# RANDose: A Region-aware Attention Network for Accurate Radiation Dose Prediction

G. Jignesh Chowdary[1], Tiezhi Zhang[2], Xin Qian[2], and Zhaozheng Yin[1]

[1] Department of Computer Science, Stony Brook University, NY, USA.
[2] Department of Radiation Oncology, Stony Brook University, NY, USA.

**Abstract.** External Radiation Therapy (ERT) is a key treatment in oncology, aiming to deliver high radiation doses to the Planned Target Volume (PTV) while minimizing exposure to surrounding healthy tissues and Organs At Risk (OARs). However, the proximity of PTVs to OARs, the presence of multiple OARs, and the time-consuming nature of manual subjective dose planning present significant challenges. While recent advancements in Deep Learning (DL) have led to various DL-based methods for dose prediction, it is still challenging to effectively capture multi-scale features and propagate essential information to related regions. In this work, we propose the Region-aware Attention Net (RANDose), which addresses these issues by integrating Multi-Scale Channel Spatial Attention (MSCSA), PTV Integration (PI), and Attention Fusion (AF) modules. Additionally, we introduce a Region-Aware Loss function to ensure accurate dose distribution within the PTV while minimizing radiation exposure to OARs. Experiments on the OpenKBP dataset demonstrate that RANDose outperforms existing models in both Dose Score and Dose Volume Histogram (DVH) Score, highlighting its superior performance. Code is available at GitHub.

**Keywords:** Radiation Dose Prediction · Deep learning · Multi-Scale Feature fusion · Attention

## 1 Introduction

External Radiation Therapy (ERT) is a critical treatment modality in oncology, designed to deliver a high radiation dose to the Planned Target Volume (PTV) while minimizing the exposure to surrounding healthy tissues and Organs At Risk (OARs) [2]. This presents a significant challenge, as PTVs are often in close proximity to OARs, and in some cases, there may be multiple OARs that require a careful planning to avoid the dose overlap. Additionally, manual dose planning is time-consuming and prone to human errors [18]. With advancements in Deep Learning (DL) [1,15,11] and its success in computer vision tasks [5,4,13,10,7,6], researchers have explored DL-based approaches for radiation dose prediction [17,14,19,8,9,16].

However, effectively capturing multi-scale features and propagating essential information throughout the network remains a challenge, both of which

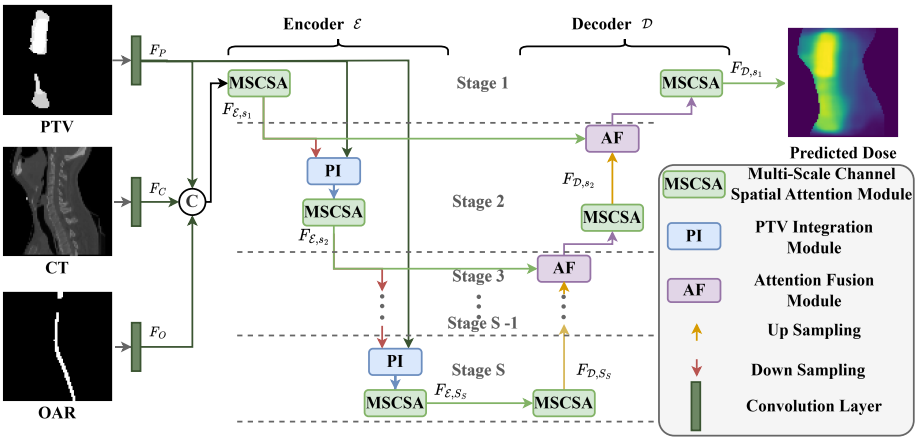


Fig. 1: Architecture of the proposed RANDose.

are crucial for predicting varying dose intensities across different regions. To address these limitations, we propose the Region-aware Attention Net (RANDose), which integrates Multi-Scale Channel Spatial Attention (MSCSA), PTV Integration (PI), and Attention Fusion (AF) modules. Additionally, to manage dose planning in relation to complex anatomical structures, we introduce the Region-Aware Loss, which ensures accurate dose distribution within the PTV while minimizing the radiation exposure to OARs. Experiments on the OpenKBP dataset show that RANDose outperforms existing models. RANDose captures multi-scale dose features, resulting in more accurate dose predictions across regions with varying dose intensities, while also protecting the OARs. Our key contributions in this work are as follows:

- We introduce Multi-Scale Channel Spatial Attention (MSCSA), which enhances feature extraction by capturing spatial and channel-wise dependencies at multiple scales.
- We propose PTV Integration (PI) to improve dose prediction accuracy by incorporating explicit information about the PTV structure.
- We develop Attention Fusion (AF) to facilitate effective information propagation across the network, refining dose estimation.
- We introduce Region-Aware Loss, which ensures accurate dose delivery to the PTV while minimizing the unnecessary exposure to OARs.

## 2 Methodology

The architecture of the proposed RANDose is shown in Figure 1. RANDose follows a U-shaped structure with an encoder ($\mathcal{E}$) and a decoder ($\mathcal{D}$), both consisting of $S$ stages, and incorporates three key modules: MSCSA, PI, and AF.

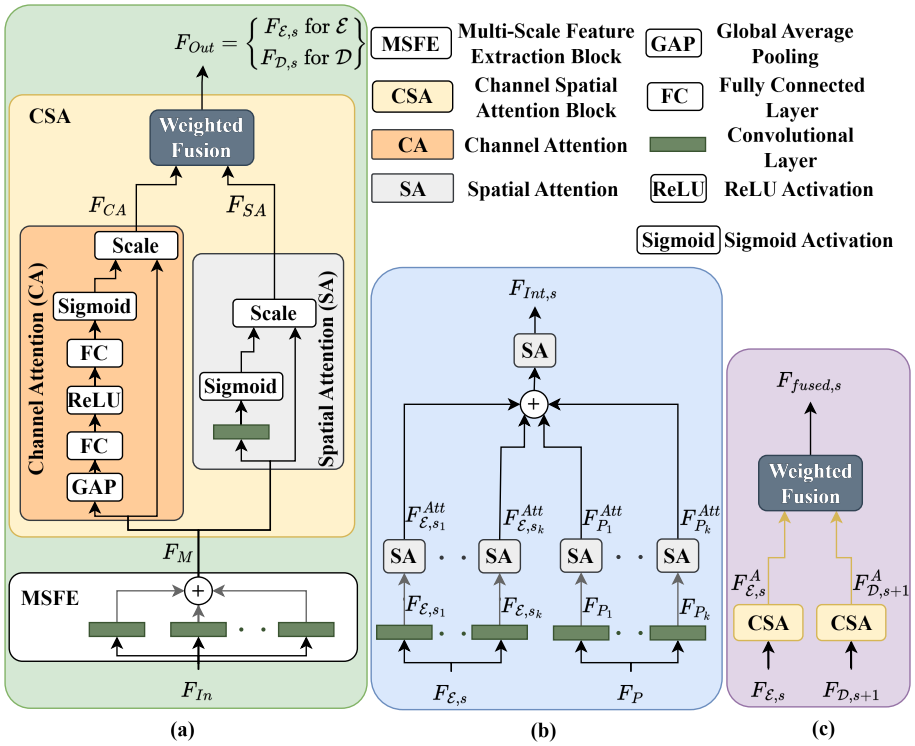


Fig. 2: Architecture of the proposed (a) MSCSA module, (b) PI module, and (c) AF module. 's' represents the stage $s$ of the network. The centered dots represent multiple layers/ blocks.

The MSCSA module enhances dose prediction by extracting multi-scale contextual information, while the PI module integrates PTV information at each stage of $\mathcal{E}$, enabling the network to focus on the PTV region by incorporating dose-relevant information alongside anatomical features. The AF module facilitates the propagation of essential information from $\mathcal{E}$ to $\mathcal{D}$, ensuring better dose estimation. The network takes three inputs: PTV, CT, and OAR, which are processed through a convolutional layer to generate feature representations $F_P$, $F_C$, and $F_O$. These features are then concatenated channel-wise and passed through the network, with $F_P$ being utilized at each stage from the second stage onward. The details of these modules are illustrated in Figure 2 and described below.

### 2.1 MSCSA module

The MSCSA architecture consists of two key components: the Multi-Scale Feature Extraction (MSFE) and the Channel Spatial Attention (CSA) blocks. The MSFE block takes an input feature map ($F_{In}$) and processes it using parallel 3D

convolutional layers with varying kernel sizes $n \times n \times n$. Smaller kernels focus on fine-grained local details, while larger kernels capture broader contextual information. The resulting feature maps from the parallel convolutions are combined via element-wise addition, forming a comprehensive multi-scale representation ($F_M$) that retains both local and global spatial dependencies.

To enhance feature relevance, the extracted multiscale features ($F_M$) are processed through the CSA block, which employs a dual attention mechanism that integrates channel-wise recalibration and spatial attention to dynamically refine feature representations. The Channel Attention (CA) mechanism begins by applying a global average pooling layer to compute the spatial average for each channel. The pooled output is then passed through a fully connected (FC) layer, followed by a ReLU activation layer to introduce non-linearity. Then the features are processed through a second FC layer, and a Sigmoid activation layer produces channel-wise attention weights that re-calibrate feature maps via scaling, enhancing the most informative channels while suppressing less relevant ones. This results in $F_{CA}$, the channel-refined features. Simultaneously, the Spatial Attention (SA) mechanism applies a $1 \times 1 \times 1$ convolution to compute a spatial attention map that highlights region-specific importance. The spatial weights are then normalized with a Sigmoid activation and broadcasted across channels, scaling the input features to enhance critical spatial regions. This produces $F_{SA}$, the spatially refined features. Finally, the outputs from both attention pathways ($F_{CA}$ and $F_{SA}$) are adaptively fused using a learnable weighted summation. This fusion produces the final enhanced feature representation: $F_{Out} = F_{CA} + \alpha \cdot F_{SA}$ where $\alpha$ is the learned parameter.

### 2.2 PI module

The PTV Integration (PI) module is introduced at each stage $s$ of $\mathcal{E}$, starting from the second stage. It effectively integrates the encoder features $F_{\mathcal{E},s}$, with the PTV features $F_P$, enhancing the model's ability to predict dose distributions by combining spatially relevant PTV information with the structural features captured by $\mathcal{E}$. The PI module begins by resizing the $F_p$ to match the spatial dimension of the $F_{\mathcal{E},s}$, then extracts multi-scale features from both $F_{\mathcal{E},s}$ and $F_P$ using parallel convolutional layers with kernel sizes $k \times k \times k$, producing feature maps $F_{\mathcal{E},s_k}$ and $F_{P_k}$. These multi-scale features are then refined using spatial attention mechanisms applied separately at each scale, generating attention-refined feature maps $F^{Att}_{\mathcal{E},s_k}$ and $F^{Att}_{P_k}$. This attention mechanism directs focus to the most critical regions for dose prediction. Finally, the refined feature maps $F^{Att}_{\mathcal{E},s_k}$ and $F^{Att}_{P_k}$ are combined to form an integrated feature map. To ensure the effective propagation of essential information, this integrated feature map undergoes a final spatial attention layer, further refining its representation by emphasizing the most significant regions. This process ultimately produces the final attention-refined integrated feature map $F_{Int,s}$, which serves as the input to the MSCSA module in $\mathcal{E}$ from the second stage onward.

### 2.3 AF module

The AF module is designed to fuse information from $\mathcal{E}$ with $\mathcal{D}$ through skip connections at each stage of the network. These skip connections enable the direct transfer of information, ensuring the preservation of crucial spatial details that might otherwise be lost during downsampling. At each stage, the feature map $F_{\mathcal{E},s}$ and the upsampled feature map $F_{\mathcal{D},s+1}$ are processed through the CSA blocks (described in Section 2.1), resulting in the refined feature maps $F^A_{\mathcal{E},s}$ and $F^A_{\mathcal{D},s+1}$, respectively. These refined feature maps are then fused to produce $F_{fused,s} = F^A_{\mathcal{E},s} + \beta \cdot F^A_{\mathcal{D},s+1}$, where $\beta$ is the trainable parameters.

### 2.4 Loss

To optimize the proposed model, we design the Region-Aware Loss, which combines standard L1 loss with region-specific penalties for the PTV and OARs. The loss function consists of three main components: the standard L1 loss ($\mathcal{L}_{L1}$), the PTV-specific loss ($\mathcal{L}_{PTV}$), and the OAR-specific loss ($\mathcal{L}_{OAR}$). The $\mathcal{L}_{L1}$ is the voxel-wise L1 norm over all voxels, expressed as:

$$\mathcal{L}_{L1} = \|\hat{y} - y\|_1$$

where $\hat{y}$ represents the predicted dose, $y$ denotes the ground truth dose, and $\|\|_1$ denotes the L1 norm. The PTV-specific loss is designed to enforce accurate dose estimation within the PTV. This term calculates the L1 loss over the voxels within the PTV region, applying a learnable weight $w_{PTV}$ to scale the loss:

$$\mathcal{L}_{PTV} = w_{PTV} \cdot \|\hat{y}_{PTV} - y_{PTV}\|_1$$

where $\hat{y}_{PTV}$ and $y_{PTV}$ are the predicted and ground truth doses, respectively, within the PTV region. The OAR-specific loss penalizes excessive dose to OARs. Similar to the PTV loss, the L1 loss is computed over the OAR region, and a learnable weight $w_{OAR}$ is applied to scale the loss:

$$\mathcal{L}_{OAR} = w_{OAR} \cdot \|\hat{y}_{OAR} - y_{OAR}\|_1$$

where $\hat{y}_{OAR}$ and $y_{OAR}$ represent the predicted and ground truth doses, respectively, within the OAR region. The total loss function is the sum of the three components:

$$\mathcal{L}_{Total} = \mathcal{L}_{L1} + \mathcal{L}_{PTV} + \mathcal{L}_{OAR}$$

By combining these terms, the Region-Aware Loss ensures that the model optimizes general dose prediction while prioritizing dose precision within the target volume and minimizing dose to OARs.

## 3 Results and Discussion

### 3.1 Dataset and Performance Metrics

We evaluate the proposed model on the OpenKBP dataset [3], a publicly available collection of CT scans from patients receiving radiation therapy for head

and neck cancer. The dataset is split into 200 training cases, 40 validation cases, and 100 test cases. We use Dose Score (DS) and Dose Volume Histogram Score (DVHS) as the evaluation metrics, following previous work [17,14,19,8,9,16].

### 3.2 Implementations Details

In the proposed model, the number of stages $S$ is set to 5, the kernel sizes $n$ in the MSFE block are set to $\{3, 5, 7\}$, and the kernel sizes $k$ in the PI module are also set to $\{3, 5, 7\}$. The model is trained and evaluated using the official dataset split, running for 80,000 iterations with a batch size of 2. Training is optimized using the Adam optimizer with a learning rate of $1 \times 10^{-4}$. All experiments are conducted on an NVIDIA H100 GPU with 80 GB of RAM.

### 3.3 Quantitative Results

The proposed model achieved a DS of 2.190, and a DVHS of 1.160. For fair comparison, we evaluate against prior state-of-the-art (SOTA) models [17,14,19,8,9,16] using DS and DVHS. As shown in Table 1, RANDose outperforms all SOTA methods, achieving the lowest scores across both metrics.

Table 1: Quantitative Evaluation of the RANDose Against SOTA Dose Prediction Models: The best performance is highlighted in bold.

| Method | DS (↓) | DVHS (↓) |
|---|---|---|
| Xu et al. [17] | 2.753 | 1.559 |
| Szalkowski et al. [14] | 2.752 | 1.555 |
| Zimmermann et al. [19] | 2.620 | 1.520 |
| Li et al. [8] | 2.367 | 1.378 |
| Lin et al. [9] | 2.357 | 1.465 |
| Wang et al. [16] | 2.276 | 1.257 |
| **RANDose** | **2.190** | **1.160** |

### 3.4 Qualitative Results

In addition to the quantitative evaluation, we present qualitative results to demonstrate the performance of the proposed RANDose model in Figure 3. Figure 3 (a) shows the dose distributions for a single patient in the axial, coronal, and sagittal views, where the PTV volumes are close to the OARs. Despite this challenge, the model predicts dose distributions that closely match the ground truth while preserving OARs, as seen in the last two rows. Figure 3 (b) displays dose distributions in the sagittal view for more patients, highlighting the model's generalizability, even in scenarios with multiple OARs. Figure 3 (c) presents a few failure cases; for larger PTVs, the predicted dose slightly extends beyond

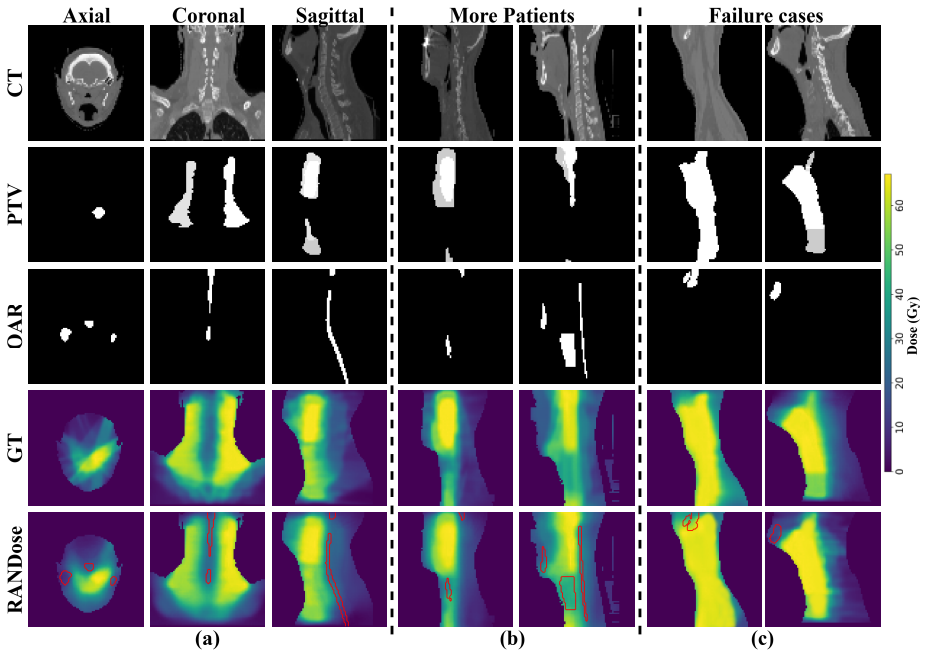


Fig. 3: Qualitative results of the proposed model in the radiation dose unit of Gray (Gy): (a) Dose predictions for a single patient in axial, coronal, and sagittal views, (b) dose predictions for more patients in sagittal view, and (c) failure cases in sagittal view.

the PTV region. This occurs because larger PTVs require more complex dose distributions, making it challenging for the model to accurately predict the dose boundaries, which may result in the dose extending beyond the intended region. However, even in these failure cases, the OARs remain unaffected, demonstrating the model's ability to maintain OAR protection despite minor inaccuracies in PTV dose boundaries.

### 3.5 Ablation Results

Furthermore, we conducted a series of ablation experiments to assess the effectiveness of various components within the network and loss function. The quantitative results of these experiments are summarized in Table 2.

- In Row 1, we present the results of the standard U-Net [12] for dose prediction, which is used as the baseline for comparison.
- In Row 2, the convolution block of the baseline is replaced with the MSFE block, resulting in improved performance compared to Row 1. This highlights the importance of extracting multi-scale features.

- Row 3 builds upon Row 2 by adding the CSA block, leading to further performance improvements, indicating that channel and spatial attention helps the model focus on essential features while minimizing irrelevant ones.
- In Row 4, we incorporated the PTV information using the PI module, which resulted in improved performance. This highlights the effectiveness of integrating PTV region information at each stage of the $\mathcal{E}$.
- In Row 5, we use the AF module to fuse the information from $\mathcal{E}$ and $\mathcal{D}$ at each stage of the network. This results in the best performance, highlighting the effectiveness of attention based fusion.

For the ablation experiments on the loss function shown in Table 2 (b), the L1 loss (i.e., Row 1) across all voxels is used as the baseline for comparison. Building upon this, Row 2 introduces the $\mathcal{L}_{PTV}$ loss, which guides the model to focus on minimizing errors within the PTV, leading to improved performance compared to Row 1. Row 3 further refines the model by incorporating all three loss functions, striking a balance between accurate dose prediction within the PTV and minimizing the dose to OARs, resulting in an additional performance enhancement.

Table 2: Results of the Ablation Experiments: The best performance is highlighted in bold.

(a) Network modules

| Row ID | MSCSA | | PI | AF | DS (↓) | DVHS (↓) |
|---|---|---|---|---|---|---|
| | MSFE | CSA | | | | |
| 1 | ✗ | ✗ | ✗ | ✗ | 3.352 | 2.954 |
| 2 | ✓ | ✗ | ✗ | ✗ | 3.110 | 2.741 |
| 3 | ✓ | ✓ | ✗ | ✗ | 2.993 | 2.353 |
| 4 | ✓ | ✓ | ✓ | ✗ | 2.589 | 1.621 |
| 5 | ✓ | ✓ | ✓ | ✓ | **2.190** | **1.160** |

(b) Loss function

| Row ID | $\mathcal{L}_{L1}$ | $\mathcal{L}_{PTV}$ | $\mathcal{L}_{OAR}$ | DS (↓) | DVHS (↓) |
|---|---|---|---|---|---|
| 1 | ✓ | ✗ | ✗ | 2.401 | 1.451 |
| 2 | ✓ | ✓ | ✗ | 2.273 | 1.284 |
| 3 | ✓ | ✓ | ✓ | **2.190** | **1.160** |

## 4 Conclusion

In this work, we introduced the Region-aware Attention Network (RANDose), a deep learning-based approach for radiation dose prediction. The model combines Multi-Scale Channel Spatial Attention (MSCSA), PTV Integration (PI), and Attention Fusion (AF) modules, along with a Region-Aware Loss function, to improve dose prediction. RANDose captures multi-scale features important for accurate dose delivery to the PTV while minimizing exposure to OARs. Experimental results on the OpenKBP dataset show that RANDose outperforms existing models. For future work, we plan to enhance the network to predict dose

without requiring OARs as input. This improvement will enable more flexible dose predictions, reduce the need for manual OAR delineation, and enhance the model's applicability in real-world clinical scenarios where OARs may not always be available or accurately defined.

**Disclosure of Interests** The authors have no competing interests to declare that are relevant to the content of this article.

## References


1. Ahmed, S.F., Alam, M.S.B., Hassan, M., Rozbu, M.R., Ishtiak, T., Rafa, N., Mofijur, M., Shawkat Ali, A., Gandomi, A.H.: Deep learning modelling techniques: current progress, applications, advantages, and challenges. Artificial Intelligence Review **56**(11), 13521–13617 (2023)
2. Atun, R., Jaffray, D.A., Barton, M.B., Bray, F., Baumann, M., Vikram, B., Hanna, T.P., Knaul, F.M., Lievens, Y., Lui, T.Y., et al.: Expanding global access to radiotherapy. The lancet oncology **16**(10), 1153–1186 (2015)
3. Babier, A., Zhang, B., Mahmood, R., Moore, K.L., Purdie, T.G., McNiven, A.L., Chan, T.C.: Openkbp: the open-access knowledge-based planning grand challenge and dataset. Medical Physics **48**(9), 5549–5561 (2021)
4. Bayoudh, K., Knani, R., Hamdaoui, F., Mtibaa, A.: A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. The Visual Computer **38**(8), 2939–2970 (2022)
5. Chai, J., Zeng, H., Li, A., Ngai, E.W.: Deep learning in computer vision: A critical review of emerging techniques and application scenarios. Machine Learning with Applications **6**, 100134 (2021)
6. Elyan, E., Vuttipittayamongkol, P., Johnston, P., Martin, K., McPherson, K., Moreno-García, C.F., Jayne, C., Sarker, M.M.K.: Computer vision and machine learning for medical image analysis: recent advances, challenges, and way forward. Artificial Intelligence Surgery **2**(1), 24–45 (2022)
7. Jiang, H., Diao, Z., Shi, T., Zhou, Y., Wang, F., Hu, W., Zhu, X., Luo, S., Tong, G., Yao, Y.D.: A review of deep learning-based multiple-lesion recognition from medical images: classification, detection and segmentation. Computers in Biology and Medicine **157**, 106726 (2023)
8. Li, F., Niu, S., Han, Y., Zhang, Y., Dong, Z., Zhu, J.: Multi-stage framework with difficulty-aware learning for progressive dose prediction. Biomedical Signal Processing and Control **82**, 104541 (2023)
9. Lin, Y., Liu, Y., Chen, H., Yang, X., Ma, K., Zheng, Y., Cheng, K.T.: Lenas: Learning-based neural architecture search and ensemble for 3-d radiotherapy dose prediction. IEEE Transactions on Cybernetics (2024)
10. Painuli, D., Bhardwaj, S., et al.: Recent advancement in cancer diagnosis using machine learning and deep learning techniques: A comprehensive review. Computers in Biology and Medicine **146**, 105580 (2022)
11. Pramod, A., Naicker, H.S., Tyagi, A.K.: Machine learning and deep learning: Open issues and future research directions for the next 10 years. Computational analysis and deep learning for medical care: Principles, methods, and applications pp. 463–490 (2021)

12. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
13. Sharma, N., Sharma, R., Jindal, N.: Machine learning and deep learning applications-a vision. Global Transitions Proceedings **2**(1), 24–28 (2021)
14. Szalkowski, G., Xu, X., Das, S., Yap, P.T., Lian, J.: Automatic treatment planning for radiation therapy: A cross-modality and protocol study. Advances in Radiation Oncology **9**(12), 101649 (2024)
15. Taye, M.M.: Understanding of machine learning with deep learning: architectures, workflow, applications and future directions. Computers **12**(5), 91 (2023)
16. Wang, B., Teng, L., Mei, L., Cui, Z., Xu, X., Feng, Q., Shen, D.: Deep learning-based head and neck radiotherapy planning dose prediction via beam-wise dose decomposition. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 575–584. Springer (2022)
17. Xu, X., Lian, C., Yap, P., Wang, A., Chera, B., Shen, C., Lian, J.: Prediction of optimal dosimetry for intensity-modulated radiotherapy with a cascaded auto-content deep learning model. International Journal of Radiation Oncology, Biology, Physics **111**(3), e113 (2021)
18. Ye, B., Tang, Q., Yao, J., Gao, W.: Collision-free path planning and delivery sequence optimization in noncoplanar radiation therapy. IEEE transactions on cybernetics **49**(1), 42–55 (2017)
19. Zimmermann, L., Faustmann, E., Ramsl, C., Georg, D., Heilemann, G.: dose prediction for radiation therapy using feature-based losses and one cycle learning. Medical physics **48**(9), 5562–5566 (2021)