Clinical and Translational Radiation Oncology 31 (2021) 50–57

Contents lists available at ScienceDirect

Clinical and Translational Radiation Oncology

journal homepage: www.sciencedirect.com/journal/clinical-and-translational-radiation-oncology

Machine learning for dose-volume histogram based clinical decision-making support system in radiation therapy plans for brain tumors

Pawel Siciarz [a,b,*], Salem Alfaifi [c], Eric Van Uytven [c], Shrinivas Rathod [c,d], Rashmi Koul [d,e], Boyd McCurdy [d,f,b]

[a] Department of Medical Physics, CancerCare Manitoba, 675 McDermot Avenue, Winnipeg, MB R3E 0V9, Canada
[b] Department of Physics and Astronomy, University of Manitoba, Allen Building, Winnipeg, MB R3T 2N2, Canada
[c] Radiation Oncology Resident, Department of Radiation Oncology, CancerCare Manitoba, 675 McDermot Avenue, Winnipeg, MB R3E 0V9, Canada
[d] Department of Radiology, University of Manitoba, GA216-820 Sherbrook Street, Winnipeg, MB R3T 2N2, Canada
[e] Medical Director and Head, Radiation Oncology Program, Department of Radiation Oncology, CancerCare Manitoba, 675 McDermot Avenue, Winnipeg, MB R3E 0V9, Canada
[f] Head of Radiation Oncology Physics Group, Department of Medical Physics, CancerCare Manitoba, 675 McDermot Avenue, Winnipeg, MB R3E 0V9, Canada

ABSTRACT

Purpose: To create and investigate a novel, clinical decision-support system using machine learning (ML).

Methods and Materials: The ML model was developed based on 79 radiotherapy plans of brain tumor patients that were prescribed a total dose of 60 Gy delivered with volumetric-modulated arc therapy (VMAT). Structures considered for analysis included planning target volume (PTV), brainstem, cochleae, and optic chiasm. The model aimed to classify the target variable that included class-0 corresponding to plans for which the PTV treatment planning objective was met and class-1 that was associated with plans for which the PTV objective was not met due to the priority trade-off to meet one or more organs-at-risk constraints. Several models were evaluated using double-nested cross-validation and an area-under-the-curve (AUC) metric, with the highest performing one selected for further investigation. The model predictions were explained with Shapely additive explanation (SHAP) interaction values.

Results: The highest-performing model was Logistic Regression achieving an accuracy of 93.8 ± 4.1% and AUC of 0.98 ± 0.02 on the testing data. The SHAP analysis indicated that the $\Delta D_{99\%}$ metric for PTV had the greatest influence on the model predictions. The least important feature was ΔD_{MAX} for the left and right cochleae.

Conclusions: The trained model achieved satisfactory accuracy and can be used by medical physicists in a data-driven quality assurance program as well as by radiation oncologists to support their decision-making process in terms of treatment plan approval and potential plan modifications. Model explanation analysis showed that the model relies on clinically valid logic when making predictions.

Introduction

In modern radiation therapy, the main steps in the treatment planning process are well established. After the treatment plan is created it is then routinely reviewed by a clinician to ensure that the treatment objectives are met and dosimetric trade-offs, when required, are at acceptable levels[1–3]. For those plans that may require a dosimetric trade-off, this sometimes complex decision-making process could benefit from the knowledge of similar plans that were developed, approved, and successfully delivered in the past to patients[4–8]. However, in practice, the extraction, analysis, and interpretation of meaningful information from relevant historical data are very time-consuming and not achievable by radiation oncology professionals in the busy clinical environment. Machine learning (ML) helps to overcome those difficulties and can be used to assist medical physicists and physicians to make better informed, data-driven decisions in the radiation therapy process.

During recent years there has been growing interest in the application of ML models to develop quality assurance (QA) tools and support the treatment planning process. For example, Hirashima et al. and Wall et al. used XGBoost and Extra-Trees methods respectively to predict the gamma passing rate for patient specific QA results for volumetric modulated radiation therapy (VMAT) plans[9,10]. Osman et al. utilized an artificial neural network for the prediction of the MLC leaf position deviations during dynamic IMRT treatment delivery using log file data [11]. ML has also been explored in many other QA applications in medical physics[12–16].

Machine learning-based enhancement of the treatment planning process has also been of strong recent interest. Knowledge-based

[*] Corresponding author at: Department of Medical Physics, CancerCare Manitoba, 675 McDermot Avenue, Winnipeg, MB R3E 0V9, Canada.
E-mail addresses: psiciarz@cancercare.mb.ca (P. Siciarz), salfaifi@cancercare.mb.ca (S. Alfaifi), evanuytven2@cancercare.mb.ca (E.V. Uytven), srathod@cancercare.mb.ca (S. Rathod), rkoul@cancercare.mb.ca (R. Koul), bmccurdy2@cancercare.mb.ca (B. McCurdy).

https://doi.org/10.1016/j.ctro.2021.09.001
Received 4 June 2021; Received in revised form 27 August 2021; Accepted 6 September 2021
Available online 15 September 2021

planning (KBP) is a commonly studied application that leverages relevant features of previous, successfully delivered treatment plans in order to predict specific treatment planning parameters or the possible attainable dose-volume histograms (DVHs) [17,18]. KBP has been successfully used across various clinical sites such as head and neck[19,20], prostate[21,22], lung[23–25], rectum[26,27], breast[28,29], pelvis [30], and brain[31].

The purpose of this study was to apply ML techniques to create a novel decision support application which has not been investigated before. Specifically, a machine learning model was trained to classify previously delivered VMAT plans of brain tumor patients into two categories. The first category contained plans that met PTV treatment planning objectives. The second category included plans for which PTV objectives were not met due to the priority given to one or more OARs (i. e. a trade-off was required); those plans, however, were still clinically acceptable and delivered. Once trained, the ML algorithm would be able to indicate which new plans required a compromise (or not). This is a novel ML application and will have a very practical impact on the decision-making process in a clinical environment (more details about clinical use and future utility of the system are included in the Discussion section). Furthermore, our study applied double nested cross-validation for model selection and tuning as well as comprehensive global and local model explainability analysis. In the literature, numerous studies use k-fold cross-validation[32], but few apply nested cross-validation[33–35], and none that are similar to our application. Model explainability is rarely performed in applications of ML in radiation therapy but has been identified as a strong need in the research community in order to properly interpret model results[36]. The explainability analysis included in this work may be particularly valuable to both medical physics and radiation oncology professionals working in a clinical environment.

Methods and materials

Treatment plans data

This study involved 79 brain tumor patients that were prescribed a total dose of 60 Gy delivered in 30 fractions, 2 Gy per fraction using two-arc VMAT plans. The data necessary to train machine learning models were derived from dose-volume histograms and anatomical contours delineated by two experienced radiation oncologists. DVHs provided dosimetric information while segmentations provided geometric information for feature extraction. Structures considered for analysis included PTV, brainstem, left and right cochlea as well as optic chiasm.

Model inputs

Dosimetric features

Fig. 1a-b summarizes the dosimetry data and deviations from the treatment objectives for each structure. Forty-one plans met all dose objectives, while 12, 13, 9, and 4 plans did not meet one, two, three, and four dose objectives, respectively. Appendix 1 (Figure A1) shows DVHs

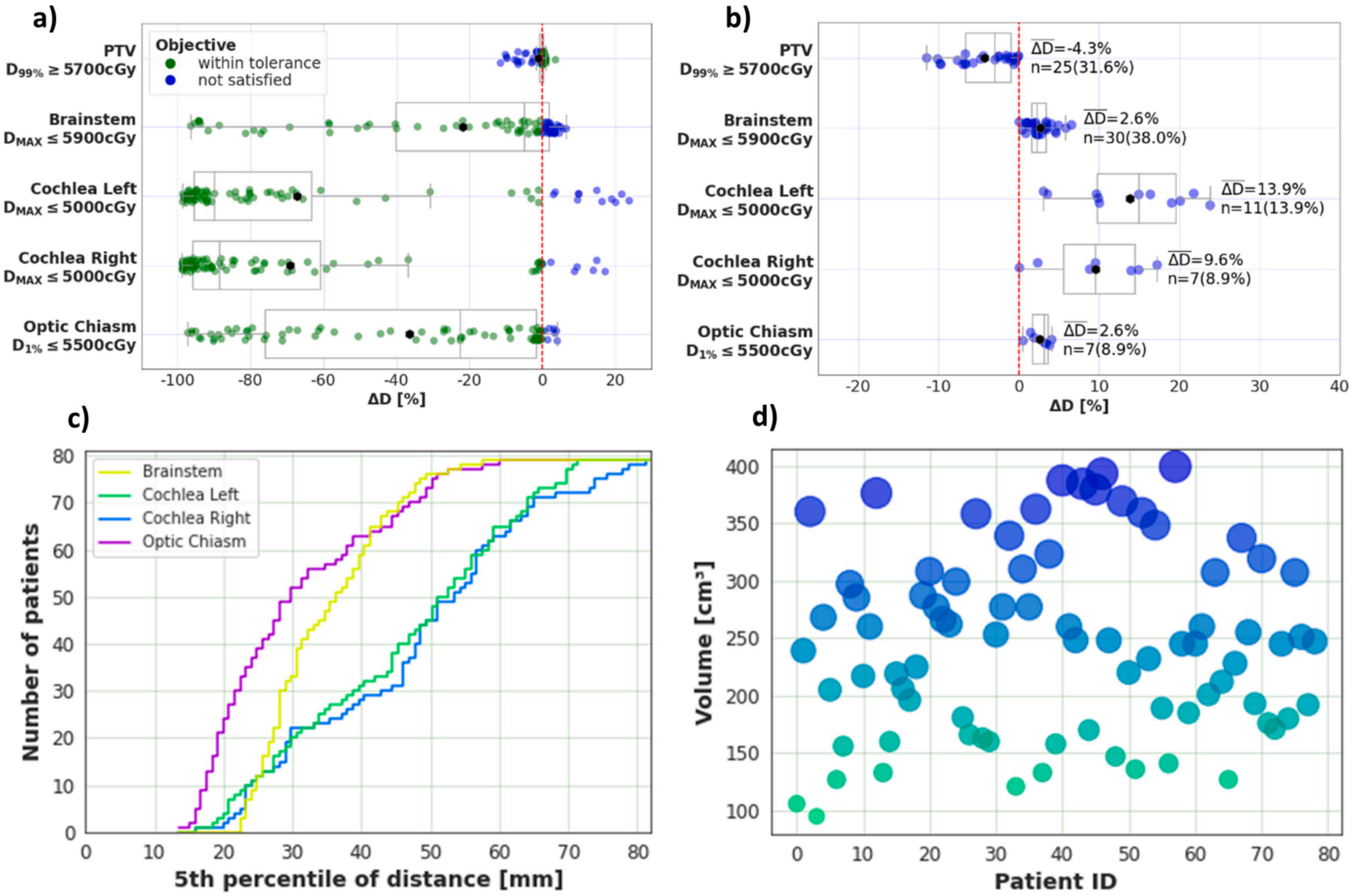

Fig. 1. **a)** Percentage deviations from treatment planning objectives ΔD for all 79 patients and associated structures, where the red dashed line indicates the boundary between positive and negative ΔD; **b)** percentage deviations from treatment planning objectives for plans which did not meet a specified objective. The percentage values of ΔD were used as dosimetric features for training the machine learning model. $\overline{\Delta D}$ correspond to the mean deviation for n plans. **c)** Cumulative distribution of the 5th percentile of Hausdorff distances between the PTV and organs-at-risk. **d)** The absolute PTV volume for each plan. The size and the colors of the markers are proportionate and correspond to the PTV volume measure. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

for all plans and structures as well as indicates plans for which a specific number of treatment planning objectives were not met.

Geometric features. The first type of geometric features included the minimum distance Δd measured in millimetres between OARs and PTV calculated by the 5th percentile of minimum Hausdorff distances[37] as shown in Fig. 1(c). The second type of geometric feature was PTV volume measured in cm^3 as presented in Fig. 1(d).

Model outputs

The models in this study were trained for the binary classification task thus the model output was represented by the binary categorical variable. The value of 'zero' corresponded to plans for which the PTV treatment planning objective was met while the value of 'one' was associated with plans for which the PTV objective was not met due to the priority trade-off to meet one or more OAR constraints. This classification was established through independent review by two radiation oncologists. The proportion of plans belonging to classes zero and one was 68.4% (54 plans) and 31.6% (25 plans) respectively. This class imbalance was not significant and was not observed to impact results.

Model and hyperparameters selection

Four models commonly used in ML were considered: Support Vector Machine Classifier, Elastic Net, Logistic Regression, and Random Forest Classifier (as implemented in the scikit-learn Python module[38]). *A priori* justification for use of particular models and their parameters is desirable from a scientific standpoint but is challenging in practice due to the large number of possible combinations of hyperparameters. Therefore, in this work, the classification algorithms were selected due to simplicity, computational efficiency, and common usage in the ML community.

Only one model was selected for further comprehensive analysis of the results on unseen testing data. The model selection process was based on the nested cross-validation that was shown to be superior relative to single cross-validation in minimizing the bias for model and hyperparameters selection as well as reducing overfitting[39,40]. The metric used for scoring the models was the area under the curve (AUC) of receiver operating characteristic (ROC)[38]. A detailed description regarding the principle of operation of double nested cross-validation is included in Figure A2 (Appendix 1).

Hyperparameters and their ranges for selected models were specified in Table A1 (Appendix 1). The search for optimal hyperparameters was performed using a grid search method that explores all the possible combinations of hyperparameters. For large data sets, it is often not a viable option, however for the sample size and number of features selected for this study, it was reasonably computationally efficient. The processor used for computations was a 2-core Intel Xeon CPU @2.3 GHz.

Model evaluation

The model with hyperparameters tuned after nested cross-validation was evaluated on the test data that accounted for 20% (15 plans) of all the data (79 plans) using single 5-fold cross-validation. For each fold, the areas under the ROC curves were reported. Additionally, the confusion matrix with true and predicted classes together with the precision, recall, and accuracy metrics (standard ROC definitions as in Pepe et al. [41]) for one of the cross-validation folds were also included in the results for more intuitive performance interpretation.

Model explainability

To better understand the predictions generated by the model we analyzed the Shapely additive explanation (SHAP) interaction values [42] for both global and local explainability. The Shapley value reflects a mean value of marginal feature contributions to the prediction and can be interpreted as a contribution to explain the difference between the average prediction of the model and the actual individual prediction.

Results

Model and hyperparameter selection

Based on the nested cross-validation, the mean and standard deviation of the ROC score was 0.9726 (±0.0059) for Support Vector Machine, 0.9986 (±0.0028) for Elastic Net, 0.9994 (±0.0012) for Logistic Regression, and 0.9979 (±0.0025) for Random Forest Classifier. The Logistic Regression model received the highest score and showed the lowest SD for inner loop evaluation metrics. The computational time of nested cross-validation for model selection was 50 s.

The hyperparameters for the best performing model (ie. Logistic Regression) also were selected with a cross-validation technique. Only two sets of hyperparameters were evaluated in this step because during model selection those two sets were associated with the best score for more than one fold. The hyperparameters were: pipeline 1 - regularization L1, C = 0.077, and liblinear solver achieving mean AUC of 0.9666 (±0.0668), and pipeline 2 - regularization L1, C = 1.668, and liblinear solver achieving mean AUC of 0.9898 (±0.0074). The computational time of nested cross-validation for hyperparameter selection was 3 s.

Model evaluation

The mean accuracy of the Logistic Regression model selected was 93.8 ± 4.1% while the mean area under the ROC curve was 0.98 ± 0.02 on the testing data. Fig. 2(a) and 2(b) shows the confusion matrix with precision, recall, and f1 scores for two classes and one (i.e. fourth) cross-validation fold.

The fourth fold was selected as an example because the resulting accuracy was approximately the same as the average accuracy of the model. The performance measures reported in Fig. 2 are similar for all the remaining folds and are included in Appendix 2 (Figure A3). Their definition and interpretation are included in Appendix 2 (Table A2). Fig. 2(c) shows the AUC values for ROC curves associated with each fold. All evaluation metrics were calculated based on the testing data.

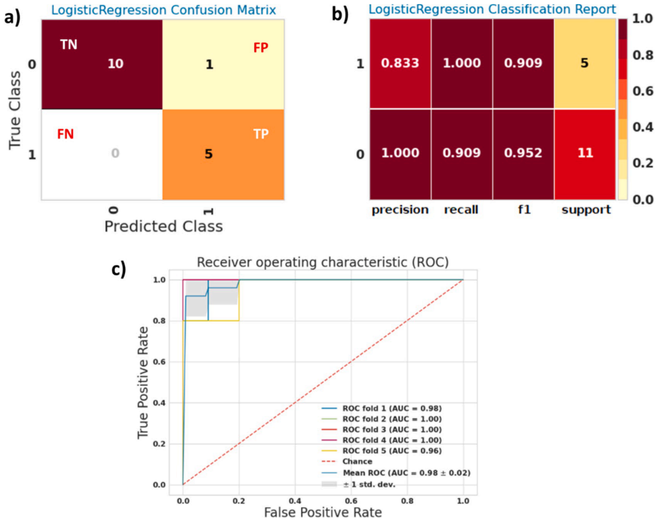

Fig. 2. Confusion matrices **a)** and Logistic Regression model evaluation metrics **b)** for fourth cross-validation fold. The meaning behind each metric was briefly summarized in Appendix 2 (Table A2); **c)** Receiver Operating Characteristic for five cross-validation folds created based on the testing data and the performance of the Logistic Regression model. The dashed diagonal line ('Chance' in the legend) represents the random assignment of classes.

Model explainability

The model explainability was addressed both globally and locally. Fig. 3 shows global explainability as an impact of each feature on the model output. Fig. 3(a) provides directional SHAP values and relative values of geometric and dosimetric model features. Fig. 3(b) shows the average impact (ie. mean SHAP) of each feature on the model output. Overall, the SHAP analysis indicates that the deviation from the $D_{99\%}$ ($\Delta D_{99\%}$) metric for PTV had the greatest influence on the model predictions and is approximately 3.5 times larger (i.e. more important) than the second most important feature – $\Delta D_{1\%}$ for optic chiasm. The least important features were percentage deviations from the maximum dose delivered to the left and right cochlea. A more in-depth interpretation of this analysis is interesting from a clinical perspective and will be discussed further in the discussion section.

It is also very practical to determine the relationship between original values (not scaled for the model training) of individual features and their global contribution to the model performance. Fig. 4 shows these relationships for both geometric and dosimetric features.

Generally, for geometric features, it is seen that the larger the distance between the PTV and the OARs and the larger the PTV volume, the lower the SHAP value. Regarding the dosimetric features, for the majority of them, there is a linear and positively correlated relationship between the feature and its contribution to the model prediction. The only exception is $\Delta D_{99\%}$ for PTV; in this case, SHAP values are linearly but negatively correlated to this feature.

The model predictions can also be interpreted locally by examining individual model predictions as shown in Fig. 5. In particular, the bar charts for each prediction show the feature importance in the form of directional SHAP values as well as the probability of the prediction belonging to class 0 and class 1 (i.e. the class predicted by the model). The determination of whether the prediction was correct (or not) is also included. It can be noticed that out of 10 model features only 7 are present in the local explainability graphs. This is because the three remaining features were not significantly contributing to these individual predictions. Appendix 3 illustrates local feature importance for all predictions.

Discussion

After executing double nested cross-validation of several models, the Logistic Regression algorithm was selected as the best performing model for further use. Double-nested cross-validation was employed instead of single-nested for model selection transparency. In terms of computational efficiency, this would also be a preferable method if the data set and the range of hyperparameters were larger. As for the Logistic Regression model, it is a relatively simple and easy-to-interpret model that can be trained and provide new predictions (model inference) quickly. The Logistic Regression model has also been of interest in recent radiation oncology research[43–46].

The model was selected based on its performance measured by the AUC, not accuracy. This is mainly because the ROC curve is insensitive to data sets with unbalanced classes and additionally reflects the classifier's performance for all values of the discrimination threshold. These characteristics make AUC a preferable metric in the evaluation of ML models[45–48]. However because our study included imbalanced data we have also provided precision-recall curves in Appendix 2. The model did not overfit the data because the model performance on testing data for each cross-validation fold is both satisfactory and consistent.

The SHAP analysis results presented in Fig. 3 requires further discussion. First, negative SHAP values do not mean that the feature importance is smaller than for positive SHAP values. Rather, SHAP values below zero drive the prediction towards class 0 while positive SHAP values drive the prediction towards class 1. This characteristic combined with the color-coded values of the particular feature delivers interesting model interpretations. For example, in the case of $\Delta D_{99\%}$ for PTV, it can be seen that what drives the predictions towards the class 0 are high values of $\Delta D_{99\%}$ that, if we look at Fig. 1, correspond to a high probability of this treatment planning objective being met. In this scenario, the fact that class 0 is associated with plans for which the PTV objective was also met, shows that the model interpretation is consistent with the clinical interpretation. Another example includes the geometric features i.e. Fig. 3(a) illustrating that larger distances Δd between PTV and OARs tend to drive model predictions towards the acceptable plans (class 0) as well. This is also a very common observation in clinical practice because with larger Δd, it is easier to create a treatment plan that would provide desirable PTV coverage and simultaneously spare OARs.

It is also interesting to note that the best performing model is logistic regression, which is less complex than SVM. We believe that it is most likely due to the simplicity of our classification task (binary classification) and because of the presence of the feature(s) with strong predictive power as shown in the global explainability chart (Fig. 3), that indicates the $\Delta D_{99\%}$ metric for PTV as the feature that contributes to the model outputs the most significantly. This observation confirms an important characteristic of ML, namely that an increase in model complexity does not always lead to an increase in model performance.

Additionally Fig. 4(a) shows that the smaller the distance between PTV and right cochlea, the higher the probability of a high dose being delivered to that organ. The same relationship can be observed between PTV-to-brainstem distance and $\Delta D_{1\%}$ for optic chiasm. By examining the partial dependency charts for dosimetric features in Fig. 4(b) it can be seen that the strongest mutual interaction of $\Delta D_{1\%}$ for optic chiasm exists with PTV-brainstem distance. The most important model feature,

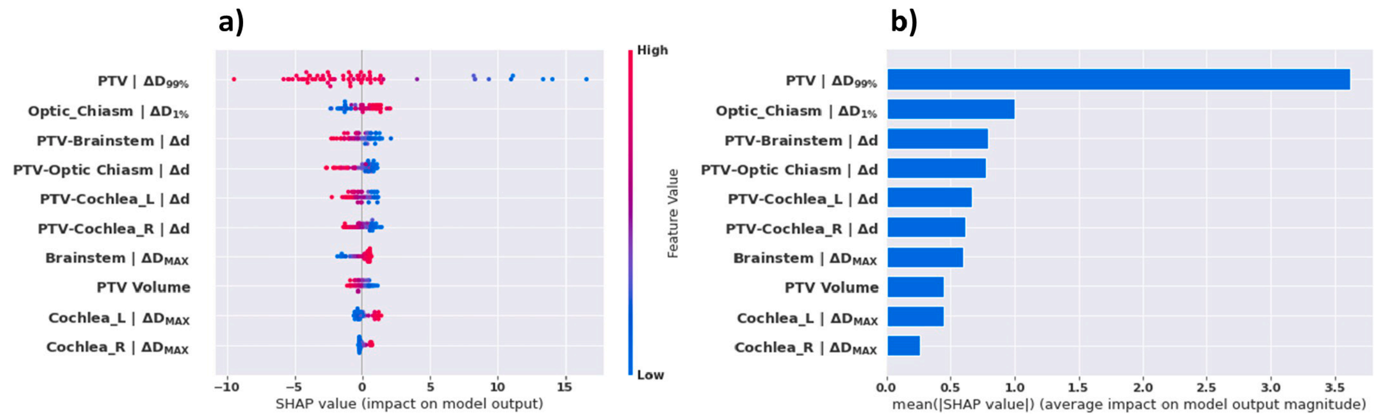

Fig. 3. **a)** Feature importance represented by the impact of directional SHAP values and particular feature values on the output of the model; **b)** The average feature contribution to the model output measured by mean absolute SHAP values.

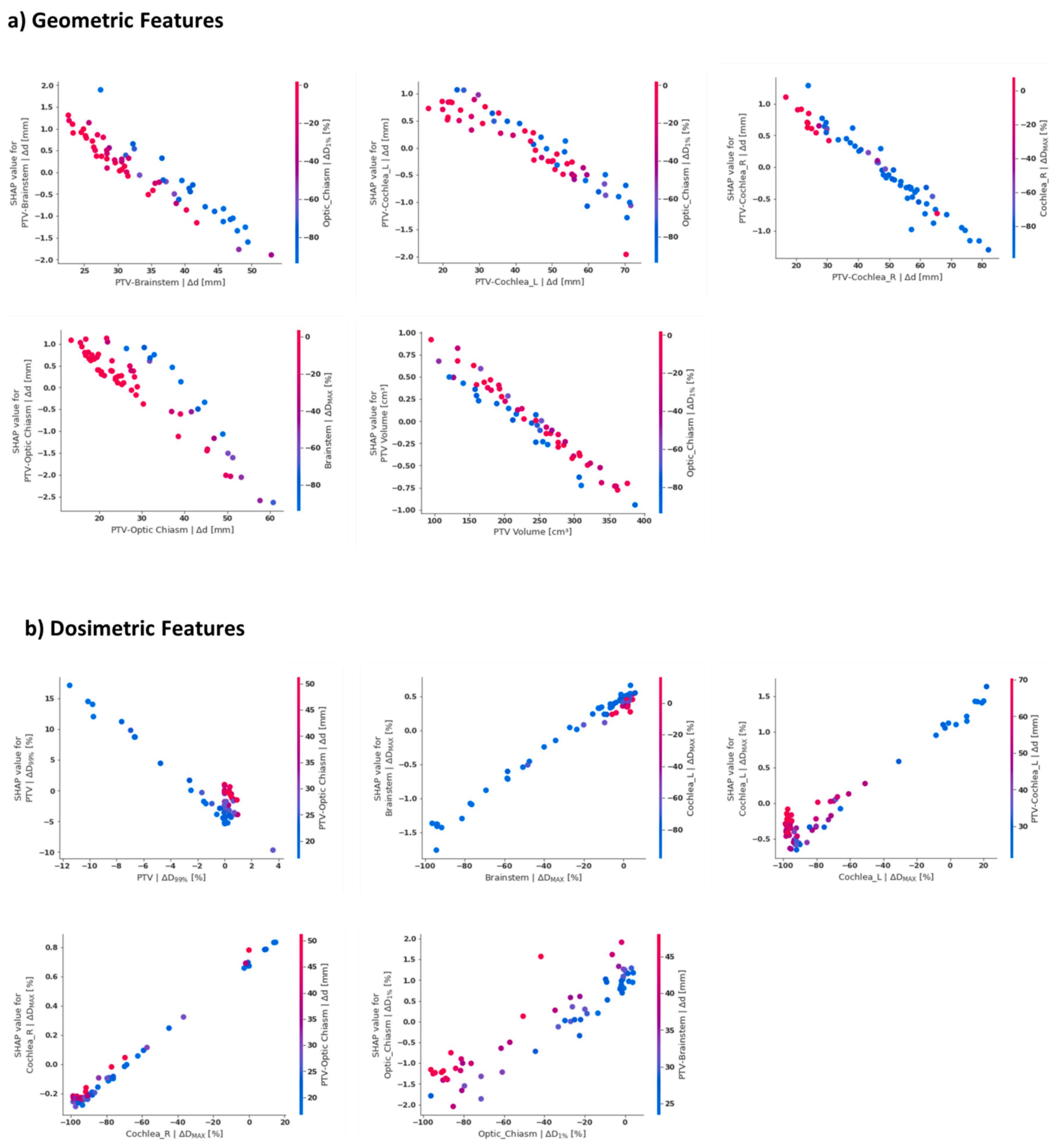

Fig. 4. Partial dependency charts for **a)** geometric features and **b)** dosimetric features. The color bars associated with each chart indicate the feature with which the evaluated feature (on the x-axis) has the strongest interaction. Specifically, the interaction indicates the influence those two features have on the model prediction. For example, if we consider a feature on the x-axis, then another feature on the color bar will be automatically selected in order to maximize the mutual impact of those two features on the model prediction. Partial dependency charts for all cross-validation folds are included in Appendix 3.

$\Delta D_{99\%}$ for PTV, had the strongest interaction with distance to optic chiasm, demonstrating that if PTV is located in the proximity of optic chiasm there is a low probability of meeting the PTV dose objectives. This is consistent with observation in clinical practice where the PTV coverage trade-off needs to be made in order to meet treatment objectives for critical structures.

The proposed ML classifier and model explainability work together to provide additional value to the clinical processes. After the plan is created and the algorithm classifies the treatment plan, the model explainability analysis (performed instantly) indicates the attributes behind the plan classification (i.e. which plan and patient-related attributes caused the plan to require trade-off). Therefore the clinician would not have to analyse the treatment plan and/or schedule a consultation with the treatment planner/dosimetrist or other radiation

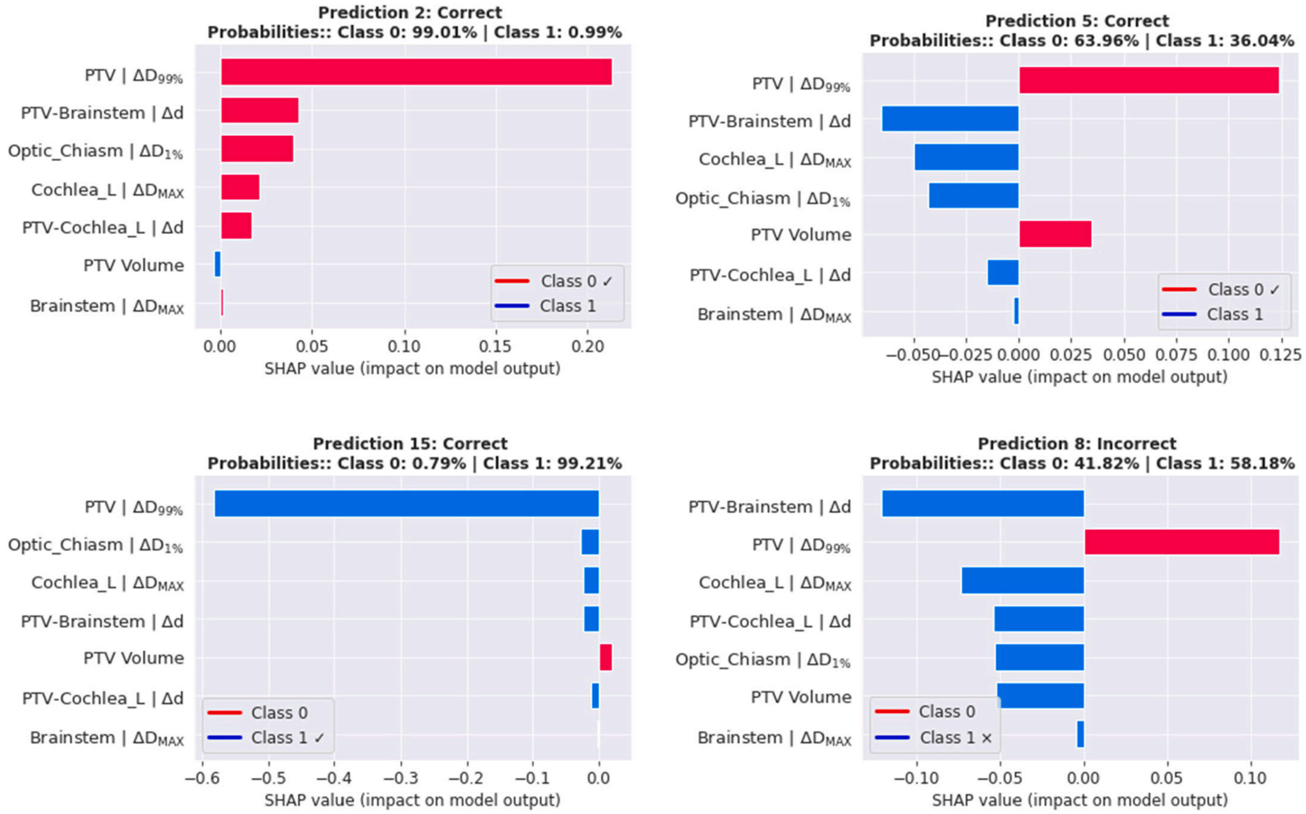

Fig. 5. Local explanations of the logistic regression model in the form of features importance and, generated by the model, binary class probabilities. For illustrative purposes, two randomly selected samples belonging to each class (including one incorrectly classified sample) were selected. The threshold for the class assignment was determined by the probability of 50%.

oncologists to find the cause. This opens up an opportunity of developing an automated notification system for clinicians (not explored in our study but possible for future work).

Furthermore for challenging plans (those with trade-offs), it is found to be useful both in our clinic as well as in the recent literature, to look at similar past plans, and make data-driven decisions regarding further steps in treatment planning (e.g. plan modifications). A new plan, once classified as a plan with trade-offs in the proposed ML system, is automatically compared to similar plans delivered in the past. The clinician does not have to search through the clinical database to find similar plans.

Potential future applications of such an ML system could expand beyond a single institution. One can envision smaller radiation therapy centers sending challenging plans (ie. those with tradeoffs) to a large experienced center for ML analysis, and thus providing an invaluable planning QA tool. The large center could incorporate the new plans in the ML training set and refine their model (ie. continual learning). This data and technology-sharing environment could potentially equalize the standard of care regardless of the resource availability of a given medical facility. This example can intuitively be expanded to broad collaboration between many large and small institutions across the world.

Another potential clinical impact of the presented system is for management of adaptive radiation therapy. Specifically, patients with dosimetric compromises (tradeoffs) might justify more accurate dose delivery and positioning. Therefore these patients could be identified as high priority for Adaptive Radiation Therapy (ART).

This study has two limitations. The first is the limited number of plan datasets, which may impact the model robustness. This is mainly due to the limitations of data availability for local brain tumor plans qualified for this study (i.e. total dose delivered, fractionation, delivery technique). Data availability is a common problem in radiation therapy studies involving the application of ML. There are many papers where ML models are trained using $<$ 150 and as little as 11 patients [9,31,49,50]. A second limitation of this study is the simplified, binary classification of plans. At the design stage of the study, we found that plans, where the priority trade-offs were made for OARs, could have been additionally divided into plans with higher and lower priority trade-offs. However, our relatively small data set would cause those two potential classes to be significantly under-sampled, therefore ultimately the classification of trade-off priorities was not pursued in this study. Both these limitations could however be addressed in future work. Additionally, a possible subsequent study could also involve testing the model using unapproved plans to further evaluate a model performance.

Conclusions

The trained ML model achieved satisfactory accuracy on the test data and can be used by medical physicists in a data-driven quality assurance program as well as by radiation oncologists to support their decision-making process in terms of treatment plan approval and potential plan modifications. Model explainability analysis facilitated a better understanding of the machine learning model reasoning for the generated predictions and showed consistency with clinical observations.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to acknowledge the financial support of the Natural Sciences and Engineering Research Council of Canada

(Discovery Grant) as well as the University of Manitoba.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ctro.2021.09.001.

References

[1] Chao KSC. Practical Essentials of Intensity Modulated Radiation Therapy. Wolters Kluwer Health; 2013.

[2] Gaya A, Mahadevan A. Stereotactic Body Radiotherapy: A Practical Guide. Springer London; 2015.

[3] Xia P, Godley A, Shah C, Gregory M. M. Videtic MDCMF, Suh J. Strategies for Radiation Therapy Treatment Planning. Springer Publishing Company; 2018.

[4] Li N, Carmona R, Sirak I, et al. Highly efficient training, refinement, and validation of a knowledge-based planning quality-control system for radiation therapy clinical trials. International Journal of Radiation Oncology* Biology* Physics. 2017;97(1): 164-172.

[5] Wang J, Hu W, Yang Z, Chen X, Wu Z, Yu X, et al. Is it possible for knowledge-based planning to improve intensity modulated radiation therapy plan quality for planners with different planning experiences in left-sided breast cancer patients? Radiation Oncol 2017;12(1). https://doi.org/10.1186/s13014-017-0822-z.

[6] Scaggion A, Fusella M, Roggio A, Bacco S, Pivato N, Rossato MA, et al. Reducing inter-and intra-planner variability in radiotherapy plan output with a commercial knowledge-based planning solution. Physica Med 2018;53:86–93.

[7] Kamima T, Ueda Y, Fukunaga J-I, Shimizu Y, Tamura M, Ishikawa K, et al. Multi-institutional evaluation of knowledge-based planning performance of volumetric modulated arc therapy (VMAT) for head and neck cancer. Physica Med 2019;64: 174–81.

[8] Yu S, Xu H, Sinclair A, Zhang X, Langner U, Mak K. Dosimetric and planning efficiency comparison for lung SBRT: CyberKnife vs VMAT vs knowledge-based VMAT. Med Dosim 2020;45(4):346–51.

[9] Wall PDH, Fontenot JD. Application and comparison of machine learning models for predicting quality assurance outcomes in radiation therapy treatment planning. Inf Med Unlocked 2020;18:100292. https://doi.org/10.1016/j.imu.2020.100292.

[10] Hirashima H, Ono T, Nakamura M, Miyabe Y, Mukumoto N, Iramina H, et al. Improvement of prediction and classification performance for gamma passing rate by using plan complexity and dosiomics features. Radiother Oncol 2020;153: 250–7.

[11] Osman AFI, Maalej NM, Jayesh K. Prediction of the individual multileaf collimator positional deviations during dynamic IMRT delivery priori with artificial neural network. Med Phys 2020;47(4):1421–30.

[12] Li Q, Chan MF. Predictive time-series modeling using artificial neural networks for Linac beam symmetry: an empirical study. Ann N Y Acad Sci 2017;1387(1):84.

[13] Guidi G, Maffei N, Meduri B, D'Angelo E, Mistretta GM, Ceroni P, et al. A machine learning tool for re-planning and adaptive RT: a multicenter cohort investigation. Phys Med 2016;32(12):1659–66.

[14] Carlson JNK, Park JM, Park S-Y, Park JI, Choi Y, Ye S-J. A machine learning approach to the accurate prediction of multi-leaf collimator positional errors. Phys Med Biol 2016;61(6):2514–31.

[15] Grewal HS, Chacko MS, Ahmad S, Jin H. Prediction of the output factor using machine and deep learning approach in uniform scanning proton therapy. J Appl Clinic Med. Phys. 2020;21(7):128–34.

[16] Valdes G, Scheuermann R, Hung CY, Olszanski A, Bellerive M, Solberg TD. A mathematical framework for virtual IMRT QA using machine learning. Med Phys 2016;43(7):4323–34.

[17] C. Wang X. Zhu J.C. Hong D. Zheng Artificial intelligence in radiotherapy treatment planning: present and future Technology in cancer research & treatment. 18 2019 153303381987392 10.1177/1533033819873922.

[18] Tambe NS, Pires IM, Moore C, Cawthorne C, Beavis AW. Validation of in-house knowledge-based planning model for advance-stage lung cancer patients treated using VMAT radiotherapy. British J Radiol 2020;93(1106):20190535. https://doi.org/10.1259/bjr.20190535.

[19] Fogliata A, Cozzi L, Reggiori G, Stravato A, Lobefalo F, Franzese C, et al. RapidPlan knowledge based planning: iterative learning process and model ability to steer planning strategies. Radiation Oncol 2019;14(1). https://doi.org/10.1186/s13014-019-1403-0.

[20] Zhang J, Ge Y, Sheng Y, et al. Knowledge-Based Tradeoff Hyperplanes for Head and Neck Treatment Planning. International Journal of Radiation Oncology* Biology* Physics. 2020.

[21] Chatterjee A, Serban M, Faria S, Souhami L, Cury F, Seuntjens J. Novel knowledge-based treatment planning model for hypofractionated radiotherapy of prostate cancer patients. Phys. Med 2020;69:36–43.

[22] van Schie MA, Janssen TM, Eekhout D, et al. Knowledge-based assessment of focal dose escalation treatment plans in prostate cancer. International Journal of Radiation Oncology* Biology* Physics. 2020.

[23] Delaney AR, Dahele M, Tol JP, Slotman BJ, Verbakel WFAR. Knowledge-based planning for stereotactic radiotherapy of peripheral early-stage lung cancer. Acta Oncol 2017;56(3):490–5.

[24] Teichert K, Currie G, Küfer K-H, Miguel-Chumacero E, Süss P, Walczak M, et al. Targeted multi-criteria optimisation in IMRT planning supplemented by knowledge based model creation. Operat Res. Health Care. 2019;23:100185. https://doi.org/10.1016/j.orhc.2019.04.003.

[25] van't Hof S, Delaney AR, Tekatli H, et al. Knowledge-based planning for identifying high-risk stereotactic ablative radiation therapy treatment plans for lung tumors larger than 5 cm. International Journal of Radiation Oncology* Biology* Physics. 2019;103(1):259-267.

[26] Shepherd Meegan, Bromley Regina, Stevens Mark, Morgia Marita, Kneebone Andrew, Hruby George, et al. Developing knowledge-based planning for gynaecological and rectal cancers: a clinical validation of RapidPlan™. J Med Radiat Sci 2020;67(3):217–24.

[27] Wang Meijiao, Li Sha, Huang Yuliang, Yue Haizhen, Li Tian, Wu Hao, et al. An interactive plan and model evolution method for knowledge-based pelvic VMAT planning. J Appl Clinic Med Phys 2018;19(5):491–8.

[28] Fan Jiawei, Wang Jiazhou, Zhang Zhen, Hu Weigang. Iterative dataset optimization in automated planning: Implementation for breast and rectal cancer radiotherapy. Med Phys 2017;44(6):2515–31.

[29] Rice Aubrie, Zoller Ian, Kocos Kevin, Weller Dannyl, DiCostanzo Dominic, Hunzeker Ashley, et al. The implementation of RapidPlan in predicting deep inspiration breath-hold candidates with left-sided breast cancer. Med Dosim 2019; 44(3):210–8.

[30] Kubo Kazuki, Monzen Hajime, Ishii Kentaro, Tamura Mikoto, Nakasaka Yuta, Kusawake Masayuki, et al. Inter-planner variation in treatment-plan quality of plans created with a knowledge-based treatment planning system. Phys Med 2019; 67:132–40.

[31] Kishi Noriko, Nakamura Mitsuhiro, Hirashima Hideaki, Mukumoto Nobutaka, Takehana Keiichi, Uto Megumi, et al. Validation of the clinical applicability of knowledge-based planning models in single-isocenter volumetric-modulated arc therapy for multiple brain metastases. J Appl Clinic Med Phys 2020;21(10): 141–50.

[32] Chan MF, Witztum A, Valdes G. Integration of AI and Machine Learning in Radiotherapy QA. Front Art Intell. 2020;3:76.

[33] Deist Timo M, Dankers Frank JWM, Valdes Gilmer, Wijsman Robin, Hsu I-Chow, Oberije Cary, et al. Machine learning algorithms for outcome prediction in (chemo) radiotherapy: An empirical comparison of classifiers. Med Phys 2018;45(7): 3449–59.

[34] Luo Yi, McShan Daniel, Ray Dipankar, Matuszak Martha, Jolly Shruti, Lawrence Theodore, et al. Development of a fully cross-validated bayesian network approach for local control prediction in lung cancer. IEEE Trans Rad Plasma Med Sci 2019;3(2):232–41.

[35] Yu Ting-ting, Lam Sai-kit, To Lok-hang, Tse Ka-yan, Cheng Nong-yi, Fan Yeuk-nam, et al. Pretreatment prediction of adaptive radiation therapy eligibility using MRI-based radiomics for advanced nasopharyngeal carcinoma patients. Front Oncol 2019;9. https://doi.org/10.3389/fonc.2019.0105010.3389/fonc.2019.01050.s001.

[36] Gilpin, Leilani H., David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. "Explaining explanations: An overview of interpretability of machine learning." In 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA), pp. 80-89. IEEE, 2018.

[37] Hausdorff F. Set Theory. Chelsea Publ.; 1991.

[38] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–30.

[39] Krstajic D, Buturovic LJ, Leahy DE, Thomas S. Cross-validation pitfalls when selecting and assessing regression and classification models. J Cheminf 2014;6(1): 1–15.

[40] Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinf 2006;7(1):91.

[41] M.S. Pepe The Statistical Evaluation of Medical Tests for Classification and Prediction 2003 OUP Oxford.

[42] Lundberg Scott M, Erion Gabriel, Chen Hugh, DeGrave Alex, Prutkin Jordan M, Nair Bala, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2(1):56–67.

[43] Boutilier Justin J, Lee Taewoo, Craig Tim, Sharpe Michael B, Chan Timothy CY. Models for predicting objective function weights in prostate cancer IMRT. Med Phys 2015;42(4):1586–95.

[44] Cooper Benjamin T, Li Xiaochun, Shin Samuel M, Modrek Aram S, Hsu Howard C, DeWyngaert JK, et al. Preplanning prediction of the left anterior descending artery maximum dose based on patient, dosimetric, and treatment planning parameters. Adv Rad Oncol 2016;1(4):373–81.

[45] Gabryś HS, Buettner F, Sterzing F, Hauswald H, Bangert M. Design and selection of machine learning methods using radiomics and dosiomics for normal tissue complication probability modeling of xerostomia. Front Oncol 2018;8:35.

[46] Arimura H, Soufi M, Kamezawa H, Ninomiya K, Yamada M. Radiomics with artificial intelligence for precision medicine in radiation therapy. J Rad Res 2019; 60(1):150–7.

[47] Huang J, Ling CX. Using AUC and accuracy in evaluating learning algorithms. IEEE Trans Knowl Data Eng 2005;17(3):299–310.

[48] Hackeling G. Mastering Machine Learning with scikit-learn. Packt Publishing Ltd 2017.

[49] Lee Taewoo, Hammad Muhannad, Chan Timothy CY, Craig Tim, Sharpe Michael B. Predicting objective function weights from patient anatomy in prostate IMRT treatment planning. Med Phys 2013;40(12):121706. https://doi.org/10.1118/1.4828841.

[50] Carolin Schubert Oliver Waletzko Christian Weiss Dirk Voelzke Sevda Toperim Arnd Roeser et al. Intercenter validation of a knowledge based model for automated planning of volumetric modulated arc therapy for prostate cancer. The

experience of the German RapidPlan Consortium PLoS One. 12 5 2017;12(5): e0178034. e0178034.