**RESEARCH**

# Integrating plan complexity and dosiomics features with deep learning in patient-specific quality assurance for volumetric modulated arc therapy

Ce Han[1†], Ji Zhang[1†], Bing Yu[1], Haoze Zheng[1], Yibo Wu[1], Zhixi Lin[1], Boda Ning[1], Jinling Yi[1], Congying Xie[1,2*] and Xiance Jin[1,3*]

## Abstract

**Purpose** To investigate the feasibility and performance of deep learning (DL) models combined with plan complexity (PC) and dosiomics features in the patient-specific quality assurance (PSQA) for patients underwent volumetric modulated arc therapy (VMAT).

**Methods** Total of 201 VMAT plans with measured PSQA results were retrospectively enrolled and divided into training and testing sets randomly at 7:3. PC metrics were calculated using house-built algorithm based on Matlab. Dosiomics features were extracted and selected using Random Forest (RF) from planning target volume (PTV) and overlap regions with 3D dose distributions. The top 50 dosiomics and 5 PC features were selected based on feature importance screening. A DL DenseNet was adapted and trained for the PSQA prediction.

**Results** The measured average gamma passing rate (GPR) of these VMAT plans was 97.94% ± 1.87%, 94.33% ± 3.22%, and 87.27% ± 4.81% at the criteria of 3%/3 mm, 3%/2 mm, and 2%/2 mm, respectively. Models with PC features alone demonstrated the lowest area under curve (AUC). The AUC and sensitivity of PC and dosiomics (D) combined model at 2%/2 mm were 0.915 and 0.833, respectively. The AUCs of DL models were improved from 0.943, 0.849, 0.841 to 0.948, 0.890, 0.942 in the combined models (PC + D + DL) at 3%/3 mm, 3%/2 mm and 2%/2 mm, respectively. A best AUC of 0.942 with a sensitivity, specificity and accuracy of 100%, 81.8%, and 83.6% was achieved with combined model (PC + D + DL) at 2%/2 mm.

**Conclusions** Integrating DL with dosiomics and PC metrics is promising in the prediction of GPRs in PSQA for patients underwent VMAT.

**Keywords** Deep learning, Dosiomics feature, Plan complexity, Classification, Patient-specific quality assurance

---

[†]Ce Han and Ji Zhang have authors contribute equally.

*Correspondence:
Congying Xie
wzxiecongying@163.com
Xiance Jin
jinxc1979@hotmail.com
Full list of author information is available at the end of the article

Han *et al. Radiation Oncology*　　(2023) 18:116

Page 2 of 9

## Introduction

Due to the inverse nature of intensity-modulated radiotherapy (IMRT) and volumetric modulated arc therapy (VMAT) planning, patients-specific quality assurance (PSQA) is an imperative step to detect potential errors resulted from an inaccurate dose calculation, a failure of the record-and-verify system, or delivery errors in the linear accelerator to ensure the accuracy of IMRT/VMAT delivery [1, 2]. Typically, PSQA is performed by measuring the radiation dose of IMRT/VMAT plans with 2D or 3D diode arrays and then comparing measured dosimetric distribution with planned one using a gamma passing rate (GPR). However, this traditional PSQA increases the overall clinical workload and usage of resources [3, 4]. Traditional PSQA also hinders the application of online adaptive radiotherapy, which requires a fast real time treatment planning and QA process [5, 6].

More sophisticated independent 3D dose calculation algorithms, such as convolution-superposition or Monte Carlo were introduced to verify the IMRT/VMAT plans virtually [7, 8]. On the other hand, studies demonstrated that treatment plan complexity (PC) and Linac performance metrics will influence the radiation therapy delivery [9]. PC metrics of modulation complexity score (MCS), leaf motion constraints, average leaf travel (LT), MCS applied to VMAT (MCSv), etc., had been investigated to assess the relation between overall PC and PSQA results [10, 11]. With the emerging and application of machine learning (ML) and deep learning (DL), more straightforward, less resource-intensive, efficient PSQA methods using treatment PC metrics and/or linac performance metrics were proposed to predicted the GPRs directly [12–15]. However, only weak correlations were reported between passing rates and these metrics as different aspects of the complexity of the plans might interact each other and associate with the failing of PSQA [16, 17].

Recently, radiomics with quantitative extracted image features had been applied to predict simulated radiotherapy errors for PSQA [18]. Gamma images resulted from IMRT plannar dose QA were evaluated to classify the presence or absence of introduced radiotherapy treatment delivery errors with convolutional neural networks (CNN), which indicates radiomic quality assurance is a promising direction for clinical radiotherapy [19]. Radiomics features extracted from dosimetric distribution (dosiomics) had been suggested to combine with PC metrics to improve the prediction and classification performance for GPR with ML [20]. Studies demonstrated that combining DL with radiomics through information fusion is able to improve the prediction ability of models [21, 22]. The purpose of this study is to investigate the feasibility and performance of DL integrated with

dosiomics and PC features in the PSQA for patients underwent VMAT.

## Materials and methods

### Study design

Figure 1 shows the flowchart for the overall study design, which consists of four-steps: (A) collection of the radiotherapy (RT) files, including RTplan, RTdose, RTstructure and RTimages, and corresponding PSQA data from each VMAT plan; (B) extraction of the complexity features of the plans, 3D dosiomics features of planning target volume (PTV) and overlapping region; (C) feature selection and modeling; (D) model evaluation and comparison.

### Patients and PSQA data

Patients underwent two-arc VMAT with measured PSQA results were retrospectively reviewed and enrolled in this study. VMAT plans were generated by commercial treatment planning system (TPS) (Monaco 5.1.1; Elekta, Crawley, UK) for a 6-MV X-ray beam with a dose grid size of $3.0 \times 3.0$ mm. Detailed optimization parameters and procedures had been reported previously [23, 24]. The PSQA measurements were conducted using a 3D diode array ArcCHECK (Model 1220) and SNC Patient (v.6.2.1; Sun Nuclear Corporation) with Elekta Synergy linac (Elekta Ltd, Crawley, UK), which was equipped with an 80-leaf multileaf collimator (MLCi2TM, Elekta Ltd, Crawley, UK). GPRs of three different acceptance criteria: 3%/3 mm, 3%/2 mm and 2%/2 mm with a 10% lower dose threshold were calculated and recorded [25, 26].

### Complexity metrics and dosiomics features

House-built algorithm based on Matlab 2016a (Mathwork Inc., USA) was built to read and calculated complexity metrics from exported DICOM-RT files from TPS, which includes RTplan, RTstructure, RTdose, RTimage etc. A total of 13 PC metrics were calculated, which includes monitor units (MUs), MU per control point (MU/CP), the proportion of CPs with MU < 3 (%MU/CP < 3), small segment area per CP (SA/CP), the percentage of CPs with segment area < 5 × 5 cm$^2$ (%SA < 5 × 5 cm$^2$), modulation complexity score of VMAT per arc (MCSv/Arc), leaf travel (LT) distance, Gantry spacing, etc., as reported in a previous study [11].

Dosiomics features were extracted from PTV and overlap regions (PTV overlapped with organs at risk) with 3D dose distributions using the PyRadiomics package (version 2.1.2) of Python (version 3.8) [20]. Figure 2 shows a typical PTV and overlap regions of a cervical cancer patient with 3D dose distribution for radiomics feature extraction, which contains the shape, image statistical values, and heterogeneity of the dose distribution. All the dose images were resampled to a
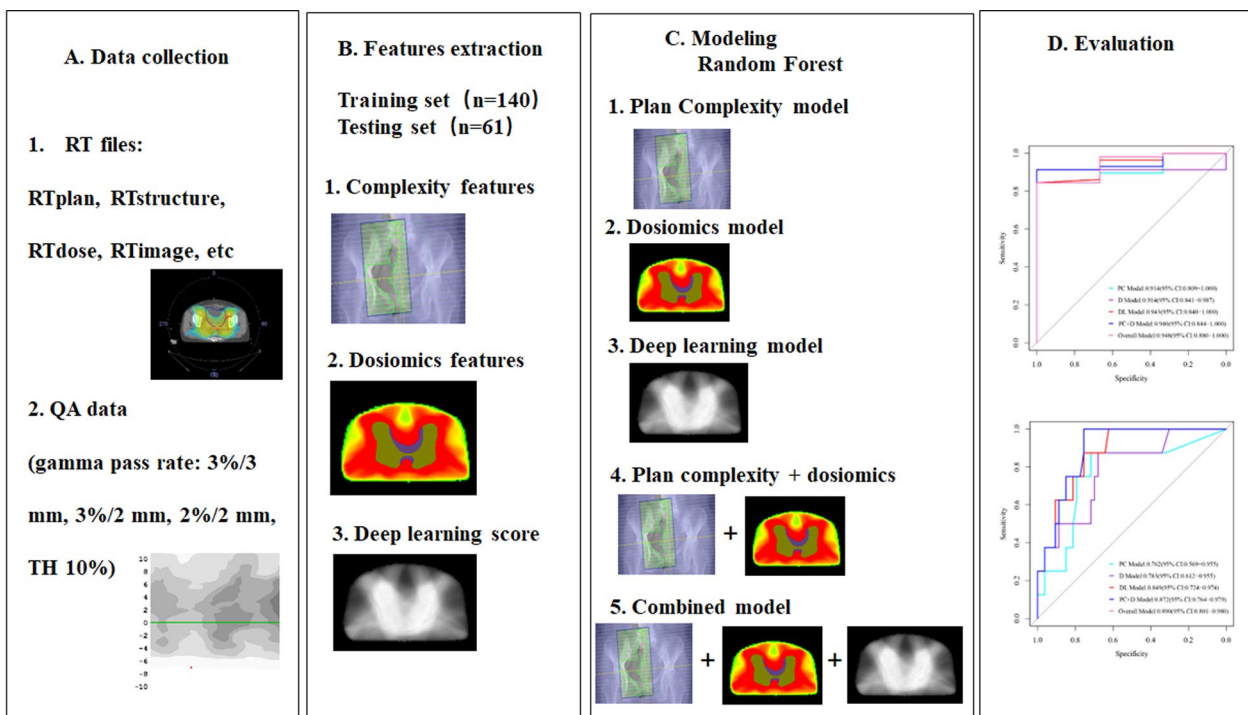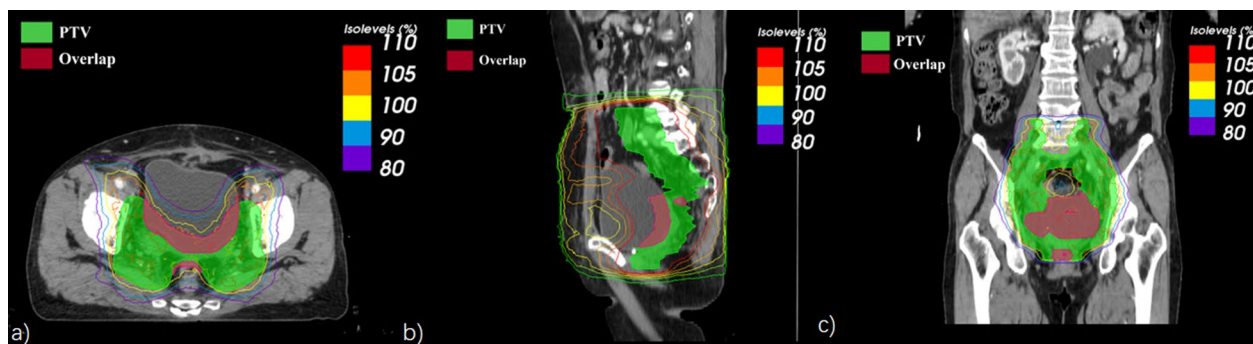
Han *et al. Radiation Oncology*      (2023) 18:116

Page 3 of 9



**Fig. 1** The flowchart for the overall study design



**Fig. 2** A typical PTV and overlap regions of a cervical cancer patient with 3D dose distribution. **a** horizontal; **b** sagittal; **c** coronal

pixel spacing of $1 \times 1 \times 1 \text{mm}^3$ with B-spline interpolation algorithm to standardized feature computation. The pixel values were discretized into equally spaced bins using a fixed bin width of 25 Hounsfield Units to eliminate the influence of different grayscale ranges and ensure better comparability. A total of 833 features were extracted, which includes 105 first-order features, and 728 s- and higher-order features of shape, gray level co-occurrence matrix (GLCM), gray level run length matrix (GLRLM), gray level size zone matrix (GLSZM), gray level dependence matrix (GLDM) and neighboring gray tone difference matrix (NGTDM) according to

the image biomarker standardization initiative (IBSI) reporting guidelines [27].

### Feature selection and modeling

The data was randomly divided into a training and testing set at a ratio of 7:3. The Random Forest (RF) algorithm was applied to select the top 50 dosiomics features based on the mean decrease accuracy in the training set [28]. Then a total of 50 dosiomics features from the PTV and overlap regions and 5 complexity features were selected in the training cohort according to the feature importance screening and applied for the construction

Han *et al. Radiation Oncology*    (2023) 18:116

Page 4 of 9

of signature via RF. During modeling, a GPR higher than 95% at 3%/3 mm, higher than 90% at 3%/2 mm, or higher than 80% at 2%/2 mm was set as the action limit of "pass", otherwise, it was "fail", respectively [29].

### Deep learning model

In the preprocessing, 3D dose distribution data of PTV and overlap were converted to images of NRRD format and normalized to $96 \times 96 \times 96$. A DenseNet 121 was adapted and trained for the PSQA prediction in the Medical Open Network for Artificial Intelligence (MONAI), an open-source framework for DL in medical imaging based on Pytorch. There are four dense blocks in the DenseNet 121. The layer between two adjacent blocks is called the transition layer, which changes the feature map size by convolution and pooling. DL models were trained for at least 200 epochs using the Adam optimizer and a learning rate of 0.00001. Models were trained from scratch with no pre-training with their last classifier layer would be a sigmoid layer capable of performing binary classification. To prevent overfitting and improve the generalization of the models, different data enhancement methods were applied during training. The DL score of the models with best performance in the training set will combined with dosiomics signature in the RF model for final prediction, as shown in Fig. 3.

### Model evaluation and statistical analysis

A total of five models were generated and compared, namely, PC models (based on PC metrics), dosiomics models (D), DL models, PC+D models (combined PC metrics and dosiomics features), and overall model (DL+PC+D, integrating DL with PC metrics and dosiomics features). The receiver operating characteristic (ROC) curves and the area under curve (AUC) were applied for the evaluation and comparison of these models using "pROC" package of R analysis platform (version

3.0.1, MathSoft). RF algorithm was based on "random-Forest" package. The classification algorithm was based in part on MONAI (version 0.8.1) and other open-source projects available at https://github.com/Project-MONAI/tutorials. Other data analysis was performed using Python 3.6.0 and custom-written software in MATLAB R2016a. For all tests, $p < 0.05$ was considered as statically significant.

## Results

A total of 201 two-arc VMAT plans were enrolled in this study with 135 pelvis plans and 66 head and neck (H&N) plans at a prescription dose to PTV of 45 Gy (1.8 Gy/fractions) and 60 Gy (2.0 Gy/ fractions), respectively. Pelvis plans includes gynecologic, rectal, and prostate cancer patients. H&N plans includes nasopharyngeal carcinoma, laryngeal carcinoma, and hypopharyngeal carcinoma. Detailed characteristics of these plans were summarized in Table 1.

Table 2 shows collected PSQA results of these VMAT plans with an average GPS of $97.94\% \pm 1.87\%$, $94.33\% \pm 3.22\%$, and $87.27\% \pm 4.81\%$ under the criteria of 3%/3 mm, 3%/2 mm, and 2%/2 mm, respectively. According to the important value of RF method, the top

**Table 1** The characteristics of patients enrolled in this study

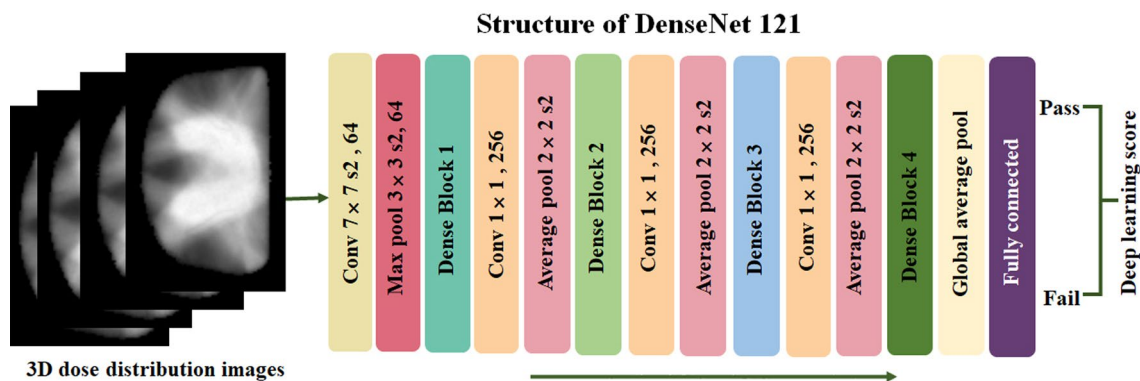| Disease site | Case number | % | Prescription |
|---|---|---|---|
| Pelvis | 135 | 67.16 | 45 Gy/25fx |
|   Gynecologic cancer | 83 | 41.29 | |
|   Rectal cancer | 44 | 21.89 | |
|   Prostate cancer | 8 | 3.98 | |
| Head and Neck | 66 | 32.84 | 60 Gy/30fx |
|   Nasopharyngeal cancer | 34 | 16.95 | |
|   Laryngeal cancer | 19 | 9.45 | |
|   Hypopharyngeal cancer | 13 | 6.47 | |



**Fig. 3** Framework for PSQA prediction based on DenseNet 121 training

**Table 2** Recorded results of patient-specific quality assurance

| Measured GPR | 3%/3 mm | 3%/2 mm | 2%/2 mm |
|---|---|---|---|
| Mean (%) | 97.94 | 94.33 | 87.27 |
| SD (%) | 1.87 | 3.22 | 4.81 |
| Case numbers | | | |
| GPR ≥ 95% | 186 | 106 | 7 |
| 90% ≤ GPR < 95% | 15 | 75 | 58 |
| 80% ≤ GPR < 90% | 0 | 22 | 115 |
| GPR < 80% | 0 | 0 | 21 |

5 complexity features and 50 dosiomics features were selected for GPR prediction with different PSQA criteria of 3%/3 mm, 3%/2 mm and 2%/2 mm, as shown in Fig. 4.

Figure 5 shows the performance of PC, D, DL, PC + D, and overall model for the PSQA criteria of 3%/3 mm, 3%/2 mm, and 2%/2 mm, respectively. The overall model achieved a best AUC of 0.948(95% CI 0.880−1), 0.890(95% CI 0.801−0.980) and 0.942(95% CI 0.856−1) at 3%/3 mm, 3%/2 mm, and 2%/2 mm, respectively. Detailed performance of these models was presented in Table 3.

## Discussion

In this study, the feasibility of combining DL with dosiomics features and PC metrics for the PSQA of patients underwent VMAT were investigated. A best AUC of 0.942 with a sensitivity, specificity and accuracy of 100%, 81.8%, and 83.6% was achieved with combined overall model at the criteria of 2%/2 mm.

PSQA is an imperative step to assess the reliability of treatment delivery of IMRT/VMAT plans and to improve the patient safety due to the increased dosimetric uncertainty resulted from inverse planning [29]. Although many studied questioned the clinical significance of GPR, gamma analysis is still the most widely applied PSQA methods [30]. GPR with the criterion of 3%/3 mm is commonly recommended and routinely applied in clinical practice for IMRT/VMAT PSQA [29, 31]. Previous studies suggested that different criteria of GPR should be applied to detect different types of errors during IMRT/VMAT delivery [32, 33]. Therefore, criteria of 3%/3 mm, 3%/2 mm, and 2%/2 mm were applied in this study for the assessment of PSQA. It was consistent with previous study, the GPR was decreased from 97.94 to 87.27% with the increased strict criteria from 3%3 mm to 2%/2 mm.

Virtual PSQA without actual measurement is desirable in the treatment planning process as to identify failing plans early in the process and to increase the efficiency of the PSQA practice [34]. Studies generally agreed that the GPRs of IMRT/VMAT plans are heavily contingent on PC [16, 35]. In this study, virtual prediction models

based on complexity metrics demonstrated that for GPRs with different criteria, the associated complex metrics were different. As shown in Fig. 4, the complexity metrics and their corresponding weights for GPRs of 3%/3 mm, 3%/2 mm and 2%/2 mm were CPs, Gantry spacing, Mean LT, MU and MU/CP; MU, CPs, Mean LT, Gantry Spacing, MU/CP < 3%; and CPs, Mean LT, Gantry Spacing, MU, MU/CP, respectively. Similarly, different complexity metrics were reported in different studies for the prediction of GPRs. Valdes et al. reported that MU/Gy, small aperture score, irregularity factor, and fraction of the plan delivered at the corners of a $40 \times 40$ cm field were the most important metrics that determines the GPRs at 3%/3 mm [12]. Shen et al. demonstrated for patients with nasopharyngeal cancer underwent two-arc VMAT treatment, complexity metrics of MU/CP and segment area (SA) per control point (SA/CP) were highly correlated with GPRs [11]. As shown in Table 3, the models based on complexity metrics demonstrated the lowest AUC for the GPRs prediction. It was consistent with previous studies that the correlations between complexity metrics and GPRs are generally weak [36, 37].

In this study, with the application of dosiomics, radiomics features extracted from dosimetric distributions, an accuracy of 91.8%, 70.5% and 78.7% was achieved in the prediction of GPRs at the criteria of 3%/3 mm, 3%/2 mm and 2%/2 mm, respectively. This is better than the reported maximum accuracy of 77.3% in the study of Nyflot et al., in which radiomics based on planar dose maps was applied for IMRT PSQA at 3%/3 mm [19]. In the prostate QA gamma deep learning prediction model, the input training data also include PTV and overlapping regions, suggesting that these areas had some important information.[15] In this study, the AUC of dosiomics features extracted from the PTV and overlapping region at 3%/2 mm, 2%/2 mm were 0.783 and 0.842, respectively, which is superior to the reported dosiomics AUCs of 0.78 and 0.81 in the study of Hirashima et al. [20]. The AUC and sensitivity of combined model with complexity features and dosiomics features at 2%/2 mm were 0.915 and 0.833, respectively, which is also comparable to the reported 0.83 and 0.90 in the study of Hirashima, et al.

Handcrafted features, such as radiomics and dosiomics, were generally the main approaches for medical imaging analysis. With the development of DL, studies indicated that combining the DL models with the handcrafted features with learned knowledge may improve the performance of these deep learning models [38, 39], In this study, the AUCs of DL models were improved from 0.943, 0.849, 0.841 to 0.948, 0.890, 0.942 in the combined overall models at the GPR criteria of 3%/3 mm, 3%/2 mm and 2%/2 mm, respectively. This is also indicated the improvement of adding DL for automatic PSQA in
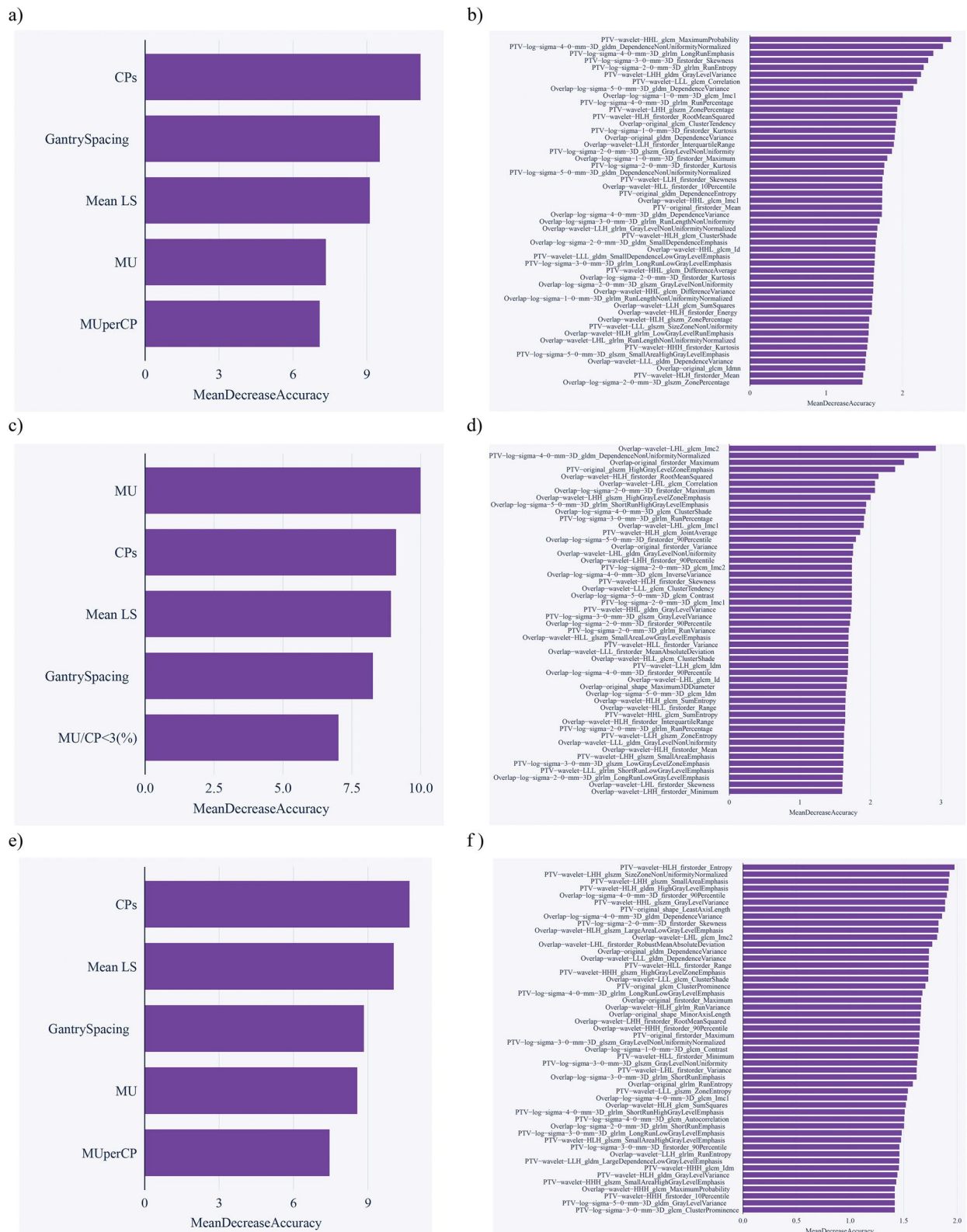
**Fig. 4** According to the important value of RF method, the top 5 complexity features and 50 dosiomics features were selected for GPR prediction with different PSQA criteria. **a** the top 5 complexity features of 3%/3 mm; **b** the top 50 dosiomics features of 3%/3 mm; **c** the top 5 complexity feature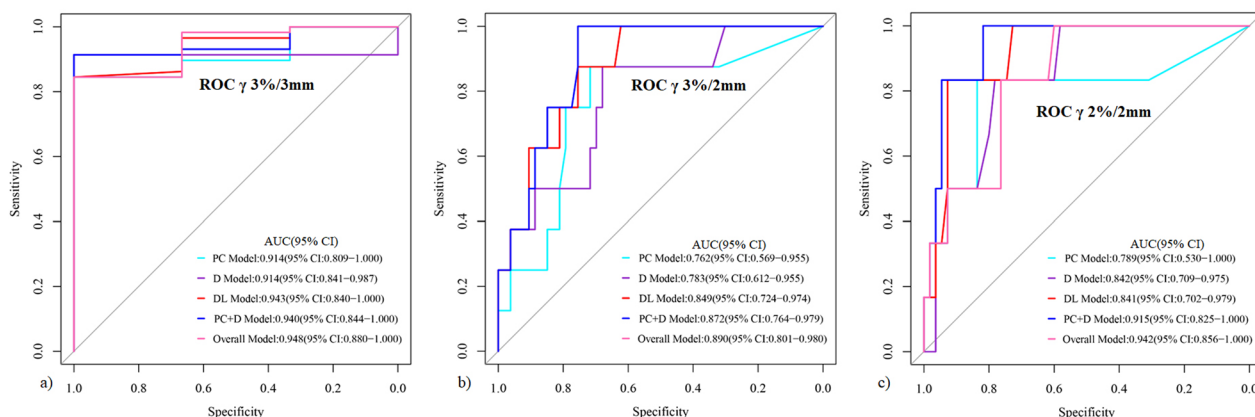s of 3%/2 mm; **d** the top 50 dosiomics features of 3%/2 mm; **e** the top 5 complexity features of 2%/2 mm; **f** the top 50 dosiomics features of 2%/2 mm

Han *et al. Radiation Oncology*     (2023) 18:116

Page 7 of 9



**Fig. 5** Performance of prediction models with plan complexity (PC), dosiomics features (D), deep learning (DP) in the testing set. **a** Prediction models of 3%/3 mm; **b** Prediction models of 3%/2 mm; **c** Prediction models of 2%/2 mm

**Table 3** Performance of different models for the prediction of patient-specific quality assurance with different percent dose difference/distance to agreement criteria

| Criteria | Models/parameters | AUC (95% CI) | Sensitivity | Specificity | Accuracy |
|---|---|---|---|---|---|
| 3%/3 mm | PC model | 0.914 (0.809–1.000) | 0.845 | 1.000 | 0.869 |
| | D model | 0.914 (0.841–0.987) | 0.914 | 1.000 | 0.918 |
| | DL model | 0.943 (0.840–1.000) | 0.845 | 1.000 | 0.853 |
| | PC + D model | 0.940 (0.844–1.000) | 0.845 | 1.000 | 0.853 |
| | Overall model | 0.948 (0.880–1.000) | 0.914 | 1.000 | 0.918 |
| 3%/2 mm | PC model | 0.762 (0.569–0.955) | 0.875 | 0.717 | 0.738 |
| | D model | 0.783 (0.612–0.955) | 0.875 | 0.679 | 0.705 |
| | DL model | 0.849 (0.724–0.974) | 1.000 | 0.566 | 0.623 |
| | PC + D model | 0.872 (0.764–0.979) | 0.875 | 0.755 | 0.771 |
| | Overall model | 0.890 (0.801–0.980) | 1.000 | 0.755 | 0.787 |
| 2%/2 mm | PC model | 0.789 (0.530–1.000) | 0.833 | 0.836 | 0.836 |
| | D model | 0.842 (0.709–0.975) | 0.833 | 0.782 | 0.787 |
| | DL model | 0.841 (0.702–0.979) | 1.000 | 0.600 | 0.639 |
| | PC + D model | 0.915 (0.825–1.000) | 0.833 | 0.927 | 0.918 |
| | Overall model | 0.942 (0.856–1.000) | 1.000 | 0.818 | 0.836 |

*PC* plan complexity, *D* dosiomics, *DL* deep learning, *AUC* area under curve

comparison with using only PC and dosiomics features in that of Hirashima et al. [20]. For the criteria of 3%/3 mm with a relatively high GPRs, the combined overall model did not show much improvement.

The cases enrolled in this study for PSQA were patients with gynecologic cancer, rectal cancer, prostate cancer and head-and-neck cancer. VMAT plans with different prescription doses were investigated. To further generalize the application of these models, VMAT plans for other site of cancer, such as esophageal, lung, and breast cancer, and from multiple institutions should be included in our future study. The reliability of GPRs is questioned, additional evaluation indices, such as individual volume-based gamma index, DVH based metrics should be further investigated in our future study.

## Conclusions

Dosiomics features were feasible for the PSQA of VMAT. Integrating DL with dosiomics and PC metrics is promising in the prediction of GPRs in PSQA for patients underwent VMAT.

**Abbreviations**
DL      Deep learning
PC      Plan complexity
PSQA      Patient-specific quality assurance
VMAT      Volumetric modulated arc therapy

Han *et al. Radiation Oncology*      (2023) 18:116

Page 8 of 9

| | |
|---|---|
| RF | Random forest |
| PTV | Planning target volume |
| GPR | Gamma passing rate |
| AUC | Area under curve |
| IMRT | Intensity-modulated radiotherapy |
| MCS | Modulation complexity score |
| LT | Leaf travel |
| ML | Machine learning |
| CNN | Convolutional neural networks |
| TPS | Treatment planning system |
| MUs | Monitor units |

## Author contributions
C.H., X.J., C.X. and J.Z. designed, supervised the project. B.Y. and H.Z. performed and analyzed most of statistical experiments. Y.W. and Z.L. wrote the manuscript. J.Y. and B.N. verified the accuracy of the data analysis. All authors have read and approved the final manuscript.

## Availability of data and materials
Research data are stored in an institutional repository and will be shared upon request to the corresponding author.

## Declarations

### Ethical approval and consent to participate
The study was conducted in accordance with the principles of the Declaration of Helsinki and approved by the institutional Ethics Committee in Clinical Research (ECCR).

### Consent for publication
Written informed consent was waived by the ECCR due to the retrospectively nature of this study (ECCR no. 2019059).

### Competing interests
The authors of this manuscript declare no conflict of interest of the article.

### Author details
[1]Department of Radiotherapy Center, 1st Affiliated Hospital of Wenzhou Medical University, Wenzhou, China. [2]Department of Medical and Radiation Oncology, 2nd Affiliated Hospital of Wenzhou Medical University, Wenzhou, China. [3]School of Basic Medical Science, Wenzhou Medical University, Wenzhou, China.

## References
1. Kerns JR, Stingo F, Followill DS, Howell RM, Melancon A, Kry SF. Treatment planning system calculation errors are present in most imaging and radiation oncology core-houston phantom failures. Int J Radiat Oncol Biol Phys. 2017;98(5):1197–203. https://doi.org/10.1016/j.ijrobp.2017.03.049.
2. Kry SF, Dromgoole L, Alvarez P, Leif J, Molineu A, Taylor P, et al. Radiation therapy deficiencies identified during on-site dosimetry visits by the imaging and radiation oncology core houston quality assurance center. Int J Radiat Oncol Biol Phys. 2017;99(5):1094–100. https://doi.org/10.1016/j.ijrobp.2017.08.013.
3. Miles EA, Clark CH, Urbano MT, Bidmead M, Dearnaley DP, Harrington KJ, et al. The impact of introducing intensity modulated radiotherapy into routine clinical practice. Radiother Oncol. 2005;77(3):241–6. https://doi.org/10.1016/j.radonc.2005.10.011.
4. Nelms BE, Chan MF, Jarry G, Lemire M, Lowden J, Hampton C, et al. Evaluating IMRT and VMAT dose accuracy: practical examples of failure to detect systematic errors when applying a commonly used metric and action levels. Med Phys. 2013;40(11):111722. https://doi.org/10.1118/1.4826166.
5. Bohoudi O, Bruynzeel AME, Senan S, Cuijpers JP, Slotman BJ, Lagerwaard FJ, et al. Fast and robust online adaptive planning in stereotactic MR-guided adaptive radiation therapy (SMART) for pancreatic cancer. Radiother Oncol. 2017;125(3):439–44. https://doi.org/10.1016/j.radonc.2017.07.028.
6. de Jong R, Crama KF, Visser J, van Wieringen N, Wiersma J, Geijsen ED, et al. Online adaptive radiotherapy compared to plan selection for rectal cancer: quantifying the benefit. Radiat Oncol. 2020;15(1):162. https://doi.org/10.1186/s13014-020-01597-1.
7. McDonald DG, Jacqmin DJ, Mart CJ, Koch NC, Peng JL, Ashenafi MS, et al. Validation of a modern second-check dosimetry system using a novel verification phantom. J Appl Clin Med Phys. 2017;18(1):170–7. https://doi.org/10.1002/acm2.12025.
8. Hoffmann L, Alber M, Söhn M, Elstrøm UV. Validation of the Acuros XB dose calculation algorithm versus Monte Carlo for clinical treatment plans. Med Phys. 2018. https://doi.org/10.1002/mp.13053.
9. Webb S. Use of a quantitative index of beam modulation to characterize dose conformality: illustration by a comparison of full beamlet IMRT, few-segment IMRT (fsIMRT) and conformal unmodulated radiotherapy. Phys Med Biol. 2003;48(14):2051–62. https://doi.org/10.1088/0031-9155/48/14/301.
10. McNiven AL, Sharpe MB, Purdie TG. A new metric for assessing IMRT modulation complexity and plan deliverability. Med Phys. 2010;37(2):505–15. https://doi.org/10.1118/1.3276775.
11. Shen L, Chen S, Zhu X, Han C, Zheng X, Deng Z, et al. Multidimensional correlation among plan complexity, quality and deliverability parameters for volumetric-modulated arc therapy using canonical correlation analysis. J Radiat Res. 2018;59(2):207–15. https://doi.org/10.1093/jrr/rrx100.
12. 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. https://doi.org/10.1118/1.4953835.
13. Valdes G, Chan MF, Lim SB, Scheuermann R, Deasy JO, Solberg TD. IMRT QA using machine learning: a multi-institutional validation. J Appl Clin Med Phys. 2017;18(5):279–84. https://doi.org/10.1002/acm2.12161.
14. Li J, Wang L, Zhang X, Liu L, Li J, Chan MF, et al. Machine learning for patient-specific quality assurance of VMAT: prediction and classification accuracy. Int J Radiat Oncol Biol Phys. 2019;105(4):893–902. https://doi.org/10.1016/j.ijrobp.2019.07.049.
15. Tomori S, Kadoya N, Takayama Y, Kajikawa T, Shima K, Narazaki K, et al. A deep learning-based prediction model for gamma evaluation in patient-specific quality assurance. Med Phys. 2018. https://doi.org/10.1002/mp.13112.
16. Du W, Cho SH, Zhang X, Hoffman KE, Kudchadker RJ. Quantification of beam complexity in intensity-modulated radiation therapy treatment plans. Med Phys. 2014;41(2):021716. https://doi.org/10.1118/1.4861821.
17. Crowe SB, Kairn T, Kenny J, Knight RT, Hill B, Langton CM, et al. Treatment plan complexity metrics for predicting IMRT pre-treatment quality assurance results. Australas Phys Eng Sci Med. 2014;37(3):475–82. https://doi.org/10.1007/s13246-014-0274-9.
18. Wootton LS, Nyflot MJ, Chaovalitwongse WA, Ford E. Error detection in intensity-modulated radiation therapy quality assurance using radiomic analysis of gamma distributions. Int J Radiat Oncol Biol Phys. 2018;102(1):219–28. https://doi.org/10.1016/j.ijrobp.2018.05.033.
19. Nyflot MJ, Thammasorn P, Wootton LS, Ford EC, Chaovalitwongse WA. Deep learning for patient-specific quality assurance: Identifying errors in radiotherapy delivery by radiomic analysis of gamma images with convolutional neural networks. Med Phys. 2019;46(2):456–64. https://doi.org/10.1002/mp.13338.
20. 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. https://doi.org/10.1016/j.radonc.2020.07.031.

Han *et al. Radiation Oncology*      (2023) 18:116

Page 9 of 9

21. Zheng X, Yao Z, Huang Y, Yu Y, Wang Y, Liu Y, et al. Deep learning radiomics can predict axillary lymph node status in early-stage breast cancer. Nat Commun. 2020;11(1):1236. https://doi.org/10.1038/s41467-020-15027-z.
22. Jiang M, Li CL, Luo XM, Chuan ZR, Lv WZ, Li X, et al. Ultrasound-based deep learning radiomics in the assessment of pathological complete response to neoadjuvant chemotherapy in locally advanced breast cancer. Eur J Cancer. 2021;147:95–105. https://doi.org/10.1016/j.ejca.2021.01.028.
23. Jin X, Yi J, Zhou Y, Yan H, Han C, Xie C. Comparison of whole-field simultaneous integrated boost VMAT and IMRT in the treatment of nasopharyngeal cancer. Med Dosim. 2013;38(4):418–23. https://doi.org/10.1016/j.meddos.2013.05.004.
24. Deng X, Han C, Chen S, Xie C, Yi J, Zhou Y, et al. Dosimetric benefits of intensity-modulated radiotherapy and volumetric-modulated arc therapy in the treatment of postoperative cervical cancer patients. J Appl Clin Med Phys. 2017;18(1):25–31. https://doi.org/10.1002/acm2.12003.
25. Jin X, Yan H, Han C, Zhou Y, Yi J, Xie C. Correlation between gamma index passing rate and clinical dosimetric difference for pre-treatment 2D and 3D volumetric modulated arc therapy dosimetric verification. Br J Radiol. 2015;88(1047):20140577. https://doi.org/10.1259/bjr.20140577.
26. Yi J, Han C, Zheng X, Zhou Y, Deng Z, Xie C, et al. Individual volume-based 3D gamma indices for pretreatment VMAT QA. J Appl Clin Med Phys. 2017;18(3):28–36. https://doi.org/10.1002/acm2.12062.
27. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res. 2017;77(21):e104–7. https://doi.org/10.1158/0008-5472.CAN-17-0339.
28. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
29. Miften M, Olch A, Mihailidis D, Moran J, Pawlicki T, Molineu A, Li H, Wijesooriya K, Shi J, Xia P, Papanikolaou N, Low DA. Tolerance limits and methodologies for IMRT measurement-based verification QA: recommendations of AAPM Task Group No. 218. Med Phys. 2018;45(4):e53–83. https://doi.org/10.1002/mp.12810.
30. Stojadinovic S, Ouyang L, Gu X, Pompoš A, Bao Q, Solberg TD. Breaking bad IMRT QA practice. J Appl Clin Med Phys. 2015;16(3):5242. https://doi.org/10.1120/jacmp.v16i3.5242.
31. Ezzell GA, Burmeister JW, Dogan N, LoSasso TJ, Mechalakos JG, Mihailidis D, et al. IMRT commissioning: multiple institution planning and dosimetry comparisons, a report from AAPM Task Group 119. Med Phys. 2009;36(11):5359–73. https://doi.org/10.1118/1.3238104.
32. Rangel A, Palte G, Dunscombe P. The sensitivity of patient specific IMRT QC to systematic MLC leaf bank offset errors. Med Phys. 2010;37(7):3862–7. https://doi.org/10.1118/1.3453576.
33. Nelms BE, Zhen H, Tomé WA. Per-beam, planar IMRT QA passing rates do not predict clinically relevant patient dose errors. Med Phys. 2011;38(2):1037–44. https://doi.org/10.1118/1.3544657.
34. Kalet AM, Luk SMH, Phillips MH. Radiation therapy quality assurance tasks and tools: the many roles of machine learning. Med Phys. 2020;47(5):e168–77. https://doi.org/10.1002/mp.13445.
35. Younge KC, Roberts D, Janes LA, Anderson C, Moran JM, Matuszak MM. Predicting deliverability of volumetric-modulated arc therapy (VMAT) plans using aperture complexity analysis. J Appl Clin Med Phys. 2016;17(4):124–31. https://doi.org/10.1120/jacmp.v17i4.6241.
36. Crowe SB, Kairn T, Middlebrook N, Sutherland B, Hill B, Kenny J, et al. Examination of the properties of IMRT and VMAT beams and evaluation against pre-treatment quality assurance results. Phys Med Biol. 2015;60(6):2587–601. https://doi.org/10.1088/0031-9155/60/6/2587.
37. Glenn MC, Hernandez V, Saez J, Followill DS, Howell RM, Pollard-Larkin JM, et al. Treatment plan complexity does not predict IROC Houston anthropomorphic head and neck phantom performance. Phys Med Biol. 2018;63(20):205015. https://doi.org/10.1088/1361-6560/aae29e.
38. Roth HR, Lu L, Liu J, Yao J, Seff A, Cherry K, et al. Improving computer-aided detection using convolutional neural networks and random view aggregation. IEEE Trans Med Imaging. 2016;35(5):1170–81. https://doi.org/10.1109/TMI.2015.2482920.
39. Antropova N, Huynh BQ, Giger ML. A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. Med Phys. 2017;44(10):5162–71. https://doi.org/10.1002/mp.12453.