cancers

MDPI

Article

Feature Importance Analysis of a Deep Learning Model for Predicting Late Bladder Toxicity Occurrence in Uterine Cervical Cancer Patients

Wonjoong Cheon [1], Mira Han [2,†], Seonghoon Jeong [1], Eun Sang Oh [1], Sung Uk Lee [1], Se Byeong Lee [1], Dongho Shin [1], Young Kyung Lim [1], Jong Hwi Jeong [1], Haksoo Kim [1,*] and Joo Young Kim [1,*]

[1] Proton Therapy Center, National Cancer Center, Goyang-si 10408, Republic of Korea; wonjoongcheon@gmail.com (W.C.)
[2] Biostatistics Collaboration Team, National Cancer Center, Goyang-si 10408, Republic of Korea
* Correspondence: haksoo.kim@ncc.re.kr (H.K.); jooyoungcasa@ncc.re.kr (J.Y.K.)
† Current address: Department of Medical Research Collaborating Center, Seoul Metropolitan Government—Seoul National University Boramae Medical Center, Seoul 07061, Republic of Korea.

Simple Summary: This study developed a prediction model for late bladder toxicity in patients with uterine cervical cancer undergoing radiation therapy. A deep learning (DL) model was trained on data from 281 patients and compared its performance with a multivariable logistic regression model. The DL model outperformed the regression model, achieving higher accuracy, recall, F1-score, and area under the receiver operating characteristic curve. Specifically, based on the feature importance analysis, the DL model identified the doses for the most exposed 2 cc volume of the bladder (BD_{2cc}), BD_{5cc}, and ICRU bladder point as high-priority features. Finally, the lightweight DL model, which was designed to focus on the top five important features, demonstrated superior predictive capabilities, highlighting its potential in improving patient outcomes and minimizing treatment-related complications with secured reliability.

Abstract: (1) In this study, we developed a deep learning (DL) model that can be used to predict late bladder toxicity. (2) We collected data obtained from 281 uterine cervical cancer patients who underwent definitive radiation therapy. The DL model was trained using 16 features, including patient, tumor, treatment, and dose parameters, and its performance was compared with that of a multivariable logistic regression model using the following metrics: accuracy, prediction, recall, F1-score, and area under the receiver operating characteristic curve (AUROC). In addition, permutation feature importance was calculated to interpret the DL model for each feature, and the lightweight DL model was designed to focus on the top five important features. (3) The DL model outperformed the multivariable logistic regression model on our dataset. It achieved an F1-score of 0.76 and an AUROC of 0.81, while the corresponding values for the multivariable logistic regression were 0.14 and 0.43, respectively. The DL model identified the doses for the most exposed 2 cc volume of the bladder (BD_{2cc}) as the most important feature, followed by BD_{5cc} and the ICRU bladder point. In the case of the lightweight DL model, the F-score and AUROC were 0.90 and 0.91, respectively. (4) The DL models exhibited superior performance in predicting late bladder toxicity compared with the statistical method. Through the interpretation of the model, it further emphasized its potential for improving patient outcomes and minimizing treatment-related complications with a high level of reliability.

Keywords: uterine cervical cancer; toxicity prediction; deep learning; feature importance; interpretable artificial intelligence

Citation: Cheon, W.; Han, M.; Jeong, S.; Oh, E.S.; Lee, S.U.; Lee, S.B.; Shin, D.; Lim, Y.K.; Jeong, J.H.; Kim, H.; et al. Feature Importance Analysis of a Deep Learning Model for Predicting Late Bladder Toxicity Occurrence in Uterine Cervical Cancer Patients. *Cancers* **2023**, *15*, 3463. https://doi.org/10.3390/cancers15133463

Academic Editors: Yudong Zhang, Juan Manuel Gorriz and Zhengchao Dong

Received: 30 May 2023
Revised: 28 June 2023
Accepted: 29 June 2023
Published: 2 July 2023

1. Introduction

Cervical cancer is the third most commonly diagnosed malignancy in women worldwide, as reported in 2022 [1]. With early detection and advances in treatment, the number of cervical cancer survivors has increased over the past 40 years. Nevertheless, most patients with cervical cancer suffer from various morbidities resulting from the disease itself or from the treatment. Depending on the treatment strategy (surgery, radiotherapy (RT), or chemotherapy), the morbidities may include symptoms associated with the gastrointestinal or urinary tract, lymphedema, and sexual dysfunctions.

External beam radiation therapy (EBRT) with concurrent chemotherapy followed by intracavitary brachytherapy represents the standard treatment for locally advanced cervical cancers [2,3]. Owing to the very high dose of radiation delivered to the pelvic area using EBRT and brachytherapy, radiation-induced toxicity in these patients is not negligible [4]. Among the irradiated pelvic organs, radiation-induced bladder toxicity is the most commonly observed morbidity in cervical cancer patients who have received curative radiotherapy, with an incidence of approximately 20% [4]. Although severe bladder toxicities have become rare with image-guided brachytherapy, they still pose a considerable risk [5].

Predicting the occurrence of bladder toxicity before treatment is crucial for improving patients' long-term health outcomes and quality of life, as it reduces the probability of treatment-related complications or interruptions. Previous studies have focused on investigating prognostic factors associated with radiation-induced bladder toxicity. These studies have examined various dose volumetric parameters, such as the doses for the most exposed 2 cc volume of the bladder (BD_{2cc}) [5] and the volume receiving a certain biologically weighted equivalent dose (EQD2) relative to the gross tumor volume (GTV) or clinical target volume (CTV), such as $V_{51.43Gy}$ [6] and $V_{8.5Gy/w}$ [7].

Statistical methods and deep learning (DL) models have been introduced to predict the toxic effects of radiation therapy on the gastrointestinal [8–15] and genitourinary systems [6,7,16–24]. Statistical methods such as univariable or multivariable linear regression, logistic regression [6,7,10,13,16,17,19], Cox regression [16,17,19,20,22], random forest [11], support vector machine [15,21,24], genetic algorithms [21], and statistical analysis [8,9,12] have been used to predict clinical outcomes. In DL methods, convolutional neural networks (CNN) and multilayer perceptrons (MLP) have been employed to predict radiation-induced toxicity [14,21].

Related studies that have used predictive models for toxicity in various cancer types and radiation therapy techniques:

- A machine-learning-based prediction model of fistula formation after interstitial brachytherapy for locally advanced gynecological malignancies achieved an accuracy of 0.901 using Support Vector Machine (SVM) [24].
- A feasibility study utilized a deep convolutional neural network (CNN) with transfer learning to predict rectum toxicity in cervical cancer radiotherapy, achieving an AUC of 0.89 [14].
- An observational study predicting radiotherapy impact on late bladder toxicity in prostate cancer patients used univariate logistic regression, achieving an AUC of 0.626 [6].
- Various studies have focused on predicting urinary toxicity in prostate cancer radiotherapy using different models, such as the international prostate symptoms score model, logistic and Cox regression, the edited nearest neighbor algorithm together with the regularized discriminant analysis classifier, and others [16,18,19].
- Predicting late organ-at-risk toxicity after prostate radiation therapy has been explored using statistical analysis, cox regression, and random forest models [11,22,23].
- In radiotherapy for cervical cancer, radiomics analysis of 3D dose distributions has been employed to predict toxicity rates, achieving AUCs ranging from 0.57 to 0.89 [13].

The differences between the statistical and DL methods can be described based on three factors: (i) model complexity, (ii) feature importance, and (iii) model transparency.

Statistical methods tend to exhibit relatively low complexity, and, compared with DL models, they operate based on clear governing principles. Statistical methods often demonstrate relatively low accuracy; however, they offer relatively high model transparency, which simplifies the interpretation of feature importance. DL models tend to outperform statistical methods. However, they are more complex [25–27] and considered "black boxes", which makes the interpretation of their results very challenging [28]. However, both accuracy and interpretability must be considered when choosing a method for predicting clinical outcomes [29–32].

Therefore, in the present study, we propose an interpretable DL model for predicting late bladder toxicity in patients with cervical cancer who have received definitive radiotherapy. We compared the performance of the statistical method and the DL model. In addition to achieving a high level of reliability, we conducted a feature-importance analysis and validated the performance of a lightweight DL model.

2. Materials and Methods

2.1. Patient Selection

We identified 545 patients with primary uterine cervical cancer who underwent definitive RT with curative intent at our institution between February 2006 and December 2017. Follow-up evaluations were performed every three months in the first two years, every four months in the third year, every six months in the fourth and fifth years, and annually thereafter. In this study, we included patients with more than three years of follow-up after treatment completion. The radiation-induced bladder toxicity was evaluated during the regular follow-up visits based on the European Organization for Research and Treatment of Cancer late radiation toxicity criteria, which represent a structured scoring schema developed by the Radiation Therapy Oncology Group (RTOG) [33].

In total, 281 patients with cervical cancer were included in this study (Figure 1). Clinical information regarding the status of the disease and treatment-related complications was collected retrospectively from patients' medical records. This study was approved by the Institutional Review Board (IRB) of the National Cancer Center, Korea: NCC2019-0166. This study adheres to the tenets of the Helsinki Declaration of 1975. The requirement for informed consent was waived owing to the retrospective nature of the study.

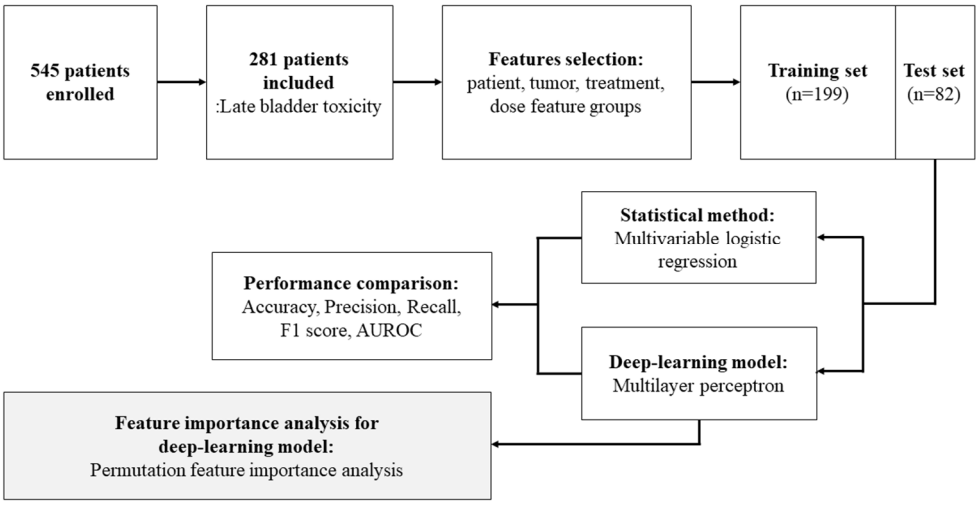

Figure 1. Flowchart of the proposed method for late bladder toxicity occurrence prediction using a deep learning model and feature importance analysis.

2.2. Treatment

The patients received concurrent chemoradiotherapy using either three-dimensional (3D) conformal EBRT or an intensity-modulated RT technique and high-dose-rate brachytherapy. The clinical target volume of EBRT included gross disease, the entire uterus, a margin

of 2.0 cm around the tumor extent at the level of the vagina, and the entire parametrial tissue. In addition, regional lymphatics, including the common, internal, and external iliac nodal regions and the presacral and para-aortic lymph nodes, were considered. Prior to 2008, brachytherapy planning was CT-based; afterwards, MRI-based planning has been employed [4]. Brachytherapy planning generally followed the recommendations of the Gynecologic Group European de Curie Therapie, European Society for Therapeutic Radiology, and Oncology Working Group [34,35].

The patients received a combination of 45–50 Gy of pelvic EBRT and 30 Gy of high-dose-rate brachytherapy with daily fractions of 5 Gy over 3 weeks. The biologically equivalent dose to point A in 2 Gy fractions was previously calculated as approximately 82 Gy for cervical tumors (α/β = 10) and 91 Gy for normal tissue (α/β = 3) [36]. Concomitant chemotherapy was administered using a weekly schedule of intravenous cisplatin at 40 mg per square meter of body surface area.

2.3. Data

We used four groups of the feature set as input data for the statistical analysis and DL method: (i) "patient" feature group (n = 2), including age and pathology; (ii) "tumor" feature group (n = 3), including the federation of gynecology and obstetrics (FIGO) stage, tumor-node-metastasis (TNM) category, and maximum tumor length on axial T2-weighted magnetic resonance image (MRI); (iii) "treatment" feature group (n = 4), including concurrent chemoradiotherapy (CCRT), CCRT regimen, number of CCRT cycles, and adjuvant chemotherapy; (iv) "dose" feature group with the equieffective dose (EQD2) at 3.0 Gy (n = 7), including the total dose of external beam radiation therapy (EBRT), the dose delivered to 100% of the primary GTV (GTV-D100), International Commission on Radiation Units and Measurements (ICRU) bladder point (BP_{ICRU}), $BD_{0.1cc}$, $BD_{1.0cc}$, $BD_{2.0cc}$, and $BD_{5.0cc}$. Features that were not continuous were dummy-coded. All features were deliberately collected in accordance with IRB guidelines and based on clinical experience and relevant literature [5,36–39]. These carefully selected features are known to play a significant role in influencing local control and survival outcomes in cervical cancer. With these selections, we aimed to facilitate meaningful comparisons with other studies that have similar diseases.

Late bladder toxicity was graded according to the scoring schema of the RTOG, with scores in the range of 0–5. The median durations from the initial start of EBRT to the occurrence of late toxicity are 37.7 months in the current study population. Owing to the imbalance of grades, the grade of late bladder toxicity was binarized by considering the occurrence of late bladder toxicity only. When late bladder toxicity is absent, the occurrence status is set to 0; otherwise, the occurrence status is set to 1.

2.4. Statistical Method

Patient characteristics were indicated in terms of mean ± standard deviation or median (min–max) for continuous variables; frequencies and percentage values were used for categorical variables. Continuous variables were compared using the *t*-test or Wilcoxon rank-sum test, whereas categorical variables were compared using the Chi-square or Fisher's exact tests. The dataset was randomly divided into two sets containing 199 and 82 cases for training and testing the statistical model, respectively. The ratios for the training and test sets were 70% and 30%, respectively. A statistical model was built using a training set with univariable and multivariable logistic regression. A univariable logistic regression was performed on the input data, considering patient, tumor, treatment, and dose features. The results of the univariable logistic regression were used as the input for the multivariable logistic regression. The target was the occurrence of binarized late toxicity. The developed multivariable logistic regression model was tested using the test dataset. Statistical analyses were performed using SAS version 9.4 (SAS Institute Inc., Cary, NC, USA) and R version 4.2.1 (R Foundation for Statistical Computing, Vienna, Austria). We assumed that statistical significance is achieved for *p*-values below 0.05.

2.5. Deep Learning Model for Permutation Feature Analysis

To implement the DL model by using Pytorch version 1.12.1 (Warsaw, Mazowieckie, Poland), we used an MLP. MLP is a type of artificial neural network that consists of multiple layers of nodes with batch normalization and nonlinearity functions such as rectified linear unit (RELU), leaky RELU, and sigmoid. Each node in one layer is connected to all nodes in the next layer, forming a fully connected network. The input layer takes the input data, and the output layer produces the final predictions. The layers between the input and output layers are called hidden layers. The DL model was built using six hidden layers. The architecture can be represented mathematically as follows:

$$h_1 = f_1(W_1 x + b_1), \tag{1}$$

$$h_2 = f_2(W_2 h_1 + b_2), \tag{2}$$

$$h_3 = f_3(W_3 h_2 + b_3), \tag{3}$$

$$h_4 = f_4(W_4 h_3 + b_4), \tag{4}$$

$$h_5 = f_5(W_5 h_4 + b_5), \tag{5}$$

$$h_6 = f_6(W_6 h_5 + b_6), \tag{6}$$

and

$$y = f_7(W_7 h_6 + b_7), \tag{7}$$

where x is the input data, y is the output data, W_i and b_i are the weight matrix and bias vector for the i-th layer, respectively, f_i is the activation function for the i-th layer, and h_i is the output of the i-th hidden layer.

As a preprocessing step, 16 features were used as input data, and binarized late bladder toxicity values were considered the output data. Z-score normalization was applied to the input data so that the DL model could rapidly converge to the optimal solution. The normalization parameters (mean and standard deviation) were determined from the training set.

The standardized input data were passed through an input layer, and output data were obtained on the output layer. In each hidden layer, except for the last one, batch normalization and dropout techniques were applied to avoid overfitting the training data. The probability of a node being zeroed for dropout was set to 0.2. In addition, a leaky ReLU was adopted as the activation function in the hidden layers to improve model complexity and performance. A sigmoid function was used as the activation function of the last layer only. These functions introduce nonlinearity into the model, allowing it to capture complex relationships within the data. The detailed architecture of the proposed model is presented in Table 1.

The loss between the ground truth and predicted toxicity occurrence was calculated using binary cross-entropy as follows:

$$Binary\ cross\ entropy\ loss = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(p_i) + (1 - y_i)\log(1 - p_i)), \tag{8}$$

where N represents the total number of samples in the batch, y_i represents the true label of the i-th sample (either 0 or 1), and p_i represents the predicted probability (between 0 and 1) of the i-th sample belonging to the positive class. The DL models were trained to minimize the loss.

Table 1. The architecture of the late bladder toxicity prediction model.

Layer Type	In Features	Out Features	Bias	Batch Normalization	Activation Function
Fully connected layer	16	36	FALSE	TRUE	* Leaky ReLU
Fully connected layer	36	72	FALSE	TRUE	Leaky ReLU
Fully connected layer	72	144	FALSE	TRUE	Leaky ReLU
Fully connected layer	144	72	FALSE	TRUE	Leaky ReLU
Fully connected layer	72	36	FALSE	TRUE	Leaky ReLU
Fully connected layer	36	1	FALSE	TRUE	Sigmoid

* Leaky ReLU: Leaky rectified linear unit.

To avoid overfitting, the synthetic minority oversampling technique (SMOTE) was adopted during the training procedure as oversampling technique and augmentation strategy, which adds random noise to the input data in every training epoch [18]. Moreover, we employed an adaptive momentum estimation optimizer with a learning rate of 0.005 and a weight decay of 0.0001. To compute the running average of the gradient, β1 and β2 were set to 0.9 and 0.999, respectively [40].

K-fold cross-validation (CV) was employed to ensure the generalizability of the DL model. The value of k was set to 5, resulting in the training set being divided into five folds. Furthermore, an early stopping strategy was implemented, where the model achieving the lowest validation loss during the 500 epochs was saved as the best-performing model. Each fold yielded independent results as separate models were trained. The final output data were determined using the voting method, with a threshold of 3 sets to obtain the ultimate decision.

2.6. Permutation Feature Importance Analysis

The permutation analysis is used to calculate the feature importance of a DL model. It involves randomly permuting the value of a single feature and evaluating the model's performance on the test dataset [41]. The analysis was performed using Scikit-learn version 1.0.2.

To measure the importance of each feature, the mean squared error (MSE) is calculated between the reference dataset and a corrupted dataset, where the value of a single feature has been randomly permuted. This is compared with the reference MSE, which is the MSE calculated on the uncorrupted reference dataset.

The variation in model performance resulting from the permutation of a feature is used to compute the importance of each feature.

$$i_j = s - \frac{1}{K}\sum_{k=1}^{K} s_{k,j} \tag{9}$$

where i is used to index the features, j represents the total number of features, i_j is the importance of j-th feature, K is the number of repetitions or permutations, S is the reference MSE, $s_{k,j}$ is the MSE calculated using a corrupted dataset for the k-th repetition, and $\frac{1}{K}\sum_{k=1}^{K} s_{k,j}$ the average MSE over K repetitions for j-th feature.

The value of K was set to 500, indicating that the permutation analysis was repeated 500 times. Finally, the permutation feature importance was calculated by averaging the feature importance values obtained from DL models trained using different training folds.

2.7. Lightweight Deep Learning Model

In order to develop a lightweight deep learning model, we utilized the importance values derived from permutation feature analysis and selected the top 5 features as input. This approach enabled us to focus on the most influential features while reducing the computational complexity of the model.

When designing the hidden layers, we adhered to the commonly employed standards of multi-layer perceptron (MLP) models. Specifically, our model was designed with a

decreasing number of neurons in each hidden layer, starting from the number of input features and gradually reducing to 5, 4, 3, 2, and finally 1 neuron.

The activation functions, preprocessing steps, loss function, optimization technique, cross-validation, and final decision-making process of the lightweight deep learning model remain the same as those of the deep learning model described previously.

2.8. Performance Comparison

To compare the performance of the DL model and multivariable logistic regression, four metrics were adopted: accuracy (Equation (10)), precision (Equation (11)), recall (Equation (12)), F1-score (Equation (13)), and the area under the receiver operating characteristic curve (AUROC) (Equation (14)). To compute these metrics, we counted the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) in the test set. Precision is defined as the number of positive predictions (i.e., those that were correctly identified by the model) expressed as a fraction of the total number of predictions. Recall indicates the fraction of positive instances that the model could identify. The F1-score is the harmonic mean of precision and recall. Therefore, a high F1-score indicates that the model exhibits a good balance between precision and recall. If a model performs well, it has a high AUROC value (close to 1.0).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{10}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{11}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{12}$$

$$\text{F1} - \text{score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{13}$$

$$\text{AUROC} = \int_0^1 \text{sensitivity}(x) \cdot \text{d}(1 - \text{specificity}(x)) \tag{14}$$

3. Results

3.1. General Study Characteristics

Patients' characteristics are summarized in Table 2. No significant differences were observed between the training and test sets in terms of baseline characteristics, such as age, pathology, FIGO stage, TNM category, maximum tumor length, CCRT, CCRT regimen, number of CCRT cycles, adjuvant chemotherapy, and the total dose of EBRT, GTV-D100, BP_{ICRU}, $BD_{0.1cc}$, $BD_{1.0cc}$, $BD_{2.0cc}$, and $BD_{5.0cc}$.

3.2. Logistic Regression Results

Univariable and multivariable logistic regression analyses were performed to identify risk predictors for adverse events in the training set (Table 3). Based on the univariable analysis, no significant features were observed in the features set. The multivariable logistic regression model returned four features with $p < 0.2$: CCRT cycle, GTV-D100, $BD_{2.0cc}$, and $BD_{5.0cc}$. The prediction metrics were calculated using a multivariable logistic regression model on the test data. Multivariable logistic regression achieved accuracy, precision, recall, F1-score, and AUROC of 0.85, 0.08, 0.5, 0.14, and 0.43, respectively. Although multivariable logistic regression yielded an accuracy of 0.85, the regression model did not exhibit acceptable performance on the test dataset.

Table 2. Baseline patients' characteristics (n = 281).

Variable			Total (N = 281)		Training Set (N = 199)		Test Set (N = 82)		*p*-Value
AGE		mean ± std	62.1	±14.0	63.4	±13.4	61.5	±14.2	0.2920 (3)
Pathology	1;	squamous	224	(79.7%)	63	(76.8%)	161	(80.9%)	0.4708 (1)
	2;	adenoca	23	(8.2%)	6	(7.3%)	17	(8.5%)	
	3;	adenosquamous	10	(3.6%)	5	(6.1%)	5	(2.5%)	
	4;	Other	24	(8.5%)	8	(9.8%)	16	(8.0%)	
FIGO stage	1;	Ia1, Ia2, Ib1, Ib2	53	(18.9%)	17	(20.7%)	36	(18.1%)	0.7285 (1)
	2;	IIa1, IIa2, Iib	161	(57.3%)	44	(53.7%)	117	(58.8%)	
	3;	IIIa, IIIb, Iva, Ivb	67	(23.8%)	21	(25.6%)	46	(23.1%)	
TNM stage	1;	T1a1, T1a2, T1b1, T1b2	56	(19.9%)	17	(20.7%)	39	(19.6%)	0.9752 (1)
	2;	T2a1, T2a2, T2b	184	(65.5%)	53	(64.6%)	131	(65.8%)	
	3;	T3a, T3b, T4	41	(14.6%)	12	(14.6%)	29	(14.6%)	
CCRT	0;	RT alone	22	(7.8%)	4	(4.9%)	18	(9.0%)	0.2372 (1)
	1;	CCRT	259	(92.2%)	78	(95.1%)	181	(91.0%)	
Concurrent chemotherapy regimen (CCRT 259 case)	1;	cisplatin	231	(89.2%)	72	(92.3%)	159	(87.8%)	0.7583 (2)
	2;	5FU + cisplatin	3	(1.2%)	0	(0.0%)	3	(1.7%)	
	3;	carboplatin	19	(7.3%)	5	(6.4%)	14	(7.7%)	
	4;	other	6	(2.3%)	1	(1.3%)	5	(2.8%)	
Number of concurrent chemotherapy cycle (CCRT 259 case)	0;	Cycle 3 or less	26	(10.0%)	7	(9.0%)	19	(10.5%)	0.7083 (1)
	1;	Cycle 3 or more	233	(90.0%)	71	(91.0%)	162	(89.5%)	
Adjuvant chemotherapy	0;	No	260	(92.5%)	79	(96.3%)	181	(91.0%)	0.1185 (1)
	1;	Yes	21	(7.5%)	3	(3.7%)	18	(9.0%)	
Tumor size (cm) (MRI axial)		median (min–max)	4.2	(1.3–10)	4.3	(2.3–8.5)	4.2	(1.3–10)	0.6955 (3)
EBRT total dose EQD2(3) (Gy)		median (min–max)	48.4	(20–73.1)	48.4	(43.2–68.3)	48.4	(20–73.1)	0.2765 (3)
GTV D100 (cGy)		median (min–max)	570.5	(112.3–1336)	552.4	(194–1187.8)	584.2	(112.3–1336)	0.1644 (3)
BPICRU EQD2(3) (Gy)		median (min–max)	23.5	(0–93.2)	26.9	(0–93.2)	22.5	(0–90.7)	0.1775 (3)
BD0.1cc EQD2(3) (Gy)		median (min–max)	58.3	(12.6–202.4)	59.7	(25–174)	57.6	(12.6–202.4)	0.2659 (3)
BD1cc EQD2(3) (Gy)		median (min–max)	46.1	(10–141.7)	48.9	(20.5–111.1)	45.6	(10–141.7)	0.1906 (3)
BD2cc EQD2(3) (Gy)		median (min–max)	41.3	(6.3–120.5)	43.6	(9.8–97.8)	39.8	(6.3–120.5)	0.2492 (3)
BD5cc EQD2(3) (Gy)		median (min–max)	33.5	(1.3–91.5)	35.2	(3.8–78.9)	33	(1.3–91.5)	0.1944 (3)

(1) Chi-square test. (2) Fisher's exact test. (3) *t*-test.

Table 3. Results of univariable and multivariable logistic regression analysis for late bladder toxicity prediction.

Variable			Univariable Analysis OR (95% CI)	*p*-Value	Multivariable Analysis OR (95% CI)	*p*-Value
AGE			0.997 (0.975–1.019)	0.803		
Pathology	1;	squamous	1 (ref)			
	2;	adenoca	0.769 (0.238–2.483)	0.661		
	3;	adenosquamous	1.667 (0.270–10.303)	0.583		
	4;	Other	0.357 (0.078–1.634)	0.185		
FIGO stage	1;	Ia1, Ia2, Ib1, Ib2	1 (ref)			
	2;	IIa1, IIa2, Iib	0.937 (0.406–2.164)	0.879		
	3;	IIIa, IIIb, Iva, Ivb	1.024 (0.388–2.706)	0.962		
TNM stage	1;	T1a1, T1a2, T1b1, T1b2	1 (ref)			
	2;	T2a1, T2a2, T2b	0.820 (0.375–1.794)	0.620		
	3;	T3a, T3b, T4	0.716 (0.241–2.127)	0.548		
CCRT	0;	RT alone	1 (ref)			
	1;	CCRT	1.336 (0.42–4.252)	0.624		

Table 3. *Cont.*

Variable			Univariable Analysis		Multivariable Analysis	
			OR (95% CI)	*p*-Value	OR (95% CI)	*p*-Value
Concurrent chemotherapy regimen (CCRT 181 case)	1;	cisplatin	1 (ref)			
	2;	5FU + cisplatin	5.066 (0.448–57.267)	0.190		
	3;	carboplatin	0.422 (0.091–1.962)	0.271		
	4;	other	0.633 (0.069–5.821)	0.687		
Concurrent chemotherapy cycle (CCRT 181 case)	0;	Cycle 3 or less	1 (ref)		1 (ref)	
	1;	Cycle 3 or more	0.481 (0.181–1.277)	0.142	0.440 (0.162–1.194)	0.107
Adjuvant chemotherapy	0;	No	1 (ref)			
	1;	Yes	1.385 (0.493–3.897)	0.537		
Tumor size (cm) (MRI axial)			1.011 (0.828–1.234)	0.918		
EBRT total dose EQD2(3) (Gy)			1.006 (0.951–1.065)	0.831		
GTV D100 (cGy)			1.001 (0.999–1.003)	0.186	1.002 (1.000–1.004)	0.136
BPICRU EQD2(3) (Gy)			1.005 (0.985–1.025)	0.651		
BD0.1cc EQD2(3) (Gy)			1.008 (0.995–1.020)	0.232		
BD1cc EQD2(3) (Gy)			1.012 (0.993–1.031)	0.213		
BD2cc EQD2(3) (Gy)			1.016 (0.996–1.036)	0.114	1.047 (0.939–1.168)	0.406
BD5cc EQD2(3) (Gy)			1.018 (0.993–1.043)	0.163	0.961 (0.841–1.100)	0.566

3.3. Permutation Feature Importance: Permutation Analysis

Permutation feature analysis determines the importance of features by quantifying the extent to which the performance of a trained model changes when the dataset is permuted (Equation (7)).

An independent permutation feature importance analysis was conducted on the DL model, separately for each training fold. To determine the overall feature importance, the mean and standard deviation of the calculated permutation feature importance were computed by averaging the importance values across folds. Figure 2 illustrates the mean and standard deviation of the permutation feature importance for all features.

Our findings reveal that the feature with the highest importance was BD_{2cc}, followed by BD_{5cc} and BP_{ICRU}. Additionally, the features with the highest importance within each group were age, TNM category, number of CCRT cycles, and BD_{2cc} for the patient, tumor, treatment, and dose feature groups, respectively.

We identified the top five features with high importance as BD_{2cc}, BD_{5cc}, BP_{ICRU}, TNM category, and tumor size. We made the decision to select tumor size instead of FIGO stage as a feature, considering the possibility of redundancy between FIGO stage and TNM category. This decision was based on the scientific rationale that tumor size provides valuable and distinct information for the prediction model [37].

3.4. Deep Learning Models

The prediction performance of the DL model for late bladder toxicity was evaluated using the voting method. Table 4 summarizes the prediction performances of the DL models and the lightweight DL models trained using the five different training folds. The means and standard deviations of prediction performance for accuracy, precision, recall, and F1-score differed between the deep learning model and the lightweight deep learning model. For the deep learning model, the values were as follows: accuracy (0.77 ± 0.06), precision (0.68 ± 0.10), recall (0.41 ± 0.12), and F1-score (0.49 ± 0.05). On the other hand, the lightweight deep learning model had different values: accuracy (0.92 ± 0.03), precision (0.97 ± 0.03), recall (0.86 ± 0.07), and F1-score (0.90 ± 0.04). In the case of the AUROC, the deep learning model achieved a value of 0.81 ± 0.04, while the lightweight deep learning model achieved a value of 0.94 ± 0.03 (Figure 3).

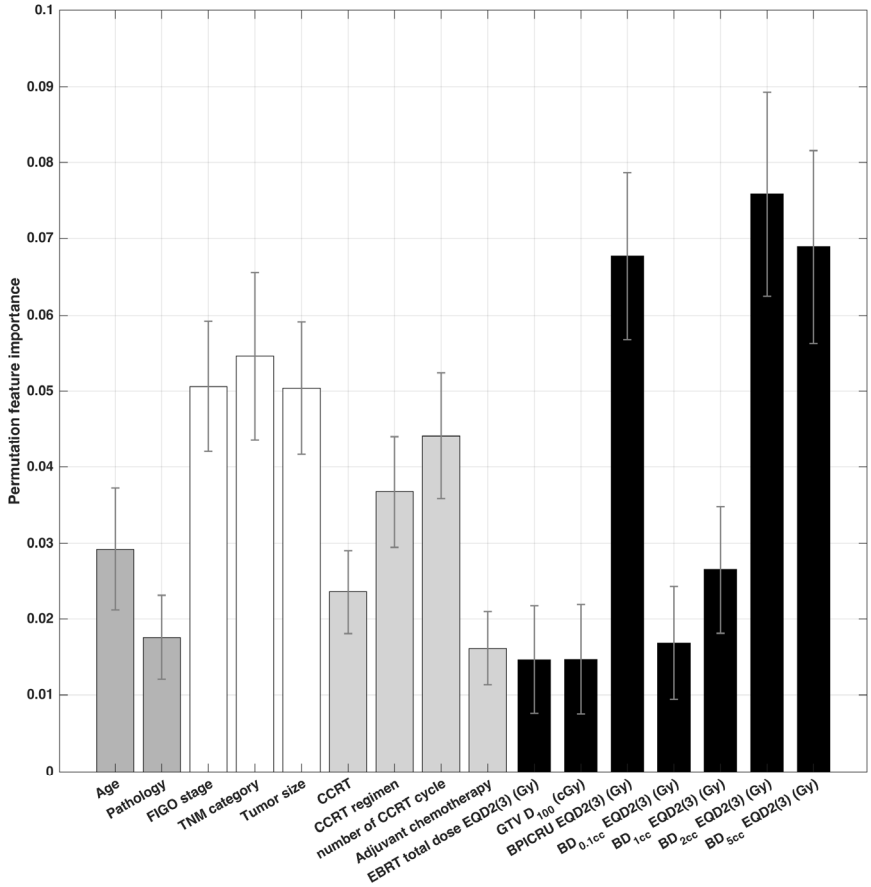

Figure 2. Permutation feature importance computed from the deep learning models.

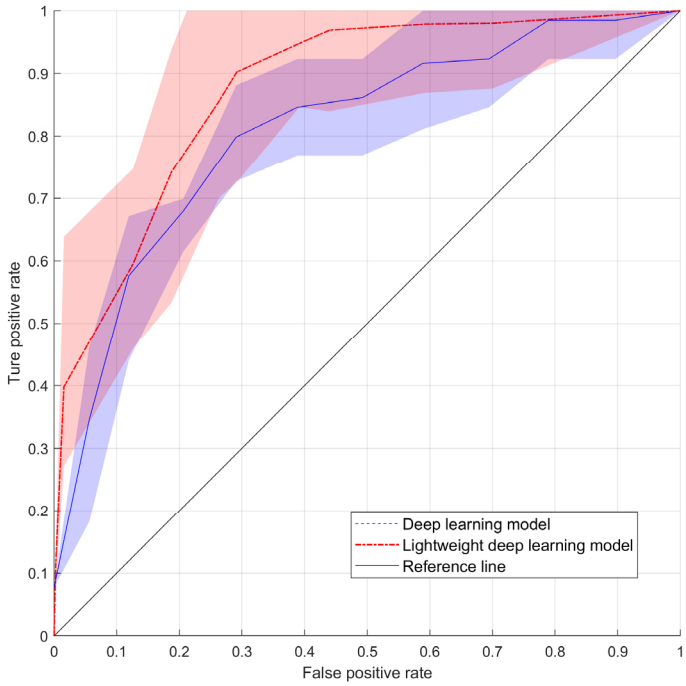

Figure 3. The area under the receiver operating curve comparison across 5-fold cross-validation with the reference line.

Table 4. Performance comparison for predicting late bladder toxicity of a statistical model and DL models by five different folds for the test data.

Model	Accuracy	Precision	Recall	F1-Score	* AUROC
Statistical model	0.85	0.08	0.5	0.14	0.43
Deep learning model: fold 1	0.78	0.62	0.38	0.47	0.76
Deep learning model: fold 2	0.73	0.85	0.35	0.50	0.86
Deep learning model: fold 3	0.76	0.69	0.36	0.47	0.84
Deep learning model: fold 4	0.70	0.69	0.30	0.42	0.83
Deep learning model: fold 5	0.88	0.54	0.64	0.58	0.77
Deep-learning model: voting method (threshold = 3)	0.91	0.85	0.69	0.76	0.81
Leightweight deep learning model: fold 1	0.93	0.99	0.83	0.92	0.93
Leightweight deep learning model: fold 2	0.88	0.94	0.81	0.87	0.88
Leightweight deep learning model: fold 3	0.90	0.99	0.81	0.89	0.90
Leightweight deep learning model: fold 4	0.97	0.94	0.99	0.98	0.97
Leightweight deep learning model: fold 5	0.91	0.98	0.85	0.86	0.89
Deep-learning model: voting method (threshold = 3)	0.93	0.94	0.88	0.90	0.91

* AUROC: area under the receiver operating characteristic curve.

Finally, both the DL model and the lightweight deep learning model were evaluated using the voting method with a threshold of 3. For the DL model, the evaluation metrics were as follows: Accuracy (0.91), Precision (0.85), Recall (0.69), F1-score (0.76), and AUROC (0.81). For the lightweight deep learning model, the evaluation metrics were as follows: Accuracy (0.93), Precision (0.94), Recall (0.88), F1-score (0.90), and AUROC (0.91).

However, because the AUROC could not be calculated from the labels predicted by each of the five models, the average of the AUROCs of all the models was considered as the overall AUROC value.

4. Discussion

This study aimed to (1) compare the ability of multivariable logistic regression and DL models to identify those patients who, having received radiation therapy, are at risk of bladder radiation toxicity, and (2) interpret the results of DL models to understand the significance of input features. To the best of our knowledge, no previous study has attempted an interpretation to ensure the reliability of a DL model in predicting the occurrence of late bladder toxicity.

Several statistical methods and DL models are available for predicting clinical outcomes, including radiation toxicity. In many instances, DL models can effectively find near-optimal solutions for nonconvex optimization problems using gradient methods and nonlinear activation functions. However, statistical methods can also achieve high accuracy and reliability in specific cases. Despite the rapid development of DL models and their relatively high performance, their clinical utility is still a topic of controversy due to concerns about their reliability, particularly in terms of how clinicians interpret the results and features.

In this study, we utilized an MLP with relatively low model complexity to enhance the interpretability of deep learning models. The performance of the MLP relies on the configuration of the hidden layer, and Muhammad et al. [42] suggested that the optimal configuration for the hidden layer is three. Furthermore, experimental findings demonstrated that having fewer than three hidden layers directly impacts the network's accuracy, while having more than three hidden layers increases the time complexity without a proportional improvement in accuracy. Determining the appropriate number of neurons in the hidden layer is still a topic of debate, and the design of the model structure should be customized to the specific problem and available computational resources while also considering the general MLP model design standard: a structure in which the number of neurons in the hidden layer continuously decreases. Therefore, we selected the top five

important features based on the results of the permutation feature importance analysis and designed the lightweight DL model, taking into account the general design standard. This approach not only improves interpretability but also offers advantages associated with the general design standard of MLP models. For instance, the continuous reduction of neurons in the hidden layer helps prevent overfitting and allows for efficient learning of complex patterns while avoiding excessive computational complexity.

Considering the experiment on our dataset, the DL model proved superior compared with the statistical model. To address the class-imbalance problem in the DL method, we adopted the SMOTE oversampling technique. Mylona et al. [18] suggested that variations in oversampling techniques, including SMOTE, increase the prediction performance of classifiers. Accordingly, our results showed that when SMOTE was not applied to the training process, the mean and standard deviation of the AUROC value on the test set were reduced to 0.52 ± 0.13.

The precision and recall of the DL model exhibited a tradeoff relationship depending on the value of the threshold used for voting, as shown in Figure 4. For the prediction of radiation-induced bladder toxicity, both precision and recall are crucial metrics for assessing the performance of a prediction model because they reflect the costs and benefits associated with false positives and false negatives. High precision is essential to minimize unnecessary interventions and potential harm for patients who are not at risk. However, a high recall is necessary to ensure that all patients at risk of developing bladder toxicity are detected and appropriate measures are taken to mitigate the associated risks. Therefore, achieving a balance between precision and recall is necessary to optimize the performance of the prediction model for clinical applications. At a threshold of 3, the precision and recall of the voting methods are indicated by markers placed inside circles denoted by asterisks.

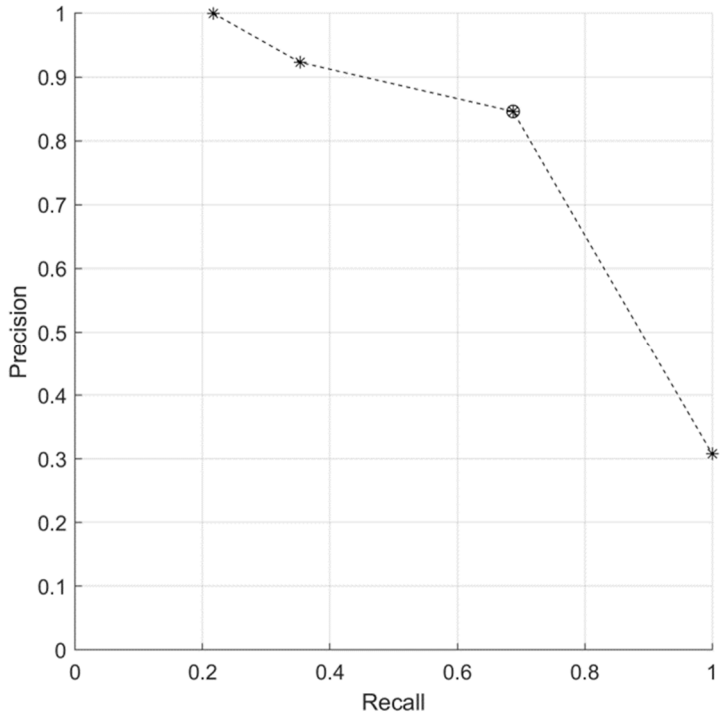

Figure 4. Precision–recall curve depending on the voting method threshold value. The asterisk in a circle indicate a threshold of 3.

Permutation feature importance is a model-agnostic technique that provides a comprehensive and computationally efficient assessment of feature importance by considering feature interactions and avoiding the bias caused by collinear or redundant features. In this study, permutation feature importance is utilized to calculate the relative importance of features while also employing a feature reduction technique like principal component

analysis [42–44]. The importance of permutation features can be sensitive to the choice of the metric used to evaluate the model's performance (see Equation (7)), leading to different feature rankings. Our findings emphasize the importance of specific features, such as BD_{2cc}, BD_{5cc}, and BP_{ICRU} for the accurate prediction of late bladder toxicity occurrence, which is also consistent with the prognostic factor for urinary toxicity suggested in previous studies, such as BD_{2cc} in [5]. Our results provide insights that could facilitate the development of more effective and personalized treatment strategies for patients undergoing radiation therapy.

Our study has certain limitations that need to be considered. Firstly, our analysis was based solely on structured data (patient and treatment dose features), which may not fully capture the complexity of bladder toxicity. Recent studies have demonstrated the potential of more sophisticated DL models that incorporate 3D dose distributions, medical images, contours, etc. to predict clinical outcomes [45,46]. Radiomics, which involves extracting quantitative features from medical images, can provide additional information about tumor characteristics and treatment response. Integrating radiomics data alongside patient data and dose volumetric parameters has shown improvements in the prediction of various types of toxicities, with potential enhancements in AUC values ranging from 0.11 to 0.16 [13]. Therefore, in future research, we aim to explore the integration of radiomics data into our DL models to enhance the prediction of late bladder toxicity. By incorporating this additional information, we anticipate improved performance and a better understanding of the underlying factors contributing to toxicity with advanced interpretation techniques.

Secondly, although the DL model outperformed the statistical model in our dataset, it is important to note that the performance of DL models can vary depending on the specific dataset and problem domain. Further validation using larger and more diverse datasets through multi-institutional studies is required to confirm the generalizability of our findings.

Thirdly, interpreting DL models remains a challenging task, especially concerning feature importance. Permutation feature analysis is a useful technique for assessing the importance of features in DL models. However, it should be noted that this method has limitations. It tends to assign higher importance to continuous variables and can produce different feature rankings depending on the choice of evaluation metric. Additionally, applying this technique to 3D input is challenging. While permutation feature analysis provides valuable insights, it needs to be supplemented with other interpretive techniques to gain a more comprehensive understanding of the behavior and functional importance of DL models. Therefore, as a further study, it is necessary to apply various analysis methods, such as LIME and its variants, input gradient-based methods, CAM and its variants, etc., to the DL model to ensure the reliability of models with relatively higher performance [47,48].

Furthermore, in order to address the limitation of a small sample size, we established an extra-validation set consisting of 17 individuals. Utilizing a lightweight DL model, we conducted predictions for late bladder toxicity. The results revealed an accuracy of 0.81%, a precision of 0.99%, a recall of 0.61%, an F1-score of 0.92, and an AUROC of 0.93. These performance metrics indicate that the lightweight DL model had limited accuracy in forecasting late bladder toxicity. While the precision was high, indicating few false positives, the recall was relatively low, meaning it missed many true positive cases.

It is crucial to exercise caution when interpreting these findings due to the relatively small size of the extra-validation set. The restricted sample may not adequately encompass the full range of late bladder toxicity in radiation therapy. For instance, the longest observed duration of late bladder toxicity in our institution was 107.7 months from the initiation of external beam radiation therapy (EBRT) to the occurrence of late toxicity. As a result, the performance metrics on the extra-validation set obtained may not truly reflect the predictive capabilities of the DL model. Thus, future research should strive to incorporate a larger patient cohort from multiple institutions to validate and enhance the predictive capabilities of DL models in accurately anticipating late bladder toxicity.

Overall, this study contributes to the ongoing discussion on the clinical utility of DL models in predicting radiation toxicity and emphasizes the importance of interpretability to enhance the reliability and practical applicability of these models. By addressing the limitations and conducting further research, we can advance the field and ultimately improve patient outcomes in radiation therapy.

5. Conclusions

In this study, we compared the performance of logistic regression and DL models for the prediction of late bladder toxicity in patients with cervical cancer. Logistic regression did not show acceptable performances, whereas the lightweight DL model achieved an accuracy of 0.92 ± 0.03 and an AUROC of 0.91 ± 0.03. Moreover, the permutation feature importance analysis identified BD_{2cc} as the most important feature for risk prediction.

Author Contributions: Conceptualization, W.C. and H.K.; methodology, W.C. and H.K.; software, W.C.; validation, E.S.O., S.U.L. and J.Y.K.; formal analysis, M.H.; investigation, S.J., J.H.J., Y.K.L., D.S. and S.B.L.; resources, E.S.O., S.U.L. and J.Y.K.; data curation, H.K. and J.Y.K.; writing—original draft preparation, W.C.; writing—review and editing, W.C. and H.K.; visualization, W.C.; supervision, H.K.; project administration, H.K.; funding acquisition, H.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Research Foundation of the Korean National Cancer Center Fund (grant number 2110610-3).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of National cancer center, Republic of Korea. (Protocol code: NCC2019-0166: and date of approval: 7 September 2019).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Some or all datasets generated during and/or analyzed during the current study are not publicly available but are available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chihikara, B.S.; Parang, K. Global cancer statistics 2022: The trends projection analysis. *Chem. Biol. Lett.* **2023**, *10*, 451.
2. Eifel, P.J.; Winter, K.; Morris, M.; Levenback, C.; Grigsby, P.W.; Cooper, J.; Rotman, M.; Gershenson, D.; Mutch, D.G. Pelvic irradiation with concurrent chemotherapy versus pelvic and para-aortic irradiation for high-risk cervical cancer: An update of Radiation Therapy Oncology Group trial (rtog) 90-01. *J. Clin. Oncol.* **2004**, *22*, 872–880. [CrossRef] [PubMed]
3. Collaboration CfCCM-A. Reducing uncertainties about the effects of chemoradiotherapy for cervical cancer: A systematic review and meta-analysis of individual patient data from 18 randomized trials. *J. Clin. Oncol.* **2008**, *26*, 5802–5812. [CrossRef] [PubMed]
4. Sung Uk, L.; Young Ae, K.; Young-Ho, Y.; Yeon-Joo, K.; Myong Cheol, L.; Sang-Yoon, P.; Sang-Soo, S.; Ji Eun, P.; Joo-Young, K. General health status of long-term cervical cancer survivors after radiotherapy. *Strahlenther. Onkol.* **2017**, *193*, 543–551. [CrossRef]
5. Manea, E.; Escande, A.; Bockel, S.; Khettab, M.; Dumas, I.; Lazarescu, I.; Fumagalli, I.; Morice, P.; Deutsch, E.; Haie-Meder, C.; et al. Risk of late urinary complications following image guided adaptive brachytherapy for locally advanced cervical cancer: Refining bladder dose-volume parameters. *Int. J. Radiat. Oncol. Biol. Phys.* **2018**, *101*, 411–420. [CrossRef]
6. Catucci, F.; Alitto, A.R.; Masciocchi, C.; Dinapoli, N.; Gatta, R.; Martino, A.; Mazzarella, C.; Fionda, B.; Frascino, V.; Piras, A.; et al. Predicting radiotherapy impact on late bladder toxicity in prostate cancer patients: An observational study. *Cancers* **2021**, *13*, 175. [CrossRef]
7. Carillo, V.; Cozzarini, C.; Rancati, T.; Avuzzi, B.; Botti, A.; Borca, V.C.; Cattari, G.; Civardi, F.; Esposti, C.D.; Franco, P.; et al. Relationships between bladder dose–volume/surface histograms and acute urinary toxicity after radiotherapy for prostate cancer. *Radiother. Oncol.* **2014**, *111*, 100–105. [CrossRef]
8. Kim, Y.; Kim, Y.J.; Kim, J.Y.; Lim, Y.K.; Jeong, C.; Jeong, J.; Kim, M.; Lim, M.C.; Seo, S.S.; Park, S.Y. Toxicities and dose–volume histogram parameters of mri-based brachytherapy for cervical cancer. *Brachytherapy* **2017**, *16*, 116–125. [CrossRef]
9. Kim, Y.J.; Kim, J.Y.; Kim, T.H.; Lim, Y.K.; Yoon, M.G.; Joo, J.N.; Park, S.Y. Dosimetric evaluation of magnetic resonance imaging-based intracavitary brachytherapy for cervical cancer. *Technol. Cancer Res. Treat.* **2014**, *13*, 243–251. [CrossRef]
10. Thor, M.; Bentzen, L.; Hysing, L.B.; Ekanger, C.; Helle, S.I.; Karlsdóttir, Á.; Muren, L.P. Prediction of rectum and bladder morbidity following radiotherapy of prostate cancer based on motion-inclusive dose distributions. *Radiother. Oncol.* **2013**, *107*, 147–152. [CrossRef]

11. Ospina, J.D.; Zhu, J.; Chira, C.; Bossi, A.; Delobel, J.B.; Beckendorf, V.; Dubray, B.; Lagrange, J.L.; Correa, J.C.; Simon, A.; et al. Random forests to predict rectal toxicity following prostate cancer radiation therapy. *Int. J. Radiat. Oncol. Biol. Phys.* **2014**, *89*, 1024–1031. [CrossRef] [PubMed]
12. Acosta, O.; Drean, G.; Ospina, J.D.; Simon, A.; Haigron, P.; Lafond, C.; de Crevoisier, R. Voxel-based population analysis for correlating local dose and rectal toxicity in prostate cancer radiotherapy. *Phys. Med. Biol.* **2013**, *58*, 2581–2595. [CrossRef] [PubMed]
13. Lucia, F.; Bourbonne, V.; Visvikis, D.; Miranda, O.; Gujral, D.M.; Gouders, D.; Dissaux, G.; Pradier, O.; Tixier, F.; Jaouen, V.; et al. Radiomics analysis of 3d dose distributions to predict toxicity of radiotherapy for cervical cancer. *J. Pers. Med.* **2021**, *11*, 398. [CrossRef]
14. Zhen, X.; Chen, J.; Zhong, Z.; Hrycushko, B.; Zhou, L.; Jiang, S.; Albuquerque, K.; Gu, X. Deep convolutional neural network with transfer learning for rectum toxicity prediction in cervical cancer radiotherapy: A feasibility study. *Phys. Med. Biol.* **2017**, *62*, 8246–8263. [CrossRef] [PubMed]
15. Chen, J.; Chen, H.; Zhong, Z.; Wang, Z.; Hrycushko, B.; Zhou, L.; Jiang, S.; Albuquerque, K.; Gu, X.; Zhen, X. Investigating rectal toxicity associated dosimetric features with deformable accumulated rectal surface dose maps for cervical cancer radiotherapy. *Radiat. Oncol.* **2018**, *13*, 125. [CrossRef]
16. Improta, I.; Palorini, F.; Cozzarini, C.; Rancati, T.; Avuzzi, B.; Franco, P.; Degli Esposti, C.; Del Mastro, E.; Girelli, G.; Iotti, C.; et al. Bladder spatial-dose descriptors correlate with acute urinary toxicity after radiation therapy for prostate cancer. *Phys. Med.* **2016**, *32*, 1681–1689. [CrossRef]
17. Mylona, E.; Acosta, O.; Lizee, T.; Lafond, C.; Crehange, G.; Magné, N.; Chiavassa, S.; Supiot, S.; Ospina Arango, J.D.; Campillo-Gimenez, B.; et al. Voxel-based analysis for identification of urethrovesical subregions predicting urinary toxicity after prostate cancer radiation therapy. *Int. J. Radiat. Oncol. Biol. Phys.* **2019**, *104*, 343–354. [CrossRef]
18. Mylona, E.; Lebreton, C.; Fontaine, P.; Supiot, S.; Magne, N.; Crehange, G.; de Crevoisier, R.; Acosta, O. Comparison of machine learning algorithms and oversampling techniques for urinary toxicity prediction after prostate cancer radiotherapy. In Proceedings of the 19th International Conference on Bioinformatics and Bioengineering (BIBE), Athens, Greece, 28–30 October 2019; IEEE Publications. IEEE: Piscataway, NJ, USA, 2019; Volume 2019, pp. 964–971. [CrossRef]
19. Mylona, E.; Cicchetti, A.; Rancati, T.; Palorini, F.; Fiorino, C.; Supiot, S.; Magne, N.; Crehange, G.; Valdagni, R.; Acosta, O.; et al. Local dose analysis to predict acute and late urinary toxicities after prostate cancer radiotherapy: Assessment of cohort and method effects. *Radiother. Oncol.* **2020**, *147*, 40–49. [CrossRef]
20. Hathout, L.; Folkert, M.R.; Kollmeier, M.A.; Yamada, Y.; Cohen, G.N.; Zelefsky, M.J. Dose to the bladder neck is the most important predictor for acute and late toxicity after low-dose-rate prostate brachytherapy: Implications for establishing new dose constraints for treatment planning. *Int. J. Radiat. Oncol. Biol. Phys.* **2014**, *90*, 312–319. [CrossRef]
21. Pella, A.; Cambria, R.; Riboldi, M.; Jereczek-Fossa, B.A.; Fodor, C.; Zerini, D.; Torshabi, A.E.; Cattani, F.; Garibaldi, C.; Pedroli, G.; et al. Use of machine learning methods for prediction of acute toxicity in organs at risk following prostate radiotherapy. *Med. Phys.* **2011**, *38*, 2859–2867. [CrossRef]
22. Ahmed, A.A.; Egleston, B.; Alcantara, P.; Li, L.; Pollack, A.; Horwitz, E.M.; Buyyounouski, M.K. A novel method for predicting late genitourinary toxicity after prostate radiation therapy and the need for age-based risk-adapted dose constraints. *Int. J. Radiat. Oncol. Biol. Phys.* **2013**, *86*, 709–715. [CrossRef] [PubMed]
23. Fleming, C.; Kelly, C.; Thirion, P.; Fitzpatrick, K.; Armstrong, J. A method for the prediction of late organ-at-risk toxicity after radiotherapy of the prostate using equivalent uniform dose. *Int. J. Radiat. Oncol. Biol. Phys.* **2011**, *80*, 608–613. [CrossRef] [PubMed]
24. Tian, Z.; Yen, A.; Zhou, Z.; Shen, C.; Albuquerque, K.; Hrycushko, B. A machine-learning–based prediction model of fistula formation after interstitial brachytherapy for locally advanced gynecological malignancies. *Brachytherapy* **2019**, *18*, 530–538. [CrossRef] [PubMed]
25. Eftekhar, B.; Mohammad, K.; Ardebili, H.E.; Ghodsi, M.; Ketabchi, E. Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data. *BMC Med. Inf. Decis. Mak.* **2005**, *5*, 3. [CrossRef] [PubMed]
26. Matsuo, K.; Purushotham, S.; Jiang, B.; Mandelbaum, R.S.; Takiuchi, T.; Liu, Y.; Roman, L.D. Survival outcome prediction in cervical cancer: Cox models vs deep-learning model. *Am. J. Obstet. Gynecol.* **2019**, *220*, 381.e1–381.e14. [CrossRef]
27. Abouzari, M.; Rashidi, A.; Zandi-Toghani, M.; Behzadi, M.; Asadollahi, M. Chronic subdural hematoma outcome prediction using logistic regression and an artificial neural network. *Neurosurg. Rev.* **2009**, *32*, 479–484. [CrossRef]
28. Tu, J.V. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J. Clin. Epidemiol.* **1996**, *49*, 1225–1231. [CrossRef]
29. Rasheed, K.; Qayyum, A.; Ghaly, M.; Al-Fuqaha, A.; Razi, A.; Qadir, J. Explainable, trustworthy, and ethical machine learning for healthcare: A survey. *Comput. Biol. Med.* **2022**, *149*, 106043. [CrossRef]
30. Moore, N.S.; McWilliam, A.; Aneja, S. Bladder cancer radiation oncology of the future: Prognostic modelling, radiomics, and treatment planning with artificial intelligence. *Semin. Radiat. Oncol.* **2023**, *33*, 70–75. [CrossRef]
31. Luo, Y.; Chen, S.; Valdes, G.J.M.P. Machine learning for radiation outcome modeling and prediction. *Med. Phys.* **2020**, *47*, e178–e184. [CrossRef]

32. Luo, Y.; Tseng, H.H.; Cui, S.; Wei, L.; Ten Haken, R.K.; El Naqa, I. Balancing accuracy and interpretability of machine learning approaches for radiation treatment outcomes modeling. *BJR Open* **2019**, *1*, 20190021. [CrossRef]
33. Cox, J.D.; Stetz, J.; Pajak, T.F. Toxicity criteria of the Radiation Therapy Oncology Group (rtog) and the European Organization for Research and Treatment of Cancer (EORTC). *Int. J. Radiat. Oncol. Biol. Phys.* **1995**, *31*, 1341–1346. [CrossRef] [PubMed]
34. Haie-Meder, C.; Pötter, R.; Van Limbergen, E.; Briot, E.; De Brabandere, M.; Dimopoulos, J.; Dumas, I.; Hellebust, T.P.; Kirisits, C.; Lang, S.; et al. Recommendations from gynaecological (GYN) gec-estro working Group (I): Concepts and terms in 3D image based 3D treatment planning in cervix cancer brachytherapy with emphasis on MRI assessment of GTV and CTV. *Radiother. Oncol.* **2005**, *74*, 235–245. [CrossRef] [PubMed]
35. Pötter, R.; Haie-Meder, C.; Van Limbergen, E.; Barillot, I.; De Brabandere, M.; Dimopoulos, J.; Dumas, I.; Erickson, B.; Lang, S.; Nulens, A.; et al. Recommendations from gynaecological (gyn) gec estro working group (ii): Concepts and terms in 3D image-based treatment planning in cervix cancer brachytherapy-3D dose volume parameters and aspects of 3D image-based anatomy, radiation physics, radiobiology. *Radiother. Oncol.* **2006**, *78*, 67–77. [CrossRef] [PubMed]
36. Kang, H.C.; Shin, K.H.; Park, S.Y.; Kim, J.Y. 3D CT-based high-dose-rate brachytherapy for cervical cancer: Clinical impact on late rectal bleeding and local control. *Radiother. Oncol.* **2010**, *97*, 507–513. [CrossRef] [PubMed]
37. Song, S.; Kim, J.-Y.; Kim, Y.-J.; Yoo, H.J.; Kim, S.H.; Kim, S.-K.; Lim, M.C.; Kang, S.; Seo, S.-S.; Park, S.-Y. The size of the metastatic lymph node is an independent prognostic factor for the patients with cervical cancer treated by definitive radiotherapy. *Radiother. Oncol.* **2013**, *108*, 168–173. [CrossRef]
38. Koom, W.S.; Sohn, D.K.; Kim, J.-Y.; Kim, J.W.; Shin, K.H.; Yoon, S.M.; Kim, D.Y.; Yoon, M.; Shin, D.; Park, S.Y. Computed tomography-based high-dose-rate intracavitary brachytherapy for uterine cervical cancer: Preliminary demonstration of correlation between dose–volume parameters and rectal mucosal changes observed by flexible sigmoidoscopy. *Int. J. Radiat. Oncol. Biol. Phys.* **2007**, *68*, 1446–1454. [CrossRef]
39. Noh, J.M.; Park, W.; Kim, Y.S.; Kim, J.-Y.; Kim, H.J.; Kim, J.; Kim, J.H.; Yoon, M.S.; Choi, J.H.; Yoon, W.S. Comparison of clinical outcomes of adenocarcinoma and adenosquamous carcinoma in uterine cervical cancer patients receiving surgical resection followed by radiotherapy: A multicenter retrospective study (KROG 13-10). *Gynecol. Oncol.* **2014**, *132*, 618–623. [CrossRef]
40. Kingma, D.P.; Adam, B.J. A method for stochastic optimization. *arXiv* **2014**, arXiv:14126980. [CrossRef]
41. Altmann, A.; Toloşi, L.; Sander, O.; Lengauer, T. Permutation importance: A corrected feature importance measure. *Bioinformatics* **2010**, *26*, 1340–1347. [CrossRef]
42. Uzair, M.; Noreen, J. Effects of hidden layers on the efficiency of neural networks. In Proceedings of the 2020 IEEE 23rd International Multitopic Conference (INMIC), Bahawalpur, Pakistan, 5–7 November 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6. [CrossRef]
43. Ljubicic, M.L.; Madsen, A.; Juul, A.; Almstrup, K.; Johannsen, T.H. The application of principal component analysis on clinical and biochemical parameters exemplified in children with congenital adrenal hyperplasia. *Front. Endocrinol.* **2021**, *12*, 652888. [CrossRef]
44. Zhang, Z.; Castelló, A. Principal components analysis in clinical studies. *Ann. Transl. Med.* **2017**, *5*, 351–357. [CrossRef]
45. Ghoshal, U.C.; Rai, S.; Kulkarni, A.; Gupta, A. Prediction of outcome of treatment of acute severe ulcerative colitis using principal component analysis and artificial intelligence. *JGH Open* **2020**, *4*, 889–897. [CrossRef] [PubMed]
46. Cheon, H.; Kim, H.; Kim, J. Deep learning in radiation oncology. *Prog. Med. Phys.* **2020**, *31*, 111–123. [CrossRef]
47. Appelt, A.L.; Elhaminia, B.; Gooya, A.; Gilbert, A.; Nix, M. Deep learning for radiotherapy outcome prediction using dose data–a review. *Clin. Oncol. (R Coll. Radiol.)* **2022**, *34*, e87–e96. [CrossRef] [PubMed]
48. Li, X.; Xiong, H.; Li, X.; Wu, X.; Zhang, X.; Liu, J.; Bian, J.; Dou, D. Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. *Knowl. Inf. Syst.* **2022**, *64*, 3197–3234. [CrossRef]