IOP Publishing Phys. Med. Biol. 65 (2020) 035014 (11pp) https://doi.org/10.1088/1361-6560/ab63b6

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

RECEIVED
16 August 2019

REVISED
29 November 2019

ACCEPTED FOR PUBLICATION
18 December 2019

PUBLISHED
4 February 2020

PAPER

A novel deep learning model using dosimetric and clinical information for grade 4 radiotherapy-induced lymphopenia prediction

Cong Zhu[1,3,6], Steven H Lin[2], Xiaoqian Jiang[4], Yang Xiang[4], Zayne Belal[5], Goo Jun[3] and Radhe Mohan[1,6]

[1] Department of Radiation Physics, The University of Texas MD Anderson Cancer Center, Houston, TX, United States of America
[2] Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, United States of America
[3] Department of Epidemiology, Human Genetics, and Environmental Sciences, The University of Texas Health Science Center at Houston, Houston, TX, United States of America
[4] School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, United States of America
[5] Texas Tech University Health Sciences Center, Lubbock, TX, United States of America
[6] Authors to whom any correspondence should be addressed.

E-mail: czhu2@mdanderson.org, cong.zhu@uth.tmc.edu and rmohan@mdanderson.org

Keywords: deep learning, recurrent neural network, radiotherapy, lymphopenia, dose volume histogram, esophageal cancer

Supplementary material for this article is available online

Abstract

Radiotherapy-induced lymphopenia has increasingly been shown to reduce cancer survivorship. We developed a novel hybrid deep learning model to efficiently integrate an entire set of dosimetric parameters of a radiation treatment plan with a patient's pre- and mid-treatment information to improve the prediction of grade 4 radiotherapy-induced lymphopenia.

We proposed a two-input channel hybrid deep learning model to process dosimetric information using a stacked bi-directional long-short term memory structure and non-dosimetric information using a multilayer perceptron structure independently before integrating the dosimetric and non-dosimetric information for final prediction. The model was trained from 505 patients and tested in 216 patients. We compared our model with other popular predictive models, including logistic regression (with and without elastic-net regularization) random forest, support vector machines, and artificial neural network.

Our hybrid deep learning model out-performed other predictive models in various evaluation metrics. It achieved the highest area under the curve at 0.831, accuracy at 0.769, F1 score at 0.631, precision at 0.670, and recall at 0.610. The hybrid deep learning model also demonstrated robustness in exploiting the value of dosimetric parameters in predictive modeling.

We demonstrated that our hybrid deep learning model with a two-input channel structure, which addressed the sequential and inter-correlated nature of dosimetric parameters, could potentially improve the prediction of radiotherapy-induced lymphopenia. Our proposed deep learning framework is flexible and transferable to other related radiotherapy-induced toxicities.

1. Introduction

Radiotherapy (RT) is an integral component of cancer treatment. It damages DNA in the cells to suppress tumor growth. Although RT is locally targeted at the tumor, it unavoidably exposes normal tissues to some radiation and causes complications (Burman *et al* 1991). The optimal balance between efficient exploitation of RT to achieve effective tumor control and acceptable risk of normal tissue complications thus is a critical component of RT planning.

One of the common side effects induced by RT is lymphopenia. This is likely due to the large low 'radiation dose bath' of traditional photon therapy, which kills the highly radiosensitive circulating lymphocytes, as well as

the unintentional exposure of lymphoid organs such as nodes, bone marrow, and the spleen to radiation (Stratton *et al* 1975, Yovino *et al* 2013). Increasing evidence has shown that a low absolute lymphocyte count (ALC), i.e. lymphopenia, reduces overall survival, disease-specific survival, and progression-free survival in cancer patients (Kitayama *et al* 2010, Tang *et al* 2014, Cho *et al* 2016a, 2016b, Davuluri *et al* 2017, Venkatesulu *et al* 2018). Therefore, reliable prediction of RT-induced lymphopenia as a function of radiation dose patterns (i.e. dosimetric factors) and patient-specific (non-dosimetric) factors represents an essential part of radiation treatment planning.

DVHs are an effective tool for treatment planning and studying the correlation of toxicity with radiation dose distributions (Hernando *et al* 2001, Michalski *et al* 2010, Tang *et al* 2014). However, because of the sequential and highly intercorrelated nature of DVHs, full use of them in statistical predictive modeling remains a challenging task. Inclusion of all dosimetric parameters may increase the risk of over-fitting the model or inaccurate prediction of outcomes due to multicollinearity among predictors. Most studies have used the variable selection procedure to keep the most significant DVH parameters in the modeling process to address these drawbacks. However, such a method might sacrifice some of the potentially valuable information from unselected dosimetric parameters, thus making the predictive model less useful in comprehensively evaluating the treatment plan as a whole and defining dose- and dose volume constraints most appropriately.

To address this dilemma, we have proposed a novel hybrid deep learning neural network that is capable of using the entire set of cumulative DVH parameters and controlling for over-fitting or multicollinearity by implementing cutting-edge model regularization techniques. The model is built with a two-channel input structure to process dosimetric and non-dosimetric information in parallel first before integrating them to predict RT-induced lymphopenia. To validate our model's robustness, we compared it with multiple popular statistical methods, including logistic regression (with and without elastic-net regularization), random forest, support vector machines, and artificial neural network.

2. Materials and methods

2.1. Inclusion and exclusion criteria

This project was approved by the institutional review board of The University of Texas MD Anderson Cancer Center with a waiver of the requirement of obtaining informed consent. The study adhered to the Health Insurance Portability and Accountability Act. We extracted data from records of patients who received concurrent chemoradiotherapy (with or without surgery) for biopsy-proven esophageal cancer between January 2004 and November 2017. Exclusion criteria included planned total radiation dose other than 50.4 Gy, radiation modality other than proton beam therapy or intensity-modulated RT, split course RT, simultaneous irradiation of a second primary tumor, tumor overall stage IV (or unknown), histologic diagnosis other than adenocarcinoma or squamous cell carcinoma, history of hematologic malignancy, endomucosal resection before chemoradiotherapy, missing records in baseline blood sample data (e.g. baseline ALC, red blood cell counts, white blood cell counts), or less than three weekly documented ALC values during the treatment.

2.2. Definition and selection of variables

The primary outcome was grade 4 RT-induced lymphopenia (G4RIL), which was defined as an ALC less than 200 cells μl^{-1} during and immediately following the course of RT. Mean lung, spleen, and heart doses were calculated using DVHs.

We selected 49 parameters as predictors on the basis of their clinical relevance, and low level of missingness (<20%). Twenty-seven of these parameters are DVH parameters that correspond to three organs at risk (OAR): lung, heart, and spleen (V5, V10…V45). The rest of the variables included RT modality (proton or photon), baseline ALC, ALC during the first week of RT, race, sex, age, body mass index (BMI), total blood volume, planning target volume (PTV), blood component profiles at baseline (red blood cells, white blood cells, and others), tumor location, tumor histologic characteristics, mean spleen/heart/lung dose, and use of induction chemotherapy. After applying inclusion and exclusion criteria and limiting the data to that of patients with complete records of all predictors, we found that the records of 721patients were eligible for the analyses.

2.3. Data preparation

The original data were split into a training and a testing set in a 7:3 ratio (505:216) using a stratified random sampling scheme to ensure balanced distribution of radiation modalities across two datasets. Min-max normalization was applied to both the training and testing sets according to equation (1):

$$X = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

where X_{min} is the minimum value of variable X vector and X_{max} is the maximum value of variable X vector. All the models were trained on training set and evaluated using the testing set.

2.4. Development of the Hybrid deep learning models

A hybrid deep learning model with two-channel input variables (dosimetric and non-dosimetric) was developed to predict patients' risk of G4RIL using the pre-selected predictors (figure 1). The first input channel processed non-dosimetric information such as the patient's age and sex, which was encoded using a multilayer perceptron (MLP) neural network with three fully connected layers. The second input branch was an independent pipeline to process the dosimetric parameters of the OARs, the primary architecture of which was a stacked bi-directional long-short term memory (LSTM) neural network (Hochreiter and Schmidhuber 1997, Hernando *et al* 2001, Schuster and Paliwal (1997)). LSTM is a variant of the recurrent neural networks model, which is designed to effectively process sequential data that contain correlations among adjacent data points. We hypothesized that the inherent mutually dependent and sequential nature of DVH parameters makes them suitable for the structure of the model in figure 1.

We also explored two additional options for the dosimetric pipeline with simpler architectures. Option 1 consisted of one simple dense layer and option 2 used a stacked unidirectional (forward) LSTM structure (supplementary figures 1 and 2(stacks.iop.org/PMB/65/035014/mmedia)).

Finally, a concatenation layer was connected to the end of two input pipelines, and this layer aggregated the processed information from both sources. Three consecutive fully connected layers were built to further encode the information. We employed sigmoid as the activation function and case classification before the final output layer. The discriminative threshold was preset as 0.5, which means a patient was classified as being at risk of developing G4RIL if the predicted probability was greater than or equal to 0.5.

In the dosimetric variables processing pipeline, 27 DVH parameters that started from V5 and ended at V45 in increments of 5 Gy for each of the three OARs were converted into the format of $\{X^1, X^2, X^3, \ldots, X^9\}$ in which each individual element X^t would be feeding into its corresponding LSTM unit (purple box that was labeled as LSTM in figure 1). $X^t \epsilon \mathbb{R}^3$ is an array of $\{X_1^t, X_2^t, X_3^t\}$ at step t. A 'step' referred to a piece of data from the whole sequential data. Specifically, the X^1 corresponds to {Spleen V5, Heart V5, Lung V5} that is the first piece(step) of the sequential DVH data, X^2 corresponds to second piece(step) of the sequential DVH data {Spleen V10, Heart V10, Lung V10}, and so on until X^9 {Spleen V45, Heart V45, Lung V45}. The direction that LSTM units processed sequential information from low dose step to high dose step ($X^1 \rightarrow X^9$) is called forward, and backward the other way around (figure 1).

2.5. Configuration of hyperparamters of the deep learning model

The deep learning model was trained with up to 220 epochs using a batch size of 40 and was implemented with Adam optimizer (Kingma and Ba 2014) with learning rate $= 10^{-3}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 10^{-8}$. The regularization of hyper-parameters of training set was adjusted empirically until the loss functions of both the training and testing set declined with similar trend without significant gap between each other. Finally, we applied kernel constraint forcing recurrent kernel weights to a magnitude of 1 (unit norm) for the first LSTM layer and a recurrent dropout of rate 0.3 to the two stacked LSTM layers since prior work showed that combination of these two techniques yielded superior performance than using drop-out alone (Srivastava *et al* 2014). An additional dropout of rate 0.2 was employed to each of the dense/fully connected layers from both input processing pipelines up to the concatenation layer.

2.6. Development of comparison models

Popular classification models, including logistic regression (Hosmer *et al* 2013), support vector machines (Scholkopf and Smola 2001), and random forest (Liaw and Wiener 2002), were developed for comparison with the deep learning model. All of these models used the same training and testing set as the hybrid deep learning model.

2.6.1. Logistic regression with and without being regularized by elastic-net

Logistic regression was first modeled as a function of all of the pre-selected variables and was then modeled using a subset of the original variable pool selected by the elastic-net regularization on the training set to adjust for potential over-fitting issues (Zou and Hastie 2005, Friedman *et al* 2010). Elastic-net regression combines L1 and L2 regularization methods to achieve a balance of good prediction performance and model simplicity:

$$\hat{\beta} = \arg\ \min||y - X\beta||^2 + \lambda_2||\beta||^2 + \lambda_2||\beta||. \quad (2)$$

Logistic regression was implemented using Adam optimizer (Kingma and Ba 2014) with learning rate $= 10^{-3}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 10^{-8}$. The model was trained up to 600 epochs with a batch size of 40. The logistic regression was set with longer training epochs to account for its potentially slower converging process.

2.6.2. Artificial neural network

The artificial neural network was developed with three hidden layers. The number of nodes in each layer was manually tuned until loss functions of the training and testing sets demonstrated a similar pattern of changes

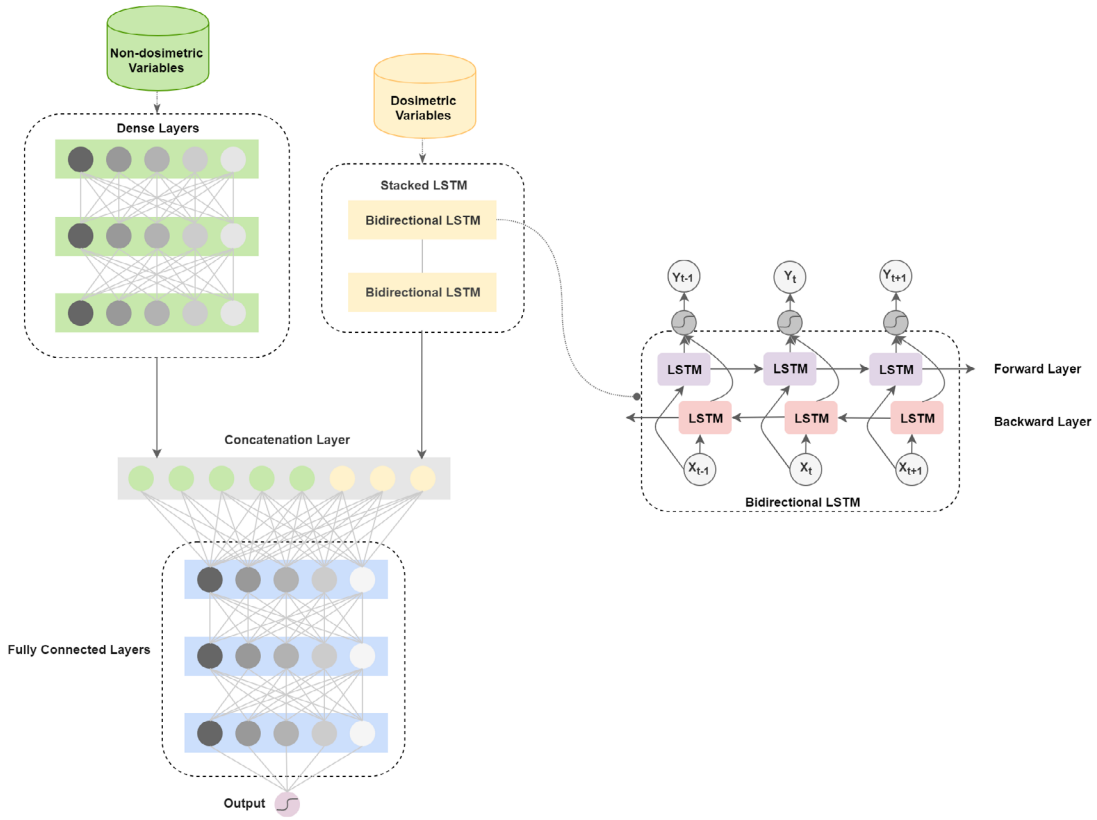

Figure 1. Architecture of the proposed hybrid deep learning model with two-channel input variables.

over epochs without a significant gap of discrepancy. Finally, the three layers were implemented with 60, 30, and 20 nodes. We implemented Adam optimizer (Kingma and Ba 2014) with learning rate $= 10^{-3}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 10^{-8}$ for this model.

2.6.3. *Random forest*

The structure of random forest was determined by a set of hyper-parameters that was identified using grid search method with three-fold cross-validation on the training set. The grid search method explored through all the possible combination of different hyper-parameters until identify a set that yields best model performance. A set of hyper-parameters of random forest included the number of trees in the forest of the model (180, 2100, 2400, 2700), maximum depth of each tree (30, 45, 60, 75), minimal number of data points allowed in a leaf node (2, 4, 6), and the minimum number of samples required to split an internal leaf node (5, 10, 15), max number of features considered for splitting a node (automatically determined).

2.6.4. *Support vector machines*

We applied same grid search technique to support vector machines. The pre-specified search space of hyper-parameters are types of kernel, which included radial, linear, and polynomial; cost parameter C (1, 10, 100, 1000, 1500, 2000) if using a radial or linear kernel; degree of polynomial (2, 3, 4, 5, 6, 7, 8, 9) if using a polynomial kernel; and value of gamma ($1 \times 10^{-3}, 1 \times 10^{-4}$) if using a radial kernel.

We used R package glmnet to perform elastic-net regression and Python based keras and Scikit-learn for constructing the neural network and performing the grid search and cross-validation (Friedman *et al* 2010, Pedregosa *et al* 2011).

2.7. Evaluation criteria for all models

Model classification performance was assessed using accuracy, recall, precision, and F1 score at a pre-specified discrimination threshold of 0.5. The evaluation metrics were defined as follows:

$$Accuracy = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{TN} + N_{FP} + N_{FN}} \tag{3}$$

$$Precision = \frac{N_{TP}}{N_{TP} + N_{FP}} \tag{4}$$

$$Recall = \frac{N_{TP}}{N_{TP} + N_{FN}} \tag{5}$$

$$F1\text{-}Score = \frac{2 * Recall * Precision}{Recall + Precision} \tag{6}$$

where TP indicates true positive; FP, false positive; FN, false negative; and TN, true negative. In addition, the receiver operating characteristics (ROC) curve was plotted to assess the model's overall diagnostic ability as the discrimination threshold was varied. The evaluation is based on the area under ROC curve (AUC).

2.8. Evaluation of robustness and the effect of DVH parameters on prediction power of all models

A robust model should demonstrate sensitivity to change in predictors. Its classification performance is desired to improve with inclusion of informative predictors and vice versa when they are dropped. We re-assessed the model's performance using the proposed five evaluation metrics by excluding one OAR's DVH from the predictors (e.g. excluding the lung DVH but retaining the spleen and heart DVHs) and then by removing all DVH parameters. This procedure was implemented for all models: the deep learning model, logistic regression, support vector machines, and random forest.

3. Results

3.1. Clinical profiles and treatment information

Among 721 patients who were eligible for the analyses, 432 (60%) developed G4RIL during RT. A significantly smaller proportion of patients received proton therapy than photon therapy in the G4RIL group (proton therapy: 183 [42.4%]; photon therapy: 249 [57.6%], $p < 0.0001$). Compared with the G4RIL group, patients in the non-G4RIL group had overall favorable clinical profiles such as younger mean age (61.79 years compared with 64.23 years), higher ALC at baseline (1.78 K μl^{-1} compared with 1.42 K μl^{-1}), and larger total blood volume (5.18 l versus 4.96 l). More details are shown in tables 1 and 2.

DVH parameters of different organs showed very similar negative correlations (~ -0.2) with ALC nadir (defined as the minimum ALC value) during the RT or immediately thereafter. Heart DVHs demonstrated the strongest average correlations with ALC nadir (mean $\pm$ standard deviation: -0.28 ± 0.01), whereas spleen and lung DVHs showed slightly weaker but similar levels of correlations (-0.22 ± 0.02; figure 3).

3.2. Model performance

As shown in figure 3, our proposed model demonstrated a superior classification performance at various discriminative thresholds (AUC = 0.831) compared with logistic regression and other machine learning approaches (random forest, AUC = 0.780; support vector machines, AUC = 0.792). Logistic regression that included all pre-selected parameters used in the other models achieved the second worst AUC (0.787). Results were not improved after applying the elastic-net regularization that selected partial clinical profiles and DVH parameters (AUC = 0.787).

Figure 4 compares classification performances among four deep learning models: artificial neural network (ANN) and our proposed two-input hybrid model with three variations (stacked bidirectional LTSM, stacked unidirectional LTSM, and MLP) in the processing pipeline of the DVHs. All proposed models outperformed ANN with lowest testing AUC of 0.811 for MLP architecture and best testing AUC of 0.831 for stacked bi-directional LSTM architecture in the dosimetric information processing pipeline.

Table 3 presents the overall performance of the model at a pre-defined 0.5 discriminative threshold. We observed the highest accuracy of approximately 77% from the deep learning model using the stacked bi-directional LSTM architecture for the dosimetric information processing pipeline. Our model also achieved the best results among the rest metrics including accuracy (76.9%), F1-score (69.5%), and precision (74%) except for recall (65.5%) which is the second best result.

The proposed deep learning model demonstrated robustness, as shown in table 4. With the exclusion of each individual OAR's DVH parameters, the hybrid deep learning model showed a slight decline across all evaluation metrics. The loss of accuracy and precision due to removing all three OAR DVH parameters (accuracy: −6.8%; precision: −12%) was approximately the linear addition of the loss due to individual OAR DVH removal (accuracy spleen: −1.3%, lung: −3.8%, heart: −3.1%; precision spleen: −7.3%, lung: −1.8%, heart: −2.4%). Such pattern was not observed in other evaluation metrics (e.g. AUC, recall) which implied the role of individual OAR's DVH in boosting prediction power is not simple linearly ensembled.

In contrast, other machine learning methods in which hyper-parameters were optimized using three-fold cross-validation grid search, including random forest (number of estimators: 2400, maximum depth: 60, minimal number of sample required at each leaf node: 6, minimal sample split: 15) and support vector machines (Gaussian kernel, C: 1500, gamma = 0.001), showed unstable results. Some of their evaluation metrics

Table 1. Patient clinical profiles and treatment information (categorical variables) for those who developed grade 4 radiotherapy-induced lymphopenia (G4RIL) and those who did not.

	No. (%)		
Variable	No G4RIL ($n = 432$)	G4RIL ($n = 289$)	P
Radiation modality			<0.0001
Photon	249 (57.6)	231 (79.9)	
Proton	183 (42.4)	58 (20.1)	
Sex			0.884
Female	51 (11.8)	36 (12.5)	
Male	381 (88.2)	253 (87.5)	
Race			0.208
Black	32 (7.4)	30 (10.4)	
White	400 (92.6)	259 (89.6)	
Histologic characteristics			0.438
Adenocarcinoma	393 (91.0)	257 (88.9)	
Squamous cell carcinoma	39 (9.0)	32 (11.1)	
Tumor location			1
Upper and middle	29 (6.7)	19 (6.6)	
Distal	403 (93.3)	270 (93.4)	
Induction chemotherapy			0.239
No	309 (71.5)	194 (67.1)	
Yes	123 (28.5)	95 (32.9)	

Table 2. Patient clinical profiles (continuous variables) for those who developed grade 4 radiotherapy-induced lymphopenia (G4RIL) and those who did not[a].

	Mean (standard deviation)		
Variable	No G4RIL ($n = 432$)	G4RIL ($n = 289$)	P
Baseline ALC, K μl^{-1}	1.78 (0.61)	1.42 (0.51)	<0.0001
RT week 1 ALC, K μl^{-1}	1.12 (0.42)	0.75 (0.34)	<0.0001
Age, years	61.79 (10.93)	64.23 (10.49)	0.003
BMI, kg m^{-2}	27.01 (5.95)	25.77 (5.41)	0.004
Total blood volume, l	5.18 (0.87)	4.96 (0.84)	0.001
PTV, cm^3	584.27 (252.32)	722.1 (272.77)	<0.0001
Baseline RBC $10^7/\mu l$	4.48 (0.51)	4.35 (0.53)	0.002
Baseline HB, g/dl	13.28 (1.62)	12.94 (1.7)	0.008
Baseline HT, %	39.5 (4.63)	38.5 (4.55)	0.005
Baseline WBC, 10^9 cell l^{-1}	7.33 (2.43)	6.79 (2.48)	0.004
Baseline ANC, K μl^{-1}	4.64 (2.13)	4.49 (2.18)	0.354
Baseline PLC, K μl^{-1}	231.78 (73.6)	232.46 (78.27)	0.908
Baseline monocyte count, K μl^{-1}	0.69 (0.24)	0.66 (0.24)	0.115
Mean lung dose, Gy	7.96 (3.36)	9.94 (3.57)	<0.0001
Mean heart dose, Gy	18.27 (7.23)	23.56 (7.91)	<0.0001
Mean spleen dose, Gy	17.34 (9.34)	21.49 (9.86)	<0.0001

[a] ALC, absolute lymphocyte count; RT, radiotherapy; BMI, body mass index; PTV, planning treatment volume; RBC, red blood cell count; HB, hemoglobin level; HT, hematocrit level; WBC, white blood cell count; ANC, absolute neutrophil count; PLC, platelet count.

contradictorily improved after removing an OAR DVH. For example, all evaluation metrics for support vector machines increased when the lung DVH was excluded from the input (accuracy: 0.699 → 0.713; F1 score: 0.663 → 0.684; precision: 0.604 → 0.615; recall: 0.736 → 0.770; AUC: 0.792 → 0.798).

4. Discussion

In the current study, we sought to develop an efficient predictive model that was capable of exploiting irradiated non-target ORAs (e.g. lung, heart, spleen) DVH's predicting power regarding G4RIL as sufficiently as possible. Full dosimetric parameters provided comprehensive information regarding dose and volume of the RT plan and thus was expected to increase model performance significantly. However such a task seemed to be challenging

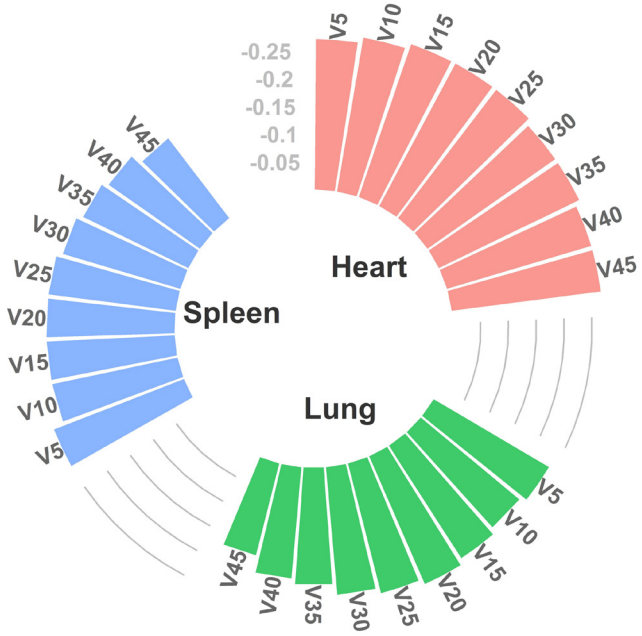

Figure 2. Pearson correlation between organ-at-risk dose volume histograms and absolute lymphocyte count nadir during radiotherapy.

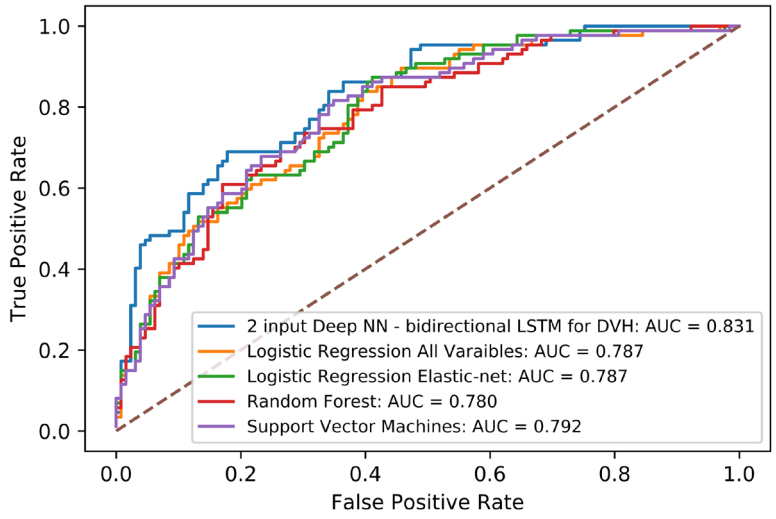

Figure 3. Classification capability of the proposed deep learning model compared with other statistical and machine learning methods, using the testing set. Deep NN, deep learning neural network model; LSTM, long-short term memory; DVH, dose volume histogram; AUC, area under the curve.

considering DHVs' high inter-dependence (VIF > 10) (Midi *et al* 2010, Allison 2012) which was shown in the supplement table 1, and their weak linear associations with the ALC nadir (figure 2). Our proposed deep learning model overcame these difficulties by showing a $AUC_{testing\text{-}set}$ 0.831 that significantly out-performed other machine learning methods. Since AUC is a performance measurement that evaluates model's classification capability at various discriminative threshold, this result indicated our model achieved superior comprehensive prediction performance than other approaches. The proposed model's performance at point-wise threshold (0.5) also outperformed other models by showing best accuracy (76.9%), F1-score (69.5%), and precision (74%) except for recall (65.5%) which is the second best result.

In addition, the proposed hybrid deep learning model showed stronger stability than other approaches in the robustness test that evaluated model's change in prediction performance after removing specific OAR DVHs. Specifically, we observed universal decline in all evaluation metrics after excluding DVHs from predictors in the proposed deep learning model. In contrast, support vector machines and random forest experienced contradictory rise in various metrics (accuracy, f1-score, precision, recall and AUC) after excluding DVH from predictors. The performance of logistic regression is less affected during this process. The model only observed slight boost

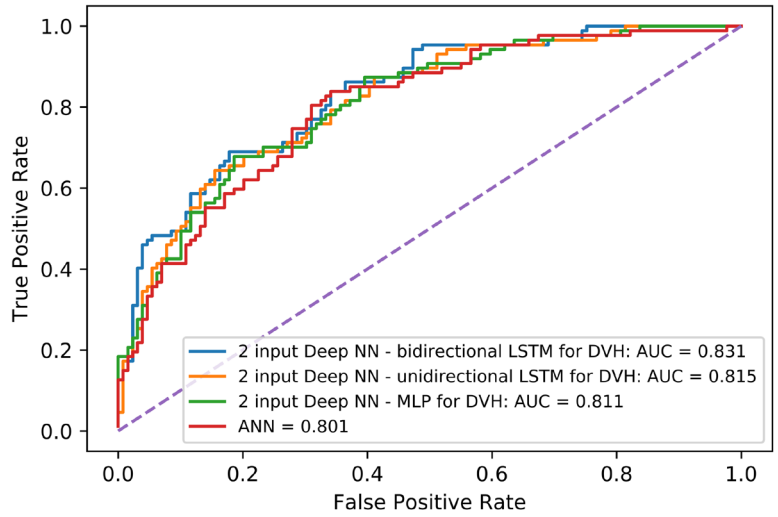

Figure 4. Classification capability of various deep learning models using the testing set. Deep NN, deep learning neural network model; LSTM, long-short term memory; DVH, dose volume histogram; AUC, area under the curve; MLP, multilayer perceptron; ANN, artificial neural network.

Table 3. Evaluation of model performance using the testing set of patients who developed grade 4 radiotherapy-induced lymphopenia.

	Hybrid deep learning model[a]	Logistic regression	Logistic regression with elastic-net regularization	Random forest	Support vector machines
Accuracy	0.769	0.717	0.722	0.718	0.699
F1 score	0.695	0.616	0.589	0.647	0.663
Precision	0.740	0.681	0.621	0.651	0.604
Recall	0.656	0.563	0.575	0.644	0.736

[a] Using stacked bi-directional long-short term memory architecture for the dose-volume histogram input.

in accuracy (+1.3%) and precision (+4.6%) after excluding spleen DVH and minor increase (+0.7%) in precision after excluding lung DVH. However logistic regression model's mediocre prediction performance (accuracy: 71.7%; F1-score: 61.6%; precision: 68.1%; recall: 56.3%; AUC 78.7%) showed a weakness in integrating all DHV for efficient G4RIL prediction. These results implied that proposed hybrid deep learning model might be more suitable than traditional statistical models when the primary goal is improving the predictive power rather than identifying the effect of an individual parameter on the outcome of interest.

We observed that DVHs of heart contributed slightly more to the prediction power of G4RIL than those of spleen, which is a major immune organ. The accuracy, F1 score, recall and AUC of testing set dropped by 3.1%, 6.0%, 8.9% and 4.6% when excluding the heart DVH from the model. In contrast, same metrics decreased by 1.3%, 3.6%, 0.0%, and 3.7% when excluding the spleen DVHs. This pattern is similar to the results from the univariate analyses, in which heart DVHs demonstrated the strongest average correlations with ALC nadir (mean $\pm$ standard deviation: -0.28 ± 0.01), whereas spleen DVHs showed slight weaker correlations (-0.22 ± 0.02). These results demonstrate that the irradiation of non-lymphatic regions such as lungs, which are not critical immune organs like the spleen, might also play a role in the risk of RT-induced lymphopenia. These results are consistent with prior work that focused on other cancer sites. Tang *et al*'s found that lung dose according to the DVH was inversely correlated with lymphocyte nadir in patients with non-small cell lung cancer (Tang *et al* 2014). This could be because the pro-tumorigenic effect is modulated by microenvironmental signaling among lymphocytes or descendants of lymphocytes that have been exposed to low radiation doses (Wright and Coates 2006, Coates *et al* 2008).

There are two possible explanations for the improvement in the model performance despite of the two challenges mentioned at the beginning: (1) The Pearson correlation only measured the linear association between individual DVH and ALC nadir, there might exist a non-linear association between dosimetric parameters and G4RIL risk. Such relationships could be efficiently absorbed by the flexible multilayer and multi-nodes structure of deep learning models. (2) The efficiency of analysis was further improved by the unique parallel analytical pipeline for dosimetric and non-dosimetric data. Specifically, the stacked bi-directional LSTM architecture for processing the dosimetric data allowed our model to efficiently exploit the backward and forward DVH parameter information to increase the prediction accuracy. In specific, we converted the DVH parameters into a sequential data format in which the sequence began with lower dose regions (e.g. {Spleen V5, Heart V5, Lung V5}), and

Table 4. Contribution of individual DVHs to each model's prediction capability, using the testing set of patients who developed grade 4 radiotherapy-induced lymphopenia[a].

	All DVHs	Exclude spleen		Exclude lung		Exclude heart		Exclude all	
		value	% change	value	% change	value	% change	value	% change
Hybrid deep learning model									
Accuracy	0.769	0.741	−1.3%	0.759	−3.8%	0.745	−3.1%	0.717	−6.8%
F1 score	0.695	0.670	−3.6%	0.683	−1.9%	0.653	−6.0%	0.647	−6.9%
Precision	0.740	0.686	−7.3%	0.727	−1.8%	0.722	−2.4%	0.651	−12%
Recall	0.655	0.655	0.0%	0.643	−1.8%	0.597	−8.9%	0.644	−2.3%
AUC	0.831	0.800	−3.7%	0.788	−5.2%	0.793	−4.6%	0.794	−4.5%
Logistic regression									
Accuracy	0.717	**0.726**	**1.3%**	0.717	0.0%	0.708	−1.3%	0.708	−1.3%
F1 score	0.616	0.614	−0.3%	0.611	−0.8%	0.588	−4.5%	0.593	−3.7%
Precision	0.681	**0.712**	**4.6%**	0.686	**0.7%**	0.681	0.0%	0.676	−0.7%
Recall	0.563	0.540	−4.1%	0.552	−2%	0.517	−8.2%	0.528	−6.2%
AUC	0.787	0.766	−2.7%	0.765	−2.8%	0.763	−3.0%	0.740	−6.0%
Support vector machines									
Accuracy	0.699	0685	−2.0%	**0.713**	**2.0%**	**0.708**	**1.3%**	0.699	0.0%
F1 score	0.663	0.638	−3.8%	**0.684**	**3.2%**	**0.674**	**1.7%**	**0.683**	**3.0%**
Precision	0.604	0.594	−1.7%	**0.615**	**1.8%**	**0.613**	**1.5%**	0.593	−1.8%
Recall	0.736	0.690	−6.3%	**0.770**	**4.6%**	**0.747**	**1.5%**	**0.805**	**9.4%**
AUC	0.792	0.780	−1.5%	**0.798**	**0.8%**	**0.793**	**0.1%**	0.786	−0.8%
Random forest									
Accuracy	0.718	0.708	−1.4%	**0.722**	**0.6%**	**0.722**	**0.6%**	0.708	−1.4%
F1 score	0.647	0.623	−3.7%	**0.655**	**1.2%**	**0.651**	**0.6%**	**0.652**	**0.8%**
Precision	0.651	0.650	−0.2%	**0.655**	**0.6%**	**0.659**	**1.2%**	0.628	−3.5%
Recall	0.644	0.598	−7.1%	**0.655**	**1.7%**	0.644	0.0%	**0.678**	**5.3%**
AUC	0.78	0.781	**0.1%**	**0.786**	**0.8%**	0.776	-0.5%	**0.785**	**0.6%**

[a] Boldface indicates an increase in accuracy. DVH, dose volume histogram; AUC, area under the curve.

ended with higher dose region. There was one forward layer that used information from the lower dose region to the higher dose region sequentially for G4RIL prediction, in which higher dose region might have higher weight during process, and vice versa for the backward layer. The bi-directional LSTM architecture thus allowed the model to integrate the role of both low dose and high dose region in predicting G4RIL. In addition, the parallel MLP input layers for the non-dosimetric data enabled the model to process patient clinical profiles and other treatment information independent of the pipeline of DVH parameters before integrating information from both channels. Finally, the proposed deep learning framework with multiple data-input processing pipelines may be applicable to predictions of other RT induced toxicities beyond G4RIL prediction. The predicted risks from modeling of other toxicities could then be included in a comprehensive evaluation of radiation treatment plans.

Our study had a few limitations. First, we applied a set of strict inclusion and exclusion criteria to the study cohort, which might decrease the external generalizability of the analyses. Although we tried to mitigate this issue by splitting the data into a training and testing set, it would be ideal to use data from a different institution to validate our approach. Although the 721 observations might be sufficient for statistical modeling, this is not an ideal sample size for a deep learning model. However, the primary goal of the study was to validate the rationale of developing a two-input channel deep learning model to improve the predictive power for RT-induced toxicities and to evaluate the entire set of DVHs. The robustness of the proposed hybrid deep learning model needs to be further evaluated using much larger data in the future. Secondly, we developed a novel and efficient approach to maximize the value of the entire set of DVHs in predicting RT-induced lymphopenia; however, DVH parameters obscure spatial information, such as the exact location in irradiated anatomical regions. Therefore, our next step is to apply a similar deep learning infrastructure with a multiple-input pipeline to the data supplemented with location information. The feasibility of using this method to study hepatobiliary toxicities was validated by a study in 2018 (Ibragimov *et al* 2018). We expect such a method to be transferrable to the prediction of RT-induced lymphopenia and to potentially advance our model's predictive power to allow us to study the effect of irradiated regions on RT-induced toxicities.

Our findings emphasized the important role of radiation exposure to OARs in esophageal cancer in predicting the risk of G4RIL. However, these results were not sufficient to provide a whole picture of the effects of RT on the lymphatic systems. In studies of patients with glioma, cranial irradiation was also associated with lymphopenia, indicating that radiation exposure to not only specific organs but also circulating lymphocytes is associated with RT-induced lymphopenia (Huang *et al* 2015).

5. Conclusion

In conclusion, our work demonstrated the use of DVH parameters for predicting RT-induced lymphopenia. The analyses showed a superior performance from the proposed deep learning model regarding predicting RT induced G4RIL. Our work also validated non-lymphatic OARs' potential critical role in RT-induced lymphopenia in addition to the primary lymphoid organs such as the spleen. Our proposed deep learning framework is flexible and transferrable to other related RT-induced toxicities.

Acknowledgments

We would like to thank Erica Goodoff and Scientific Publications Services team at the University of Texas MD Anderson Cancer Center for providing constructive editing suggestions for the manuscript.

Conflict of interest statement

None declared.

Funding

This work was supported by National Cancer Institute U19-CA021239.

Author contributions

CZ designed the deep learning model and statistical analyses plan, implemented the analyses, interpreted results and drafted the manuscript. SL, XJ, YX revised the analyses plan, interpreted the results, contributed critical revisions of the manuscript's intellectual content. ZB downloaded and prepared the data, performed data quality assurance and contributed critical of the manuscript's intellectual content. GJ contributed critical revisions of the manuscript's intellectual content and organized the contents. RM led the project, designed the experiment, revised the analyses plan and contributed critical revisions of the manuscript's intellectual content. All authors substantially contributed to the study, engaged in writing and approved the manuscript. All authors agreed to be accountable for all aspects of the work.

ORCID iDs

Cong Zhu https://orcid.org/0000-0001-9687-6653

References

Allison P D 2012 *Logistic Regression Using SAS: Theory and Application* (Cary, NC: SAS Institute) p 30

Burman C, Kutcher G J, Emami B and Goitein M 1991 Fitting of normal tissue tolerance data to an analytic function *Int. J. Radiat. Oncol. Biol. Phys.* **21** 123–35

Cho O, Oh Y T, Chun M, Noh O K and Lee H W 2016b Radiation-related lymphopenia as a new prognostic factor in limited-stage small cell lung cancer *Tumor Biol.* **37** 971–8

Cho O, Oh Y T, Chun M, Noh O K, Hoe J S and Kim H 2016a Minimum absolute lymphocyte count during radiotherapy as a new prognostic factor for nasopharyngeal cancer *Head & Neck* **38** E1061–7

Coates P J, Rundle J K, Lorimore S A and Wright E G 2008 Indirect macrophage responses to ionizing radiation: implications for genotype-dependent bystander signaling *Cancer Res.* **68** 450–6

Davuluri R *et al* 2017 Lymphocyte nadir and esophageal cancer survival outcomes after chemoradiation therapy *Int. J. Radiat. Oncol. Biol. Phys.* **99** 128–35

Friedman J, Hastie T and Tibshirani R 2010 Regularization paths for generalized linear models via coordinate descent *J. Stat. Softw.* **33** 1

Hernando M L *et al* 2001 Radiation-induced pulmonary toxicity: a dose-volume histogram analysis in 201 patients with lung cancer *Int. J. Radiat. Oncol. Biol. Phys.* **51** 650–9

Hochreiter S and Schmidhuber J 1997 Long short-term memory *Neural Comput.* **9** 1735–80

Hosmer D W Jr, Lemeshow S and Sturdivant R X 2013 *Applied Logistic Regression* (Hoboken, NJ: Wiley) pp 1–14

Huang J *et al* 2015 Clinical and dosimetric predictors of acute severe lymphopenia during radiation therapy and concurrent temozolomide for high-grade glioma *Int. J. Radiat. Oncol. Biol. Phys.* **92** 1000–7

Ibragimov B, Toesca D, Chang D, Yuan Y, Koong A and Xing L 2018 Development of deep neural network for individualized hepatobiliary toxicity prediction after liver SBRT *Med. Phys.* **45** 4763–74

Kingma D P and Ba J 2014 Adam: a method for stochastic optimization (arXiv:1412.6980v9)

Kitayama J, Yasuda K, Kawai K, Sunami E and Nagawa H 2010 Circulating lymphocyte number has a positive association with tumor response in neoadjuvant chemoradiotherapy for advanced rectal cancer *Radiat. Oncol.* **5** 47

Liaw A and Wiener M 2002 Classification and regression by random forest *R News* **2** 18–22

Michalski J M, Gay H, Jackson A, Tucker S L and Deasy J O 2010 Radiation dose–volume effects in radiation-induced rectal injury *Int. J. Radiat. Oncol. Biol. Phys.* **76** S123–9

Midi H, Sarkar S K and Rana S 2010 Collinearity diagnostics of binary logistic regression model *J. Interdiscip. Math.* **13** 253–67

Pedregosa F *et al* 2011 Scikit-learn: machine learning in Python *J. Mach. Learn. Res.* **12** 2825–30

Scholkopf B and Smola A J 2001 *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond* (Cambridge, MA: MIT Press)

Schuster M and Paliwal K K 1997 Bidirectional recurrent neural networks *IEEE Trans. Signal Process.* **45** 2673–81

Srivastava N, Hinton G, Krizhevsky A, Sutskever I and Salakhutdinov R 2014 Dropout: a simple way to prevent neural networks from overfitting *J. Mach. Learn. Res.* **15** 1929–58

Stratton J A, Byfield P E, Byfield J E, Small R C, Benfield J and Pilch Y 1975 A comparison of the acute effects of radiation therapy, including or excluding the thymus, on the lymphocyte subpopulations of cancer patients *J. Clin. Invest.* **56** 88–97

Tang C, Liao Z, Gomez D, Levy L, Zhuang Y, Gebremichael R A, Hong D S, Komaki R and Welsh J W 2014 Lymphopenia association with gross tumor volume and lung V5 and its effects on non-small cell lung cancer patient outcomes *Int. J. Radiat. Oncol. Biol. Phys.* **89** 1084–91

Venkatesulu B P, Mallick S, Lin S H and Krishnan S 2018 A systematic review of the influence of radiation-induced lymphopenia on survival outcomes in solid tumors *Crit. Rev. Oncol./Hematol.* **123** 42–51

Wright E G and Coates P J 2006 Untargeted effects of ionizing radiation: implications for radiation pathology *Mutation Res./Fundam. Mol. Mech. Mutagen.* **597** 119–32

Yovino S, Kleinberg L, Grossman S A, Narayanan M and Ford E 2013 The etiology of treatment-related lymphopenia in patients with malignant gliomas: modeling radiation dose to circulating lymphocytes explains clinical observations and suggests methods of modifying the impact of radiation on immune cells *Cancer Invest.* **31** 140–4

Zou H and Hastie T 2005 Regularization and variable selection via the elastic net *J. R. Stat. Soc.* B **67** 301–20