Physics and Imaging in Radiation Oncology 29 (2024) 100526

Contents lists available at ScienceDirect

Physics and Imaging in Radiation Oncology

journal homepage: www.sciencedirect.com/journal/physics-and-imaging-in-radiation-oncology

Correspondence

Harnessing uncertainty in radiotherapy auto-segmentation quality assurance

We have read with great interest the article by Outeiral et al. [1], in which the authors propose a simple metric for optimizing quality assurance in deep learning (DL) auto-segmentation workflows. We commend the authors for their insightful analysis using two meticulously curated MRI datasets. Through this study, the authors have probed into the relatively underexplored yet clinically relevant domain of uncertainty estimation in auto-segmentation.

One of the key contributions of this study is the reappropriation of standard DL outputs as a quality indicator to identify cases that clinicians should review further. The authors achieve this by applying an empirically derived threshold to the softmax output of their DL network, computing the mean of the thresholded score map (termed the HiS metric), and correlating it with standard geometric quality indices. When juxtaposed with a mean entropy — a commonly used measure of model output uncertainty — HiS consistently demonstrated a stronger correlation with the geometric indices, suggesting its superior ability to stratify cases needing additional review. We applaud the authors' efforts for their novel contributions and would like to note some potential caveats that could pave the way for future research directions.

Conventional large DL networks often yield overconfident predictions which can result in poor model calibration [2], meaning the predicted probabilities do not align with the true underlying data. This discrepancy could undermine the reliability of these outputs in detecting out-of-distribution data, a critical aspect of quality assurance systems. Notably, the direct use of softmax outputs as measures of model uncertainty is a point of contention within the DL community [3,4]. However, moderately sized standard DL networks have the potential to exhibit well-calibrated performance [5]. Therefore, it is unclear whether calibration had a major impact on Outeiral et al.'s analysis using a standard nnU-Net architecture. In contrast, Bayesian DL approaches have been observed to be well-calibrated and may circumvent these issues [6]. Specifically, the application of approximate Bayesian techniques, such as Monte Carlo dropout [7] or deep ensembles [8] (Fig. 1), is relatively simple compared to conventional solutions. While these methods demand a slightly higher computational cost, they could be considered for investigating HiS in future studies. Importantly, ensembling (e.g., through cross-validation schemes) is becoming increasingly common for many DL solutions [9]. We have previously benchmarked ensembling under a U-net framework for uncertainty estimation in oropharyngeal cancer auto-segmentation and have shown its efficacy [10]. Interestingly, Outeiral et al. use cross-validation within their study for robustness analysis; merging their cross-validation outputs into an ensemble could have improved calibration when employing their HiS metric. Of note, alternative methods that allow for calibrated uncertainty estimates, such as conformal prediction [11], may also show promise for auto-segmentation and should be further investigated.

Finally, we would like to note that the proposed HiS metric, if used to measure uncertainty, may be unable to disentangle epistemic uncertainty (i.e., intrinsic model uncertainty) and aleatoric uncertainty (i.e., extrinsic statistical uncertainty) [12]. While the same can be said of general measures of entropy, there exist alternative entropy-related uncertainty metrics, like expected entropy and mutual information, that could distinguish the source of the uncertainty when combined with an approximate Bayesian approach [10,13]. Moreover, when the distribution of DL network parameters is assumed to be a delta distribution, e.g., in a conventional DL network, the epistemic uncertainty is implicitly assumed to be non-existent. Therefore, depending on the specific auto-segmentation use case, alternative uncertainty metrics, or combinations of uncertainty metrics, may be more suitable.

An increasing number of studies have begun to apply uncertainty estimation to the quality assurance of radiotherapy-related auto-segmentation [10,14,15–23]. The study by Outeiral et al. serves as a cornerstone contribution to this crucial literature. We eagerly anticipate further advances in this clinically significant field of work.

1. Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT (GPT-4 architecture; ChatGPT October 17, 2023 Version) to improve the grammatical accuracy and semantic structure of portions of the text. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

2. Funding Acknowledgements

Kareem A. Wahid is supported by the NCI NRSA Image Guided Cancer Therapy Training Program (T32CA261856). The work of Joel Jaskari, Jaakko Sahlsten, and Kimmo K. Kaski was supported in part by the Academy of Finland under Project 345449. Antti Mäkitie is supported in part by a grant from the Finnish Society of Sciences and Letters. Benjamin H. Kann is supported by an NIH/National Institute for Dental and Craniofacial Research (NIDCR) K08 Grant (K08DE030216). Clifton D. Fuller receives related grant support from the NCI NRSA Image Guided Cancer Therapy Training Program (T32CA261856), as well as additional unrelated salary/effort support from NIH institutes. Clifton D. Fuller also receives grant and infrastructure support from MD Anderson Cancer Center via: the Charles and Daneen Stiefel Center for Head and Neck Cancer Oropharyngeal Cancer Research Program; the Program in Image-guided Cancer Therapy; and the NIH/NCI Cancer

DOI of original article: https://doi.org/10.1016/j.phro.2023.100500.

https://doi.org/10.1016/j.phro.2023.100526
Received 5 November 2023; Accepted 13 December 2023
Available online 19 December 2023

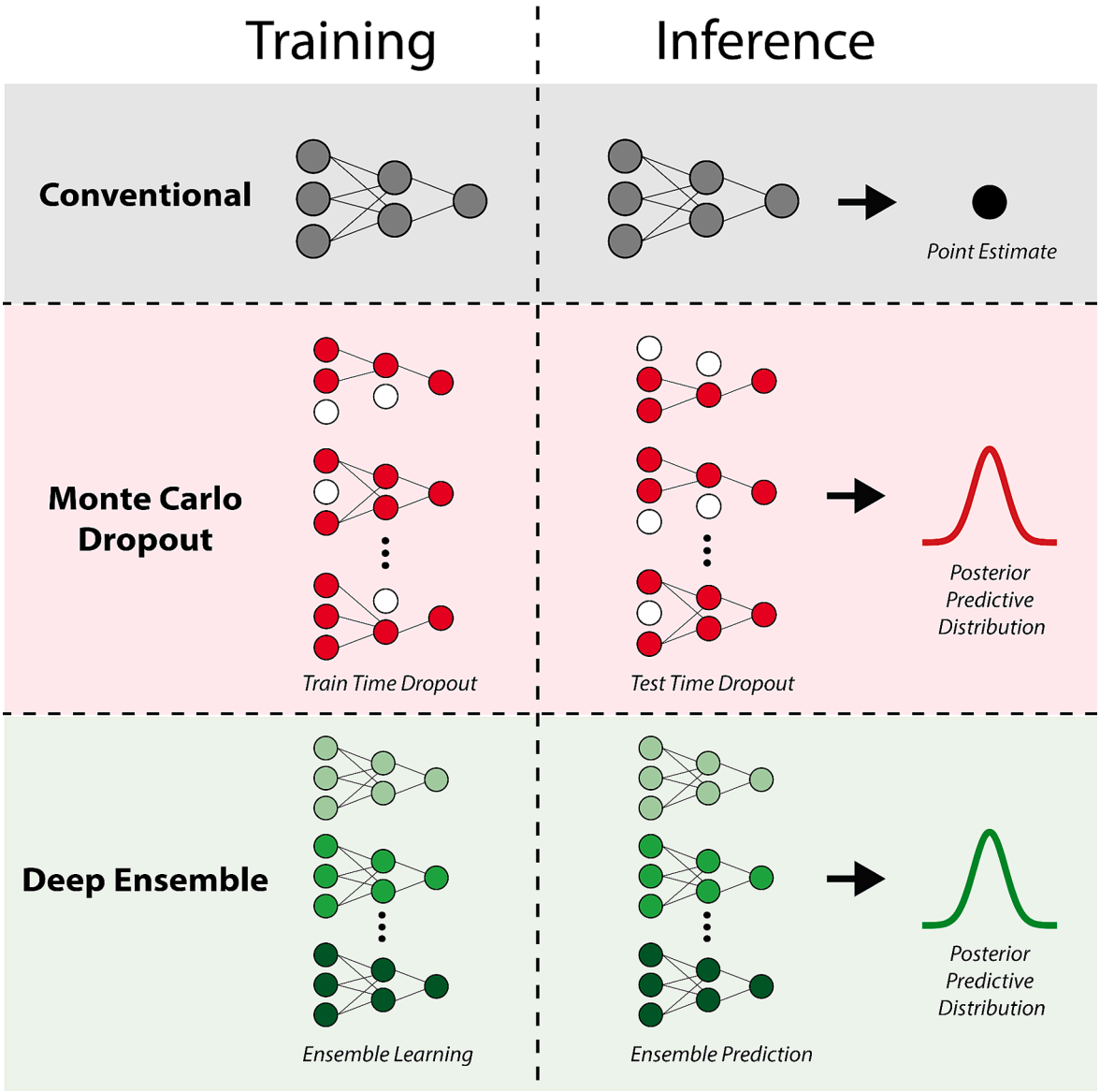

Fig. 1. Comparison between conventional and approximate Bayesian deep learning approaches. Conventional deep learning methods generate point estimates and are often poorly calibrated, while approximate Bayesian methods, e.g., Monte Carlo dropout and deep ensemble, generate posterior predictive distributions that are often better calibrated. The Monte Carlo dropout method consists of randomly removing nodes from the network during the training and inference procedures. The deep ensemble method trains submodels with different random initializations of network parameters and, optionally, varying subsets of training data, then combines their predictions. This figure is loosely inspired by figures from van den Berg and Meliadò [14].

Center Support Grant (CCSG) Radiation Oncology and Cancer Imaging Program (P30CA016672). David Fuentes was supported by R01CA195524.

Declaration of competing interest

Clifton D. Fuller has received unrelated direct industry grant/in-kind support, honoraria, and travel funding from Elekta AB.

References

[1] Rodríguez Outeiral R, Ferreira Silvério N, González PJ, Schaake EE, Janssen T, van der Heide UA, et al. A network score-based metric to optimize the quality assurance of automatic radiotherapy target segmentations. Phys Imaging Radiat Oncol 2023; 28:100500.

[2] Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: Precup D, Teh YW, editors. Proceedings of the 34th international conference on machine learning, vol. 70, PMLR; 06–11 Aug 2017, p. 1321–30.

[3] Holm AN, Wright D, Augenstein I. Revisiting softmax for uncertainty approximation in text classification. Information 2023;14:420.

[4] Pearce T, Brintrup A, Zhu J. Understanding softmax confidence and uncertainty. arXiv [csLG] 2021.

[5] Jaskari J, Sahlsten J, Damoulas T, Knoblauch J, Särkkä S, Kärkkäinen L, et al. Uncertainty-aware deep learning methods for robust diabetic retinopathy classification. IEEE Access 2022;10:76669–81.

[6] Izmailov P, Vikram S, Hoffman MD, Wilson AGG. What are Bayesian neural network posteriors really like? In: Meila M, Zhang T, editors. Proceedings of the 38th international conference on machine learning, vol. 139, PMLR; 18–24 Jul 2021, p. 4629–40.

[7] Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: Balcan MF, Weinberger KQ, editors. Proceedings of The 33rd international conference on machine learning, vol. 48, New York, New York, USA: PMLR; 20–22 Jun 2016, p. 1050–9.

[8] Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. Adv Neural Inf Process Syst 2017;30.

[9] Eisenmann M, Reinke A, Weru V, Tizabi MD, Isensee F, Adler TJ, et al. Why is the winner the best? 2023 IEEE/CVF Conference on computer vision and pattern recognition (CVPR), IEEE; 2023, p. 19955–66.

[10] Sahlsten J, Jaskari J, Wahid KA, Ahmed S, Glerean E, He R, et al. Application of simultaneous uncertainty quantification for image segmentation with probabilistic deep learning: Performance benchmarking of oropharyngeal cancer target

delineation as a use-case. medRxiv 2023. https://doi.org/10.1101/2023.02.20.23286188.
[11] Angelopoulos AN, Bates S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv [csLG] 2021.
[12] Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach Learn 2021;110:457–506.
[13] Band N, Rudner TGJ, Feng Q, Filos A, Nado Z, Dusenberry MW, et al. Benchmarking Bayesian deep learning on diabetic retinopathy detection tasks. arXiv [statML] 2022.
[14] van den Berg CAT, Meliadò EF. Uncertainty assessment for deep learning radiotherapy applications. Semin Radiat Oncol 2022;32:304–18.
[15] Tang P, Yang P, Nie D, Wu X, Zhou J, Wang Y. Unified medical image segmentation by learning from uncertainty in an end-to-end manner. Knowl-Based Syst 2022;241:108215.
[16] De Biase A, Sijtsema NM, van Dijk L, Langendijk JA, van Ooijen PMA. Deep learning aided oropharyngeal cancer segmentation with adaptive thresholding for predicted tumor probability in FDG PET and CT images. Phys Med Biol 2023. https://doi.org/10.1088/1361-6560/acb9cf.
[17] Balagopal A, Nguyen D, Morgan H, Weng Y, Dohopolski M, Lin M-H, et al. A deep learning-based framework for segmenting invisible clinical target volumes with estimated uncertainties for post-operative prostate cancer radiotherapy. Med Image Anal 2021;72:102101.
[18] Li X, Bagher-Ebadian H, Gardner S, Kim J, Elshaikh M, Movsas B, et al. An uncertainty-aware deep learning architecture with outlier mitigation for prostate gland segmentation in radiotherapy treatment planning. Med Phys 2023;50:311–22.
[19] Bragman FJS, Tanno R, Eaton-Rosen Z, Li W, Hawkes DJ, Ourselin S, et al. Uncertainty in multitask learning: joint representations for probabilistic MR-only radiotherapy planning. Medical image computing and computer assisted intervention – MICCAI 2018, Springer International Publishing; 2018, p. 3–11.
[20] Chen X, Men K, Chen B, Tang Y, Zhang T, Wang S, et al. CNN-based quality assurance for automatic segmentation of breast cancer in radiotherapy. Front Oncol 2020;10:524.
[21] Cubero L, Serrano J, Castelli J, De Crevoisier R, Acosta O, Exploring PJ, et al. IEEE 20th international symposium on biomedical imaging (ISBI). IEEE 2023;2023:1–4.
[22] Min H, Dowling J, Jameson MG, Cloak K, Faustino J, Sidhom M, et al. Clinical target volume delineation quality assurance for MRI-guided prostate radiotherapy using deep learning with uncertainty estimation. Radiother Oncol 2023;186:109794.
[23] Lei W, Mei H, Sun Z, Ye S, Gu R, Wang H, et al. Automatic segmentation of organs-at-risk from head-and-neck CT using separable convolutional neural network with hard-region-weighted loss. Neurocomputing 2021;442:184–99.

Kareem A. Wahid[a,b,*], Jaakko Sahlsten[c], Joel Jaskari[c], Michael J. Dohopolski[d], Kimmo Kaski[c], Renjie He[b], Enrico Glerean[e], Benjamin H. Kann[f], Antti Mäkitie[g], Clifton D. Fuller[b], Mohamed A. Naser[b], David Fuentes[a]

[a] *Department of Imaging Physics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA*
[b] *Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA*
[c] *Department of Computer Science, Aalto University School of Science, Espoo, Finland*
[d] *Department of Radiation Oncology, The University of Texas Southwestern Medical Center, Dallas, TX, USA*
[e] *Department of Neuroscience and Biomedical Engineering, Aalto University School of Science, Espoo, Finland*
[f] *Artificial Intelligence in Medicine Program, Brigham and Women's Hospital, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA*
[g] *Department of Otorhinolaryngology, Head and Neck Surgery, University of Helsinki and Helsinki University Hospital, Research Program in Systems Oncology, University of Helsinki, Helsinki, Finland*

[*] Corresponding author at: Department of Imaging Physics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.
E-mail address: kawahid@mdanderson.org (K.A. Wahid).