wang, Y., 2022. Unified medical image segmentation by learning from uncertainty in an end-to-end manner. Knowledge-Based Systems 241, 108215.
Tanno, R., Worrall, D.E., Ghosh, A., Kaden, E., Sotiropoulos, S.N., Criminisi, A., Alexander, D.C., 2017. Bayesian image quality transfer with cnns: exploring uncertainty in dmri super-resolution, in: Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20, Springer. pp. 611–619.
Tardy, M., Scheffer, B., Mateus, D., 2019. Uncertainty measurements for the reliable classification of mammograms, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 495–503.
Tavakoli, F., Ghasemi, J., 2018. Brain mri segmentation by combining different mri modalities using dempster–shafer theory. IET Image Processing 12, 1322–1330.
Teye, M., Azizpour, H., Smith, K., 2018. Bayesian uncertainty estimation for batch normalized deep networks, in: International Conference on Machine Learning, PMLR. pp. 4907–4916.
Thagaard, J., Hauberg, S., van der Vegt, B., Ebstrup, T., Hansen, J.D., Dahl, A.B., 2020. Can you trust predictive uncertainty under real dataset shifts in digital pathology?, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23, Springer. pp. 824–833.
Thiagarajan, P., Khairnar, P., Ghosh, S., 2021. Explanation and use of uncertainty quantified by bayesian neural network classifiers for breast histopathology images. IEEE transactions on medical imaging 41, 815–825.
Tran, D., Dusenberry, M., van der Wilk, M., Hafner, D., 2019. Bayesian layers: A module for neural network uncertainty. Advances in neural information processing systems 32.
Ulmer, D., Cinà, G., 2021. Know your limits: Uncertainty estimation with relu classifiers fails at reliable ood detection, in: Uncertainty in Artificial Intelligence, PMLR. pp. 1766–1776.
Valen, J., Balki, I., Mendez, M., Qu, W., Levman, J., Bilbily, A., Tyrrell, P.N., 2022. Quantifying uncertainty in machine learning classifiers for medical imaging. International Journal of Computer Assisted Radiology and Surgery 17, 711–718.
Venturini, L., Papageorghiou, A.T., Noble, J.A., Namburete, A.I., 2020. Uncertainty estimates as data selection criteria to boost omni-supervised learning, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23, Springer. pp. 689–698.
Vinod, S.K., Min, M., Jameson, M.G., Holloway, L.C., 2016. A review of interventions to reduce inter-observer variability in volume delineation in radiation oncology. Journal of medical imaging and radiation oncology 60, 393–406.
Vlašić, T., Matulić, T., Seršić, D., 2023. Estimating uncertainty in pet image reconstruction via deep posterior sampling. arXiv preprint arXiv:2306.04664 .
Wachinger, C., Golland, P., Reuter, M., Wells, W., 2014. Gaussian process interpolation for uncertainty estimation in image registration, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14-18, 2014, Proceedings, Part I 17, Springer. pp. 267–274.
Wallman, M., Smith, N.P., Rodriguez, B., 2014. Computational methods to reduce uncertainty in the estimation of cardiac conduction properties from electroanatomical recordings. Medical image analysis 18, 228–240.
Wang, C., Lv, X., Shao, M., Qian, Y., Zhang, Y., 2023a. A novel fuzzy hierarchical fusion attention convolution neural network for medical image super-resolution reconstruction. Information Sciences 622, 424–436.
Wang, G., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., Zhang, S., 2020a. Uncertainty-guided efficient interactive refinement of fetal brain segmentation from stacks of mri slices, in: Medical Image

Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23, Springer. pp. 279–288.

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., 2018a. Interactive medical image segmentation using deep learning with image-specific fine tuning. IEEE transactions on medical imaging 37, 1562–1573.

Wang, J., Wells, W.M., Golland, P., Zhang, M., 2018b. Efficient laplace approximation for bayesian registration uncertainty quantification, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I, Springer. pp. 880–888.

Wang, K., Zhan, B., Zu, C., Wu, X., Zhou, J., Zhou, L., Wang, Y., 2021a. Tripled-uncertainty guided mean teacher model for semi-supervised medical image segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24, Springer. pp. 450–460.

Wang, K., Zhan, B., Zu, C., Wu, X., Zhou, J., Zhou, L., Wang, Y., 2022a. Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning. Medical Image Analysis 79, 102447.

Wang, L., Ju, L., Zhang, D., Wang, X., He, W., Huang, Y., Yang, Z., Yao, X., Zhao, X., Ye, X., et al., 2021b. Medical matting: a new perspective on medical segmentation with uncertainty, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24, Springer. pp. 573–583.

Wang, L., Ye, X., Ju, L., He, W., Zhang, D., Wang, X., Huang, Y., Feng, W., Song, K., Ge, Z., 2023b. Medical matting: Medical image segmentation with uncertainty from the matting perspective. Computers in Biology and Medicine 158, 106714.

Wang, S., Zhu, Y., Lee, S., Elton, D.C., Shen, T.C., Tang, Y., Peng, Y., Lu, Z., Summers, R.M., 2022b. Global-local attention network with multi-task uncertainty loss for abnormal lymph node detection in mr images. Medical Image Analysis 77, 102345.

Wang, T., Lu, J., Lai, Z., Wen, J., Kong, H., 2022c. Uncertainty-guided pixel contrastive learning for semi-supervised medical image segmentation, in: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI, pp. 1444–1450.

Wang, X., Gao, S., Jiang, K., Zhang, H., Wang, L., Chen, F., Yu, J., Yang, F., 2023c. Multi-level uncertainty aware learning for semi-supervised dental panoramic caries segmentation. Neurocomputing 540, 126208.

Wang, Y., Zhang, Y., Tian, J., Zhong, C., Shi, Z., Zhang, Y., He, Z., 2020b. Double-uncertainty weighted method for semi-supervised learning, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23, Springer. pp. 542–551.

Warfield, S.K., Zou, K.H., Wells, W.M., 2002. Validation of image segmentation and expert quality with an expectation-maximization algorithm, in: Medical Image Computing and Computer-Assisted Intervention—MICCAI 2002: 5th International Conference Tokyo, Japan, September 25–28, 2002 Proceedings, Part I 5, Springer. pp. 298–306.

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.

Wu, J., Gu, R., Lu, T., Zhang, S., Wang, G., 2023. Upl-tta: Uncertainty-aware pseudo label guided fully test time adaptation for fetal brain segmentation, in: International Conference on Information Processing in Medical Imaging, Springer. pp. 237–249.

Wu, J., Lian, C., Ruan, S., Mazur, T.R., Mutic, S., Anastasio, M.A., Grigsby, P.W., Vera, P., Li, H., 2018.

Treatment outcome prediction for cancer patients based on radiomics and belief function theory. IEEE transactions on radiation and plasma medical sciences 3, 216–224.

Wu, S., Chen, C., Xiong, Z., Chen, X., Sun, X., 2021a. Uncertainty-aware label rectification for domain adaptive mitochondria segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24, Springer. pp. 191–200.

Wu, Z., Yang, Y., Gu, J., Tresp, V., 2021b. Quantifying predictive uncertainty in medical image analysis with deep kernel learning, in: 2021 IEEE 9th International Conference on Healthcare Informatics (ICHI), IEEE. pp. 63–72.

Xia, Y., Liu, F., Yang, D., Cai, J., Yu, L., Zhu, Z., Xu, D., Yuille, A., Roth, H., 2020a. 3d semi-supervised learning with uncertainty-aware multi-view co-training, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3646–3655.

Xia, Y., Yang, D., Yu, Z., Liu, F., Cai, J., Yu, L., Zhu, Z., Xu, D., Yuille, A., Roth, H., 2020b. 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, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 481–491.

Xiao, R., Ding, H., Zhai, F., Zhao, T., Zhou, W., Wang, G., 2017. Vascular segmentation of head phase-contrast magnetic resonance angiograms using grayscale and shape features. Computer Methods and Programs in Biomedicine 142, 157–166.

Xiao, Z., Su, Y., Deng, Z., Zhang, W., 2022. Efficient combination of cnn and transformer for dual-teacher uncertainty-guided semi-supervised medical image segmentation. Computer Methods and Programs in Biomedicine 226, 107099.

Xie, Y., Liao, H., Zhang, D., Chen, F., 2022. Uncertainty-aware cascade network for ultrasound image segmentation with ambiguous boundary, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 268–278.

Xu, C., Yang, Y., Xia, Z., Wang, B., Zhang, D., Zhang, Y., Zhao, S., 2023. Dual uncertainty-guided mixing consistency for semi-supervised 3d medical image segmentation. IEEE Transactions on Big Data .

Xu, N., Price, B., Cohen, S., Yang, J., Huang, T., 2016. Deep interactive object selection, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society. pp. 373–381.

Xu, S., Chen, Y., Ma, C., Yue, X., 2022a. Deep evidential fusion network for medical image classification. International Journal of Approximate Reasoning 150, 188–198.

Xu, X., Sanford, T., Turkbey, B., Xu, S., Wood, B.J., Yan, P., 2022b. Polar transform network for prostate ultrasound segmentation with uncertainty estimation. Medical Image Analysis 78, 102418.

Xu, Z., Luo, J., Lu, D., Yan, J., Frisken, S., Jagadeesan, J., Wells III, W.M., Li, X., Zheng, Y., Tong, R.K.y., 2022c. Double-uncertainty guided spatial and temporal consistency regularization weighting for learning-based abdominal registration, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 14–24.

Yager, R.R., Zadeh, L.A., 2012. An introduction to fuzzy logic applications in intelligent systems. volume 165. Springer Science & Business Media.

Yang, C.I., Li, Y.P., 2023. Explainable uncertainty quantifications for deep learning-based molecular property prediction. Journal of Cheminformatics 15, 13.

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

Yang, S., Fevens, T., 2021. Uncertainty quantification and estimation in medical image classification, in: Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part III 30, Springer. pp. 671–683.

Yang, X., Niethammer, M., 2015. Uncertainty quantification for lddmm using a low-rank hessian ap-

proximation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part II 18, Springer. pp. 289–296.

Ye, C., Li, Y., Zeng, X., 2020. An improved deep network for tissue microstructure estimation with uncertainty quantification. Medical image analysis 61, 101650.

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, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22, Springer. pp. 605–613.

Yuan, B., Yue, X., Lv, Y., Denoeux, T., 2020. Evidential deep neural networks for uncertain data classification, in: Knowledge Science, Engineering and Management: 13th International Conference, KSEM 2020, Hangzhou, China, August 28–30, 2020, Proceedings, Part II 13, Springer. pp. 427–437.

Zadeh, L.A., 1965. Fuzzy sets. Information and control 8, 338–353.

Zhang, Y., Jiao, R., Liao, Q., Li, D., Zhang, J., 2023a. Uncertainty-guided mutual consistency learning for semi-supervised medical image segmentation. Artificial Intelligence in Medicine 138, 102476.

Zhang, Y., Peng, C., Peng, L., Huang, H., Tong, R., Lin, L., Li, J., Chen, Y.W., Chen, Q., Hu, H., et al., 2021. Multi-phase liver tumor segmentation with spatial aggregation and uncertain region inpainting, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24, Springer. pp. 68–77.

Zhang, Y., Peng, C., Tong, R., Lin, L., Chen, Y.W., Chen, Q., Hu, H., Zhou, S.K., 2023b. Multi-modal tumor segmentation with deformable aggregation and uncertain region inpainting. IEEE Transactions on Medical Imaging .

Zhang, Z., Romero, A., Muckley, M.J., Vincent, P., Yang, L., Drozdzal, M., 2019. Reducing uncertainty in undersampled mri reconstruction with active acquisition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2049–2058.

Zhang, Z., Ye, S., Liu, Z., Wang, H., Ding, W., 2023c. Deep hyperspherical clustering for skin lesion medical image segmentation. IEEE Journal of Biomedical and Health Informatics .

Zhao, C., Li, D., Feng, C., Li, S., 2021. Of-umrn: Uncertainty-guided multitask regression network aided by optical flow for fully automated comprehensive analysis of carotid artery. Medical Image Analysis 70, 101982.

Zhao, Y., Yang, C., Schweidtmann, A., Tao, Q., 2022. Efficient bayesian uncertainty estimation for nnu-net, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 535–544.

Zheng, E., Yu, Q., Li, R., Shi, P., Haake, A., 2021. A continual learning framework for uncertainty-aware interactive image segmentation, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 6030–6038.

Zheng, H., Chen, Y., Yue, X., Ma, C., Liu, X., Yang, P., Lu, J., 2020a. Deep pancreas segmentation with uncertain regions of shadowed sets. Magnetic Resonance Imaging 68, 45–52.

Zheng, H., Motch Perrine, S.M., Pitirri, M.K., Kawasaki, K., Wang, C., Richtsmeier, J.T., Chen, D.Z., 2020b. Cartilage segmentation in high-resolution 3d micro-ct images via uncertainty-guided self-training with very sparse annotation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23, Springer. pp. 802–812.

Zheng, X., Fu, C., Xie, H., Chen, J., Wang, X., Sham, C.W., 2022. Uncertainty-aware deep co-training for semi-supervised medical image segmentation. Computers in Biology and Medicine 149, 106051.

Zhou, Q., Yu, T., Zhang, X., Li, J., 2020. Bayesian inference and uncertainty quantification for medical image reconstruction with poisson data. SIAM Journal on Imaging Sciences 13, 29–52.

Zhou, Z.H., 2012. Ensemble methods: foundations and algorithms. CRC press.

Zimmer, V.A., Gomez, A., Skelton, E., Wright, R., Wheeler, G., Deng, S., Ghavami, N., Lloyd, K., Matthew, J., Kainz, B., et al., 2023. Placenta segmentation in ultrasound imaging: Addressing sources of uncertainty

and limited field-of-view. Medical Image Analysis 83, 102639.
Zou, K., Yuan, X., Shen, X., Chen, Y., Wang, M., Goh, R.S.M., Liu, Y., Fu, H., 2023. Evidencecap: Towards trustworthy medical image segmentation via evidential identity cap. arXiv preprint arXiv:2301.00349 .

Supplementary Material A

Bayesian inference

Probabilistic Distribution (PD). In Bayesian inference, probabilistic distribution, such as Gaussian distribution (the most commonly used one), Beta distribution, Poisson Distribution, Exponential distribution, and Dirichlet distribution, are usually used to generate distributions over predictions rather than point estimates (Wallman et al., 2014; Liao et al., 2019; Islam and Glocker, 2021). The parameters of the posterior probabilistic distribution provide estimates of the parameter of interest, and the posterior covariance matrix gives the parameters' uncertainties. The diagonal elements of the covariance matrix correspond to the variances of the estimated parameters.

Gaussian Process (GP). GP is a non-parametric approach used to model functions as probability distributions over possible functions (Wachinger et al., 2014; Wu et al., 2021b; Peter et al., 2021). GP provides not only point predictions but also the associated uncertainty estimates at every point in the input space, making them valuable for regression, interpolation, and optimization tasks where uncertainty needs to be considered.

Bayesian Neural Networks (BNNs). With the success of neural networks (NNs), Bayesian inference is also integrated into neural networks to contract a BNN for uncertainty estimation (Blundell et al., 2015; Bian et al., 2020; Li et al., 2021b; Krygier et al., 2021). In BNN, each weight w of the NN is replaced by placing a prior distribution over the neural network weights rather than having a single fixed value. A prior distribution $p(w)$ is first initialized over the NN weights and the model learns the posterior distribution $p(w|D)$ given the training dataset D and the prior distribution during training. The trained BNN is akin to a virtually infinite ensemble of NNs, where each instance has weights drawn from the learned posterior distribution.

MC methods

MC sampling. MC sampling (Zheng et al., 2021; Ghoshal and Tucker, 2021)is a general interpretation of methods that estimates uncertainty by drawing random samples from a given distribution (normally Gaussian distribution), estimating quantities of interest, and characterizing uncertainty using the obtained samples. Two basic sampling types: simple sampling which draws independent samples from the distribution of interest and importance sampling which draws samples from a different, easier-to-sample distribution and uses weights to adjust for the difference between the true distribution and the sampling distribution are used. Advanced techniques such as Latin hypercube sampling and Jackknife resampling, are also employed to enhance the efficiency of MC methods and reduce the number of required samples.

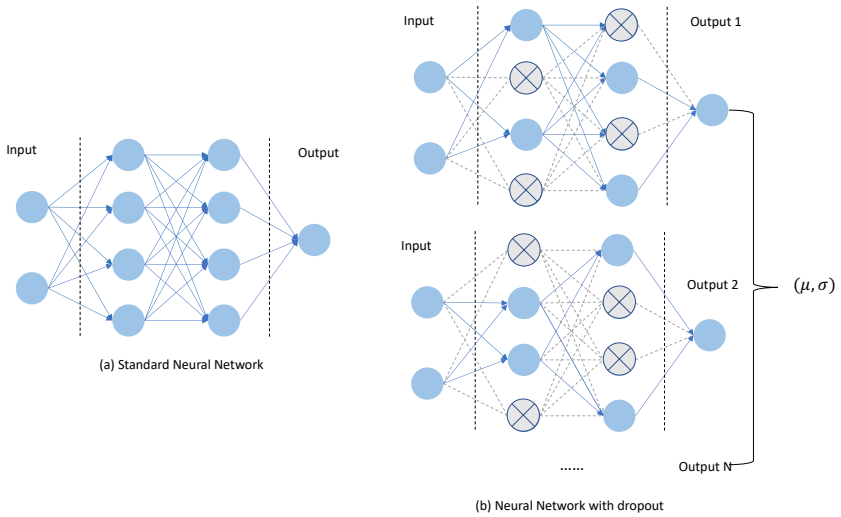


Figure 7: Example of (a) standard Neural Network and (b) Neural Network with Dropout. The dropped neurons were marked in grey and linked by a dotted line. Parameter μ is the mean or expectation of N distributions, and σ is its standard deviation.

Test-Time Dropout (TTD). Dropout (Srivastava et al., 2014) is primarily a regularization technique used during training to prevent overfitting in neural networks. However, it can also be adapted for uncertainty quantification during the test or inference phase. Test-time dropout (TTD) is commonly used in various machine learning applications to estimate predictive uncertainty and make probabilistic predictions. Figure 7 shows an example of a standard Neural Network (left) and a Neural Network with Dropout (right), where the dropped neurons were marked in grey and linked by a dotted line. By applying TTD, the model generates different predictions for the same input data, and these predictions reflect the uncertainty associated with the model's weights and architecture.

Monte Carlo dropout (MCD). In Gal and Ghahramani (2016), the authors demonstrated that an NN trained with dropout operation 7(b) is able to efficiently approximate Bayesian inference that sampling from a variational family (Gaussian Mixture) and approximate the true deep Gaussian process posterior without the associated prohibitive computational cost. Based on this principle, MCD, a SOTA technique for estimating uncertainty in predictions, is proposed. In MCD, dropout is applied at both training and test time. During test time, multiple forward passes are performed with dropout instead of using a single forward pass, resulting in a collection of different predictions for each input.

Markov Chain Monte Carlo (MCMC). MCMC methods use Markov chains to generate dependent data samples. The basic idea is to build such Markov chains, which are easy to

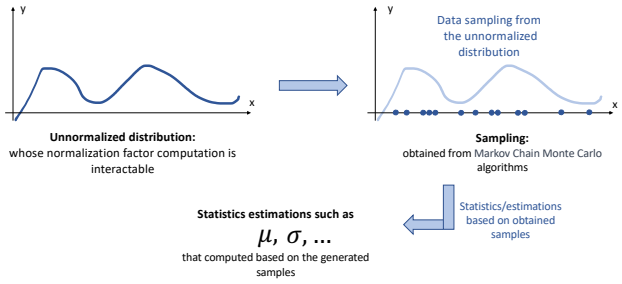


Figure 8: Markov Chain Monte Carlo reasoning process. Parameter μ is the mean or expectation of all sampling samples, σ is the corresponding standard deviation.

sample from and whose stationary distribution is the target distribution, such that when following them, in the limit, we obtain samples from the target distribution (Christophe et al., 2023). MCMC methods, such as the Metropolis-Hastings algorithm (Chib and Greenberg, 1995), Gibbs (Kozumi and Kobayashi, 2011) or slice sampling (Neal, 2003), are used to sample from probability distributions. These methods are particularly useful when analytical solutions are not available. Figure 8 shows the MCMC reasoning process.

Bootstrap. Bootstrap (Efron, 1992; Davison and Hinkley, 1997), a statistical technique used for uncertainty quantification by estimating the sampling variability of a statistical estimator or model, also belongs to the broader category of MC sampling. It involves resampling the observed data (with replacement) to create multiple bootstrap samples. Those samples are then used to estimate the uncertainty by calculating statistics such as the standard deviation, confidence intervals, or percentile intervals of interest.

- Step 1 (Sampling): Randomly select a bootstrap sample of size N (with replacement) from the original dataset.
- Step 2 (Estimation): Apply the desired estimation or modeling procedure to the bootstrap sample to obtain an estimate of interest.
- Step 3: Repeat Steps 1 and 2 N times (typically, $N \gg D$), each time generating a new bootstrap sample and computing the corresponding estimation.
- Step 4 (Uncertainty calculation): Analyze the distribution or variability of the obtained estimates across the N bootstrap samples.

Deep ensemble

The idea of deep ensemble is that N neural networks are trained independently to collect N deterministic predictions. The variability in predictions across ensemble members is then

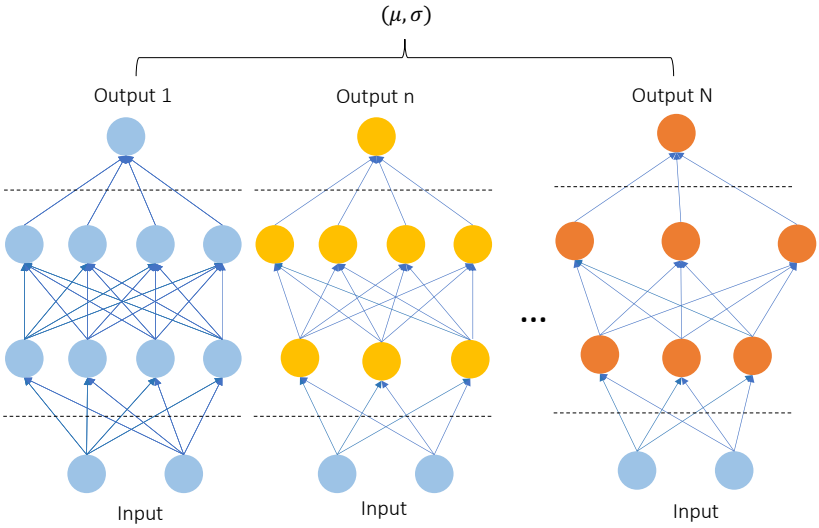


Figure 9: Example of an ensemble model with multiple neural networks. Parameter μ is the mean or expectation of N distributions, σ is the standard deviation.

used to estimate uncertainty (Guo et al., 2022; Zhang et al., 2023b). Figure 9 shows an example of an ensemble model with multiple neural networks, where epistemic uncertainty is captured as different models in the ensemble may have different learned representations, reflecting uncertainty about the true model structure.

Dempster-Shafer Theory

Let $\Omega = \{\omega_1, \omega_2, ..., \omega_C\}$ be a finite set of all possible hypotheses about some problem, called a frame of discernment. Evidence about a variable ω taking values in Ω can be represented by mass function m, from the power set 2^Ω to $[0, 1]$, such that

$$\sum_{A \subseteq \Omega} m(A) = 1, \tag{9a}$$

$$m(\emptyset) = 0. \tag{9b}$$

Each subset $A \subseteq \Omega$ and $m(A)$ is called a focal set of m. The uncertainty (total ignorance) of the problem can be represented as $m(\Omega)$. In DST, the belief about a certain item is elaborated by aggregating different belief functions over the same frame of discernment.

Shafer's model. Assuming that conditional density functions $f(x \mid \omega_c)$ are known, then the conditional likelihood associated with the pattern X is defined by $\ell(\omega_c \mid x) = f(x \mid \omega_c)$. The mass functions are defined according to the knowledge of all hypotheses $\omega_1, \ldots, \omega_C$. Firstly, the plausibility of a simple hypothesis ω_c is proportional to its likelihood. The plausibility is, thus, given by

$$Pl(\{\omega_c\}) = \hbar \cdot \ell(\omega_c \mid x), \quad \forall \omega_c \in \Omega, \tag{10}$$

where $\hbar$ is a normalization factor with $\hbar = 1/\max_{\omega\in\Omega} \ell(\omega|x)$. The plausibility of a set A is, thus, given by

$$Pl(A) = \hbar \cdot \max_{\omega_c \in A} \ell(\omega_c \mid x). \tag{11}$$

Dempster's rule. Given two mass functions m_1 and m_2 derived from two independent items of evidence, the final belief that supports A can be obtained by combining m_1 and m_2 with Dempster's rule (Shafer, 1976), which is defined as

$$(m_1 \oplus m_2)(A) = \frac{1}{1-\kappa} \sum_{B\cap D=A} m_1(B)m_2(D), \tag{12}$$

for all $A \subseteq \Omega, A \neq \emptyset$, and $(m_1 \oplus m_2)(\emptyset) = 0$. The coefficient κ is the degree of conflict between m_1 and m_2 that

$$\kappa = \sum_{B\cap D=\emptyset} m_1(B)m_2(D). \tag{13}$$

Evidential K-Nearest Neighbor (EKNN) rule. Let $N_K(x)$ denote the set of the K nearest neighbors of x in a learning set. Each $x_i \in N_K(x)$ is considered as a piece of evidence regarding the class label of x. The strength of evidence decreases with the distance between x and x_i. The evidence of (x_i, y_i) support class c is represented by

$$m_i(\{\omega_c\}) = \varphi_c(d_i)y_{ic}, \quad 1 \leq c \leq C, \tag{14a}$$

$$m_i(\Omega) = 1 - \varphi_c(d_i), \tag{14b}$$

where d_i is the distance between x and x_i, which can be the Euclidean or other distance function; and $y_{ic} = 1$ if $y_i = \omega_c$ and $y_{ic} = 0$ otherwise. Function φ_c is defined as

$$\varphi_c(d) = \alpha \exp(-\gamma d^2), \tag{15}$$

where α and γ are two tuning parameters. The evidence of the K nearest neighbors of x is fused by Dempster's rule:

$$m = \bigoplus_{x_i \in N_K(x)} m_i. \tag{16}$$

The final decision is made according to maximum plausibility.

Evidential C-Means (ECM). In (Denœux and Masson, 2004), Denoeux et al. proposed an evidential clustering algorithm that extends the notion of fuzzy partition with *Credal partition*, which extends the existing concepts of hard, fuzzy (probabilistic), and possibilistic partition by allocating each object a 'mass of belief,' not only to single clusters but also to any subsets of $\Omega = \{\omega_1, ..., \omega_C\}$. Based on the credal partition, Evidential C-Means (ECM) (Masson and Denoeux, 2008) was introduced to generate mass functions. In ECM, a cluster is represented by a prototype p_c. For each non-empty set $A_j \subseteq \Omega$, a prototype $\bar{p_j}$ is defined as the center of mass of the prototypes p_c such that $\omega_c \in A_j$. Then the non-empty focal set is defined as $F = \{A_1, ..., A_f\} \subseteq 2^\Omega \setminus \{\emptyset\}$. Deriving a credal partition from object data implies determining, for each object x_i, the quantities $m_{ij} = m_i(A_j), A_i \neq \emptyset, A_j \subseteq \Omega$. The distance between an object and any nonempty subset of Ω has thus to be defined by

$$d_{ij}^2 = \|x_i - \bar{p_j}\|^2 . \tag{17}$$

Evidential Neural Network (ENN). In (Denœux, 2000), Denœux proposed an Evidential Neural Network (ENN) classifier in which mass functions are computed based on distances to prototypes. The ENN classifier is composed of an input layer of H neurons, two hidden layers, and an output layer. The first input layer is composed of I units, whose weights vectors are prototypes $p_1, \ldots, p_I$ in input space. The activation of unit i in the prototype layer is

$$s_i = \alpha_i \exp(-\gamma_i d_i^2), \tag{18}$$

where $d_i = \|x - p_i\|$ is the Euclidean distance between input vector x and prototype p_i, $\gamma_i > 0$ is a scale parameter, and $\alpha_i \in [0, 1]$ is an additional parameter. The second hidden layer computes mass functions m_i representing the evidence of each prototype p_i, using the following equations:

$$m_i(\{\omega_c\}) = u_{ic} s_i, \quad c = 1, ..., C \tag{19a}$$

$$m_i(\Omega) = 1 - s_i, \tag{19b}$$

where u_{ic} is the membership degree of prototype i to class ω_c, and $\sum_{c=1}^{C} u_{ic} = 1$. Finally, using Dempster's rule, the third layer combines the I mass functions $m_1, \ldots, m_I$. The output mass function $m = \bigoplus_{i=1}^{I} m_i$ is a discounted Bayesian mass function that summarizes the evidence of the I prototypes.

Subjective Logic (SL). Subjective logic (Josang et al., 2006; Jøsang, 2016) extends DST by introducing additional concepts and principles to handle subjective judgments and uncertainty. It incorporates degrees of belief, disbelief, and uncertainty to capture subjective opinions and incomplete information. Arguments in SL are subjective opinions about state variables that can take values from a domain (aka state space), where a state value can be thought of as a proposition that can be true or false. A binomial opinion applies to a binary state variable and can be represented as a Beta PDF (Probability Density Function) (Kotz et al., 2004). A multinomial opinion applies to a state variable of multiple possible values and can be represented as a Dirichlet PDF (Probability Density Function) (Olkin and Rubin, 1964). For each input X_n, the SL provides belief mass b_c for different classes (Assuming C classes here) and an uncertainty mass U for whole classes. Accordingly,

$$\sum_{c=1}^{C} b_c + u = 1, \tag{20}$$

where $b_c \geq 1$ and $u \geq 1$ denote the probability of the input X_n for the c^{th} class and the input's global ignorance (uncertainty). The evidence $e^n = [e_1^n, ..., e_C^n]$ for the classification result is acquired by an activation function layer softplus and $e_c^n \geq 0$. Then the Dirichlet distribution can be parameterized by $\alpha^n = [\alpha_1^n, ..., \alpha_C^n]$, which associated with the evidence e_c^n, i.e., $\alpha_c^n = e_c^n + 1$. In the end, the image-level belief mass and the uncertainty mass of the classification can be calculated by:

$$b_c^n = \frac{e_C^n}{S^n} = \frac{\alpha_c^n - 1}{S^n}, \tag{21}$$

Table 13: Label uncertainty modeling& analysis in medical image analysis

	Publications	Methods to uncertain label analysis	New dataset	Clinical applications
Label uncertainty modeling	Kohl et al. (2018)	Plausible sets	No	Lung abnormalities seg
	Liao et al. (2019)	PD	No	2D echo quality assess
	Czolbe et al. (2021)	Ensemble, MCD, TTA	No	Skin lesion& lung canc
	Pham et al. (2021)	Soft label	No	Thoracic diseases class
	Redekop and Chernyavskiy (2021)	TTD	No	Skin lesion and liver se
	Islam and Glocker (2021)	PD	No	Brain tumors, prostate and lung nodules segm
	Peter et al. (2021)	PD	No	chest CT scan registra
	Khawaled and Freiman (2022)	PD	No	Brain MRI registration
	Adiga Vasudeva et al. (2022)	PD	No	Left atrium segmentat
	Wu et al. (2021a)	MCD	No	Mitochondria segment
	Ghoshal and Tucker (2022)	MCD	No	COVID-19 detection
	Aljuhani et al. (2022)	MCD	No	Tumor region classifica
	Javadi et al. (2022)	TTA, TTD	No	Prostate cancer detect
	Wu et al. (2023)	TTA	No	Fetal brain Segmentati
	Islam et al. (2023)	PD	No	Breast segmentation
	Del Amor et al. (2023)	Soft label	No	Histology image classif
Uncertain label fusion	Jungo et al. (2018b)	STAPLE, vote, intersection, union	No	Brain tumor segmenta
	Li et al. (2022b)	multi-rater label fusion	No	Breast tumor cellularit
	Lemay et al. (2022)	STAPLE, average, sampling	No	Spinal cord gray matte
New dataset	Irvin et al. (2019)	PD	Yes	Chest radiograph inter
	Ju et al. (2022)	MCD	Yes	Skin lesions, prostate c and retinal disease cla

and

$$U^n = \frac{C}{S^n}, \tag{22}$$

where $S^n = \sum_{c=1}^{C} \alpha_c^n = \sum_{c=1}^{C} e_c^n + 1$ represents the Dirichlet strength.

Supplementary Material B

Table 13 lists the related works that focus on medical image labeling uncertainty analysis.

Image label uncertainty modeling

To deal with the uncertainty of image labels, the straightforward way is to model it with a label distribution map using fuzzy concepts. It can be achieved by introducing probabilistic uncertainty modeling algorithms such as prediction variability (Liao et al., 2019) or non-probabilistic algorithms such as fuzzy predictions (Kohl et al., 2018; Adiga Vasudeva et al., 2022) and label smoothing strategies (Del Amor et al., 2023; Islam and Glocker, 2021; Pham et al., 2021).

In 2018, Kohl et al. approximated the uncertain expert label distribution using generative neural networks in MIA task (Kohl et al., 2018). They proposed a generative segmentation

model based on a combination of a U-Net with a conditional variational autoencoder that is capable of efficiently producing an unlimited number of plausible sets.

In 2019, Liao et al. proposed a method to model the intra-observer variability in echo quality assessment as an aleatoric uncertainty modeling regression problem with Cumulative Density Function (CDF) Probability (Liao et al., 2019). It addressed the observer variability as aleatoric uncertainty, which models experts' opinions as Laplace or Gaussian distributions over the regression space.

In 2021, Czolbe et al. considered four established strategies, i.e., U-Net with Softmax Output, Ensemble Methods, MCD and Probabilistic U-Net to address the inter-observer variability or intra-observer variability (Czolbe et al., 2021). In the same year, Pham et al. presented a multi-label classification framework based on deep CNNs for predicting the presence of 14 common thoracic diseases and observations (Pham et al., 2021). They trained several state-of-the-art CNNs that exploit hierarchical dependencies among abnormality labels using the label smoothing technique to handle uncertain samples. Redekop and Chernyavskiy proposed to train binary segmentation DCNNs using sets of unreliable pixel-level annotations (Redekop and Chernyavskiy, 2021). Islam et al. proposed a spatially varying label smoothing mechanism for incorporating structural label uncertainty by capturing ambiguity about object boundaries in expert segmentation maps in (Islam and Glocker, 2021).

In 2022, Adiga et al. proposed to estimate the pixel-level uncertainty by leveraging the labeling representation into a set of plausible masks and estimating the uncertainty with a single inference from the labeling representation (Adiga Vasudeva et al., 2022). In the same year, Aljuhani et al. presented an importance-based sampling framework with MCD-based approximate inference for robust histopathology image analysis (Aljuhani et al., 2022). Ghoshal et al. extended the approximate inference for the loss-calibrated Bayesian framework to drop weights-based Bayesian neural networks by maximizing expected utility over a model posterior to calibration uncertainty in deep learning (Ghoshal and Tucker, 2022).

In 2023, Del Amor et al. designed an uncertainty-driven labeling strategy to generate soft labels from 10 non-expert annotators for multi-class skin cancer classification (Del Amor et al., 2023). Based on the soft annotations, they proposed an uncertainty estimation framework to handle these noisy labels with a novel formulation using a dual-branch min–max entropy calibration to penalize inexact labels during the training.

Fusion of uncertain image labels

Research on the fusion of uncertain image labels mainly focuses on modeling and addressing the conflicts or ambiguities among labels. This part of the study deals only with the post-processing of uncertain labels, therefore, we do not distinguish between probabilistic and non-probabilistic methods. In 2018, Jun et al. analyzed the effect of common image label fusion techniques with uncertain labels: (a) no fusion, (b) majority vote, (c) STAPLE (Warfield et al., 2002), (d) intersection and (e) union of all observers, and then analysis model's capability to learn the inter-observer variability into the estimation of segmentation uncertainty regardless of the image content in (Jungo et al., 2018b). An interesting

finding is that the obtained results highlighted the negative effect of fusion methods applied in deep learning to obtain reliable estimates of segmentation uncertainty and showed that the learned observers' uncertainty can be combined with current MCD-based models to characterize the uncertainty of the model's parameters.

In 2022, Lemay et al. compared three label fusion methods: STAPLE, average of the rater's prediction, and random sampling of each rater's prediction (Lemay et al., 2022). The results indicated conventional models trained with a Dice loss, with binary inputs and sigmoid/softmax final activate, were overconfident and underestimated the uncertainty associated with inter-rater variability. Conversely, fusing labels by averaging with the soft prediction framework led to underconfident outputs and overestimation of the rater disagreement.

To efficiently leverage the label ambiguities, in 2022, Li et al. proposed an uncertainty-aware label distribution learning framework (Li et al., 2022b) by converting single-value labels to discrete label distributions and modeling the ambiguity among all possible labels. The framework then learned label distributions by minimizing the KL divergence between the predicted and ground-truth label distributions and mimicked the multi-rater fusion process in clinical practice with a multi-branch feature fusion module to further explore the uncertainties of labels.

New image dataset with uncertainty annotation

In addition to modeling or analyzing the label uncertainty in the existing open public dataset, there are some researchers who contribute to larger-scale medical datasets with uncertainty annotation. For example, in 2019, Irvin et al. presented a large dataset of chest radiographs called CheXpert, which features uncertainty labels and radiologist-labeled reference standard evaluation sets. This dataset consists of 224,316 chest radiographs of 65,240 patients labeled for the presence of 14 common chest radiographic observations (Irvin et al., 2019). To our knowledge, this is the first dataset that provided both accuracy and uncertainty annotations. It helps the development and validation of chest radiograph interpretation models towards improving healthcare access and delivery worldwide. In 2022, Ju et al. released a large re-engineered database that consists of annotations from more than ten ophthalmologists with an unbiased golden standard dataset for evaluation and benchmarking (Ju et al., 2022).

Those label uncertainty analysis methods could have a high impact in real-world applications, such as being used as clinical decision-making algorithms accounting for multiple plausible semantic segmentation hypotheses to provide possible diagnoses and recommend further actions to resolve the present ambiguities.