Physics in Medicine &
Biology

On factors that influence deep learning-based dose prediction of head and neck tumors

To cite this article before publication: Ruochen Gao *et al* 2025 *Phys. Med. Biol.* in press https://doi.org/10.1088/1361-6560/adcfeb

On Factors that Influence Deep Learning-Based Dose Prediction of Head and Neck Tumors

Ruochen Gao[1], Prerak Mody[1], Chinmay Rao[1], Frank Dankers[2] and Marius Staring[1,2]

[1] Division of Image Processing, Department of Radiology, Leiden University Medical Center, Leiden, the Netherlands

[2] Department of Radiation Oncology, Leiden University Medical Center, Leiden, The Netherlands

E-mail: r.gao@lumc.nl

Abstract.

Objective. This study investigates key factors influencing deep learning-based dose prediction models for head and neck cancer radiation therapy (RT). The goal is to evaluate model accuracy, robustness, and computational efficiency, and to identify key components necessary for optimal performance.

Approach. We systematically analyze the impact of input and dose grid resolution, input type, loss function, model architecture, and noise on model performance. Two datasets are used: a public dataset (OpenKBP) and an in-house clinical dataset (LUMC). Model performance is primarily evaluated using two metrics: dose score and dose-volume histogram (DVH) score.

Main results. High-resolution inputs improve prediction accuracy (dose score and DVH score) by 8.6–13.5% compared to low resolution. Using a combination of CT, planning target volumes (PTVs), and organs-at-risk (OARs) as input significantly enhances accuracy, with improvements of 57.4–86.8% over using CT alone. Integrating mean absolute error (MAE) loss with value-based and criteria-based DVH loss functions further boosts DVH score by 7.2–7.5% compared to MAE loss alone. In the robustness analysis, most models show minimal degradation under Poisson noise (0–0.3 Gy) but are more susceptible to adversarial noise (0.2–7.8 Gy). Notably, certain models, such as SwinUNETR, demonstrate superior robustness against adversarial perturbations.

Significance. These findings highlight the importance of optimizing deep learning models and provide valuable guidance for achieving more accurate and reliable radiotherapy dose prediction.

1. Introduction

Radiation therapy (RT) is one of the primary treatment modalities for head and neck cancers, aiming to deliver high doses of radiation to the planning target volume (PTV) while minimizing exposure to surrounding organs at risk (OARs). Due to the irregular shape of PTV, the multiple dose levels of PTV (such as primary tumors and lymph nodes) and the close proximity of OARs, head and neck cancer is considered one of the most complex clinical treatment areas (Morgan & Sher 2020).

Modern RT techniques, such as Intensity-Modulated Radiotherapy (IMRT) and Volumetric Modulated Arc Therapy (VMAT), rely on an inverse planning process. This process, which involves iterative adjustment of treatment objectives for the tumor and OARs, can be time-consuming and heavily dependent on the planner's expertise, leading to significant variability in plan quality (Nelms et al. 2012).

To enhance the consistency, quality, and efficiency of RT plans, Knowledge-Based Planning (KBP) has been introduced to automate the process of RT plan generation. The KBP approach leverages knowledge from prior high-quality, clinically accepted plans to predict optimal RT plans for new patients. Traditional KBP models use anatomical and geometric features to predict the Dose-Volume Histogram (DVH) and the dose distribution (Babier et al. 2018, Jiao et al. 2019, Bai et al. 2020). However, with the rise of deep learning in recent years, deep learning-based dose prediction methods no longer require manually designed features as input. Instead, they can learn directly from raw image data, marking a new stage in the development of KBP (Momin et al. 2021). Typically, the dose distribution predicted by deep learning-based methods is then mimicked to create a clinically deliverable plan (Bakx et al. 2021).

Segmentation models based on the encoder-decoder architecture have become dominant in medical imaging segmentation tasks (Milletari et al. 2016, Li et al. 2018, Isensee et al. 2021, Hatamizadeh et al. 2021, Cao et al. 2022). These models take inputs, such as CT or MRI images, and output segmentation masks with matching dimensions, a feature that also makes them suitable for dose prediction tasks. This compatibility has led many researchers to adopt the encoder-decoder structure for dose prediction. For example, Kearney et al. (2018) introduced DoseNet, a fully convolutional volumetric network, to predict the dose distribution for prostate cancer. Nguyen et al. (2019) proposed a hierarchically densely connected U-Net (HDUNet), which combines U-Net and DenseNet to predict the dose distribution for the head and neck. In 2020, the American Association of Physicists in Medicine (AAPM) launched the Open Knowledge-Based Planning Challenge (OpenKBP), significantly advancing the field of radiation dose prediction by providing a standardized open source dataset for head and neck cancer (OpenKBP dataset) along with standardized evaluation metrics (Babier et al. 2021). This initiative has spurred the development of various innovative methods, broadly categorized into two main approaches: (1) U-Net-based methods and (2) Transformer-based methods. In the first category, numerous studies have leveraged U-Net and its variants to predict dose distributions. Gronberg et al. (2021) proposed a

three-dimensional (3D) densely connected U-Net with dilated convolutions to capture complex spatial relationships in the data. Liu et al. (2021) presented a 3D cascaded U-Net model (C3D) with a knowledge distillation technique, achieving first place in the OpenKBP Challenge. Additionally, Wang et al. (2022) proposed a beam-wise dose decomposition 3D cascaded network, while Chandran et al. (2023) developed MemU-net, which integrates MemNet memory blocks within a U-Net framework to improve network efficiency and prediction accuracy. Recently, Lin et al. (2024) introduced the LENAS framework, an approach that incorporates Neural Architecture Search (NAS) and knowledge distillation to optimize the search for effective components in U-Net. In the second category, transformer architectures have recently emerged as a promising alternative in head and neck dose prediction, capitalizing on their ability to capture long-range dependencies. Hu et al. (2023) introduced the TrDosePred framework, pioneering the use of Transformer architectures for this task. Following this, Gheshlaghi et al. (2024) proposed DOSE-PYFER, a cascaded Transformer-based model specifically designed to predict the dose distribution. However, most of these studies focus mainly on proposing a new model structure to improve accuracy.

In addition to innovations in model architecture, only a few studies have explored other factors that influence the accuracy of dose prediction. For example, Nguyen et al. (2020) studied the impact of the loss function, suggesting that incorporating a domain-specific DVH loss can further enhance the model performance. Gu et al. (2023) investigated the impact of model input on performance, finding that CT and PTV are sufficient to predict dose distribution. However, these investigations are often isolated and based on private datasets, which limits the generalizability of their findings. Consequently, a comprehensive analysis of the factors that influence deep learning-based dose prediction across various datasets is still needed.

To facilitate clinical translation, research should go beyond accuracy to rigorously evaluate both robustness and computational efficiency. Robustness is crucial for managing various forms of noise, including inherent image noise and adversarial disturbances. In particular, adversarial noise, potentially introduced by malicious attacks, such as ransomware, poses a significant threat to hospital and patient safety (Finlayson et al. 2019). However, computational efficiency, which encompasses hardware requirements and runtime, is equally critical to ensure that models can be feasibly deployed within clinical settings. Although accuracy often takes precedence in most studies, there is a noticeable gap in assessing how models handle noise and perform efficiently. Bridging this gap is essential to develop dose prediction models that are not only accurate but also robust and practically viable in a clinical context.

Figure 1 provides an overview of the deep learning-based dose prediction pipeline and highlights the key factors analyzed in this study. In this paper, we make two primary contributions:

- **Comprehensive Analysis of Influential Factors:** We conduct a systematic investigation into various factors that influence deep learning-based dose prediction for head and neck tumors. These factors include input resolution, dose grid

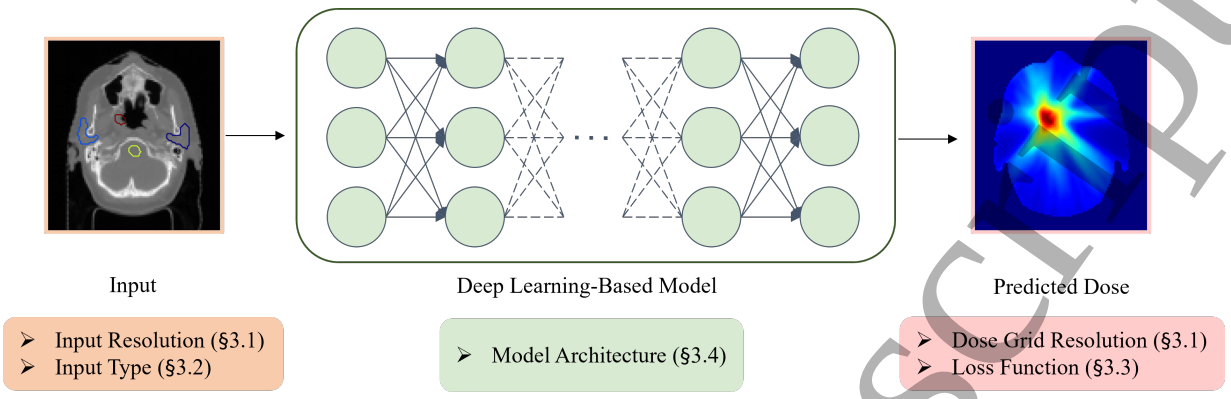

Figure 1: Pipeline of deep learning-based dose prediction and focus of this paper.

resolution, input type, loss function, and model structure. By analyzing these factors using both a public dataset and an in-house clinical dataset, we provide a holistic understanding of how each factor impacts the performance of dose prediction.

- **Evaluation of Model Robustness and Efficiency:** In addition to focusing on the accuracy of the prediction, we evaluate the robustness of different model structures against Poisson noise and adversarial noise. Furthermore, we assess the computational efficiency of these models, including GPU memory consumption and runtime. This dual focus on robustness and efficiency addresses practical concerns that are critical for clinical applications.

2. Materials and method

2.1. Patient dataset

In this study, we used two datasets. The first is the OpenKBP dataset, which contains data from 340 head and neck cancer patients: 200 for training, 40 for validation, and 100 for testing (Babier et al. 2021). The second is an in-house dataset (referred to as LUMC) of 104 patients treated for oropharyngeal and hypopharyngeal cancer at Leiden University Medical Center between 2017 and 2024. The study was approved by the Medical Ethics Committee of Leiden, The Hague, and Delft (G21.142, October 15, 2021). Patient consent was waived due to the retrospective nature of the study.

The two datasets are very similar in composition. Each patient's data includes a planning CT, PTV contours, OAR contours, and the corresponding 3D dose distribution. There are several differences. For OpenKBP, the dose distribution was generated from a 9-beam equidistant coplanar IMRT set-up with 6 MV fields. For LUMC, the dose distribution was generated from dual-arc full rotation VMAT beams. The PTVs in OpenKBP include three dose levels: PTV_{56}, PTV_{63}, and PTV_{70}, representing 56 Gy, 63 Gy, and 70 Gy, respectively. In contrast, LUMC's PTVs include two dose levels: $\text{PTV}_{54.25}$ and PTV_{70}, representing 54.25 Gy and 70 Gy. In addition, the OpenKBP

Table 1: Comparison between the OpenKBP and LUMC dataset

Dataset	Plan Type	PTVs	OARs	Dose Grid Resolution
OpenKBP	IMRT	PTV_{70}, PTV_{63}, PTV_{56}	Brainstem, Spinal cord, Right parotid, Left parotid, Esophagus, Larynx, and Mandible	(3.9, 3.9, 2.5)
LUMC	VMAT	PTV_{70}, $PTV_{54.25}$	Brainstem, Spinal cord, Right parotid, Left parotid, Esophagus, and Larynx	(2, 2, 2)

dataset has a lower dose grid resolution compared to the LUMC dataset. For detailed comparisons, see table 1.

2.2. Data preprocessing and training strategy

We used the same data preprocessing and training strategies for both the OpenKBP and LUMC datasets. First, during data pre-processing, we followed Liu et al. (2021) to clip the intensity of the CT images to the range $[-1024, 1500]$ Hounsfield Units (HU). We then normalized the values by dividing by 1000 HU. For the dose distribution, we performed normalization by dividing by 70 Gy. For dataset splitting, we adhered to the original settings for the OpenKBP dataset. For the LUMC dataset, we used three-fold cross-validation.

During training, we used online data augmentation to avoid overfitting, including random translation, random flipping, and random rotation along the cranio-caudal axis. We used the AdamW optimizer with an initial learning rate of 3×10^{-4}. All models presented in this study were trained for a maximum of 1000 epochs, with an early stopping employed to prevent overfitting.

2.3. Evaluation metrics

2.3.1. Accuracy metric Following the OpenKBP-2020 AAPM Grand Challenge, we evaluate the accuracy of the model using the Dose score and the DVH score (Babier et al. 2021). The Dose score (in Gy) is measured by the mean absolute error (MAE) between the actual and predicted doses. The DVH score (in Gy), specific to radiation therapy, includes criteria for both PTVs and OARs. For PTVs, the DVH score covers three criteria: $D_{1\%}$, $D_{95\%}$, and $D_{99\%}$, indicating the minimum doses received by 1%, 95%, and 99% of the volume, respectively. For OARs, the DVH score involves calculating D_{mean} (the mean dose received by volume) and $D_{0.1cc}$ (the near maximum dose received by 0.1 cc of volume). The Dose score and the DVH score are defined as follows:

$$\text{Dose score} = \frac{1}{N} \sum_{n} \left| \hat{D}_n - D_n \right|, \tag{1}$$

where $\hat{D}_n$ and D_n denote the predicted dose and the ground truth dose for the n-th voxel, respectively, and N is the total number of voxels.

$$\text{DVH score} = \frac{1}{3P+2O}\left(\sum_{p=1}^{P}\left(\left|\hat{D}^p_{1\%} - D^p_{1\%}\right| + \left|\hat{D}^p_{95\%} - D^p_{95\%}\right| + \left|\hat{D}^p_{99\%} - D^p_{99\%}\right|\right) + \sum_{o=1}^{O}\left(\left|\hat{D}^o_{\text{mean}} - D^o_{\text{mean}}\right| + \left|\hat{D}^o_{0.1cc} - D^o_{0.1cc}\right|\right)\right) \tag{2}$$

where P is the total number of PTVs and O is the total number of OARs.

In addition to using the aggregate Dose Score and DVH Score as the main accuracy metrics in this study, we also use individual clinical dose evaluation metrics for comparison with the clinical plan, including $V_{95\%}$ (the percentage of a volume that receives at least 95% of the prescribed dose), D_{mean} and $D_{0.03cc}$ (the near maximum dose received by 0.03 cc of volume). It is important to note that the OpenKBP dataset does not include a clinically approved plan; therefore, clinical dose evaluation metrics are only calculated for the LUMC dataset.

2.3.2. Robustness metric To evaluate the robustness of the model, we measure the change in accuracy of the model before and after introducing noise. To note, the noise is added during the model inference stage without retraining the model. Robustness is quantified using ΔDose score and ΔDVH score, which represent the differences in the accuracy metrics between noise-free conditions and those under noisy conditions. Higher values of ΔDose score and ΔDVH score indicate poorer robustness, reflecting higher sensitivity to noise. They are defined as follows:

$$\Delta\text{Dose score} = \text{Dose score}^* - \text{Dose score}, \tag{3}$$

$$\Delta\text{DVH score} = \text{DVH score}^* - \text{DVH score}. \tag{4}$$

Here, the asterisk (*) denotes the scores calculated under noisy conditions, while the unmarked scores represent those obtained under noise-free (normal) conditions.

2.3.3. Computational efficiency metric To evaluate the computational efficiency of the model, we prioritize two key metrics: GPU memory usage and GPU runtime, rather than focusing on the number of model parameters and floating point operations (FLOPs). This choice is made because GPU memory usage and runtime are more relevant for clinical deployment scenarios, as they accurately represent the hardware resources needed and the actual computation time. Lower GPU memory usage and shorter runtime indicate a more computationally efficient model. To note, our evaluation focuses specifically on the model inference, without including data loading or preprocessing, to better reflect the model's computational efficiency. All tests were performed on an NVIDIA A100 40GB GPU, and we report the average usage of GPU memory and runtime to predict a 3D dose distribution.

2.4. Input and dose grid resolution

Typically, the input resolution is set to match the dose grid resolution. This means that we determine the input resolution by defining the dose grid resolution. However, due to the large size of the 3D dose distribution, many studies employ interpolation algorithms to down-sample the dose distribution to use as ground truth (Babier et al. 2021). As a result, the model predicts a low-resolution dose distribution, which should be upsampled to the original resolution for clinical use. Although this approach reduces hardware requirements and speeds up training, it introduces the risk of potential interpolation-related errors.

To evaluate the influence of dose grid resolution on dose prediction, we use a cubic-spline interpolation algorithm to resample the dose distribution in different resolutions. Since the resolution of the OpenKBP dataset is low by default, we conducted this experiment using only the LUMC dataset, which is of higher resolution. We evaluated dose grid resolutions at three levels: $2 \times 2 \times 2$ mm^3 (high and original resolution), $3 \times 3 \times 3$ mm^3 (medium resolution), and $4 \times 4 \times 4$ mm^3 (low resolution).

2.5. Input type

Selecting the right input is essential for deep learning-based dose prediction models. Before developing an RT plan, we typically have anatomical information such as the patient's planning CT, PTV contours, and OARs contours. This information identifies critical areas to treat and high-risk organs to avoid. Using these data as input enables the neural network to learn complex spatial relationships and dose distribution patterns.

To assess how different types of input affect dose prediction accuracy, we chose four types of input: (1) CT only, (2) PTVs and OARs, (3) CT combined with PTVs, and (4) CT combined with PTVs and OARs.

2.6. Loss function

The loss function of neural networks plays a crucial role during training because it quantifies the difference between the predicted output and the ground truth. By minimizing the loss function, the model continuously adjusts its parameters, thereby enhancing its prediction accuracy. Most prior work uses the mean squared error (MSE) or MAE loss (Kearney et al. 2018, Nguyen et al. 2019, Liu et al. 2021, Chandran et al. 2023, Gheshlaghi et al. 2024). Nguyen et al. (2020) proposed making the DVH a differentiable target and introduced a value-based DVH loss function. However, this approach requires a high computational overhead. Wang et al. (2022) proposed two more efficient DVH-based loss functions, including a value-based DVH loss function ($\mathcal{L}_{vDVH}$) and a criteria-based DVH loss function ($\mathcal{L}_{cDVH}$), defined as follows:

$$\mathcal{L}_{vDVH} = \sum_{s=1}^{P+O} \frac{1}{N_s} \sum_{n=1}^{N_s} \left| R(\hat{D} \cdot W_s)_n - R(D \cdot W_s)_n \right|, \tag{5}$$

where N_s is the number of voxels in the s-th ROI (W_s), and $R(\cdot)$ denotes the sorting operation,

$$\mathcal{L}_{cDVH} = \frac{1}{3P+2O}\left(\sum_{p=1}^{P}\left(\left|\hat{D}^p_{1\%} - D^p_{1\%}\right| + \left|\hat{D}^p_{95\%} - D^p_{95\%}\right| + \left|\hat{D}^p_{99\%} - D^p_{99\%}\right|\right) + \sum_{o=1}^{O}\left(\left|\hat{D}^o_{\text{mean}} - D^o_{\text{mean}}\right| + \left|\hat{D}^o_{0.1cc} - D^o_{0.1cc}\right|\right)\right), \tag{6}$$

where these DVH criteria are calculated by the sorted dose values. We interpret $\mathcal{L}_{vDVH}$ as penalizing the discrepancy between the predicted and ground-truth DVH curves, while $\mathcal{L}_{cDVH}$ focuses on penalizing differences specifically at critical points of the DVH curve.

To evaluate the influence of different loss functions on the dose prediction task, we conducted three sets of experiments: (1) using $\mathcal{L}_{MAE}$ alone, (2) using $\mathcal{L}_{MAE} + \mathcal{L}_{vDVH}$, and (3) using $\mathcal{L}_{MAE} + \mathcal{L}_{vDVH} + \mathcal{L}_{cDVH}$. For (2), the loss weight parameters are set to 1 and 1, and for (3), they are set to 1, 0.5, and 0.5.

2.7. Model structure

For model structure selection, we focus on 3D neural networks proposed in the existing literature, as 2D neural networks often overlook spatial information between slices. There are several representative works in the literature based on UNet for dose prediction, such as DoseNet (Kearney et al. 2018), HDUNet (Nguyen et al. 2019), C3D (the champion of the OpenKBP challenge, (Liu et al. 2021)), and U-NAS (Lin et al. 2024). In recent years, transformer-based architectures have also been investigated, such as SwinUNETR (Hatamizadeh et al. 2021) and Dose-PYFER (Gheshlaghi et al. 2024), which we also include. We selected these networks because of their superior performance in dose prediction tasks, making them strong candidates for this study. It is important to note that C3D and Dose-PYFER have cascaded structures, while U-NAS is available in both single and cascaded setups.

2.8. Noise

To evaluate the robustness of the model structure, we introduce noise, including Poisson noise and adversarial noise, into the input CT image.

2.8.1. Poisson noise Poisson noise frequently appears in CT images due to the nature of the photon detection process (Thanh et al. 2019). In CT imaging, the emitted X-ray photons are detected after passing through the body. However, the number of detected photons follows a Poisson distribution, which leads to statistical fluctuations in the image. The probability mass function (PMF) of a Poisson distribution is given by $P(k;\lambda) = \lambda^k e^{-\lambda}/k!$, where λ represents the average rate (i.e., the mean number of

detected photons), and k is a discrete random variable representing the actual number of photons detected. In this study, we set the value of $\lambda = 20$ based on visual inspection.

2.8.2. Adversarial noise Research indicates that deep learning models are susceptible to adversarial noise (Madry et al. 2018). By adding a small, carefully designed perturbation to the original image, which may not be perceptible to humans, the model can produce outputs with a certain degree of error, leading to a significant decline in its accuracy. In this study, we focus on two commonly used methods to generate adversarial noise.

Projected Gradient Descent (PGD) (Kurakin et al. 2018) is an iterative method that generates an adversarial example by taking multiple small steps. For a single step, given input data x with its true dose distribution D, the adversarial example $\overline{x}$ is generated by perturbing the input in the direction that maximizes the loss with respect to D. The perturbation is updated by projecting it onto a specified ϵ-ball (a small region around the original data controlled by the parameter ϵ) while ensuring it maximizes the loss function:

$$\overline{x}^{(t+1)} = \text{Clip}_{x,\epsilon}\left(\overline{x}^{(t)} + \alpha \cdot \text{sign}(\nabla_x J(\theta, \overline{x}^{(t)}, D))\right), \tag{7}$$

where t is the iteration index, α is the step size, $J(\theta, \overline{x}^{(t)}, D)$ is the loss function, and $\text{Clip}_{x,\epsilon}$ projects $\overline{x}$ back into the ϵ-neighborhood of x.

Momentum Iterative FGSM (Mi-FGSM) (Dong et al. 2018) improves on FGSM (Goodfellow et al. 2015) by introducing momentum in gradient updates, helping to escape local optima more effectively. The update procedure becomes:

$$g^{(t+1)} = \mu \cdot g^{(t)} + \frac{\nabla_x J(\theta, \overline{x}^{(t)}, D)}{\|\nabla_x J(\theta, \overline{x}^{(t)}, D)\|_1}, \tag{8}$$

$$\overline{x}^{(t+1)} = \text{Clip}_{x,\epsilon}\left(\overline{x}^{(t)} + \alpha \cdot \text{sign}(g^{(t+1)})\right), \tag{9}$$

where $g^{(t)}$ represents the accumulated gradient momentum up to iteration t, μ is the momentum parameter that controls the influence of previous gradients, and $\|\cdot\|_1$ denotes the L1 norm, which sums the absolute values of the gradient components. In this study, we set $\epsilon = 16$ HU based on visual inspection.

3. Results

3.1. Input and dose grid resolution

For the experiment evaluating the impact of different input and dose grid resolutions on dose prediction, we used the C3D model, the winning solution of the OpenKBP Challenge. The model input consisted of CT, PTVs, and OARs, and we employed the MAE loss function. Table 2 reports the accuracy on the LUMC dataset. The results indicate that using high and original resolution input ($2 \times 2 \times 2$ mm^3) yielded the lowest Dose score and DVH score. Therefore, we selected this resolution for subsequent experiments on the LUMC dataset.

Table 2: Accuracy comparison of different input and dose grid resolution

Resolution	LUMC	
	Dose score (Gy) ↓	DVH score (Gy) ↓
$4 \times 4 \times 4$ mm^3 (low resolution)	1.39 ± 0.28	1.85 ± 0.45
$3 \times 3 \times 3$ mm^3 (medium resolution)	1.37 ± 0.26	1.67 ± 0.49
$2 \times 2 \times 2$ mm^3 (high and original resolution)	$\mathbf{1.27 \pm 0.26}$	$\mathbf{1.60 \pm 0.55}$

Table 3: Accuracy comparison of different input types

Network input	OpenKBP		LUMC	
	Dose score (Gy) ↓	DVH score (Gy) ↓	Dose score (Gy) ↓	DVH score (Gy) ↓
CT	5.91 ± 2.20	8.82 ± 3.80	3.10 ± 0.87	12.09 ± 3.84
PTVs + OARs	2.68 ± 1.06	1.77 ± 1.23	1.31 ± 0.26	1.69 ± 0.65
CT + PTVs	2.73 ± 1.15	1.86 ± 1.14	1.28 ± 0.25	1.63 ± 0.49
CT + PTVs + OARs	$\mathbf{2.52 \pm 1.00}$	$\mathbf{1.52 \pm 1.13}$	$\mathbf{1.27 \pm 0.26}$	$\mathbf{1.60 \pm 0.55}$

Table 4: Accuracy comparison of different loss functions

Loss	OpenKBP		LUMC	
	Dose score (Gy) ↓	DVH score (Gy) ↓	Dose score (Gy) ↓	DVH score (Gy) ↓
$\mathcal{L}_{MAE}$	$\mathbf{2.52 \pm 1.00}$	1.52 ± 1.13	$\mathbf{1.27 \pm 0.26}$	1.60 ± 0.55
$\mathcal{L}_{MAE} + \mathcal{L}_{vDVH}$	2.60 ± 1.08	1.48 ± 1.11	1.30 ± 0.26	1.59 ± 0.55
$\mathcal{L}_{MAE} + \mathcal{L}_{vDVH} + \mathcal{L}_{cDVH}$	2.59 ± 1.09	$\mathbf{1.41 \pm 1.10}$	1.29 ± 0.27	$\mathbf{1.48 \pm 0.50}$

3.2. Input type

For the experiment that evaluated the impact of different input types on dose prediction, we used the C3D model and the MAE loss function. Table 3 reports the accuracy on the OpenKBP and LUMC datasets. The results show that the incorporation of CT, PTVs, and OARs resulted in the lowest Dose score and DVH score for both datasets. In contrast, using CT alone resulted in the worst scores.

3.3. Loss function

For the experiment that assessed the impact of different loss functions on dose prediction, we used the C3D model with input consisting of CT, PTVs and OARs. As shown in table 4, for the OpenKBP dataset, the $\mathcal{L}_{MAE}$ loss achieved the lowest Dose score, while the $\mathcal{L}_{MAE}+\mathcal{L}_{vDVH}+\mathcal{L}_{cDVH}$ loss achieved the lowest DVH score along with a competitive Dose score. A similar trend was observed in the LUMC dataset. This pattern suggests that while the $\mathcal{L}_{MAE}$ loss alone provided the best Dose score, $\mathcal{L}_{MAE} + \mathcal{L}_{vDVH} + \mathcal{L}_{cDVH}$ offered the best DVH score and a more balanced overall improvement (that is, better Dose and DVH scores compared to $\mathcal{L}_{MAE} + \mathcal{L}_{vDVH}$). Given the high clinical relevance of the DVH score, we selected $\mathcal{L}_{MAE} + \mathcal{L}_{vDVH} + \mathcal{L}_{cDVH}$ for subsequent experiments.

Table 5: Accuracy comparison of different model structures

	Model Structure	OpenKBP		LUMC	
		Dose score (Gy) ↓	DVH score (Gy) ↓	Dose score (Gy) ↓	DVH score (Gy) ↓
Single	DoseNet (Kearney et al. 2018)	3.06 ± 1.22	1.67 ± 1.17	1.49 ± 0.27	1.56 ± 0.55
Single	HDUNet (Nguyen et al. 2019)	2.70 ± 1.05	1.46 ± 1.10	1.34 ± 0.26	1.57 ± 0.59
Single	SwinUNETR (Hatamizadeh et al. 2021)	2.86 ± 1.16	1.50 ± 1.25	1.40 ± 0.26	1.54 ± 0.54
Single	U-NAS (Lin et al. 2024)	2.72 ± 1.17	1.45 ± 1.22	1.32 ± 0.26	1.51 ± 0.53
Cascade	C3D (Liu et al. 2021)	$\mathbf{2.59 \pm 1.09}$	1.41 ± 1.10	$\mathbf{1.29 \pm 0.27}$	1.48 ± 0.50
Cascade	DOSE-PYFER (Gheshlaghi et al. 2024)	2.75 ± 1.18	1.49 ± 1.23	1.32 ± 0.24	1.54 ± 0.53
Cascade	U-NAS (Lin et al. 2024)	2.65 ± 1.05	$\mathbf{1.37 \pm 1.12}$	1.31 ± 0.27	$\mathbf{1.45 \pm 0.51}$

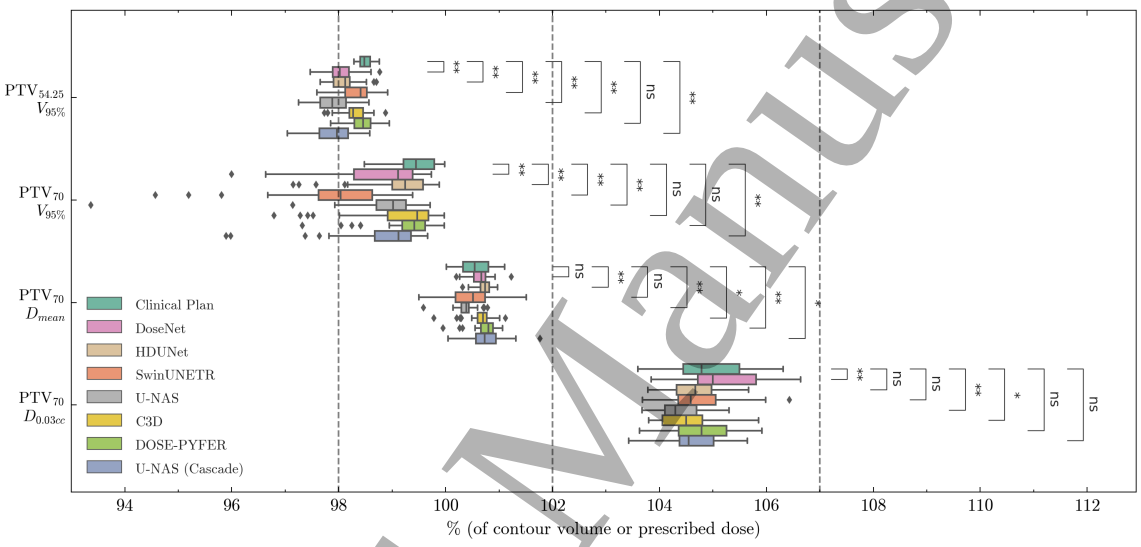

Figure 2: Boxplots with clinical dose evaluation metrics for targets on the LUMC dataset. For $\text{PTV}_{54.25}$, the clinical goal is $V_{95\%} \geq 98\%$. For PTV_{70}, the clinical goals are $V_{95\%} \geq 98\%$, $D_{\text{mean}} \leq 102\%$ (of the prescribed dose), and $D_{0.03\text{cc}} \leq 107\%$ (of the prescribed dose). Statistical significance was tested using a two-tailed Wilcoxon signed-rank test. **: $p \leq 0.01$, *: $0.01 < p \leq 0.05$, ns = not significant.

3.4. Model structure

In this section, we used the input consisting of CT, PTVs, and OARs, and the loss function was $\mathcal{L}_{MAE} + \mathcal{L}_{vDVH} + \mathcal{L}_{cDVH}$. Table 5 shows that in the single model setup, U-NAS performed well on both datasets. In the cascade model setup, C3D achieved the best dose scores in both datasets, while U-NAS achieved the best DVH scores. In the LUMC dataset, the accuracy differences between all models are not great, being less than or equal to 0.2 Gy. However, for some of the clinical dose evaluation metrics shown in figures 2 and 3, there remain significant differences between predictions and the clinical plan (mainly with respect to $V_{95\%}$ of $\text{PTV}_{54.25}$, D_{mean} of PTV_{70}, $D_{0.03\text{cc}}$ of Spinal Cord, and D_{mean} of the contralateral submandibular gland).

Under Poisson noise (table 6), all models demonstrated strong robustness, with most models having ΔDose and ΔDVH scores below 0.1 Gy. Figure 4 also illustrates this point: even in extreme conditions with $\lambda = 20$, although the quality of the CT

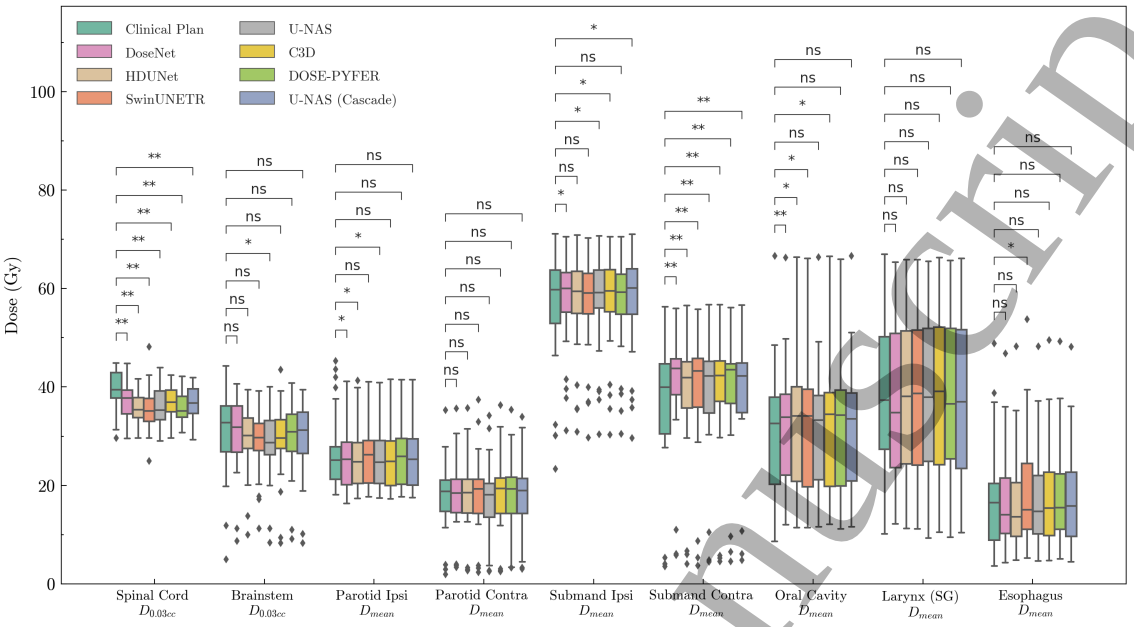

Figure 3: Boxplots with clinical dose evaluation metrics for OARs on the LUMC dataset. Ipsi = ipsilateral, Contra = contralateral, Submand = submandibular gland, SG = supraglottis. Statistical significance was tested using a two-tailed Wilcoxon signed-rank test. **: $p \leq 0.01$, *: $0.01 < p \leq 0.05$, ns = not significant.

Table 6: Robustness comparison of different model structures against Poisson noise ($\lambda = 20$)

	Model Structure	OpenKBP		LUMC	
		ΔDose score (Gy) ↓	ΔDVH score (Gy) ↓	ΔDose score (Gy) ↓	ΔDVH score (Gy) ↓
Single	DoseNet (Kearney et al. 2018)	0.01 ± 0.02	0.01 ± 0.10	0.03 ± 0.01	0.02 ± 0.05
Single	HDUNet (Nguyen et al. 2019)	0.28 ± 0.27	0.10 ± 0.32	0.08 ± 0.06	0.03 ± 0.21
Single	SwinUNETR (Hatamizadeh et al. 2021)	0.04 ± 0.10	0.09 ± 0.18	0.12 ± 0.05	0.16 ± 0.17
Single	U-NAS (Lin et al. 2024)	$\mathbf{0.00 \pm 0.03}$	$\mathbf{0.00 \pm 0.03}$	$\mathbf{0.00 \pm 0.01}$	0.01 ± 0.03
Cascade	C3D (Liu et al. 2021)	0.00 ± 0.04	0.00 ± 0.06	0.02 ± 0.02	0.00 ± 0.08
Cascade	DOSE-PYFER (Gheshlaghi et al. 2024)	0.00 ± 0.07	0.00 ± 0.07	0.01 ± 0.01	0.00 ± 0.06
Cascade	U-NAS (Lin et al. 2024)	0.00 ± 0.04	0.00 ± 0.04	0.01 ± 0.01	$\mathbf{0.00 \pm 0.02}$

Table 7: Robustness comparison of different model structures against PGD attack ($\epsilon = 16$ HU)

	Model Structure	OpenKBP		LUMC	
		ΔDose score (Gy) ↓	ΔDVH score (Gy) ↓	ΔDose score (Gy) ↓	ΔDVH score (Gy) ↓
Single	DoseNet (Kearney et al. 2018)	3.63 ± 1.85	6.41 ± 1.87	0.66 ± 0.23	0.64 ± 0.42
Single	HDUNet (Nguyen et al. 2019)	0.20 ± 0.30	0.23 ± 0.43	0.22 ± 0.14	0.18 ± 0.26
Single	SwinUNETR (Hatamizadeh et al. 2021)	$\mathbf{0.06 \pm 0.19}$	$\mathbf{0.17 \pm 0.37}$	0.18 ± 0.12	$\mathbf{0.01 \pm 0.17}$
Single	U-NAS (Lin et al. 2024)	0.35 ± 0.43	0.31 ± 0.54	0.67 ± 0.20	0.46 ± 0.29
Cascade	C3D (Liu et al. 2021)	0.14 ± 0.27	0.24 ± 0.40	$\mathbf{0.08 \pm 0.07}$	0.13 ± 0.20
Cascade	DOSE-PYFER (Gheshlaghi et al. 2024)	0.29 ± 0.41	0.55 ± 0.65	0.11 ± 0.09	0.12 ± 0.21
Cascade	U-NAS (Lin et al. 2024)	0.35 ± 0.39	0.31 ± 0.49	0.30 ± 0.14	0.31 ± 0.26

images had severely degraded, the dose prediction remained almost unchanged.

The robustness of different model structures was further evaluated under adversarial

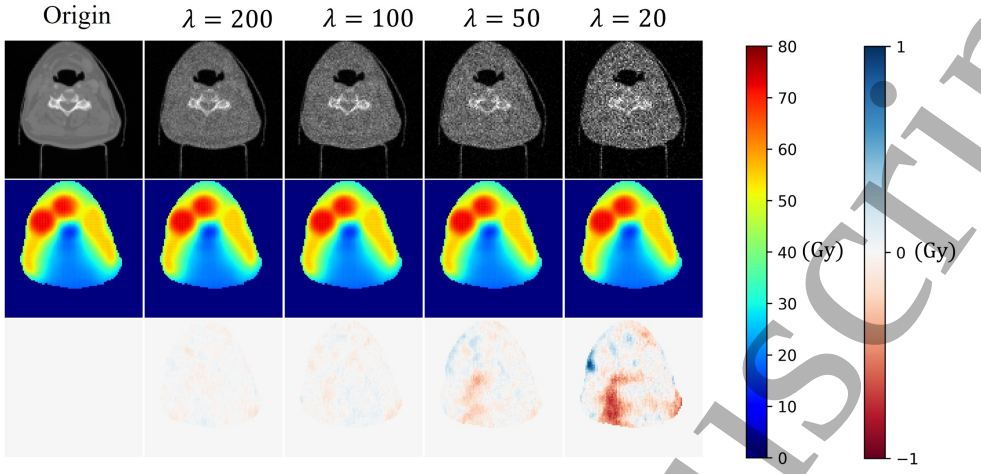

Figure 4: Impact of different Poisson noise levels (λ) on U-NAS (cascade) dose prediction, using a randomly selected slice from the LUMC dataset. The first row shows CT images with increasing noise levels. The second and third rows depict U-NAS (cascade) dose predictions and the differences map from the original dose prediction (noise-free), respectively. The color bars represent dose values (0 to 80 Gy) and differences map from the original prediction (-1 to 1 Gy).

Table 8: Robustness comparison of different model structures against Mi-FGSM attack ($\epsilon = 16$ HU)

	Model Structure	OpenKBP		LUMC	
		ΔDose score (Gy) ↓	ΔDVH score (Gy) ↓	ΔDose score (Gy) ↓	ΔDVH score (Gy) ↓
Single	DoseNet (Kearney et al. 2018)	4.42 ± 1.72	7.77 ± 1.87	0.78 ± 0.23	0.50 ± 0.26
Single	HDUNet (Nguyen et al. 2019)	0.63 ± 0.24	0.90 ± 0.35	0.48 ± 0.10	0.43 ± 0.20
Single	SwinUNETR (Hatamizadeh et al. 2021)	$\mathbf{0.24 \pm 0.13}$	$\mathbf{0.52 \pm 0.24}$	$\mathbf{0.20 \pm 0.07}$	$\mathbf{0.22 \pm 0.18}$
Single	U-NAS (Lin et al. 2024)	0.93 ± 0.56	0.93 ± 0.51	0.70 ± 0.27	0.41 ± 0.38
Cascade	C3D (Liu et al. 2021)	0.51 ± 0.52	0.63 ± 0.61	0.23 ± 0.07	0.28 ± 0.31
Cascade	DOSE-PYFER (Gheshlaghi et al. 2024)	0.63 ± 0.53	1.11 ± 0.69	0.31 ± 0.17	0.40 ± 0.37
Cascade	U-NAS (Lin et al. 2024)	0.99 ± 0.73	1.28 ± 0.66	0.32 ± 0.09	0.31 ± 0.20

noise by PGD and Mi-FGSM attacks (table 7, table 8). The tables reveal that SwinUNETR consistently demonstrated superior resilience across both the OpenKBP and LUMC datasets when subjected to these adversarial noises. Notably, U-NAS (cascade), which had performed well under Poisson noise, struggled considerably under adversarial noise, especially on the OpenKBP dataset. Figure 5 visually aligns with these findings. As the attack strength increases, both U-NAS (cascade) and SwinUNETR show increasing degradation from the original dose predictions. However, SwinUNETR maintains greater stability, with less pronounced deviations compared to U-NAS (cascade), particularly as the attack strength becomes stronger.

Table 9 compares the computational efficiency of various model structures. DoseNet was the most computationally efficient among the single models, having the lowest GPU memory usage and runtime. Among the cascade models, U-NAS was the most efficient compared to others. SwinUNETR, among all models, demanded the most resources

Table 9: Computational efficiency comparison of different model structures

	Model Structure	OpenKBP		LUMC	
		GPU memory (MB)	GPU runtime (ms)	GPU memory (MB)	GPU runtime (ms)
Single	DoseNet (Kearney et al. 2018)	297	11	708	31
Single	HDUNet (Nguyen et al. 2019)	2589	109	9437	179
Single	SwinUNETR (Hatamizadeh et al. 2021)	3593	126	11359	440
Single	U-NAS (Lin et al. 2024)	1412	58	4778	142
Cascade	C3D (Liu et al. 2021)	2254	77	8036	241
Cascade	DOSE-PYFER (Gheshlaghi et al. 2024)	2309	186	6798	512
Cascade	U-NAS (Lin et al. 2024)	1550	115	5150	356

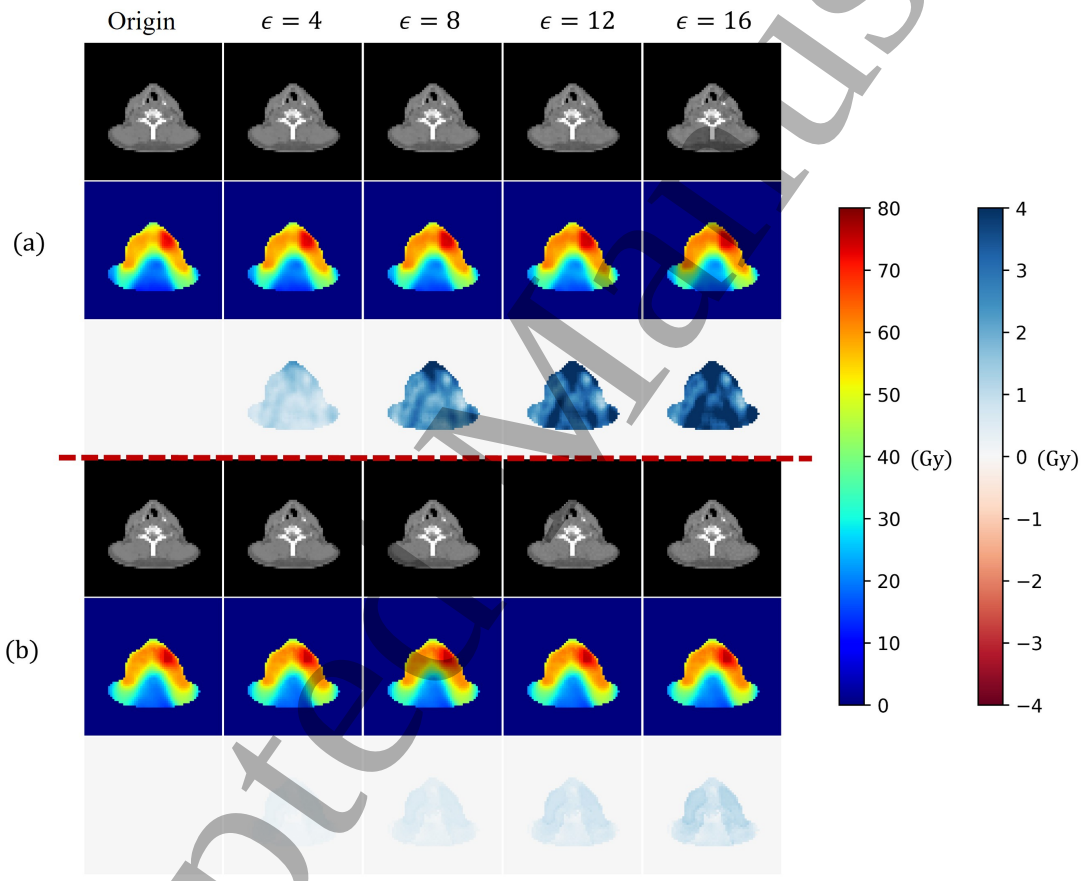

Figure 5: Impact of different attack strengths (ϵ) of the MI-FGSM attack on U-NAS (cascade) and SwinUNETR, using a randomly selected slice from the OpenKBP dataset. Panel (a) showcases the results on U-NAS (cascade) : the first row displays the original and perturbed CT images, the second and third rows depict the corresponding dose predictions and the differences map from the original dose prediction (noise-free), respectively. Panel (b) presents similar sequence results for SwinUNETR. The color bars represent dose values (0 to 80 Gy) and differences map from the original prediction (-4 to 4 Gy).

and had the longest runtimes. The results also indicated that the LUMC dataset, which contains higher resolution images, led to higher GPU memory usage and longer runtimes

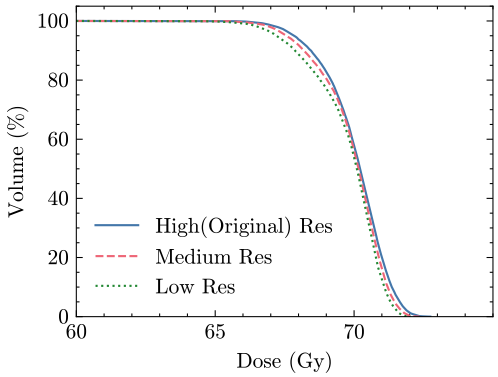

Figure 6: Comparison of the DVH curves of PTV_{70} under high (original), medium, and low resolutions.

compared to the OpenKBP dataset. Note that all models achieved GPU runtimes of less than 1 second in both datasets.

4. Discussion

In this paper, we explore the factors that affect the performance of deep learning-based dose prediction using the publicly available OpenKBP dataset and in-house LUMC dataset. From a dataset perspective, the OpenKBP dataset has certain limitations. The data originates from The Cancer Imaging Archive (TCIA), has been re-annotated and processed; the resolution of the CT image and dose grid is lower than clinical standard. In addition, it does not include a clinically approved RT plan. However, the LUMC dataset compensates for these shortcomings well. Furthermore, the IMRT-based OpenKBP and VMAT-based LUMC datasets well represent current treatment planning techniques in the field of radiation therapy. Our findings indicate that most conclusions are applicable to both datasets, suggesting that the results discussed in this paper have considerable generalizability.

To ensure a fair comparison, all models were either re-implemented based on their original publications or adapted from publicly available GitHub repositories (e.g., C3D, DOSE-PYFER, U-NAS). For models without released code, we re-implemented the architectures using the MONAI‡ framework and closely followed the original publication. Although there may be implementation differences, we standardized the training settings across all models as described in Section 2.2 (e.g., same optimizer, learning rate, and maximum training epoch). These settings were chosen because we found them to be robust across different model architectures and generally allowed the models to reach performance levels comparable to those reported in the original papers. We acknowledge that further hyperparameter tuning could improve individual model performance. However, our goal was to benchmark under a unified and reproducible

‡ https://www.monai.io

training setup. All code and configurations used in this study are publicly available at https://github.com/RuochenGao/HaN-DosePrediction.

As shown in figure 6, the differences observed in the dose fall-off region of the DVH plot are due to interpolation effects. The primary goal of RT planning is to deliver a high radiation dose to the tumor while minimizing exposure to surrounding healthy tissues, creating a sharp dose gradient visible as a steep decline in the DVH curve. However, interpolation algorithms can smooth out this gradient between high-dose and low-dose regions, introducing deviations from the actual dose distribution. Therefore, we recommend using high-resolution input and dose grid to preserve the integrity of the gradient.

From table 3, we can see that the model performs poorly when using only CT images. This is because CT scans do not provide the neural network with explicit information about the tumor's location, shape, and size, and the neural network must therefore learn to extract this information from the CT data directly, without the use of supervision. This significantly increases the difficulty of the overall learning task for dose prediction, as tumors in the head and neck region exhibit irregular shapes and sizes. In contrast, when using only PTVs + OARs, we observe that the model performed well. This indicates that the spatial and anatomical relationships between targets and OARs contain the most critical information for the dose prediction task. While CT (HU values) play a key role in traditional dose calculation engines by enabling photon attenuation modeling, deep learning models appear to rely more heavily on explicit geometric representations provided by PTV and OAR contours. However, the combination of CT + PTVs + OARs achieved the highest accuracy, and therefore we still recommend using this combination.

The selection of a loss function is closely related to the desired output. In RT planning, clinicians are more concerned with ensuring adequate dose coverage for PTVs and minimizing the mean and maximum dose to OARs, which is related to the DVH score in our study. Previous work has demonstrated the benefits of incorporating a value-based DVH loss (Nguyen et al. 2020). As demonstrated in table 4, we found that combining MAE, value-based DVH loss, and criteria-based DVH loss functions improves the DVH score even further. Moreover, the criteria-based DVH loss can be tailored to meet specific clinical requirements in different scenarios. Thus, we recommend using the combined loss function $\mathcal{L}_{MAE} + \mathcal{L}_{vDVH} + \mathcal{L}_{cDVH}$ to achieve optimal results.

From table 5, we observe that models with a cascade architecture often outperformed single models, achieving lower Dose and DVH scores, indicating improved accuracy in dose prediction. This suggests that the progressive refinement inherent in cascade architectures enables more precise adjustments during the prediction process, resulting in better accuracy in capturing complex patterns within CT scans, combined with PTVs and OARs. However, as shown in table 9, the GPU runtimes for cascade models were generally longer than those for single models, indicating a trade-off in computational efficiency. Since the inference time remained below 1 second, this extended runtime may not pose a significant issue. We believe that for the complex

task of dose prediction, a cascade architecture may be necessary to achieve optimal performance. Furthermore, it should be noted that U-NAS (cascade) achieved the best results on the DVH score. However, as shown in figure 2, U-NAS (cascade) shows significant differences ($p \leq 0.01$) compared to the clinical plan in terms of $V_{95\%}$ for $PTV_{54.25}$ and PTV_{70}. In contrast, Dose-PYFER performs better on these metrics, showing no significant differences from the clinical plan ($p > 0.05$). This discrepancy arises because the DVH score is an aggregated metric composed of several DVH metrics. As such, it does not necessarily reflect good performance in all individual DVH metrics. Furthermore, U-NAS focuses on finding the optimal neural network structure based on the current loss function, which does not include metrics such as $V_{95\%}$.

In the experiments on model robustness, we found that for Poisson noise, which commonly occurs in CT scans, all models exhibited strong robustness (table 6). As shown in figure 4, where the quality of CT image had been significantly degraded at $\lambda = 20$, the dose prediction showed almost no change. However, when facing adversarial noise, where imperceptible noise is added to the original CT image (shown in figure 5), the accuracy of the model can be affected (table 7, table 8). This is mainly because adversarial noise is generated based on the model's gradient, maximizing the model's prediction error. However, we observed that SwinUNETR demonstrated stronger robustness against adversarial noise. This may be due to SwinUNETR divides the input data into non-overlapping patches and processes these patches individually. This patch-based approach can act as a form of regularization, reducing the impact of adversarial noise by ensuring that localized perturbations in each patch do not significantly affect the overall representation. However, SwinUNETR does not perform as well as other models, such as C3D and U-NAS, in terms of the accuracy metric. Furthermore, it has the highest GPU memory consumption among the evaluated models. Therefore, for practical clinical deployment, it is essential to choose an appropriate model structure that balances accuracy, robustness, and computational efficiency.

One aspect not explored in this study is the incorporation of prior knowledge information such as beam configuration. As the beam setup is fixed in our dataset, the influence of beam configuration was not investigated in this study. However, recent studies, such as Gao et al. (2023), have shown that including beam-related information (e.g., beam angles and beam plates) can improve dose prediction accuracy, particularly in more diverse clinical scenarios. Although this was beyond the scope of the current work, incorporating this prior knowledge is a promising direction for future research to improve model generalizability.

Another limitation is that we add noise to CT images to verify the robustness of the model structure. However, in clinical practice, the uncertainty introduced by inter-physician variability in the contouring of tumors and OARs (Guzene et al. 2023) represents another critical element affecting model robustness. Future work will focus on investigating the impact of this uncertainty on dose prediction. We suggest that in dose prediction tasks, it is essential not only to focus on the accuracy of the model but also to consider the robustness of the model, including its ability to counter noise and handle

uncertainty. These aspects have significant practical value in clinical applications.

Finally, our study focuses on photon-based radiotherapy (IMRT and VMAT), while proton therapy is also currently used for head and neck cancer treatment. Proton therapy introduces additional challenges in dose prediction due to its higher sensitivity to variations in CT HU values and its sharper dose gradients. While our findings suggest that photon-based deep learning models benefit in a relatively small amount from CT as an input (table 3), this input may be more relevant for proton therapy dose prediction. In the future, a dedicated evaluation using proton therapy datasets will be needed to determine if our findings can be generalized to proton radiotherapy.

5. Conclusion

This study presents a comprehensive analysis of key factors that influence deep learning-based dose prediction models for head and neck cancer radiotherapy. By systematically examining input and dose grid resolution, input type, loss function, and model architecture using both public and in-house clinical datasets, we demonstrate their significant effects on model accuracy, robustness, and computational efficiency. Our findings show that high-resolution inputs, specifically CT images with PTVs and OARs, combined with a hybrid loss function that incorporates MAE and value-based and criteria-based DVH components, substantially improve prediction accuracy. Robustness testing reveals that while most models exhibit greater resistance to Poisson noise than adversarial noise, certain models, such as SwinUNETR, demonstrate superior robustness against adversarial perturbations. These insights provide valuable guidance for optimizing deep learning-based dose prediction models, contributing to more precise and reliable radiotherapy planning.

Acknowledgments

This work was supported by the China Scholarship Council (No. 202207720085) and utilized the Dutch national e-infrastructure with the support of the SURF Cooperative using grant No. EINF-6458.

Appendix A. Boxplots of absolute differences in clinical dose evaluation metrics

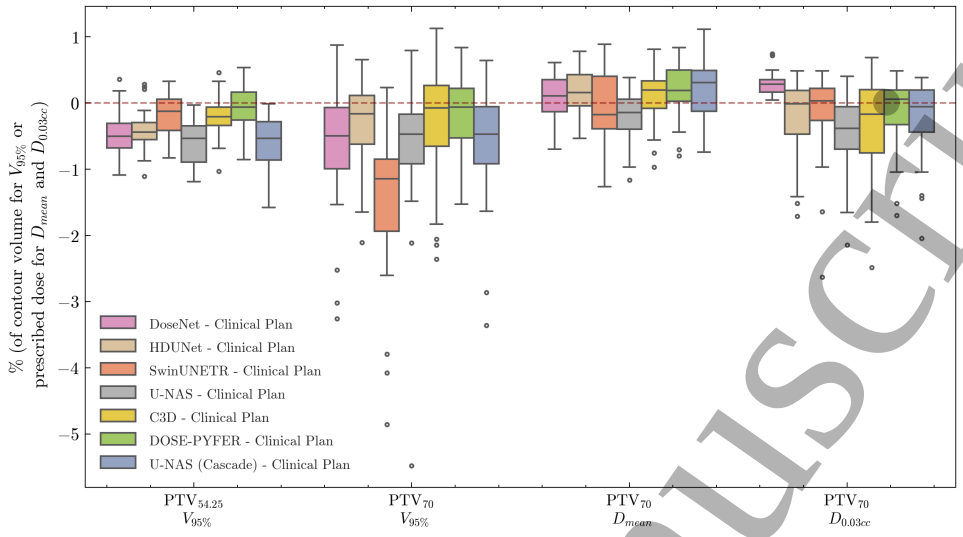

Figure A1: Boxplots of absolute differences in clinical dose evaluation metrics for targets on the LUMC dataset, comparing different models with the clinical plan.

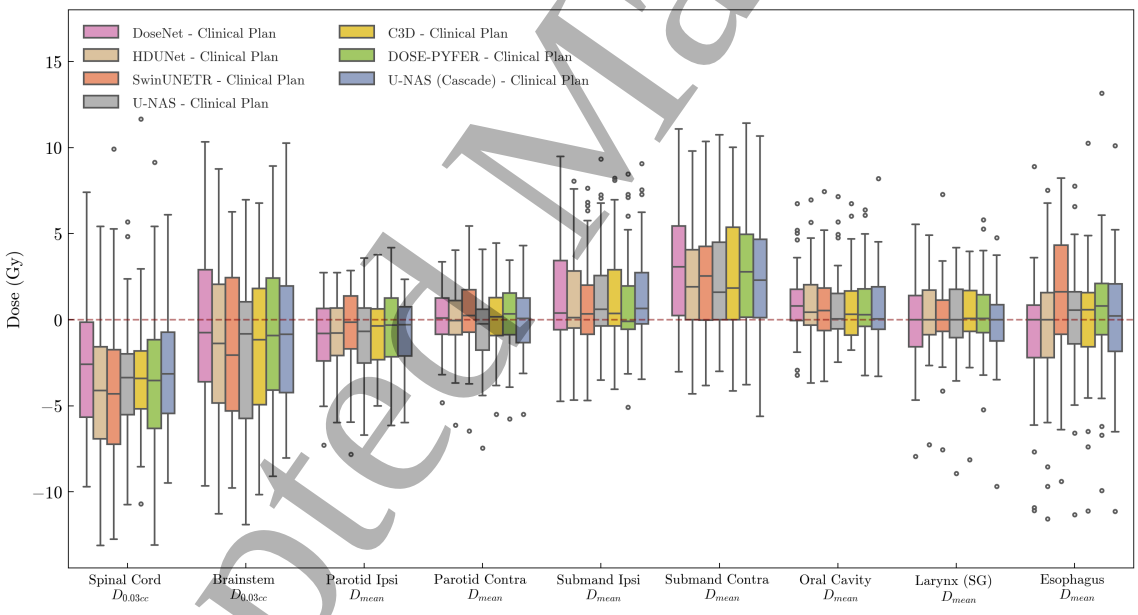

Figure A2: Boxplots of absolute differences in clinical dose evaluation metrics for OARs on the LUMC dataset, comparing different models with the clinical plan. Ipsi = ipsilateral, Contra = contralateral, Submand = submandibular gland, SG = supraglottis.

References

Babier, A., Boutilier, J. J., McNiven, A. L. & Chan, T. C. (2018). Knowledge-based automated planning for oropharyngeal cancer, *Medical physics* **45**(7): 2875–2883.

Babier, A., Zhang, B., Mahmood, R., Moore, K. L., Purdie, T. G., McNiven, A. L. & Chan, T. C. (2021). Openkbp: the open-access knowledge-based planning grand challenge and dataset,

Medical Physics **48**(9): 5549–5561.

Bai, P., Weng, X., Quan, K., Chen, J., Dai, Y., Xu, Y., Lin, F., Zhong, J., Wu, T. & Chen, C. (2020). A knowledge-based intensity-modulated radiation therapy treatment planning technique for locally advanced nasopharyngeal carcinoma radiotherapy, *Radiation Oncology* **15**: 1–10.

Bakx, N., Bluemink, H., Hagelaar, E., van der Sangen, M., Theuws, J. & Hurkmans, C. (2021). Development and evaluation of radiotherapy deep learning dose prediction models for breast cancer, *Physics and imaging in radiation oncology* **17**: 65–70.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q. & Wang, M. (2022). Swin-UNet: UNet-like pure transformer for medical image segmentation, *European conference on computer vision*, Springer, pp. 205–218.

Chandran, L. P., KA, A. N., Puzhakkal, N. & Makuny, D. (2023). MemU-Net: A new volumetric dose prediction model using deep learning techniques in radiation treatment planning, *Biomedical Signal Processing and Control* **85**: 104940.

Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X. & Li, J. (2018). Boosting adversarial attacks with momentum, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9185–9193.

Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L. & Kohane, I. S. (2019). Adversarial attacks on medical machine learning, *Science* **363**(6433): 1287–1289.

Gao, R., Lou, B., Xu, Z., Comaniciu, D. & Kamen, A. (2023). Flexible-cm gan: Towards precise 3d dose prediction in radiotherapy, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 715–725.

Gheshlaghi, T., Nabavi, S., Shirzadikia, S., Moghaddam, M. E. & Rostampour, N. (2024). A cascade transformer-based model for 3d dose distribution prediction in head and neck cancer radiotherapy, *Physics in Medicine & Biology* **69**(4): 045010.

Goodfellow, I. J., Shlens, J. & Szegedy, C. (2015). Explaining and harnessing adversarial examples, *Proceedings of the 3rd International Conference on Learning Representations*.

Gronberg, M. P., Gay, S. S., Netherton, T. J., Rhee, D. J., Court, L. E. & Cardenas, C. E. (2021). Dose prediction for head and neck radiotherapy using a three-dimensional dense dilated U-Net architecture, *Medical physics* **48**(9): 5567–5573.

Gu, X., Strijbis, V. I., Slotman, B. J., Dahele, M. R. & Verbakel, W. F. (2023). Dose distribution prediction for head-and-neck cancer radiotherapy using a generative adversarial network: influence of input data, *Frontiers in Oncology* **13**: 1251132.

Guzene, L., Beddok, A., Nioche, C., Modzelewski, R., Loiseau, C., Salleron, J. & Thariat, J. (2023). Assessing interobserver variability in the delineation of structures in radiation oncology: A systematic review, *International Journal of Radiation Oncology* Biology* Physics* **115**(5): 1047–1060.

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H. R. & Xu, D. (2021). Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images, *International MICCAI Brainlesion Workshop*, Springer, pp. 272–284.

Hu, C., Wang, H., Zhang, W., Xie, Y., Jiao, L. & Cui, S. (2023). TrDosePred: A deep learning dose prediction algorithm based on transformers for head and neck cancer radiotherapy, *Journal of Applied Clinical Medical Physics* **24**(7): e13942.

Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. & Maier-Hein, K. H. (2021). nnUNet: a self-configuring method for deep learning-based biomedical image segmentation, *Nature methods* **18**(2): 203–211.

Jiao, S.-X., Chen, L.-X., Zhu, J.-H., Wang, M.-L. & Liu, X.-W. (2019). Prediction of dose-volume histograms in nasopharyngeal cancer IMRT using geometric and dosimetric information, *Physics in Medicine & Biology* **64**(23): 23NT04.

Kearney, V., Chan, J. W., Haaf, S., Descovich, M. & Solberg, T. D. (2018). DoseNet: a volumetric dose prediction algorithm using 3d fully-convolutional neural networks, *Physics in Medicine*

& Biology **63**(23): 235022.

Kurakin, A., Goodfellow, I., Bengio, S., Dong, Y., Liao, F., Liang, M., Pang, T., Zhu, J., Hu, X., Xie, C. et al. (2018). Adversarial attacks and defences competition, *The NIPS'17 Competition: Building Intelligent Systems*, Springer, pp. 195–231.

Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.-W. & Heng, P.-A. (2018). H-DenseUNet: hybrid densely connected unet for liver and tumor segmentation from CT volumes, *IEEE transactions on medical imaging* **37**(12): 2663–2674.

Lin, Y., Liu, Y., Chen, H., Yang, X., Ma, K., Zheng, Y. & Cheng, K.-T. (2024). LENAS: Learning-based neural architecture search and ensemble for 3-d radiotherapy dose prediction, *IEEE Transactions on Cybernetics* .

Liu, S., Zhang, J., Li, T., Yan, H. & Liu, J. (2021). A cascade 3d U-Net for dose prediction in radiotherapy, *Medical physics* **48**(9): 5574–5582.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D. & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks.

Milletari, F., Navab, N. & Ahmadi, S.-A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation, *2016 fourth international conference on 3D vision (3DV)*, Ieee, pp. 565–571.

Momin, S., Fu, Y., Lei, Y., Roper, J., Bradley, J. D., Curran, W. J., Liu, T. & Yang, X. (2021). Knowledge-based radiation treatment planning: a data-driven method survey, *Journal of applied clinical medical physics* **22**(8): 16–44.

Morgan, H. E. & Sher, D. J. (2020). Adaptive radiotherapy for head and neck cancer, *Cancers of the head & neck* **5**: 1–16.

Nelms, B. E., Robinson, G., Markham, J., Velasco, K., Boyd, S., Narayan, S., Wheeler, J. & Sobczak, M. L. (2012). Variation in external beam treatment plan quality: an inter-institutional study of planners and planning systems, *Practical radiation oncology* **2**(4): 296–305.

Nguyen, D., Jia, X., Sher, D., Lin, M.-H., Iqbal, Z., Liu, H. & Jiang, S. (2019). 3d radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected U-Net deep learning architecture, *Physics in medicine & Biology* **64**(6): 065020.

Nguyen, D., McBeth, R., Sadeghnejad Barkousaraie, A., Bohara, G., Shen, C., Jia, X. & Jiang, S. (2020). Incorporating human and learned domain knowledge into training deep neural networks: a differentiable dose-volume histogram and adversarial inspired framework for generating Pareto optimal dose distributions in radiation therapy, *Medical physics* **47**(3): 837–849.

Thanh, D., Surya, P. et al. (2019). A review on CT and X-ray images denoising methods, *Informatica* **43**(2).

Wang, B., Teng, L., Mei, L., Cui, Z., Xu, X., Feng, Q. & Shen, D. (2022). Deep learning-based head and neck radiotherapy planning dose prediction via beam-wise dose decomposition, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, pp. 575–584.