Editorial | Uncover This Tech Term

eISSN 2005-8330
https://doi.org/10.3348/kjr.2024.0108
Korean J Radiol 2024;25(4):395-398

Korean Journal of Radiology
KJR

Uncover This Tech Term: Uncertainty Quantification for Deep Learning

Shahriar Faghani, Cooper Gamble, Bradley J. Erickson

Artificial Intelligence Laboratory, Department of Radiology, Mayo Clinic, Rochester, MN, USA

See the invited Editorial "Caveats in Using Abnormality/Probability Scores from Artificial Intelligence Algorithms: Neither True Probability nor Level of Trustworthiness" at https://doi.org/10.3348/kjr.2024.0144.

Keywords: Uncertainty quantification; Deep learning; Radiology; Calibration

What is Uncertainty Quantification?

Deep learning (DL) has been recognized for its potential in radiology, yet concerns regarding its reliability in clinical workflows limit its adoption. This has become a greater challenge in radiology societies following the recent awareness of hallucinations in large language models [1]. These arise from predictions made without an estimate of the trustworthiness of DL models [2]. DL models, including computer vision and language models, generate numerical outputs that resemble probability. However, these outputs are used for training the model and are not indicative of the actual likelihood of a specific outcome, because they lack calibration [3]. Consequently, if a model outputs a value of 0.8 for a particular diagnosis, it does not necessarily imply that there is an 80% chance of the diagnosis being correct. Although methods exist to calibrate these probability-like model outputs either during or after training, such calibration alone does not address the underlying uncertainty of the model. This is because each output represents a point estimate from a distribution, and it is the spread of this distribution that reflects the level of uncertainty [2]. As a result, even when two predictions have the same calibrated probabilities, they can still exhibit different degrees of uncertainty, as shown in Figure 1. Additionally, in the medical literature, uncertainty is usually conveyed through intervals derived from population data rather than from individual samples [2]. These underscore the necessity for techniques that offer insights into a model's uncertainty for each prediction, going beyond mere calibration. To this end, we will explore the key categories of uncertainty quantification (UQ) methods.

Types of Uncertainty Quantification

Frequentist Approaches

Frequentist approaches focus on data distribution within a target population without incorporating prior beliefs or subjective probabilities. Conformal prediction (CP), the most prominent method in this category and often referred to as distribution-free UQ, creates trustworthy prediction sets for each prediction with a statistical guarantee that such sets contain the ground truth at a user-specified error rate [4]. When it comes to suggesting a diagnosis, CP provides a trustworthy differential diagnosis list that guarantees the correct answer is included in the list as opposed to a single diagnosis. CP uses a '*calibration dataset*,' which is a subset of the target population, to capture the model's uncertainty based on the target population. CP is based on the principle of conformality, which gauges the '*conformity*' of new data

Received: January 29, 2024 **Revised:** February 5, 2024
Accepted: February 6, 2024
Corresponding author: Bradley J. Erickson, MD, PhD, Artificial Intelligence Laboratory, Department of Radiology, Mayo Clinic, 200 1st St. SW, Rochester, MN 55905, USA
• E-mail: E-mail: bje@mayo.edu

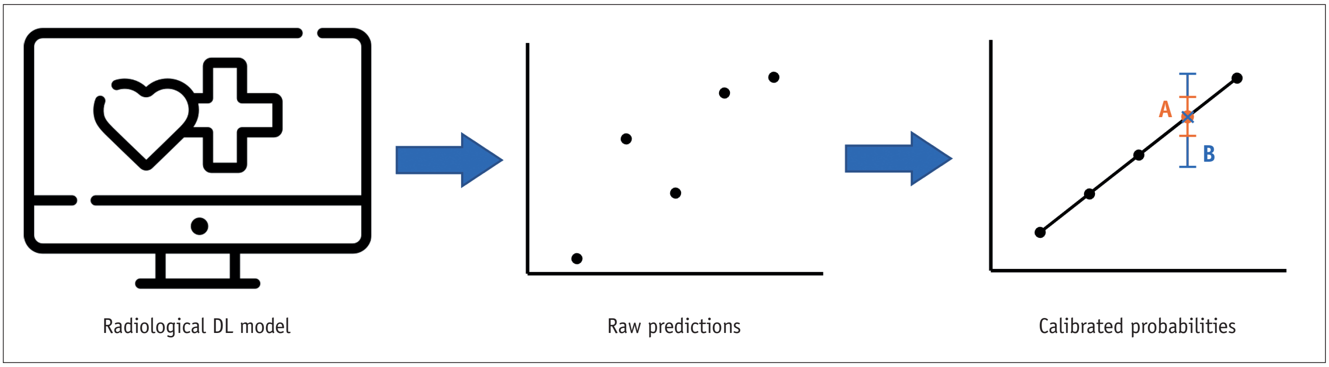

Fig. 1. Example of a case in which two predictions with identical calibrated probabilities may have different uncertainty levels due to the varying spread of their respective distributions. Two predictions, A (orange dot) and B (blue X mark), have the same calibrated probability but have different uncertainty levels represented by the respective whiskers.

points to the learned patterns in the training set. This approach ensures the existence of a correct diagnosis in the differential diagnosis list, irrespective of the underlying data distribution. Essentially, the '*calibration dataset*' is used to refine the differential diagnosis list, which may expand as each sample deviates more from the training data distribution. One notable disadvantage of CP is the requirement of an additional data subset, exacerbating the challenge of data scarcity in radiology DL. This is especially true for rare diagnoses or underrepresented groups, as setting aside data for calibration can further constrain the already limited training dataset.

Probabilistic Methods

Probabilistic UQ assigns different values to the model parameters or variables based on a prior (assumption).

Ensemble Method

The ensemble method relies on the concept that the level of disagreement among various model outputs for a sample indicates the uncertainty associated with that sample [5]. As an analogy, consider a panel of radiologists who read studies. If half of the radiologists identified a pathology but the other half did not, the study would be considered uncertain. However, unanimous agreement among the readers indicates low uncertainty. The uncertainty value is derived using a measure of spread (e.g., variance, range, etc.) across the predictions made by various models. A significant limitation of this method is its high computational cost. Additionally, the final measure of spread only indicates a correlation with the uncertainty of the prediction rather than providing a direct interpretation of uncertainty.

Bayesian Methods

Bayesian methods run a model several times on an input while slightly changing the model parameters, with each inference assessing the spread of the prediction [6]. This is akin to a radiologist reviewing the same study multiple times under different circumstances. If the radiologist consistently finds the same result in different scenarios, the case is considered certain, and varying conclusions indicate uncertainty. Monte Carlo Dropout, a popular approach in this category, uses a single model and transforms it at inference time by randomly deactivating nodes based on probabilities specified '*a priori*' [7]. This creates multiple unique models. Although straightforward to apply, this approach requires users to specify the probability of deactivating nodes (known prior distribution), and similar to the ensemble method, it can only indicate a correlation with the uncertainty of the prediction.

Evidential Deep Learning

Evidential DL (EDL) methods gather category-specific features from images as "evidence" to determine prediction certainty [8,9]. More evidence leads to higher confidence in the prediction. While EDL is less resource-demanding and has a robust theoretical foundation, its outputs lack statistical guarantees and require conversion into measures that are comprehensible to humans, a process that presents significant challenges.

Applications in Radiology

UQ can be used in various DL applications, including classification, detection, segmentation, and generation. In classification and detection, UQ enables models to not

Table 1. Additional use cases for UQ in radiological deep learning

Application	Definition	Example
Active learning	UQ can be employed to select uncertain samples from the training dataset and continuously pass them through the model to force it to learn the features of those samples and improve its performance on uncertain samples overall	Hemmer et al. (2022) [13] used uncertainty values for active learning in pneumonia detection on chest X-rays images
Out-of-domain detection	UQ can help detect out-of-domain samples among a given test dataset because samples that are further away from a model's training domain will have higher uncertainty, so a threshold or selection process can be used to identify samples which may be out-of-domain based only on their uncertainty values	Lakara and Valdenegro-Toro (2022) [14] showed that uncertainty can be used to detect out-of-domain inputs
Bias detection	UQ can identify potential biases in a model because if a subpopulation of the data repeatedly yields high uncertainty values, this could be due to bias in the model	Faghani et al. (2022) [15] explain how UQ can highlight model biases
Data drift/shift detection	UQ can monitor the uncertainty associated with model predictions as the model is exposed to new data over time. An increase in prediction uncertainty might indicate that the model is facing data points that are significantly different from what it was trained on, suggesting potential data drift	Baier et al. (2021) [16] employed an uncertainty-based approach to detect data drift in neural networks

UQ = uncertainty quantification

only localize and detect findings but also offer differential diagnoses ranked by confidence levels [10,11]. This serves as a "safety net," allowing radiologists to concentrate on cases with high uncertainty or a broad differential, thus improving the diagnostic accuracy and patient outcomes. In segmentation, UQ highlights areas of low confidence within the segmentation map. For instance, in the DL segmentation of glioblastoma tumor areas for radiotherapy planning [12], this feature enables experts to review and possibly revise uncertain regions before finalizing treatment plans. Finally, UQ assists in ensuring the accuracy of generative models in radiology, allowing radiologists to determine the reliability of details in a synthetic image and identify trustworthy regions, thereby avoiding the introduction of artificial anomalies that could mimic pathological findings. Table 1 summarizes the other main use cases of UQ in radiology [13-16].

CONCLUSION

The potential of UQ to improve the reliability and trustworthiness of radiology DL applications is evident. However, integrating UQ into clinical settings involves navigating extensive research, regulatory approvals, and practical considerations [17,18]. UQ will likely become an essential component of medical DL tools because it provides insight into the certainty of predictions, which is vital for patient care given the complexity of understanding DL decision-making. Stakeholders should remain informed of these developments and actively participate in the dialogue and experimentation that will shape the future of UQ in medical artificial intelligence applications.

Conflicts of Interest

The authors have no potential conflicts of interest to disclose.

Author Contributions

Conceptualization: all authors. Supervision: Bradley J. Erickson. Writing—original draft: Shahriar Faghani, Cooper Gamble. Writing—review & editing: all authors.

ORCID IDs

Shahriar Faghani
https://orcid.org/0000-0003-3275-2971
Cooper Gamble
https://orcid.org/0009-0009-5139-4875
Bradley J. Erickson
https://orcid.org/0000-0001-7926-6095

Funding Statement

None

REFERENCES

1. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, et al. ChatGPT and other large language models are double-edged swords. *Radiology* 2023;307:e230163
2. Faghani S, Moassefi M, Rouzrokh P, Khosravi B, Baffour FI, Ringler MD, et al. Quantifying uncertainty in deep learning of radiologic images. *Radiology* 2023;308:e222217
3. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks [accessed on January 1, 2024]. Available at: http://proceedings.mlr.press/v70/guo17a.html
4. Angelopoulos AN, Bates S. Conformal prediction: a gentle introduction. *Found Trends Mach Learn* 2023;16:494-591
5. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles [accessed on January 1, 2024]. Available at: https://proceedings.neurips.cc/paper_files/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html
6. Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, et al. A review of uncertainty quantification in deep learning: techniques, applications and challenges. *Inf Fusion* 2021;76:243-297
7. Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning [accessed on January 1, 2024]. Available at: https://proceedings.mlr.press/v48/gal16.html?trk=public_post_comment-text
8. Sensoy M, Kaplan L, Kandemir M. Evidential deep learning to quantify classification uncertainty [accessed on January 1, 2024]. Available at: https://proceedings.neurips.cc/paper/2018/hash/a981f2b708044d6fb4a71a1463242520-Abstract.html
9. Khosravi B, Faghani S, Ashraf-Ganjouei A. Uncertainty quantification in COVID-19 detection using evidential deep learning. medRxiv [Preprint]. 2022 [accessed on January 1, 2024]. Available at: https://doi.org/10.1101/2022.05.29.22275732
10. Gamble C, Faghani S, Erickson BJ. Toward clinically trustworthy deep learning: applying conformal prediction to intracranial hemorrhage detection. arXiv [Preprint]. 2024 [accessed on January 1, 2024]. Available at: https://doi.org/10.48550/arXiv.2401.08058
11. Alves N, Bosma JS, Venkadesh KV, Jacobs C, Saghir Z, de Rooij M, et al. Prediction variability to identify reduced AI performance in cancer diagnosis at MRI and CT. *Radiology* 2023;308:e230275
12. 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:e210031
13. 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, volume 3. Singapore: Springer, 2022:171-192
14. Lakara K, Valdenegro-Toro M. Disentangled uncertainty and out of distribution detection in medical generative models. arXiv [Preprint]. 2022 [accessed on January 1, 2024]. Available at: https://doi.org/10.48550/arXiv.2211.06250
15. Faghani S, Khosravi B, Zhang K, Moassefi M, Jagtap JM, Nugen F, et al. Mitigating bias in radiology machine learning: 3. Performance metrics. *Radiol Artif Intell* 2022;4:e220061
16. Baier L, Schlör T, Schöffer J, Kühl N. Detecting concept drift with neural network model uncertainty. arXiv [Preprint]. 2021 [accessed on January 1, 2024]. Available at: https://doi.org/10.48550/arXiv.2107.01873
17. Zhang K, Khosravi B, Vahdati S, Erickson BJ. FDA review of radiologic AI algorithms: process and challenges. *Radiology* 2024;310:e230242
18. Brady AP, Allen B, Chong J, Kotter E, Kottler N, Mongan J, et al. Developing, purchasing, implementing and monitoring AI tools in radiology: practical considerations. A multi-society statement from the ACR, CAR, ESR, RANZCR and RSNA. *Radiol Artif Intell* 2024;6:e230513