

The impact of training dataset size and model size on the accuracy of deep learning dose prediction

Joep van Genderingen¹, Dan Nguyen², Franziska Knuth¹, Hazem Nomer¹, Luca Incrocci¹, Abdul Sharfo¹, Uwe Oelfke³, Steve Jiang², Linda Rossi¹, Ben Heijmen¹, and Sebastiaan Breedveld¹

¹Erasmus MC Cancer Institute, University Medical Center Rotterdam, Department of Radiotherapy, Rotterdam, The Netherlands.

²UT Southwestern Medical Center, Medical Artificial Intelligence and Automation (MAIA) Laboratory, Department of Radiation Oncology, Dallas, USA.

³The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust, Joint Department of Physics, London, United Kingdom.

Abstract Deep learning has demonstrated to be capable of predicting realistic radiotherapy dose distributions. Collecting and curating clinical data is a labour-intensive and time-consuming task, making large, curated datasets scarce for training deep learning dose prediction models. Partly due to the lack of training data, model size has never been fully explored in previous studies. In this study, we investigated how training dataset size and model size impact 3D dose prediction performance for prostate cancer. For a curated dataset consisting of 1250 contoured planning CT-scans, ground truth dose distributions were fully automatically generated for each patient using Erasmus-iCycle, guaranteeing consistent trade-offs. Three HD U-Nets with two, four and six pooling layers were trained from scratch with different sized datasets (50/250/1000 plans) for prediction of dose distributions. Mixed precision was applied to optimize GPU VRAM usage. Prediction accuracy was lowest for the 2-level model, while differences between 4-levels and 6-levels were small. For all model sizes, agreement improved with increasing training dataset size, suggesting that more training patients could further increase prediction accuracy.

1 Introduction

Treatment planning for RT is a fragmented, labor-intensive and time-consuming collaboration between physicians, medical physicist and dosimetrists. There is a growing need for shorter treatment planning times, as cancer incidences and staff shortages are increasing [1], while the RT field is also moving towards (online) adaptive treatment strategies and same-day treatment [2]. Although multi-criterial optimization based automated treatment planning as applied in our center has shown to drastically reduce hands-on treatment planning workload in the clinic while producing high quality plans [3, 4], substantial computational time is required. Deep learning has been proposed to predict the patient-specific 3D dose in the order of seconds [5, 6, 7, 8]. The predicted dose can either be used as a directive for the physician [9] or converted to machine parameters to obtain a deliverable treatment plan [10, 11]. From previous studies it is expected that the dosimetric performance of deep learning dose prediction depends on the size and quality of the dataset used for training [12, 13]. However, no consistent dataset size study has been performed to assess whether larger datasets could improve prediction accuracy. In addition, model size has never been fully explored in previous works.

In this study, we systematically investigated the impact of training dataset size and model size on 3D dose prediction performance for prostate cancer, using a curated dataset of

1250 contoured planning CT-scans. All ground truth (GT) plans used for the analyses were generated with Erasmus-iCycle [14], a system for fully automated multi-criterial plan generation, guaranteeing consistent trade-offs between all treatment objectives. Many studies have demonstrated superiority of Erasmus-iCycle plans compared to manual planning [3, 4, 14, 15, 16, 17].

2 Materials and Methods

2.1 Patient Cohort

Patients with low, intermediate or high-risk prostate cancer without metal hip prostheses were included in this study. The data was combined from two sources: 719 patients were included from the multicenter phase 3 randomized trial on hypofractionation for prostate cancer (HYPRO) [15, 18], which was complemented with 531 recent patients from the Erasmus University Medical Center Rotterdam.

Erasmus-iCycle was used to automatically generate for each patient a 10 MV 23-beam GT IMRT plan, mimicking VMAT dose distributions, using original CT resolutions. The clinically established and applied delineations of the target volumes, rectum, anus, femoral heads, bladder and body were also used in this study. The patient cohort consisted of three groups, based on the risk of involvement of the seminal vesicles. For Group 1 patients, a dose of 78 Gy (PTV high) was prescribed to the prostate. For Group 2, a dose of 78 Gy was prescribed to the prostate and 72.2 Gy (PTV low) to the seminal vesicles. In Group 3 patients, prescribed dose for both prostate and seminal vesicles was 78 Gy.

2.2 Experimental setup for Deep Learning

The CT scan, binary masks of the structure delineations and GT dose distributions were downsampled to a grid size of 256 x 256 x 128, reflecting a patient resolution 1.95 x 1.52 x (1.5-3) mm³, where the original slice distance was conserved. The data was divided into 1000 training, 100 validation, and 150 test patients using stratified sampling. Stratification was based on the patient Groups (1-3), size of the patient in the isocenter slice, and the volume of the high

dose target. For training, target volumes were combined into one target map where each voxel value was set to the corresponding prescription dose. Generated models had seven input channels: the CT scan, target map, and binary masks of the anus, bladder, rectum, combined femoral heads and patient contour. The Hounsfield units in the CT were clipped to a range of $[-1000, 1500]$ to improve training convergence, following Liu et al.[19]. All model inputs were left unscaled for training.

The dose prediction models for the investigated dataset sizes and model sizes were based on the Hierarchically Densely Connected U-Net (HD U-Net) architecture proposed by Nguyen et al. [20]. We extended the HD U-Net with a mixed precision learning strategy to optimize VRAM usage. Three HD U-Nets with two, four and six pooling layers were separately trained on randomly stratified subsets of the training data, $N=50/250/1000$ (5/25/100%), resulting in nine trained models. All model training was performed with the Adam optimizer with a learning rate of 10^{-4} , batch size of 1 and mean squared error (MSE) as the loss function. We trained each experiment with four different patience parameters (5, 10, 15 and 20) for the reduce on plateau learning rate scheduler, and early stopping with a threshold of 25 epochs was used to reduce overfitting. The model with the lowest validation loss was used to assess the test data. Training of all models was performed on a cluster with 15 NVIDIA A40 GPU cards with 48 GiB of VRAM.

Predicted dose distributions were interpolated back to their original CT resolutions. All models were evaluated on the same test set without normalization of doses. The performance of the models was evaluated using dose metrics, spatial dose distributions, mean absolute error (MAE) and MSE. To evaluate target dose, we computed $V_{95\%}$ for PTV high and PTV low, and $V_{107\%}$ for PTV high for the predicted dose distributions and GT. For the rectum, we computed V_{75Gy} , V_{60Gy} and Dmean. The bladder was evaluated for V_{65Gy} and Dmean, and anus for Dmean. The near maximum dose, $D_{0.03cc}$, was used to evaluate dose delivered to the femoral heads, and V_{20Gy} was computed for the patient contour to evaluate the prediction quality in the low to intermediate dose regions. Statistical significance of differences between predicted and GT parameters was assessed with two-sided Wilcoxon signed rank tests, using $p<0.05$ for significance.

3 Results

Figure 1 shows predicted and GT dose distributions for an axial slice of a random patient from the test set. Where the 2-level model with 50 training patients only gives an impression of the dose distribution, details start showing up with 4-levels and 250 training patients, with highest correspondence to GT for the 6-level model and 1000

training patients. For the 4-level and 6-level models, the accuracy of the predicted dose distributions clearly increased with more training data, especially regarding the distinct spikes in the intermediate dose regions (20-30 Gy). In figure 2, the population MAE and MSE show the same trend. For all model sizes, a reduction in both metrics was found with increasing training dataset size, indicating that more plans are needed for training to reach convergence. For each training dataset size, the 4- and 6-level models clearly outperformed the 2-level model in both MAE and MSE, while differences between the 4-level and 6-level models were minor for all dataset sizes.

Figure 3 shows box-and-whisker plots for the selected dose metrics. Overall, all models showed good performance for targets and OARs. For PTV high $V_{95\%}$, PTV low $V_{95\%}$, rectum V_{75Gy} , rectum V_{60Gy} , bladder V_{65Gy} , bladder Dmean, anus Dmean and patient V_{20} , increases in agreement between predictions and GT were observed with increasing training dataset size. For fixed dataset sizes, no noticeable differences were observed between model sizes, except for bladder Dmean and V_{20} , where the 2-level models had larger errors. The 6-level model trained on 1000 patients had overall the smallest dose metrics differences between predictions and GT.

4 Discussion

The purpose of this study was to systematically investigate the effect of training dataset size and model size on deep learning 3D dose prediction accuracy for prostate cancer, based on a large curated dataset of contoured planning CT-scans and corresponding highly consistent, automatically generated GT plans.

The 3D dose distributions showed a consistent trend in dosimetric detail when using more data and/or deeper models (Fig. 1). The MAE and MSE (Fig. 2) were clearly highest for the 2-level model, while differences between 4-levels and 6-levels were small. For all model sizes, MSE and MAE went down with increasing training dataset size, suggesting that more patients could further increase prediction accuracy. Differences in the 20-30 Gy areas (spikes in Fig. 1) contributed substantially to the found MSE and MAE differences. A study by Kandalan et al. [21] also showed issues with prediction accuracy in these dose regions. Probably, there is more dosimetric variation between patients for intermediate and low doses.

Contrary, for the evaluated (high dose) plan parameters, there was overall no clear dependency of prediction accuracy on model size, while predictions based on 1000 training patients were slightly more accurate than those for 250 or 50 training patients (Fig. 3).

Given the observations, evaluation of deep learning dose prediction accuracy based only on target and OAR plan parameters appears insufficient for a full assessment.

This study was based on a large dataset with highly consistent, automatically generated GT plans. It remains to be investigated how the conclusions from this work will translate to datasets with clinically/manually generated plans. Due to well-known variability in quality and trade-offs in manual plans, we do expect that convergence will already be reached with smaller training datasets, and differences between model sizes will become smaller. This study has some limitations. We have only investigated model size in terms of model depth. U-Net based architectures are capable of capturing global information. That capability is influenced by the number of trainable parameters. The number of trainable parameters of a model can be increased in more ways than only in depth. Further research is needed to investigate if model width can compensate for a lack of depth. This study investigated prostate dose prediction. In a next study, we will investigate dataset size and model size for head-and-neck cancer, with more anatomical variability and more complex treatment plans than the prostate site. Currently, data is gathered to create a large size head-and-neck database.

5 Conclusions

Using a large set of fully-automatically generated prostate cancer plans with consistent trade-offs, increasing the training dataset size from 50 to 250 to 1000 patients resulted for (high dose) PTV and OAR plan parameters in relatively small reductions in differences between deep learning predictions and GT. More substantial differences were found for dose spikes in the 20-30 Gy dose range around the PTV. For all model sizes, MAE and MSE decreased with dataset size, possibly indicating that further enhancing the training dataset size could further increase prediction accuracy.

References

- [1] ECIS, E. C. I. S.-. 2022. Estimated incidence by year - summary [Online]. Available: [https://ecis.jrc.ec.europa.eu/explorer.php?50-451-AE28E\\$4-1,2\\$3-0\\$6-0,85\\$5-All\\$77\\$21-0\\$2-All\\$CLongtermChart3_1\\$X0_14- \\$CLongtermChart3_2\\$X1_14-](https://ecis.jrc.ec.europa.eu/explorer.php?50-451-AE28E$4-1,2$3-0$6-0,85$5-All$77$21-0$2-All$CLongtermChart3_1$X0_14- $CLongtermChart3_2$X1_14-) [Accessed Jan 26th 2024].
- [2] N. Hoffmans-Holtzer, L. Smolenaers, R. Peeters, et al. "Robust scheduling for a One Stop Shop palliative radiotherapy clinic using genetic algorithms". *Radiotherapy and Oncology* 170 (2022), S874–S876
- [3] L. Rossi, S. Breedveld, S. Aluwini, et al. "Noncoplanar beam angle class solutions to replace time-consuming patient-specific beam angle optimization in robotic prostate stereotactic body radiation therapy". *92.4* (2015), pp. 762–770.
- [4] P. W. Voet, M. L. Dirkx, S. Breedveld, et al. "Toward fully automated multicriterial plan generation: a prospective clinical study". *International Journal of Radiation Oncology* Biology* Physics* 85.3 (2013), pp. 866–872.
- [5] V. Kearney, J. W. Chan, S. Haaf, et al. "DoseNet: a volumetric dose prediction algorithm using 3D fully-convolutional neural networks". *Physics in Medicine & Biology* 63.23 (2018), p. 235022.
- [6] A. Babier, R. Mahmood, A. L. McNiven, et al. "Knowledge-based automated planning with three-dimensional generative adversarial networks". *Medical Physics* 47.2 (2020), pp. 297–306.
- [7] V. Kearney, J. W. Chan, T. Wang, et al. "DoseGAN: a generative adversarial network for synthetic dose prediction using attention-gated discrimination and generation". *Scientific reports* 10.1 (2020), p. 11073.
- [8] B. Zhan, J. Xiao, C. Cao, et al. "Multi-constraint generative adversarial network for dose prediction in radiotherapy". *Medical Image Analysis* 77 (2022), p. 102339.
- [9] M. Mashayekhi, R. McBeth, D. Nguyen, et al. "Artificial intelligence guided physician directive improves head and neck planning quality and practice Uniformity: A prospective study". *Clinical and Translational Radiation Oncology* 40 (2023).
- [10] A. Babier, B. Zhang, R. Mahmood, et al. "OpenKBP: the open-access knowledge-based planning grand challenge and dataset". *Medical Physics* 48.9 (2021), pp. 5549–5561.
- [11] A. Babier, R. Mahmood, A. L. McNiven, et al. "The importance of evaluating the complete automated knowledge-based planning pipeline". *Physica Medica* 72 (2020), pp. 73–79.
- [12] A. M. Barragán-Montero, M. Thomas, G. Defraene, et al. "Deep learning dose prediction for IMRT of esophageal cancer: the effect of data quality and quantity on model performance". *Physica Medica* 83 (2021), pp. 52–63.
- [13] D. Nguyen, M.- H. Lin, D. Sher, et al. "Advances in Automated Treatment Planning". *Seminars in Radiation Oncology*. Vol. 32. 4. Elsevier. 2022, pp. 343–350.
- [14] S. Breedveld, P. R. Storchi, P. W. Voet, et al. "iCycle: Integrated, multicriterial beam angle, and profile optimization for generation of coplanar and noncoplanar IMRT plans". *Medical physics* 39.2 (2012), pp. 951–963.
- [15] A. W. M. Sharfo, M. L. Dirkx, R. G. Bijman, et al. "Late toxicity in the randomized multicenter HYPRO trial for prostate cancer analyzed with automated treatment planning". *Radiotherapy and Oncology* 128.2 (2018), pp. 349–356.
- [16] S. Breedveld, P. R. Storchi, and B. J. Heijmen. "The equivalence of multi-criteria methods for radiotherapy plan optimization". *Physics in Medicine & Biology* 54.23 (2009), p. 7199.
- [17] B. Heijmen, P. Voet, D. Fransen, et al. "Fully automated, multicriterial planning for Volumetric Modulated Arc Therapy—An international multi-center validation for prostate cancer". *Radiotherapy and Oncology* 128.2 (2018), pp. 343–348.
- [18] S. Aluwini, F. Pos, E. Schimmel, et al. "Hypofractionated versus conventionally fractionated radiotherapy for patients with prostate cancer (HYPRO): acute toxicity results from a randomised non-inferiority phase 3 trial". *The Lancet Oncology* 16.3 (2015), pp. 274–283.
- [19] S. Liu, J. Zhang, T. Li, et al. "A cascade 3D U-Net for dose prediction in radiotherapy". *Medical Physics* 48.9 (2021), pp. 5574–5582.
- [20] D. Nguyen, X. Jia, D. Sher, et al. "3D radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected U-net deep learning architecture". *Physics in medicine & Biology* 64.6 (2019), p. 065020.
- [21] R. N. Kandalan, D. Nguyen, N. H. Rezaeian, et al. "Dose prediction with deep learning for prostate cancer radiation therapy: model adaptation to different treatment planning practices". *Radiotherapy and Oncology* 153 (2020), pp. 228–235.

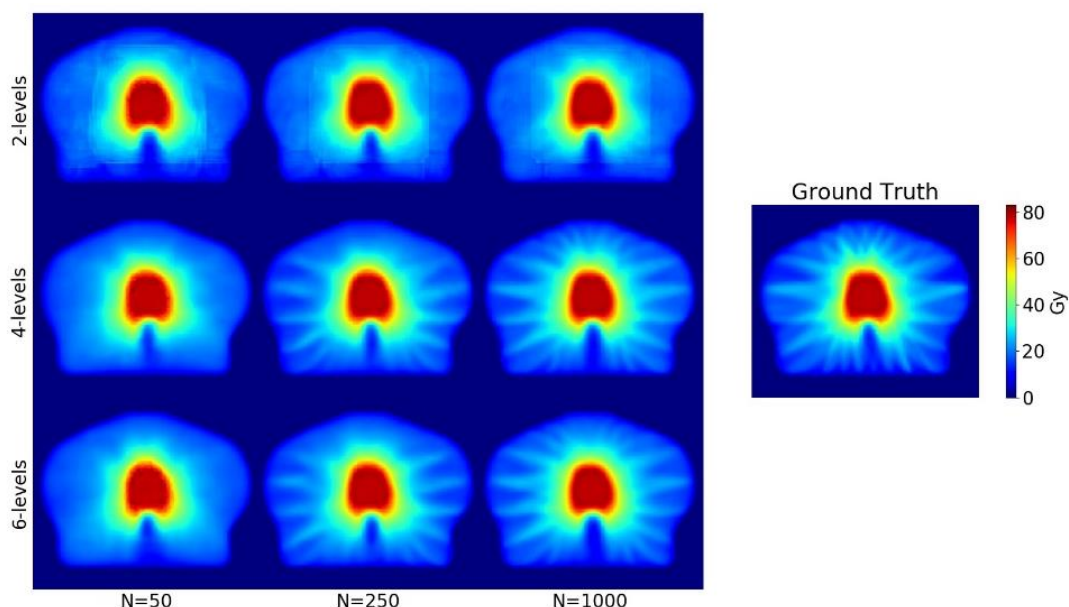


Figure 1: Axial slice of the predicted 3D dose distributions of a randomly selected patient from the test set (left) compared to ground truth (right).

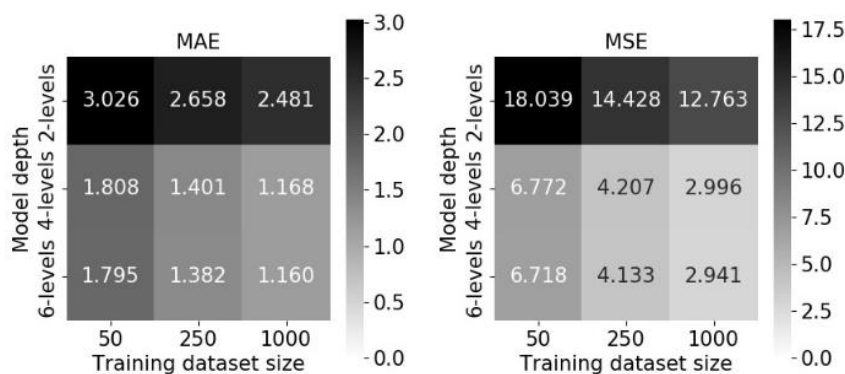


Figure 2: MAE and MSE for the test set. MAE and MSE were computed over voxels with a dose higher than 1 Gy.

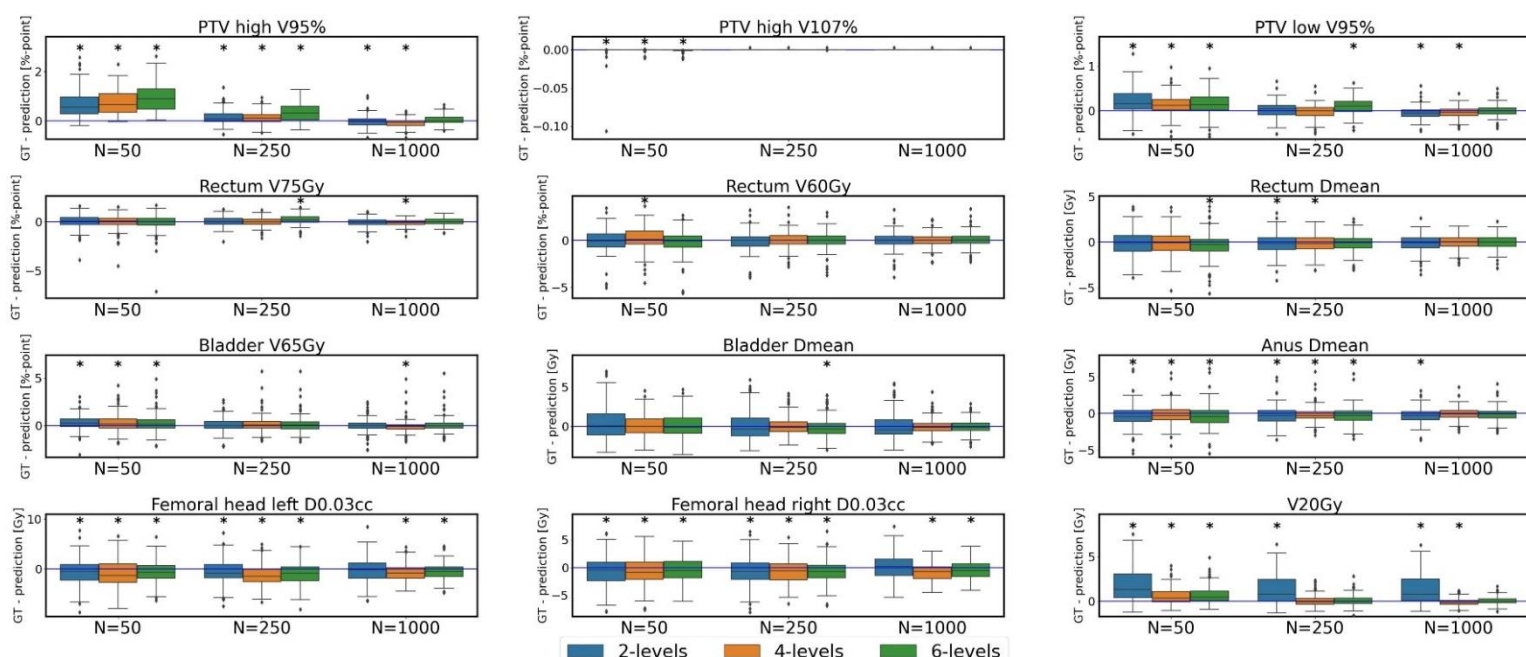


Figure 3: Box-and-whisker plots of the differences between the predicted and ground truth (GT) dose metrics. * denotes statistical significance difference to GT as a result of Wilcoxon signed rank test ($p < 0.05$).