Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

PAPER

Performance assessment of variant UNet-based deep-learning dose engines for MR-Linac-based prostate IMRT plans

To cite this article: Wenchih Tseng *et al* 2023 *Phys. Med. Biol.* **68** 175004

IOP Publishing Phys. Med. Biol. 68 (2023) 175004 https://doi.org/10.1088/1361-6560/aceb2c

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

PAPER

RECEIVED
10 March 2023

REVISED
4 July 2023

ACCEPTED FOR PUBLICATION
27 July 2023

PUBLISHED
11 August 2023

Performance assessment of variant UNet-based deep-learning dose engines for MR-Linac-based prostate IMRT plans

Wenchih Tseng[1], **Hongcheng Liu**[2,*], **Yu Yang**[2], **Chihray Liu**[1], **Keith Furutani**[3], **Chris Beltran**[3] and **Bo Lu**[1,3,*]

[1] Department of Radiation Oncology, University of Florida, Gainesville, FL 32610-0385, United States of America
[2] Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611-6595, United States of America
[3] Department of Radiation Oncology, Mayo Clinic, Jacksonville, FL 32224-0001, United States of America
[*] Authors to whom any correspondence should be addressed.

E-mail: hliu@ise.ufl.edu **and** lu.bo@Mayo.edu

Keywords: deep learning, dose calculation engine, IMRT, MR-linac

Abstract

Objective. UNet-based deep-learning (DL) architectures are promising dose engines for traditional linear accelerator (Linac) models. Current UNet-based engines, however, were designed differently with various strategies, making it challenging to fairly compare the results from different studies. The objective of this study is to thoroughly evaluate the performance of UNet-based models on magnetic-resonance (MR)-Linac-based intensity-modulated radiation therapy (IMRT) dose calculations. *Approach*. The UNet-based models, including the standard-UNet, cascaded-UNet, dense-dilated-UNet, residual-UNet, HD-UNet, and attention-aware-UNet, were implemented. The model input is patient CT and IMRT field dose in water, and the output is patient dose calculated by DL model. The reference dose was calculated by the Monaco Monte Carlo module. Twenty training and ten test cases of prostate patients were included. The accuracy of the DL-calculated doses was measured using gamma analysis, and the calculation efficiency was evaluated by inference time. *Results*. All the studied models effectively corrected low-accuracy doses in water to high-accuracy patient doses in a magnetic field. The gamma passing rates between reference and DL-calculated doses were over 86% (1%/1 mm), 98% (2%/2 mm), and 99% (3%/3 mm) for all the models. The inference times ranged from 0.03 (graphics processing unit) to 7.5 (central processing unit) seconds. Each model demonstrated different strengths in calculation accuracy and efficiency; Res-UNet achieved the highest accuracy, HD-UNet offered high accuracy with the fewest parameters but the longest inference, dense-dilated-UNet was consistently accurate regardless of model levels, standard-UNet had the shortest inference but relatively lower accuracy, and the others showed average performance. Therefore, the best-performing model would depend on the specific clinical needs and available computational resources. *Significance*. The feasibility of using common UNet-based models for MR-Linac-based dose calculations has been explored in this study. By using the same model input type, patient training data, and computing environment, a fair assessment of the models' performance was present.

1. Introduction

The integration of magnetic resonance imaging (MRI) with a linear accelerator (Linac) in the MR-Linac system (such as Elekta Unity (Elekta AB, Stockholm, Sweden)) offers a significant advance in image-guided radiotherapy compared to traditional computed tomography (CT)-guided radiotherapy. The fully integrated MRI unit provides superior visualization of soft tissues and enables nearly real-time tumor-tracking using non-ionizing radiation (Green *et al* 2018, Eccles *et al* 2019). This results in improved capturing of patient anatomical variations during a course of radiation treatment and the possibility of conducting personalized and adaptive

magnetic resonance-guided radiotherapy (MRgRT) on a daily basis for better targeting tumors and sparing organs-at-risk. Theses advancements are demonstrated in various studies (Pathmanathan *et al* 2018, Kurz *et al* 2020).

A fast and accurate dose calculation engine is essential for effective online adaptive planning of the MR-Linac system, as it enables practical efficiency and ensures high plan quality through frequent calculations for guidance and verification during the planning process (Shepard *et al* 2002). The external magnetic field of an MR-Linac system can cause a large deflection of the secondary electrons at the interface of inhomogeneous media, such as the air-tissue boundary, due to the Lorentz force, a phenomenon known as the 'electron return effect (ERE)' (Costa *et al* 2018, Shortall *et al* 2020). The ERE can result in considerable dose inaccuracy in model-based fast computational algorithms, such as pencil beam (PB) algorithm and convolution/superposition algorithm, as the pre-computed kernels cannot accurately reflect the true radiological dose depositions under the influence of the magnetic field (Pfaffenberger 2013, Kurz *et al* 2020, Chu *et al* 2021). For an accurate representation of ERE, Monte Carlo (MC) simulation, which models complete particle transports, is considered the most suitable method for dose calculation in MR-Linac-based treatment planning (Hissoiny *et al* 2011, Kurz *et al* 2020). Despite its accuracy, MC simulation requires intensive computation due to the vast number of particle simulations, making it challenging to achieve practical calculation times while maintaining low statistical noise. To address the computation challenges in MC simulation, various efforts have been made to expedite the calculation through parallel processing techniques using multiple central processing units (CPUs) or graphics processing units (GPUs) (Hissoiny *et al* 2011, Jia *et al* 2011, Ziegenhein *et al* 2015). While this has led to a faster MC computation (e.g. a few minutes), it may still be insufficient for clinical use and a limiting factor in the real-time adaptive treatment process.

Recently, deep-learning (DL) approaches have emerged as a promising alternative for fast and accurate dose calculations in both the traditional-Linac-based intensity-modulated radiation therapy (IMRT) and volumetric modulated arc therapy plans. The efficacy of these approaches has been demonstrated in several studies, including those by Peng *et al* (2019), Fu *et al* (2020), Kontaxis *et al* (2020), Xing *et al* (2020a, 2020b), Bai *et al* (2021), Tseng *et al* (2022). These approaches have also been extended to the MR-Linac-based IMRT dose calculation by some research groups (Tsekas *et al* 2021, Song *et al* 2022). Those results show promising prediction accuracy with impressive computation efficiency. The average gamma passing rate, using a 2%/2 mm and 10% dose threshold, ranges from 97% to 99%, and the total computation time is just a matter of seconds with the use of powerful GPU devices like the NVIDIA RTX Titan GPU. This suggests that these DL models can offer highly accurate and fast dose calculation support for real-time adaptive MRgRT. However, these engines are commonly based on the standard 3D UNet (Çiçek *et al* 2016) architecture with various modifications, such as the implementation of residual, dense, and attention-gated modules, making it challenging to fairly compare the performance of different models due to the differences in training data and computing devices used by different research groups.

In this study, our goal is to compare and evaluate the performance of different UNet variants in terms of prediction accuracy and computation efficiency, with a focus on the MR-Linac-based IMRT plans. By doing so, we hope to establish a benchmark for these models and advance their applications in the field of radiation therapy dose calculation. To comprehensively assess their feasibility of various approaches, we first developed DL-based dose engines from the original to the recently developed UNet variants, including the standard-UNet, cascade-UNet (Liu *et al* 2021), dense-dilated-UNet (Zhang *et al* 2020), residual-UNet (Zhou *et al* 2020) (Res-UNet), hierarchically densely connected-UNet (Nguyen *et al* 2019) (HD-UNet), and attention-aware-UNet (Osman and Tamam 2022). These engines were subsequently trained by the same kind of training data and the same computing devices, so the dosimetric performance and inference efficiency of the aforementioned engines can be fairly compared. Detailed information on the model implementations and comparison metrics is presented in the Methods and Materials section. A thorough assessment of the engines is conducted and discussed.

2. Methods and materials

In this section, we begin by introducing the patient and plan database in section 2.1. Section 2.2 outlines the generation of training data, while section 2.3 details the UNet-based DL architecture. The implementation and training of the model are described in section 2.4. Finally, section 2.5 presents the comparison metrics used to evaluate the performance of the UNet variants.

2.1. Patient and plan database

Thirty cases of intermediate-risk prostate adenocarcinoma (stage II) were collected, with sixteen, four, and ten randomly chosen as the training, validation, and test sets, respectively. The patients underwent nine-field IMRT

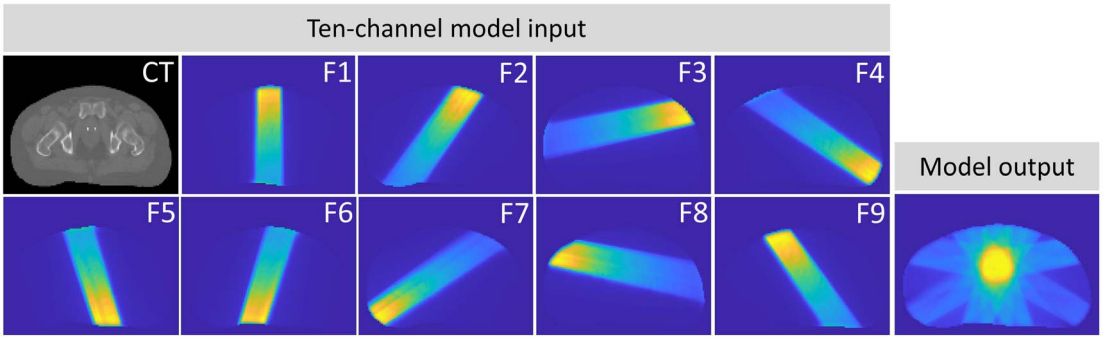

Figure 1. Visual illustrations of the ten-channel model input and model output. F1 to F9 represent the IMRT field one to field nine.

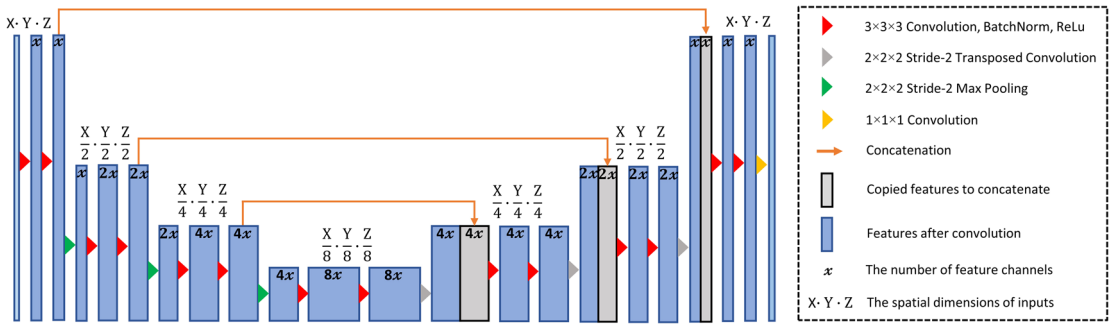

Figure 2. An illustration of the 4-level standard -UNet DL architecture.

with 6 MV beams at 40° gantry intervals, delivered on an Elekta Unity system. The target dose was 60 Gy in fractions of 3 Gy, delivered to the planning target volume (PTV). To increase the size of the training and validation datasets, the recycling strategy (Tseng *et al* 2022) was applied. Each plan was applied to different patient CT scans to calculate the corresponding doses, resulting in 320 samples in the training dataset and 80 samples in the validation dataset.

2.2. Training data generation

Ten input channels were designed for DL-based dose calculation, as shown in figure 1. The first channel is the patient CT images with a Hounsfield Unit range of 0–3000. The other nine channels are the 3D doses for each IMRT field, calculated using a simple PB convolution on a water phantom with the patient's external contour (details can be found in Tseng *et al* (2022)). For each set of inputs, the corresponding output is the composite patient dose calculated by the GPU-MC dose engine (GPUMCD) (Hissoiny *et al* 2011) of the Monaco TPS (version 5.51.11) on the Unity model, using a dose grid of $0.3 \times 0.3 \times 0.3\ \text{cm}^3$ and a statistical uncertainty of 1% per segment. The dose calculations were performed on a server equipped with dual Intel(R) 2.59 GHz Xeon (R) Gold 6240 CPUs with 128 GB RAM and an NVIDIA Tesla V100 GPU (32 GB). Before training the model, all data, including the patient CT images, 3D doses in water, and reference MC patient doses, were cropped to the same size of $144 \times 96 \times 48$ voxels with a resolution of $0.3 \times 0.3 \times 0.3\ \text{cm}^3$.

2.3. UNet-based DL architecture

The UNet-based DL architectures have been used as the primary engine for DL-based dose calculation tasks, as demonstrated by several studies (Peng *et al* 2019, Fu *et al* 2020, Kontaxis *et al* 2020, Xing *et al* 2020a, 2020b, Bai *et al* 2021, Neph *et al* 2021, Tsekas *et al* 2021, Song *et al* 2022, Tseng *et al* 2022, Xiao *et al* 2022). Figure 2 illustrates the 4-level standard-UNet DL architecture. This architecture consists of an encoder path and a decoder path. Each path has four spatial resolution levels: $X \cdot Y \cdot Z$, $\frac{X}{2} \cdot \frac{Y}{2} \cdot \frac{Z}{2}$, $\frac{X}{4} \cdot \frac{Y}{4} \cdot \frac{Z}{4}$, and $\frac{X}{8} \cdot \frac{Y}{8} \cdot \frac{Z}{8}$. where $X \cdot Y \cdot Z$ are the spatial dimensions of the 3D inputs, which are $144 \times 96 \times 48$ in this study. The levels in both paths are composed of two 3D convolutional layers with a $3 \times 3 \times 3$ kernel size and zero padding, followed by both the batch normalization and rectified linear units (ReLu). A stride-two $2 \times 2 \times 2$ max pooling layer and a stride-two 3D transposed convolutional layers with a kernel size of $2 \times 2 \times 2$ are applied to down-sample the features in the encoder path and to up-sample the features in the decoder path, respectively. The skip connections reuse

and concatenate the features from the encoder path to the decoder path for preserving fine-grained details. The last 3D convolutional layer with a kernel size of $1 \times 1 \times 1$ produces a single feature channel for final voxel-wise operation. The number of feature channels is doubled when the spatial resolution is down-sampled by a factor of two in the encoder path, whereas the number of feature channels is halved when the spatial resolution is up-sampled by a factor of two in the decoder path.

Unlike the standard-UNet, UNet variants were typically developed by implementing additional neural network modules (e.g. residual module, dense module, etc) with diverse configurations to the original UNet architecture for achieving superior model performance. These additional modules are primarily designed to capture more complex feature representations and enhance the information flow throughout the model. To thoroughly study their capabilities for the MR-Linac-based dose calculation task, the widely-used UNet-based DL architectures in the field, including the standard-UNet, cascaded-UNet (Liu *et al* 2021), dense-dilated-UNet (Zhang *et al* 2020), Res-UNet (Zhou *et al* 2020), HD-UNet (Nguyen *et al* 2019), and attention-aware-UNet (Osman and Tamam 2022), were all constructed based on their original designs. Detailed infrastructures of UNet variants can be found in the listed references and will not be repeated here.

2.4. Model implementation and training

Each of the studied models was implemented with the PyTorch DL framework (version 1.10) and individually trained on an NVIDIA A100 SXM4 GPU with 80 GB dedicated RAM. To fully understand the performance limits of these models, we increased the number of trainable parameters for each model by gradually increasing the number of feature channels and the number of resolution levels, until no further improvement was observed. The starting number of feature channels and the number of resolution levels for testing are 16 and 3, respectively. The scale expanding stopped at 64 and 5 for channels and resolution, respectively. The Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, and epsilon $= 10^{-8}$) was selected to minimize the loss function, with an initial learning rate of 0.01. A learning rate decay approach was used to reduce the learning rate by 50% if the validation loss did not improve by 10% over 50 epochs. In addition, an early stopping technique was adopted to terminate the training session when the validation loss failed to improve by 10% over 100 epochs to avoid potential overfitting. The mean square error between the DL-calculated and the reference MC doses was used as the loss function for optimization. The training mini-batch size and epoch number were set to 2 and 500, respectively. The training settings were applied consistently across all studied models.

2.5. Performance assessment

To evaluate the final performance of the UNet variants studied, we compared their dosimetric accuracy and inference efficiency for DL-based dose calculation to that of the standard UNet. First, to determine the overall dosimetric agreements between the DL-calculated and the reference MC doses on the test patient cases, a global gamma analysis was performed at 1%/1 mm, 2%/2 mm, and 3%/3 mm criteria with a 10% low-dose threshold. Paired samples t-test (one-tailed) was conducted to statistically compare the dosimetric accuracy among the studied models. Additionally, the dosimetric performance of the studied models on ERE modeling and tissue heterogeneity corrections was further evaluated using a patient case with a gas-filled rectum. Second, the inference efficiency of each model was assessed by performing dose calculations on various computing devices, including an NVIDIA A100 SXM4 GPU, an NVIDIA GTX 3080 GPU with 10 GB dedicated RAM, and an Intel (R) Core (TM) 3.5 GHz i9-11900KF CPU with 64 GB dedicated RAM. To analyze the overall performance and limitations of each model, the experimental results of the top model from each UNet-based architecture with different model resolution levels were presented and compared.

3. Results

3.1. Dosimetric results of the studied models

3.1.1. Overall dosimetric performance

Table 1 lists the gamma analysis between the DL-calculated and the reference MC doses on the test patient cases. All the 5-level DL models achieved over 86% (1%/1 mm), 98% (2%/2 mm), and 99% (3%/3 mm) average gamma passing rates. However, the calculation accuracy dropped when a lower number of model resolution levels was implemented.

Table 2 presents the test-statistic values and *p*-values for the gamma passing rate comparisons of the studied models from table 1. From a statistical point of view, the differences in dosimetric performance were generally more pronounced between models with lower-level, compared to those with higher-level. This suggests that the dosimetric accuracy among the higher-level models was more statistically consistent in comparison to the lower-level models.

Table 1. The gamma passing rates between the reference MC doses and DL-calculated doses of the studied models on the test patients.

Model	Resolution levels	Number of trainable parameters (unit: millions)	Average gamma passing rate and standard deviation (%)		
			1%/1 mm	2%/2 mm	3%/3 mm
Standard-UNet[a]	3	5.4	76.93 ± 4.76	93.66 ± 2.55	97.70 ± 1.38
	4	22.4	83.01 ± 4.24	96.84 ± 1.64	99.27 ± 0.73
	5	90.3	86.31 ± 3.32	98.29 ± 0.79	99.67 ± 0.22
Cascaded-UNet[b]	3	6.9	81.88 ± 4.61	96.19 ± 1.91	99.01 ± 1.08
	4	28.1	85.71 ± 5.27	97.40 ± 1.48	99.37 ± 0.64
	5	112	88.24 ± 3.15	98.51 ± 0.88	99.79 ± 0.27
Dense-dilated-UNet[c]	3	32.4	85.99 ± 4.27	98.04 ± 1.55	99.64 ± 0.53
	4	130	86.56 ± 4.57	98.33 ± 0.78	99.77 ± 0.18
	5	521	87.95 ± 3.03	98.46 ± 0.79	99.78 ± 0.19
Res-UNet[d]	3	11.3	83.54 ± 4.42	97.14 ± 1.53	99.34 ± 0.70
	4	46	87.02 ± 2.85	98.62 ± 0.57	99.88 ± 0.08
	5	184	88.59 ± 3.42	98.70 ± 0.73	99.88 ± 0.10
HD-UNet[e]	3	7.3	80.60 ± 6.78	95.32 ± 3.02	98.61 ± 1.54
	4	14.1	85.19 ± 4.59	97.02 ± 1.34	99.23 ± 0.52
	5	22.8	88.51 ± 3.02	98.68 ± 0.85	99.86 ± 0.29
Attention-aware-UNet[a]	3	5.5	77.10 ± 5.41	93.79 ± 2.40	97.81 ± 1.30
	4	22.5	84.20 ± 4.32	97.13 ± 1.24	99.38 ± 0.57
	5	90.7	88.42 ± 2.72	98.54 ± 0.51	99.83 ± 0.10

[a] Initial number of feature channel was 64.
[b] Initial number of feature channel was 32 and 64 for the first and second UNets, respectively.
[c] Five dilated convolutional layers (dilation rates: 2, 3, 5, 7, 9) were implemented in the bottleneck level of the dense-dilated-UNet; initial number of feature channel was 64.
[d] Two residual blocks at each level were implemented in both the encoder and decoder; initial number of feature channel was 64.
[e] All convolutional layers have a feature channel of 64.

Table 2. The statistical analysis (one-tailed paired sample t-test) of the dosimetric comparisons using the gamma passing rate results from table 1. The *p*-values that are considered statistically significant (*p*-value < 0.05) are presented in boldface.

Resolution level	Model comparisons	Paired samples t-test (one-tailed)					
		1%/1 mm		2%/2 mm		3%/3 mm	
		Test-statistic	*p*-value	Test-statistic	*p*-value	Test-statistic	*p*-value
3	Dense-dilated-UNet versus Res-UNet	3.3761	**0.0041**	2.0085	**0.0378**	1.0301	0.1649
	Dense-dilated-UNet versus Cascaded-UNet	4.7681	**0.0006**	2.7175	**0.0119**	1.5918	0.0729
	Dense-dilated-UNet versus HD-UNet	3.2156	**0.0053**	3.0616	**0.0068**	2.0951	**0.0328**
	Dense-dilated-UNet versus attention-aware-UNet	8.0711	**0.000 01**	4.9575	**0.0004**	3.8345	**0.0020**
	Dense-dilated-UNet versus standard-UNet	7.0641	**0.000 03**	5.2985	**0.0003**	4.1670	**0.0012**
	Res-UNet versus cascaded-UNet	1.4718	0.0876	1.8020	0.0525	1.0008	0.1715
	Res-UNet versus HD-UNet	1.9099	**0.0442**	2.3281	**0.0224**	1.3945	0.0983
	Res-UNet versus attention-aware-UNet	4.8036	**0.0005**	4.4420	**0.0008**	3.5586	**0.0031**
	Res-UNet versus standard-UNet	4.4676	**0.0008**	4.2886	**0.0010**	3.5667	**0.0030**
	Cascaded-UNet versus HD-UNet	1.0643	0.1575	1.8559	**0.0482**	1.4707	0.0877
	Cascaded-UNet versus attention-aware-UNet	8.1488	**0.000 01**	6.3539	**0.0001**	6.0394	**0.0001**
	Cascaded-UNet versus standard-UNet	6.1337	**0.0001**	4.5818	**0.0007**	5.2816	**0.0003**
	HD-UNet versus attention-aware-UNet	2.9253	**0.0084**	2.4856	**0.0173**	2.7346	**0.0115**
	HD-UNet versus standard-UNet	2.8923	**0.0089**	2.3615	**0.0212**	3.0006	**0.0075**
	Attention-aware-UNet versus standard-UNet	0.2064	0.4205	0.2840	0.3914	0.4540	0.3303
4	Res-UNet versus dense-dilated-UNet	0.4437	0.3339	1.2051	0.1295	2.1361	**0.0307**
	Res-UNet versus cascaded-UNet	1.1161	0.1467	2.9554	**0.0080**	2.5038	**0.0168**
	Res-UNet versus HD-UNet	1.6682	0.0648	4.5344	**0.0007**	2.8933	**0.0089**
	Res-UNet versus attention-aware-UNet	3.4074	**0.0039**	4.4869	**0.0008**	2.6974	**0.0123**
	Res-UNet versus standard-UNet	4.3249	**0.0010**	3.9321	**0.0017**	2.6720	**0.0128**
	Dense-dilated-UNet versus cascaded-UNet	1.3639	0.1029	3.0816	**0.0066**	1.8658	**0.0475**
	Dense-dilated-UNet versus HD-UNet	1.2995	0.1130	3.6695	**0.0026**	2.7457	**0.0113**
	Dense-dilated-UNet versus attention-aware-UNet	3.7404	**0.0023**	4.0959	**0.0013**	1.9330	**0.0426**
	Dense-dilated-UNet versus standard-UNet	6.2475	**0.0001**	4.2603	**0.0011**	2.2645	**0.0249**
	Cascaded-UNet versus HD-UNet	0.5603	0.2945	0.9632	0.1803	0.8356	0.2125
	Cascaded-UNet versus attention-aware-UNet	2.0046	**0.0380**	0.9605	0.1810	0.1766	0.4319
	Cascaded-UNet versus standard-UNet	4.0056	**0.0015**	1.9094	**0.0443**	0.7294	0.2422
	HD-UNet versus attention-aware-UNet	1.0488	0.1608	−0.3672	0.3611	−1.1749	0.1351
	HD-UNet versus standard-UNet	2.1160	**0.0317**	0.3380	0.3716	−0.1763	0.4320
	Attention-aware-UNet versus standard-UNet	1.8279	0.0504	0.7423	0.2384	0.7260	0.2431
5	Res-UNet versus HD-UNet	0.0805	0.4688	0.7876	0.2256	0.1117	0.4568

Table 2. (Continued.)

Resolution level	Model comparisons	Paired samples t-test (one-tailed)					
		1%/1 mm		2%/2 mm		3%/3 mm	
		Test-statistic	*p*-value	Test-statistic	*p*-value	Test-statistic	*p*-value
	Res-UNet versus attention-aware-UNet	0.2855	0.3909	1.0456	0.1615	1.0060	0.1704
	Res-UNet versus cascaded-UNet	0.5452	0.2994	1.1585	0.1382	1.4693	0.0879
	Res-UNet versus dense-dilated-UNet	1.6602	0.0656	2.7617	**0.0110**	3.2652	**0.0049**
	Res-UNet versus standard-UNet	3.1723	**0.0057**	2.2806	**0.0243**	3.8544	**0.0019**
	HD-UNet versus attention-aware-UNet	0.0934	0.4638	0.5172	0.3087	0.1524	0.4411
	HD-UNet versus cascaded-UNet	0.3575	0.3645	0.5932	0.2838	0.3497	0.3673
	HD-UNet versus dense-dilated-UNet	0.5588	0.2950	0.8347	0.2127	0.4585	0.3287
	HD-UNet versus standard-UNet	2.1760	**0.0288**	2.1772	**0.0287**	1.0767	0.1548
	Attention-aware-UNet versus cascaded-UNet	0.2580	0.4011	0.1168	0.4548	1.0580	0.1588
	Attention-aware-UNet versus dense-dilated-UNet	1.2207	0.1266	0.6015	0.2812	1.8075	0.0521
	Attention-aware-UNet versus standard-UNet	2.7551	**0.0112**	1.1246	0.1449	3.5102	**0.0033**
	Cascaded-UNet versus dense-dilated-UNet	0.4633	0.3271	0.2684	0.3972	0.4315	0.3381
	Cascaded-UNet versus standard-UNet	2.8377	**0.0097**	1.2239	0.1260	1.8977	**0.0451**
	Dense-dilated-UNet versus standard-UNet	2.6566	**0.0131**	1.8899	**0.0457**	2.6300	**0.0137**

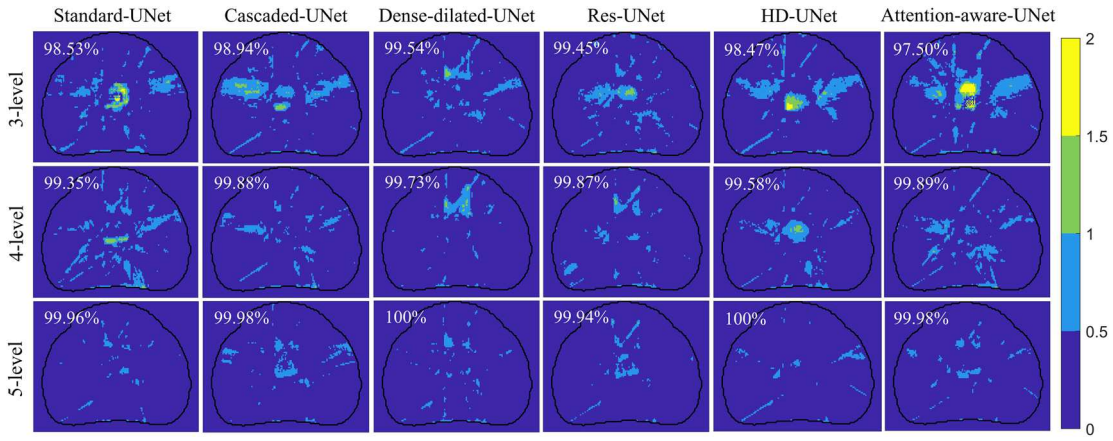

Figure 3. Gamma index map (3%/3 mm) comparisons of the central axial slice for a test patient case using the studied models with 3, 4, and 5 resolution levels. The percentage at the top left corner on each 2D map represents the gamma passing rate of the 2D slice. The color bar shows the gamma index. Gamma test fails when the gamma index is larger than 1.

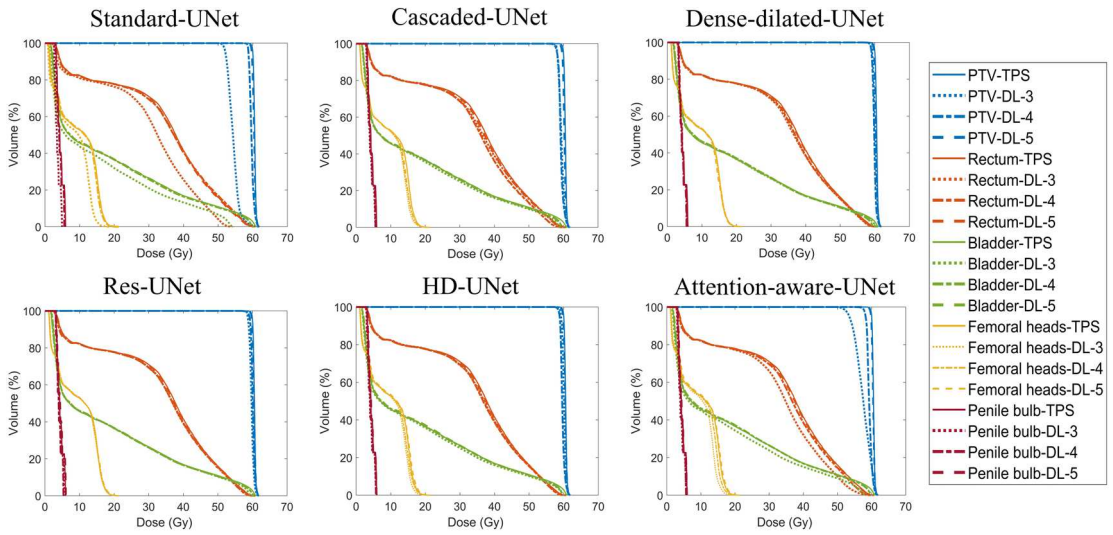

Figure 4. DVH plots of a test patient case between the reference MC and DL-calculated doses for the PTV, rectum, bladder, femoral heads, and penile bulb structures using the studied models with 3, 4, and 5 resolution levels. *-DL-3, *-DL-4, and *-DL-5 are the DVH curves of the DL-calculated doses from the model with 3-level, 4-level, and 5-level, respectively. *-TPS is the DVH curves of the MC reference doses from TPS.

Figure 3 provides visual comparisons of the 2D axial gamma index map (3%/3 mm) between the DL-calculated and the MC reference doses at the center of the PTV for a test patient case using all the studied models. Large regions (presented in the color green and yellow) that failed the gamma test can be observed for all the 3-level models. The regions were noticeably reduced as the number of model resolution levels increased.

Figure 4 presents the dose-volume histogram (DVH) curve comparisons for a test patient case. The DVH curves for the DL-calculated doses of all the 5-level models, in general, were matched well with those of the MC reference doses, while noticeable discrepancies can be seen between the DVH curves of MC reference doses and the DL-calculated doses of the 3-level and 4-level models.

3.1.2. Tissue heterogeneity correction and ERE modeling

Figure 5 provides visual comparisons of the 2D isodose distributions between DL-calculated and MC reference doses using a test patient case with a gas-filled rectum. The isodose lines of the DL-calculated doses computed by the higher-level models mostly aligned with those of the reference MC doses in both areas with tissue heterogeneity and air-tissue interfaces. While relatively larger differences were observed in the comparisons between DL-calculated (3-level models) and reference MC doses, the isodose lines of DL-calculated doses still

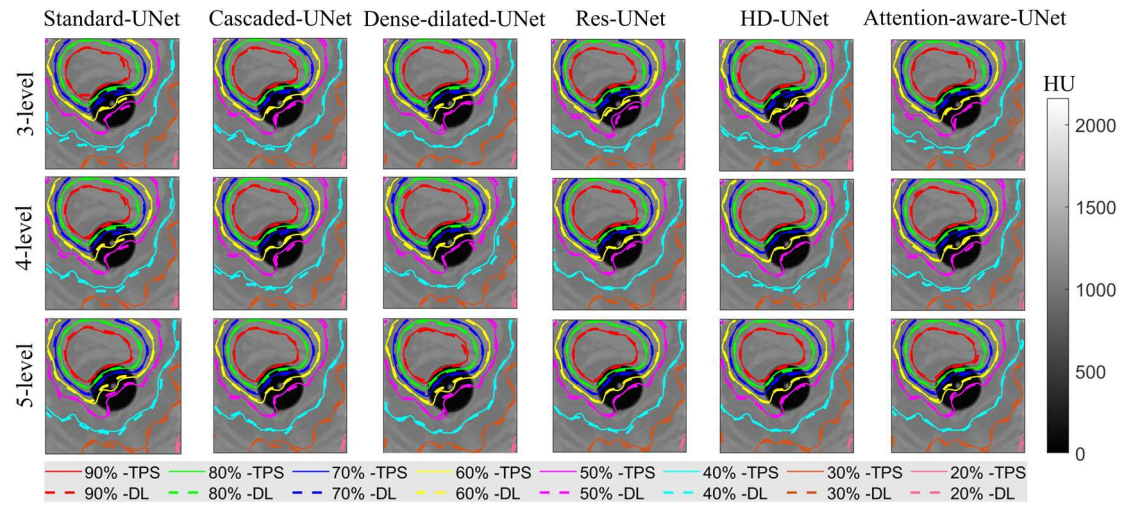

Figure 5. Comparisons of 2D dose distributions (20%–90%) between the reference MC (solid isodose lines) and DL-calculated doses (dashed isodose lines) for a test patient case with a gas-filled rectum. All the dose comparison maps and CT images were cropped to a size of 8 × 8 cm^2 to enlarge details.

Table 3. The inference times of the studied models from table 1. Batch size of 1 was used for DL-based dose calculations.

Model	Resolution levels	Average inference time per plan (unit: seconds)		
		GPU		CPU[a]
		A100[a]	RTX 3080[a]	
Standard-UNet	3	0.0266 ± 0.0004	0.0834 ± 0.0005	2.3440 ± 0.0149
	4	0.0316 ± 0.0010	0.0932 ± 0.0006	2.5918 ± 0.0266
	5	0.0440 ± 0.0014	0.1054 ± 0.0003	2.7947 ± 0.0539
Cascaded-UNet	3	0.0374 ± 0.0007	0.1186 ± 0.0003	3.3072 ± 0.0212
	4	0.0424 ± 0.0004	0.1324 ± 0.0008	3.6151 ± 0.0280
	5	0.0560 ± 0.0013	0.1485 ± 0.0006	3.8293 ± 0.0399
Dense-dilated-UNet	3	0.0363 ± 0.0005	0.1120 ± 0.0004	2.8724 ± 0.0220
	4	0.0501 ± 0.0006	0.1201 ± 0.0006	3.1391 ± 0.0232
	5	0.1066 ± 0.0013	0.1742 ± 0.0004	3.2552 ± 0.0288
Res-UNet	3	0.0442 ± 0.0005	0.1529 ± 0.0006	4.4094 ± 0.0052
	4	0.0513 ± 0.0006	0.1692 ± 0.0005	4.8604 ± 0.0136
	5	0.0737 ± 0.0026	0.1958 ± 0.0008	5.3040 ± 0.0174
HD-UNet	3	0.0802 ± 0.0003	0.2606 ± 0.0007	7.0563 ± 0.0447
	4	0.0858 ± 0.0002	0.2737 ± 0.0008	7.2603 ± 0.0477
	5	0.0896 ± 0.0004	0.2792 ± 0.0009	7.3070 ± 0.0703
Attention-aware-UNet	3	0.0310 ± 0.0005	0.0880 ± 0.0002	2.6439 ± 0.0210
	4	0.0354 ± 0.0006	0.0984 ± 0.0004	2.8226 ± 0.0099
	5	0.0477 ± 0.0011	0.1117 ± 0.0003	3.0545 ± 0.0148

[a] The computational capacity of the A100, RTX 3080, and CPU computing devices used in this study are 312, 59.5, and 0.9 trillion floating-point operations per second (TFLOPS), respectively.

followed a similar trend to the reference MC isodose lines. These findings indicate that the studied DL models are effective in handling tissue heterogeneity corrections as well as ERE modeling.

3.2. Model inference efficiency

Table 3 reports the average inference time of the studied DL models for the nine-field IMRT dose calculation. All the DL models with different model resolution levels can offer inference times approximately shorter than 0.1 s, 0.3 s, and 7.5 s per plan on the A100 GPU, RTX 3080 GPU, and CPU computing devices, respectively. In addition, all the models with a lower number of resolution levels yielded a shorter inference on average than the ones with a higher number of resolution levels.

4. Discussion

4.1. The overall performance of the UNet-based DL models

The MR-Linac-based dose calculation task was extensively tackled in this work, utilizing commonly used UNet-based DL models. Table 1 presents the statistical results of the gamma analysis, which demonstrate the ability of all studied models to accurately manage radiological dose depositions in patient anatomy for prostate IMRT plans, even under the influence of an external magnetic field. While slight differences in gamma passing rates were observed among the studied models, all of them are effective in performing DL-based dose calculations with clinically acceptable accuracy, as evidenced by a gamma passing rate of over 90% at the 3%/3 mm criterion. The dosimetric comparisons demonstrated in figure 5 further suggest that the studied models are capable of managing both the ERE modeling and tissue heterogeneity correction in the presence of significant inhomogeneous patient anatomy with clinically acceptable calculation accuracy. In addition, table 3 reports inference time comparisons that demonstrate the ability of all models to provide fast calculation speeds, taking only seconds per IMRT plan dose calculation on both GPU and CPU computing devices. Taken together, these results suggest that all studied UNet-based models are a feasible option for highly accurate and efficient dose calculations in MR-Linac-based prostate IMRT plans. Implementation of these fast and accurate dose engines in clinical practice can offer significant benefits, particularly for supporting real-time/online adaptive MRgRT.

4.2. The analysis of the studied models

To ensure a fair performance comparison of the UNet-based models under study, the experiments were designed to strictly utilize training data from the same group of patient cases, the same types of model inputs, and identical computing environment settings. To investigate the effects of network resolution on dose accuracy, three different layer structures were utilized for dose estimation. As reported in table 1, all the UNet variants achieved higher average gamma passing rates on the test patient cases than the standard-UNet for every model resolution. The implementation of enhanced neural network modules into the standard UNet architecture improved the model's calculation accuracy. Interestingly, less performance improvements were observed as the resolution of the network structure increased. For 3-level models, up to a 9% average gamma passing rate difference was observed at a 1%/1 mm criterion, whereas for 5-level models, the differences were only 2% or less. Figure 3, which includes 2D gamma index map comparisons, demonstrates that dosimetric agreement improves as the number of model resolution levels increases for all studied models. This suggests that deeper, more complex models are better able to learn accurate dose correction from water to patient anatomy in a magnetic field.

The 5-level HD-UNet performed the best among the studied models with less than 25 million trainable parameters. Other models required a much greater number of trainable parameters to achieve comparable accuracy to that of the HD-UNet. The dense-dilated-UNet demonstrated comparatively stable calculation accuracy regardless of the model resolution levels, in contrast to the larger performance gaps observed between levels for the standard-UNet and attention-aware-UNet. This can be observed in figure 4, where substantial deviations in DVH curves between different model levels are apparent for the standard-UNet and attention-aware-UNet, while the DVH curves for the other models with different levels are much closer to each other.

Table 3 shows that all UNet variants take longer to perform inference on both GPU and CPU computing devices compared to the standard-UNet, indicating reduced model inference efficiency. The inference times were approximately 1.1 times (attention-aware-UNet), 1.5 times (cascaded-UNet and dense-dilated-UNet), 2 times (Res-UNet), and 3 times (HD-UNet) longer than that of the standard-UNet. This is due to the additional DL modules implemented into the standard-UNet architecture that considerably increase the computational burden, resulting in more intensive computation. Despite decreased inference efficiency as model complexity increases, all studied models with different resolution levels provide superior calculation efficiency compared to MC simulation, such as 400 s for a nine-field IMRT plan (72 segments) on Monaco TPS using GPUMCD with a 1% statistical uncertainty per segment.

Ranking the performance of the studied models for MR-Linac-based dose calculation is challenging due to slight differences in both calculation accuracy and inference efficiency. The choice of the most suitable model for a clinic largely depends on clinical needs and available computational resources. Given that 5-level models achieve higher calculation accuracy than 3-level models, a deeper model could be clinically preferable despite its longer inference time, which is still much shorter than MC simulation. From table 2, compared to better consistency found among the 5-level UNet variants, the dosimetric accuracy differences between the 5-level standard-UNet and each of the 5-level UNet variants, overall, were found statistically significant, which reveals that the 5-level standard-UNet offers inferior dosimetric accuracy from a statistical standpoint. Despite the fact that all UNet variants outperform the standard-UNet in dosimetric accuracy, the standard-UNet has a shorter

inference time while achieving a similar level of calculation accuracy as UNet variants. Therefore, it is beneficial and applicable in scenarios where only resource-limited computing devices (e.g. CPUs) are available in a clinic.

In summary, we found that each UNet-based architecture studied is effective for MR-Linac-based prostate IMRT dose calculations with promising results. It is important to note that additional DL modules and components of the UNet variants could be beneficial in scenarios that require even higher accuracy, such as larger density gradients of media. Therefore, further investigation of their performance in dose calculations for other treatment sites, such as head and neck or lung, would be valuable for future studies.

4.3. The feasible applications of the studied models in clinical practice

The implementation of these UNet-based models in the MR-Linac treatment planning process can potentially reduce the dependency on conventional dose calculation algorithms, such as the PB convolution algorithm, which is known to have limitations in accurately accounting for magnetic fields and tissue heterogeneities. The use of DL models can, therefore, provide a more accurate and efficient alternative for dose calculation in the MR-Linac treatment planning process. However, further studies are needed to validate the clinical efficacy of these models in a larger patient cohort and to ensure their safe and reliable implementation in routine clinical practice.

5. Conclusion

Overall, this study demonstrates the potential of using UNet-based DL models as an alternative to traditional MC dose engines for MR-Linac-based dose calculations. The use of DL models could considerably improve calculation efficiency while maintaining high accuracy, making it a viable option for clinical implementation. Further studies are needed to evaluate the performance of these models for other treatment sites and to optimize the model architecture for specific clinical needs.

Acknowledgments

This retrospective study has received partial support from NSF grant CMMI-2016571 and has obtained approval from UF IRB 202002754.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary information files). Data will be available from 10 March 2023.

Conflict of interest

The authors declare no conflicts of interest for this publication.

ORCID iDs

Yu Yang https://orcid.org/0000-0002-0502-7603
Keith Furutani https://orcid.org/0000-0002-6794-2540

References

Bai T, Wang B, Nguyen D and Jiang S 2021 Deep dose plugin: towards real-time Monte Carlo dose calculation through a deep learning-based denoising algorithm *Mach. Learn.: Sci. Technol.* **2** 025033

Chu V W S, Kan M W K, Lee L K Y, Wong K C W, Tong M and Chan A T C 2021 The effect of the magnetic fields from three different configurations of the MRIgRT systems on the dose deposition from lateral opposing photon beams in a laryngeal geometry—a Monte Carlo study *Radiat. Med. Prot.* **2** 103–11

Çiçek Ö, Abdulkadir A, Lienkamp S S, Brox T and Ronneberger O 2016 3D U-Net: learning dense volumetric segmentation from sparse annotation *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*

Costa F, Doran S J, Hanson I M, Nill S, Billas I, Shipley D, Duane S, Adamovics J and Oelfke U 2018 Investigating the effect of a magnetic field on dose distributions at phantom-air interfaces using PRESAGE(®) 3D dosimeter and Monte Carlo simulations *Phys. Med. Biol.* **63** 05nt01

Eccles C L *et al* 2019 Magnetic resonance imaging sequence evaluation of an MR Linac system; early clinical experience *Tech. Innov. Patient Support Radiat. Oncol.* **12** 56–63

Fu J, Bai J, Liu Y and Ni C 2020 Fast Monte Carlo dose calculation based on deep learning *2020 13th Int. Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)* vol 2020*(17–19 October)*, pp 721–6

Green O L *et al* 2018 First clinical implementation of real-time, real anatomy tracking and radiation beam control *Med. Phys.* **45** 722–39
Hissoiny S, Raaijmakers A J, Ozell B, Després P and Raaymakers B W 2011 Fast dose calculation in magnetic fields with GPUMCD *Phys. Med. Biol.* **56** 5119–29
Jia X, Gu X, Graves Y J, Folkerts M and Jiang S B 2011 GPU-based fast Monte Carlo simulation for radiotherapy dose calculation *Phys. Med. Biol.* **56** 7017–31
Kontaxis C, Bol G H, Lagendijk J J W and Raaymakers B W 2020 DeepDose: towards a fast dose calculation engine for radiation therapy using deep learning *Phys. Med. Biol.* **65** 075013
Kurz C *et al* 2020 Medical physics challenges in clinical MR-guided radiotherapy *Radiat. Oncol.* **15** 1–16
Liu S, Zhang J, Li T, Yan H and Liu J 2021 Technical note: a cascade 3D U-net for dose prediction in radiotherapy *Med. Phys.* **48** 5574–82
Neph R, Lyu Q, Huang Y, Yang Y M and Sheng K 2021 DeepMC: a deep learning method for efficient Monte Carlo beamlet dose calculation by predictive denoising in magnetic resonance-guided radiotherapy *Phys. Med. Biol.* **66** 035022
Nguyen D, Jia X, Sher D, Lin M H, Iqbal Z, Liu H and Jiang S 2019 3D radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected U-net deep learning architecture *Phys. Med. Biol.* **64** 065020
Osman A F I and Tamam N M 2022 Attention-aware 3D U-Net convolutional neural network for knowledge-based planning 3D dose distribution prediction of head-and-neck cancer *J. Appl. Clin. Med. Phys.* **23** e13630
Pathmanathan A U *et al* 2018 Magnetic resonance imaging-guided adaptive radiation therapy: a 'game changer' for prostate treatment? *Int. J. Radiat. Oncol. Biol. Phys.* **100** 361–73
Peng Z, Shan H, Liu T, Pei X, Zhou J, Wang G and Xu X G 2019 Deep learning for accelerating Monte Carlo radiation transport simulation in intensity-modulated radiation therapy arXiv:1910.07735
Pfaffenberger A 2013 *Dose calculation algorithms for radiation therapy with an MI-integrated radiation device* Heidelberg University
Shepard D M, Earl M A, Li X A, Naqvi S and Yu C 2002 Direct aperture optimization: a turnkey solution for step-and-shoot IMRT *Med. Phys.* **29** 1007–18
Shortall J *et al* 2020 Experimental verification the electron return effect around spherical air cavities for the MR-Linac using Monte Carlo calculation *Med. Phys.* **47** 2506–15
Song T, Zhou L and Li Y 2022 Cross-engine transformation-based fast dose calculation for MRI-Linac online treatment planning *Med. Phys.* **50** 2429–37
Tsekas G, Bol G H, Raaymakers B W and Kontaxis C 2021 Deep dose: a robust deep learning-based dose engine for abdominal tumours in a 1.5 T MRI radiotherapy system *Phys. Med. Biol.* **66** 065017
Tseng W, Liu H, Yang Y, Liu C and Lu B 2022 An ultra-fast deep-learning-based dose engine for prostate VMAT via knowledge distillation framework with limited patient data *Phys. Med. Biol.* **68** 015002
Xiao F, Cai J, Zhou X, Zhou L, Song T and Li Y 2022 TransDose: a transformer-based UNet model for fast and accurate dose calculation for MR-LINACs *Phys. Med. Biol.* **67** 125013
Xing Y, Nguyen D, Lu W, Yang M and Jiang S 2020a Technical Note: a feasibility study on deep learning-based radiotherapy dose calculation *Med. Phys.* **47** 753–8
Xing Y, Zhang Y, Nguyen D, Lin M H, Lu W and Jiang S 2020b Boosting radiotherapy dose calculation accuracy with deep learning *J. Appl. Clin. Med. Phys.* **21** 149–59
Zhang J, Liu S, Yan H, Li T, Mao R and Liu J 2020 Predicting voxel-level dose distributions for esophageal radiotherapy using densely connected network with dilated convolutions *Phys. Med. Biol.* **65** 205013
Zhou J, Peng Z, Song Y, Chang Y, Pei X, Sheng L and Xu X G 2020 A method of using deep learning to predict three-dimensional dose distributions for intensity-modulated radiotherapy of rectal cancer *J. Appl. Clin. Med. Phys.* **21** 26–37
Ziegenhein P, Pirner S, Ph Kamerling C and Oelfke U 2015 Fast CPU-based Monte Carlo simulation for radiotherapy dose calculation *Phys. Med. Biol.* **60** 6097–111