Building a VMAT Dose Prediction Engine

Braeden Cairns

Abstract

Volumetric modulated arc therapy (VMAT) is the standard of care in radiation oncology. Despite this, accurate dose calculation remains computationally expensive, often requiring tens of minutes per plan using clinical algorithms. This work presents and compares two deep learning-based VMAT dose prediction engines developed at Roswell Park Comprehensive Cancer Center, both trained and evaluated on the same dataset of real clinical patient data. The first engine is based on the improved Dose Transformer Algorithm (iDoTA), a convolutional U-Net architecture with a transformer backbone that models beam propagation as a sequence of 2D cross-sections along the beam axis. The second adapts the MedNeXt architecture, a fully convolutional U-Net without any transformer component, originally designed for medical image segmentation. Both engines take as input a patient CT volume and a ray-traced first-principles dose estimate per control point, and output a predicted 3D dose distribution. Networks were trained on 2,838 control points from 16 patients across four treatment sites: brain, lung, head-and-neck, and prostate, using Monte Carlo (MC)-simulated doses as ground truth. Three dose matrix sizes were investigated for each architecture: 96×96×64, 144×144×80, and 192×192×112, with the largest size yielding the best results for both networks. At the largest matrix size, iDoTA achieved a mean full-plan 2%2mm Gamma Pass Rate (GPR) of 96.9±2.7% with end-to-end plan inference in 9.5±2.6 seconds, while MedNeXt achieved 98.9±0.6% with inference in 23.2±8.1 seconds. Both engines represent improvements over conventional TPS workflows by one to two orders of magnitude in speed while maintaining state of the art clinical accuracy, and both show promising generalization to an abdominal treatment site unseen during training.

1 Introduction

Accurate dose calculation is a fundamental requirement of radiation therapy treatment planning. In VMAT, a linear accelerator rotates continuously around the patient while simultaneously modulating the beam energy, shape, and intensity. This allows for highly conformal, three-dimensional dose sculpting that can target cancerous tissue while sparing surrounding healthy organs at risk. However, the physical complexity of VMAT delivery makes dose computation inherently expensive (1). Clinical treatment planning systems (TPS), such as Varian's Eclipse, currently manage this by using a fast but approximate algorithm during iterative plan optimization, then switching to a more rigorous algorithm, typically AcurosXB (AXB). This AXB solves the linear Boltzmann transport equation, for final dose evaluation. AXB produces highly accurate dose distributions but requires several minutes of runtime on a clinical workstation. Monte Carlo (MC) simulation is considered the gold standard for dose accuracy, modeling individual photon-electron interactions stochastically, but historically has been far too slow for routine clinical use (1). The demand for rapid, accurate dose calculation extends well beyond final plan generation. Secondary dose verification, as recommended by AAPM Task Group 219 (2), requires an independent volumetric calculation, not merely a point-based check. Adaptive radiotherapy workflows, where plan re-optimization occurs while the patient remains on the treatment table, require full dose recalculation within a clinically acceptable time frame which is typically only a few minutes on systems such as Varian's Ethos. A dose engine capable of delivering MC-level accuracy in seconds would therefore be of immediate clinical value across plan verification, adaptive re-planning, and quality assurance applications.

Deep learning-based dose engines have emerged as a promising path toward this goal. An early class of methods, promoted through benchmark challenges such as OpenKBP (**?**) and GDP-HMM (3), directly predicts full-plan dose from CT images, structure contours, and prescribed dose. While useful for plan optimization at the organ level, these approaches lack the beam physics inputs, such as percent depth dose (PDD) and off-axis ratio (OAR), needed for voxel-level accuracy or control-point-level dose prediction. A more physics-aware class of engines trains neural networks to predict MC- or AXB-calculated doses for individual beams or control points, using CT data and beam geometry encodings as input. Kontaxis et al. pioneered this approach with a U-Net that predicted single-beam doses in under one second (4). Witte et al. extended this to full VMAT plans using 2D convolutions along the beam path combined with recurrent units, achieving strong Gamma pass rates versus MC in approximately 20 seconds (5). A key limitation common to these works is the absence of explicit PDD and OAR information, which places a ceiling on the physical accuracy achievable in principle. Pastor-Serrano et al. addressed this with the improved Dose Transformer Algorithm (iDoTA), which incorporates a first-principles ray-traced dose estimate based on the TG-71 formalism as a direct model input (6). The iDoTA architecture treats the CT and ray-traced dose volumes as sequences of 2D slices along the beam axis, passing them through a convolutional encoder, a transformer enabling long-range attention along the beam path, and a convolutional decoder. Despite being trained on static beams, iDoTA achieves state-of-the-art Gamma pass rates versus AXB for full VMAT plans in roughly 8 seconds on average. The present work builds directly on the iDoTA framework and introduces a second architecture, MedNeXt (7), an adapted medical segmentation network. Both engines share the same data pipeline and training dataset, enabling a direct architectural comparison. Critically, both are extended to support multiple dose matrix sizes and are trained on 2-degree sub-arc doses between consecutive control points, directly modeling the continuous gantry motion of VMAT rather than approximating it with static beam predictions.

2 Dataset

The dataset used for this work was a non-publicly available dataset comprised of real patient data from the Roswell Park Eclipse clinical treatment system. This was made up of the patients CT images in DICOM file format, as well as RT, RD, and RS files that contain the patients therapy plan, radiation dosages, and body structures respectively. The RP and RS files are useful in the preprocessing and raytracing steps defined below. The eclipse dose, found in the RD file, is not used in the current pipeline as we instead use Monte Carlo (MC) simulated doses as ground truth as this is viewed as the gold standard in dose calculation (1).

Previous efforts into using deep learning for radiation therapy dose prediction have mainly centered around a single treatment site. With the iDoTa paper (6) being the first to really introduce a viable multi-site network. Following that breakthrough this work is also multi-site, but with an expanded training on four separate treatment sites. With those being head-and-neck (HN), brain (BN), lung (LG), and prostate (PR). There was a total of 2838 total CPs from 16 patients used as training data. Each patient, and therefore the overall training data, had an 85-15 training-validation data split. The training patients were broken down by site as follows: 2 HN patients with 555 total CPs, 4 BN patients with 685 total CPs, 7 LG patients with 998 total CPs, and 3 PR patients with 600 total CPs. This training and testing patient data breakdown is highlighted below in Tab. 1. The network inputs were made up of the patients CT volume plus a hand-calculated raytraced dose intensity map per control point with the MC calculated dose as the ground truth. Below I will discuss this raytracing step and other preprocessing more in depth.

Table 1: Training and evaluation data by treatment site

Site	Train Sub-arcs	Train Patients	Test Patients	Test Arcs
Brain	685	4	2	4
Lung	998	7	6	15
Head/Neck	555	2	2	5
Prostate	600	3	3	8
Abdominal	0	0	1	4
Total	**2838**	**16**	**14**	**36**

2.1 Raytracing and Preprocessing

As mentioned above, the input to this network is a 3D 2-channel tensor in the shape $(2, D, H, W)$ where D, H, W represent the crop size height, depth, and width respectively. The first channel is made up of the CT volume density array read from DICOM file. This volume gets cropped or padded when necessary to have a static input and output size for the network. We utilize three different sizes with a comparison of how that effects performance found in Results below. These three crop sizes are $96 \times 96 \times 64$, $144 \times 144 \times 80$, and $192 \times 192 \times 112$. As highlighted below in Fig. 1, these crop sizes correspond directly to the output dose matrix dimensions, and therefore the anatomical coverage of the beam. Each dimension represents its respective voxel count, with these voxels being sized at 2mm$\times$2mm$\times$2.5mm each. The second channel in input is made up of a raytraced body mask that uses the patient CT image as well as the clinical RT Dose, Plan, and Structure files. These clinical therapy files give physical coordinate and orientation information for the dose array, and defines the isocenter position given the beam configuration. All of these files are used to generate the control point MLC Aperture's BEV mapping in the patient's body. Voxels inside the body are assigned a hand calculated value in cGy/MU, while voxels outside the body are dropped to 0. The resulting body mask is then cropped in the same manner as the CT density array.

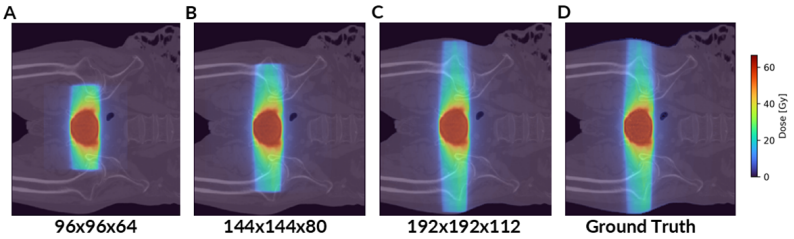

Figure 1: Dose matrix crop size breakdown

3 Model Descriptions

A VMAT Dose Prediction engine aims to beat our current clinical software, the Eclipse treatment planning system and dose algorithm, in terms of speed while maintaining accurate dosage levels. This prediction uses the patients CT image along with a raytraced body mask that captures the treatment area from the beams eye view. These input files are generated using the CT image, therapy plan data, and the LINAC beam data for the hardware being used. The resulting input files are concatenated and fed as our input to the dose prediction network, with this process highlighted below in Fig. 2. At its core, this is simply a 3D regression task. This dose engine also allows for plug-and-play use of the main prediction network. As mentioned above, two different architectures were used and compared against each other with these being the iDoTa architecture and the MedNeXt architecture. A more in depth discussion of the two models can be found below.

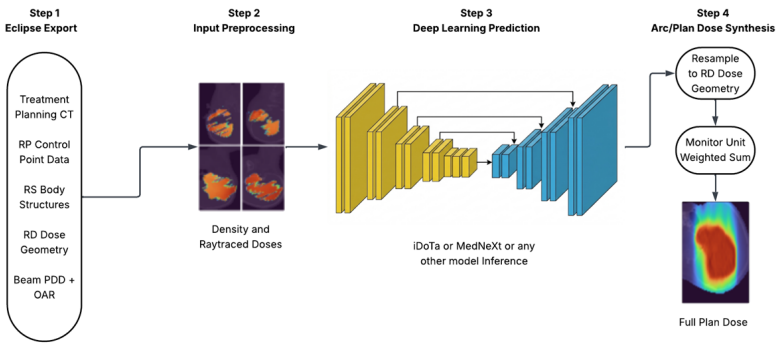

Figure 2: Dose prediction engine pipeline illustration

3.1 iDoTa Baseline

The baseline architecture my research group started from was built off of the iDoTa paper (6). This model was a very innovative multi-site dose prediction network that introduced a powerful Transformer-based architecture to learn beam-body interactions for three-dimensional therapy planning. To go more into model specifics this architecture is centered around the 'improved Dose Transformer Algorithm' which map arbitrary patient geometries and beam information to the corresponding 3D dose distribution (6). It uses the input data described above, patient CT volumes along with a raytraced intensity map that shows the beams eye view of the treatment, to output the 3D dose matrix at whatever the desired output size is. The architecture includes a series of down-sampling convolutional blocks to extract local features into tokens sequence. These token sequences then go through the Transformer backbone which route information between extracted features along the depth of the entire volume. This Transformer block uses pre-Layer Normalization (8) to ensure stable gradient flow and prevent instabilities caused by deep attention stacks. Finally, a series of upsampling convolutional blocks are used to convert back from the token sequence to the desired output dose volume. This encompasses a standard U-net style architecture with an attention backbone. Below in Fig. 3 you will see a visualization of this network architecture.

As previously mentioned, the dataset was tested at three separate crop sizes, $96 \times 96 \times 64$, $144 \times 144 \times 80$, and $192 \times 192 \times 112$. For this architecture the output size is directly tied to the parameter size of the network, so a total of 3 models were trained. This means the baseline $96 \times 96 \times 64$ crop size network has 2.23m parameters, while the $144 \times 144 \times 80$ crop size model rises to 5.02m parameters and the largest $192 \times 192 \times 112$ crop size model reaches all the way up to 10.6m parameters. These crop sizes also had an affect on training time, with the seconds per epoch rising as the corp size rose. This makes sense as not only does this increase raise the number of learnable parameters but it also increases the number of computations in both the forward and backward pass due to increased spatial resolutions. Another thing to note when training the largest crop size was that the batch size had to be reduced during training for times sake. The typical batch size used for training was four, however with the largest crop size this proved to be incredibly resource and time inefficient and a batch size of 2 was used, which proved to have no adverse affects as you'll see below in the results.

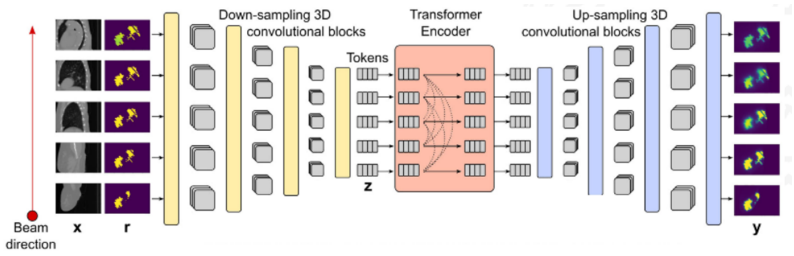

Figure 3: iDoTa Dose prediction architecture

3.2 MedNeXt

The network used was an adapted version of the MedNeXt architecture originally produced by the German Cancer Research Center (DKFZ), Division of Medical Imaging Computing (MIC) (7). This architecture is mainly used as a powerful medical image segmentation tool, but can also be adapted to a dense regression task such as 3D VMAT dose prediction. This network also follows a typical convolutional U-net encoder-decoder paradigm, however notably without the Transformer backbone. This is appealing as the removal of the Transformer should theoretically allow for better control over network and input scalability given the lack of the quadratically scaling attention mechanism. However, this model is built in an effort to mirror the attention-mechanism with convolutions, so it unsurprisingly does not show increased speeds in this pipeline. In fact, this work build directly off of the ConvNeXt work (9) to adapt the powerful convolution framework specifically for medical imaging domain tasks. To be specific, a MedNeXt block is made up of three layers that mirror the typical Transformer block as first introduced in 2017 (10). This mirrored architecture allows for Transformer-like information routing while remaining fully convolutional. The blocks are made up of a depthwise convolution layer to replicate the large attention window found in a Swin-Transformer (11) while keeping the computational cost down by removing the attention backbone. An expansion

layer is used to effectively decouple width scaling from receptive field scaling to allow for increased capacity without an increase in computations. Finally, a compression layer which performs a channel-wise compression of the feature maps to allow for compact representations that preserve important. The overall network has 4 encoder-decoder layers with a bottleneck layer and residual connections between them to best leverage the benefits of an inverted Transformer-like bottleneck. This means the model is able to preserve information well in lower spatial resolutions across all components (7). The architecture details are illustrated below in Fig. 4.

Similar to above there were networks trained and configured at all three crop sizes for the data, $96 \times 96 \times 64$, $144 \times 144 \times 80$, and $192 \times 192 \times 112$. However, as the MedNeXt models do not scale with data size the weights were static for all three crop sizes. This led me to investigate different sizes networks, I used both the 'small' and 'baseline' configurations provided for the MedNeXt network. The 'small' configuration has 5.5m parameters, while the 'baseline' configuration has 10.5m parameters, while both utilize a kernel size of 3×3. This means there were 6 total MedNeXt models created and trained, 2 models sizes x 3 crop sizes. The main adaptation point to shift this segmentation network to a regression network was altering the number of output classes and the loss function used. Simply setting the number of output classes to 1 and utilizing Mean Squared Error (MSE) loss as opposed to the Cross-Entropy loss used previously shifted the network to output a 3D dose distribution of continuous values. A more in depth discussion of the optimization process, hyperparameters used, and hardware used can be found below in 3.3.

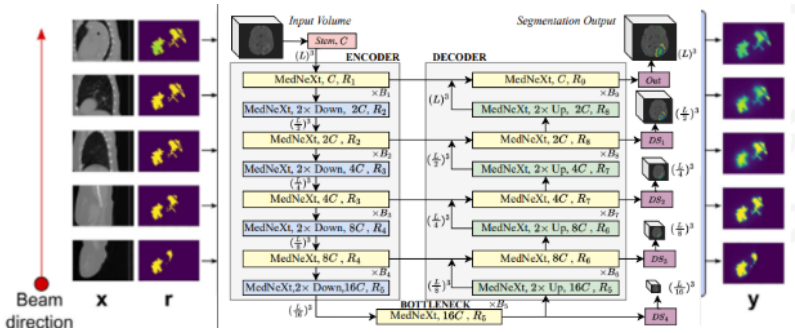

Figure 4: Adapted MedNeXt architecture

3.3 Training Details

Both models were trained using MSE loss, $L = \frac{1}{N}\Sigma_{i=1}^{N}(y_i - \hat{y}_i)^2$, appropriate for this regression task. A hybrid dose loss combining L1, masked weighted L1, moment, and dual sharp L2 terms was also briefly evaluated for MedNeXt but produced no meaningful improvement in training or convergence, so MSE was retained for all configurations.

The AdamW optimizer (12) was used for all models, preferred over standard Adam (13) for its decoupled weight decay, which provides smoother gradient flow and better generalization. Training followed a supervised mini-batch scheme with epoch-level validation, scheduling, and checkpointing. All hyperparameters were held constant across all nine network configurations: learning rate of 0.001 decaying by 0.5 every 15 epochs, weight decay of 0.0001, and 120 training epochs. The typical batch size was 4, reduced to 2 for the largest crop size to manage memory overhead. Validation loss was used for model selection. Training was performed on an NVIDIA H100 GPU.

4 Metrics and Experimental Results

The Gamma Passing Rate of a radiation therapy plan is the standard evaluation metric for dose predictions. This metric ensures the planning target volume (PTV), or cancerous tissue, receives radiation while the organs at risk (OAR), or healthy nearby tissue, does not receive such radiation. The common dose percentages and distance thresholds used are 3%, 3mm, 2%, 2mm, and 1%, 1mm all with a lower 10% dose cutoff. For this work and other similar ones I focused mainly on 2%, 2mm as that is the clinical standard. For this configuration (2%2mm) any result above 90% is considered clinically acceptable, though obviously the strongest possible results are desired.

There were two styles of evaluations performed at varying levels of the plan. We evaluated the models performance on an arc-by-arc basis as well as the overall full plan dose. Evaluation was done on an

arc-by-arc basis as this is useful in clinical plan assurance and adaptive treatment planning. As this model predicts on a CP-by-CP basis, this means the arc dose is gathered by summing all CP doses for a given arc. It was also done on a full-plan basis as the main focus and goal of a dose prediction engine is to quickly predict the full plan dose for a given patient. Similar to above, as this engine predicts on a CP-by-CP basis, the full plan dose is gathered by summing all arc doses for a given patient. Furthermore, the speed was measured in seconds for both arc-by-arc prediction and full plan prediction. When evaluating speed we aimed to reproduce inference conditions, so the raytracing and pre-processing steps were performed during inference. The raytracing step proved to be by far the most time consuming portion of inference time. Testing was performed on a total of 4121 CPs from 10 patients. We evaluated on 5 treatment sites, HN, BN, LG, PR, with the added abdominal (AB) region as shown above in Tab. 1. The patients were broken down by site as follows: 2 HN patients with 575 total CPs, 2 BN patients with 712 total CPs, 3 LG patients with 892 total CPs, 2 PR pat with 703 total CPs, and 1 AB patient with 703 total CPs.

I initially hypothesized that the MedNeXt network would not only achieve higher GPR accuracies, but also have faster inference and throughout times as it removes the costly Transformer backbone found in the iDoTa baseline. As you'll see below I was not entirely correct, while the MedNeXt network yielded improved accuracies it actually ended up having slower inference and throughput times. The increased inference time of the fully convolutional network is likely due to the computational cost of repeated high-resolution convolution operations, whereas the transformer-based architecture benefits from more efficient parallel feature processing and reduced spatial computations. Another reason I suspect is because the model can have up to 4x the parameters of the iDoTa baseline. The fact the time does not scale accordingly to the increased parameter count is a good indication that it is not terribly inefficient. Also, the speeds are still much faster compared to the clinical Eclipse software, so slightly losing out on speed to the iDoTa network is not a massive pitfall.

4.1 Results

As seen in Fig. 5, both architectures showed improved accuracy as the dose matrix size increased, a consequence of increased anatomical coverage and richer physical information in both the inputs and MC ground-truth targets. The MedNeXt engine outperformed iDoTA on full-plan GPR across almost all of the test patients, and achieved a higher mean GPR in both evaluations performed. On the full-plan front MedNeXt achieves a mean GPR of 98.9% vs. iDoTa's 96.8% at the largest matrix size. At the arc level this advantage holds as well, with MedNeXt reaching 99.3±0.7% compared to 97.8±1.9% for iDoTA. The iDoTA engine, however, was substantially faster: approximately 9.5 seconds per plan versus approximately 23–26 seconds for MedNeXt. As discussed above, this is likely due to the lack of parallel feature processing in fully convolutional networks. Despite its slower inference, MedNeXt's lower variance across patients suggests it produces more consistent predictions, which is an important quality for a system intended for clinical use. Regardless of comparative results, both engines are one to two orders of magnitude faster than conventional clinical dose algorithms such as AXB. The key scalability advantage of MedNeXt is that further increasing the matrix size does not increase its parameter count, while iDoTA's parameter count grows quadratically with matrix resolution. Site-level performance was consistent with training data volume: lung (the most represented site) and brain performed best overall, while head-and-neck (fewest training patients) showed the most variability. A summary of key results at the largest matrix size is given below in Tab. 2.

Table 2: Evaluation comparison of the best network configurations for both the iDoTa and MedNeXt architectures. Both used the largest crop size of 192x192x112 and this is the larger 10.5m parameter MedNeXt model.

Model	Mean Full Plan GPR (2%/2mm)	Mean Arc-by-Arc GPR (2%/2mm)	Mean Full Plan Inference Time	Mean Single Arc Inference Time
iDoTa	$96.9 \pm 2.7\%$	$97.8 \pm 1.9\%$	$\mathbf{9.5 \pm 2.6s}$	$\mathbf{4.9 \pm 0.5s}$
MedNeXt	$\mathbf{98.9 \pm 0.6}\%$	$\mathbf{99.3 \pm 0.7}\%$	$23.2 \pm 8.1s$	$8.6 \pm 2.1s$

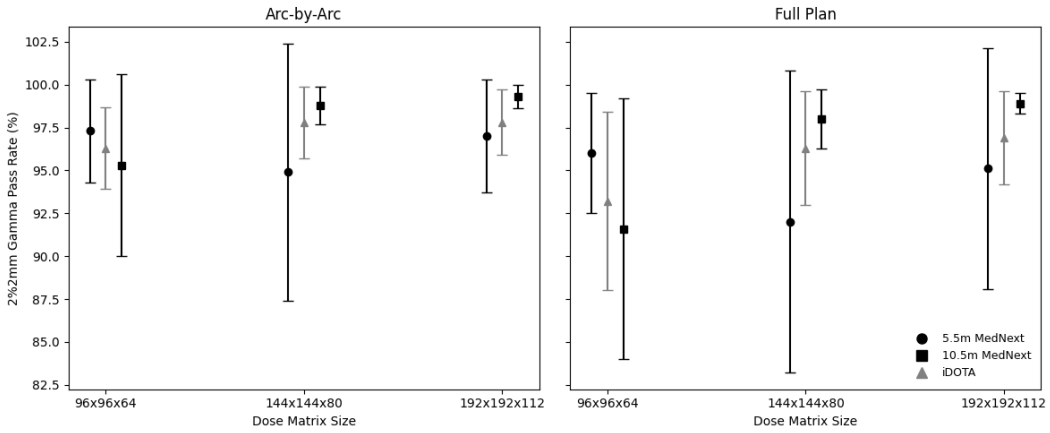

Figure 5: Arc-by-arc and full plan 2%2mm GPR evaluation results broken down by architecture, crop size, as well as parameter count for MedNeXt only.

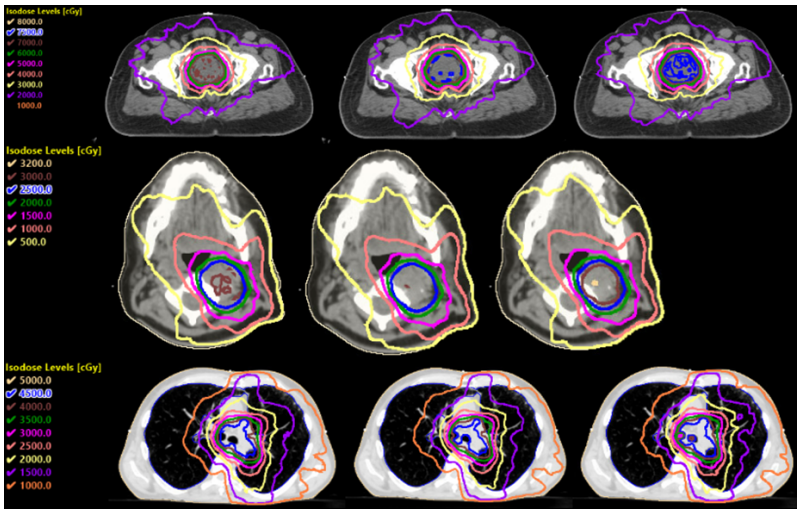

Figure 6: Full Plan isodose line comparison between the Eclipse true plan (left), our iDoTa prediction (middle), and our MedNeXt prediction (right) for a test prostate (top), head-and-neck (middle), and lung (bottom) patient.

4.2 Discussion

Above you can see the same pattern being highlighted in both the iDoTa and MedNeXt models in both arc-by-arc and full plan cases. As the parameter count and crop sizes increase so does model performance. The best performing iDoTa model was the $192 \times 192 \times 112$ crop size model, and this sentiment was echoed in the MedNeXt evaluations. With the best performing model being the 10.5m parameter model at the largest crop size as well. This highlights how an increased anatomical coverage corresponds with improved planning results. However, there has not been enough conclusive testing to figure out at what point increasing the crop size fails to return better results. This will be investigated in the future as this work continues. In terms of MedNeXt performance I can confirm that increasing the parameter count to 17.5m did not show a large leap in performance, so I suspect the iDoTa results will be similar. The variability in person from site to site is linked to two elements. First, the different sites come with different size and tissue challenges. Second, the training dataset is not properly balanced due to data restrictions. When comparing the performance of the strongest iDoTa and MedNeXt models against each other in terms of full plan predictions you see the MedNeXt model outperforms its iDoTa counterpart on all patients. The MedNeXt models strong performance on the unseen AB test site further emphasizes its powerful generalization qualities. The most interesting

dichotomy I found was that the iDoTa network beats out its MedNeXt competitors in terms of inference speeds at all crop sizes, yet can not match the GPR performance in almost any test patient case.

5 Conclusion

This work demonstrates that both the iDoTA transformer-based architecture and the MedNeXt fully convolutional architecture can serve as effective VMAT dose prediction engines, each capable of producing full-plan dose distributions approaching MC-level accuracy in well under one minute. Using the same dataset, preprocessing pipeline, and evaluation criteria for both, several clear conclusions emerge. First, increasing the dose matrix size consistently improved accuracy for both architectures, confirming that greater anatomical coverage and richer input information are beneficial. Second, MedNeXt achieved higher overall GPR than iDoTA on all fronts and showed superior scalability. Because its parameter count does not grow with matrix size, it is better suited to further increases in anatomical coverage. Third, iDoTA maintained a substantial speed advantage, making it preferable in contexts where latency is the primary constraint. Fourth, both engines demonstrated meaningful generalization to the abdominal site not seen during training, with MedNeXt achieving 98.2% GPR on this patient and iDoTA achieving 97.7%, a finding that supports the clinical robustness of both approaches. Neither engine is intended to fully replace the clinical TPS at this current point in time. Both are best positioned as secondary dose verification tools, rapid, independent checks on clinical treatment plans, and as components of adaptive radiotherapy workflows where speed is critical and MC-level accuracy is desirable.

6 Future Work

Several important questions remain. First, both training and evaluation datasets were limited to flattened 6 MV beams delivered on a single linear accelerator at one institution. The generalizability of both engines to different beam energies (e.g., 10 MV, flattening-filter-free beams), different linac models, and external patient cohorts has not yet been established. Expanding training data to cover this diversity is a clear next step. Second, the relationship between matrix size and accuracy has not yet reached a saturation point. A 224×224×160 matrix size was briefly attempted for MedNeXt but encountered memory and computational constraints; distributed multi-GPU training is being investigated to push this boundary further. This avenue is not viable for iDoTA without a corresponding and computationally costly increase in model parameters. Third, training data balance across treatment sites warrants attention. The observed differences in per-site accuracy were correlated with the amount of training data per site, suggesting that a more balanced dataset would reduce these disparities. Collecting additional patient data, particularly for head-and-neck cases, is ongoing. Finally, integration into the clinical Eclipse workflow for use as a secondary quality assurance tool is actively in progress, and both engines are candidates for future prospective clinical validation.

References

[1] A. Pant, N. Miri, S. Bhagroo, J. A. Matthews, and D. P. Nazareth, "Monitor unit verification for varian truebeam vmat plans using monte carlo (mc) calculations and phase space data," 2023.

[2] Z. T.C. and et al., "Report of aapm task group 219 on independent calculation-based dose/mu verification for imrt," 2021.

[3] Q. Labs, "Generalizable dose prediction for heterogenous multi-cohort and multi-site radiotherapy planning (gdp-hmm) grand challenge," 2025.

[4] K. C., B. G.H., L. J.J.W., and R. B.W., "Deepdose: Towards a fast dose calculation engine for radiation therapy using deep learning," 2020.

[5] W. M. and S. J.J., "A deep learning based dynamic arc radiotherapy photon dose engine trained on monte carlo dose distributions," 2024.

[6] O. Pastor-Serrano, P. Dong, C. Huang, L. Xing, and Z. Perkó, "Sub-second photon dose prediction via transformer neural networks," *Medical Physics*, vol. 50, no. 5, p. 3159–3171, Feb. 2023. [Online]. Available: http://dx.doi.org/10.1002/mp.16231

[7] S. Roy, G. Koehler, C. Ulrich, M. Baumgartner, J. Petersen, F. Isensee, P. F. Jaeger, and K. Maier-Hein, "Mednext: Transformer-driven scaling of convnets for medical image segmentation," 2024. [Online]. Available: https://arxiv.org/abs/2303.09975

[8] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T.-Y. Liu, "On layer normalization in the transformer architecture," 2020. [Online]. Available: https://arxiv.org/abs/2002.04745

[9] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," 2022. [Online]. Available: https://arxiv.org/abs/2201.03545

[10] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[11] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," 2021. [Online]. Available: https://arxiv.org/abs/2103.14030

[12] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019. [Online]. Available: https://arxiv.org/abs/1711.05101

[13] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017. [Online]. Available: https://arxiv.org/abs/1412.6980

7 Appendix

Appendix A: iDoTa arc-by-arc results

Table 3: Arc-by-arc 2%, 2mm GPR for all three iDoTa configurations on all test patients

Patient	GPR Type	96x96x64	144x144x80	192x192x112
AB1	Arc 1	96.30	98.18	99.12
AB1	Arc 2	92.65	96.96	99.03
AB1	Arc 3	95.50	99.24	99.42
AB1	Arc 4	94.74	98.30	99.07
BN1	Arc 1	96.47	99.70	99.79
BN1	Arc 2	96.52	99.34	99.68
BN2	Arc 1	99.20	98.98	99.79
BN2	Arc 2	97.84	96.67	98.58
LG1	Arc 1	97.06	98.83	97.03
LG1	Arc 2	97.28	98.84	97.59
LG1	Arc 3	97.03	98.79	94.61
LG2	Arc 7	99.47	92.36	97.47
LG2	Arc 8	98.18	91.91	96.62
LG2	Arc 9	99.41	90.81	96.84
LG3	Arc 1	97.18	99.17	99.16
LG3	Arc 2	98.43	99.01	99.33
HN1	Arc 1	96.39	98.78	98.67
HN1	Arc 2	92.60	97.54	94.97
HN1	Arc 3	93.34	97.88	96.51
HN2	Arc 3	95.79	97.45	97.39
HN2	Arc 4	94.70	97.35	94.58
PR1	Arc 1	91.51	99.47	99.93
PR1	Arc 2	94.40	99.78	99.97
PR1	Arc 3	93.88	99.40	99.79
LG4	Arc 1	99.42	99.07	97.13
LG4	Arc 2	99.27	99.15	97.20
LG4	Arc 3	98.75	99.08	94.89
LG5	Arc 1	97.85	99.04	98.57
LG5	Arc 2	98.00	98.91	98.52
LG6	Arc 1	95.21	97.84	93.42
LG6	Arc 2	96.69	96.49	93.63
PR2	Arc 1	92.97	98.61	98.89
PR2	Arc 2	91.57	96.99	97.44
PR2	Arc 3	93.57	99.12	99.20
PR3	Arc 4	99.08	99.81	99.13
PR3	Arc 5	99.31	99.70	99.56

Appendix B: iDoTa full plan results

Table 4: Full Plan 2%, 2mm GPR for all three iDoTa configurations on all test patients

Patient	GPR Type	96x96x64	144x144x80	192x192x112
AB1	Full Plan	83.13	94.11	97.75
BN1	Full Plan	92.73	98.98	99.54
BN2	Full Plan	95.09	94.46	97.82
LG1	Full Plan	94.26	98.03	91.90
LG2	Full Plan	97.80	86.95	96.45
LG3	Full Plan	96.81	98.67	99.02
HN1	Full Plan	93.18	96.43	94.35
HN2	Full Plan	94.91	96.81	95.27
PR1	Full Plan	86.18	98.54	99.87
LG4	Full Plan	99.07	98.98	96.28
LG5	Full Plan	97.19	98.74	98.06
LG6	Full Plan	93.96	96.45	92.36
PR2	Full Plan	83.82	94.59	94.80
PR3	Full Plan	97.25	99.49	98.70

Appendix C: iDoTa inference timing results

Table 5: Mean inference times in seconds for all three iDoTa configurations on both evaluation styles

	96x96x64	144x144x80	192x192x112
Arc by Arc	2.7 ± 0.9	3.2 ± 0.9	4.9 ± 0.5
Full Plan	8.4 ± 3.7	8.6 ± 3.7	9.5 ± 2.6

Appendix D: MedNeXt arc-by-arc results

Table 6: Arc-by-arc 2%, 2mm GPR for all six MedNeXt configurations on all test patients

		5.5m model	5.5m model	5.5m model	10.5m model	10.5m model	10.5m model
Patient	GPR Type	96x96x64	144x144x80	192x192x112	96x96x64	144x144x80	192x192x112
AB1	Arc 1	98.31	77.70	98.82	96.30	98.09	99.50
AB1	Arc 2	97.126	66.594	98.364	97.813	95.976	98.799
AB1	Arc 3	98.237	97.978	98.361	94.617	99.391	99.158
AB1	Arc 4	97.185	98.659	94.802	89.361	99.075	96.16
BN1	Arc 1	99.313	98.535	99.723	99.754	99.801	99.685
BN1	Arc 2	98.983	97.797	99.662	99.474	99.659	99.534
BN2	Arc 1	99.915	99.439	99.794	99.737	99.279	99.846
BN2	Arc 2	99.576	98.93	99.921	99.898	99.925	99.703
LG1	Arc 1	95.973	94.641	96.825	98.254	98.406	99.098
LG1	Arc 2	94.761	92.115	97.161	97.847	98.282	98.934
LG1	Arc 3	96.346	95.515	96.49	97.779	98.544	98.874
LG2	Arc 7	98.796	99.572	99.833	94	99.964	99.812
LG2	Arc 8	97.72	98.912	99.566	91.421	99.882	99.487
LG2	Arc 9	97.66	99.409	99.688	93.023	99.852	99.685
LG3	Arc 1	99.203	88.296	98.834	96.151	99.289	99.721
LG3	Arc 2	99.118	87.256	99.163	94.634	99.259	99.758
HN1	Arc 1	98.135	99.702	99.883	99.797	99.955	99.79
HN1	Arc 2	96.304	98.261	99.545	98.61	99.741	99.767
HN1	Arc 3	97.372	99.34	99.698	98.126	99.885	99.672
HN2	Arc 3	87.996	96.275	96.922	93.642	97.73	97.89
HN2	Arc 4	85.775	96.229	96.618	92.455	97.779	98.885
PR1	Arc 1	99.686	98.879	99.436	98.652	99.843	99.464
PR1	Arc 2	99.067	98.972	99.717	98.974	99.922	99.62
PR1	Arc 3	99.686	99.515	99.335	97.219	99.858	99.254
LG4	Arc 1	97.567	98.954	91.884	74.051	98.216	99.184
LG4	Arc 2	98.016	99.096	95.788	91.667	98.179	99.148
LG4	Arc 3	96.804	99.005	91.361	80.57	97.48	99.138
LG5	Arc 1	95.295	81.65	93.462	92.803	97.733	99.64
LG5	Arc 2	96.168	80.984	95.081	95.303	97.559	99.569
LG6	Arc 1	95.253	92.911	93.712	97.44	97.739	98.661
LG6	Arc 2	96.1	94.294	94.467	96.228	96.464	98.502
PR2	Arc 1	98.678	97.767	94.676	99	99.623	99.989
PR2	Arc 2	97.584	96.398	93.365	97.364	99.391	99.672
PR2	Arc 3	99.542	98.924	95.909	99.898	99.91	99.777
PR3	Arc 4	99.822	99.572	86.188	95.592	98.018	99.853
PR3	Arc 5	99.449	99.578	90.899	93.65	98.457	99.556

Appendix E: MedNeXt full plan results

Table 7: Full Plan 2%, 2mm GPR for all six MedNeXt configurations on all test patients

		5.5m model	5.5m model	5.5m model	10.5m model	10.5m model	10.5m model
Patient	GPR Type	96x96x64	144x144x80	192x192x112	96x96x64	144x144x80	192x192x112
AB1	Full Plan	96.46	68.569	98.026	87.072	96.986	98.788
BN1	Full Plan	98.34	96.356	99.402	99.254	99.427	99.278
BN2	Full Plan	98.73	96.57	99.232	99.144	99.343	99.224
LG1	Full Plan	93.39	89.601	94.813	96.923	97.595	98.775
LG2	Full Plan	95.38	98.433	98.411	83.113	99.476	98.087
LG3	Full Plan	98.94	84.941	98.558	93.294	98.966	99.661
HN1	Full Plan	96.22	98.412	99.32	97.126	99.397	99.223
HN2	Full Plan	85.89	95.536	95.968	92.074	96.98	97.267
PR1	Full Plan	99.20	97.387	98.59	94.608	99.277	98.614
LG4	Full Plan	97.35	99.126	85.923	71.889	98.309	98.728
LG5	Full Plan	94.60	80.134	92.991	92.218	97.323	99.497
LG6	Full Plan	94.91	91.072	92.564	95.558	95.394	98.3
PR2	Full Plan	94.74	92.422	89.057	96.134	99.286	99.764
PR3	Full Plan	99.34	98.728	74.404	84.554	93.797	99.251

Appendix F: MedNeXt inference timing results

Table 8: Mean inference times in seconds for all six MedNeXt configurations on both evaluation styles

	5.5m model	5.5m model	5.5m model	10.5m model	10.5m model	10.5m model
	96x96x64	144x144x80	192x192x112	96x96x64	144x144x80	192x192x112
Arc by Arc	3.1 ± 0.9	4.7 ± 1.3	8.4 ± 2.1	3.1 ± 0.8	4.8 ± 1.2	8.6 ± 2.1
Full Plan	9.2 ± 3.9	13.7 ± 5.3	22.9 ± 8.4	9.2 ± 3.5	14.1 ± 6.0	23.2 ± 8.1