Toxicity Prediction in Pelvic Radiotherapy Using Multiple Instance Learning and Cascaded Attention Layers

Behnaz Elhaminia, Alexandra Gilbert, John Lilley, Moloud Abdar, Alejandro F Frangi, Andrew Scarsbrook, Ane Appelt, and Ali Gooya.

***Abstract*— Modern radiotherapy delivers treatment plans optimised on an individual patient level, using CT-based 3D models of patient anatomy. This optimisation is fundamentally based on simple assumptions about the relationship between radiation dose delivered to the cancer (increased dose will increase cancer control) and normal tissue (increased dose will increase rate of side effects). The details of these relationships are still not well understood, especially for radiation-induced toxicity. We propose a convolutional neural network based on multiple instance learning to analyse toxicity relationships for patients receiving pelvic radiotherapy. A dataset comprising of 315 patients were included in this study; with 3D dose distributions, pre-treatment CT scans with annotated abdominal structures, and patient-reported toxicity scores provided for each participant. In addition, we propose a novel mechanism for segregating the attentions over space and dose/imaging features independently for a better understanding of the anatomical distribution of toxicity. Quantitative and qualitative experiments were performed to evaluate the network performance. The proposed network could predict toxicity with 80% accuracy. Attention analysis over space demonstrated that there was a significant association between radiation dose to the anterior and right iliac of the abdomen and patient-reported toxicity. Experimental results showed that the proposed network had outstanding performance for toxicity prediction, localisation and explanation with the ability of generalisation for an unseen dataset.**

A.Go. is supported by the Startup Grant from the University of Leeds, CRUK Early Detection Sandpit awards (EDIFICE, A2902), AI and MR Physics Simulation (RG.COMP.118477). A.A. is supported by Yorkshire Cancer Research Fellowship (L389AA). A.S. is supported by a Cancer Imaging Grant from the Leeds Hospitals Charity (9R01-1403). A.F.F. is supported by a Royal Academy of Engineering Chair in Emerging Technologies (CiET1919/19).

BE and AFF are with Centre for Computational Imaging and Simulation Technologies in Biomedicine (CISTIB), Schools of Computing and Medicine, University of Leeds, Leeds, UK. AFF and AGo are with the Alan Turing Institute, London, UK. AFF is also with the Department of Electrical Engineering (ESAT), KU Leuven, Leuven, Belgium (e-mail: b.elhaminia1@leeds.ac.uk; a.frangi@leeds.ac.uk;).

AGo is with the School of Computing Science (IDA-Section) at the University of Glasgow, Scotland, UK. (email: ali.gooya@glasgow.ac.uk

AA, AGi, and AS are with the Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK (e-mail:a.l.appelt@leeds.ac.uk; a.gilbert@leeds.ac.uk; a.f.scarsbrook@leeds.ac.uk)

JL and AA are with the Department of Medical Physics, Leeds Cancer Centre, St James's University Hospitals, UK (email:johnlilley@nhs.net)

MA is with the Institute for Intelligent Systems Research and Innovation (IISRI), Deakin University, Australia. (e-mail: m.abdar1987@gmail.com).

***Index Terms*— deep learning, multiple instance learning, radiotherapy, outcome prediction, toxicity map**

I. Introduction

RADIATION Therapy (RT), used in the majority of cancer treatments, aims to eliminate cancerous tissue using ionising radiation. However, radiation can also damage normal tissues in organs around the tumour (organs at risks, OARs), which may lead to treatment induced toxicity, both in the short- and long-term. Every course of RT is carefully designed for the individual patient, to minimise irradiation of normal tissue and limit the risk of toxicity. Consequently, accurate assessment of the possible toxicity risks is a key step in RT treatment planning. However, the correlation between dose delivered to normal tissue and the risk of late toxicity is not well understood for many organs.

In recent years, medical image analysis has been leveraging machine learning and particularly deep learning techniques [1]. A number of studies have addressed the radiotherapy toxicity prediction problem with the help of artificial neural network-based methods [2]. Most of the previously published studies have been based on 3D convolutional neural networks (CNN), with a smaller number of authors using 2D neural networks. Zhen X et al. [3] proposed a 2D CNN model for toxicity prediction where the input is a 2D dose surface map constructed by unfolding the 3D dose distribution on the rectum. Although unfolding 3D dose can be applied to the rectum, it cannot be extrapolated to most of the other organs; the rectum is a hollow structure and approximating that with an unfolded 2D surface does not lead to information loss. Amongst studies exploring the correlation between treatment features and toxicity following RT with the help of 3D neural networks, some focused only on the importance of dose distributions while others included different information (for example CT images). Amongst the former, Ibragimov et al. [4], Liang et al. [5] and Yang et al. [6] proposed different 3D CNNs to investigate dose distributions for prediction of radiotherapy-induced toxicity. Ibragimov et al. compared the prediction power of their CNN with conventional dose volume histograms (DVH)-based predictors, and found almost two times fewer false-positives using the 3D CNN model.

To increase the prediction accuracy, some studies took other features into account in addition to dose distributions. In [7], the authors proposed a multi-path network; one path for the input of 3D dose and the other path for treatment features (patients' demographics, OAR properties, tumour size, etc.). Bin et al. in [8] used a ventilation image and the functional dose image (obtained by weighting the dose distribution with the ventilation image) in addition to dose distribution as inputs for their proposed CNN. A concatenated pair of 3D CT image and 3D dose distribution was considered as the input of the network in [9]. Men et al. [10] employed a 3D residual CNN with three inputs of CT, contour images and dose plans in order to predict xerostomia toxicity. Other input features such as PET imaging [11], and a combination of MRI and cone-beam CT scans [12] have also been studied.

There are some challenges with previous approaches, however. Medical applications require models with interpretability and clinicians need to understand which regions of dose distribution treatments have an impact on the outcome of RT, in order to be able to guide optimal treatment for the individual patient. Due to the complexity of neural networks, it can be difficult to explain the behaviour of the network. To overcome the problem, many of the previously discussed studies presented various visual explanations for the predicted outcome. Gradient-weighted class activation mapping (Grad-CAM) [13], utilised in [3], [5], [14]–[16], is a method for convolutional networks that localises the features which are important for the prediction made by the network. Grad-CAM uses the gradients of the feature maps of the network's last layer with respect to the decision. In [10] Men et al. explained the feature importance by visualising feature maps of the different layers of the network, although it is unclear exactly how these should be interpreted. The other challenge that limits the use of neural networks for real-world medical imaging problems is their training complexity. 3D medical data are generally of considerable size and require a network with millions of parameters, with resulting time and memory complexities. In addition, collecting medical datasets requires professional expertise for contouring and labelling, and therefore most medical datasets are small. This makes the learning process challenging. Using transfer learning [3]–[5], [7]–[9], oversampling and data synthesising [6], are among the approaches to overcome the data sparsity problem.

In this work we introduce a convolutional model to explore both spatial dose distribution information and patients' organ structure in order to predict grade ≥ 2 bowel urgency toxicity in patients treated with pelvic radiotherapy. The novelty of this work is twofold: firstly, using two attention modules provides a patient-specific explanation to understand which regions are contributing to risk, and allows for assessment of the relative contribution of CT and dose features to the final prediction. Secondly, dividing input data into smaller cubes, using multiple instance learning, noticeably reduces the number of network parameters and results in lower time and memory complexities. Additionally, we constructed an atlas to summarise correlation between anatomical regions and the toxicity. Quantitative and qualitative experiments are conducted to evaluate the proposed framework as a clinically convincing tool for outcome prediction.

The remainder of this paper is structured as follows: Section II describes the methodology for modelling toxicity prediction and, Section III presents the experimental results and discussion. Section IV provides an overview of the study and conclusions.

II. Materials and method

A. Patient Cohort

The cross-sectional study cohort comprised 315 patients with anal, rectal, cervical and endometrial cancer treated between 2009 and 2014 at Leeds Cancer Centre, UK with curative 3D-conformal (n=307) or Intensity Modulated Radiotherapy (n=8; IMRT) external beam radiotherapy to the pelvis. The National Research Ethics Service Leeds East Committee approved the original data collection study following ethical review (reference 13-YH-0156). Further use of data for the current project was provided by LeedsCAT research database (reference 19-YH-0300). Median duration of follow up from RT was 2 years (IQR: 1.4-3.5 years). The dataset consisted of radiotherapy dose distributions, CT scans and structure set files for each patient. All patients had their intestinal cavity structure contoured ('Bowel Bag') as per RTOG guidelines [17]. Patient-reported bowel urgency was assessed using the validated European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life questionnaire; "When you felt the urge to move your bowels, did you have to hurry to get to the toilet?". Responses use an ordinal scale: 0: "not at all"; 1: "a little"; 2: "quite a bit"; 3: "very much". For the purpose of this work, we classified patients into two categories; grade$\geq$ 2 for moderate/severe toxicity (85 patients) and grade$<$ 2 for no/mild toxicity (164 patients). For the sake of simplicity, we refer patients with grade$\geq$ 2 as patients with bowel urgency and grade$<$ 2 as patients without bowel urgency toxicity. 66 patients were excluded from the study due to missing data or having a stoma (item not relevant).

B. Data Pre-processing

Input data were preprocessed prior to network training. For 168 patients, due to the different treatment planning systems, there were multiple dose distributions per patient, each representing a single radiation beam out of the 2-7 used for treatment delivery. We combined these beam dose distributions into a single dose distribution to calculate the full delivered treatment for the individual patient. As all dose distributions had the same coordinate system, the combination was computed simply by summing over all of the single dose treatments. For radiobiological effect correction, for each voxel in the dose distribution with dose per fraction d, we recalculated the dose to the equivalent dose in 2 Gy fractions (EQD2) [18] by: $EQD2 = n * d * (d + \alpha/\beta)/(2 + \alpha/\beta)$, where n is the number of treatment fractions for patient's RT treatment and α/β is a constant controlling the fraction-size sensitivity (here set to 3 [19]). Both CT and dose images were spatially re-sampled with linear interpolation to voxel size of $0.97mm \times 0.97mm$ and thickness of $5mm$. All of the CT and dose volumes were rigidly registered to a reference image and a mask based on the bowel bag structure - the region of interest for bowel urgency toxicity - was applied to them.

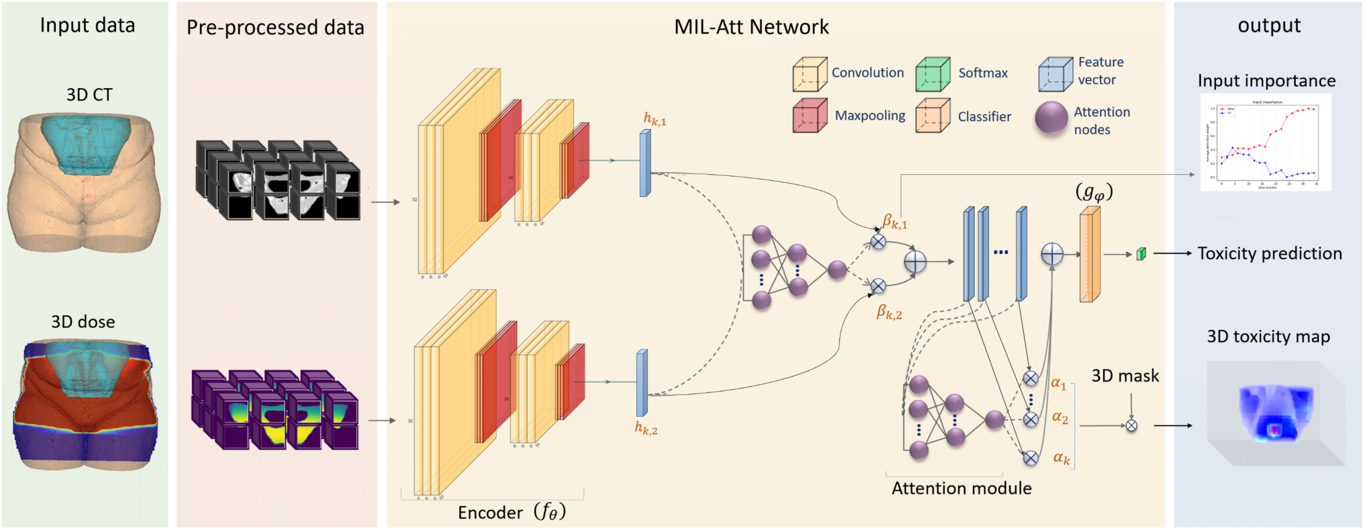

Fig. 1. The schematic illustration of the proposed model. 3D input images are pre-processed and fed into the Attention-MIL network. The output of the network is a binary variable defining toxicity prediction. Attention weights $\boldsymbol{\alpha_1, .., \alpha_k}$are utilised to generate toxicity maps; and weights of the first attention module, $\boldsymbol{\beta_1, ..., \beta_k}$, are extracted to analyse the impact of each input on the network's decision.

C. MIL Convolutional Neural Network with Attention Layers for Outcome Modelling

We developed a Multiple Instance Learning (MIL)-based convolutional neural network with attention mechanism (MIL-Att) to explore the existing relationship between the patient's dose distribution and CT images, and bowel urgency toxicity severity. The network consists of two input encoders, each extracting significant features from CT and dose distributions separately. Two attention modules are attached to explain the network behaviour, and one classification module is included to predict the outcome of toxicity. The general architecture of the network is depicted in Fig. 1.

1) MIL Convolutional Neural Network: In binary classification using CNN, one aims to find a model that assigns (predicts) a label $y \in \{0, 1\}$ for a given input image x. In the MIL model, however, the goal is to find a label for multiple instances from a same category. Let $X^{(K)} = \{x_1, x_2, ..., x_K\}$ denote a bag of K instances, the MIL model predicts the label $y^{(K)}$ for the entire bag. The problem of poorly annotated data makes MIL well suited to various medical imaging tasks [20], where the input image can be divided into multiple patches. We employed MIL for two reasons; firstly, medical data are large 3D images and having two different 3D inputs makes it impractical to train a deep neural network. We used MIL and divided each input (CT and dose) into smaller 3D cubes; this significantly reduces time and memory complexities and provides a model that exploits both modalities. Secondly, a key challenge in toxicity prediction is detecting the anatomical regions involved in toxicity. With MIL and attention-aggregation we can discover which instances in the bag (regions in the image) trigger the final decision.

2) Cascaded Attention Modules and Interpretation of Toxicity: In addition to outcome prediction, we aimed to provide an explanation for the network's decision. To meet this aim, we proposed a novel cascaded synergy of two attention modules that disentangle the attention weights over the feature and imaging space. This enables us not only to discover the locations of anatomical areas contributing to toxicity with ease, but also identify the separate contributions of CT and dose channels to toxicity prediction. Hence, for each patient, the network generates two toxicity risk maps; one highlights the overall high-risk areas, where the combinations of the patient's OAR and dose received drive the toxicity risk, and the other explains how CT and dose trigger the network's decision.

In a typical MIL classification problem, the features extracted from all of the instances are passed through a fully-connected layer with an equal weight for classification decision. To generate toxicity risk maps, similar to [21] and [22], we used the concept of attention mechanism [23], which is to compute the weighted average of the features (in this case, the extracted features do not have the same weight). In our network, these weights were computed by feed-forward neural networks - called attention modules - that are jointly trained along with other modules of the network.

3) Model Formulation: Consider input $X^{(K)}$ as a bag of K instances, $X^{(K)} = \{\mathsf{X}_{1,i}, \mathsf{X}_{2,i}, ..., \mathsf{X}_{K,i}\}$, where $\mathsf{X}_{k,i}$ represents the k^{th} instance from i^{th} input of the bag. For the sake of simplicity, we use the notation X and y instead of $X^{(K)}$and $y^{(K)}$ for each bag and label respectively. All the notations are summarised in Table I.

The toxicity prediction problem can be formulated as predicting the bag label y as a posterior probability obtained by:

$$y = \Phi_\Omega(X), \quad y \in [0, 1], \tag{1}$$

where Ω are the parameters of the model. Considering each instance $\mathsf{X}_{k,i}$ within the bag as a 3D tensor, we can define the encoder part of the network with f_{θ_i} which is a neural network with parameter θ_i and the output of $\boldsymbol{h}_{k,i}$, i.e, $\boldsymbol{h}_{k,i} = f_{\theta_i}(\mathsf{X}_{k,i})$. Having $\boldsymbol{h}_{k,i}$ as a feature vector extracted from cube k in input i, the attention weights signifying the importance of each input

Notation	Description	Value
k	Instance number	$1 \leq k \leq K$ (K can be different for each bag)
i	Input number	$i \in \{1, 2\}$, CT: $i = 1$, dose: $i = 2$
$\boldsymbol{h}_{k,i}$	Feature vector for instance k and input i	$\boldsymbol{h}_{k,i} \in \mathbb{R}^{1\times l}$
$\boldsymbol{w}, \mathbf{V}$	Weights for input attention	$\boldsymbol{w} \in \mathbb{R}^{d\times 1}, \mathbf{V} \in \mathbb{R}^{d\times l}$
$\beta_{k,i}$	Attention weights for instance k, input i	$\beta_{k,i} \in \mathbb{R}\vert\ \ 0 \leq \beta_{k,i} \leq 1$
$\boldsymbol{z}_k$	Weighted feature for instance k	$\boldsymbol{z}_k \in \mathbb{R}^{1\times l}$
$\boldsymbol{q}, \mathbf{R}$	Weights for region attention	$\boldsymbol{q} \in \mathbb{R}^{p\times 1}, \mathbf{R} \in \mathbb{R}^{p\times l}$
α_k	Attention weights for instance k	$\alpha_k \in \mathbb{R}\vert\ \ 0 \leq \alpha_k \leq 1$
$\boldsymbol{s}$	Weighted feature for input bag	$\boldsymbol{s} \in \mathbb{R}^{1\times l}$

TABLE I
SUMMARY OF THE NOTATIONS.

channel (CT vs dose) are computed as:

$$\beta_{k,i} = \frac{exp\{\boldsymbol{w}^T tanh(\mathbf{V}\boldsymbol{h}_{k,i}^T)\}}{\sum\limits_{j=1}^{2} exp\{\boldsymbol{w}^T tanh(\mathbf{V}\boldsymbol{h}_{k,j}^T)\}} \tag{2}$$

where $\mathbf{V}$ and $\boldsymbol{w}$ are weights matrix and vector, respectively, for a feed-forward neural network (attention module). We compute the total feature vector extracted for the cube k as:

$$\boldsymbol{z}_k = \sum_{i=1}^{2} \beta_{k,i} * \boldsymbol{h}_{k,i}. \tag{3}$$

Then the attention weights for each cube k is computed by the second attention module using:

$$\alpha_k = \frac{exp\{\boldsymbol{q}^T tanh(\mathbf{R}\boldsymbol{z}_k^T)\}}{\sum\limits_{j=1}^{K} exp\{\boldsymbol{q}^T tanh(\mathbf{R}\boldsymbol{z}_j^T)\}}, \tag{4}$$

where $\mathbf{R}$ and $\boldsymbol{q}$ are the weight parameters for the second attention module. We compute the ultimate feature vector fed to the classification module via:

$$\boldsymbol{s} = \sum_{k=1}^{K} \alpha_k * \boldsymbol{z}_k. \tag{5}$$

Considering (1), the toxicity classification problem can be written as:

$$\begin{aligned} &y = \Phi_\Omega(X) = g_\varphi(\boldsymbol{s}), \\ &g_\varphi : \boldsymbol{s} \mapsto [0, 1], \quad \Omega = \{\varphi, \theta\,, \boldsymbol{w}, \mathbf{V}, \boldsymbol{q}, \mathbf{R}\}. \end{aligned} \tag{6}$$

Let $t \in \{0, 1\}$ be the target class for X, training of the model can be performed by minimising the negative log-likelihood as the loss function L as:

$$L(t, \Phi_\Omega) = -tlog(\Phi_\Omega) - (1 - t)log(1 - \Phi_\Omega) \tag{7}$$

The log-likelihood is summed over all input bags from all the training set and minimisation is performed w.r.t. Ω parameters.

III. EXPERIMENTS AND DISCUSSION

A. Implementation Details

In our experiments we registered both CT and dose images using a 3D rigid transformation (SimpleITK [24], version 2.0.1) to a reference image with the dimension of $[35, 512, 512]$voxels (we selected the patient with the minimum number of CT slices for bowel bag as a reference). Each pixel had the dimension of $5mm \times 0.97mm \times 0.97mm$. We divided each input data into smaller cubes with the dimension of $[6, 32, 32]$voxels. The encoder consisted of two 3D convolutional layers each followed by maxpooling and batch normalisation layers. The two convolutional layers used 30 and 50 convolution filters with the kernel size of $(2, 3, 3)$. Adam optimisation with the learning rate of $1e^{-4}$ was employed. We set up $p, d = 512$ (number of neurons in attention modules, see Table I) for both attention modules. All convolutional layers were activated using ReLU functions and the last fully-connected layer was activated using Sigmoid function.

B. Training Strategy

We divided our dataset into training and test sets as follows: 20 patients with and 20 patients without bowel urgency were randomly selected and left aside as our test set and the rest (209 patients) were used for training. Due to the highly imbalanced training data (144 non-toxicity patients, 65 toxicity patients), we applied data augmentation for the minority class on the training data. Additive Gaussian noise with zero mean and 0.1 standard deviations and smoothing recursive Gaussian noise with 5 mm sigma across each axis were randomly applied to 65 patients with toxicity. Both filters were adopted from SimpleITK Python toolbox [24] (Version 2.1.1) for 3D image analysis. The augmentations were applied to CT and dose distributions in the original resolution (before data preprocessing). We randomly selected 20 patients (approximately 10% of the training dataset) as a validation set and in each epoch, we assessed network performance on this validation set. We did not apply cross-validation for two reasons: firstly with imbalanced data, dividing the dataset into smaller folds (subset) may result in some folds with no positive label. Therefore, evaluations with metrics such as accuracy and area under receiver operating characteristic curve (AUC) may not reflect the performance of the classifier correctly. Secondly, using data augmentation, validation must be applied only to non-augmented data. This implies that the data augmentation must be repeated in each fold, which is computationally expensive.

To increase the network generalisation while avoiding overfitting, we used transfer learning. Two autoencoder (AE) networks, sharing the same structure of MIL-Att encoders, were independently trained using the CT images and dose distributions, resulting in AE-CT and AE-dose, respectively. We trained AEs with all patients' data except the 40 left-out test cohort. We tested our MIL-Att framework with various

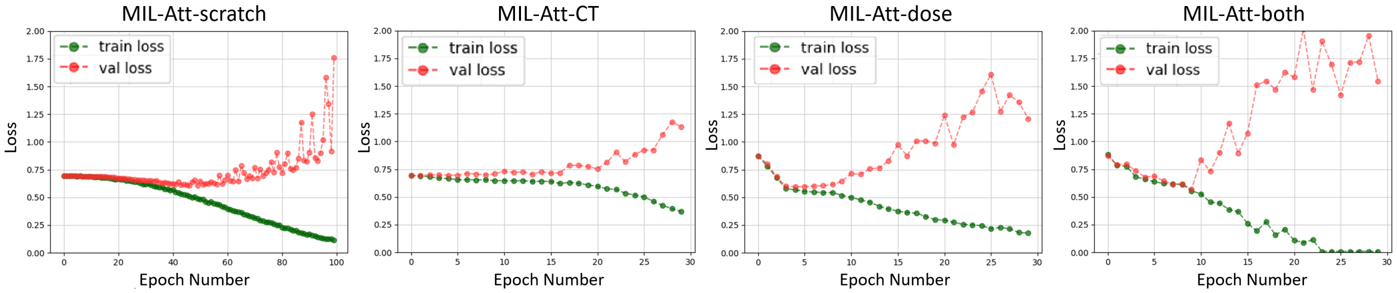

Fig. 2. Training and validation loss for four modes of MIL-Att network. Training from scratch converged at higher epochs while for pretrained modes it converged at lower epochs.

training strategies; firstly, we trained the network from scratch (MIL-Att-scratch), i.e. all the weights of the network were randomly initialised. Secondly, to investigate how transferred weights of CT and dose improve network performance, we trained the network in the following settings: both MIL-Att encoders were initialised with AE-CT weights (MIL-Att-CT), both encoders were initialised with AE-dose weights (MIL-Att-dose), and CT and dose encoders were initialised using AE-CT and AE-dose weights, respectively (MIL-Att-both). The comparison of training and validation losses are illustrated in Fig. 2. We trained all networks for several epochs, and for each epoch, we computed training and validation losses to evaluate the behaviour of our networks. For all training modes, the training loss and validation loss both decrease and stabilise at a specific point that indicates an optimal fit. The best model was selected based on this point (lowest validation loss). For training without transfer learning (MIL-Att-scratch), we trained the network for 100 epochs. Both training and validation loss decreased until epoch number 47 (see Fig.2 MIL-Att-scratch). After epoch 47, the training loss drastically decreased while the validation loss increased -the network was overfitted at this point. For transfer learning modes (MIL-Att-CT, MIL-Att-dose, MIL-Att-both), all networks converged in lower epochs (epochs $<$ 16) and we stopped training after 30 epochs (we reduced the learning rate for transfer learning modes to $1e^{-3}$); after epochs > 16, the validation loss was significantly greater than the training loss indicating the network was overfitted. Comparing different transfer learning modes, MIL-Att-both resulted in lower validation loss. Fig.2 shows how transfer learning improves performance and the complexity of different modes of training.

C. Prediction Performance

We quantified our network performance by calculating five evaluation metrics: accuracy, sensitivity, specificity, F1-score, and AUC; and compared the results with three existing models proposed by Yang et al. [6] , Ibragimov et al. [4] and Liang et al. [5], briefly reviewed here. Yang et al. [6] proposed a 3D CNN (CT-dose-CNN) with two input channels for CT and dose images to predict toxicity in prostate radiotherapy. They used transfer learning by training an autoencoder on their augmented image data and then employing the encoder part for their proposed network. In a similar vein, Ibragimov et al. [4] employed a 3D CNN (Dose-CNN) with three convolutional layers to predict liver toxicity. Their proposed network includes only one input channel to analyse dose distribution plan. Liang et al. [5] transferred the C3D proposed and learned in [25] to predict toxicity after lung RT. They trained the network in two settings; training all layers (C3D-FT) and just fine tuning the last layer (C3D-FE). We implemented all models in Python 3.7 and followed the described procedures to train them on our own dataset. Table II and Fig. 3 summarise the comparison results.

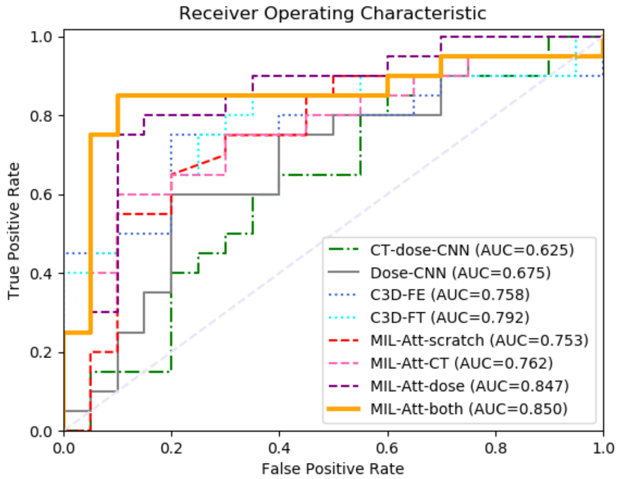

Fig. 3. Receiver operating curve analysis for toxicity prediction using the test set.

Method	Parameter	ACC	SPE	SEN	F1-score
MIL-Att	scratch	0.65	0.70	0.60	0.64
	CT	0.65	0.60	0.70	0.64
	dose	0.75	0.75	0.75	0.80
	both	**0.80**	0.80	**0.80**	**0.82**
C3D [5]	FE	0.65	**1.0**	0.30	0.65
	FT	0.72	0.75	0.70	0.75
CT-dose-CNN [6]	-	0.60	0.55	0.65	0.60
Dose-CNN [4]	-	0.60	0.85	0.35	0.46

***Abbreviation:**: Accuracy (ACC), Specificity (SPE), Sensitivity (SEN)

TABLE II
COMPARISON OF PREDICTION PERFORMANCE ACROSS DIFFERENT METHODS. BEST PERFORMANCE IN EACH METRIC IS SHOWN IN BOLD.

As seen in Table II, the prediction performance of MIL-Att-both reached the highest values for accuracy, sensitivity and F1-score, and for specificity, the C3D-FE model gained the greatest value (1.0). The specificity metric summarises how well the negative class is predicted and sensitivity is the complement to it. The low value of sensitivity (0.3) for C3D-FE, implies that the C3D-FE network is biased towards predicting lack of toxicity and does not learn the data distribution for the positive class (with toxicity). With the same inference, we can conclude that Dose-CNN is also biased to the negative class. The next best specificity metric is achieved by our model when both encoders are pre-trained. This is supported by ROC analysis (Fig.3) that shows the best AUC value is achieved by the MIL-Att-both method. To investigate significant statistical differences, AUC for all methods are compared using DeLong's test [26]. (see Fig. 4). AUC is significantly improved by using both CT and dose pre-trained networks (p-value< 0.05) compared to training from scratch. Performance of pre-training with CT did not differ statistically compared to not pre-training (p-value$= 0.913$). Furthermore, pre-training the network with both CT and dose does not result in better performance compared to a network pre-trained only with dose (p-value=0.958). These results suggest that pre-training the network with only dose data can improve the model performance. This might be attributable to the structure of dose data which is more difficult to be learnt; dose data is a grayscale image where the intensity changes very gradually while the CT data contains various edges, corners, ridges and interested points.

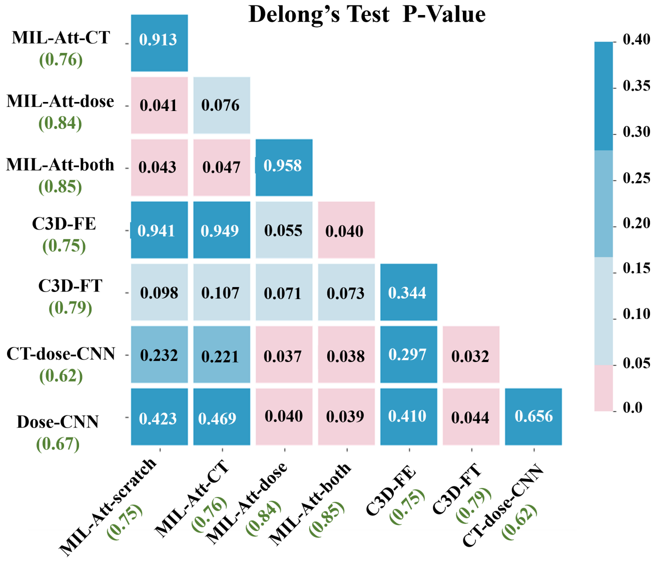

Fig. 4. AUC comparison using DeLong's test. Smaller p-values demonstrate significant differences. AUC values are in parentheses.

When we compare MIL-Att and C3D with the other networks, we see that they both perform better than CT-dose-CNN and Dose-CNN. This might be attributable to the depth of the networks or how they extract the features; in both networks (CT-dose-CNN and Dose-CNN), three convolutional layers extract features from the input data. In comparison, C3D transforms the input to the features within 8 convolutional layers. Although MIL-Att network extracts the features using two encoders each with two convolutional layers, this is achieved locally for every single cube. Considering that the average number of cubes for each input was 209, the network represents each input with 209 high-dimensional features. We can conclude that for CT-dose-CNN and Dose-CNN, the layers' architecture may not be adequate to capture the inherent pattern of the data distribution in our dataset.

The performance of C3D is the most similar to MIL-Att. Both learn deep feature maps either directly from the whole input or by investigating local regions in the data. However, the performance of MIL-Att is superior to C3D on our test set overall on the evaluation metrics. This is because the MIL-Att network analyses both CT and dose data, while C3D only analyses dose. This observation emphasises that exploring CT in addition to dose plans can provide more useful information and consequently improve prediction performance.

Comparing network architecture, multiple instance learning with attention employs fewer convolutional layers than C3D. This reduces the number of parameters that the model must learn; C3D has approximately 78 million parameters that require gradient, while this number is nearly 11 million for MIL-Att. We should consider that MIL-Att analyses two 3D input data and C3D explores only one input. The developed software of these models will be eventually used in hospitals and it needs to be stored in typical computers (not all hospitals have GPUs with large memories). Therefore, we can conclude that multiple instance learning can perform more efficiently in terms of time and resource usage when we have large data. Note that the idea of MIL might be similar to the other efficient models that explore local or adjacent patches/slices (for example, 2.5D networks), but the difference is that in MIL, the input is still the whole volume and the output is computed based on it (not only for each stack/cube). In classification problems, where the output is based on the whole input data (despite 3D segmentation), MIL can be memory/GPU usage efficient.

Comparison between CT-dose-CNN and dose-CNN shows that CT-dose-CNN achieved a lower AUC despite analysing both CT and dose. This could be caused by the fact that in Dose-CNN the input is rescaled and cropped into the size of [19,19,19] image, while CT-dose-CNN applies convolutional filters on the original size of CT and dose. This suggests that the latent space dimensionally utilised in the CT-dose-CNN does not project the inherent structure of the CT and dose data of our dataset.

We summarise the above discussion with three general points as follows:

1) The highest performance is achieved when both CT and dose images are adequately explored.
2) When memory is constrained and input data are large 3D volumes, multiple instance learning can encode the input more efficiently than the traditional deep learning models.
3) Identifying the most discriminating features is a pivotal argument in the toxicity problem.

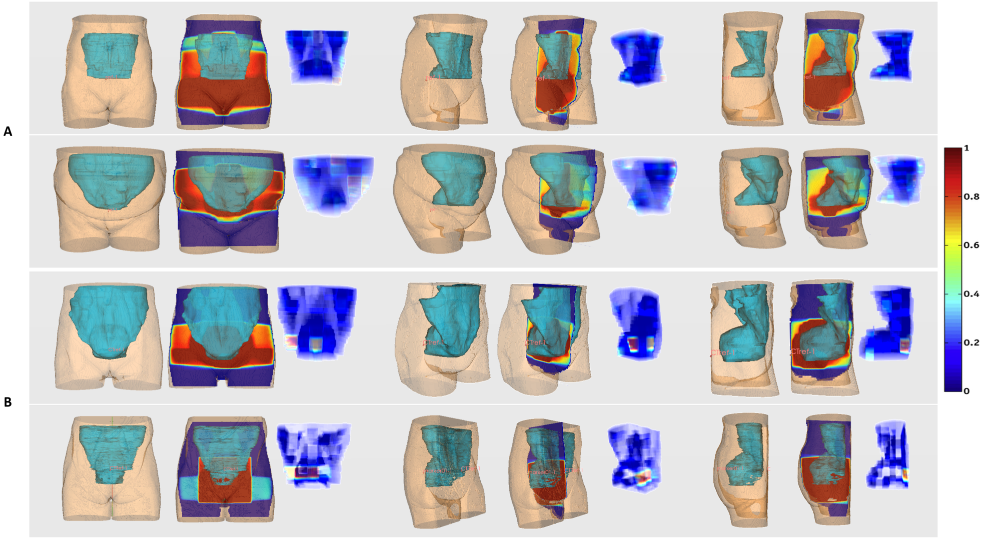

Fig. 5. Results of the toxicity map generated by our proposed method. A: two example patients without bowel urgency. B: Two examples of patients with bowel urgency. For each view, from left to right, first and second images are the patient's bowel bag structure and the radiation dose distribution, respectively. The third image is the toxicity map generated by the proposed model. Higher value of toxicity map indicates higher risk of toxicity.

D. Visualisation of Attention Maps

Another important aspect of the proposed MIL-Att model is that we can employ the attention mechanism to provide a visual explanation for the predicted outcome. In MIL-Att, the features are extracted from local areas and the attention module investigates how they are important for the final decision. Finding critical regions is important for clinicians to understand the network's decision and it can help with optimal dose planning. After training was completed, for each patient in test dataset, we extracted the second attention weights (α_k) to generate risk map. Fig. 5 shows the 3D visualisation of the toxicity maps for two patients without and two patients with reported bowel urgency toxicity, in rows A and B respectively. To allow better visualisation, we plotted the results in three different views. Higher values in the toxicity map show more decisive regions in the predicted toxicities. For both patients with toxicity, the attention weights are concentrated in anterior and iliac fossa anatomical regions of the bowel bag, suggesting that these regions are associated with higher-risk of bowel urgency toxicity. On the other hand, for patients without toxicity, the attention weights are scattered across the whole bowel bag and not localised to a specific region.

E. Comparison with Grad-CAM

As mentioned in Section I, many studies have used Grad-CAM to generate risk maps. To further evaluate our derived attention toxicity maps, we compared our best network (MIL-Att-both) results with Grad-CAM based on C3D-FT network. We used PyTorch-Grad-CAM library [27] and modified it for 3D data. Fig.6 shows the results of the comparison for two patients A and B, without and with toxicity, respectively. 3D comparison, for patient A without toxicity attention weights are spread within the bowel bag volume while for patient B with toxicity, all the attention is in the anterior aspect of the bowel bag. Grad-CAM also generated similar results; for patient A, gradients of the features are spread within the cube while for patient B, highest gradients are located anteriorly. For 2D comparison of the attention maps we plotted three slices covering different sections from bottom to top of the patients' abdomen. It can be seen that as slice number increases, the activation map generated by Grad-CAM changes slightly. This is because for the last convolution layer, the size of feature map is smaller than the size of input image and to generate an activation map the feature tensor must be up-sampled. This generates an approximation for critical regions. In the C3D network, the image dimension is [18,112,112] voxels, and features in the last convolution layer are with the dimension of [2,7,7] voxels. Furthermore, up-sampling on a grid, the Grad-CAM activation map does not comply with the morphology of the patient's bowel bag anatomy. In contrast, as attention weights using our MIL-Att are generated individually for each cube in the bowel bag, toxicity localisation is more reliable than Grad-CAM; in 3D Grad-CAM (Fig.6), the whole anterior portion of the activation maps are highlighted as critical regions.

F. Input Importance

Previously discussed in Section I, additional to dose distribution, recently developed deep learning methods applied other input factors (e.g, CT, PET) for their network. Men et al. [10] demonstrated that prediction performance increased when the network was trained with both CT and dose distributions. To the best of our knowledge none of the existing works explored the relative importance of the inputs on toxicity prediction. With the first attention module, we evaluated the significance of each input regarding the network decision. Fig. 7 shows the average of attention weights for each CT and dose slice within the bowel bag. For the inferior slices of the bowel bag (slice number< 10), both CT and dose slices gained high importance. This implies that the anatomical information (from the CT) and irradiated dose are both associated with the risk of toxicity when analysing the caudal part of the bowel

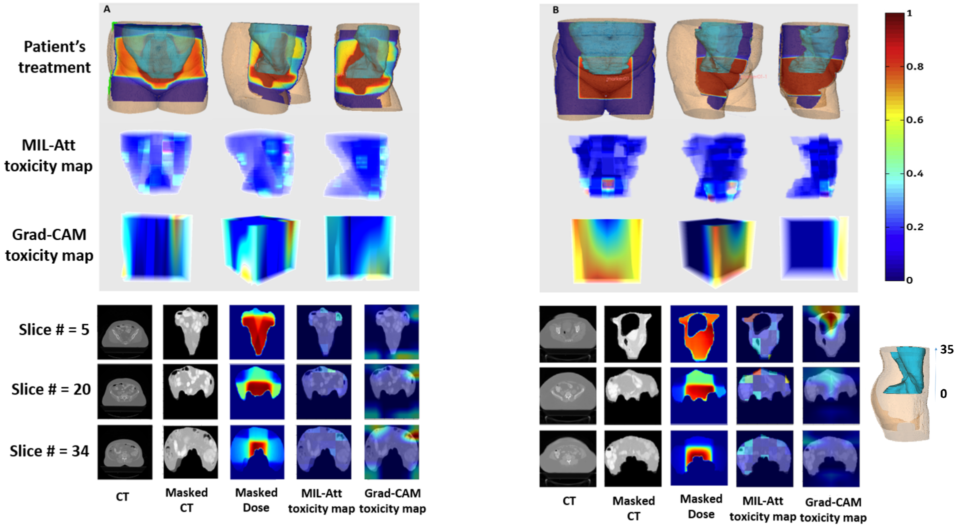

Fig. 6. 3D/2D comparison of generated toxicity map by MIL-Att and Grad-CAM for two patients A and B without and with bowel urgency, respectively.

bag. For the more cranial part (slice number> 10), the dose distribution becomes more important for toxicity prediction than the CT features. This makes sense from a clinical point of view: The amount of bowel present in the lower (caudal) part of the pelvis can vary significantly from patient to patient (e.g. due to bladder size and location); and the risk of bowel-related toxicity will depend not only on the dose delivered to that area, but also on whether the individual patient has low-lying small bowel loops. Conversely, the presence of bowel in the upper (cranial) part of the pelvis is much less varied (i.e., all patients will have small and large bowel in this area), and the dose delivered becomes the sole differentiating factor between patients with and without bowel urgency toxicity. As an example, see Fig. 6, where the structure of bowel bag in the slices numbered 20 and 34 are nearly similar for patient A and B, whereas, for slice number 5 -inferior aspect of the bowel bag- patient A and B have different shapes. We computed the average of the attention weights across all the slices. The average weight computed for CT and dose input were 0.17 and 0.53, respectively. Normalising overall weights, the dose image had a 76% association with bowel urgency toxicity whilst the CT image had only a 24% association.

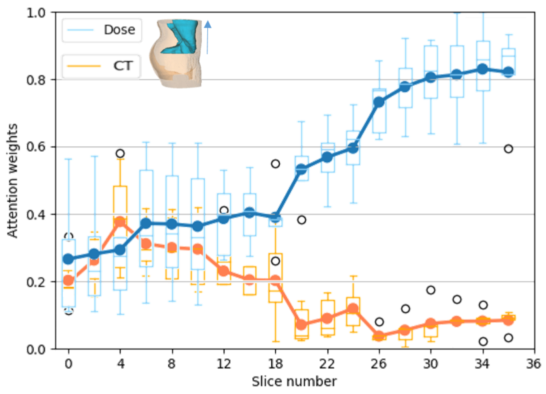

Fig. 7. Quantitative evaluation of input association with toxicity. Higher value of attention weights shows higher impact on toxicity prediction.

G. Toxicity Atlas Over All Patient Population

The toxicity risk maps vary between individuals and this is due to patients having organs of different sizes and shapes. To investigate correlation between anatomical regions and bowel urgency toxicity, rather than relying on individual risk maps, we constructed a toxicity atlas from co-registered maps generated by the proposed network. The data of CT images for 85 patients with toxicity were co-registered to a reference patient using 3D Diffeomorphic Demons registration algorithm [28]. We also registered dose distributions using the computed transformation of their corresponding CT. The average of all the registered images were computed in order to construct the atlas. Considering attention atlas in Fig. 8, we conclude the irradiated dose received by the anterior and right iliac fossa anatomical regions of the bowel bag is related to bowel urgency toxicity. Clinically, these findings may allow application in RT treatment planning to avoid dose to these more sensitive anatomical regions. The anterior dose region is most likely related to small bowel loops within the radiotherapy field. With increasing use of IMRT (where RT dose may be sculpted to more readily avoid OAR structures) within clinical practice the dose to this region in current clinical practice is likely to be lower than within this series, where the majority of patients received 3D-conformal RT. In comparison, the right iliac fossa region most likely relates to dose to the terminal ileum. This area is commonly affected in Crohn's disease and known to be associated with significant bowel symptoms, including urgency and diarrhoea. Damage to this area may be related to nutritional deficiencies such as bile acid malabsorption, which has been identified as a known side effect of pelvic radiotherapy [29]. This anatomical region is not currently included as an avoidance structure in routine radiotherapy practice and should be an area of future work.

IV. Conclusion

We proposed a novel deep learning architecture based on multiple instance learning and attention mechanism for the prediction of patient-reported bowel urgency toxicity in

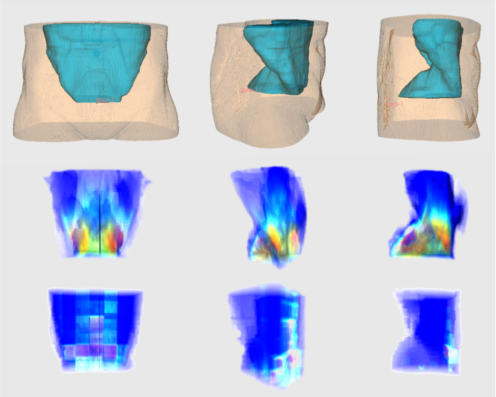

Fig. 8. Toxicity model. From top to bottom: reference patient, average of irradiated dose and atlas for toxicity. Generated atlas localises high-risk regions for toxicity with higher values.

patients with pelvic radiotherapy. Additionally, the proposed model provides visual explanation to create a comprehensive understanding of the network operation for predicting the bowel-related toxicity. Two attention modules are incorporated into the network to explain: (1) which anatomical regions are associated with high risk of toxicity and (2) how CT and dose images are impacting the network's prediction. Furthermore, we construct a toxicity atlas to integrate information from multiple toxicity risk maps and localise, visualise and summarise toxicity based on the bowel bag structure. The comparative experiment demonstrated the framework offers clinically convincing tools for radiotherapy toxicity prediction.

References

[1] J. R. Burt, N. Torosdagli, N. Khosravan, H. RaviPrakash, A. Mortazi, F. Tissavirasingham, S. Hussein, and U. Bagci, "Deep learning beyond cats and dogs: recent advances in diagnosing breast cancer with deep neural networks," *Br J Radiol.*, vol. 91, no. 1089, p. 20170545, 2018.

[2] A. Appelt, B. Elhaminia, A. Gooya, A. Gilbert, and M. Nix, "Deep learning for radiotherapy outcome prediction using dose data–a review," *Clin. Oncol (R Coll Radiol).*, vol. 34, no. 2, pp. e87–e96, 2022.

[3] X. Zhen, J. Chen, Z. Zhong, B. Hrycushko, L. Zhou, S. Jiang, K. Albuquerque, and X. Gu, "Deep convolutional neural network with transfer learning for rectum toxicity prediction in cervical cancer radiotherapy: a feasibility study," *Phys. Med. Biol.*, vol. 62, no. 21, pp. 8246–8263, 2017.

[4] B. Ibragimov, D. Toesca, D. Chang, Y. Yuan, A. Koong, and L. Xing, "Development of deep neural network for individualized hepatobiliary toxicity prediction after liver SBRT," *Med. Phys.*, vol. 45, no. 10, pp. 4763–4774, 2018.

[5] B. Liang, Y. Tian, X. Chen, H. Yan, L. Yan, T. Zhang, Z. Zhou, L. Wang, and J. Dai, "Prediction of radiation pneumonitis with dose distribution: a convolutional neural network (CNN) based model," *Front. Oncol.*, vol. 9, p. 1500, 2020.

[6] Z. Yang, D. Olszewski, C. He, G. Pintea, J. Lian, T. Chou, R. C. Chen, and B. Shtylla, "Machine learning and statistical prediction of patient quality-of-life after prostate radiation therapy," *Comput. Biol. Med.*, vol. 129, p. 104127, 2021.

[7] B. Ibragimov, D. A. Toesca, Y. Yuan, A. C. Koong, D. T. Chang, and L. Xing, "Neural networks for deep radiotherapy dose analysis and prediction of liver SBRT outcomes," *IEEE J Biomed. Health. Inform.*, vol. 23, no. 5, pp. 1821–1833, 2019.

[8] L. Bin, T. Yuan, S. Zhaohui, R. Wenting, L. Zhiqiang, H. Peng, Y. Shuying, D. Lei, W. Jianyang, W. Jingbo, *et al.*, "A deep learning-based dual-omics prediction model for radiation pneumonitis," *Med. Phys.*, vol. 48, no. 10, pp. 6247–6256, 2021.

[9] B. Ibragimov, D. A. Toesca, D. T. Chang, Y. Yuan, A. C. Koong, L. Xing, and I. R. Vogelius, "Deep learning for identification of critical regions associated with toxicities after liver stereotactic body radiation therapy," *Med. phys.*, vol. 47, no. 8, pp. 3721–3731, 2020.

[10] K. Men, H. Geng, H. Zhong, Y. Fan, A. Lin, and Y. Xiao, "A deep learning model for predicting xerostomia due to radiation therapy for head and neck squamous cell carcinoma in the RTOG 0522 clinical trial," *Int. J. Radiat. Oncol. Biol. Phys*, vol. 105, no. 2, pp. 440–447, 2019.

[11] C. Wang, C. Liu, Y. Chang, K. Lafata, Y. Cui, J. Zhang, Y. Sheng, Y. Mowery, D. Brizel, and F.-F. Yin, "Dose-distribution-driven PET image-based outcome prediction (DDD-PIOP): A deep learning study for oropharyngeal cancer IMRT application," *Front. Oncol.*, vol. 10, p. 1592, 2020.

[12] C. Wang, S. R. Alam, S. Zhang, Y.-C. Hu, S. Nadeem, N. Tyagi, A. Rimner, W. Lu, M. Thor, and P. Zhang, "Predicting spatial esophageal changes in a multimodal longitudinal imaging study via a convolutional recurrent neural network," *Phys. Med. Biol.*, vol. 65, no. 23, p. 235027, 2020.

[13] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Conf. Comput. Vision*, pp. 618–626, 2017.

[14] A. Diamant, A. Chatterjee, M. Vallières, G. Shenouda, and J. Seuntjens, "Deep learning in head and neck cancer outcome prediction," *Sci. Rep.*, vol. 9, no. 1, pp. 1–10, 2019.

[15] Y. Jiang, C. Jin, H. Yu, J. Wu, C. Chen, Q. Yuan, W. Huang, Y. Hu, Y. Xu, Z. Zhou, *et al.*, "Development and validation of a deep learning CT signature to predict survival and chemotherapy benefit in gastric cancer: a multicenter, retrospective study," *Ann. Surg.*, vol. 274, no. 6, pp. e1153–e1161, 2021.

[16] S. Cui, R. K. Ten Haken, and I. El Naqa, "Integrating multiomics information in deep learning architectures for joint actuarial outcome prediction in non-small cell lung cancer patients after radiation therapy," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 110, no. 3, pp. 893–904, 2021.

[17] H. A. Gay, H. J. Barthold, E. O'Meara, W. R. Bosch, I. El Naqa, R. Al-Lozi, S. A. Rosenthal, C. Lawton, W. R. Lee, H. Sandler, *et al.*, "Pelvic normal tissue contouring guidelines for radiation therapy: a radiation therapy oncology group consensus panel atlas," *Int. J. Radiat. Oncol. Biol. Phys*, vol. 83, no. 3, pp. e353–e362, 2012.

[18] S. M. Bentzen, W. Dörr, R. Gahbauer, R. W. Howell, M. C. Joiner, B. Jones, D. T. Jones, A. J. Van Der Kogel, A. Wambersie, and G. Whitmore, "Bioeffect modeling and equieffective dose concepts in radiation oncology–terminology, quantities and units," *Radiother. Oncol.*, vol. 105, no. 2, pp. 266–268, 2012.

[19] M. A. Benadjaoud, P. Blanchard, B. Schwartz, J. Champoudry, R. Bouaita, D. Lefkopoulos, E. Deutsch, I. Diallo, H. Cardot, and F. de Vathaire, "Functional data analysis in NTCP modeling: a new method to explore the radiation dose-volume effects," *Int. J. Radiat. Oncol. Biol. Phys*, vol. 90, no. 3, pp. 654–663, 2014.

[20] G. Quellec, G. Cazuguel, B. Cochener, and M. Lamard, "Multiple-instance learning for medical image and video analysis," *IEEE Rev. Biomed. Eng.*, vol. 10, pp. 213–234, 2017.

[21] A. Sadafi, A. Makhro, A. Bogdanova, N. Navab, T. Peng, S. Albarqouni, and C. Marr, "Attention based multiple instance learning for classification of blood cell disorders," in *MICCAI*, pp. 246–256, 2020.

[22] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proc. Int. Conf. Mach. Learn.*, pp. 2127–2136, 2018.

[23] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint:1409.0473*, 2014.

[24] B. C. Lowekamp, D. T. Chen, L. Ibáñez, and D. Blezek, "The design of simpleitk," *Front. Neuroinform.*, vol. 7, p. 45, 2013.

[25] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE In. Conf. Comput. Vision*, pp. 4489–4497, 2015.

[26] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics*, pp. 837–845, 1988.

[27] J. Gildenblat and contributors, "PyTorch library for CAM methods." url: https://github.com/jacobgil/pytorch-grad-cam, 2021.

[28] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, "Diffeomorphic demons: Efficient non-parametric image registration," *NeuroImage*, vol. 45, no. 1, pp. S61–S72, 2009.

[29] J. Andreyev, "Gastrointestinal symptoms after pelvic radiotherapy: a new understanding to improve management of symptomatic patients," *Lancet. Oncol.*, vol. 8, no. 11, pp. 1007–1017, 2007.