Radiology

REVIEWS AND COMMENTARY • REVIEW

Quantifying Uncertainty in Deep Learning of Radiologic Images

Shahriar Faghani, MD • Mana Moassefi, MD • Pouria Rouzrokh, MD, MPH, MHPE • Bardia Khosravi, MD, MPH, MHPE • Francis I. Baffour, MD • Michael D. Ringler, MD • Bradley J. Erickson, MD, PhD

From the Artificial Intelligence Laboratory (S.F., M.M., P.R., B.K., B.J.E.) and Division of Musculoskeletal Radiology (F.I.B., M.D.R.), Department of Radiology, Mayo Clinic, 200 1st St SW, Rochester, MN 55905. Received September 4, 2022; revision requested October 18; revision received March 30, 2023; accepted April 11. **Address correspondence to** B.J.E. (email: *bje@mayo.edu*).

Conflicts of interest are listed at the end of this article.

Radiology 2023; 308(2):e222217 • https://doi.org/10.1148/radiol.222217 • Content code: AI

In recent years, deep learning (DL) has shown impressive performance in radiologic image analysis. However, for a DL model to be useful in a real-world setting, its confidence in a prediction must also be known. Each DL model's output has an estimated probability, and these estimated probabilities are not always reliable. Uncertainty represents the trustworthiness (validity) of estimated probabilities. The higher the uncertainty, the lower the validity. Uncertainty quantification (UQ) methods determine the uncertainty level of each prediction. Predictions made without UQ methods are generally not trustworthy. By implementing UQ in medical DL models, users can be alerted when a model does not have enough information to make a confident decision. Consequently, a medical expert could reevaluate the uncertain cases, which would eventually lead to gaining more trust when using a model. This review focuses on recent trends using UQ methods in DL radiologic image analysis within a conceptual framework. Also discussed in this review are potential applications, challenges, and future directions of UQ in DL radiologic image analysis.

Machine learning is defined as the ability of machines to identify patterns and automatically learn from data to predict unseen (future) data. Recent years have seen a rise in deep learning (DL) as a machine learning subfield due to its ability to perform automated feature extraction from data, ease of training, and superior performance over other machine learning algorithms (1).

These DL models can be used for different tasks in radiology, including classification (Fig 1A), segmentation (Fig 1B), regression (Fig 1C), object detection (Fig 1D), and image generation (Fig 1E) (Table 1) (2–5).

While the potential of DL to assist radiologists with automated decision support and thus reduce workload is widely recognized, there is increasing concern about its potentially negative consequences, such as lack of trustworthiness, incorrect decisions, and unfairness, which could degrade health care (6–8). To put machine learning models into practical use, it is crucial to ensure that they are trustworthy (9). To accomplish this, it is essential to include an indication of predictive confidence for each prediction, and this becomes even more important for DL models (10). DL models are sophisticated mathematical functions with millions to billions of parameters, which makes understanding the basis for their predictions challenging.

When making a prediction, probability (ie, how likely something is to happen) is associated with an outcome. Ideally, the reported probability for an outcome predicted by DL should be calibrated, which means that the probability value reflects the actual frequency of that outcome. For example, if an 80% probability of pneumonia is reported for an input chest radiograph but 70% of patients with that chest radiograph have pneumonia, then the model is not calibrated.

During training, a DL model learns how to predict the outcome based on the features it extracts from training data. Still, even a well-trained and calibrated model may see examples in real-use cases that were not well represented in training data. In such scenarios, predictions the model makes on these poorly represented samples are not reliable. These are high uncertainty predictions. Furthermore, the training data of DL models are rarely perfectly labeled, resulting in another source of uncertainty for the model's performance. Because of these factors, the output probability of a given DL model alone, without the uncertainty component, may be misleading and result in inappropriate clinical management. For instance, a fracture detection model for radiography that has been trained on an adult population may detect the growth plate of a pediatric radiograph as a fracture with high probability. However, a model that could output uncertainty for each prediction would identify this example as an uncertain sample despite its high probability. Uncertainty quantification (UQ) provides the DL equivalent of radiologist confidence in the diagnosis and is, therefore, a critical component of the model output.

UQ techniques measure the uncertainty of a model, depicting the trustworthiness of its predictions. Despite its importance, there is a paucity of literature describing UQ techniques in DL, especially for radiologic image analysis. Thus, this review discusses types of uncertainty in a prediction, general ways to report uncertainty, and how to assess and perform calibration. The review then summarizes the literature describing different UQ methods, applications of UQ in radiologic image analysis, and challenges and opportunities in quantifying uncertainty in radiologic DL models.

Abbreviations

DL = deep learning, UQ = uncertainty quantification

Summary

A calibrated deep learning model for analyzing medical images that outputs an uncertainty level for each prediction provides better understanding and trust in stand-alone artificial intelligence models deployed in the clinical workflow.

Essentials

- For a deep learning (DL) model to be useful in a real-world setting, its level of uncertainty when making a prediction must be known, as uncertainty represents the validity of an estimated probability.
- Uncertainty quantification (UQ) techniques measure the trustworthiness of each prediction, allowing end users to be aware of the DL model's prediction reliability, thus affecting radiologists' decision-making.
- Models are not always reliable; because the probability is reliant on the model's output, an outcome's probability does not reflect the validity of the prediction.
- Model calibration aims to maximize the agreement between the prediction value and the true (actual) probability of a specific class (diagnosis) and, thus, model calibration is a necessary step to ensure a trustworthy model by matching the estimated probability with the observed frequency of an outcome; however, calibration alone does not measure uncertainty.
- UQ improves model performance metrics (eg, sensitivity, specificity, accuracy).

Probability and Uncertainty

A simple definition of probability is how likely something is to happen. The formal definition of probability, from a frequentist point of view, is the relative frequency of occurrence of an outcome when the experiment is repeated an infinite number of times (8). For instance, when the probability of having a rainy day tomorrow is reported as 70%, it means that if we experience an infinite number of tomorrows, we will experience a rainy day on 70% of them. Similarly, when we say the probability of having pneumonia based on a chest radiograph is 70%, it means that, hypothetically, from an infinite number of patients with the exact same chest radiograph, 70% would have pneumonia. This is because some other diagnoses may mimic pneumonia findings at chest radiography.

For a DL model to be useful in a real-world setting, its level of uncertainty, or confidence, when making a prediction must be known. The probability of an outcome does not reflect the uncertainty level of a prediction. Hence, in addition to reporting an outcome probability, disclosing the prediction uncertainty is essential. A DL model's level of uncertainty reflects the degree of validity in its outputs (estimated probabilities). In other words, uncertainty represents the probability distribution of probabilities.

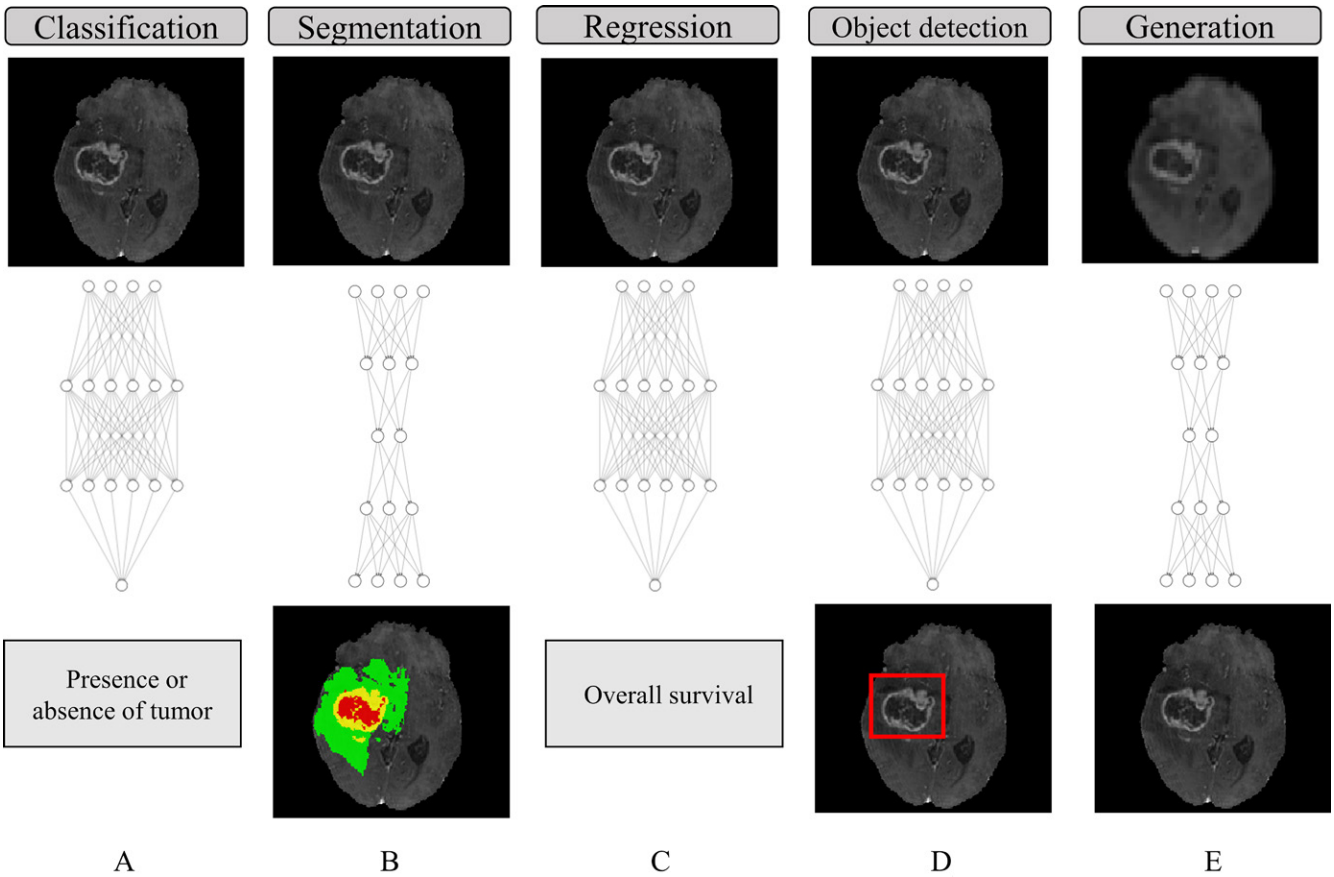

Figure 1: Schematic illustrations show five different deep learning tasks, using MRI as an example. **(A)** A classification task determines the presence or absence of a brain tumor in a specific MRI section. **(B)** A segmentation task partitions a brain tumor into different regions, with green, yellow, and red areas depicting edema, contrast enhancement, and necrotic core area, respectively. **(C)** A regression task predicts the overall survival of a patient with a brain tumor. **(D)** An object detection task detects a brain tumor, then creates a red bounding box around the tumor. **(E)** An image generation task denoises the image (and can also be used to synthesize images).

Table 1: Glossary of Terms

Term	Definition
Deep learning	A subset of machine learning algorithms that makes the feature extraction process automated
Classification task	Assigning a class label to input examples (eg, deciding whether a mammogram has benign or malignant findings)
Segmentation task	Dividing different regions of an image by accurate delineation (eg, delineating liver and spleen on an abdominal CT image)
Object detection task	Creating bounding boxes around the object of interest (eg, drawing bounding boxes around lung nodules in chest radiographs)
Regression task	Predicting a continuous numerical value (eg, predicting bone age from pediatric hand radiographs)
Image generation task	Denoising or synthesizing images that do not exist (eg, generating thin-section MRI scans from thick-section MRI scans)
Probability	How likely something is to happen; the formal definition of probability from a frequentist point of view is the relative frequency of occurrence of an experiment's outcome when the experiment is repeated infinitely (eg, flipping a coin infinitely and observing heads in half of the experiments, then the probability of the head is 0.5)
Calibration	Process of matching the estimated probability of a model with the observed frequencies; for instance, if a model assigns 0.6 probability to outcome X, but it is observed that the frequency of outcome X is 0.7, then the model is not calibrated and, hence, the assigned probabilities are not valid and need calibration
Uncertainty	After a model predicts an outcome probability, uncertainty shows the trustworthiness of the estimated probability
Aleatoric uncertainty	The part of uncertainty that stems from data, which happens when there are similar features but different labels (eg, two similar hyperintensities on brain MRI scans might have different diagnoses based on the patient's history, laboratory value, physical examination); this implies that based on stand-alone imaging always, there would be uncertainty in the predictions
Epistemic uncertainty	The part of uncertainty that stems from the "model's knowledge," which happens when model knowledge is not complete for decision-making (eg, a model trained on adult radiographs for bone fracture detection has not seen any pediatric greenstick fracture during training, thus would be uncertain for pediatric fractures)
Uncertainty quantification	A group of methods that can assess the level of uncertainty in each model's prediction
Bayesian methods	For each prediction, the input image will pass multiple times through the model and each time the model's parameters will change slightly, resulting in slightly different predictions; the distribution of predictions can be used as a measure of uncertainty for that specific input
Ensemble methods	In the case of having different models for a single task, if the input image passes through all the models, then the distribution of predictions can be used as a measure of uncertainty for that specific input
Evidential deep learning methods	Tries to collect features belonging to different classes from images, which are called "evidence" in this method, and based on the amount of collected evidence, assigns an uncertainty measure to each prediction
Uncertainty map	Used for reporting uncertainty for segmentation and image generation tasks, where the amount of uncertainty is reported for each pixel and/or voxel; this alerts the reader to which part of the segmented or generated image is trustworthy
Uncertainty interval	A way of reporting uncertainty for each prediction by providing a lower bound and an upper bound for the prediction (eg, 0.6–0.7 as the probability of having pneumonia based on a chest radiograph); a wider interval implies higher uncertainty

Types of Uncertainty in Machine Learning

Uncertainty can be broadly divided into two categories—aleatoric and epistemic. Aleatoric uncertainty, or data uncertainty, can be due to the presence of label errors (label noise, also known as noisy labels) or the ambiguity of the data. For example, take a data set of chest radiographs with pneumonia or pulmonary thromboembolism. If some of the pneumonia chest radiographs are mislabeled or subjectively labeled as pulmonary thromboembolism, this is an example of aleatoric uncertainty due to label errors. But if the data set contains two chest radiographs with the exact same characteristics, one with pneumonia and the other with pulmonary thromboembolism, then the aleatoric uncertainty would be due to the ambiguity of characteristics for that chest radiograph and thus is irreducible regardless of how the model is trained (11). Part of the aleatoric uncertainty caused by mislabeled data can be improved by gathering higher quality (with more accurate labels) training data. In contrast, epistemic uncertainty, or knowledge uncertainty, refers to the uncertainty of a model due to a lack of knowledge. Epistemic uncertainty often arises due to a lack of enough training data or less-than-perfect performance of the DL model and thus can be reduced by gathering more training data and/or improving the model's architecture or training strategy. For example, a model trained on a data set with typical radiographic characteristics of COVID-19 for pneumonia detection would be highly uncertain on atypical COVID-19 cases because the model has not "seen" any atypical cases during training; however, by providing atypical examples in the data set and retraining the model, the epistemic uncertainty of the model will be decreased (Fig 2). When not specified, the term "model uncertainty" implies epistemic uncertainty, which is the focus of this review.

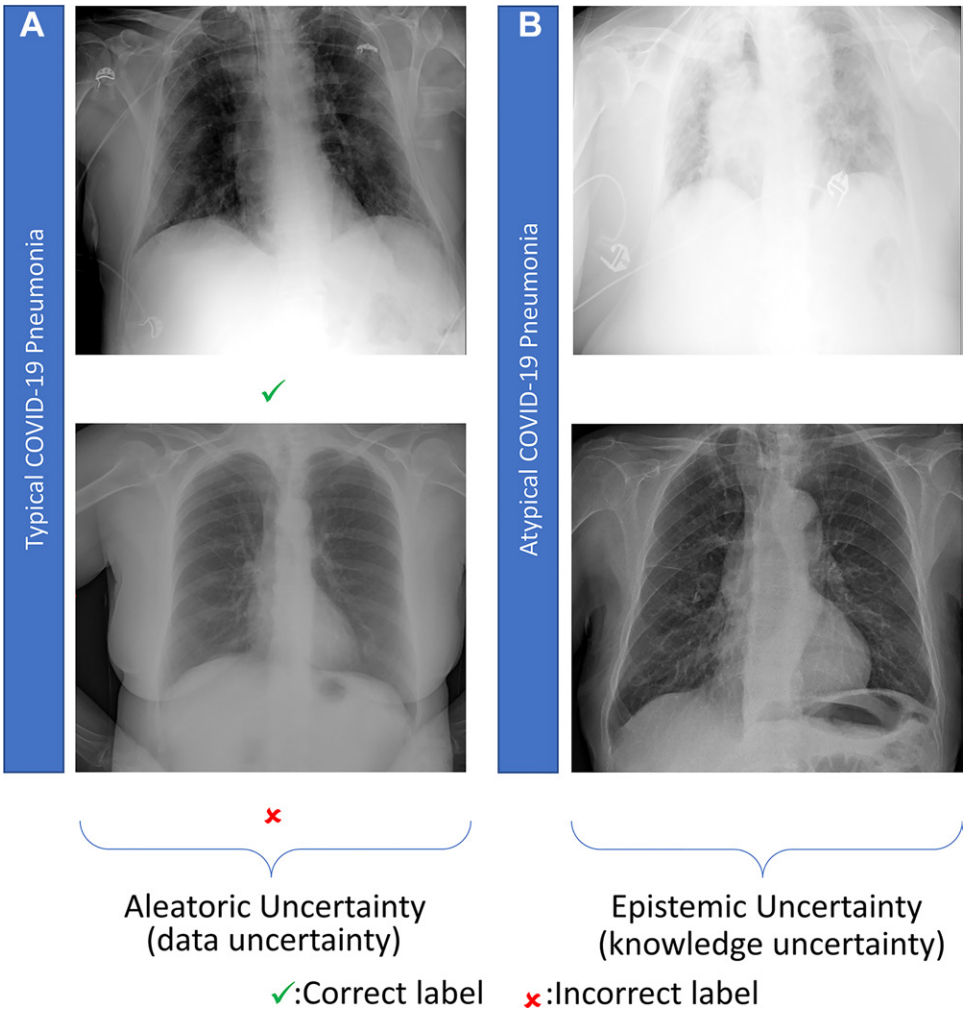

Figure 2: Schematic illustrations show **(A)** aleatoric uncertainty (data uncertainty; ie, noise in raw data) and **(B)** epistemic uncertainty (knowledge uncertainty; ie, lack of diverse data), using radiography as an example. With aleatoric uncertainty **(A)**, all the chest radiographs should indicate typical COVID-19 pneumonia, but some are normal radiographs that are mislabeled as typical COVID-19 pneumonia. With epistemic uncertainty **(B)**, the model has not been trained on atypical COVID-19 pneumonia radiographs and, thus, the lack of diverse data results in unreliable predictions of outliers.

Reporting Uncertainty

Uncertainty is inversely related to the trustworthiness of the model outputs. The higher the uncertainty, the less the outputs are trustworthy. Thus, it is critical to know the model's uncertainty for each output in the clinical decision-making process.

Three ways to report uncertainty are reporting a single number, assigning an uncertainty interval, or plotting an uncertainty map. Reporting a single number represents the magnitude of uncertainty for each prediction (eg, a prediction with 0.2 uncertainty is more trustable than a prediction with 0.7 uncertainty). Thiagarajan et al (12) reported uncertainty in this way for the diagnosis of pneumonia using chest radiography. Assigning an interval to each prediction represents the lower and upper bounds of confidence levels to each prediction (eg, [72,76] for age prediction using brain MRI indicates 72 and 76 years as the lower and upper bounds of age). Finally, for certain tasks, such as segmentation and image generation, uncertainty can be reported for each predicted pixel (voxel), resulting in an uncertainty map (Fig 3). For example, Zabihollahy et al (13) reported a fully automated DL algorithm that uses MRI for segmenting clinical target volume in uterine cervical cancer for radiation therapy planning. By applying UQ techniques, their proposed method generates an uncertainty map along with clinical target volume segmentation, helping clinicians to identify areas with high uncertainty for manual correction, thereby potentially decreasing contour variability and enhancing the precision of clinical segmentation.

While the majority of studies did not use a measure of aleatoric and epistemic uncertainty in their models' performance, some studies have reported them separately (14–16). When these two uncertainties are reported separately, it allows uncertainty source identification, which may highlight ways to reduce uncertainty. A part of aleatoric uncertainty can be mitigated by using more accurately labeled data, whereas high epistemic uncertainty can be reduced by providing more diverse data in the training examples. An example of reducing the aleatoric uncertainty in a brain hemorrhage detection model is by correcting mislabeled training instances with parenchymal calcification marked as intraparenchymal hemorrhage. To decrease epistemic uncertainty, the model can be trained with more diverse cases, such as those with hypercellular neoplasms showing similar CT characteristics to brain hemorrhage.

Calibration

Model calibration aims to maximize the agreement between the estimated probability and the true (actual) frequency of a specific class (diagnosis). In real-world clinical scenarios for each prediction, in addition to diagnosis, it is of interest to know the probability of the diagnosis. But the output of binary classification (ie, presence vs absence of tumor) will be "0" or "1" after applying the threshold (usually 0.5) on estimated probability; thus, the probability is unclear when using binary classification.

Calibration is defined as adjusting the output values to accurately represent the true probabilities of the various diagnostic options. Research has shown that most of the probabilities from DL models have poor calibration. Poor calibration means that a model has low clinical value, as it cannot be trusted for

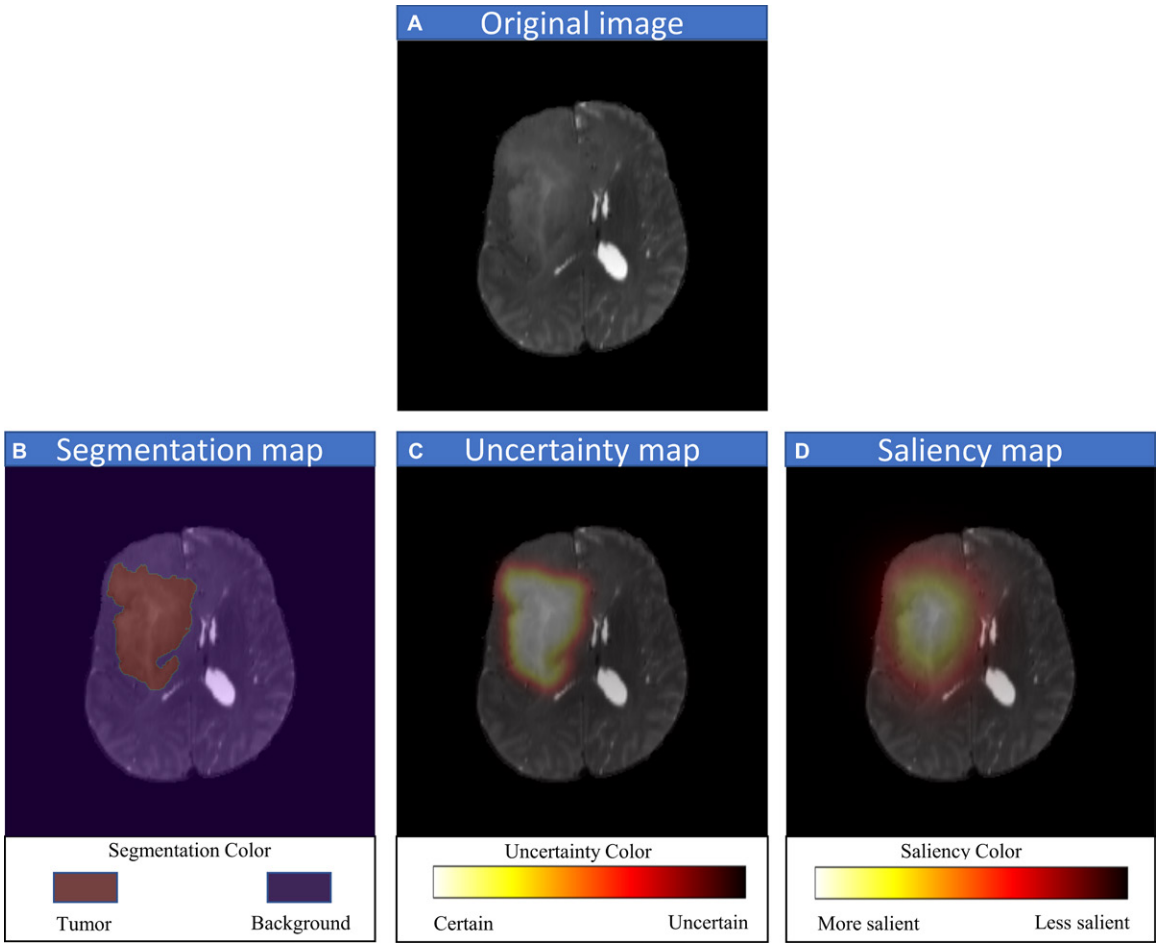

Figure 3: Comparison of uncertainty map and saliency map for brain segmentation and classification at MRI. **(A)** T2-weighted brain MRI scan shows glioblastoma. **(B)** Segmentation map of the same MRI scan shows the ground truth mask containing tumoral tissue. **(C)** Uncertainty map of tumoral tissue from the same MRI scan shows the predicted uncertainty of a segmentation deep learning model, which indicates higher uncertainty for tumor margins than the central area. **(D)** Saliency map of the same MRI scan shows classification of tumor versus no tumor. Although the uncertainty and saliency maps look similar, they serve different purposes; the uncertainty map **(C)** displays the confidence of the predicted segmentation mask for each pixel, whereas the saliency map **(D)** shows where the salient image features are for the classification problem at hand.

clinical decision-making due to incorrect prediction of probabilities (17,18). Rajaraman et al (19) showed how an uncalibrated tuberculosis detection model performs inferiorly in comparison to a calibrated model, and thus is unreliable. Also note that a poorly calibrated model can still have high accuracy due to the effect of thresholding, which raises the need for assessment of model calibration besides conventional performance metrics (eg, accuracy, sensitivity, specificity).

The calibration of models can be assessed using different techniques, the most intuitive of which is using a calibration curve. In a calibration curve, the estimated probability of a particular class is plotted against its observed frequency. For instance, for pneumonia versus a normal chest radiograph classifier, the estimated probabilities of pneumonia in the test set are plotted against the true frequency of pneumonia. The ideal calibration curve would be a straight line at a 45° angle, with a perfect match between estimated probabilities and observed frequencies, and thus a calibrated probability (20). Without this calibration, an uncalibrated model may overestimate or underestimate the probabilities for all or some output values.

It is worth mentioning that model calibration depends on the pretest probability of disease. For example, if a model is calibrated using the data of a population with 20% disease prevalence, the model needs recalibration prior to applying it to a new population with 60% disease prevalence. This implies that calibration assessment does not generalize to all circumstances, particularly when applying the model to a distinct population with different disease prevalence. Therefore, a calibration assessment should be performed on the target population.

Although poor calibration is a source of uncertainty, good calibration does not reflect certainty. In other words, regardless of calibration, all reported probabilities of a DL model are point estimates of the predicted outcome, and the model can be certain or uncertain about each of those estimated probabilities. Imagine two well-calibrated pneumonia detection models, A and B, trained to predict the presence or absence of bacterial pneumonia, regardless of etiology. These were trained on two different data sets, including data set A, with many instances of *Klebsiella* pneumonia, and data set B, with no instances of *Klebsiella* pneumonia. Model A may output a 70% probability

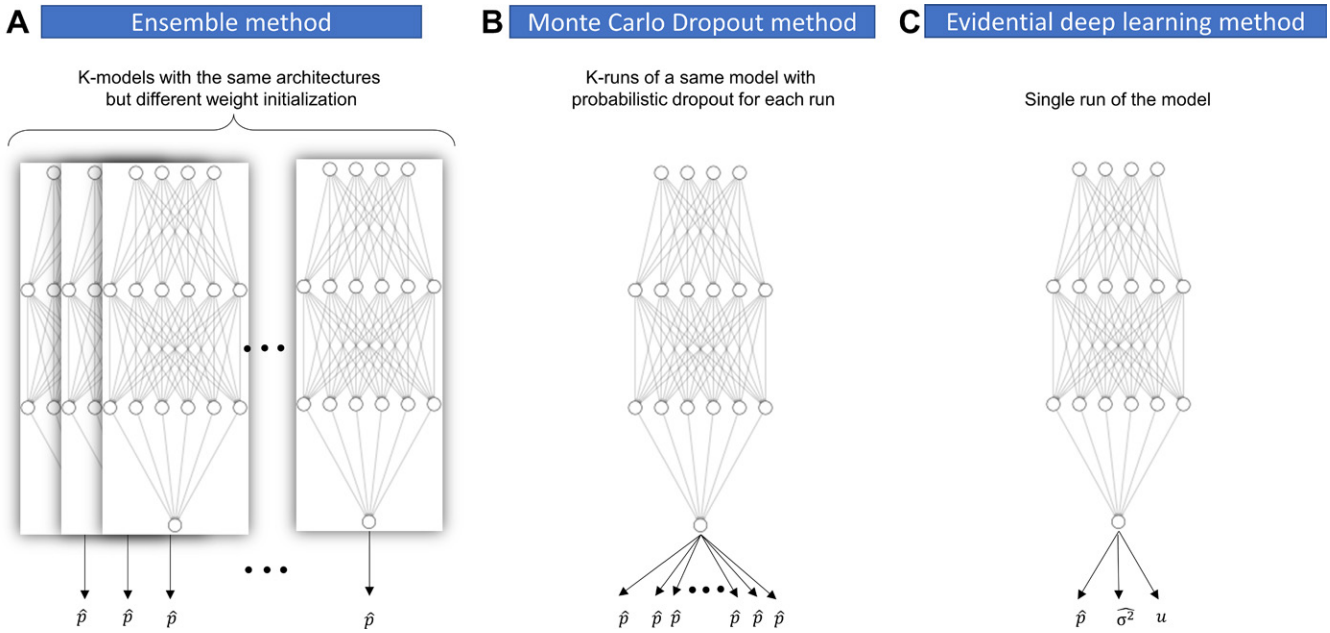

Figure 4: Schematic illustrations show characteristics of different uncertainty quantification methods. **(A)** The ensemble method is performed by training *k* models with the same architectures but different weight initialization. **(B)** The Monte Carlo dropout method is performed by running the same model *k* times with probabilistic dropout for each run. **(C)** The evidential deep learning (DL) method is performed by training a model with evidential DL loss function. Each of these methods has advantages and disadvantages, and the best method will depend on the circumstances. $\hat{p}$ = estimated probability, $\widehat{\sigma^2}$ = estimated variance, u = uncertainty. (Adapted, with permission, from reference 21.)

of pneumonia on an unseen chest radiograph with *Klebsiella* pneumonia findings with an uncertainty level of 5%. Model B might report a 70% probability but with 20% uncertainty for that particular chest radiograph because model B is less familiar with *Klebsiella* pneumonia, and thus is more uncertain. This illustrates the need for understanding the model's uncertainty and its calibration, both of which show the trustworthiness of a prediction.

Although calibration is a necessary step to having a trustworthy DL model, calibration alone does not measure uncertainty. There are many ways to perform model calibration during or after model training, which is beyond the scope of this review. However, some calibration techniques are closely tied to UQ methods, as discussed hereafter. In other words, some UQ methods, in addition to quantifying the uncertainty of predictions, will calibrate DL models without the need to perform any separate calibration.

UQ Methods

As defined previously, UQ measures the uncertainty of a model, depicting the trustworthiness of its predictions. UQ can demonstrate biases resulting in overconfidence or underconfidence in the model's predictions (21). UQ identifies when a DL model has insufficient information to make reliable decisions, thus putting lower trust in the model's predictions. For instance, in the case of a screening tool for stroke based on brain CT scans, if a model predicts stroke but with high uncertainty due to patient motion, then this should also cause uncertainty for the radiologist. Thus, re-examination by a radiologist is required to ensure accurate diagnosis, or at least as accurate as possible, because the source of uncertainty (patient motion) for the model also causes uncertainty for the radiologist. Such communication between the model and medical experts is likely to increase trust in the model's high-certainty predictions as long as those are usually correct.

Although several UQ techniques are discussed in the DL literature, the more popular techniques may be classified into three methodologic groups as follows (Fig 4):

1. Ensemble methods (ie, an ensemble of models) use more than one DL model for prediction. Thus, instead of a single prediction, the distribution of prediction can be reported (22). For example, consider three different bone age prediction DL models. As each model differs slightly, each may predict a different age for a particular pediatric hand radiograph, thus the SD of the predictions can be considered as an uncertainty measure.

2. Bayesian methods approximate the uncertainty of predictions by varying the DL model's weights and running the model multiple times on each instance. The Monte Carlo dropout method is one of the most popular of these that randomly zeros some model's weight during each run (23).

3. Evidential DL methods try to collect features belonging to different classes from images, which are called "evidence" in this method. Based on the amount of collected evidence, an uncertainty measure is assigned to each prediction (higher collected evidence results in lower uncertainty of prediction and vice versa) (24,25). For instance, in a pneumonia detection model, the presence of pneumonia-related characteristics (eg, consolidation or pleural effusion) is considered as evidence. The more evidence the model gathers, the lower its prediction uncertainty becomes.

Each of these methods has advantages and disadvantages, and the best method will depend on circumstances that are beyond the scope of this review.

Applications of UQ

UQ allows the user to gauge the trustworthiness of each model's prediction. In other words, it makes the user aware of when the model's prediction is reliable. In this way, the DL model communicates when a prediction can be relied upon by the radiologist versus when the model is uncertain. Dohopolski et al (26) created a DL model to diagnose lymph node metastasis in patients with oropharyngeal cancer by using PET/CT, with sensitivity and specificity of 0.94 and 0.90. They used the Monte Carlo dropout method to quantify prediction uncertainty, and divided cases into certain and uncertain groups based on the uncertainties. The results showed that sensitivity and specificity were 1.0 and 0.98 for certain cases, and 0.67 and 0.41 for uncertain cases, respectively. This study highlights the potential of UQ in distinguishing reliable predictions from unreliable ones that require review by radiologists.

Furthermore, Laves et al (27) proposed a UQ-based DL model for CT image reconstruction by generating an uncertainty map for the image so that radiologists were aware of which anatomic regions were reliably reconstructed in the generated image.

Some UQ methods can also calibrate the model, improving its performance on a specific task. For example, one method may increase the Dice similarity coefficient in segmentation tasks because a threshold (typically 0.5) must be used to decide if a pixel is inside or outside the segmented region. If the model is not calibrated, then it may include or exclude more pixels than it should. Lee et al (14) demonstrated the efficacy of UQ in improving the performance of DL models in brain tumor segmentation on MRI scans. Their results showed that the Dice coefficient score of enhancing tumor and necrotic tumor increased by 3.15% and 0.58%, respectively. These improvements in defining actual brain tumors, particularly in enhancing tumor and necrotic tumor, are crucial for treatment planning and treatment response follow-up. This study emphasizes the potential of UQ in enhancing DL model performance, leading to improved patient care, to produce an uncertainty map for the segmented region. Using those uncertainty maps, the authors were able to produce more accurate segmentation results.

Khawaled and Freiman (28) developed a DL algorithm that used an uncertainty map to improve brain image coregistration in 390 brain MRI pairs, resulting in 7.3% improvement in registration accuracy and 18% improvement in registration smoothness. This improvement is noteworthy because image registration is a crucial task in various image-guided clinical applications, such as image-based surgical planning, motion compensation, and longitudinal analysis, among others. The ability of the algorithm to enhance the accuracy and smoothness of registration can lead to better clinical decision-making and more precise treatment planning for patients. In another study, Ozdemir et al (29) applied UQ for lung cancer detection on low-dose CT scans to measure the uncertainty of each prediction and to calibrate estimated probabilities for identifying which prediction needed to be reviewed by radiologists. Thus, the authors demonstrated improved overall performance by using uncertainty scores in a simulated workflow.

Finally, UQ allows efficient active learning, whereby the model is continuously trained, updated, and improved with new data (30). Having a model that outputs uncertainty for each prediction allows selection of the uncertain samples for model training instead of using all the available new data, resulting in improved time and computational efficiency (31). For instance, Liu et al (32) used UQ to report uncertainty for pancreas segmentation on abdominal CT scans for iterative refinement of predicted masks and to improve the autosegmentation results.

UQ and Bias Mitigation

Bias due to an algorithm's performance being against or in favor of a population subgroup can arise at any step of DL creation, including data handling, model development, and performance evaluation. This bias often arises when there is a small amount of data for a particular population (eg, in the African American population) (33). Performance evaluation toolboxes, including performance metrics, performance interpretation maps, and UQ, can be used to help detect bias (21).

Proper performance metrics can reveal the existence of bias in a DL model, while each of the performance interpretation maps and UQ methods could potentially highlight a part of the bias source. Figure 5 demonstrates how UQ can unveil model bias where interpretation maps (eg, saliency maps) are unable to detect bias.

Performance interpretation maps usually demonstrate the salient features of the decision-making process, but UQ identifies examples where the model is overconfident or underconfident, thus bringing awareness to bias stemming from the uncertainty of the model and, if so, whether it is due to aleatoric or epistemic uncertainty. By knowing the type of uncertainty in the prediction, developers can reduce bias by providing higher quality data (reducing aleatoric uncertainty) or more diverse data (reducing epistemic uncertainty; eg, including a more heterogeneous patient population in the training data set).

Misconceptions about UQ and CIs

Even without UQ, a CI for a DL model's performance metrics can still be reported, such as with accuracy, sensitivity, and specificity. However, as opposed to UQ, which measures the uncertainty of each model's prediction, a CI does not reflect the uncertainty of a single prediction because CIs are intended for use with population-based metrics (eg, accuracy, sensitivity, specificity). Such measures are not as useful as uncertainty measures when assessing the trustworthiness of a model's performance on individual data points. Also, these measures likely would not perform well on data that contain examples that are "out of distribution," which refers to examples that are very different from the training examples (eg, lung cancer when the training data are for pneumonia, or worse, when a CT image is given and the training data were all radiographs).

UQ versus Radiologist Uncertainty

A radiologist typically reports their findings with some adjectives that reflect the subjective level of uncertainty. In contrast, UQ produces an objective measure of uncertainty. Some reporting tools, such as the Breast Imaging Reporting and Data System

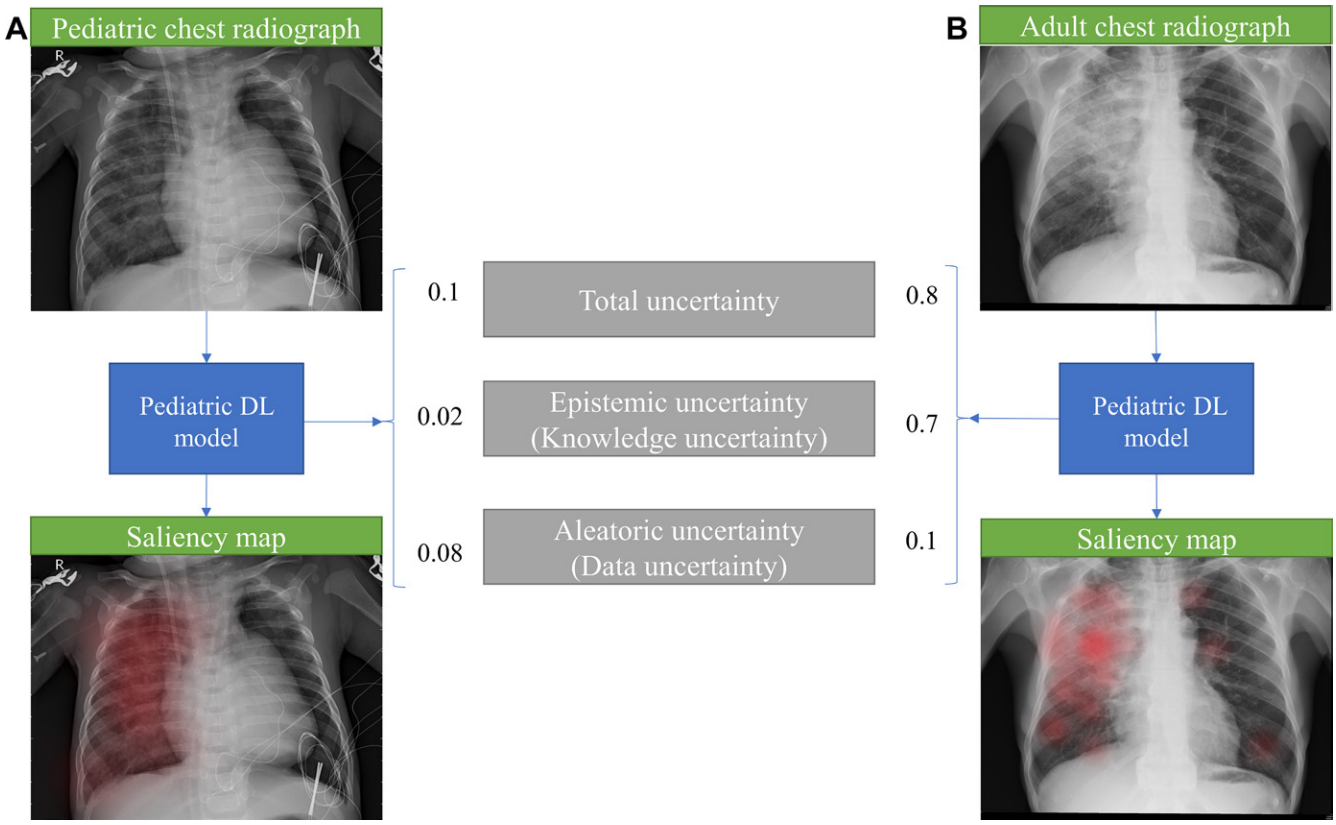

Figure 5: Schematic illustrations show radiographic examples of how uncertainty quantification reveals model bias that is not detectable using interpretation maps (ie, saliency maps). **(A)** A deep learning (DL) model trained on pediatric chest radiographs for pneumonia detection outputs low total uncertainty (0.1) and correctly identifies the affected area on a pediatric chest radiograph. **(B)** The same DL model outputs high total uncertainty (0.8), particularly high epistemic (knowledge) uncertainty (0.7), that indicates model bias when presented with an adult chest radiograph not seen during training. However, because the saliency map correctly identifies the affected area in the radiograph, it is not able to identify bias.

(BI-RADS), use adjectives to describe probabilities. For instance, classification of a probably benign finding (BI-RADS 3) is defined as less than 2% probability of malignancy (34). However, there is less uncertainty in classification of BI-RADS 3 reported by a breast radiologist than that by a trainee radiologist.

Challenges and Future Directions in Performing UQ

To have a trustworthy DL model, the uncertainty level for each output (prediction) must be quantified. Still, the implementation of a DL model that reports uncertainty for its outputs can be a challenging task for the following reasons. First, it is a math-heavy concept and implementation of each method has its own complexity. Second, there are many techniques that have not been compared with each other. Third, the absolute uncertainty value for each prediction depends on the training data set. In other words, the absolute uncertainty value does not necessarily translate into the same meaning for different models, or even the same models with different training sets. Thus, easy-to-use UQ tools would be helpful.

Despite the growing number of publications about UQ, there are no easily applied UQ tools for DL radiologic image analysis that report the uncertainty level for each output, assess model calibration, and perform model calibration if not calibrated. The lack of user-friendly UQ tools poses unique challenges to their adoption. To mitigate this predicament, and to develop a framework for routine inclusion of UQ into artificial intelligence products, some issues need to be addressed.

First, unlike other DL medical imaging analysis tasks (eg, classification, segmentation), there is lack of an established reference standard for the assessment of UQ methods, such as a benchmark data set with known levels of aleatoric and epistemic uncertainty. An appropriate set of medical images for applying UQ should have a known and controllable level of uncertainty in the data and their labels. For many years, the number of "artificial intelligence–ready" annotated medical imaging data sets lagged behind the demand from the machine learning community (35).

Second, very few studies (ie, interoperability studies) map the relationship between the reported uncertainty of radiologists and objective UQ of DL models. A related problem is that terms often used in radiology reports, such as "consistent with" or "likely to represent," have different certainty levels between radiologists.

Finally, although UQ methods have been applied to natural images, such as animals, vehicles, and fruits (36,37), there is a lack of studies that compare the performance of different UQ methods in medical images (ie, radiologic image research). This is a critical shortcoming, as medical images have unique characteristics (eg, the number of channels, dimensions, resolutions,

Table 2: Summary of Challenges in Performing UQ and Possible Solutions

Challenge	Solution
Uncertainty is an underinvestigated but essential component in predictive models	*(a)* Educational papers on the topic *(b)* Devoting sessions in conferences to the topic *(c)* Journals that publish medical imaging DL; ask authors to report a measure of uncertainty along with other performance metrics *(d)* Having medical imaging DL competitions on UQ
Lack of benchmark data sets for UQ	Creating publicly available data sets with engineered levels of uncertainty
Lack of comparison across UQ methods	Testing different UQ methods on benchmark data sets
Requires strong mathematical and statistical background	Developing user-friendly programming packages for developers to easily add UQ methods to their current pipeline
Most UQ methods provide a measure of uncertainty without any clinical correspondence	Perform sensitivity analysis on each UQ-enabled model to find a threshold that can be translated into clinical workflow; in this way, if a prediction exceeds a threshold, it is flagged as suspicious

Note.—DL = deep learning, UQ = uncertainty quantification.

and features) that make them fundamentally different from natural images (38,39).

Without reliable radiologic UQ data sets, a trained DL model will remain a black box, making the performance comparison of different UQ and model calibration methods even more challenging. A model's uncertainty output needs to be understandable, which can be accomplished in several ways. For example, a threshold can be determined based on sensitivity analysis so that if a model produces a prediction with an uncertainty below a certain threshold, the result can be trusted. However, if a model prediction has uncertainty above the threshold, the radiologist should carefully consider whether the prediction is correct. An alternative approach would be to map the numerical uncertainty values to descriptive adjectives. Table 2 summarizes the challenges in performing UQ and potential solutions.

Conclusion

To have trustworthy and unbiased deep learning (DL) models for real-world medical imaging, calibrated models that can output the uncertainty of each prediction are needed. For a DL model to be useful, understanding both its level of uncertainty and its calibration is essential for trustworthy and reliable predictions. uncertainty quantification (UQ) may also assist calibration. Thus, it is imperative that radiologists understand uncertainty and how it can be measured and quantified to inform appropriate clinical decision-making. By unveiling the recent trends of UQ in DL for radiologic imaging analysis, a conceptual framework for understanding and taxonomy of UQ for radiologists has been provided. Next, it is necessary to develop publicly available annotated data sets with known levels of aleatoric and epistemic uncertainty that may be used by researchers or vendors as benchmarks for the assessment of UQ methods.

Disclosures of conflicts of interest: S.F. *Radiology: Artificial Intelligence* trainee editorial board member. **M.M.** No relevant relationships. **P.R.** *RadioGraphics* trainee editorial board member. **B.K.** *Radiology: Artificial Intelligence* trainee editorial board member. **F.I.B.** No relevant relationships. **M.D.R.** No relevant relationships. **B.J.E.** Co-chair of Society for Imaging Informatics in Medicine Research Committee; consultant to the editor for *Radiology: Artificial Intelligence*.

References

1. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine Learning for Medical Imaging. RadioGraphics 2017;37(2):505–515.
2. Moassefi M, Faghani S, Conte GM, et al. A deep learning model for discriminating true progression from pseudoprogression in glioblastoma patients. J Neurooncol 2022;159(2):447–455.
3. Faghani S, Baffour FI, Ringler MD, et al. A deep learning algorithm for detecting lytic bone lesions of multiple myeloma on CT. Skeletal Radiol 2023;52(1):91–98.
4. Rouzrokh P, Wyles CC, Philbrick KA, et al. A Deep Learning Tool for Automated Radiographic Measurement of Acetabular Component Inclination and Version After Total Hip Arthroplasty. J Arthroplasty 2021;36(7):2510–2517.e6.
5. Zhang K, Hu H, Philbrick K, et al. SOUP-GAN: Super-Resolution MRI Using Generative Adversarial Networks. Tomography 2022;8(2):905–919.
6. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med 2019;25(9):1337–1340. [Published correction appears in Nat Med 2019;25(10):1627.]
7. Rouzrokh P, Khosravi B, Faghani S, et al. Mitigating Bias in Radiology Machine Learning: 1. Data Handling. Radiol Artif Intell 2022;4(5):e210290.
8. Hasani N, Morris MA, Rhamim A, et al. Trustworthy Artificial Intelligence in Medical Imaging. PET Clin 2022;17(1):1–12.
9. Chokshi FH, Flanders AE, Prevedello LM, Langlotz CP. Fostering a Healthy AI Ecosystem for Radiology: Conclusions of the 2018 RSNA Summit on AI in Radiology. Radiol Artif Intell 2019;1(2):190021.
10. McCrindle B, Zukotynski K, Doyle TE, Noseworthy MD. A Radiology-focused Review of Predictive Uncertainty for AI Interpretability in Computer-assisted Segmentation. Radiol Artif Intell 2021;3(6):e210031.
11. Abdar M, Pourpanah F, Hussain S, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Inf Fusion 2021;76:243–297.
12. Thiagarajan JJ, Thopalli K, Rajan D, Turaga P. Training calibration-based counterfactual explainers for deep learning models in medical image analysis. Sci Rep 2022;12(1):597.
13. Zabihollahy F, Viswanathan AN, Schmidt EJ, Lee J. Fully automated segmentation of clinical target volume in cervical cancer from magnetic resonance imaging with convolutional neural network. J Appl Clin Med Phys 2022;23(9):e13725.
14. Lee J, Shin D, Oh SH, Kim H. Method to Minimize the Errors of AI: Quantifying and Exploiting Uncertainty of Deep Learning in Brain Tumor Segmentation. Sensors (Basel) 2022;22(6):2406.
15. Rajaraman S, Zamzmi G, Yang F, Xue Z, Jaeger S, Antani SK. Uncertainty Quantification in Segmenting Tuberculosis-Consistent Findings in Frontal Chest X-rays. Biomedicines 2022;10(6):1323.
16. Adams J, Bhalodia R, Elhabian S. Uncertain-DeepSSM: From Images to Probabilistic Shape Models. Shape Med Imaging (2020) 2020;12474:57–72.

17. Guo C, Pleiss G, Sun Y, Weinberger KQ. On Calibration of Modern Neural Networks. In: Precup D, Teh YW, eds. Proceedings of the 34th International Conference on Machine Learning. PMLR; 06–11 Aug 2017; 1321–1330.
18. Henne M, Schwaiger A, Roscher K, Weiss G. Benchmarking Uncertainty Estimation Methods for Deep Learning With Safety-Related Metrics. SafeAI@AAAI. https://www.semanticscholar.org/paper/e22cc1995f69e8ef4afe451c6fea3649884c2992. Published 2020. Accessed August 28, 2022.
19. Rajaraman S, Ganesan P, Antani S. Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks. PLoS One 2022;17(1):e0262838.
20. Vuk M, Curk T. ROC curve, lift chart and calibration plot. Adv Methodol Stat 2006;3(1):89–108.
21. Faghani S, Khosravi B, Zhang K, et al. Mitigating Bias in Radiology Machine Learning: 3. Performance Metrics. Radiol Artif Intell 2022;4(5): e220061.
22. Lakshminarayanan B, Pritzel A, Blundell C. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. arXiv:1612.01474 [preprint]. https://arxiv.org/abs/1612.01474. Posted December 5, 2016. Accessed August 1, 2022.
23. Gal Y, Ghahramani Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In: Balcan MF, Weinberger KQ, eds. Proceedings of the 33rd International Conference on Machine Learning. New York, NY: PMLR; 20–22 Jun 2016; 1050–1059.
24. Sensoy M, Kaplan L, Kandemir M. Evidential deep learning to quantify classification uncertainty. Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates, 2018; 3183–3193.
25. Khosravi B, Faghani S, Ashraf-Ganjouei A. Uncertainty quantification in COVID-19 detection using evidential deep learning. bioRxiv. https://doi.org/10.1101/2022.05.29.22275732. Posted May 29, 2022. Accessed August 1, 2022.
26. Dohopolski M, Chen L, Sher D, Wang J. Predicting lymph node metastasis in patients with oropharyngeal cancer by using a convolutional neural network with associated epistemic and aleatoric uncertainty. Phys Med Biol 2020;65(22):225002.
27. Laves M-H, Tölle M, Schlaefer A, Engelhardt S. Posterior temperature optimized Bayesian models for inverse problems in medical imaging. Med Image Anal 2022;78:102382.
28. Khawaled S, Freiman M. NPBDREG: Uncertainty assessment in diffeomorphic brain MRI registration using a non-parametric Bayesian deep-learning based approach. Comput Med Imaging Graph 2022;99:102087.
29. Ozdemir O, Russell RL, Berlin AA. A 3D Probabilistic Deep Learning System for Detection and Diagnosis of Lung Cancer Using Low-Dose CT Scans. IEEE Trans Med Imaging 2020;39(5):1419–1429.
30. Pianykh OS, Langs G, Dewey M, et al. Continuous Learning AI in Radiology: Implementation Principles and Early Applications. Radiology 2020;297(1):6–14.
31. Hemmer P, Kühl N, Schöffer J. DEAL: Deep Evidential Active Learning for Image Classification. In: Wani MA, Raj B, Luo F, Dou D, eds. Deep Learning Applications, Vol 3. Singapore: Springer Singapore, 2022; 171–192.
32. Liu S, Liang S, Huang X, Yuan X, Zhong T, Zhang Y. Graph-enhanced U-Net for semi-supervised segmentation of pancreas from abdomen CT scan. Phys Med Biol 2022;67(15):155017.
33. Tejani AS, Retson TA, Moy L, Cook TS. Detecting Common Sources of AI Bias: Questions to Ask When Procuring an AI Solution. Radiology 2023. 10.1148/radiol.230580. Published online March 21, 2023.
34. Fowler EE, Sellers TA, Lu B, Heine JJ. Breast Imaging Reporting and Data System (BI-RADS) breast composition descriptors: automated measurement development for full field digital mammography. Med Phys 2013;40(11):113502.
35. Shad R, Cunningham JP, Ashley EA, Langlotz CP, Hiesinger W. Designing clinically translatable artificial intelligence systems for high-dimensional medical imaging. Nat Mach Intell 2021;3(11):929–935.
36. Caldeira J, Nord B. Deeply Uncertain: Comparing Methods of Uncertainty Quantification in Deep Learning Algorithms. arXiv:2004.10710 [preprint]. https://arxiv.org/abs/2004.10710. Posted April 22, 2020. Accessed August 1, 2022.
37. Psaros AF, Meng X, Zou Z, Guo L, Karniadakis GE. Uncertainty Quantification in Scientific Machine Learning: Methods, Metrics, and Comparisons. arXiv:2201.07766 [preprint]. https://arxiv.org/abs/2201.07766. Posted January 19, 2022. Accessed August 1, 2022.
38. Morra L, Piano L, Lamberti F, Tommasi T. Bridging the gap between Natural and Medical Images through Deep Colorization. arXiv:2005.10589 [preprint]. https://arxiv.org/abs/2005.10589. Posted May 21, 2020. Accessed August 1, 2022.
39. Singh SP, Wang L, Gupta S, Goli H, Padmanabhan P, Gulyás B. 3D Deep Learning on Medical Images: A Review. Sensors (Basel) 2020;20(18):5097.