Received: 25 June 2024 | Revised: 28 December 2024 | Accepted: 3 January 2025

DOI: 10.1002/mp.17628

RESEARCH ARTICLE

MEDICAL PHYSICS

Deep learning-based Monte Carlo dose prediction for heavy-ion online adaptive radiotherapy and fast quality assurance: A feasibility study

Rui He[1,2,3,4] | Jian Wang[1,3,4,5] | Wei Wu[1,3,4,5] | Hui Zhang[1,3,4] | Yinuo Liu[1,3,4,6] | Ying Luo[1,3,4,5] | Xinyang Zhang[1,3,4,5] | Yuanyuan Ma[1,3,4] | Xinguo Liu[1,3,4] | Yazhou Li[1,3,4,5,7] | Haibo Peng[2] | Pengbo He[1,3,4] | Qiang Li[1,3,4,5,8]

[1]Institute of Modern Physics, Chinese Academy of Sciences, Lanzhou, China

[2]School of Nuclear Science and Technology, Lanzhou University, Lanzhou, China

[3]Key Laboratory of Heavy Ion Radiation Biology and Medicine of Chinese Academy of Sciences, Lanzhou, China

[4]Key Laboratory of Basic Research on Heavy Ion Radiation Application in Medicine, Gansu Province, Lanzhou, China

[5]University of Chinese Academy of Sciences, Beijing, China

[6]School of Future Technology, Xi'an Jiaotong University, Xi'an, China

[7]Gansu Provincial Hospital, Lanzhou, China

[8]Putian Lanhai Nuclear Medicine Research Center, Putian, China

Correspondence
Qiang Li and Pengbo He, Institute of Modern Physics, Chinese Academy of Sciences, 509 Nanchang Road, Lanzhou 730000, Gansu Province, China.
Email: liqiang@impcas.ac.cn and hepengbo@impcas.ac.cn

Funding information
National Key Research and Development Program of China, Grant/Award Number: 2022YFC2401503; Top Talents Project of Gansu Province, Grant/Award Number: 2023-9; Key Research and Development Program of Gansu Province, Grant/Award Number: 23YFFA0010; Special Project of Science and Technology Cooperation between Hubei Province and Chinese Academy of Sciences, Grant/Award Number: 42000021817T300000050

Abstract

Background: Online adaptive radiotherapy (OART) and rapid quality assurance (QA) are essential for effective heavy ion therapy (HIT). However, there is a shortage of deep learning (DL) models and workflows for predicting Monte Carlo (MC) doses in such treatments.

Purpose: This study seeks to address this gap by developing a DL model for independent MC dose (MCDose) prediction, aiming to facilitate OART and rapid QA implementation for HIT.

Methods and Materials: A MC dose prediction DL model called CAM-CHD U-Net for HIT was introduced, based on the GATE/Geant4 MC simulation platform. The proposed model improved upon the original CHD U-Net by adding a Channel Attention Mechanism (CAM). Two experiments were conducted, one with CHD U-Net (Experiment 1) and another with CAM-CHD U-Net (Experiment 2), and involved data from 120 head and neck cancer patients. Using patient CT images, three-dimensional energy matrices, and ray-masks as inputs, the model completed the entire MC dose prediction process within a few seconds.

Results: In Experiment 2, within the Planned Target Volume (PTV) region, the average gamma passing rate (3%/3 mm) between the predicted dose and true MC dose reached 99.31%, and 96.48% across all body voxels. Experiment 2 demonstrated a 46.15% reduction in the mean absolute difference in D_5 in organs at risk compared to Experiment 1.

Conclusions: By extracting relevant parameters of radiotherapy plans, the CAM-CHD U-Net model can directly and accurately predict independent MC dose, and has a high gamma passing rate with the ground truth dose (the dose obtained after a complete MC simulation). Our workflow enables the implementation of heavy ion OART, and the predicted MCDose can be used for rapid QA of HIT.

KEYWORDS
deep learning, dose prediction, heavy ion radiotherapy, Monte Carlo simulation, online adaptive radiotherapy, rapid quality assurance

Rui He and Jian Wang contributed equally to this work and should be considered co-first authors.

1 | INTRODUCTION

In heavy ion therapy (HIT), the distinct physical properties and greater biological effectiveness of heavy ion beam allowed for precise tumor eradication with minimal damage to surrounding tissues.[1–4] Yet, daily physiological changes in tumors and normal tissues challenged the accuracy of treatment plans. This necessitates real-time or near-real-time adaptations in HIT to maintain treatment precision and efficacy. Consequently, adaptive radiotherapy (ART) and rapid quality assurance (QA) were crucial.[5–8]

ART often failed to capture a patient's real-time condition,[9–11] whereas online adaptive radiotherapy (OART) enabled real-time monitoring and adjustments, enhancing treatment precision, adaptability, efficiency, safety, and customization.[5,7] However, OART required more technical expertise and resources.[12]

Deep learning (DL) had advanced significantly in medicine, especially in image processing, pattern recognition, and predictive modeling. It excelled in image segmentation and automated contouring for rapid, high-precision tasks. Yet, a key challenge in OART implementation was the time-consuming nature of dose calculations, especially with Monte Carlo (MC) simulations used for robust optimization.[5,13]

MC simulation methods were crucial in medicine for accurately simulating nuclear interactions and Coulomb scattering, serving as the "gold standard" for independent dose calculations.[14–19] However, their stochastic nature introduced significant uncertainty in dose distributions, requiring exponentially more particles to mitigate this, which greatly increases simulation time and computational demands, making it impractical for clinical OART.[20,21]

The primary challenge was to rapidly and accurately predict dose distributions using independent MC simulations to facilitate quick radiotherapy treatment planning and QA. To address the long time consumption of MC simulations, researchers have developed accelerated MC engines using CPUs or GPUs. These engines, by simplifying some physical processes and employing parallel computing, could complete simulations in minutes. However, this solution may be out of reach for those with limited hardware resources.[22–25]

DL had shown strong capabilities in dose prediction, particularly in denoising MC-simulated dose distributions (MCDose) across various radiation therapies. Researchers employed various DL models to clarify MCDose results.[26–28] Despite these advances, generating high-noise MCDose still takes several minutes to hours. To streamline this, some researchers are refining DL models or using readily available data as inputs to predict MCDose for photons or protons more quickly. For instance, Zhang et al. predicted proton MCDose for pencil beam scanning proton therapy using beam masks,[29] and Pastor-Serrano et al. used the DoTA model to predict MCDose for single-energy proton beams in milliseconds.[30] The MC simulation platforms used in these studies are mostly MCsquare,[31] which is specifically designed for proton radiotherapy. It simplified some physical processes, accelerates proton transport simulation, and achieves results close to GATE.[32] To our knowledge, there was currently no DL model for heavy-ion independent dose calculation that can achieve heavy-ion MCDose prediction from obtaining model input information to completion in the millisecond range.

In this study, we developed an innovative DL model, called CAM-CHD U-Net, using the GATE 9.1/Geant4 MC simulation platform. This model enhanced the classic 3D U-Net architecture by incorporating Dense blocks and Channel Attention Blocks, and it utilized a cascaded dual-level DL network approach for end-to-end independent dose prediction. The input data to the model, including computed tomography (CT) images, a three-dimensional (3D) energy matrix of the radiation therapy plan, and a ray mask, were obtained in a few seconds through the script, but these inputs did not directly contain dose information, thus maintaining the independence of the model in the dose prediction task. The CAM-CHD U-Net model could predict the heavy ion MCDose distribution within 2 s, showing great potential for achieving heavy ion OART and rapid QA.This approach could potentially replace traditional measurement and computation-based validation methods, representing a significant technological advancement in the field of heavy-ion OART.

2 | MATERIALS AND METHODS

2.1 | Data composition and workflow

In order to achieve OART and fast QA of HIT, we proposed a new idea and workflow. The traditional radiotherapy plan formulation was shown in the leftmost column of Figure 1. The dose distribution obtained by the treatment planning system (TPS) could be QA by measurement or MC simulation calculation. The traditional MC simulation process was shown in the middle column of Figure 1. The MC dose obtained by MC simulation was regarded as the "gold standard" for dose calculation through the patient's CT image and the set relevant treatment parameters (such as angle, prescription dose, etc.). However, the MC simulation process was extremely time-consuming and could not be widely used in clinical practice.

We proposed an independent MC dose prediction process based on the DL model, as shown in the rightmost column of Figure 1. Only the patient's CT image and some radiotherapy plan parameters were needed to

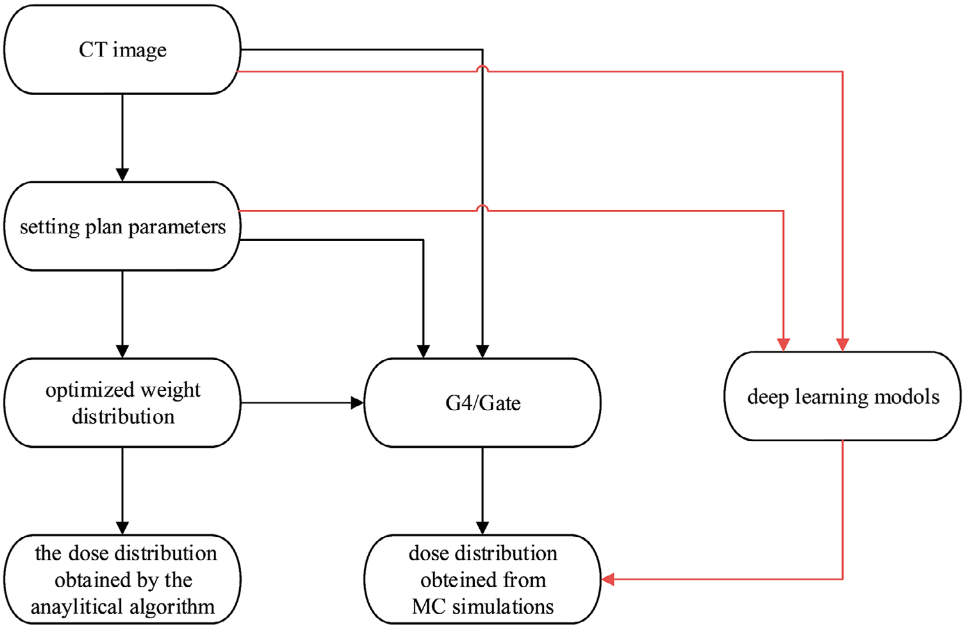
CT image
setting plan parameters
optimized weight distribution
G4/Gate
deep learning modols
the dose distribution obtained by the anaylitical algorithm
dose distribution obteined from MC simulations

FIGURE 1 Schematic diagram of the dose calculation of heavy-ion radiotherapy treatment planning system, MC simulation by GATE and dose distribution prediction by deep learning model.

directly predict the MC dose distribution. The whole process was completed by computer, taking only a dozen seconds from extracting the input information of the DL model to completing the MC dose prediction. It greatly saved time. If the predicted dose was highly consistent with the dose distribution obtained by MC simulation in gamma analysis, then the dose distribution predicted by the DL model could replace the dose distribution of traditional MC simulation for OART and fast QA of HIT.

In this study, we retrospectively selected 120 patients with head-and-neck tumors from the Wuwei Heavy Ion Hospital, China, with the research protocols approved by the institutional review board of the hospital and the Academic Committee of the Institute of Modern Physics, Chinese Academy of Sciences. The patients' CT data were randomly divided into training, validation, and testing groups in a 10:1:1 ratio. Experienced clinicians delineated the tumor target volumes and organs at risk (OARs) for all subjects. Physicists designed HIT treatment plans with single field uniform dose (SFUD) strategy for each patient, using beam angles of 90° or 270°. These plans were developed utilizing the spot-scanning beamline database of the Heavy Ion Medical Machine (HIMM) situated in Lanzhou, China. The matRad open-source research treatment planning system (TPS) was utilized to design the HIT treatment plans. All treatment plans were optimized based on the physical dose, adhering to a total prescribed dose of 30 Gy delivered in 30 fractions.[33,34] These plans were created for research purposes only and were not used in actual patient treatments.

MC simulations were conducted using the Geant4 10.7.4/Gate 9.1 platform, aligning simulation inputs with calculations from matRad, including scanning point positions, energies, and weights. Consistency in Hounsfield values and stopping power conversion factors was maintained in the GATE simulations, setting particle generation thresholds for gamma, electron, and positron at 1 m, 1 mm, and 0.1 mm, respectively, to optimize the balance between simulation time and accuracy. The QGSP_BERT_HP_EMY physics list, covering relevant hadron and electromagnetic processes for carbon ion therapy, was used. A total of 10^8 particles were tracked across the simulations performed on a 2000-core computing platform equipped with Intel(R) Xeon(R) Gold 6330 CPUs at 2.00 GHz. MC noise levels were consistently maintained below 1%, with an average of approximately 0.3%.

The specific workflow for the matRad open-source TPS, GATE MC simulation, and DL model prediction of MCDose is illustrated in Figure 1.

2.2 | Construction of 3D energy matrix

In current clinical treatments, the pencil beam algorithm is widely used for carbon-ion dose calculation. Our study aimed to directly predict the MC simulated dose distribution using treatment parameters set before the pencil beam algorithm's dose distribution calculation. To capture this information, we used the carbon ion beam dose algorithm model of our research center. Through an our custom-developed program, we extracted the relevant information from the basic calculations of the pencil beam algorithm in seconds. Therefore, we used the standard pencil beam algorithm from matRad to calculate the dose distribution for carbon ion beams.[35] This algorithm calculates dose by multiplying two key components: the depth and lateral dose distributions. Specifically, matRad uses predefined depth dose curves to describe how particle energy attenuates with depth. For the lateral dose, it models the beam's spread with a Gaussian function, where the Gaussian standard deviation (σ value) changes with depth, effectively representing the beam's energy distribution across different energy levels.

In our study, we used the carbon_Generic data files which contain the essential information like depth dose curves at various energy levels and lateral beam width data represented by Gaussian σ values. The data were obtained from the HIMM's spot-scanning beam delivery database in Lanzhou, China.[34]

We first determined the entry points of the pencil beam into the patient's body based on the vertical projection of the scanning points on the patient's body, which corresponds to the *X* and *Y* coordinates of the 3D energy matrix. The depth and energy of the pencil beam within the patient, as well as the *Z* coordinate of the 3D energy matrix and the energy value at each coordinate point, were determined using the carbon_Generic data file from the point scanning beam transmission database of HIMM in Lanzhou, China. This process can be quickly completed within seconds using a script.

In matRad, the voxel values of the patient's CT images were first converted to physical coordinates before subsequent calculations. Therefore, we converted the physical coordinates of the 3D energy matrix to match the voxel size of the CT images according to matRad's coordinate-voxel conversion formula. The position and energy magnitude of the energy in a specific layer within the patient were illustrated in Figure 2.

2.3 | Construction of ray-mask

In HIT, the incident direction of a pencil beam was critical for accurate particle deposition within the patient's body, but the 3D energy matrix could not represent these directional properties.[36] To address this, our study extracted the specific path information of a pencil beam from the treatment plan and created a binary representation: voxels along the beam path were assigned a value of 1, while those outside were set to 0. We maintained data continuity by keeping the interval between scanning points at 3 mm.

To enhance the utility of the binary mask, now referred to as the ray-mask, we expanded the contour of the ray paths by 2 mm to bridge gaps between paths. This method provides a more precise geometric description that better reflects and incorporates the beam's incidence direction into dose calculations and simulations. The construction of this ray-mask was detailed in Figure 3.

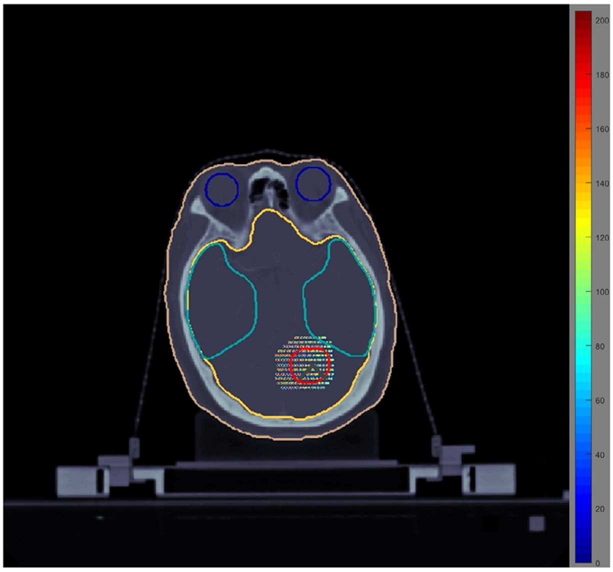

FIGURE 2 Schematic diagram of the energy distribution in a certain layer of the three-dimensional energy matrix in a human body. (The red circle is the PTV area, and the other color circles are the outline of organ segmentation.).

2.4 | Channel attention mechanism

As the number of input channels increases, in the CHD U-Net model (Cascaded Hierarchically Densely Connection 3D U-Net), Our CHD U-Net model, derived from the improved Hierarchical Dense Connectivity U-Net (HD U-Net), employed a cascading mechanism and hierarchical dense connections to enhance performance and prediction. Cascaded networks captureed and learned richer information than single-level networks. The initial input for model one (CT images, 3D energy matrix, ray-mask) and the output from the final upsampling convolution were used as the initial inputs for the model two. To further enhance its ability to process high-dimensional and complex data, we introduced a 1 × 1 × 1 convolutional layer with learnable parameters to map each channel of the input tensor to an attention weight. The calculated channel attention weights were normalized using the Sigmoid activation function to map the attention weights to the range (0, 1). The output dimension of this convolutional layer remained channels. Each original channel of the input tensor was multiplied by the corresponding attention weight to obtain the weighted channel features. The output was the feature tensor weighted by the channel attention mechanism (CAM), with the same shape as the input. In the downsampling process, the channel attention mechanism module was added after each dense convolutional module in each layer.[37–40]

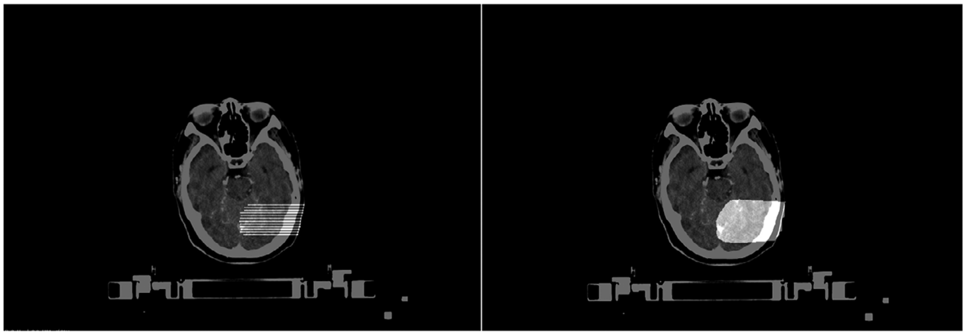

FIGURE 3 Schematic diagram of the ray-mask distribution in a certain layer in the human body (Left panel: without 2 mm ray-mask margin; right panel: with 2 mm ray-mask margin).

The CAM was a type of attention mechanism used in deep neural networks, aimed at weighting the channels of the input tensor to enable the network to focus more on important channels. Assuming the input tensor was $X \in \mathbb{R}^{C\times H\times W\times D}$, where C was the number of channels, and H, W, and D were spatial dimensions. We applied a 1 × 1 × 1 convolutional layer (linear transformation) to weight each channel, introducing learnable parameters $W \in \mathbb{R}^{C\times C}$ and bias $b \in \mathbb{R}^{C}$. The weighted result obtained was denoted as $U \in \mathbb{R}^{C\times H\times W\times D}$.

$$U = \text{Conv}1\times\ 1\ \times 1\,(X, W, b)$$

After applying global average pooling to U, we obtain the average weight on each channel:

$$\alpha_c = \frac{1}{H\times W\times D}\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{k=1}^{D} U_{cijk}$$

We then normalize the average weights using the Sigmoid activation function. Each channel of the input tensor X was multiplied by the corresponding attention weight to obtain the weighted channel features, and the result was output.

The CAM in the CAM-CHD U-Net allowed the network to autonomously prioritize important channels, enhancing its discriminative power with fewer parameters compared to traditional attention mechanisms. This significantly reduced computational and storage demands. The CAM was versatile, applicable to various DL tasks like image classification, object detection, and semantic segmentation, effectively handling high-dimensional data and improving adaptability to complex structures. The enhanced network, named CAM-CHD U-Net, is illustrated in Figure 4.

The CAM-CHD U-Net consisted of a two-stage improved HD U-Net. During the U-Net downsampling process, it included dense convolution modules, dense downsampling modules, and channel attention mechanism modules. In the upsampling process, it incorporated corresponding downsampled feature mapping modules, upsampling modules, and standard convolution modules. The dense convolution module was composed of a 3 × 3 × 3 convolutional kernel with a stride of 1, followed by InstanceNorm and ReLU activation, with the output connected to the previous feature map. The Dense Downsampling module included a 3 × 3 × 3 convolutional kernel with a stride of 1, InstanceNorm, ReLU activation, and a 2 × 2 × 2 max pooling operation with a stride of 2. The outputs of the two operations were concatenated, with the aforementioned channel attention mechanism module added at the end. The standard convolution module contained a 3 × 3 × 3 convolutional kernel with a stride of 1, followed by InstanceNorm and ReLU activation. Each level of the network performed five downsampling and five upsampling operations. During downsampling, each layer included two dense convolution modules and a channel attention mechanism module. The five layers indicated that the same dense convolution and channel attention mechanism convolution were performed five times in total. In the upsampling process, to improve efficiency, two standard convolution modules replaced the original dense convolution modules. Skip connections were used to link each upsampled feature map with its corresponding downsampled feature map. A 1 × 1 × 1 convolutional kernel was applied in the final layer to generate the ultimate three-dimensional dose prediction.

Copied features refer to features that are passed from the previous layer to the current layer. The newly calculated features are the feature maps obtained by convolving the feature maps of the previous layer. The growth rate refers to the number of feature maps (or channels) added at each layer.

2.5 | Model training details and hyperparameter settings

In our study, we conducted two experiments using the CHD U-Net architecture on CT scan images, 3D energy

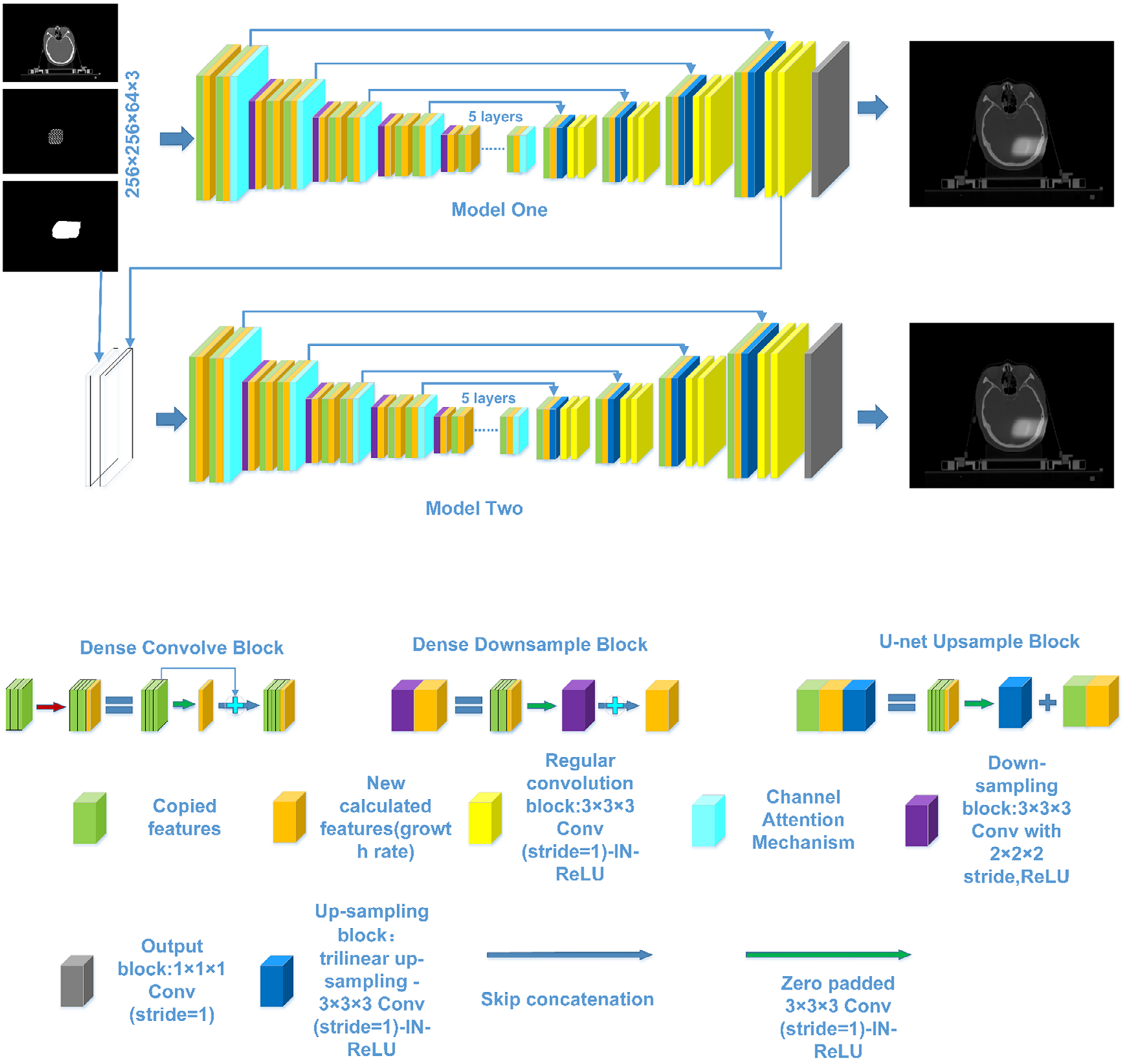
256×256×64×3
5 layers
Model One
5 layers
Model Two
Dense Convolve Block
Dense Downsample Block
U-net Upsample Block
Copied features
New calculated features(growth rate)
Regular convolution block:3×3×3 Conv (stride=1)-IN-ReLU
Channel Attention Mechanism
Down-sampling block:3×3×3 Conv with 2×2×2 stride,ReLU
Output block:1×1×1 Conv (stride=1)
Up-sampling block: trilinear up-sampling - 3×3×3 Conv (stride=1)-IN-ReLU
Skip concatenation
Zero padded 3×3×3 Conv (stride=1)-IN-ReLU

FIGURE 4 Schematic diagram of the CAM-CHD U-Net model structure.

matrices, and ray-mask images, resampled to 256 × 256 × 64 to optimize GPU usage. Experiment 1 utilized the CHD U-Net without the CAM, and Experiment 2 incorporated CAM into the CHD U-Net. Both models were developed on PyTorch, trained on an NVIDIA A6000TU GPU, and used the identical experimental conditions to ensure comparability. To ensure consistency, normalization operations were applied to CT images, cropping them to −1024 to 3500 HU, dividing by 3500 HU, and scaling to [0,1]. Each 3D energy matrix was normalized to a range between [0,1] by dividing it by its maximum energy. The mean absolute error (MAE) was used as the loss function.

To enhance the prediction accuracy of the proposed model at each level during the training process, random flips, random rotations, and random panning techniques were incorporated. These online data augmentation techniques introduce variations to the training samples, enabling the model to better handle different orientations and positions of the input data.[25] Random flips (around x and z axes, probability = 0.7) involved randomly mirroring the input data along different axes, providing additional training samples with flipped orientations. Random rotations (around the z-axis, degrees were randomly selected from {0, 40, 80, 120, 160, 200, 240, 280, 320}, probability = 0.3) introduced random rotations of the input data, simulating different angles of view. Random panning (along all three axes, probability = 0.8) involved shifting the input data within the image plane, creating variations in the spatial location.

For efficient training, both networks employed a deep supervision strategy, optimizing not just the output layer but also intermediate layers.[41] We initialized model weights using the MSRA method and employed the Adam optimizer with an initial learning rate of 0.0003 and a weight decay of 0.0001.[42] A learning rate of 0.0003 is typically a conservative choice suitable for the Adam optimizer, aiding in stable convergence. In our study, we found that a learning rate too high could cause instability, while a rate too low might lead to slow convergence. The weight decay of 0.0001 (L2 regularization) helped prevent overfitting by penalizing large weights, enhancing the model's generalization ability. A weight decay value too large could lead to underfitting due to excessive restriction on the model's learning capacity, while a value too small might not effectively prevent overfitting. Based on references, peer discussions, and our research experience, 0.0001 is a common starting value, adjustable through experiments to suit different datasets and model complexities. To avoid local optima and enhance convergence, a cosine annealing strategy was used to periodically adjust the learning rate.

Both models were trained for 60,000 iterations, with the stopping point determined by model convergence and performance. The total training time for the model was approximately 14.5 h. No underfitting or overfitting was observed during the experiments.

3 | RESULTS

In this study, we used the Dice coefficient to assess the similarity between the dose distributions predicted by the DL model and calculated with the Monte Carlo simulation (MCDose). This symmetric measure is particularly useful in medical imaging for distinguishing between the region of interest and other areas.

We analyzed the regions receiving more than 10% and 90% of the prescribed dose. The results showed that the dice coefficient for the regions over 10% of the prescribed dose was 0.93 ± 0.0073 in Experiment 1 and improved to 0.97 ± 0.0048 in Experiment 2. For the regions receiving more than 90% of the dose, the coefficient was 0.97 ± 0.0065 in Experiment 1 and increased to 0.99 ± 0.0043 in Experiment 2.

In this study, the dosimetric evaluation included the 3D gamma passing rate (GPR), dose-volume histogram (DVH) curves, and DVH indices. The DVH curves, displayed in Figure 5, show that in both experiments 1 and 2, the predicted dose model's predictions closely matched the actual MCDose across both the planned target volume (PTV) and OARs.

For quantitative analysis, we calculated the D_{95} metric for the PTV. The mean absolute difference in the PTV region was 0.0025 ± 0.0029 Gy for Experiment 1 and improved to 0.0018 ± 0.0014 Gy for Experiment 2, reflecting a 28% reduction in difference and a significant improvement in Experiment 2, as detailed in Table 1.

Regarding the OARs, we used the D_5 metric. The absolute differences in D_5 for the brainstem, left and right eyeballs, and optic nerves were 0.00013 ± 0.00011 Gy in Experiment 1 and reduced to 0.000071 ± 0.00004 Gy in Experiment 2. This represents a significant 46.15% decrease in the dose received by OARs in Experiment 2 compared to Experiment 1, demonstrating an enhanced protection of the sensitive structures.

In this study, we evaluated the dosimetric indices D_{mean} and D_{max} for HIT. The mean absolute differences in D_{mean} between the predicted dose and MCDose across the entire body improved from 0.000097 ± 0.000051 Gy in Experiment 1 to 0.000084 ± 0.000046 Gy in Experiment 2, representing a 14.43% reduction in prediction error. For D_{max}, the improvement was from 0.076 ± 0.049 Gy to 0.052 ± 0.033 Gy, showing an increase of 32.89% in accuracy.

As shown in Table 2, the 3D GPR also showed enhancements. Under the 3%/3 mm criterion for HIT, the GPR in the PTV region increased from 98.30% ± 1.34% in Experiment 1 to 99.31% ± 0.89% in Experiment 2. Across the entire body, it rose from 92.20% ± 2.24% to 96.48% ± 1.10%. Figure 6 visually illustrates these differences, highlighting the enhanced performance of the CAM-CHD U-Net in Experiment 2.

4 | DISCUSSION

In OART, achieving rapid and precise dose distribution is essential for effective treatment. Traditional heavy-ion TPS typically involved a time-consuming optimization process with multiple iterative calculations to determine the optimal beam intensity distribution. Each iteration often required a complete dose calculation, prolonging the overall process. Using MC simulations for dose calculation in OART was usually impractical due to the significant time required.

The proposed CAM-CHD U-Net model dramatically streamlined the dose prediction process in OART. It rapidly acquired input data in approximately 10 s and completed MCDose prediction in about 2 s. Therefore, the entire process, from obtaining DL model input information to dose prediction, was completed within a dozen seconds. This represented a significant improvement compared to traditional methods and other studies in this field.This rapid processing, combined with maintained accuracy in dose prediction, provides robust

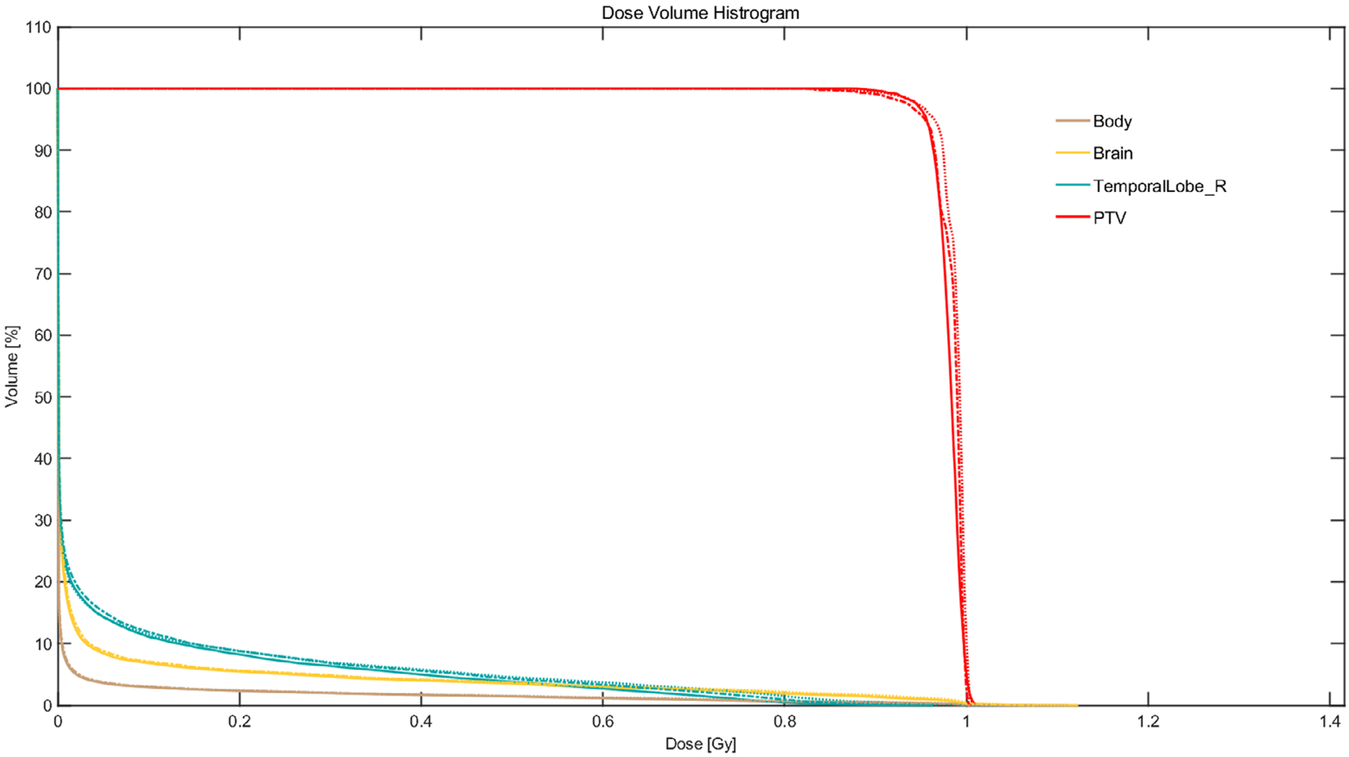

FIGURE 5 DVH curve of a test patient (the solid line is MCDose, the dotted line is predicted dose in Experiment 1, and the dotted solid line is predicted dose in Experiment 2).

TABLE 1 The absolute difference between the predicted dose and the reference dose (MCDose) for all patients in the test set in the experiments (mean ± standard deviation).

Experiment category	D_{95} (Gy)	D_5 (Gy)	D_{mean} (Gy)	$D_{\max}$ (Gy)
Experiment 1	0.0025 ± 0.0028	0.00013 ± 0.00011	0.000097 ± 0.00005	0.076 ± 0.048
Experiment 2	0.0018 ± 0.0014	0.00007 ± 0.00004	0.000083 ± 0.00004	0.051 ± 0.035
Relative difference(%)	28.00%	46.15%	14.43%	32.89%

Abbreviation = MCDose: Three-dimensional dose distribution from MC simulations, the relative difference is calculated by dividing the absolute difference between the mean of Experiment 1 and Experiment 2 by the mean of Experiment 1.

TABLE 2 Average gamma passing rate (3%/3 mm) of the predicted dose and reference dose (MCDose) for all patients in the test set in the experiments (mean ± standard deviation).

	Gamma passing rate(%)	
Experiment category	PTV	the entire body voxel area
Experiment 1	98.32 ± 1.34	92.20 ± 2.24
Experiment 2	99.31 ± 0.89	96.48 ± 1.10

Abbreviations = PTV, planned target volume; MCDose: three-dimensional dose distribution from MC simulation, the dose cut-off level was 3%.

technical support for OART, significantly enhancing the efficiency and effectiveness of treatment planning and delivery.

The technical evaluation results from Experiment 2, which incorporated the CAM module, demonstrated significant improvements in the model's performance. Specifically, the Dice coefficients, which measure the similarity between two datasets, exceeded 95% in both the 10% and 90% regions of the prescribed dose. Impressively, in the region exceeding 90% of the prescription dose, the Dice coefficient reached 0.99, indicating an extremely high degree of similarity between the predicted dose and the MCDose.

A high Dice coefficient indicates that the predicted dose distribution closely matches the dose distribution obtained through physical simulation, affirming the model's effectiveness in accurately predicting treatment doses. The integration of the CAM into the model significantly enhanced its precision in processing complex dose distribution data, thereby increasing its clinical applicability and reliability.

Experiment 2, which analyzed DVH indices such as D95 and D5, demonstrated significant improvements in dose prediction accuracy compared to Experiment 1. Specifically, the difference in D95 was reduced by

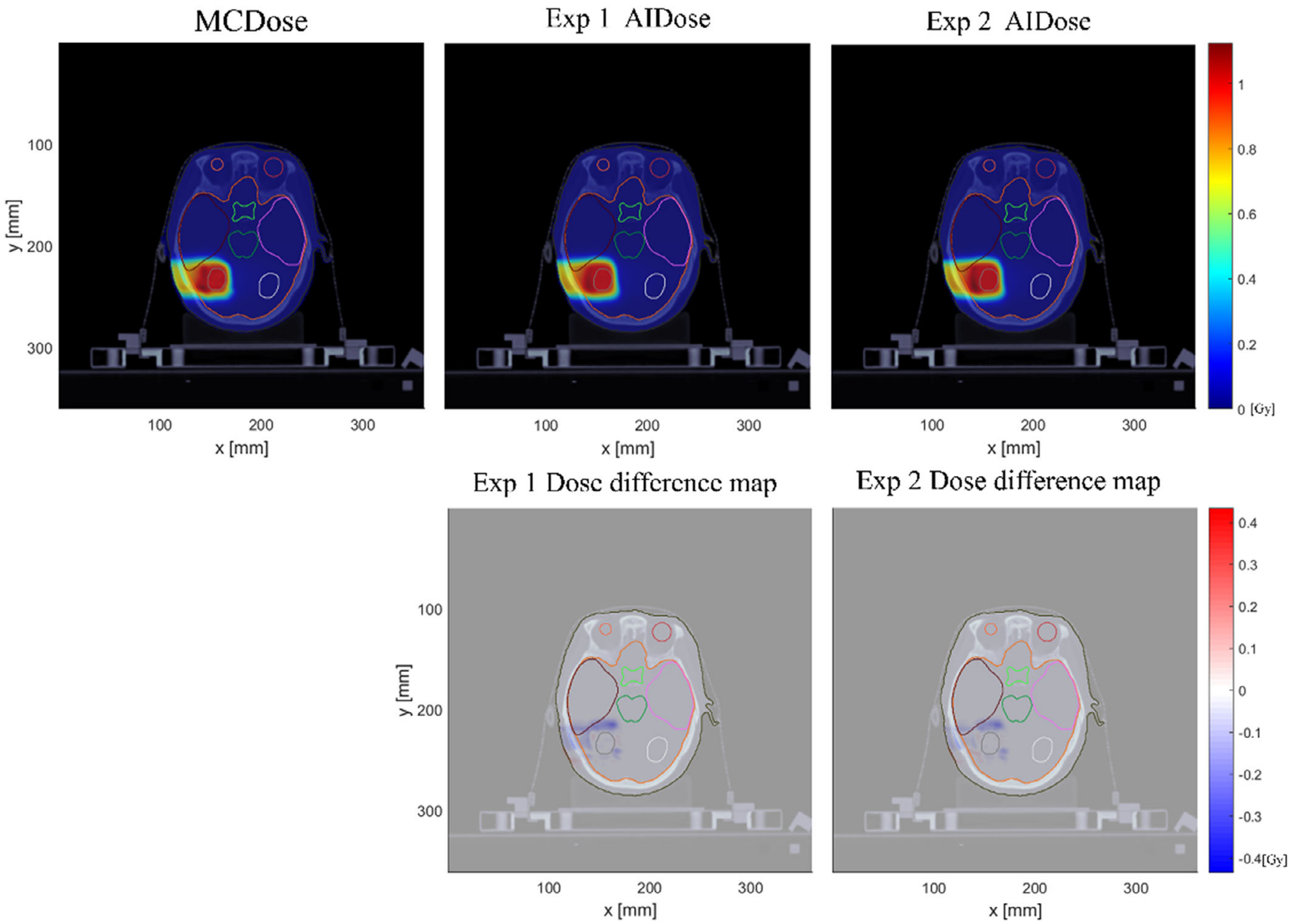

FIGURE 6 Difference between the predicted and real doses in Experiment 1 and Experiment 2.

28%, and the difference in D5 by 46.15%. These results indicated that the enhanced model offered more precise predictions of dose distributions in OARs, highlighting a substantial performance enhancement.

Further, the improvements in D_{mean} and D_{max} in Experiment 2 underscored this advancement. Notably, the prediction accuracy for D_{max} improved by 32.89% due to the incorporation of the CAM. This enhancement significantly boosted the model's sensitivity and accuracy in predicting doses to OARs.

From a clinical perspective, these improvements were crucial for protecting OARs. Minimizing radiation exposure to these areas was vital in radiation therapy, as it could substantially reduce treatment-related side effects and improve patient outcomes. The increased accuracy in dose prediction facilitated by DL models not only showcased technological progress but also played a direct role in advancing safer and more effective patient treatment plans.

In HIT treatment planning, traditional measurement-based QA methods struggled with complex human structures and were prone to errors, while MC simulation-based QA offered precise simulations but was too slow for OART. To address these issues, we introduced the CAM-CHD U-Net model, utilizing DL to quickly predict MC dose distributions, enabling rapid QA for heavy ion OART. We used the 3%/3 mm GPR, a clinical standard, to assess the consistency between predicted dose and MCDose. The results showed that in the PTV, both Experiment 1 and Experiment 2 achieved an average 3D GPR of over 95%, with Experiment 2 reaching up to 99% GPR, indicating high model accuracy. Experiment 2 also achieved over 95% GPR across all body voxels, meeting the clinical HIT standards. This model's rapid and precise dose prediction enhanced OART efficiency and enabled personalized treatment without sacrificing quality, improving patient safety and effectiveness.

This study has several limitations, the study did not incorporate Relative Biological Effectiveness (RBE) in treatment planning optimization, relying solely on physical absorbed dose, and set a uniform prescription dose of 30 Gy/30 fractions for the target volumes without considering dose constraints for OARs. Although predicted dose and MCDose showed high consistency in OARs, the model's performance was suboptimal in marginal regions outside the PTV, potentially explaining why the GPR for the entire body voxel region did not surpass 99%. Additionally, the study's generalizability is limited by difficulties in accessing patient data and a dataset containing only one tumor type. Our future study will aim to diversify the data and enhance the model accuracy to better meet clinical needs.

In conclusion, real-time and accurate dose prediction was crucial for enhancing the efficiency of OART applications. The present study developed a novel independent dose prediction DL model—the CAM-CHD U-Net model—specifically designed to directly predict MCDose. Therefore, this dose prediction method not only provided technical support for clinical OART in HIT but also opened up new possibilities for achieving rapid QA processes.

ACKNOWLEDGMENTS

This work was jointly supported by the National Key Research and Development Program of China (grant number 2022YFC2401503), the Top Talents Project of Gansu Province (grant number 2023-9), the Key Research and Development Program of Gansu Province (grant number 23YFFA0010), and the Special Project of Science and Technology Cooperation between Hubei Province and Chinese Academy of Sciences (grant number 42000021817T300000050).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

REFERENCES

1. Kraft G. Tumor therapy with heavy charged particles. *Progr Part Nucl Phys*. 2000;45:S473-S544.
2. Tsujii H, Kamada T, Shirai T, et al. *Carbon-ion radiotherapy*. Springer; 2014.
3. Jäkel O, Krämer M, Schulz-Ertner D, et al. Treatment planning for carbon ion radiotherapy in Germany: review of clinical trials and treatment planning studies. *Radiother Oncol*. 2004;73:S86-S91.
4. Shirai K, Kawashima M, Saitoh J, et al. Clinical outcomes using carbon-ion radiotherapy and dose-volume histogram comparison between carbon-ion radiotherapy and photon therapy for T2b-4N0M0 non-small cell lung cancer—a pilot study. *PLoS One*. 2017;12(4):e0175589.
5. Feng H, Patel SH, Wong WW, et al. GPU-accelerated Monte Carlo-based online adaptive proton therapy: a feasibility study. *Med Phys*. 2022;49(6):3550-3563.
6. Yan D, Vicini F, Wong J, et al. Adaptive radiation therapy. *Phys Med Biol*. 1997;42(1):123.
7. Ding Y, Feng H, Yang Y, et al. Deep-learning based fast and accurate 3D CT deformable image registration in lung cancer. *Med Phys*. 2023;50(11):6864-6880.
8. Miften M. TH-A-BRC-03: aAPM TG218: measurement methods and tolerance levels for patient-specific IMRT verification QA. *Med Phys*. 2016;43(6Part43):3852-3853.
9. Wu QJ, Li T, Wu Q, et al. Adaptive radiation therapy: technical components and clinical applications. *Cancer J*. 2011;17(3):182-189.
10. Sonke JJ, Aznar M, Rasch C. Adaptive radiotherapy for anatomical changes[C]//Seminars in radiation oncology. *WB Saunders*. 2019;29(3):245-257.
11. Wu C, Jeraj R, Olivera GH, et al. Re-optimization in adaptive radiotherapy. *Phys Med Biol*. 2002;47(17):3181.
12. Lim-Reinders S, Keller BM, Al-Ward S, et al. Online adaptive radiation therapy. *Int J Radiat Oncol Biol Phys*. 2017;99(4):994-1003.
13. Rigaud B, Anderson BM, Zhiqian HY, et al. Automatic segmentation using deep learning to enable online dose optimization during adaptive radiation therapy of cervical cancer. *Int J Radiat Oncol Biol Phys*. 2021;109(4):1096-1110.
14. Meier G, Besson R, Nanz A, et al. Independent dose calculations for commissioning, quality assurance and dose reconstruction of PBS proton therapy. *Phys Med Biol*. 2015;60(7):2819.
15. Matter M, Nenoff L, Meier G, et al. Alternatives to patient specific verification measurements in proton therapy: a comparative experimental study with intentional errors. *Phys Med Biol*. 2018;63(20):205014.
16. Johnson JE, Beltran C, Wan Chan Tseung H, et al. Highly efficient and sensitive patient-specific quality assurance for spot-scanned proton therapy. *PLoS One*. 2019;14(2):e0212412.
17. Paganetti H. Range uncertainties in proton therapy and the role of Monte Carlo simulations. *Phys Med Biol*. 2012;57(11):R99.
18. Sorriaux J, Testa M, Paganetti H, et al. Experimental assessment of proton dose calculation accuracy in inhomogeneous media. *Phys Med*. 2017;38:10-15.
19. Yang J, Li J, Chen L, et al. Dosimetric verification of IMRT treatment planning using Monte Carlo simulations for prostate cancer. *Phys Med Biol*. 2005;50(5):869.
20. Chetty IJ, Curran B, Cygler JE, et al. Report of the AAPM task group no. *Med Phys*. 2007;34(105):4818-4853.
21. Buffa FM, Nahum AE. Monte Carlo dose calculations and radiobiological modelling: analysis of the effect of the statistical noise of the dose distribution on the probability of tumour control. *Phys Med Biol*. 2000;45(10):3009.
22. Qin N, Pinto M, Tian Z, et al. Initial development of goCMC: a GPU-oriented fast cross-platform Monte Carlo engine for carbon ion therapy. *Phys Med Biol*. 2017;62(9):3682.
23. De Simoni M, Battistoni G, De Gregorio A, et al. A data-driven fragmentation model for carbon therapy gpu-accelerated monte-carlo dose recalculation. *Front Oncol*. 2022;12:780784.
24. Choi KD, Mein SB, Kopp B, et al. FRoG—a new calculation engine for clinical investigations with proton and carbon ion beams at CNAO. *Cancers*. 2018;10(11):395.
25. Souris K, Lee JA, Sterpin E. Fast multipurpose Monte Carlo simulation for proton therapy using multi-and many-core CPU architectures. *Med Phys*. 2016;43(4):1700-1712.
26. Bai T, Wang B, Nguyen D, et al. Deep dose plugin: towards real-time Monte Carlo dose calculation through a deep learning-based denoising algorithm. *Mach Learn Sci Technol*. 2021(2):025033.
27. Zhang X, Zhang H, Wang J, et al. Deep learning-based fast denoising of Monte Carlo dose calculation in carbon ion radiotherapy. *Med Phys*. 2023;50(12):7314-7323.
28. Zhang G, Chen X, Dai J, et al. A plan verification platform for online adaptive proton therapy using deep learning-based Monte–Carlo denoising. *Phys Med*. 2022;103:18-25.
29. Zhang L, Holmes JM, Liu Z, et al. Beam mask and sliding window-facilitated deep learning-based accurate and efficient dose prediction for pencil beam scanning proton therapy. *Med Phys*. 2024;51(2):1484-1498.
30. Pastor-Serrano O, Perkó Z. Millisecond speed deep learning based proton dose calculation with Monte Carlo accuracy. *Phys Med Biol*. 2022;67(10):105006.
31. Souris K, Lee JA, Sterpin E. Fast multipurpose Monte Carlo simulation for proton therapy using multi-and many-core CPU architectures. *Med Phys*. 2016;43(4):1700-1712.

32. Jan S, Santin G, Strul D, et al. GATE: a simulation toolkit for PET and SPECT. *Phys Med Biol*. 2004;49(19):4543.
33. Wieser HP, Cisternas E, Wahl N, et al. Development of the open-source dose calculation and optimization toolkit matRad. *Med Phys*. 2017;44(6):2556-2568.
34. Zhang H, Li Q, Liu X, et al. Validation and testing of a novel pencil-beam model derived from Monte Carlo simulations in carbon-ion treatment planning for different scenarios. *Phys Med*. 2022;99:1-9.
35. Hong L, Goitein M, Bucciolini M, et al. A pencil beam algorithm for proton dose calculations. *Phys Med Biol*. 1996;41(8):1305.
36. Peng Y, Liu Y, Chen Z, et al. Accuracy improvement method based on characteristic database classification for imrt dose prediction in cervical cancer: scientifically training data selection. *Front Oncol*. 2022;12:808580.
37. Li H, Qiu K, Chen L, et al. SCAttNet: semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images. *IEEE Geosci Rem Sens Lett*. 2020;18(5):905-909.
38. Huang G, Zhu J, Li J, et al. Channel-attention U-Net: channel attention mechanism for semantic segmentation of esophagus and esophageal cancer. *IEEE Access*. 2020;8:122798-122810.
39. Wang Q, Wu B, Zhu P, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020:11534-11542.
40. Niu Z, Zhong G, Yu H. A review on the attention mechanism of deep learning. *Neurocomputing*. 2021;452:48-62.
41. Kingma DP, Adam BaJ, : A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.
42. Huang H, Flynn NM, King JH, et al. Comparisons of community-associated methicillin-resistant *Staphylococcus aureus* (MRSA) and hospital-associated MSRA infections in Sacramento, California. *J Clin Microbiol*. 2006;44(7):2423-2427.

How to cite this article: He R, Wang J, Wu W, et al. Deep learning-based Monte Carlo dose prediction for heavy-ion online adaptive radiotherapy and fast quality assurance: A feasibility study. *Med Phys.* 2025;52:2570–2580. https://doi.org/10.1002/mp.17628