Machine learning-based predictions of gamma passing rates for virtual specific-plan verification based on modulation maps, monitor unit profiles, and composite dose images

QUINTERO, Paulo, BENOIT, David, CHENG, Yongqiang, MOORE, Craig and BEAVIS, Andrew

IOP Publishing Phys. Med. Biol. 67 (2022) 245001 https://doi.org/10.1088/1361-6560/aca38a

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

OPEN ACCESS

RECEIVED
24 September 2022

REVISED
31 October 2022

ACCEPTED FOR PUBLICATION
16 November 2022

PUBLISHED
6 December 2022

PAPER

Machine learning-based predictions of gamma passing rates for virtual specific-plan verification based on modulation maps, monitor unit profiles, and composite dose images

Paulo Quintero[1,2,*], **David Benoit**[1], **Yongqiang Cheng**[1], **Craig Moore**[2] and **Andrew Beavis**[2,3,4]

[1] Faculty of Science and Engineering, University of Hull, Hull, United Kingdom
[2] Medical Physics Department, Queen's Centre for Oncology, Hull University Teaching Hospitals NHS Trust, Cottingham, United Kingdom
[3] Faculty of Health and Wellbeing, Sheffield Hallam University, Sheffield, United Kingdom
[4] Faculty of Health Sciences, University of Hull, Hull, United Kingdom
[*] Author to whom any correspondence should be addressed.

E-mail: pquinterome@gmail.com

Keywords: machine-learning, radiotherapy, CNN, gamma-passing-rates

Supplementary material for this article is available online

Abstract

Machine learning (ML) methods have been implemented in radiotherapy to aid virtual specific-plan verification protocols, predicting gamma passing rates (GPR) based on calculated modulation complexity metrics because of their direct relation to dose deliverability. Nevertheless, these metrics might not comprehensively represent the modulation complexity, and automatically extracted features from alternative predictors associated with modulation complexity are needed. For this reason, three convolutional neural networks (CNN) based models were trained to predict GPR values (regression and classification), using respectively three predictors: (1) the modulation maps (MM) from the multi-leaf collimator, (2) the relative monitor units per control point profile (MUcp), and (3) the composite dose image (CDI) used for portal dosimetry, from 1024 anonymized prostate plans. The models' performance was assessed for classification and regression by the area under the receiver operator characteristic curve (AUC_ROC) and Spearman's correlation coefficient (r). Finally, four hybrid models were designed using all possible combinations of the three predictors. The prediction performance for the CNN-models using single predictors (MM, MUcp, and CDI) were AUC_ROC $= 0.84 \pm 0.03, 0.77 \pm 0.07, 0.75 \pm 0.04$, and $r = 0.6, 0.5, 0.7$. Contrastingly, the hybrid models (MM + MUcp, MM + CDI, MUcp+CDI, MM + MUcp+CDI) performance were AUC_ROC $= 0.94 \pm 0.03, 0.85 \pm 0.06, 0.89 \pm 0.06, 0.91 \pm 0.03$, and $r = 0.7, 0.5, 0.6, 0.7$. The MP, MUcp, and CDI are suitable predictors for dose deliverability models implementing ML methods. Additionally, hybrid models are susceptible to improving their prediction performance, including two or more input predictors.

Introduction

Artificial intelligence (AI) methods have been applied in radiotherapy, supporting the contouring of the target and organs at risk volumes (Lustberg *et al* 2018, Meyer *et al* 2018, Sahiner *et al* 2019, el Naqa and Das 2020), the prediction of clinical outcomes (el Naqa *et al* 2009, J *et al* 2015, Nguyen *et al* 2017), the dose distribution predictions (Campbell *et al* 2017, Nguyen *et al* 2017, Liu *et al* 2019), synthetic image reconstructions (Han 2017, Trullo *et al* 2017, Wolterink *et al* 2017, Zhao *et al* 2017, Xiang *et al* 2018), and the dose deliverability prediction (Tomori *et al* 2018, Ono *et al* 2019), among others (Workshop on unsupervised, transfer Learning PB-I, undefined 2012, Cha *et al* 2016, Yan *et al* 2016, Ibragimov *et al* 2017, Hesamian *et al* 2019). Certainly, in the past six years, machine learning (ML) methods dedicated to quality assurance (QA) predictions of

intensity-modulated radiotherapy (IMRT) and volumetric modulated arc therapy (VMAT) treatments have increasingly been studied (Hussein *et al* 2017, Chan *et al* 2020a, Osman and Maalej 2021). The most common ML models implemented in this matter are Poisson regression (Valdes *et al* 2016, 2017, Li *et al* 2019), decision trees-based models (e.g. random forest or gradient boosting models) (Lam *et al* 2019, Hirashima *et al* 2020), support vector machine (Valdes *et al* 2016, Granville *et al* 2019), and artificial neural networks or convolutional neural networks (CNN) (Interian *et al* 2018, Tomori *et al* 2018, 2020). The CNN-based models, which were being less explored in QA predictions, are characterized commonly by convolution plus pooling layers arranged consecutively, ending with fully connected layers and a *Softmax* activated dense layer for classification or a *Linear* activated dense layer for regression (Payer *et al* 2016). The convolution operations intend to detect patterns from the input images using specific filters and reducing their dimensions. Then, these newly detected features are processed by the pooling layers, weighting the found features and their nearby values to be the input of the next convolutional-pooling layer arrangement, filtering intricated 'hidden' features that will potentially be associated with the predicted output.

From the specific-plan verification perspective, models dedicated to QA prediction were implemented generally to detect potential treatment errors (Ezzell *et al* 2009, Miften *et al* 2018) and predict gamma passing rate (GPR) values (Low *et al* 1998). The GPRs account for the dosimetric regions in agreement with the gamma index analysis between the calculated and the measured dose distributions (Low *et al* 1998, Hussein *et al* 2013). In turn, the gamma index is a metric that evaluates the coincidence between both dose distributions, calculating the dose difference (DD) and the distance to agreement (DTA) (Hussein *et al* 2013). Commonly, a verified treatment is suitable for delivery if the GPR is higher than one reference value, selecting the DD/DTA criteria defined in each institution and per the expert recommendations (Miften *et al* 2018). For instance, a specific treatment might be considered appropriate if its GPR is equal to or higher than 98% based on 3%/2 mm criteria. Nevertheless, although this metric has been studied and implemented widely, some gaps have been identified in detecting errors with clinical impact or retrieving information needed to detect specific discrepancies regarding treatment parameters (Zhen *et al* 2011, Hussein *et al* 2017, Park *et al* 2018). Hence, the GPR evaluation and the modelled predictions should be considered complementary tests to other assessment protocols (e.g. dose-volume histogram changes evaluation) rather than one exclusive verification method.

Consequently, a useful GPR prediction model based on ML methods should be able to provide additional information to complement and explain the expected dose deliverability evaluation results, featuring the predominant predictors and achieving a more robust evaluation of the treatment parameters. Similarly, it might be beneficial to track possible 'problematic' treatment features, as suggested by Park *et al (*Park *et al* 2015, Carlson *et al* 2016), McNivell *et al (*McNiven *et al* 2010), Petroccia *et al (*Petroccia *et al* 2019), and Chiavassa *et al (*Chiavassa *et al* 2019*)*, using modulation complexity metrics and plan parameters. However, the reported models using automatic-extracted features methods (e.g. CNN-based models) are based mainly on dose distributions (Osman and Maalej 2021), and predictor features associated with the plan parameters cannot be extracted. In contrast, other input features, such as modulation maps (MM) given by the multi-leaf collimator (MLC) trajectories per control points (CP), gantry speed variations, or monitor units (MU) variations profiles, have not been explored, and it might help to complement the dose deliverability evaluation because their direct relation to specific treatment conditions.

In terms of the studied features for GPR predictions using ML models, classification or regression solutions have been proposed based on IMRT beam fluencies (Interian *et al* 2018, Hirashima *et al* 2020), planar dose images plus organs at risk volumes and total MU values (Tomori *et al* 2018, 2020), radiomic features from the dose distribution images (Nyflot *et al* 2018, Hirashima *et al* 2020), and various calculated modulation complexity metrics (Valdes *et al* 2016, Ono *et al* 2019, Chan *et al* 2020a). In fact, benefits on prediction performance have been reported when more than one input feature category is implemented (i.e. hybrid datasets or hybrid models) (Tomori *et al* 2018, Hirashima *et al* 2020). However, considering that complexity metrics and features related to MLC movements are the most relevant features for GPR predictions (Park *et al* 2018, Lam *et al* 2019, Park *et al* 2019, Wall and Fontenot 2020), it is necessary to contemplate the MM and the MU per CP (MUcp) variations as potential GPR predictors, implementing automatic-feature extraction methods and avoiding in this way the use of conventional complexity formulas (McNiven *et al* 2010, Masi *et al* 2013, Tamura *et al* 2020) that might limit the amount of information extracted.

Considering the abovementioned, this study aims to explore features directly related to treatment unit parameters to predict GPR values based on CNN models, contributing to the inclusion and evaluation of additional treatment parameters that might facilitate the designs of more robust dose deliverability evaluation protocols. For this reason, the primary objective of this study was to evaluate the potential utility of MM and MUcp as input features for GPR predictions. Consequently, since our GPR values were calculated using electronic portal imaging devices (EPID) measurements, we decided to include the calculated composite dose image (CDI) as a third evaluated input feature (i.e. dosimetric input feature). The second objective was to verify

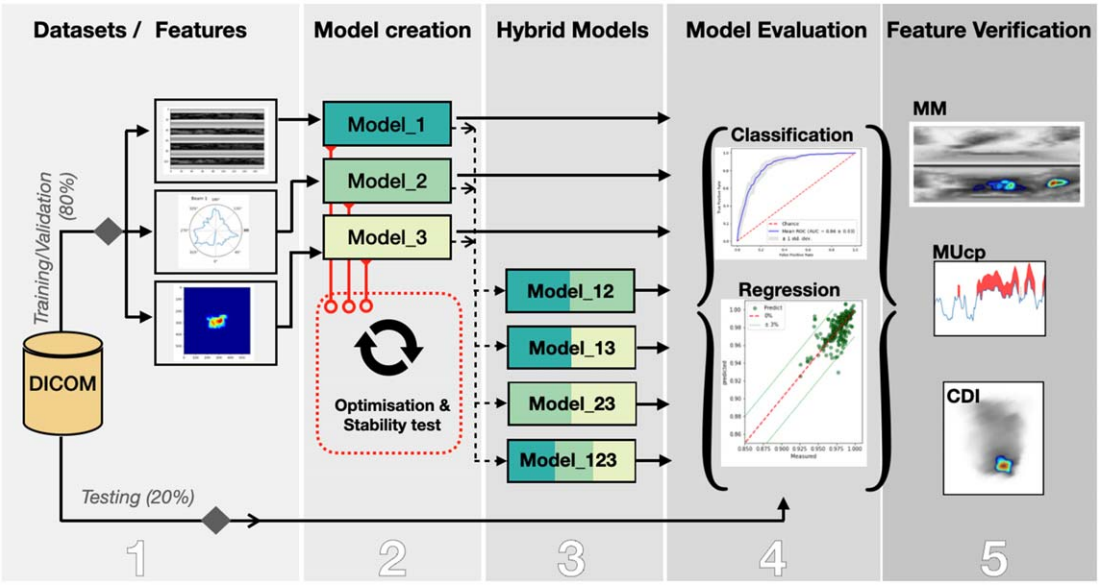

Figure 1. Workflow of the present study, including the (1) dataset creation, (2) the corresponding designed main models (M_1, M_2, and M_3) plus their optimization and stability evaluation, (3) the design of the assembled hybrid models, (4) the prediction performance evaluation, for the training and testing sub-datasets, and (5) the verification of activated features.

whether concatenated models presented an improved GPR prediction performance or not. Furthermore, we aimed to evaluate the model stability in terms of the quality of the learned features extracted by each model.

Methods

Workflow

The five-step workflow followed in this study is illustrated in figure 1. (I) From 1024 DICOM-RT files, the MM, MUcp profiles, and CDI were retrieved and classified to form three specific datasets representing each feature category. (II) An independent CNN model was designed for each input dataset to predict GPRs (classification and regression). The architecture optimization, the hyper-parameter tunning, and stability tests were performed with TensorFlow (Dillon *et al* 2017). (III) In addition, four hybrid models based on all possible previous models' combinations were proposed to verify if the GPR prediction improves concatenating two or more models. (IV) Next, the ROC-AUC and the accuracy were calculated to evaluate the prediction performance of classification models, and the MAE, RMSE, and Spearman correlation coefficients were calculated for regression models. (V) Finally, the activation maps for randomly chosen plans were extracted to verify the relevance of the trained features.

Dataset

A total of 1024 anonymized DICOM-RT files from 746 prostate plans, retrospectively treated in our institution, were retrieved to extract the MM, the MUcp, and the CDI features by Python scripting (Quintero 2020). The treatments were planned with Eclipse version 15.6 (Varian Medical Systems, Palo Alto, CA), 2 degrees per CP configuration, and 6 MV beam energy in two Varian treatment units (TrueBeam and Halcyon-v2) available in our institution with the same EPID model (aS1200) and calibrated under the same reference conditions. Both treatment units have 5 mm of nominal resolution at the isocentre with Millennium 120 MLC (TrueBeam) and dual-layer MLC (Halcyon-v2) models and a maximum leaf speed of 25 mm s^{-1} and 50 mm s^{-1}, respectively. Furthermore, the dataset was divided into 80% for training and validation sub-datasets (80%/20% in turn, $N = 819$) and 20% for the testing sub-dataset ($N = 205$), as it is illustrated in figure 2. The treatment plan conditions are summarised in table 1.

The GPRs were calculated from gamma analysis evaluation (Low *et al* 1998) based on EPID measurements and a global 2% dose and 1 mm distance differences criteria (2%/1 mm). For classification models, the VMAT dose distributions with a GPR $\geqslant$ 98% were labelled as 'pass' ($N = 49\%$); otherwise, they were labelled as fail ($N = 51\%$). This 2%/1 mm reference value was chosen considering both treatment units and one evaluation threshold able to discriminate potential errors that might affect the planned dose distributions, in accordance with the AAPM-TG 218 recommendations (Miften *et al* 2018). However, this value also promoted the

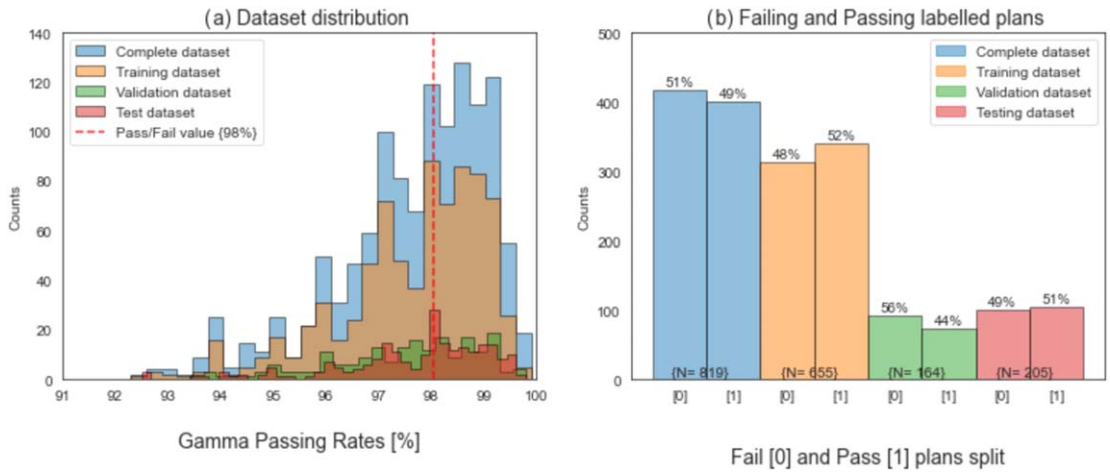

Figure 2. Distribution of the number of plans (counts) for (a) all GPR values with the GPR criteria of 98% (dotted red line), and (b) the representation of all sub-dataset splits. Plans labelled as 'fail' were represented with [0] and plans labelled as 'pass' were represented with (Lustberg *et al* 2018).

Table 1. Summary of planning conditions for prostate dataset considering.

Treatment unit	Energy mode	Number of arcs	Dose per fraction [Gy]	Number of plans	Number of inputs	%	
TrueBeam	6 MV FF	1	2	85	85	8.3	46.6
			2.7	70	70	6.8	
			3	236	236	23.0	
		2	2	43	86	8.4	
Halcyon	6 MV FFF	1	3	77	77	7.5	53.4
		2	3	235	470	45.9	

Abbreviations: Flattening filter, FF. Flattening filter free, FFF.

best-balanced conditions in GPR terms when the datasets were divided into sub-datasets (figure 2(b)), avoiding unreliable classification modelling and overfitting effects (Chen *et al* 2020). As it is registered in the supplementary material 1.1, most measured plans evaluated with 3%/3 mm, 3%/2 mm, 2%/3 mm and 2%/2 mm criteria presented GPR values of 100%, generating highly unbalanced datasets.

Input features

- The MM input feature from a single VMAT-arc is a two-dimensional image created with all MLC positions per cp (figure 3(a)). The leaf number indicated on the *y*-axis includes both MLC banks (four in the case of Halcyon-v2), and the displacements were normalized to take values from zero to one. Additionally, to optimize the model's 'learning process,' the static leaves were removed, keeping just the active ones during the treatment (figure 3(b)).
- The MUcp is one-dimensional data containing all MU contributions per cp during one VMAT-arc trajectory, normalized from zero to one based on the total MU values (figure 3(c)). It is extracted from the dose contribution coefficient within the DICOM-RT tag [300A,010C] labelled *CumulativeDoseReferenceCoefficient*.
- The CDI is a two-dimensional image created with the superposition of all calculated dose fluencies during the VMAT-arc trajectory over a gantry perpendicular common plane. It is calculated by the portal dosimetry image prediction algorithm (Berger *et al* 2006, Esch *et al* 2013) integrated into Eclipse (figure 2(d)) and is used to be compared to the dose measured by the EPID to perform the gamma analysis. For modelling purposes, the CDIs were normalized from zero to one.

Models

The designed models for MM, MUcp, and CDI features were noted as M_1, M_2, and M_3, respectively. An *r* or *c* character was included at the end of the notation to differentiate between regression and classification models

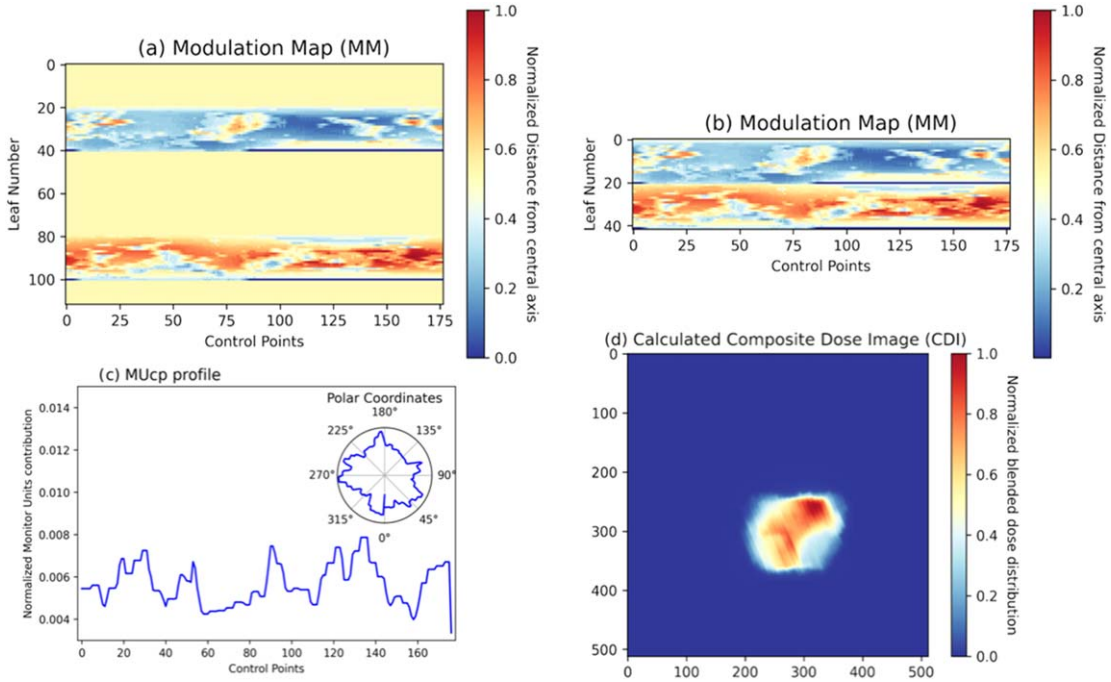

Figure 3. Representation of the three features used in this study. (a) The full modulation map (MM) and (b) the edited MM removing the static leaves. (c) The monitor units per control point (MUcp) profile and its representation in polar coordinates. (d) Composite dose image (CDI) calculated by the portal dosimetry tools in the treatment planning system.

(e.g. M_1r for regression and M_1c for classification). Additionally, four hybrid models were created from the three main previous models and were noted as M_12, M_13, M_23, and M_123, indicating the included concatenated models with their indexed notation. Furthermore, five-fold cross-validation was applied and 'Horizontal Flip' was the only data augmentation explored in this study to ensure that all input features keep accurate physical representation within training modelling. Accordingly, all models implemented in this study were based on CNN architectures and were designed using the most straightforward possible architectures, establishing the minimum optimal number of CNN-Maxpool layers and filters for each type of input category. This direction might help to control overfitting events, track specific features from each input increasing the model reliability, and reduce the predictions predominated by random features with no physical context (Chauhan *et al* 2018, Chen *et al* 2020, Kimura *et al* 2020).

After the models were designed and optimized, the three main models, M_1c, M_2c, and M_3c were modified, including drop-out layers after each convolution/max-pooling layer arrangement to evaluate their performance stability as the drop-out rate increases systematically. This test is proposed to verify the minimum number of nodes needed to extract features that correlate to GPRs and simultaneously evaluate the contribution of the random extracted features created by the convolutions.

Evaluation

The prediction performance for regression models were evaluated measuring the mean absolute error (MAE), the root mean squared error (RMSE), and the Spearman's correlation coefficient (r) between the measured and the predicted GPR values. High, moderate, and lower correlations were defined for $r < 0.4$, $0.4 \leqslant r \leqslant 0.7$, and $r > 0.7$ values, respectively. Furthermore, the classification model performance was assessed calculating the area under the receiver operating characteristic curve (ROC_AUC), accuracy, specificity, and sensitivity (table 2).

Activation maps

The activation maps of six plans from the testing datasets were generated to verify if the trained features correspond to regions of interest associated with dose deliverability (e.g. demanding hardware conditions) that might help in further decision support tools implementations. Three cases were randomly selected from the correctly classified plans labelled as 'Pass', and three plans correctly labelled as 'fail'.

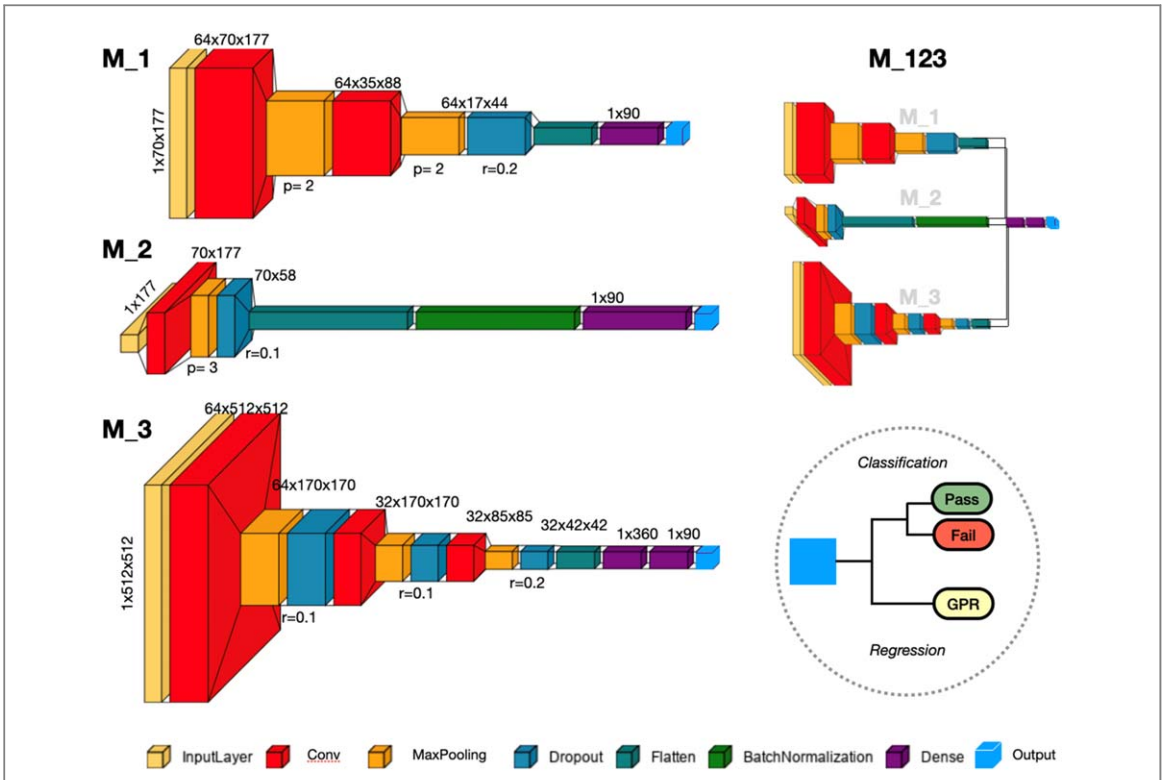

Figure 4. Convolutional Neural network architectures corresponding to the models M_1, M_2, M_3, and M_123. The output is also represented as a dual output for classification (*pass-fail*) and a single output for regression.

Table 2. Evaluation metrics implemented in this study.

Model Prediction	Metric	Equation
Regression	MAE	$MAE = \sum\|(y_i - y_p)\|/n$
	RMSE	$RMSE = \sqrt{\sum(y_i - y_p)^2/n}$
	r	—
Classification	Accuracy	$A = (TP + TN)/(TP + TN + FP + FN)$
	Specificity (Sp)	$Sp = TN / (TN + FP)$
	Sensitivity (Se)	$Se = TP / (TP + FN)$
	ROC_AUC	—

Abbreviation: MAE, mean absolute error. RMSE, root mean square error. *yi*, actual value. *yp*, predicted value. *n*, number of observations. TP, true positives. TN, true negatives. FP, false positives. FN, false negatives. *r*, Spearman's correlation coefficient. ROC_AUC, area under the receiver operating characteristic curve. A, accuracy. Sp, specificity. Se, sensitivity, also known as Recall.

Results

Model architecture

The M_1, M_2, and M_3 models were designed independently using *HParam* tool in TensorBoard, optimizing for each model the number of layers, number of filters, kernel size, drop-out rate, and activation functions. A brief representation of the resulting models' architecture is displayed in figure 4 and a detailed description is available in the supplementary material 1.2.

Architecture stability

The results for the model stability test are represented in figure 5. The models M_1, M_2, and M_3 presented more stability with up to 50% activated nodes (Drop-Out rate of 0.5) of each convolution layer, indicating that the remaining extracted features are still enough for GPR predictions. These results are consistent with the original models' performances, however, is it clear that M_2 is more susceptible to reduce the accuracy compared to M_1, which represent a more robust prediction based on the remaining features.

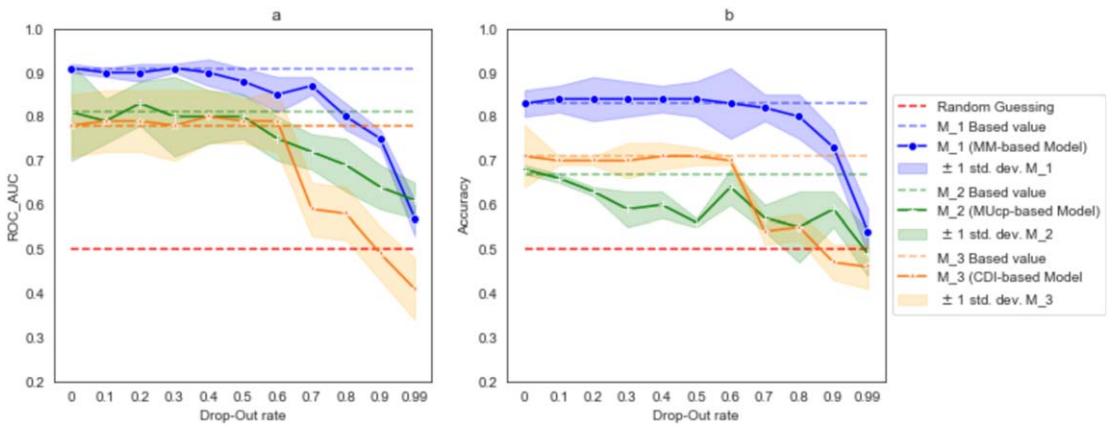

Figure 5. Model stability test of ROC_AUC and accuracy for models M_1c, M_2c, and M_3c.

Modelling performance

The modelling classification and regression performances for all models were summarised in table 3, and in figures 6 and 7.

The activation maps for the classified 'passing' and 'failing' plans are summarised in figures 8 and 9, respectively. For MM, passing plans activated static leaf regions while failing plans detected specific regions associated with demanding variations of leaf positions. For MUcp profiles, no distinctive regions were detected. Finally, for CDI, the high dose regions were identified in both failing and passing scenarios. The information of the other plans is available in supplementary material 1.3.

Discussion

Our study investigated the suitability of MM, MUcp, and CDI for GPR predictions implementing ML models. We used these three input features to explore new treatment-plan information apart from the already studied dose distributions and reported complexity metrics (McNiven *et al* 2010, Masi *et al* 2013, Chiavassa *et al* 2019, Tamura *et al* 2020). Indeed, the MM and MUcp can be considered high-dimensional modulation complexity features directly related to the treatment unit performance, which correlates to the dose deliverability (Park *et al* 2015, Chiavassa *et al* 2019, Park *et al* 2019). Hence, we intended to predict GPRs based on practical physical aspects involved in the treatment delivery, avoiding calculating limited complexity metrics from empirical equations. Furthermore, we also evaluated the CDI as an additional predictor feature because the GPR values in this study were calculated from EPID measurements, and these dose images might contain information associated with demanding linac conditions (Agnew *et al* 2014, Miri *et al* 2016, Lam *et al* 2019). In addition to this exploratory study, we also evaluated and confirmed the potential benefit of including more than one kind of treatment feature within the GPR prediction process (figure 6). Certainly, we believe that a GPR prediction model should consider all possible physical aspects involved in the treatment simultaneously, whether dosimetric or mechanic features, to achieve a more robust performance based on all variables that intervene in each treatment plan delivery. Considering the above, the goal of this study was not to propose the more efficient and complex CNN-based models but to (1) implement straightforward architecture models to evaluate the potential utility of MM, MUcp, and CDI features in GPR predictions, (2) verify if concatenated models increase the GPR prediction performance, and (3) asses the quality of the learned features extracted by each model in GPR predictions.

This study is the first reported evaluation of the MM, MUcp, and CDI as potential GPR predictors using ML methods (Chan *et al* 2020a, Osman and Maalej 2021). Previous works have implemented regression models based on modulation complexity metrics and dosimetric parameters, reporting mean prediction errors between 2.2% and 4.5% (Valdes *et al* 2016, Lam *et al* 2019, Li *et al* 2019, Kimura *et al* 2020). Similarly, MAE values between 0.74 and 4.2, RMSE = 1.54–5.6, and $r = 0.38$–0.73 have been reported from models using: one VGG-16 adapted architecture model based on 2D IMRT fluencies (Interian *et al* 2018); one CNN-based hybrid model based on planar (sagittal) dose images, volumes data, and MU values (Tomori *et al* 2018); one gradient-boosting model based on radiomic features, clinical parameters, and modulation complexity metrics (Hirashima *et al* 2020); and one support vector machine based on complexity metrics and plan parameters (Wall and Fontenot 2020). Likewise, using the same input features, reported classification models presented ROC_AUC values between 0.7 and 0.88 (Granville *et al* 2019, Hirashima *et al* 2020). In contrast, this study's MAE, RMSE, *r*,

Table 3. Evaluation metrics results for classification and regression models.

	Metric		M_1	M_2	M_3	M_12	M_13	M_23	M_123
Classification	ROC_AUC	Val.	0.91 ± 0.01	0.81 ± 0.05	0.78 ± 0.03	0.95 ± 0.01	0.89 ± 0.04	0.93 ± 0.01	0.93 ± 0.02
		Test	0.84 ± 0.03	0.77 ± 0.07	0.75 ± 0.04	0.94 ± 0.03	0.85 ± 0.06	0.89 ± 0.06	0.91 ± 0.03
	Accuracy	Val.	0.83 ± 0.09	0.68 ± 0.04	0.71 ± 0.07	0.87 ± 0.10	0.91 ± 0.02	0.82 ± 0.13	0.87 ± 0.02
		Test	0.81 ± 0.03	0.66 ± 0.10	0.68 ± 0.03	0.83 ± 0.04	0.90 ± 0.02	0.78 ± 0.05	0.88 ± 0.03
Regression	MAE [%]	Val.	1.11 ± 0.33	2.02 ± 0.23	1.09 ± 0.29	1.05 ± 0.81	1.03 ± 0.12	1.40 ± 0.12	1.12 ± 0.13
		Test	1.41 ± 0.23	2.31 ± 0.43	1.12 ± 0.23	1.08 ± 0.32	1.41 ± 0.29	1.81 ± 0.46	1.71 ± 0.11
	RMSE [%]	Val.	2.13 ± 0.01	2.66 ± 0.01	2.05 ± 0.01	2.02 ± 0.01	3.02 ± 0.01	2.11 ± 0.02	2.41 ± 0.12
		Test	2.61 ± 0.03	3.01 ± 0.02	2.11 ± 0.03	2.71 ± 0.33	3.11 ± 0.12	3.07 ± 0.05	3.16 ± 0.08
	r	Val.	0.62	0.46	0.65	0.66	0.53	0.58	0.68
	spear corr.	Test	0.61	0.33	0.61	0.58	0.42	0.49	0.59

Abbreviations. ROC_AUC, area under the receiver operating characteristic curve. MAE, mean absolute error. RMSE, root mean square error.

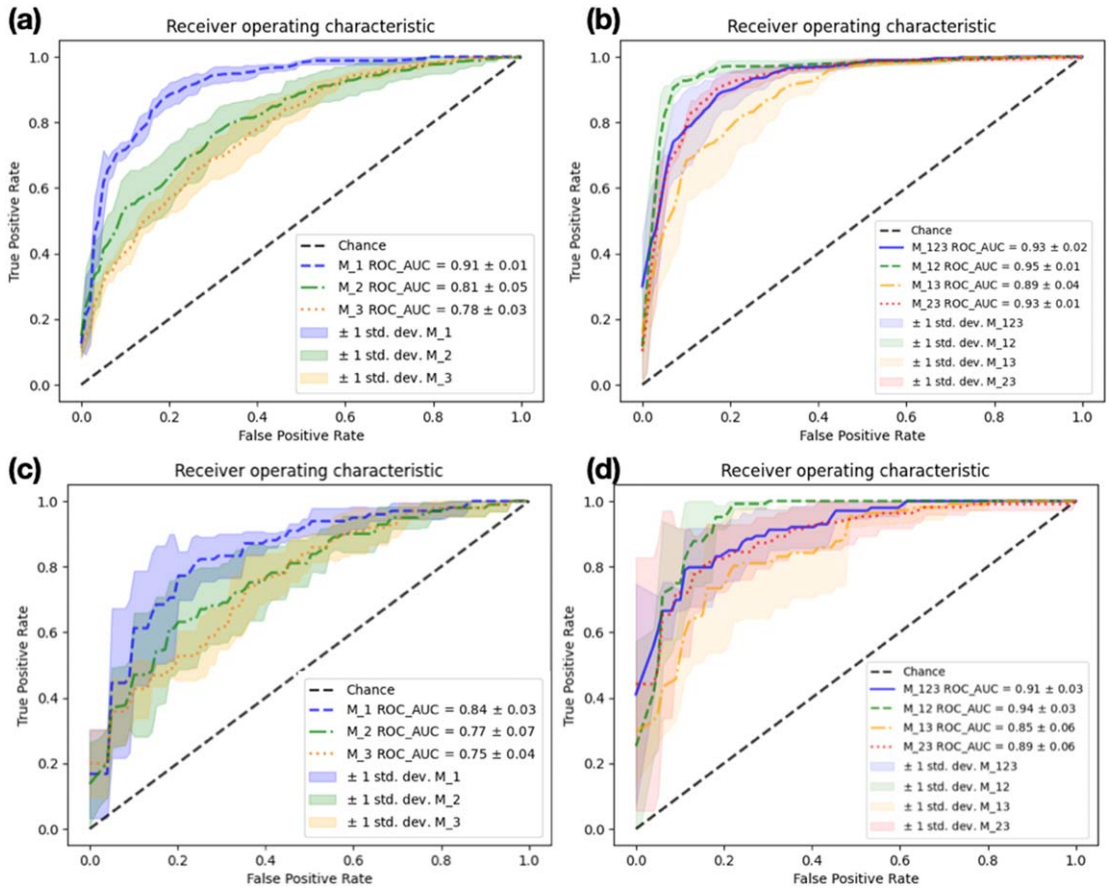

Figure 6. ROC plots and ROC_AUC values of the main models (M_1c, M_2c, and M_3), and the hybrid models (M_12c, M_13c, M_23c, M_123c) for validation (figures 6(a), (b)) and testing sub-datasets (figures 6(c), (d)).

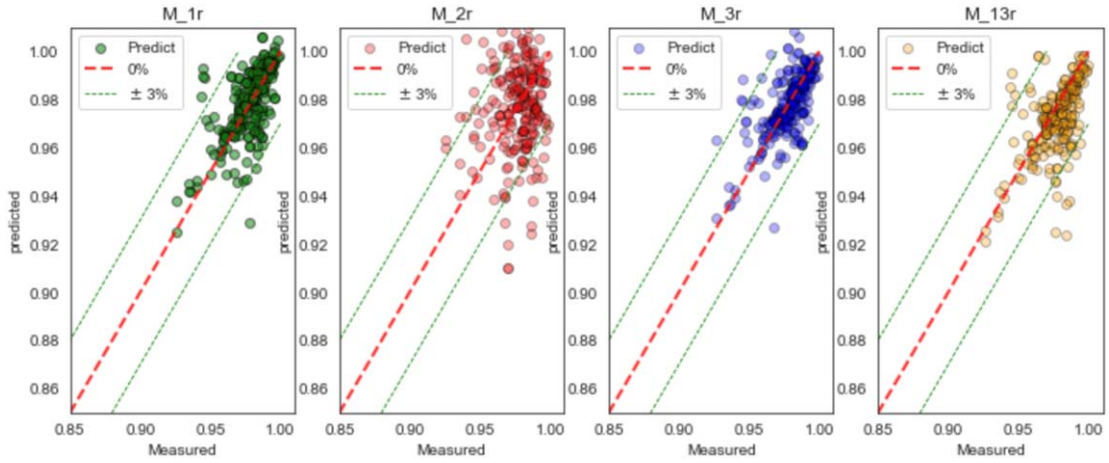

Figure 7. Regression results for the models M_1r, M_2r, M _3r, and M_13r with a 3% deviation (dotted green lines) from the ideal GPR distribution represented by the red line.

and ROC_AUC values presented comparable results for all models (table 3), demonstrating the potential benefits of these features for GPRs prediction. Indeed, for model classification, the models designed in this study demonstrated outstanding performance with similar or higher ROC_AUC values than the reported studies. However, while many published models did not report the model performance with the validation tests (Chan *et al* 2020a, Osman and Maalej 2021), the results obtained in this study using the validation dataset are also comparable (ROC_AUC values of 0.84 ± 0.01, 0.77 ± 0.05, and 0.75 ± 0.03 for M_1, M_2, and M_3 respectively). These results demonstrate the present models' suitability since the validation results are one of the

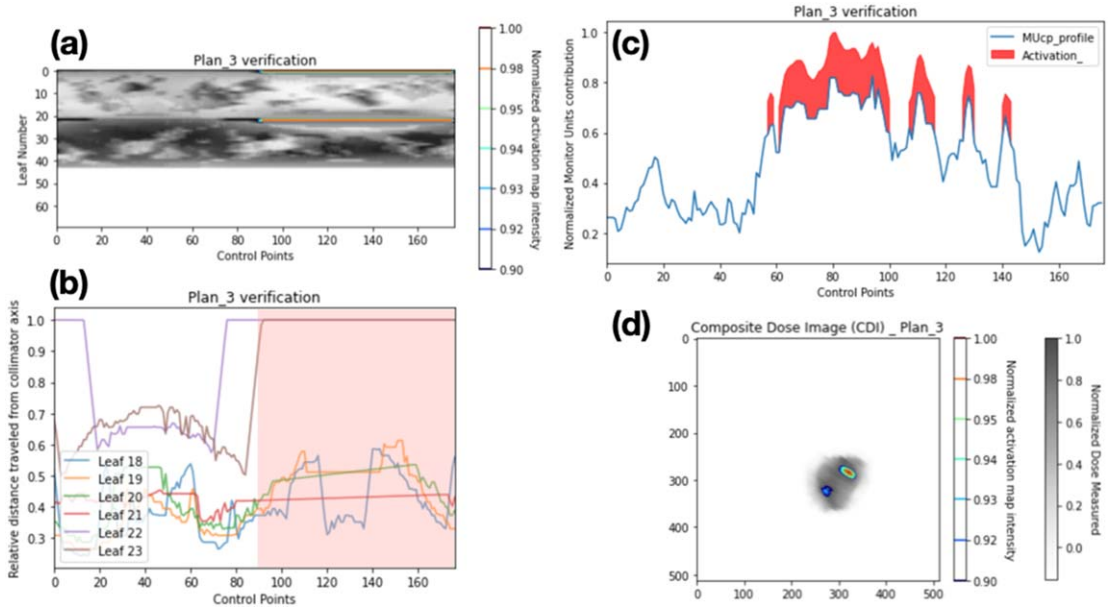

Figure 8. The activation maps of model M_1, M_2, and M_3 applied to features extracted from the 'passing' plan Plan_3. (a) Activation map from model M_1 applied to the modulation map. (b) Leaf trajectories corresponding to the activated regions, highlighting in red the control points of interest. (c) Activated regions, in red, from model M_2 applied to the respective MUcp profile. (d) Activation map from model M_3 applied to the CDI.

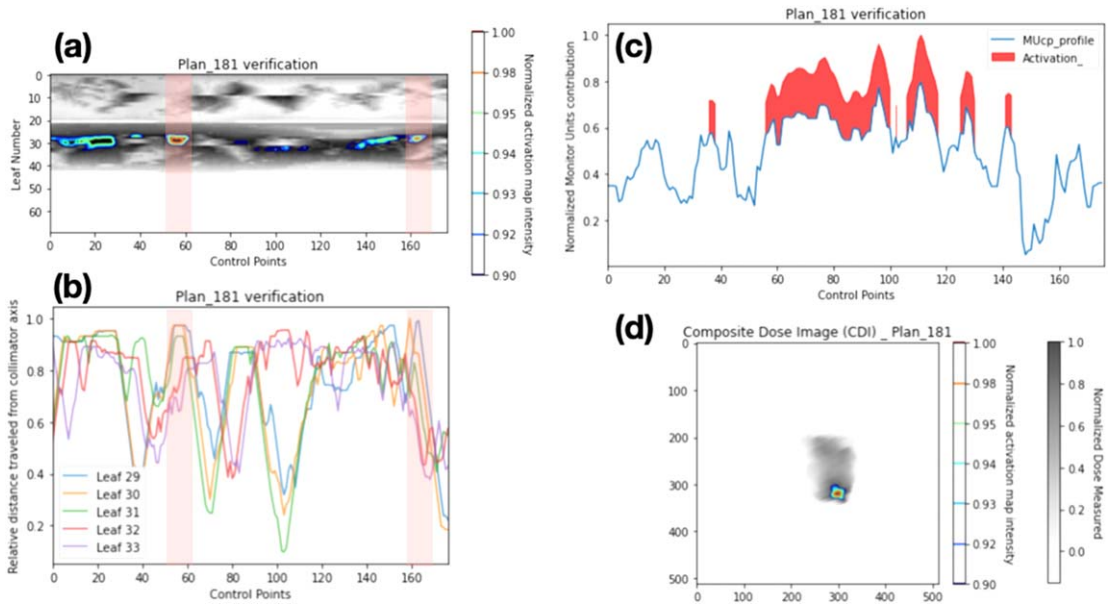

Figure 9. The activation maps of model M_1, M_2, and M_3 applied to features extracted from the 'failing' plan Plan_181. (a) Activation map from model M_1 applied to the modulation map. (b) Leaf trajectories corresponding to the activated regions, highlighting in red the control points of interest. (c) Activated regions, in red, from model M_2 applied to the respective MUcp profile. (d) Activation map from model M_3 applied to the CDI.

main approaches to verify the model generalization and the overfitting level; consequently, it is usual that these values are lower than those obtained by the training-testing dataset.

Following the already reported works (Tomori *et al* 2018, Hirashima *et al* 2020) and the discussion regarding model evaluation, we also confirm the improving effects of concatenating models using more than one feature category, especially from the validation dataset point of view, combining MM and CDI for model M_13 having ROC_AUC value of 0.91 ± 0.02 (figures 6, 7). However, the general improvement effects of concatenated models are still a field not completely explored and should be evaluated independently in each case because of the different origins and dimensions of the predictor features (Shin *et al* 2016, Li *et al* 2017). Furthermore, although the benefits of concatenating various multi-scale features have been reported, even in radiotherapy

(Hirashima *et al* 2020, Tomori *et al* 2020), concatenating too many features might compromise the model's performance and the training model (Li *et al* 2017). However, using concatenated models and controlling the different types of inputs might represent a technical advantage in mitigating premature or suboptimal gradient optimization (Tomori *et al* 2018), plus the benefit of implementing additional treatment plan features that describe treatment plan parameters related to dose deliverability during the same control points.

From the dataset conformation point of view, it is important to notice that the GPRs and modulation metrics ranges are susceptible to change between treatment units and anatomic regions (Wall and Fontenot 2020, 2021, Jin *et al* 2015). Thus, the previously reported models trained with their respective datasets (having a heterogeneous number of anatomic regions, beam energies, treatment units, and unbalanced GPR values) might potentially experience low data generalization and overfitting events (Payer *et al* 2016, Chen *et al* 2020), heading suboptimal predictions. Therefore, we deem that our datasets were designed using treatment plans for one single pathology (prostate), planned for two different treatment units (46.6% TB and 53.4% Halcyon, table 1), and ensuring that the passing and failing plans contribute equally to the dataset. Furthermore, with this dataset design and adopting the most straightforward CNN architectures, we intended that the extracted features by the CNNs correspond mainly to specific treatment conditions and, in turn, be able to associate physical or mechanical aspects to the final prediction. Consequently, we only explore horizontal flip for data augmentation. This rationale, from a practical point of view, might procure more robust models since the predicting process is highly focused on features with a real physical meaning and does not rely completely on random weighted feature extractions. Eventually (with further studies), tools like activation maps (Payer *et al* 2016) might be used to narrow specific treatment moments susceptible to contributing to a '*fail*' or lower GPR prediction, or to assist onboard adaptative therapy strategies. Accordingly, similar insights will be beneficial to develop ML solutions from a closer medical physics perspective, contemplating potential strategies to evaluate the model's reliability and consistency of in-house or commercial models dedicated to dose deliverability predictions. In this study, we proposed to evaluate the architecture model stability and the relevance of the 'learned' (extracted) features in the prediction performance, increasing systematically drop-out rates after each CNN layer (figure 5). With this method, we implicitly estimated for each model (1) the proportion of the minimum active nodes (i.e. remaining features) to maintain comparable prediction performances, and subsequently, (2) the potential random features extracted by the model that not necessarily contributes to the prediction.

From the model interpretability point of view, the reported CNN-based models dedicated to GPR predictions (Tomori *et al* 2018, 2020) do not offer straightforward ways to retrieve or identify the features associated with the predictions (Feng *et al* 2018, Chan *et al* 2020b), limiting the understanding and evaluation of the model quality because they were developed using dose distribution regions as predictors (Osman and Maalej 2021). These inputs do not provide enough explanatory parameters for plan deliverability analysis; hence, ML models considering high dimensional treatment parameters are also needed to contemplate the utility of retrieving the activation maps pinpointing specific hardware or dosimetric aspects that might influence the dose deliverability in a particular treatment moment (i.e. control point). Accordingly, and considering the mentioned utility of activation maps, figures 8 and 9 are a clear representation of the retrieved plan information associated with the prediction. However, despite the failing and passing activation maps localized distinctive regions, mainly for MMs, further studies are needed to verify that these highlighted changes in MLC position represent actual demanding hardware scenarios that might compromise the dose deliverability. Furthermore, this information might potentially support the setting of hardware tolerance limits for MLC trajectories or configuring TPS tools associated with the MLC sequencing algorithms (Varian Medical Systems 2018).

The GPR evaluation is widely used as a deliverability metric and is one of the worldwide standard tests for specific treatment verification (Miften *et al* 2018). However, it has been thoroughly questioned because of its arguable sensitivity to reflect or discriminate plan errors with potential clinical implications (Hussein *et al* 2017). Nevertheless, this study, rather than predicting just on metric, shows the promising opportunity to explore more treatment-associated parameters that can be part of an integral evaluation method of dose deliverability evaluation. We consider that this evaluation does not have to be enclosed by one single metric; hence, ML-based models in this matter will have to explore how to include new treatment parameters to predict relevant features contributing to a multiple-factors analysis to decide if the deliverability of a specific plan is acceptable or not. Additionally, we note that ML-based applications within treatment verification protocols are not intended to replace the quality assurance evaluation. Instead, ML models are recommended as part of decision-making tools to ease the evaluation workflow and reduce the number of dose measurements from suboptimal plans.

We acknowledge that this study was performed with limitations also identified in previously reported works. First, the dataset size is a fundamental factor related to ML model performance, especially for CNN-based models (Tomori *et al* 2018, 2020). However, considering that our dataset size is similar to or higher than others reported, our principal aim was to explore the suitability of three treatment features, and our results were consistent, encouraging further investigations. Similarly, we acknowledge that the extracted datasets were based on treatment plan information from one institution, and external verifications will be necessary to perform

further validations. Finally, we acknowledge that further studies are necessary to explore and evaluate the effects of including the intrinsic uncertainty of the dose detectors, the dose calculation, and mainly the uncertainty from the model itself (el Naqa and Murphy 2015, Avanzo *et al* 2020, el Naqa and Das 2020). We consider that including different sources of uncertainty in ML algorithm design is an essential field to be explored, which might increase the model's robustness and reliability, mainly if it is intended to be implemented in practice.

In summary, with this research, we aimed to contribute to three main gaps within the ML models predicting dose deliverability using CNN-based models. First, the implementation of new treatment features, especially with potential physical factors traceable by the activation maps. Also, the use of multiple feature inputs to increase the prediction performance. And finally, to opening the discussion about how to develop and understand ML applications in radiotherapy that might help to design new strategies to evaluate dose deliverability.

Conclusions

The MP, MUcp, and CDI are convenient features for dose deliverability predictive models implementing ML methods. Additionally, hybrid models including two or more input features are susceptible to improving the prediction performance compared to models with single features. Besides, decision-making strategies based on ML models might help to support new methodologies to evaluate dose deliverability within the patient-specific treatment verification protocols.

ORCID iDs

Paulo Quintero https://orcid.org/0000-0001-6574-1828
Craig Moore https://orcid.org/0000-0001-7409-8387
Andrew Beavis https://orcid.org/0000-0002-2519-0205

References

Agnew A, Agnew C E, Grattan M W D, Hounsell A R and McGarry C K 2014 Monitoring daily MLC positional errors using trajectory log files and EPID measurements for IMRT and VMAT deliveries *Phys. Med. Biol.* **59** N49–N63

Avanzo M *et al* 2020 Machine and deep learning methods for radiomics *Med. Phys.* **47** e185–e202

Berger L, François P, Gaboriaud G and Rosenwald J-C 2006 Performance optimization of the Varian aS500 EPID system *J Appl Clin Med Phys.* **7** 105–14

Campbell W, Olsen L A, Miften M, Goodman K A, Schefter T and Jones B L 2017 Using machine learning to predict physician-approved dose distributions for pancreatic SBRT *Int. J. Radiat. Oncol.***Biol.***Phys.* **99** S174 Eposter Session

Carlson J N K, Park J M I, Park S-Y, Park J M I, Choi Y and Ye S-J 2016 A machine learning approach to the accurate prediction of multi-leaf collimator positional errors *Phys. Med. Biol.* **61** 2514–31

Cha K H, Hadjiiski L, Samala R K, Chan H-P, Caoili E M and Cohan R H 2016 Urinary bladder segmentation in CT urography using deep-learning convolutional neural network and level sets *Med. Phys.* **43** 1882–96

Chan M F, Witztum A and Valdes G 2020a Integration of AI and machine learning in radiotherapy QA *Front Artif Intell.* **3** 76–84

Chan M F, Witztum A and Valdes G 2020b Integration of AI and machine learning in radiotherapy QA *Front Artif Intell.* **3** 76–84

Chauhan R, Ghanshala K K and Joshi R C 2018 Convolutional neural network (CNN) for image detection and recognition *ICSCCC 2018 - 1st Int. Conf. on Secure Cyber Computing and Communications* (Institute of Electrical and Electronics Engineers Inc.) pp 278–82

Chen R C, Dewi C, Huang S W and Caraka R E 2020 Selecting critical features for data classification based on machine learning methods *J Big Data* **7** 1–26

Chiavassa S, Bessieres I, Edouard M, Mathot M and Moignier A 2019 Complexity metrics for IMRT and VMAT plans: a review of current literature and applications *Br. J. Radiol.* **92** 20190270

Dillon J V *et al* 2017 TensorFlow Distributions [cited 2022 Sep 24]; Available from: https://arxiv.org/abs/1711.10604v1

el Naqa I, Bradley J D, Lindsay P E, Hope A J and Deasy J O 2009 Predicting radiotherapy outcomes using statistical learning techniques *Phys. Med. Biol.* **54** S9–30

el Naqa I and Das S 2020 The role of machine and deep learning in modern medical physics *Med. Phys.* **47** e125–6

el Naqa I and Murphy M J 2015 *What Is Machine Learning? Machine Learning in Radiation Oncology* (Cham: Springer) pp 3–11

Esch A, van, Huyskens D P, Hirschi L, Scheib S and Baltes C 2013 Optimized Varian aSi portal dosimetry: development of datasets for collective use *J Appl Clin Med Phys.* **14** 82–99

Ezzell G A *et al* 2009 IMRT commissioning: multiple institution planning and dosimetry comparisons, a report from AAPM Task Group 119 *Med. Phys.* **36** 5359–73

Feng M, Valdes G, Dixit N and Solberg T D 2018 Machine learning in radiation oncology: opportunities, requirements, and needs *Front Oncol* **8** 1–7

Granville D A, Sutherland J G, Belec J G and la Russa D J 2019 Predicting VMAT patient-specific QA results using a support vector classifier trained on treatment plan characteristics and linac QC metrics *Phys. Med. Biol.* **64** 095017

Han X 2017 MR-based synthetic CT generation using a deep convolutional neural network method *Med. Phys.* **44** 1408–19

Hesamian M H, Jia W, He X and Kennedy P 2019 Deep learning techniques for medical image segmentation: achievements and challenges *J Digit Imaging* **32** 582–96

Hirashima H *et al* 2020 Improvement of prediction and classification performance for gamma passing rate by using plan complexity and dosiomics features *Radiother. Oncol.* **153** 250–57

Hussein M, Clark C H and Nisbet A 2017 Challenges in calculation of the gamma index in radiotherapy—Towards good practice *Phys. Med.* **36** 1–11 Elsevier
Hussein M, Rowshanfarzad P, Ebert M A, Nisbet A and Clark C H 2013 A comparison of the gamma index analysis in various commercial IMRT/VMAT QA systems *Radiother. Oncol.* **109** 370–6 Elsevier
Ibragimov B, Toesca D A S, Chang D T, Koong A C and Xing L 2017 Deep learning-based autosegmentation of portal vein for prediction of central liver toxicity after SBRT *Int. J. Radiat. Oncol.***Biol.***Phys.* **99** E672
Interian Y *et al* 2018 Deep nets versus expert designed features in medical physics: An IMRT QA case study *Med. Phys.* **45** 2672–80
J K, R S, J F and S B 2015 Machine learning approaches for predicting radiation therapy outcomes: a clinician's perspective *Int. J. Radiat. Oncol. Biol. Phys.* **93** 1127–35
Jin X, Yan H, Han C, Zhou Y, Yi J and Xie C 2015 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.* **88** 20140577
Kimura Y, Kadoya N, Tomori S, Oku Y and Jingu K 2020 Error detection using a convolutional neural network with dose difference maps in patient-specific quality assurance for volumetric modulated arc therapy *Phys. Med. Phys. Med.* **73** 57–64
Lam D *et al* 2019 Predicting gamma passing rates for portal dosimetry-based IMRT QA using machine learning *Med. Phys.* **46** 4666–75
Li Y, Zhang T, Liu Z and Hu H 2017 *A Concatenating Framework of Shortcut Convolutional Neural Networks* (https://doi.org/10.48550/arXiv.1710.00974)
Li J *et al* 2019 Machine learning for patient-specific quality assurance of VMAT: prediction and classification accuracy *Int. J. Radiat. Oncol. Biol. Phys.* **105** 893–902
Liu Z *et al* 2019 A deep learning method for prediction of three-dimensional dose distribution of helical tomotherapy *Med. Phys.* **46** 1972–83
Low D A, Harms W B, Mutic S and Purdy J A 1998 A technique for the quantitative evaluation of dose distributions *Med. Phys.* **25** 656–61
Lustberg T *et al* 2018 Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer *Radiother. Oncol.* **126** 312–7
Masi L, Doro R, Favuzza V, Cipressi S and Livi L 2013 Impact of plan parameters on the dosimetric accuracy of volumetric modulated arc therapy *Med. Phys.* **40** 071718
McNiven A L, Sharpe M B and Purdie T G 2010 A new metric for assessing IMRT modulation complexity and plan deliverability *Med. Phys.* **37** 505–15
Meyer P, Noblet V, Mazzara C and Lallement A 2018 Survey on deep learning for radiotherapy *Comput. Biol. Med.* **98** 126–46
Miften M *et al* 2018 Tolerance limits and methodologies for IMRT measurement-based verification QA: recommendations of AAPM Task Group No. 218 *Med Phys.* **45** e53–83
Miri N, Keller P, Zwan B J and Greer P 2016 EPID-based dosimetry to verify IMRT planar dose distribution for the aS1200 EPID and FFF beams *J. Appl. Clin. Med. Phys.* **17** 292–304
Nguyen D *et al* 2017 A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning [cited 2019 Dec 16]; Available from: http://arxiv.org/abs/1709.09233
Nyflot M J, Thammasorn P, Wootton L S, Ford E C and Chaovalitwongse W A 2018 Deep learning for patient-specific quality assurance: identifying errors in radiotherapy delivery by radiomic analysis of gamma images with convolutional neural networks *Med. Phys.* [cited 2019 Dec 16];mp.13338. Available from: **46** 456–64
Ono T *et al* 2019 Prediction of dosimetric accuracy for VMAT plans using plan complexity parameters via machine learning *Med. Phys.* **46** 3823–32
Osman A F I and Maalej N M 2021 Applications of machine and deep learning to patient-specific IMRT/VMAT quality assurance *J. Appl. Clin. Med. Phys.* **22** 20–36
Park J M, Kim J and Park S 2019 Modulation indices and plan delivery accuracy of volumetric modulated arc therapy *J. Appl. Clin. Med. Phys.* **20** 12–22
Park J M, Kim J I, Park S Y, Oh D H and Kim S T 2018 Reliability of the gamma index analysis as a verification method of volumetric modulated arc therapy plans *Radiat. Oncol.* **13** 1–14
Park J M, Park S-Y and Kim H 2015 Modulation index for VMAT considering both mechanical and dose calculation uncertainties *Phys. Med. Biol.* **60** 7101–25
Payer C, Štern D, Bischof H and Urschler M 2016 *Regressing Heatmaps for Multiple Landmark Localization Using CNNs* (Cham: Springer) pp 230–8
Petroccia H M *et al* 2019 Spine SBRT with halcyon plan quality, modulation complexity, delivery accuracy, and speed *Front Oncol.* **9** 319–327
Quintero P 2020 pquinterome/MCS-calculation: Calculating the MCS for VMAT based on:' Masi*et al.* : Plan parameters and VMAT dosimetric accuracy - 2013'. Github. [cited 2020 Jul 27]. Available from: https://github.com/pquinterome/MCS-calculation
Sahiner B *et al* 2019 Deep learning in medical imaging and radiation therapy *Med. Phys.* **46** e1–36
Shin H-C *et al* 2016 Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning *IEEE Trans. Med. Imaging* **35** 1285–98
Tamura M, Matsumoto K, Otsuka M and Monzen H 2020 Plan complexity quantification of dual-layer multi-leaf collimator for volumetric modulated arc therapy with Halcyon linac *Phys. Eng. Sci. Med.* **43** 947–57
Tomori S *et al* 2018 A deep learning-based prediction model for gamma evaluation in patient-specific quality assurance *Med. Phys.* **45** 4055–65
Tomori S *et al* 2020 Systematic method for a deep learning-based prediction model for gamma evaluation in patient-specific quality assurance of volumetric modulated arc therapy *Med. Phys.* **48** 1003–18
Trullo R, Petitjean C, Nie D, Shen D and Ruan S 2017 *Joint Segmentation of Multiple Thoracic Organs in CT Images with Two Collaborative Deep Architectures* (Cham: Springer) pp 21–9
Valdes G *et al* 2017 Clinical decision support of radiotherapy treatment planning: a data-driven machine learning strategy for patient-specific dosimetric decision making *Radiother. Oncol.* **125** 392–7
Valdes G, Scheuermann R, Hung C Y, Olszanski A, Bellerive M and Solberg T D 2016 A mathematical framework for virtual IMRT QA using machine learning *Med Phys. AAPM - Am. Assoc. Phys. Med.* **43** 4323–34
2018 TPS New Features Workbook v15.6Varian Medical Systems - Manual User
Wall P D H and Fontenot J D 2020 Application and comparison of machine learning models for predicting quality assurance outcomes in radiation therapy treatment planning *Inform Med. Unlocked.* **18** 1–12
Wall P D H and Fontenot J D 2021 Quality assurance-based optimization (QAO): Towards improving patient-specific quality assurance in volumetric modulated arc therapy plans using machine learning *Phys Med.* **87** 136–43
Workshop on unsupervised, transfer Learning PB-I, undefined 2012 Learning PB-I workshop on unsupervised and transfer, 2012 undefined. Autoencoders, unsupervised learning, and deep architectures. jmlr.org. [cited 2019 Oct 20]; Available from:http://jmlr.org/proceedings/papers/v27/baldi12a/baldi12a.pdf

Wolterink J M, Dinkla A M, Savenije M H F, Seevinck P R, van den Berg C A T and Išgum I 2017 *Deep MR to CT Synthesis Using Unpaired Data* (Cham: Springer) pp 14–23

Xiang L *et al* 2018 Deep embedding convolutional neural network for synthesizing CT image from T1-Weighted MR image *Med. Image Anal.* **47** 31–44

Yan Z *et al* 2016 Multi-Instance deep learning: discover discriminative local anatomies for bodypart recognition *IEEE Trans. Med. Imaging* **35** 1332–43

Zhao C, Carass A, Lee J, He Y and Prince J L 2017 *Whole Brain Segmentation and Labeling from CT Using Synthetic MR Images* (Cham: Springer) pp 291–8

Zhen H, Nelms B E and Tomé W A 2011 Moving from gamma passing rates to patient DVH-based QA metrics in pretreatment dose QA *Med. Phys.* **38** 5477–89