UNIVERSITY OF HULL

Virtual patient-specific treatment verification using Machine Learning methods to assist the dose deliverability evaluation of radiotherapy prostate plans

being a thesis submitted in fulfilment of the

requirements for the degree of

Doctor of

Philosophy

in the University of Hull

by

Paulo Alejandro Quintero Mejia, M.Sc.

October 2022

Acknowledgements

First, I would like to thank my supervisors for their high commitment, compromise, and kind help with this research during all the time in this journey, especially during the lockdown times. Thank you for motivating me and making me feel enthusiastic about my work.

In the same way, I would like to thank my beloved family and friends for helping me and supporting me with compassion, understanding, and a great sense of humour.

Finally, I would like to thank the University of Hull and the NHS because all the financial and scientific support needed to develop this research.

Publications and Conferences

Manuscripts:

- Quintero P, Cheng Y, Benoit D, Moore C, Beavis A. *Effect of treatment planning system parameters on beam modulation complexity for treatment plans with single-layer multi-leaf collimator and dual-layer stacked multi-leaf collimator*. British Journal of Radiology. https://doi.org/101259/bjr20201011
- Quintero P, Benoit D, Cheng Y, Moore C, Beavis A. *Evaluation of the dataset quality in gamma passing rate predictions using machine learning methods.* British Journal of Radiology (Accepted, March 2022)
- Quintero P, Benoit D, Cheng Y, Moore C, Beavis A. *Machine learning-based predictions of gamma passing rates for virtual specific-plan verification based on modulation maps, monitor unit profiles, and composite dose images.* Journal of Physics in Medicine and Biology (Accepted, September 2022)
- Quintero P, Cheng Y, Benoit D, Moore C, Beavis A. *Implementation of one support-decision tool for patient-specific treatment verification based on ML models*. (In progress)

Conferences and Posters:

- Quintero P, Cheng Y, Benoit D, Moore C, Beavis A. *Gamma Passing Rates prediction of prostate treatments based on machine learning methods.* [Poster]. AI in practice, January 2021. British Institute of Radiology.
- Quintero P, Cheng Y, Benoit D, Moore C, Beavis A. *Challenges in Gamma Passing Rates Prediction using artificial intelligence methods. A comprehensive analysis of the current state of art.* [Presentation]. BIR Annual Radiotherapy and Oncology meeting, March 2021. British Institute of Radiology.
- Quintero P, Benoit D, Cheng Y, Moore C, Beavis A. *High-dimensional beam modulation complexity for gamma passing rates prediction using AI methods*. [Poster]. ESTRO-2021 meeting, August 2021. European Society for Radiotherapy and Oncology
- Quintero P, Benoit D, Cheng Y, Moore C, Beavis A. *The dataset heterogeneity matters: A machine learning study of dataset conformation effects on model performance for dose deliverability prediction*. [Poster]. 64th AAPM-2022, July 2022. American Association of Physicists in Medicine.
- Quintero P, Benoit D, Cheng Y, Moore C, Beavis A. *Gamma passing rate predictions based on automatic feature extraction of modulation maps and monitor unit profiles: A machine learning approach for virtual specific-plan verification*. [Poster]. 64th AAPM-2022, July 2022. American Association of Physicists in Medicine.
- Quintero P. *Physical aspects of Machine Learning applications in radiotherapy*. [Presentation]. National Conference in Medica Physics, October 2022. Colombian Association of Medical Physics.

Abstract

Machine Learning (ML) methods represent a potential tool to support and optimize virtual patient-specific plan verifications within radiotherapy workflows. However, previously reported applications did not consider the actual physical implications in the predictor's quality and model performance and did not report the implementation pertinence nor their limitations. Therefore, the main goal of this thesis was to predict dose deliverability using different ML models and input predictor features, analysing the physical aspects involved in the predictions to propose a reliable decision-support tool for virtual patient-specific plan verification protocols.

Among the principal predictors explored in this thesis, numerical and high-dimensional features based on modulation complexity, treatment-unit parameters, and dosimetric plan parameters were all implemented by designing random forest (RF), extreme gradient boosting (XG-Boost), neural networks (NN), and convolutional neural networks (CNN) models to predict gamma passing rates (GPR) for prostate treatments. Accordingly, this research highlights three principal findings. (1) The dataset composition's heterogeneity directly impacts the quality of the predictor features and, subsequently, the model performance. (2) The models based on automatic extracted features methods (CNN models) of multi-leaf-collimator modulation maps (MM) presented a more independent and transferable prediction performance. Furthermore, (3) ML algorithms incorporated in radiotherapy workflows for virtual plan verification are required to retrieve treatment plan parameters associated with the prediction to support the model's reliability and stability. Finally, this thesis presents how the most relevant automatically extracted features from the activation maps were considered to suggest an alternative decision support tool to comprehensively evaluate the causes of the predicted dose deliverability.

Contents

List of Figures

Acronyms

- RT: Radiation Therapy
- ML: Machine Learning
- GPR: Gamma Passing Rates
- MLC: Multi-Leaf Collimator
- VMAT: Volumetric Modulated Arc Therapy
- TPS: Treatment Planning System
- Linac: linear accelerator
- RF: Random Forest
- XG-Boost: Extreme Gradient Boost
- CNN: Convolutional Neural Network
- NN: Neural Network
- MM: Modulation Map
- MU: Monitor Units
- CP: control points
- MUcp: Monitor Units per control point
- CDI: Composite Dose Image
- MAE: Mean Absolute Error
- RMSE: Roost Mean Square Error

Chapter 1 Introdcution

Machine learning (ML) applications in radiation therapy (RT) dedicated to patient-specific plan verification require studies to define what model-reliability involves and what are the actual applications of those physical aspects, given by the predictors, linked to the dose deliverability. While other ML applications for organ contouring or detection can be intuitively assessed and corrected during the RT plan design-optimization process [1], the dose deliverability predictions based on ML models face several hidden challenges in their actual application due to the implicit uncertainty of the ground truth definition of a predicted 'passing' or 'failing' plan [2,3]. Specifically, the no control of physical aspects within the dataset, such as the dose detection device, the treatment unit hardware, the dose optimization/calculation software, and the clinical configurations established in each RT facility, impact the minimum conditions needed to decide if a specific treatment is suitable for delivering to the patient [4–6]. Thus, the dose deliverability analysis supported by ML models should become a more customized protocol to attempt more robust prediction.

Considering already reported ML models predicting gamma passing rates (GPRs) (Section 2.3.1), the main technical aspects considered in their designs were: the kind of predictors, the ML algorithms, and the dataset size [2]. However, their potential applications in practice, the predictors' quality, and their technical limitations have not been thoroughly discussed, which generates important gaps in the reliability of the published ML models [3]. Therefore, this research addressed these aspects through a series of studies designing an ML model that predicts GPR values, oriented to transfer modeled features to physical parameters from the treatment delivery, promoting reliable ways to verify the prediction quality and model stability, and suggesting potential indicators of technical tolerance limits for further plan designing.

Accordingly, this thesis is oriented to contribute to the medical physics field, implementing ML models to support the RT virtual dose deliverability evaluation and defining more comprehensive challenges and limitations of these models predicting GPRs. Thus, this research was developed in four main steps, (I) extracting and calculating all the complexity metrics and plan parameters as predictor features (Chapter 4), (II) verifying the effect of the dataset assembling conditions on model performance (Chapter 5), (III) implementing high-dimensional features to avoid the numeric calculated predictors (Chapter 6), and (IV) proposing the minimum aspects needed to implement these ML models in practice (Chapter 7). Correspondingly, the main contributions of this thesis are (I) the new modulation complexity metrics for treatments based on dual-layer MLC models, (II) the demonstration of dataset-composition effects in prediction quality, (III) the suitability of high-dimensional features implementation to predict

GPRs, and (IV) the verification of model reliability extracting the activated maps from the high-dimensional features, and including them within an RT plan verification workflow.

Considering the above-mentioned and the available resources provided by the University of Hull and the Hull University Teaching Hospitals NHS Trust, the whole research developed in this thesis was oriented to demonstrating the following hypothesis: "***It is possible to use ML models to support virtual patient-specific treatment verification in prostate radiotherapy, retrieving critical physical aspects involved in dose deliverability.***" Moreover, this hypothesis encompasses the dataset quality, the virtual patient-specific treatment verification workflow design, and the verification of model reliability by including the physical plan parameters associated with the prediction.

Considering these three aspects associated with the hypothesis, the principal aim of this thesis was **to explore the best dataset and model configuration to predict GPR values retrieving specific features corresponding to physical aspects involved in the dose deliverability**. Accordingly, this objective was developed intending to answer the following research questions:

1. Which input features are more convenient for GPR predictions using ML models?
2. What dataset configuration is optimal for a reliable GPR modelling performance?
3. Are the ML models based on high-dimensional input features suitable for GPR predictions?
4. What decision-support strategy for virtual plan verification might be beneficial in practice?

The structure of this thesis, intending to address the research questions, is as follows**. Chapter 2** gives a literature review with the main generalities and technical aspects of RT, ML, and ML applications in RT. Chapter 2 includes the background necessary to understand the related state of the art and the different technical aspects referenced in this thesis that will be described in the introduction sections of each following chapter.

In **Chapter 3**, a four-steps workflow of this thesis was included describing the methods and materials implemented. This section summarises the research pathway followed in this study, from the predictor features extraction and the dataset assembling evaluation to the predictions based on high-dimensional features to design a decision-support rationale for virtual patient-specific treatment verification. Also, additional information regarding the materials used in this thesis were summarised in this section.

Chapter 4 describes the modulation complexity metrics, feature extraction procedures, and all the predictors considered in this thesis. Additionally, since many plans from the available dataset were designed and delivered in a relatively new treatment unit (Halcyon-v2) with a different multi-leaf collimator (MLC) design, a study was included in this chapter (Section 4.2) to propose and validate modulation complexity metrics tailored to the new hardware conditions. Consequently, in this chapter, the first question regarding the suitable inputs for GPR predictions was initially addressed, exploring the origin of the modulation complexity metrics, their representation, and their impact on GPR values or dose deliverability.

In **Chapter 5**, the effects of the dataset configuration on the ML model performance were explored, addressing the first and second research question. Once the predictors were discussed previously, this section intended to explore and demonstrate the potential effects of the dataset composition on the model prediction performance. Consequently, the results included in this section opened the discussion about the variation of the potential predictors that the model relies on to perform a prediction, suggesting a poor interpretability and control of physical aspects linked to any GPR prediction based on ML methods. Accordingly, new modelling approaches were proposed using high-dimensional features, as shown in Chapter 6.

Chapter 6 addressed the third research question, demonstrating the suitability of high-dimensional features associated with modulation complexity to predict GPR values. In this section, straightforward model architecture designs were considered to control the changes in the input features, which were extracted from the variations of treatment unit parameters associated with the MLC movements, the gantry speed, and the dose rate. This section demonstrates that features directly represented by physical aspects of the treatment unit performance during the treatment time are suitable predictors and might improve the model's reliability and interpretability since it provides an understanding of actual patterns or hardware conditions related to dose deliverability.

In **Chapter 7**, a practical application of the model designed in Chapter 6 is proposed to generate a decision-support strategy for virtual plan verification, according to the last (4th) research question. This practical application was designed to explore the potential advantages of using the model's activated feature maps, retrieving specific treatment physical aspects that might contribute in three main technical instances: (I) the model interpretability and reliability, understanding the potential technical properties related to the prediction, (II) the identification of critical hardware conditions to set new tolerance limits in the dosimetric planning or re-planing scenarios, and (III) the understanding of the model limitations and obsolescence scenarios in practice.

Finally, in **Chapter 8**, the conclusions from all previous chapters were summarised and oriented to consider the feasibility and requirements to develop *ML models to support virtual patient-specific treatment verification in prostate radiotherapy, retrieving critical physical aspects involved in dose deliverability.*

Chapter 2 Background

This section describes the general aspects of one standard RT treatment unit, including the multi-leaf collimator (MLC) as a beam modulator hardware device (**Section2.1**). Indeed, the MLC trajectories were used to calculate the complexity metrics used in this thesis as GPR predictors. Additionally, an overview of the treatment modality (volumetric modulated arc therapy - VMAT) applied to all dataset's plans and the software (treatment planning system - TPS) used to generate those treatments are also described. Next, a general RT workflow is described to contextualize the RT area to which this thesis aims to contribute. On the other hand, from the ML point of view, **Section 2.2** summarises the basic concepts of ML, including a description of the algorithms used in this thesis. Finally, **Section 2.3** describes the ML applications in RT predicting GPRs.

2.1 Radiotherapy

RT is the therapy dedicated to cancer treatments using ionizing radiation. The World Health Organization and the International Atomic Energy Agency have considered it one of the principal cancer treatment branches since up to 48% of the patients should receive RT as part of their treatment for palliative or curative purposes [7,8]. RT can be classified as external or internal, depending on the radiation source location. External radiotherapy, the most common type, is delivered by machines that provide conformed radiation energy from outside the patient, whereas internal radiotherapy can be classified as brachytherapy and radionuclide therapy. Respectively, brachytherapy consists of treatments using solid radioactive materials to be locate within the natural body cavities (*i.e.,* gynaecological regions), within the body interstice (*i.e.,* prostate), or superficially. The sealed radiation sources can be located through hollow applicators and needles, or by contact. Contrastingly, the radionuclide therapy consists of injection of liquid radioactive material linked to biomarkers needed to take the radionuclide to the tumoral region. In this thesis, the treatment plans retrieved to conform the dataset were planned with external radiotherapy.

2.1.1 Treatment Unit

The most common treatment unit in external radiotherapy is a linear accelerator (linac) of electrons, which are generated by the thermionic effect and are ejected by the electron gun. The ejected electron bunches are accelerated by radiofrequency pulses through one waveguide system to be subsequently impacted against a high atomic-number target (usually tungsten), generating X-ray photons by the Bremsstrahlung effect [9]. Finally, the radiation is collimated and modulated by jaws and mechanical devices controlled by computer systems delivering to the patient the planned dose, using as a position reference guide one common geometrical point

located between the rotation axes of the gantry, the collimator, and the couch, known as the isocentre (Figure 2.1).

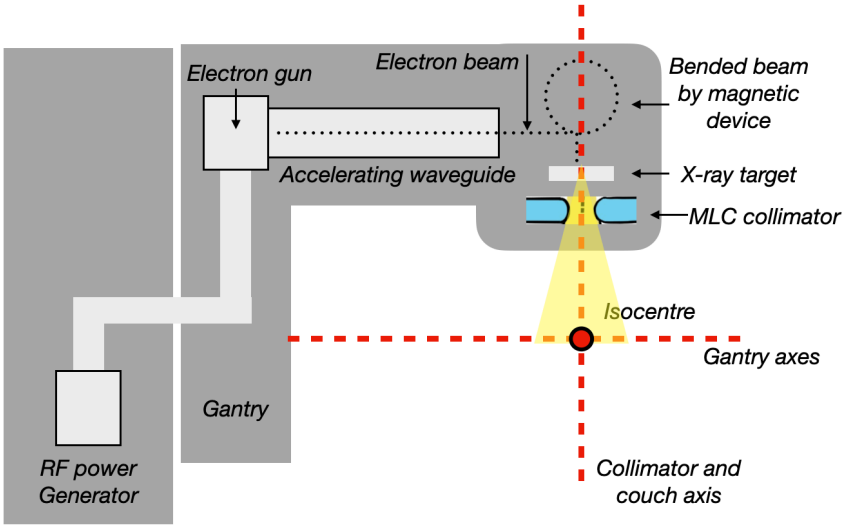


Figure 2.1 Schematic diagram of one radiotherapy treatment unit known as c-arm linac. Here are represented the radiofrequency (RF) generator, the gantry with the waveguide, and the unit head with the bending magnet, target, and multi-leaf collimator (MLC) array. Adapted from [9].

2.1.2 The Multi-Ceaf collimator - MLC

The MLC is a linac component that sets the shape of the radiation beam and modulates it to deliver a specific dose distribution through the calculated pattern of leaves trajectories. The shape and model of this device influence the RT performance in terms of dosimetric factors such as dose distribution complexity (dose map resolution), dose fall-off, and dose transmission trough and between the leaves [10–13]. The MLC of Varian linear accelerators (Varian Medical Systems - Palo Alto, USA) has rounded tips, as shown in Figure 2.2, generating a dose region named dosimetric leaf gap (DLG) [13,14]. The DLG is a factor measured in the machine commissioning process included in TPS and is related, along with faster and thinner MLC, to the dose fall-off and conformity within complex geometries [15–17].

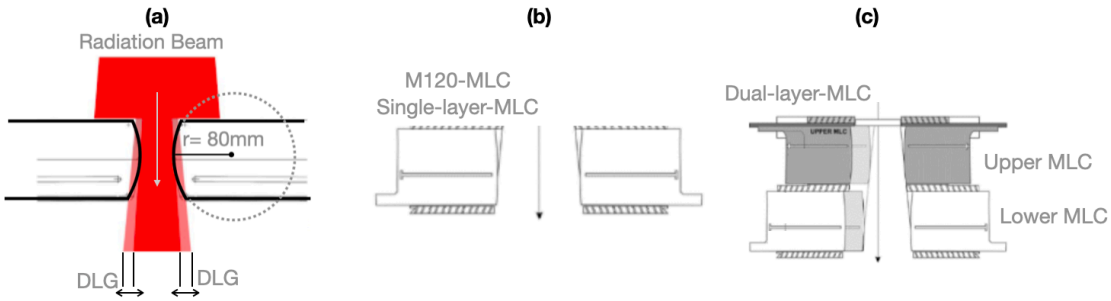


Figure 2.2 (a) Image representation of DLG, (b) Leaf model of Millennium 120-MLC (M120-MLC) with radius curvature of 80 mm and thickness of 67 mm, (c) Leaf model of dual-layer MLC configuration with a radius of curvature of 234 mm and thickness of 77 mm. Taken and adapted from [25, 26].

In this thesis two treatment units were implemented, the TrueBeam and Halcyon-v2 (Section3.2.1). The Varian TrueBeam is a conventional c-arm linac model with the option to have the millenium-120 MLC with 120 leaves of tungsten, each with a radius curvature of 8 cm (Figure 1.2.a). The maximum conformed radiation fields by this MLC model are up to 40 cm wide and long, and the centred 32 pair-leaves have a width of 5 mm at the isocentre while the remaining 28 pairs are 10 mm wide. In contrast, the Halcyon-v2 (Varian Medical Systems - Palo Alto, US) is a ring-linac with jaw-free mode that has a stacked-staggered MLC with two layers of leaves (distal and proximal to the linac target) offset by 5 mm (Figure 1.2.b). As described in the work of Cozzi et al. [18], each leaf has a 10 mm width projected at isocentre and has an effective conformity-resolution of 5 mm because of the overlap arrangement. Furthermore, the Halcyon-v2 allows independent displacements of the proximal and distal layers simultaneously, resulting in more modulation possibilities [18–23].

2.1.3 Volumetric Modulated Arc Therapy - VMAT

VMAT is an RT modality that continually delivers a series of beam fluencies with variations on the beam intensity (*i.e.,* dose rate) modulated by the MLC during the gantry rotation across the patient. These variations in beam fluence and dose rate during the arc trajectory conform and modulate volumetric dose distributions to cover the target volume with the desired dose and simultaneously avoid the surrounding organs at risk (Figure 1.3) [24–26]. Consequently, this treatment modality can deliver a highly conformed dose in a short time, representing lower patient toxicity [27] and reducing the potential errors by patient motion during the treatment delivery [9]. However, this technique implies demanding hardware conditions, especially for the MLC movements and the gantry speed, that might compromise the accuracy of the dose delivered according to the planned treatment. For this reason, empiric2al modulation metrics based on these hardware parameters have been explored as dose deliverability predictors (Section 4.1) [28].

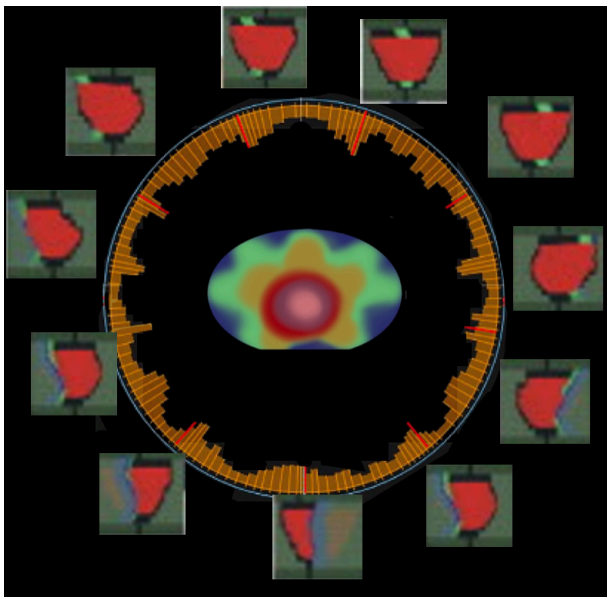
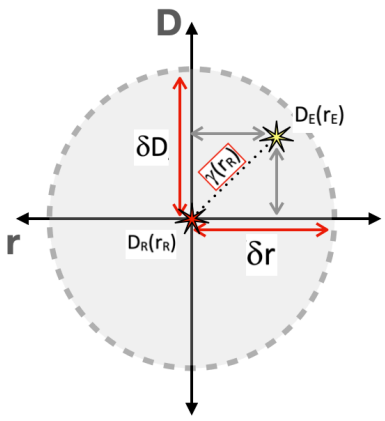


Figure 2.3 VMAT representation with the different beam intensities across the arc surrounding the patient. The dose conformation is achieved by the dose rate variation and the multi-leaf collimator, which is changing their leave positions to modulate the beam fluency. The dose is delivered to conform the target volume and to reduce the dose contributions to the healthy tissue.

2.1.4 Gamma Index and Gamma Passing Rate

In RT, the dose deliverability evaluation has been studied using different dose measurement protocols, being the gamma index evaluation the most implemented metric worldwide [5]. The gamma index (γ), developed by Low et al. [29], is a geometric approach to comparing the displacement between two dose-distributions points in a region of interest. As is displayed in Figure 2.4, γ is the minimum distance between one reference dose point and one measured dose point, given by the dose and geometric variations considered as evaluation criteria ($\delta D/\delta r$) [5]. Consequently, the GPR is the percentage of all measured points within the dose distribution representing γ values lower than 1. In other words, is the proportion of all dose points that are acceptably close enough according to the dose and distance criteria.



$$\Delta D(r_R, r_E) = D_E(r_E) - D_R(r_R)$$

$$\Gamma(\mathbf{r_R}, \mathbf{r_E}) = \sqrt{\frac{\Delta r^2(\mathbf{r}_R, \mathbf{r}_E)}{\delta r^2} + \frac{\Delta D^2(\mathbf{r}_R, \mathbf{r}_E)}{\delta D^2}}$$

$$\gamma(\mathbf{r}_R) = min\{\Gamma(\mathbf{r_R}, \mathbf{r_E})\} \forall \{\mathbf{r_E}\}$$

Figure 2.4 Representation of gamma Index calculation. The x and y axis represent distance and dose, respectively. The reference point (r_R) is marked with the centred red cross, representing an evaluated point to be compared to the measured dose points (yellow points, r_E) represented in the blue line. The δD and δr are the dose/distance acceptance criteria (e.g., $\delta D/\delta r$ = 3%/3 mm), creating an acceptance ellipse around the reference point. Adapted from [5,29].

2.1.5 Treatment Planning System - TPS

The TPS is the software dedicated to designing and calculating RT treatments, having a 2D/3D image visualization interface suitable for target volume contouring, irradiation beam designing, and dose calculation [30] (Figure 2.5). For the latter, the TPS uses dose calculation algorithms based on electron transport approximations and the electron densities extracted from the CT images given by the Hounsfield Units (HU) [30,31]. Additionally, the TPS is dedicated to the inverse optimisation process needed to generate VMAT treatment plans. This process sets the desired gantry arc trajectory, the target volume dose, and the organs at risk dose.

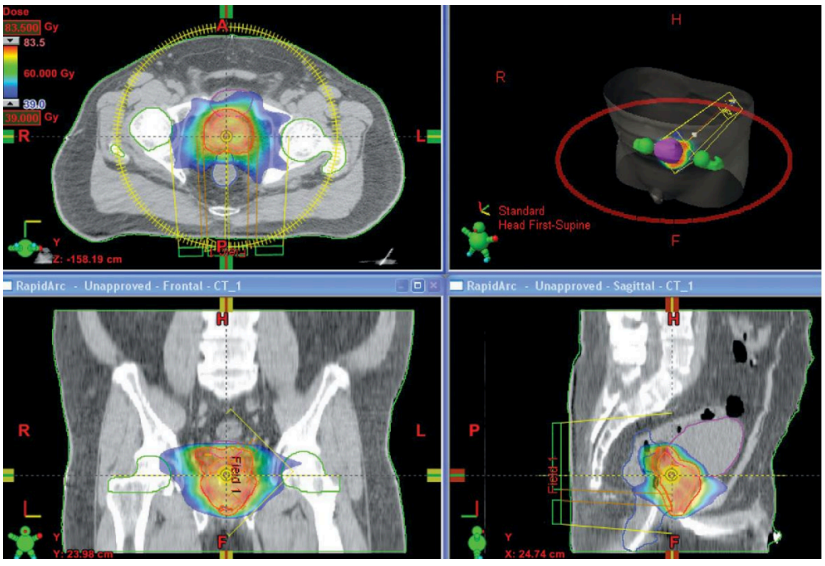


Figure 2.5 TPS visualization with the 2D/3D interface of the axial, coronal, and sagittal view of the CT images corresponding to the patient anatomy with the target volume delimited, the field array, and the dose distribution.

Eclipse-v15.6 TPS (Varian Medical Systems - Palo Alto, U.S.) performs the inverse optimisation of VMAT plans with the Photon Optimiser (PO) algorithm [32]. The PO, based on a direct aperture optimization process, uses a multi-resolution (MR) approach with fast and periodical calculations of the dose distribution, starting with a lower number of dose calculation segments and initial MLC positions conforming to the target volume. When this optimisation is continued, and the MR level increases, the dose calculation segments also increase, interpolating the MLC positions to obtain new leaf apertures that correspond to the improved dose distribution. During the MR optimisation, the dose calculation accuracy increases as the number of dose segments increases with a maximum separation of 2-4 degrees, depending on the arc span [33].

The different TPS options have parameters that impact the final dose fluence by the hardware setting parameters such as MLC velocity, gantry speed, and dose rate [34–36], as it will be discussed in Section 4.2. Consequently, if these TPS parameters are not handled properly, treatments with challenging dose requirements may bring unrealistic or high demanding machine conditions (i.e., highly modulated plans), reducing the accuracy of dose delivery [37,38].

2.1.6 RT Workflow

The radiotherapy workflow (Figure 2.6) consists of one assessment moment directed by a Physician that decides if the patient is or is not a candidate for radiotherapy treatment based on laboratory information (*i.e.,* pathology examination, or genetic tests, such as Oncogene) and previous clinical examination. If the patient requires radiotherapy, they need to have a computerized tomography (CT) in a specific position, ensuring the setup reproducibility and the same immobilization conditions as in the treatment room. With this CT and possibly other modality images, such as Positron Emission Tomography (PET) or Magnetic Resonance Imaging (MRI), the physician contours the organs at risk (OARs) and the gross tumour volume (GTV), defined as the structures detected by visual changes [39,40]. With those volumes, the Medical Physicist and dosimetrists calculates the optimal dose distribution to get better dose coverage and OARs sparing results. These calculations might be verified by computational and experimental protocols involving radiation detectors and alternative dose calculation systems. Subsequently, the original set-up of the patient is verified through images in the treatment room. If all technical conditions are achieved successfully, radiation treatment is delivered, and verification images evaluate the pertinence of a possible adaptation to the therapy during the whole treatment. Finally, after fulfilling all the programmed schedules, the patient is sent for follow-up [1].

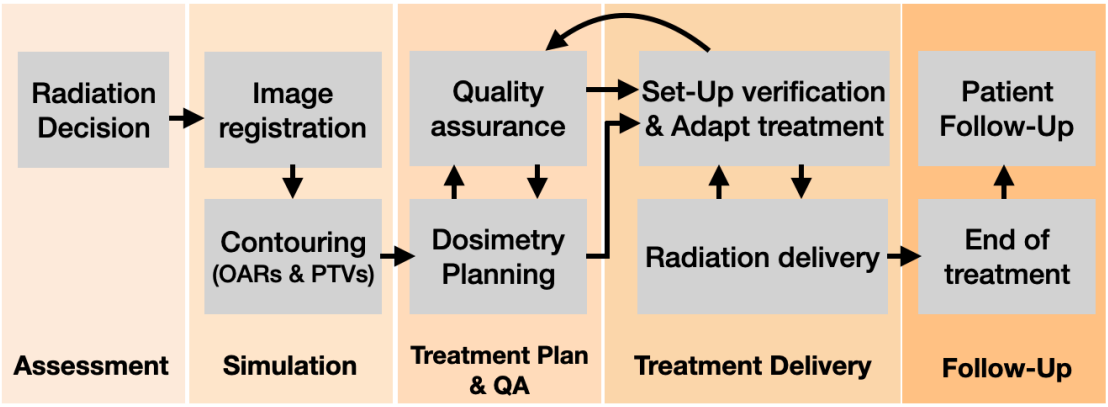


Figure 2.6 Radiotherapy standard workflow displaying each step where ML applications can be implemented, assisting the treatment prognosis (Assessment). Synthetic image generation or image reconstruction, Image detection, or segmentation (Simulation). Error prediction and dose deliverability evaluation (Treatment Delivery). Survival or end-points prediction (Follow-Up).

Considering this general RT workflow, it is essential to note that this thesis intends to improve the "Treatment Plan & QA" section, developing virtual patient-specific plan verification tools to avoid the dose measurements of plans that might not pass the physical tests, making this step more efficient. However, this virtual verification tool might also result in high interest in the "Treatment Delivery" step due to the onboard plan adaptation therapies that have been increasingly implemented [41–43]. Here, new adapted plans are automatically generated if significant changes in the patient are evident through the set-up verification images. Hence, while the patient is positioned in the treatment room, a new plan is generated, intending to be delivered immediately to avoid changes in the patient's condition. Thus, a virtual verification of these automatically created plans might help make a fast evaluation protocol to support the accuracy of treatment delivery evaluation.

2.2 Machine Learning Algorithms

Artificial Intelligence (AI) is an active field of computer sciences that comprehends the models and computer processes based on ‘knowledge learned’ by previous input features applied to new information scenarios, emulating how humans learn [44]. The name of AI was proposed by John McCarthy in 1956 [45,46], settling a starting point of several developments based on this rationale [47–49], such as the first Machine Learning (ML) publication in 1959 about one program that plays checkers using previously gained information [49].

ML is an AI division dedicated to data processing using training information to predict accurate outcomes from limited experienced datasets, making data generalizations without direct programmed instructions [49]. In the middle of the '90s and the beginning of the '00s, ML

becomes an accessible tool for image recognition and medical purposes because of its outstanding performance in this area [46,50]. Furthermore, although ML demands high expertise in features representation modeling, it has been a field with a high and fast development in the Radiation Therapy (RT) area [1].

ML comprehends a series of highly developed programming methods dedicated to data analysis. Its applications include pattern recognition, computer vision, spacecraft engineering, economy, social media, sentiment analysis, computational biology, and biomedical applications [51]. Generally, an ML algorithm does not require extensive computer codes to resolve a particular task. Instead, its programming architecture parameters improve through iterative periodical events, adapting the desired outcomes based on minimizing objective functions. This process is called training, in which the input data are simultaneously provided with the desired outcomes. Then, the algorithm 'learns' (through various statistical methods) to optimize its performance and generalize the predicted outcomes to new (or unseen) input data [1,52,53].

The ML models can be categorized in terms of how the 'learning task' is performed as unsupervised, supervised, reinforced, and transferred [46,54]. In supervised learning, the model algorithm must have training inputs and known desired outputs (e.g., categorizing functions by labels). In contrast, in unsupervised learning, the outcomes are groups of structures instead of referenced known outputs (e.g., clustering functions for image recognition). Reinforcement learning is determined by the interaction with its dynamic environment to predict specific long-term responses (e.g., exploring possible scenarios and responses of cellular growth). Finally, the transferred learning use information from other training sets to perform another task improving the data processing, and it is applicated when incomplete or low information is available.

This thesis implemented a series of ML algorithms to predict GPR values using numeric and high-dimensional predictor inputs. The numeric predictors were used to train the models: random forest (RF), XG-Boost, and a neural network to verify the effects of the dataset composition. Contrastingly, the high-dimensional predictors were dedicated to training convolutional neural network (CNN) models to be incorporated in an RT workflow plan verification. The models' descriptions are explained below.

2.2.1 Random Forest

Random forest (RF) has been used in RT applications as a decision support tool to verify outcome and toxicity predictors in different treatment modalities, especially pneumonia and xerostomia for breast, lung, and head and neck treatments [54,55]. Similarly, RF has also supported patient-specific plan verifications as it will be described in Section **Error! Reference source not found.**

and Section **Error! Reference source not found.**. In general terms, RF is an ensemble-based model that combines the prediction results from several decision trees using a bagging method [56]. To explain more in detail this model, a deifinition of decision tree and bagging is included below.

2.2.1.1 Decision Tree

A decision tree is a model generated through different boolean operations applied to attributes or input features to explain a predicted condition with a series of disjunctive hypothesis. This model generates an flowchart or a tree-ike stucture weightening the entropy of each tested hypothesis by classifying all possible scenarios given by the dataset. Thus, each dcission tree will start with a root node representing the whole dataset that will subsequently be divided in two or more sets or sub-trees. Next a decision node or internal node set the dominant condition founded that dominates the attributes description. Finally, a leaf node is the final output from the specific branch.

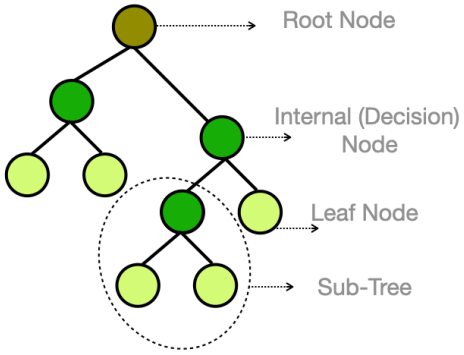


Figure 2.7 Decision Tree representation including the root node, decision/internal nodes, leaf nodes, and sub-tree sections.

The root and internal nodes are organized by measuring each decision node's entropy changes or dataset-description impurity. The purity is the inverse concept of entropy and is related to how balanced is the dataset split in the decision denoted by the node. For a decision node with not a complete segregation of the options given by the dataset, the impurity degree is usually calculated with the Gini index (GI) that is given by $GI = 1 - \sum_j P_j^2$, where P is the probability of a specific condition provided by the number of splits (j) that the decision node generates.

During the decision tree optimization using the training dataset, the decision node splitting and organization are improved by measuring the impurity generated by each sub-tree and the information gained by the model. This “information gained” is given by subtracting all the sub-nodes entropy from the decision node with all attributes' entropy; particularly, the information gained measured the changes in entropy after the segmentation or node splitting of a dataset

based on a specific attribute. Finally, a tree branch stops growing when a node achieves a completely "pure" data splitting or all the attributes have been used.

2.2.1.2 Random Forest

RF is a meta-algorithm that ensembles the learning of several decision trees. This ensembling procedure is performed by a bagging technique known as bootstrap aggregation. In this method, random samples from the training dataset are selected with the possibility of being chosen repeatedly (*i.e.,* sample with replacement) for the different generated decision trees that are subsequently trained independently. Once all models are trained, the average results (for regression) or the majority class predicted (for classification) generally yield a more accurate estimate than a single decision tree prediction. This approach is commonly used to reduce variance within a noisy dataset without increase the bias (Figure 2.8) [57].

After training, a simplified mathematical representation of one prediction of one unseen event (x') by one decision tree f_b trained with bagging B times is displayed in Equation 1.1 (being $b = 1, \ldots, B$). This mathematical average of several trees predictions implies the no sensitivity on its total prediction against the prediction of a single tree.

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x') \tag{1.1}$$

The main hyperparameters in RF models are the number of trees, the maximum depth of trees (*i.e.,* the number of nodes in the tree), and the number of features sampled.

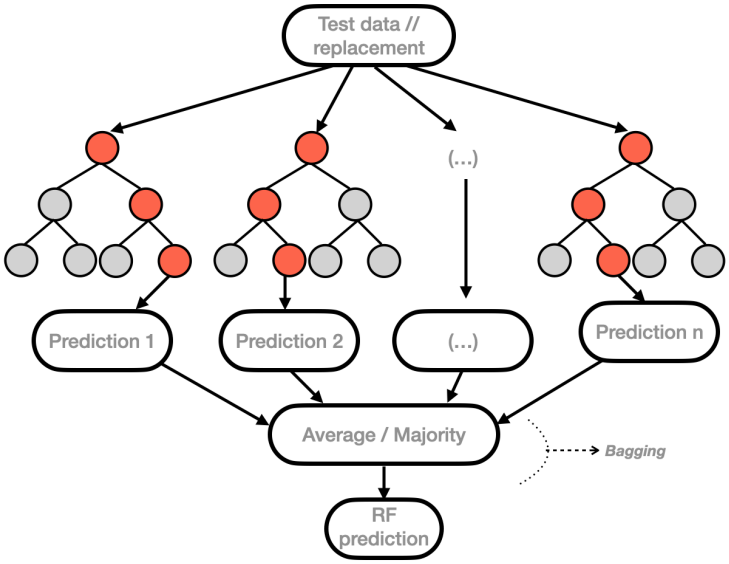


Figure 2.8 Random Forest representation with the bagging of the independent trained decision trees based on different and interchangeable sub-datasets randomly created (replacements).

2.2.2 XG-Boost

XG-Boost, which stands for *extreme gradient boosting*, is a model based on the gradient-boosted decision tree method known as boosting [56]. Unlike the "bagging" ensembling method, which follows a parallel-rationale flow because of the simultaneous and independent training of each tree, boosting might be considered a serial assembling method since the prediction of the initially formed tree influence the prediction performance of the following decision tree until an objective function reaches the minimum (Figure 2.9). Commonly, this model tends to be more efficient by consuming fewer machine resources since the decision tree calculations are not performed at once. Moreover, this boosting model is also considered more effective because it plays a crucial role in dealing with bias-variance trade-offs; unlike bagging algorithms, which only control for high variance in a model, boosting controls both the aspects (bias & variance) and is considered to be more effective.

Gradient boosting intends to minimize a loss function by adding weak learners using a gradient descent optimization algorithm. The loss function is given by the difference between the expected and predicted value, estimating how the model is better in making predictions with the given data, depending on the type of problem (regression or classification). For regression, the loss function is a sum of all squared residuals calculated between the observation and the predicted value by the initial model. In contrast, for classification, is a sum of the log-likelihood functions taking values between 0 and 1. Gradient boost is the first-order gradient of the calculated loss function, also known as the gradient descent algorithm, and is widely used to direct and optimize the next added weak model.

In XG-Boost, the second-order gradients are calculated from the loss function based on the Taylor expansion to implement a mathematical method (Newton–Raphson [58]) to reach the loss function minimum. In addition to this, XG-Boost implements a loss function containing regularisation ('penalty') terms for adding new decision tree leaves to the model ('pruning') with a penalty proportional to the size of the leaf weights, preventing overfitting by avoiding unnecessary longer trees. Similarly, XG-Boost introduces another regularization strategy introducing randomness within the fitting/training process, selecting a random part of the training data. Mathematically, the model prediction $\widehat{y_i}$ can be expressed as Equation 1.2, where K is the number of trees, f_k is a function in the $\mathcal{F}$ functional space with al possible set of classification and regression trees.

$$\widehat{y_i} = \sum_{k=1}^{K} f_k(x_i),\ f_k \in \mathcal{F} \tag{1.2}$$

Consequently, the respective objective function is given by Equation 1.3, where $\omega(f_k)$ is the complexity of te tree (f_k).

$$obj\ (\theta) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{k=1}^{K} \omega(f_k) \tag{1.3}$$

The boosting or additive training can be expressed by the prediction value at a step t as $\hat{y}_i^{(t)}$ given by Equation 1.4.

$$\begin{aligned}
&\hat{y}_i^{(0)} = 0 \\
&\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\
&\hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\
&\ldots \\
&\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)
\end{aligned} \tag{1.4}$$

In summary, the XG-Boost model combines the results of each independent tree from the beginning, adjusting the weights of the dynamic cost function in every iteration (residuals), improving the next tree's performance by fitting new predictors to the previous tree using the residuals to calculate new regularization parameters. The common XG-Boost hyperparameters are the number of trees, the maximum depth of trees, and the learning rate that controls the weighted contribution of each tree in the overall result (Figure 2.9).

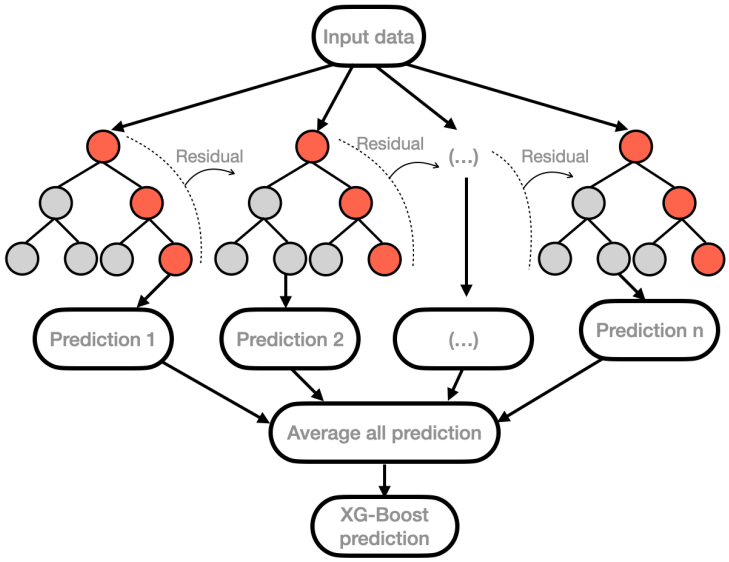


Figure 2.9 XG-Boost representation. The residual of each tree is used to improve the prediction of the next tree, calculating a new regularization parameter to be included in the next aggregated tree.

2.2.3 Deep Learning

Deep Learning (DL) is a set of Machine Learning (ML) models that use layers of mathematical process functions to generate different data representations with multiple levels of features from one data set. Generally, each layer of a DL model has different numbers of transformation nodes named neurons. As it is represented in Figure 2.10, each neuron summarizes the input feature including a weighted factor (*w*) and a bias parameter (*b*), mimicking biological neurons and synapsis processes. The calculated sum of the weighted input features parameters and the bias coefficients of the data is computed by a logistic function (activation function, *f*) that defines if a neuron is activated or not, generating an output (y) [46,50]. This function controls the data flow to the output or to the next layer of neurons trough non-linear functions, depending on the model architecture. The elemental mathematical expression of the predicted output (y) is noted in Equation 1.5, where n is the number of inputs.

$$y = f(xw + b) = f\left(\sum_{j=1}^{n} x_i\, w_i + b\right) \tag{1.5}$$

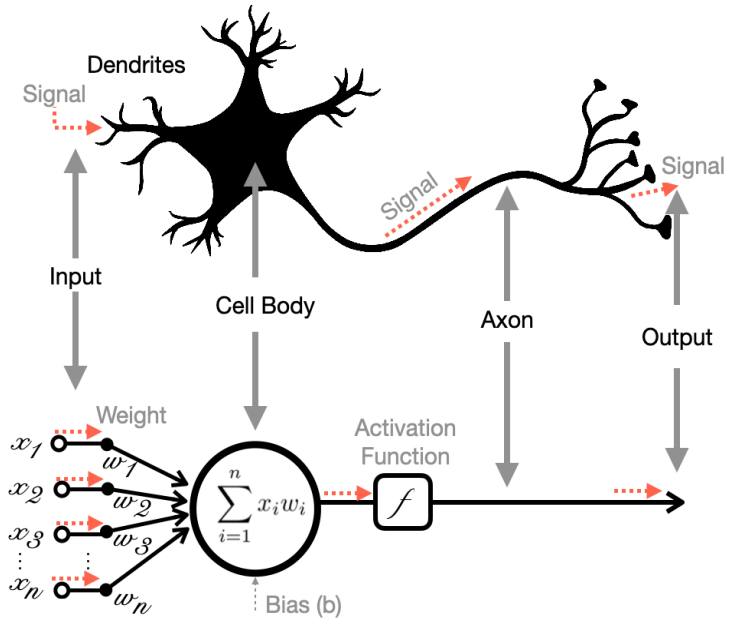


Figure 2.10 Workflow representation of one "artificial" neuron compared to one biological neuron identifying the similarities between dendrites, cell body, and axons, and the input signal nodes, the optimization-activation function, and the output, respectively. Adapted from [46]

2.2.4 Neural Networks

Neural Network (NN) is a DL model based on connections between several neurons [1,46,59,60]. Depending on the designed architecture, each randomly weighted hidden layer inputs comes from the input dataset or another neuron output based on the activation functions to create their respective output (Figure 2.11). This action is repeated over the model's training process to reduce the error function, between the predicted and the expected output [1]. The NN hyperparameters commonly tunned are the number of layers, the number of neurons per layer, and the loss function.

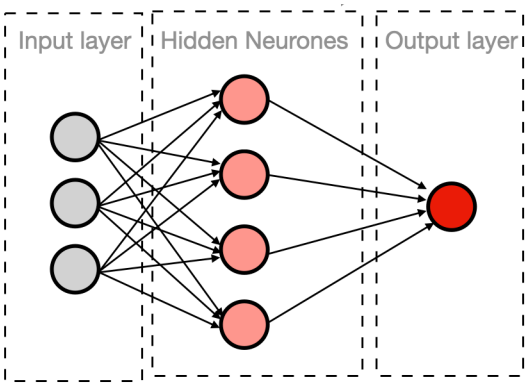


Figure 2.11 Representation of a neural network with three inputs in the input layer, four neurons in the single hidden layer, and one output.

In NN, forward propagation (FP) process is known when a loss function (LF) evaluates the outputs of the model, calculating how different are these values from the expected results (e.g., the error in the prediction), following the normal pathway of one neural network (Figure 2.12). On the contrary, the backward propagation (BP) occurs on the inverse direction, optimizing the calculated LF by a series of derivative functions to minimize the error and find more accurate weighted parameters, starting a "learning process" loop (i.e., training) [60].

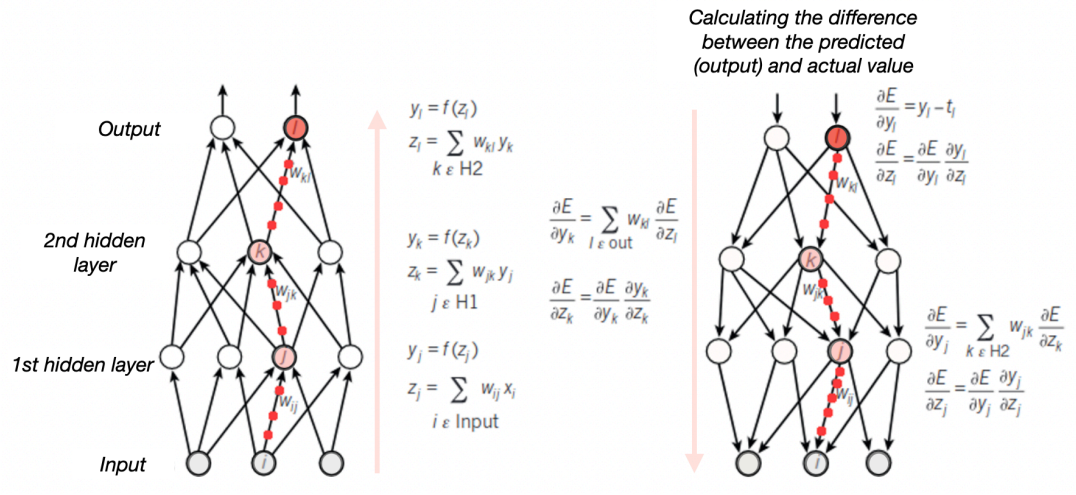


Figure 2.12 Forward (left) and backward (right) propagation scheme. The equations used for computing the forward pass in a NN can be backpropagate gradients. These gradients can be used, through the chain law of derivative equation, to calculate the better weighted parameters to reduce the difference between the predicted and real output, optimizing the process again until the minimum is reached. This way to work lets the model to learn progressively. Adapted from [60].

2.2.4.1 Convolutional Neural Networks

Convolutional Neural Network (CNN) is the DL architecture most popular implemented in video and image analysis. It consists of several successive arranged layers of convolution data processing plus pooling layers, given by a series of filters used to extract features associated with the prediction desired (Figure 2.13) [61]. CNN-based models might be designed for data processing following unsupervised or supervised approach and is widely used in the medical context because it processes information organized originally in multiple dimensional arrays. The 1D arrays are normally signals sequences data, and 2D or 3D arrays are mostly images, video, and volumetric data.

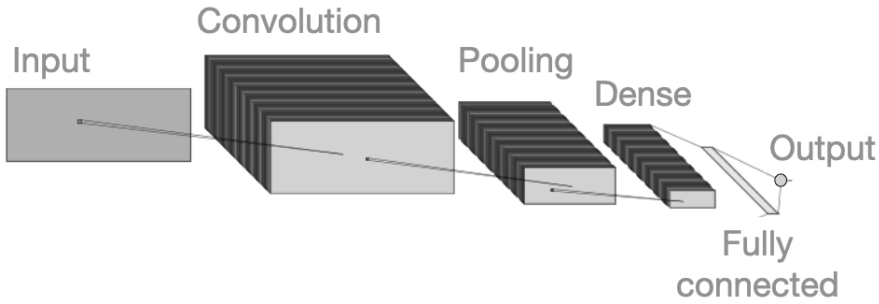


Figure 2.13 Convolutional Neural Network representation with two convolution layers, two pooling layers and one layer that connect and assemble the processed data to take it to the output. Adapted from [50].

2.3 ML applications in RT

Publications in ML dedicated to improving dedicated aspects of radiotherapy workflow have increased dramatically in the past ten years [2,3,46,50,62–65]. Indeed, as shown in Figure 2.14,Meyer et. al. [46] and Sahiner et. al. [50] demonstrated the increasing contributions of DL to contouring, segmentation, and detection of critical structures in medical imaging oriented to RT. Actually, after following and reviewing these major publications references, it was found that more published contributions were dedicated to the organs brain, breast, hearth, and lungs (Figure 2.15). In the case of the brain, the dominant programming architecture is CNN (Figure 2.15.a) and was mostly orientated to GTV segmentation tasks (Figure 2.15.b). For breast, the leading architecture is also CNN, oriented to GTV detection. Whereas, for heart, OAR segmentation was the principal application using autoencoder-based architectures (AE). Finally, for lungs, the dominant architecture was CNN, which was oriented mainly to GTV detection.

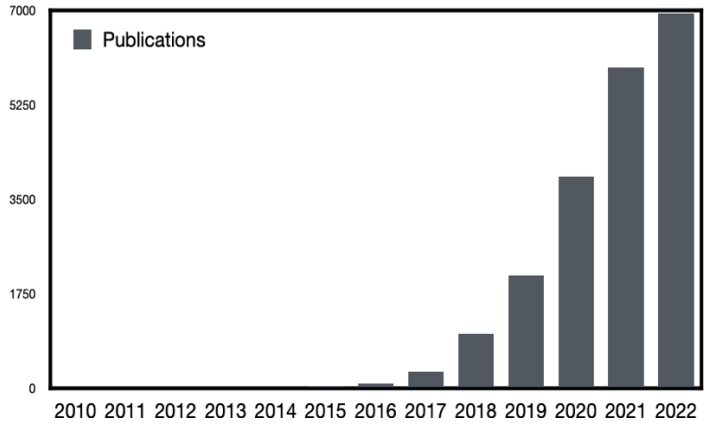


Figure 2.14 Main contributions in deep learning (DL) to imaging and radiation therapy (RT). Updated to 2022 the number of publications by PubMed for the search phrases with the terms: (“deep learning” OR “deep neural net- work” OR deep conv# OR “shift-invariant artificial neural network”) AND (radiography OR x-ray OR mammography OR CT OR MRI OR PET OR ultrasound OR therapy OR radiology OR MR OR mammogram OR SPECT). Adapted from Sahiner et al. [50].

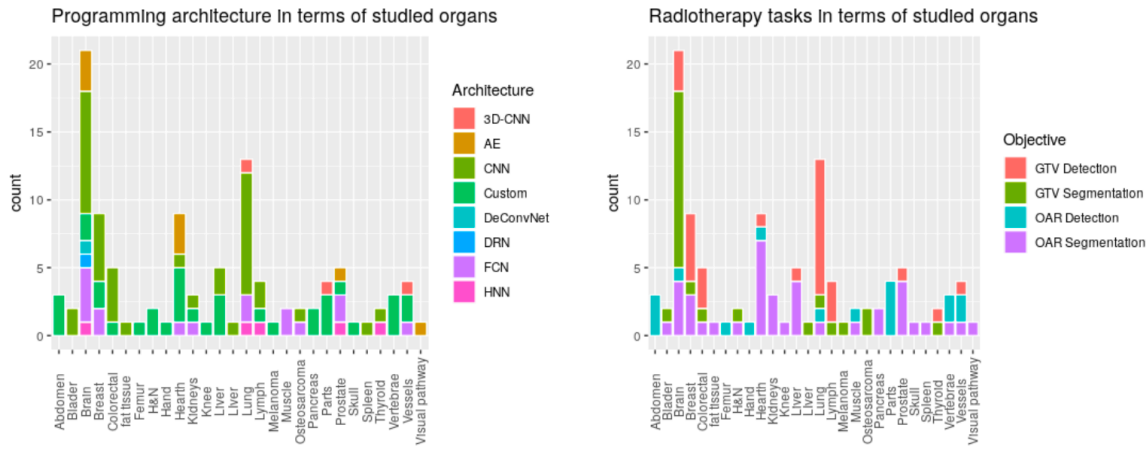


Figure 2.15 Bar-plot representation of the main DL contributions to RT considering the model architecture (Left) and the model objective Right).

Further to imaging processes, ML methods has been applied in radiation toxicity modelling associated to RT treatments using Support Vector Machine (SVM), K-Means Clustering (k-means), and Linear Regression (LR) models to predict pneumonitis [66], esophagitis-pneumonitis-xerostomia [67], and tumor control probability (TCP) or normal tissue complication probability (NTCP) [68], respectively. Additionally, other ML models has been implemented in lung cancer prognosis [69], genitourinary toxicity prediction [70], decision-tool for melanoma indications [71], local control prediction of non-small cell lung cancer (NSCLC) [72], prostate RT side effects modelling [73], and adaptative RT applications [42].

2.3.1 ML methods applied in QA evaluation or dose deliverability

In recent 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 [2–4,74,75]. The most common ML models implemented in this matter are Poisson regression [76], decision trees-based models (e.g., random forest or gradient boosting models) [77], support vector machine (SVM) [63], and artificial neural networks (ANN) or convolutional neural networks (CNN) [78–81]. These 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 [82]. 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[60,83].

From the specific-plan verification perspective, models dedicated to QA prediction were implemented generally to detect potential treatment errors [84–87] and predict gamma passing rate (GPR) values [3,63,78,88,89]. The GPRs account for the dosimetric regions in agreement with the gamma index analysis between the calculated and the measured dose distributions [4,5,29]. 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) [29,89]. 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 [4,75]. 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 [5,6,90]. 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. [91,92], McNivell et al. [93], and Chiavassa et al. [94] 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 [77,79,81], and predictor features associated with the plan parameters cannot be extracted. In contrast, other input features, such as modulation maps (MM) given by the 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 [78], planar dose images plus organs at risk volumes and total MU values [81], radiomic features from the dose distribution images [77], and various calculated modulation complexity metrics [63,80]. 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) [77,81]. However, considering that complexity metrics and features related to MLC movements are the most relevant features for GPR predictions [74,93,95–97], it is necessary to contemplate complementary features, such as 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 [74,93,97,98] that might limit the amount of information extracted.

Chapter 3 Materials and Methods

In RT, the dose deliverability evaluation of specific-treatment plans includes mechanical and dosimetric tests to verify potential differences between one calculated treatment plan and its corresponding dose distribution delivered by the linac [4]. Thus, dosimetric protocols to measure the delivered absorbed dose and evaluate the spatial distribution differences between the calculated and the measured plans are usually implemented in each RT facility [6,99].

As it is mentioned and explained in Section 2.1.4 and Section 3.2.2, the more accepted and implemented dose verification test is the gamma index evaluation, where one treatment plan can be considered suitable for clinical delivery if the GPR is higher than one defined value (e.g., GPR $\geq$ 98%) under one specific DD/DTA criteria (e.g., 3%/2 mm, considering each institution protocols). Accordingly, predicting GPR values based on ML applications is a straightforward step to support virtual specific-treatment verification protocols, mainly focused on reducing the unnecessary irradiation time of treatment plans with a high probability of presenting unaccepted GPR values. However, the reported GPR prediction models based on dose distributions or calculated modulation complexity metrics are not focused on recognizing concrete physical treatment parameters implied in the GPR value prediction, limiting the models' reliability.

This chapter describes the general research methodology following a workflow map (Figure 3.1) that encloses all the technical aspects needed to address the formulated central thesis hypothesis and research questions (Chapter 1). Additionally, the implemented materials, including the datasets, treatment units, detector specifications, and ML models, are also specified. However, specific materials or additional information needed for each chapter are included in their respective introduction sections.

3.1 Methods

The workflow of this research, represented in Figure 3.1, was designed to verify if *"It is possible to use ML models to aid virtual specific-treatment verifications in prostate radiotherapy, retrieving critical physical aspects involved in dose deliverability."* Hence, the main four tasks or steps followed were: (I) the features extraction, (II) the dataset assembling, (III) the GPR modelling, and (IV) the retrieving of the specific predictors involved in the GPR predictions.

Due to the widely reported correlation between modulation complexity and the GPRs [91,93,94,96], the *feature extraction* section was dedicated to calculating reported complexity metrics, retrieving dosimetric quality parameters, extracting new high-dimensional (2D)

features related to modulation complexity, and proposing new modulation complexity metrics tailored to the treatment unit with dual-layer MLC (Halcyon-v2, Section 3.2.1) since a major part of the treatments used were designed and delivered in this linac. Then, with these metrics, a study was developed to find the optimal *dataset assembling* conditions to improve the GPR prediction performance using the RF, XG-Boost, and NN algorithms. Next, considering the optimal dataset conditions and the low information provided by the numeric predictors, the *GPR modelling* was implemented using high-dimensional data and CNN-based models. Finally, the *models' activated features were retrieved* to assist a decision-support protocol and evaluate the dose deliverability of individual RT treatments. Besides, a model verification with an external dataset was performed to evaluate the model generalization.

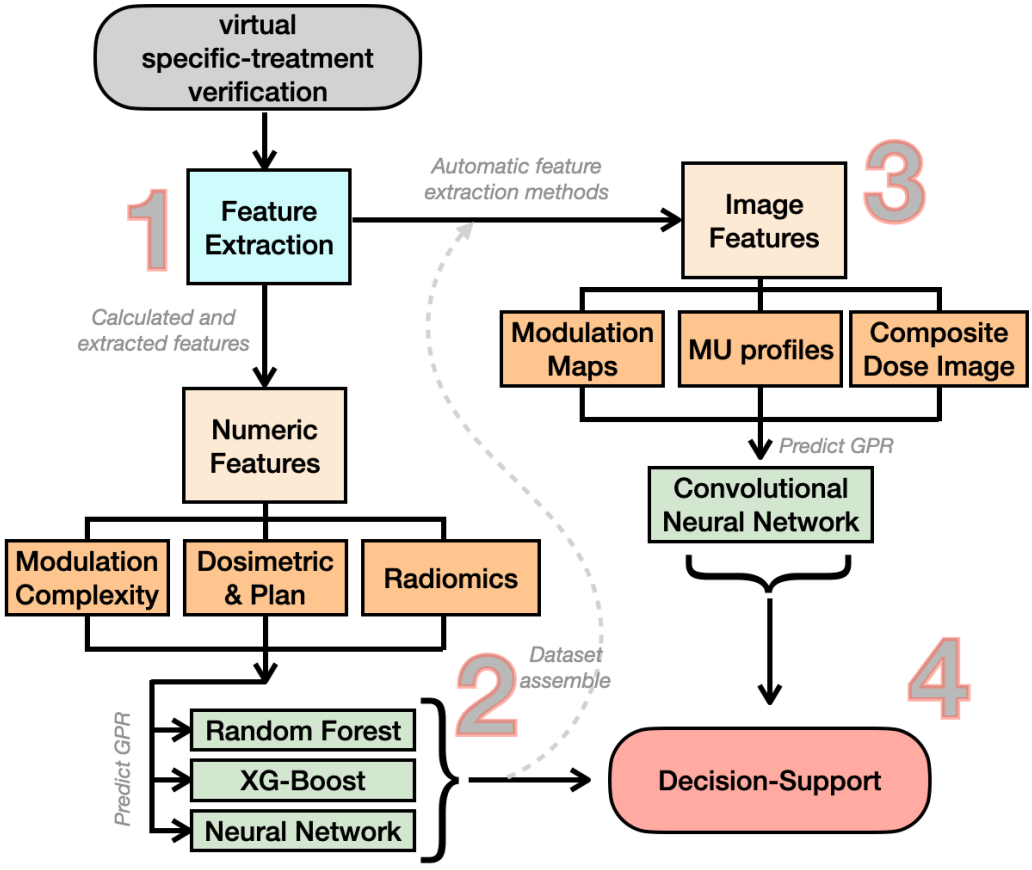


Figure 3.1 Workflow map of the research pathway following four main research steps: (1) Extracting, calculating, and creating the input features based on modulation complexities. (2) Analysing the best dataset configuration. (3) Modelling GPRs using automatic extracted features. (4) Implementing decision-support tools to evaluate data deliverability using the previous information generated.

3.1.1 Features Extraction

The input features (Section 4.1, Section 4.3, and Section 4.4) were extracted from anonymized treatment plans optimized and calculated using the Eclipse v15.6 TPS (Section 2.1.5). The dosimetric and quality plan parameters (Section 4.2.1.4) were retrieved using the Eclipse scripting application programming interface (ESAPI), which is an interface based on the C# programming language, and it is incorporated in Eclipse to ease the development of automatized tools needed to retrieve, organize, and systematically collect treatment information managed in the TPS [100]. In contrast, the information needed to calculate the modulation complexity metrics and the high dimensional features were extracted from the anonymized files saved in the Digital Imaging and Communications in Medicine (DICOM) protocol in RT (DICOM-RT) format, implementing Python scripting [101,102]. Moreover, the DICOM-RT plan, DICOM-RT dose, and DICOM-RT image files contain specific hardware parameters (such as the MLC trajectories and irradiation beam geometry during the treatment), the dose distribution, and the composite dose images for dose verification, respectively.

Since the reported modulation complexity scores [94] were originally developed and created for conventional treatment units with single-layer MLC models. Two new complexity metrics (plus one adapted metric) were proposed and evaluated for Halcyon-v.2 to incorporate additional complexity features related to dose deliverability (Section 4.1.5, Section4.1.6).

3.1.2 Dataset Assembling

The dataset's composition effects on the model performance were studied to evaluate how specific heterogeneities within the datasets might compromise the model's reliability. The motivation of this study was based on the several reported ML models predicting GPR values that were commonly trained using unbalanced datasets with heterogeneous treatment conditions, such as different anatomic regions, dose per fraction, number of beams, different treatment units, or different beam energy, among others [2,3]. Indeed, these treatment conditions have been previously associated with variations in GPR values [76,89,92,103,104], suggesting that special care might be necessary for GPR modelling, considering the representation of each treatment condition within the dataset to ease the model data generalization.

Consequently, the prediction performance of models (RF, XG-Boost, and NN) trained with various datasets assembled with controlled variations of treatment plans (having different treatment conditions) was evaluated. Additionally, the main features of each dataset were retrieved to verify if the predictors correspond to actual physical aspects of the treatment that might help the dose deliverability evaluation.

3.1.3 GPR Modelling

Considering the dataset assembling evaluation and the more convenient dataset configuration, the GPR values were predicted using the numeric extracted and calculated features using RF, XG-Boost, and NN models (Section 2.2), whereas CNN-based models were implemented with the MU per control point MUcp profiles, the modulation maps (MM) images generated with the MLC movements, and the composite dose images (CDI) generated to perform portal dosimetry. The ML models design and implementation were developed with Python with scikit-learn, Keras, and TensorFlow as main ML libraries [105].

3.1.4 Decision Support Workflow

The decision-support protocol was based on the retrieved predictor features and the features extracted by the models as potential technical plan parameters involved in the GPR prediction. A reliable evaluation method should include different physical properties linked to the dose deliverability prediction to formulate and establish reference hardware or dosimetric acceptance criteria, verifying if the predicted GPR value reflects realistic scenarios that might help to adjust or design new treatment plans with more accurate dose deliverability. For this reason, the proposed protocol is more oriented toward understanding the prediction causes rather than just a predicted value or a classification.

3.2 Materials

3.2.1 Treatment Unit

The linac models implemented in this research are the TrueBeam and Halcyon-v2 (Varian Medical Systems - Palo Alto, USA) with c-arm and ring-gantry architecture, respectively (Figure 3.2). The treatment unit were three Varian linear accelerators calibrated at the same reference conditions, two dosimetrically matched TrueBeams (TB) and one Halcyon-v2 (HL), with the same nominal resolution at isocentre (5 mm) and same electronic portal imaging device (EPID) model (Section 2.3.3). Both TBs have a Millennium 120 multi-leaf collimator (MLC) with a maximum leaf speed of 25 mm/s, 6 MV flattened filter (FF) photon beam, and dose rate at isocentre of 800 Gy/ min. In contrast, the HL has dual-layer MLC with a maximum leaf speed of 50 mm/s, 6 MV flattening filter-free (FFF) photon beam, and 740Gy/min dose rate at isocentre (Figure 3.2).

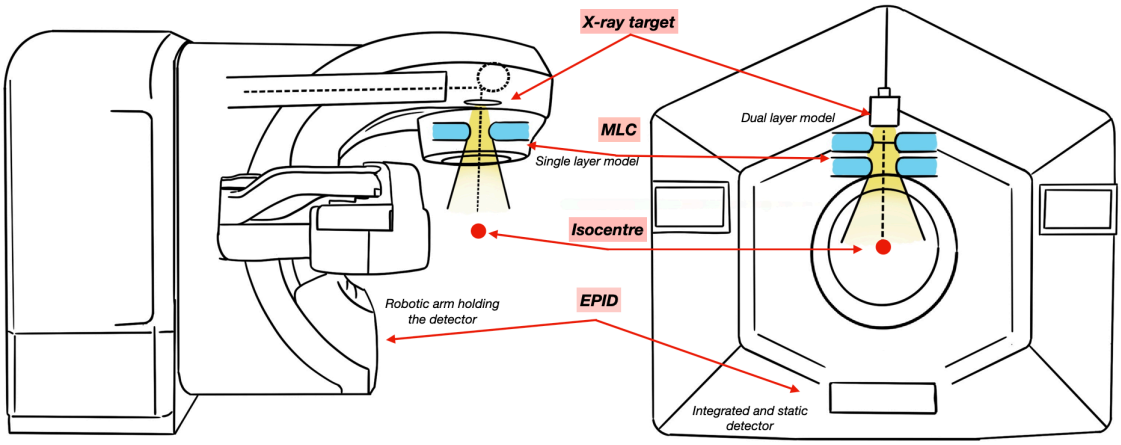


Figure 3.2 Comparison between the treatment units TrueBeam and Halcyon-v2. The plot shows the equivalent aspects highlighting the beam target, where is located the multi-leaf collimator, the isocentre, and the EPID device.

3.2.2 Gamma Index

In this research, the GPR values considering the 3%/3 mm, 3%/2 mm, 2%/3 mm, 2%/2 mm, and 2%/1 mm criteria were retrieved from the measurements related to the clinical plans. These GPR values were used to identify a GPR threshold (percentage of evaluated points passing the gamma index < 1) that grants balanced distributions within the dataset (i.e., ideally, 50% pass and 50% fail). Simultaneously, the chosen GPR criteria and threshold had to be suitable for detecting potential clinical errors [106], avoiding unbalanced datasets [81,104,107], and excluding this potential bias within the model performance. The balanced distributions on the dataset were necessary to ensure similar representation of both conditions and avoid overfitting during the model training.

3.2.3 Electronic Portal Dosimetry Device

The dose measurements needed for the gamma index evaluation and obtaining the specific-treatment GPR value were performed using the same EPID model (aS1200) attached to each treatment unit and calibrated in the same reference conditions. This EPID is a silicon-based detector with a resolution of 1280x1280 pixels, 0.34 mm/pixel at the panel and 0.22 mm/pixel at the isoplane, and a panel size of 43 cm x 43 cm [108]. The image is created because one incident photon interacts against the electron's shells from a copper plate which releases an electron to a scintillator, generating one light photon that in turn generates a hole-electron pair within an amorphous silicon layer, creating and storing the charge on the intrinsic photodiodes' capacitors (Figure 3.3). Consequently, an image pixel corresponds to a single light-sensitive photodiode plus one thin film transistor configuration, generating a discrete electric signal based on the charge held by the photodiode switched by the transistor.

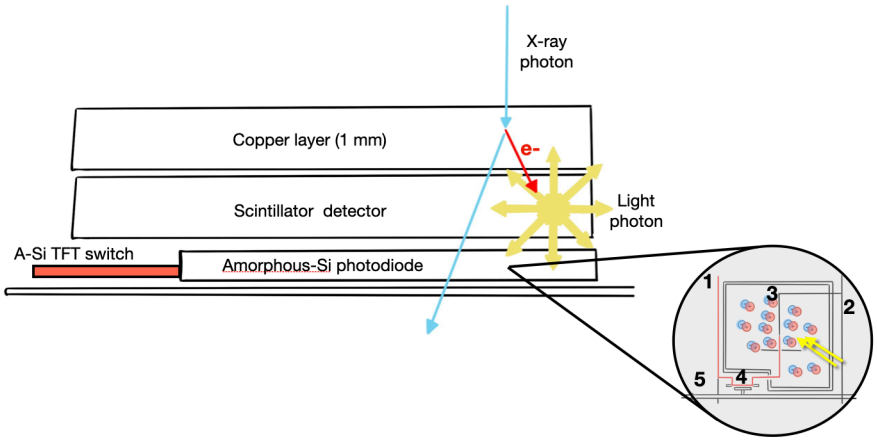


Figure 3.3 Dose detection process for electronic portal imaging devices based on amorphous silicon detectors. For the pixel image: (1) Data Line, (2) Bias Line, (3) Photodiode, (4) Thin Film Transistor (TFT), and (5) Gate Line.

3.2.4 ML Model Evaluation

Different evaluation metrics might be implemented to evaluate the prediction performance depending on the ML model task (classification or regression). The metrics used over all this study are summarised in Table 3.1

Table 3.1 Evaluation metrics implemented in this study

Model Prediction	Metric	Description
Regression	MAE	$\mathrm{MAE} = \sum \lvert(y_i - y_p)\rvert / n$
	RMSE	$\mathrm{RMSE} = \sqrt{\sum (y_i - y_p)^2 \,/\, n}$
	r (Spearman's correlation coefficient)	High, moderate, and lower correlations were defined for $r<0.4$, $0.4 \le r \le 0.7$, and $r>0.7$ values, respectively
Classification	Accuracy	$A = (TP + TN)/(TP + TN + FP + FN)$
	Specificity (Sp)	$Sp = TN \,/\, (TN + FP)$
	Sensitivity (Se)	$Se = TP \,/\, (TP + FN)$
	ROC_AUC	Area under the receiver operating characteristic curve

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, or recall.

Chapter 4 Modulation Complexity and Feature Extraction

Beam modulation is a principal feature in advanced RT techniques using static field Intensity Modulation or Volumetric Modulated Arc-Therapy (VMAT). Due to the synchronised motion of the MLC leaves, the radiation dose can be conformed to complex planning target volume (PTV) shapes, increasing the treatment effectiveness and keeping the adverse effects as low as possible by avoiding organs at risk (OARs) [109,110]. Nevertheless, high modulation levels or complex conformity scenarios might demand challenging or unrealistic treatment unit performances, compromising the dose deliverability. Consequently, the modulation complexity has been studied widely on linacs with single-layer MLC architecture to predict dose deliverability, using metrics such as modulation index (MI) [96], modulation complexity score (MCS) [93], texture methods [111], dimensional fractal analysis [95], and aperture-based methods [112]. These complexity analyses have proven to help compare linac performances between treatment techniques [20], evaluate the best plan parameters in specific planning scenarios [74,112], establish reference values for dosimetry audits [113], and predict delivery accuracy in terms of GPRs [76,114].

GPR predictions applying ML methods [77,81,115,116] have been explored using modulation complexity metrics as the principal predictor variables due to their reported relationships with dose deliverability [76,94,114,117]. However, two main gaps related to the modulation complexity calculations and the need for new modulation features were addressed in this chapter. (1) The lack of information about complexity metrics dedicated to dual-layer MLC linacs and (2) the nonexistence of high dimensional features that might reflect the modulation complexity more comprehensively.

Considering the above, this chapter is dedicated to featuring the main modulation complexity metrics implemented in this thesis and demonstrating the suitability of the new proposed dual-layer MLC complexity metrics. Additionally, the consideration and extraction methods of the high dimensional modulation complexity features are included.

4.1 Modulation Complexity metrics

The modulation complexity metrics have been comprehensively reported by Park et al. [92] and Antoine et al. [28] covering various modulation indices dedicated to predicting the plan-delivery accuracy in VMAT treatments. Similarly, the literature review by Chiavassa et al. [94] includes all

current complexity indices and the relevance of each metric. However, all these metrics based on MLC movements (i.e., beam fluence modulation), were explored for treatment units having conventional single-layer MLC. Contrastingly, Tamura et al. [97] propose a modulation metric dedicated to dual-layer MLC architecture (for Halcyon-v2), considering a weighted method taking the distal and proximal layer contributions in the field conformation, adapting the metric originally suggested by McNiven et al. in 2010 [93]. Nevertheless, this metric does not consider technical aspects implied in the stacked MLC arrangement, and a new way to measure modulation complexity in this treatment unit model are needed. For this reason, two new modulation complexity metrics (plus one adapted metric) were proposed in this section to improve the further ML modelling performances implemented in this research.

The complexity metrics implemented in this studio were calculated using Python scripting [98], processing the information from DICOM-RT plan files [118,119], and reading the leaves positions per control point (CP) using the Pydicom library [119]. The calculated or extracted complexity metrics were the number of MU (**MU**) [97], the average MU increment by CP (**MUcp**) [97], the modulation complexity score (MCS) for VMAT treatments (**MCSv**) [74], and the weighted MCSv for DL-MLC architecture (**MCSw**) [97]. Additionally, two new complexity metrics and one adapted metric were proposed: the uncovered-layer score (**UL)**, the number of peaks score (**NP**), and the MCSw weighted by UL (**MCSUL)**, respectively.

4.1.1 MU

The total monitor units (MU) planned to be delivered in the treatment corresponds to the dose delivered by the linac in terms of output energy. Depending on the linac calibration, 1 MU represents 1 cGy at a reference depth in a reference phantom at a specific distance from the linac source having a reference field size [75].

4.1.2 MUcp

The averaged monitor unit increment by control point (MUcp) is the factor that accounts for all the MU variations between two adjacent CPs (j, j+1) for the total MU of the plan (4-1).

$$MU_{cp} = \frac{MU_{j,j+1}}{MU} \tag{4-1}$$

4.1.3 MCSv

The modulation complexity score is defined in 4-2, where j is the control point number, $AAVj$ is the Aperture area variability (4-3), $Max - pos_a$ is the maximum aperture of each leaf bank, A is

the number of leaves in the arc, LSV is the leaf sequence variability (4-4), and N is the number of moving leaves [74].

$$MCS_v = \sum_j \left[\frac{AAV_j + AAV_{j+1}}{2} \times \frac{LSV_j + LSV_{j+1}}{2} \times \frac{MU_{j,j+1}}{MU} \right] \tag{4-2}$$

$$AAV_j = \frac{\sum_{a=1}^{A} \left((pos_a)_{leftbank} - (pos_a)_{rightbank} \right)}{\sum_{a=1}^{A} \left((Max - pos_a)_{leftban\,\epsilon\,arc} - (Max - pos_a)_{rightbank\,\epsilon\,arc} \right)} \tag{4-3}$$

$$LSV_j = \left[\frac{\sum_{n=1}^{N-1} pos_{max} - |pos_n - pos_{n+1}|}{(N-1) \times pos_{max}} \right]_{leftbank} \times \left[\frac{\sum_{n=1}^{N-1} pos_{max} - |pos_n - pos_{n+1}|}{(N-1) \times pos_{max}} \right]_{rightbank} \tag{4-4}$$

4.1.4 MCSw

The weighted MCS for a dual-layer MLC configuration is defined in

(4-5 adapted from MCSv (4-4). The $pMCSw$ is the proximal weighted modulation complexity score ((4-6), and the $dMCSw$ is the distal weighted modulation complexity score ((4-7). The w_p and w_d are the proximal and distal weighted factors for each MLC layer, respectively, and are estimated from the contributions of each layer in the field conformation ($w_p + w_d = 1$) [97].

$$MCS_w = pMCS_w + dMCS_w \tag{4-5}$$

$$pMCS_w = \sum_j \left[\frac{AAV_j + AAV_{j+1}}{2} \times \frac{LSV_j + LSV_{j+1}}{2} \times \frac{MU_{j,j+1}}{MU} \times w_{pj,j+1} \right] \tag{4-6}$$

$$dMCS_w = \sum_j \left[\frac{AAV_j + AAV_{j+1}}{2} \times \frac{LSV_j + LSV_{j+1}}{2} \times \frac{MU_{j,j+1}}{MU} \times w_{dj,j+1} \right] \tag{4-7}$$

4.1.5 UL

The uncovered layer UL considers all leaf-pairs uncovered by their respective leaves from the complementary MLC layer (above or below). Figure 4.1 shows an example from an MLC sequence where the proximal layer leaves do not cover a distal leaf-pair section, creating an uncovered region that might increase the interleave-dose transmission, and thus, it might be related with dose measurement discrepancies due to the reported incomplete attenuation of the beam [19]. This metric is calculated by summing the number of uncovered gaps per CP considering both, the distal and proximal layers. This sum is weighted by the relative fraction of MU in that CP ((4-8).

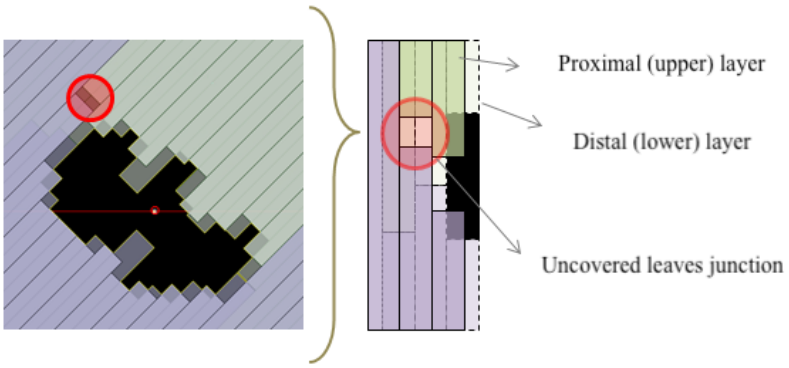


Figure 4.1 Uncovered leaves junction of the distal multi-leaf collimator (MLC) layer by the proximal MLC layer. This figure shows a conformed field in one control point (CP) from a specific treatment. The green and purple leaves differentiate the MLC banks (right and left).

For the UC defined in (4-8, j is the CP number, ${}_{p}ul_j$ is the number of uncovered spots in the proximal layer at j, ${}_{d}ul_j$ is the number of uncovered spots in the distal layer at j, MUj is the fraction of MU at j, and MU is the total number of monitor units.

$$UL = \sum_{j} \left[\left({}_{p}ul_j + {}_{d}ul_j \right) \times (MU_j / MU) \right] \tag{4-8}$$

4.1.6 NP

The number of peaks NP, accounts for the modulation complexity of both MLC models, calculating the average number of peaks presented in the trajectory profiles of all moving leaves in a VMAT treatment. As is shown in Figure 4.2, the position at each CP of a single leaf can be visualized within a trajectory profile, where the peaks represent significant changes in leaf speed and position. These variations can be associated with demanding hardware conditions that may generate dose delivery inaccuracies [13,120].

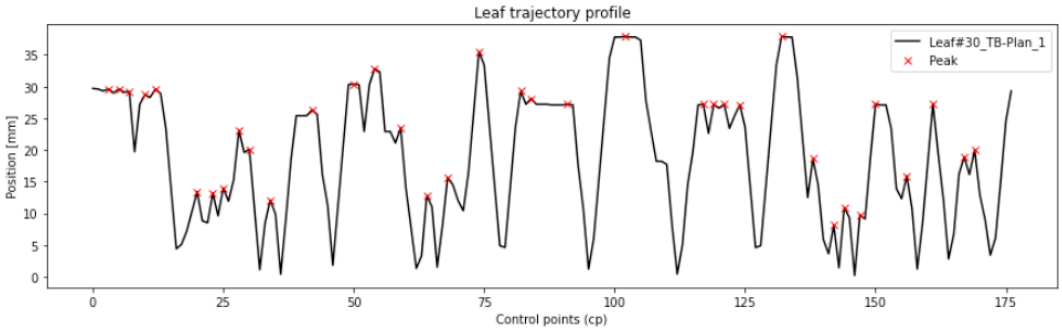


Figure 4.2 Trajectory profile of the 30th leaf of TrueBeam (TB) from a prostate treatment plan labelled TB-plan_1. The red marks indicate the number of detected peaks using the function *find_peaks* from SciPy (41).

The NP is defined in Equation 4-9, where np_i is calculated using the Python function *scipy.signal.find_peaks* [38] with the default parameters values, and N is the number of CPs.

$$NP = \frac{\sum_i^N np_i}{N}$$

Equation 4-9

4.1.7 MCS_{UL}

The metric, MCSUL, is an adapted version of MCSw, including UL as an additional factor to be considered in the complexity score of each MLC layer. Its calculation is described in Equation 4-10, Equation 4-11, and Equation 4-12, where $pMCS_{ul}$ is the modulation complexity of proximal layer including the average number of uncover spots in the proximal (same to distal layer) between two CPs ($pul_{j,j+1}$). AAV and LSV are calculated to its respective layer considering previous metrics' calculations.

$$MCS_{UL} =_p MCS_{UL} +_d MCS_{UL}$$

Equation 4-10

$${}_pMCS_{UL} = \sum_j \left[\frac{AAV_j + AAV_{j+1}}{2} \times \frac{LSV_j + LSV_{j+1}}{2} \times \frac{MU_{j,j+1}}{MU} \times w_{pj,j+1} \times_p ul_{j,j+1} \right]$$

Equation 4-11

$${}_dMCS_{UL} = \sum_j \left[\frac{AAV_j + AAV_{j+1}}{2} \times \frac{LSV_j + LSV_{j+1}}{2} \times \frac{MU_{j,j+1}}{MU} \times w_{dj,j+1} \times_d ul_{j,j+1} \right]$$

Equation 4-12

4.2 Validation of Complexity Metrics

This section describes the study dedicated to verifying the correlation between the reported complexity metrics with GPR values and validating the proposed novel metrics for dual-layer MLC linac models. Therefore, 192 VMAT plans were calculated using one virtual prostate phantom (avoiding volume effects) considering three main settings:

(1) Three TPS-parameters (Convergence; Aperture Shape Controller, ASC; and Dose Calculation Resolution, DCR) selected from Eclipse v15.6 to modify systematically the TPS conditions within the inverse optimization to verify if these parameters are related to dose deliverability.
(2) Four levels of dose-sparing priority for organs at risk (OAR), to emulate four levels of hardware demanding conditions.
(3) Two treatment units with same nominal conformity resolution and different MLC architectures (Halcyon-v2 dual-layer MLC and TrueBeam single-layer MLC).
(4) Seven complexity metrics to evaluate the modulation complexity, including two new metrics and one adapted metric for dual-layer MLC, assessed by their correlation with gamma passing rate (GPR) analysis.

4.2.1 Methods

4.2.1.1 Treatment plan configuration

Ninety-six VMAT plans were generated in one institution[1] with Eclipse 15.6 using a single prostate patient dataset as a virtual phantom to deliver 2Gy per fraction in one full arc. The linac configuration was Halcyon-v2 with dual-layer MLC, maximum leaf speed of 50 mm/s, 6 MV flattening filter-free (FFF) photon beam, and a dose rate of 740 Gy/min. The same plans were replicated using the TrueBeam linac configuration with single-layer MLC Millennium-120, maximum leaf speed of 25 mm/s, 6 MV FFF photon beam, a dose rate of 800 Gy/min, with jaw tracking mode turned off. For both cases, the plans were calculated with the anisotropic analytical algorithm (AAA) and were optimised with the PO algorithm, applying automatic mode for normal tissue objective and a structure optimisation resolution of 2.5 mm. Both treatment units were calibrated at the same reference conditions [121].

4.2.1.2 TPS parameters

[1] Queen's Centre for Oncology and Haematology - Castle Hill Hospital. Hull University Teaching Hospitals NHS Trust.

The three studied parameters from Eclipse TPS features were Convergence (Conv), Aperture Shape Controller (ASC), and Dose Calculation Resolution (DCR), and their respective modes were Conv{off; on; extended}, ASC{off; low; moderate; very_high}, and DCR{normal; high}.

1. The Conv parameter controls the internal schedule of the transitions between and within the different multi-resolution (MR) levels of the PO. These changes in the transition times expect improved optimization results in dose fluence because the number of iterations increase, when modes= on/extended (respect the mode= off) by a factor of 2.5/11.2 on MR-1, 2.0/17.8 on MR-2, 1.0/17 on MR-3, and 1.0/15 on MR-4 for modes On/Extended respectively. However, the MU values may increase, and the optimisation time rises 1.2 - 3-fold for On mode, and a few hours for Extended mode [31].

2. The ASC parameter is a tool of the leaf-motion sequencer of the PO that penalises the leaf position deviations with respect to the adjacent leaves in the same continuous target projection. This penalty is introduced in the optimisation process, and its magnitude depends on the selected mode (Off, Very_low, Low, Moderate, high, and Very_high). Controlling the size and shape of the field with ASC may help to reduce the MU, the dose delivery inaccuracies, and the control quality failures [31]. For single-layer MLC architecture, Binny et al [122] found that ASC may be useful to improve the distribution of MU per degree throughout the treatment time, but it requires to evaluate its potential impact on treatment time. In this study, the modes explored were off, low, moderate, and very_high, evaluating the impact of the parameter and differences in the obtained results between extremes (off and very_high) and small changes (low and moderate).

3. The DCR is a dose optimisation parameter related to the grid resolution of the internal dose calculation engine of PO [31]. The modes High (1.25 mm) and Normal (2.50 mm) of DCR change the internal grid size within each MR dose calculation, influencing the pre-calculated dose resolution, which impacts directly in the leaf sequencing, the dose rate, the MU/deg, and thus, the final dose distribution within the optimisation process.

4.2.1.3 Dose sparing Priority

To simplify the planning process, the OARs (OAR1: rectum, OAR2: bladder) were considered as independent structures to be avoided with no clinical differentiation between them. The avoidance was controlled by reducing their mean dose using the optimization objective upper_gEUD (from generalized Equivalent Uniform Dose)[123]. This optimization tool tries to reduce the volume that receives mid-dose levels (mean dose) using the parameter '*a*' that is set

as 1 for parallel organs (following the rationale of Lyman-Kutcher-Burman NTCP model) [124–126]. This parameter *α* can take values up to 40 for serial organs minimizing the maximum dose contributions to the OAR. In this experiment, it was assumed both OARs as parallel organs using *α*=1.

Parallel organs are the functional human structures that can remain functional, even if a specific part has been affected by considerable damage (i.e., high absorbed radiation dose). Some examples are the liver or lungs. In contrast, serial organs are functional structures that cannot remain functional when part of it is affected by considerable damage. Some examples are the spinal cord or the optic nerve.

To counter the dependence of the same-patient dataset [127] and to consider possible effects of the TPS parameters over various dose-sparing scenarios, four levels of dose-sparing priorities for OARs were implemented within the optimization process. Priority values of 20, 40, 60, and 80 were selected to be applied with the upper_gEUD parameter, representing lower, moderate, high and very-high dose sparing conditions respectively. Contrastingly, a priority value of 100 was used with the dose coverage (100% of the prescription dose) and maximum dose (105% of prescription dose) parameters for the PTV, and for the maximum dose constraint for the whole-body structure. In summary, 96 plans were produced, covering all permutations of the three TPS-parameters mode settings (four for ASC, three for Conv, and two for DCR), and four optimisation priority settings for the OAR mean dose constraint.

4.2.1.4 Plan quality indices

The metrics used to evaluate the plan quality were based on the recommendations of the International Commission on Radiation Units & Measurements (ICRU) Report 83 [40]. It was selected: the conformity index (CI), defined as the ratio between the volume that enclose the prescription dose (Vp) and the volume of PTV (VPTV), {CI= Vp / VPTV}; and the homogeneity index (HI), defined as the ratio between the dose difference that covers 98% and 2% of the volume (D98% and D2% respectively) and the prescription dose (Dp), {HI= (D2%-D98%) / Dp)}. Additionally, it was recorded the mean dose of the PTV (mD-PTV), the volume enclosed by the 50% isodose (V50%) as a dose spillage metric, and the mean dose of OAR1 and OAR2 (mD-OAR*n*).

4.2.1.5 Complexity metrics

The complexity metrics used in this study are summarised and explained in Section 4.1.

4.2.1.6 Complexity metrics validation

The new complexity metrics were introduced in this study to investigate the deliverability and quality of the plans produced with DL-MLC. To assess the value of these, they were compared with the gamma passing rate (GPR) calculated using gamma analysis [29].

To analyse the correlation of the new complexity metrics with GPR, the prostate plans for Halcyon-v2 were measured with the integrated EPID (Section 3.2.3). Furthermore, the accuracy of dose delivery was evaluated with gamma analysis (γ) [29] using various levels for global dose difference (DD) of prescribed dose and distance to agreement (DTA) for at least 98% of all pixels. The DD/DTA criteria were 3%/3 mm, 3%/2 mm, 2%/3 mm, 2%/2 mm, and 2%/1 mm. The images were processed using the portal dosimetry tool available in Eclipse 15.6, with the absolute absorbed dose correction and the improved gamma evaluation mode utilised.

In contrast with the Halcyon-v2, portal dosimetry on the TrueBeam with 6 MV-FFF mode is not possible in the treatment institution due to the detector saturation and lack of an image prediction algorithm, depending on linac model used. For this reason and based on the reported correlation between the MU values and the dose deliverability (18-29), the MU was selected as a reference index to compare the performance of each calculated complexity metric.

4.2.1.7 Statistical Analysis

The statistical significance of the correlations between the TPS-parameters, the complexity metrics and the gamma analysis were evaluated using Spearman's rank correlation coefficient (r) with a threshold of $p<0.05$ [92]. The low, moderate and high correlations were considered for values of $|r|<0.4$, $0.4\leq|r|\leq0.7$, and $|r|>0.7$ respectively [97,128]. The correlation between the modes of each TPS-parameter were tested for significance ($p<0.05$) using Wilcoxon signed-rank test.

4.2.2 Results

After calculating the VMAT plans as described earlier, three main aspects were assessed for this study. First, as a general overview, the modulation complexity metrics and plan quality indices calculated for both linacs were compared. Second, the impact of each TPS-parameter mode on modulation complexity and plan quality were evaluated, considering the MLC architecture. Finally, to verify the implications in plan deliverability, the correlations between the complexity metrics and MU, and between GPRs and the novel metrics for dual-layer MLC were evaluated.

To compare the performance between the two linacs-MLC designs, Figure 4.3 presents the boxplots of all complexity metrics and plan quality indices that demonstrated a significant difference ($p<0.05$) between Halcyon-v2 and TrueBeam plans. It was found that Halcyon-v2 plans demonstrated lower values of V50%, mD-OAR1 (rectum), CI, MUcp, MCSv, and NP,

compared to TB plans. Additionally, it was noticed that TB plans presented more outliers, indicating less consistent results.

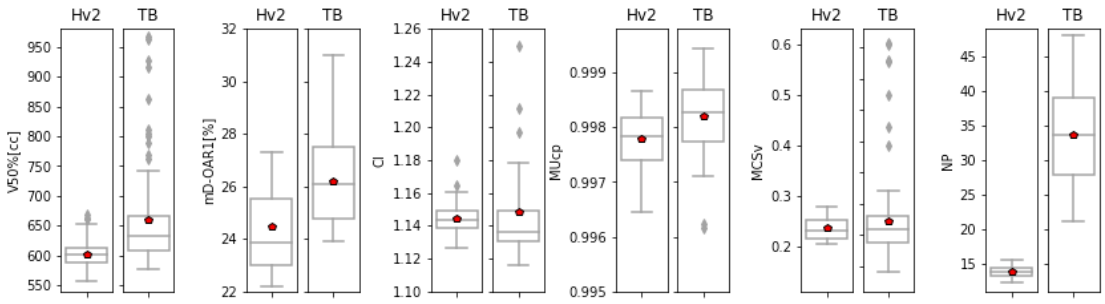


Figure 4.3 Boxplots of complexity metrics and plan quality indices that presented a significant difference between Halcyon-v2 (Hv2) and TrueBeam (TB) plans. The boxplot displays the minimum and maximum values of the data distribution indicated by the end of the whiskers; the lower and upper box limits are the first and third quartile; the horizontal line indicates the median value, and the red dot represents the mean value. Any additional point outside is considered as an outlier).

For each combination of TPS-parameter modes, the complexity scores and plan quality indices were compared using the Wilcoxon signed-rank test. Table 4.1 summarizes the parameter modes where significant changes were found. For Halcyon-v2 plans (dual-layer MLC) with Conv{off} were associated with slightly lower V50% values than Conv{extended}. However, the other TPS-parameters combinations did not influence the complexity nor the plan quality metrics significantly. In TrueBeam plans (single-layer MLC), the CI, HI, mD-PTV, and V50%, demonstrated significant differences for parameters combinations including ASC and DCR (Table 4.1). Furthermore, significantly lower values of MU were presented with ASC{off} compared to ASC{moderate}.

Table 4.1 Significant differences between the TPS-parameter modes on plan quality indices and complexity metrics for Hv2 and TB plans.

linac	Metric	Sample Size	TPS-parameter	Mean ± SD	p
Halcyon	V50% [cc]	32	Conv{Off}	597 ± 18	0.04
		32	Conv{Ext}	603 ± 24	
TrueBeam	CI	24	ASC{off}	1.14 ± 0.01	<0.01
		24	ASC{very_high}	1.15 ± 0.05	

		48	DCR{normal}	1.16 ± 0.06	<0.01
		48	DCR{high}	1.13 ± 0.02	
	HI	24	ASC{off}	0.10 ± 0.03	0.04
		24	ASC{very_high}	0.11 ± 0.05	
	mD-PTV	24	ASC{off}	105 ± 2	0.04
		24	ASC{moderate}	104 ± 2	
		24	ASC{low}	104 ± 2	<0.01
		24	ASC{very_high}	105 ±3	
	V50% [cc]	48	DCR{normal}	678 ± 101	<0.01
		48	DCR{high}	641 ± 54	
	MU	24	ASC{off}	802 ± 149	0.04
		24	ASC{moderate}	880 ± 134	
Abbreviations: TPS treatment planning system, Hv2 Halcyon-v2, TB TrueBeam, CI conformity index, HI homogeneity index, mD-PTV means dose of planning target volume, mD-OARn mean dose of OARn, V50% volume enclosed by the 50% isodose, ASC aperture shape controller, DCR dose calculation resolution, Conv convergence, SD standard deviation.					

Figures 4.4 and 4.5 present scatterplots of all the complexity scores against required MU for Hv2 and TB plans, respectively. For Hv2 plans, required MU showed a high correlation to MCSv (I*r*I= 0.97), MCSw (I*r*I= 0.96), MUcp (I*r*I= 0.78), and NP (I*r*I= 0.76); and a moderate correlation to UL (I*r*I= 0.69) and MCSUL (I*r*I= 0.58). For TB plans, MU showed high correlation only to MCSv (I*r*I= 0.92). Additionally, a remarkable data clustering by the upper_gEUD priority values was demonstrated for Hv2 plans (Figure 4.4), which is not present in the case of TB (Figure 4.5)

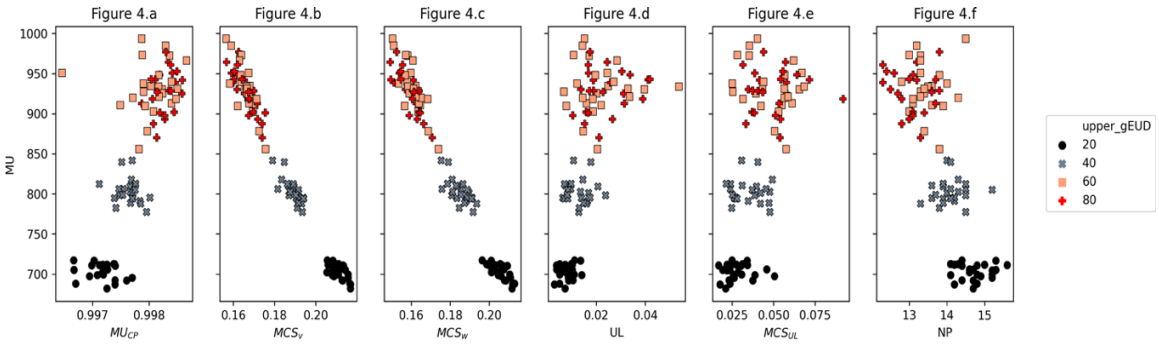


Figure 4.4 Scatterplot of all complexity metrics for Hv2 plans using the MU values as the reference score and considering the effect of different levels of dose sparing priorities (upper_gEUD values). Abbreviations: Hv2 Halcyon-v2, MU monitor units, MUcp average MU increment by control point, MCSv modulation complexity score for volumetric modulated arc therapy, MCSw the weighted MCSv for dual-layer multi-leaf collimator architecture, UL uncover layer score, MCSUL weighted MCSw by UL, NP number of peaks

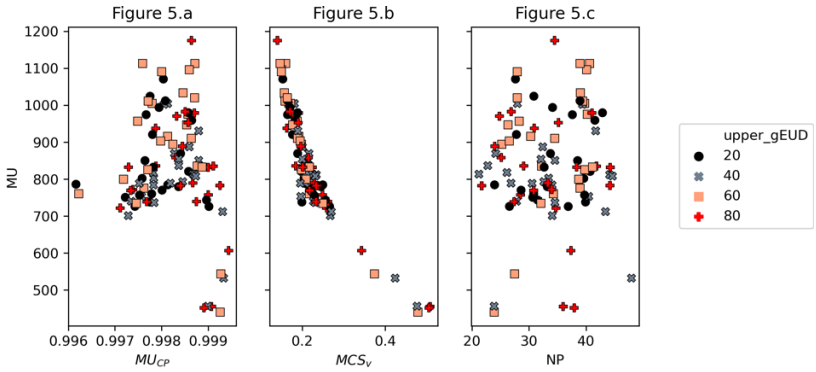


Figure 4.5 Scatterplot of all complexity metrics for TB plans using the MU values as the reference score and considering the effect of different levels of dose sparing priorities (upper_gEUD values). Abbreviations: TB TrueBeam, MU monitor units, MUcp average MU increment by CP, MCSv modulation complexity score for volumetric modulated arc therapy, NP number of peaks.

The GPR's for evaluation criteria of 3%/3 mm, 3%/2 mm, and 2%/ 3mm were always 100% for all cases and thus, were not considered in the analysis. The mean value and standard deviation (SD) for GPR with 2%/1 mm criteria were 96.3% and 1.7% respectively. Figure 4.6 shows the scatterplot of the complexity metrics against GPR, again plotted to indicate the associated upper_gEUD priority values. The GPR presented high correlation to MU, MCSv, and MCSw (|r|= 0.74, 0.74, and 0.72); moderate correlation to MUcp, UL, and NP (|r|= 0.66, 0.48, and 0.63); and low correlation to MCSUL. Additionally, the GPR present a similar clustering data effect as seen in Figure 4, with less differentiation for upper_gEUD priority values of 40, 60, and 80, compared to upper_gEUD values of 20.

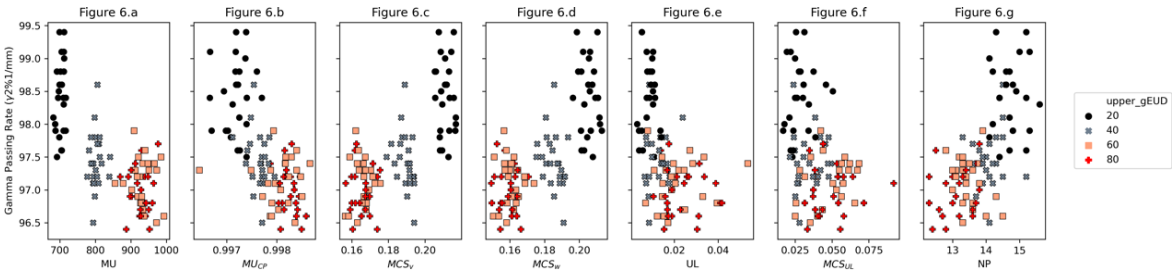


Figure 4.6 Scatterplot of all complexity metrics from 96 prostate plans delivered on Halcyon-v2 (Hv2), considering the gamma passing rate (GPR) values.

4.2.3 Discussion

This *in silico* study investigated the possible effects of the selected TPS-parameters on different plan quality and mainly in the modulation complexity metrics. At the same time, it was intended to verify and validate the modulation complexity metrics needed to train the further ML models designed to predict GPR values. Accordingly, three main aspects were considered to develop this research, (1) the TPS-parameters of ASC, DCR, and Conv were chosen because of their possible effects on the final dose fluence (*i.e.*, MLC movements) [31] , (2) the selected linacs Halcyon-v2 with dual-layer MLC and TrueBeam with single-layer MLC (since all the treatments from the ML models' datasets were planned and delivered in this linacs), and (3) one prostate CT data set was used as a virtual phantom to control any effects attributable to differences in anatomy or planning volumes [127].

Figure 4.3 summarises the statistically significant differences observed in the plan quality indices and modulation complexity, comparing Halcyon-v2 and TrueBeam plans. It was found that Halcyon-v2 plans were associated with a higher median value of CI, better dose-sparing contributions (lower V50%), and lower mean dose values in OAR1 (mD-OAR1) ($p<0.05$). As it is described in previous reports [20,108,128,129], these results can be attributed to the Halcyon-v2 features of lower penumbra (due to the leaf tip shape), higher leaf speed, lower dosimetric leaf gap, and higher gantry speed, compared to TrueBeam with Millennium-120 MLC. In the same way, it is important to note that these differences in features (hardware and beam modelling) make it impossible to directly compare the complexity metrics between the two linacs [127], and *these differences are part of a fundamental aspect needed to be addressed in the next chapter (Chapter 5) regarding the real impact of dataset heterogeneities (datasets with various treatments having different treatment units, dose, anatomic region, etc.) on the ML modelling of GPR values.*

From Figure 4.3, it is also clear that metrics from Halcyon-v2 plans demonstrate less spread or variation than the data from TrueBeam plans. Furthermore, the TrueBeam data exhibits considerable outliers in the CI, V50%, and MCSv. It was inferred from this observation that the Halcyon-v2 plans show more consistent outcomes or less sensitivity to the TPS parameters, than those for the TrueBeam configuration with SL-MLC. As we move towards the era of on-table adaptation (47-49), this reduced sensitivity to parameter variation may be an important feature regarding the requirement for rapid (high pressured) re-planning using either manual or automatic techniques, given both require some oversight and quality control (QC).

The results summarized in Table 4.1 demonstrate the selected TPS-parameters combinations do not impact the modulation complexity of plans with DL-MLC. Contrastingly, plans with SL-MLC

presented lower MU values for treatments with ASC{off} decreasing the plan complexity (Figure 4.5). For Halcyon-v2 plans, only the comparison between Conv{off} and Conv{extended} demonstrated a statistically significant difference in the V50% metric. Interestingly, with the mode set to "off", a lower mean V50% value was obtained; however, this reflected the narrower range of values achieved for this parameter settings compared to the "extended" mode. Thus, although the difference was evident, it is important to note that these variations may not represent considerable clinical differences.

For TrueBeam plans, the same scenario (low variations) happened to the CI, HI, and mD-PTV metrics. Moreover, lower values of V50% (achieved by DCR{high}) and MU (achieved by ASC{off}) presented relevant changes that might impact the plan quality and dose deliverability (26, 27). Nevertheless, the statistical significance needs to be careful considered in each particular case because each mode has different plans depending on their respective TPS-parameter. For instance, ASC with four modes has 24 plans each, whilst DCR (two modes) has 48 plans.

Figure 4.4 shows the correlation of all modulation complexity metrics with MU for Halcyon-v2 plans. Aside from the strong correlation seen in this data, a clear grouping level is evident with the priority settings used with the upper_gEUD optimisation constraint. For each of the plots (the different modulation) the data groups to the lower (20), moderate (40) and high/very-high (60/80) priority settings for the dose sparing parameter. These well-differentiated regions suggest a strong dependence between the modulation complexity degree (measured in 7 different ways), and the priority levels used to reduce the mean dose of OARs in the optimization process, therefore, providing an opportunity to "pre-select" the required range of solution in terms of acceptable complexity. *These results showed that high demanding dose sparing conditions might generate plans with higher MU values, with more complex modulation (lower values of MCSv and MCSw), with higher number of uncovered leaves junction per CP (UL), but at the same time with a lower number of demanding changes in the leaf position throughout the modulation process (NP), albeit a small effect of the latter*.

Figure 4.5, showing the same analysis for the TrueBeam plans, does not show a strong correlation, nor grouping. It is likely that the latter reflects the weaker overall correlation and the wider range of plan metrics previously highlighted. Comparison of the corresponding plots in Figure 4.4 and Figure 4.5 suggests again that the variation of the treatment planning parameter modes has a smaller effect on the plans produced for the Halcyon-v2 model over that for the TrueBeam. This is particularly apparent in the behaviour seen in Figure 4.5c, where a much greater heterogeneity is seen in the data. The large variation in Numbers of Peaks seen in the leaf trajectories suggest an 'unstable' relationship between the leaf sequences generated

and parameter variation. In turn this indicates the 'TrueBeam' optimisation search space is far more complex and poorly behaved, with many local minima, leading to these spreads of 'optimal' solutions. This should not be taken as a reason to distrust the algorithms; however, it does emphasize the need for caution, QC and oversight of the planning process.

The different behaviour shown in Figure 4.4 and Figure 4.5, suggests that PO algorithm might work differently for the two Linac/ MLC models when the optimization priorities are used to reduce the OAR mean dose. In general terms, it was expected that more demanding plans (with higher dose sparing priorities) would require more complex beam modulation with higher MU values. This was evident in the results seen for Halcyon-v2 cases, however for TrueBeam plans, it seems to be uncorrelated; suggesting that a common optimization template could not be expected to produce similar results for the different Linac/MLC models. Nevertheless, this behaviour needs to be analysed in additional investigations considering other optimization parameters used to control the dose of OARs and the potential impact on dose deliverability.

Considering previous publications that explored various modulation complexity metrics [92,94,97], it is important to note the contrasting results regarding their correlation to GPR values. While this study found high and moderate correlations for specific metrics using the 2%/1 mm criteria (high correlation: MU (|r|= 0.74), MCSv (|r|= 0.74), and MCSw (|r|= 0.72), moderate correlation: MUcp (|r|= 0.66), UL (|r|=0.48), and NP (|r|=0.63)), studies performed by Park *et al.* [92] and Tamura *et al*. [97] found, respectively, low correlation to GPRs analysing the TrueBeam results (MCSv (|r|= 0.21)) with 40 prostate plans, and low correlation analysing the Halcyon-v2 results (MCSw (|r|=0.0122), MUcp (|r|=0.0084), MCS_5 (|r|=0.2131)) with 15 prostate plans. These remarkable results might be attributed to different treatment unit models, TPS, and the detectors, used in the previous studies. In contrast, as it was studied before with a standard phantom [127], constant target volume geometry and constant treatment conditions should be considered as a factor that improves the correlation between the dose deliverability and modulation complexity. In other words, analysing together the GPR values of a series of patients treated with different technology (hardware, dosimetry, and software) and with different plan conditions (static beams, VMAT plans with one, two, or more arcs) are always prone to introduce additional noise to the data representation since those treatment parameters should not be comparable because of their physical representation in practice.

Finally, and more directly connected to the goals of the present thesis exploring new complexity metrics, the correlation between the novel modulation complexity scores (UL, NP) and the GPR showed a moderate correlation (Figure 4.6). In line with the thesis objectives regarding the need to track the specific treatment parameters associated with the predicted dose deliverability

evaluation, these new complexity metrics designed for dual-layer MLC account for traceable physical aspects that may impact the delivered dose, while other published metrics cannot, being valuable to include them in a treatment verification programs. However, a clear clustering of the data with dose-limiting priority value is evident and suggests a simple connection between driving the optimiser harder (higher priority) and obtaining more complex solutions (higher MU and MUCP, and lower MCSV, MCSW, NP), which intuitively challenge attaining a maximum GPR.

4.3 Radiomic Features

The radiomic features are a set of mathematical extracted metrics based on texture analysis performed mainly in high-dimensional datasets [130,131]. In oncology and medicine, the images have been studied with these features implementing statistical tests contributing to genomics, protein sequencing, metabolomics, and medical images analysis for treatment outcome predictions [54,83,130]. In RT, specifically, the analysis of the dose distribution images is called *dosiomics*, and it has been implemented to predict lung toxicity, overall survival, and linac performance [77,132,133].

The radiomic features were calculated with Pyradiomics [131] using the 3D dose distribution (*i.e.,* dosiomics), and are summarised in Table 4.2. Additionally, it was proposed two new radiomic features sets. First, using the 2D image created with all the MLC movements through each control point per arc (modulation maps, MM), extracting additional features related to modulation complexity that might not be calculated using conventional equations. And secondly, it was extracted the radiomic features using the calculated blended image per arc used for portal dosimetry evaluations to consider the final composite dose distribution of each beam used for gamma index analysis

Table 4.2 Radiomic features and sub-features [131]

Main Radiomic Feature	Num. of sub-features	Sub-features	Description
First Order Statistics	19	Energy, Total Energy, Entropy, Minimum, 10^{th} percent, 90^{th} percent, Maximum, Mean, Median, Interquartile range, range, Mean abs deviation, Robust mean abs dev, Root mean square, Standard deviation, Skewness, Kurtosis, Variance, Uniformity.	Based on the distribution of voxel intensities from the image
Shape-Based (3D)	16	Mesh volume, Voxel volume, Surface area, Sphericity, Compactness 1, Compactness 2, Surface area to volume ratio, Spherical disproportion, Max 3D diameter, Max 2D diameter (slice), Max 2D diameter (column), Max 2D diameter (row), Mayor axis length, Minor Axis Length, Least Axis Length, Elongation, Flatness.	Descriptors of three-dimensional shape and size calculated on the non-derived image

Shape-Based (2D)	10	Mesh Surface, Pixel Surface, Perimeter, Perimeter to Surface ratio, Sphericity, Spherical Disproportion, Maximum 2D diameter, Major Axis Length, Minor Axis Length, Elongation	Descriptors of two-dimensional shape and size calculated on the non-derived image
Gray Level Cooccurrence Matrix (GLCM)	24	Autocorrelation, Joint Average, Cluster Prominence, Cluster Shade, Cluster Tendency, Contrast, Correlation, Difference Average, Difference Entropy, Difference Variance, Dissimilarity, Joint Energy, Joint Entropy, Homogeneity 1, Homogeneity 2, Informational Measure of Correlation (IMC) 1, Informational Measure of Correlation (IMC) 2, Inverse Difference Moment (IDM), Maximal Correlation Coefficient (MCC), Inverse Difference Moment Normalized (IDMN), Inverse Difference (ID), Inverse Difference Normalized (IDN), Inverse Variance, Maximum Probability, Sum Average, Sum Variance, Sum Entropy, Sum of Squares.	GLCM describes the second-oriented joint probability function of an image contrasted with a mask
Gray Level Run Length Matrix (GLSZM)	16	Small Area Emphasis, Large Area Emphasis, Gray Level Non-Uniformity, Gray Level Non-Uniformity Normalized, Size-Zone Non-Uniformity, Size-Zone Non-Uniformity Normalized, Zone Percentage, Gray Level Variance, Zone Variance, Zone Entropy, Low Gray Level Zone Emphasis, High Gray Level Zone Emphasis, Small Area Low Gray Level Emphasis, Small Area High Gray Level Emphasis, Large Area Low Gray Level Emphasis, Large Area High Gray Level Emphasis	Quantify grey level zones in an image
Gray Level Size Zone Matrix (GLRLM)	16	Short Run Emphasis, Long Run Emphasis, Gray Level Non-Uniformity, Gray Level Non-Uniformity Normalized, Run Length Non-Uniformity, Run Length Non-Uniformity Normalized, Run Percentage, Gray Level Variance, Run Variance, Run Entropy, Low Gray Level Run Emphasis, High Gray Level Run Emphasis, Short Run Low Gray Level Emphasis, Short Run High Gray Level Emphasis, Long Run Low Gray Level Emphasis, Long Run High Gray Level Emphasis.	Quantifies the grey level runs, which are defined as the length in number of pixels, of consecutive pixels that have the same grey level value
Neighbouring Gray Tone Difference Matrix (NGTDM)	5	Contrast Feature Value, Coarseness Feature Value, Busyness Feature Value, Complexity Feature Value, Strength Feature Value.	Quantifies the difference between a grey value and the average grey value of its neighbours within a defined distance
Gray Level Dependence Matrix (GLDM)	14	Small Dependence Emphasis, Large Dependence Emphasis, Gray Level Non-Uniformity, Gray Level Non-Uniformity Normalized, Dependence Non-Uniformity, Dependence Non-Uniformity Normalized, Gray Level Variance, Dependence Variance, Dependence Entropy, Dependence Percentage, Low Gray Level Emphasis, High Gray Level Emphasis, Small	quantifies grey level dependencies in an image. A grey level dependency is defined as the number of

		Dependence Low Gray Level Emphasis, Small Dependence High Gray Level Emphasis, Large Dependence Low Gray Level Emphasis, Large Dependence High Gray Level Emphasis	connected voxels within distance d that are dependent on the centre voxel.

4.4 High-dimensional Complexity Features

The input features considered for CNN-based models to predict GPR were selected because its direct relation with the conventional modulation complexity scores. They are:

4.4.1 Modulation Maps - MM

The modulation maps (MM) input feature from a single VMAT-arc is a two-dimensional image created with all MLC positions per cp (Figure 4.7.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 4.7.b).

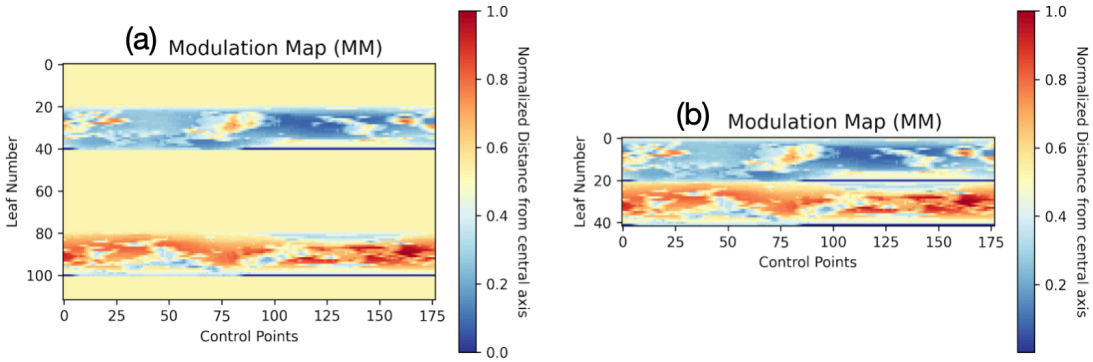


Figure 4.7 (a) Modulation maps (MM) of one prostate plan VMAT-arc including both MLC banks representing each leaf trajectory throughout 180 control points. (b) MM of the same treatment removing the static fields

4.4.2 MUcp_profile

The MUcp_profile 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 4.8). It is extracted from the dose contribution coefficient within the DICOM-RT tag [300A,010C] labelled *CumulativeDoseReferenceCoefficient*.

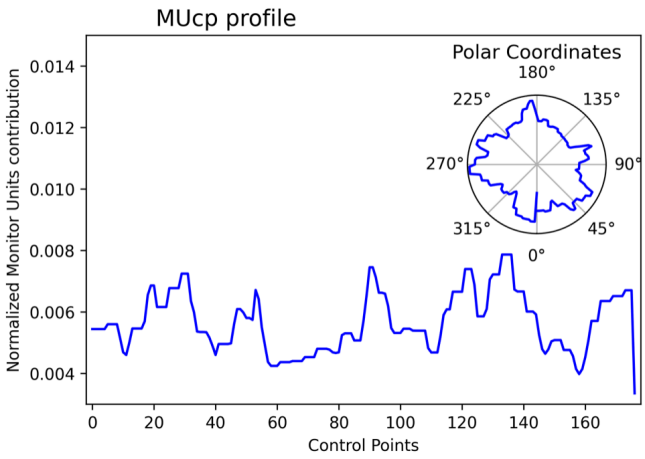


Figure 4.8 Normalized monitor units per control point profiles (MUcp_profile) throughout 180 control points from one prostate VMAT-arc plan. Additionally, a polar plot is integrated to represent the MU contribution in each VMAT-arc section.

4.4.3 Composite Dose Image - CDI

The composite dose image (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 [31,134] 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.

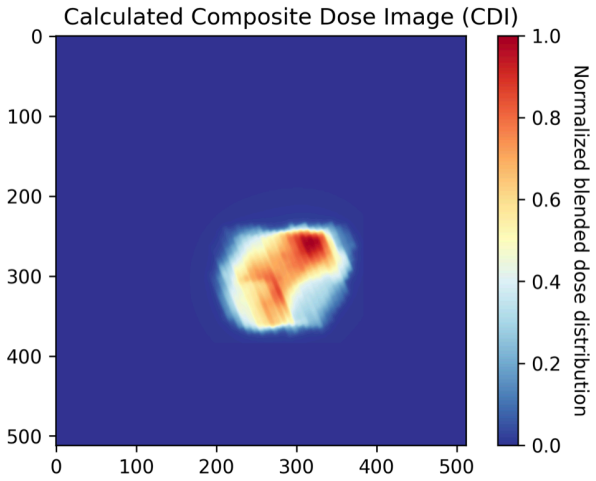


Figure 4.9 Composite dose image created with all dose fluences delivered per each control point from one prostate VMAT-arc. This image is calculated by the TPS and is used to perform portal dosimetry evaluation, comparing this dose distribution with the measured by the EPID.

4.5 Conclusions

Metrics dedicated to accounting for new ways to measure modulation complexity for dual-layer MLC models were developed and validated, generating new numeric predictors that might improve the prediction performance of GPR modelling using ML algorithms. However, the study followed to validate these metrics also highlights two main aspects. First, the variations of hardware and planning optimization parameters influence the range of possible modulation complexity scores, even for similar clinical conditions. And secondly, the conventional modulation complexity scores do not provide exact information about the hardware planning parameters linked to a specific predicted GPR value.

The modulation complexity score variations for treatments with different optimization conditions or treatment unit hardware are an essential starting point to evaluate the potential technical effects in modelling GPRs using datasets with diverse treatment conditions (as it will be studied in the next chapter, Chapter 5). Explicitly, due to the modulation complexity corresponds to specific treatment conditions, an ML model trained with a dataset with not enough data representation of specific treatment conditions (a heterogeneous dataset), might present a prediction weighted by random processes rather than physical parameters involved in the treatment deliverability.

Since the conventional complexity metrics do not provide information about specific plan parameters, high dimensional metrics associated with modulation complexity and its variations throughout the time (control points) are a natural step to be explored by implementing CNN-based models (Chapter 6). For this reason, it was extracted the MM and MUcp profile (Section 4.4.1), which might support a more comprehensive analysis of dose deliverability, rather than predict one GPR value, retrieving specific plan delivery moments more relevant to the GPR modelling.

In summary, the proposed extracted metrics for Halcyon-v2 and the other retrieved or calculated metrics suggest two main aspects that will be addressed in the following chapters. First, the effect of the dataset composition using conventional GPR predictors (Chapter 5) since it was demonstrated that changes in hardware and software conditions influence the GPR results. And second, the poor utility of these metrics in understanding the model prediction reliability and the lack of specific plan parameters representation associated with dose deliverability. For this reason, models with high-dimensional features were also explored (Chapter 6).

Chapter 5 Dataset effects

As it has been discussed previously (Chapter 4), GPR predictions applying ML methods [77,81,115,116] have been explored using complexity metrics as the principal predictor variables due to its reported relationships with dose deliverability [94]. One of the first reported studies in this field, performed by Valdes et al.[116], proposed a generalized linear model (GLM) with Poisson regression and Lasso regularization. They predicted GPR values based on 498 intensity-modulated radiation therapy (IMRT) plans for six anatomical sites[2], 78 complexity metrics, and diode-array detector measurements, reporting prediction error of up to 3%. These results were validated later[63] using an external dataset consisting of 139 plans (no anatomical sites specified), 90 complexity metrics, and EPID measurements, reporting an overall error of up to 3.5%. In their report, Valdes et al. suggested the need to explore potential effects in model performance with datasets having various anatomical regions, treatment unit models, or detector types. Simultaneously and based on the same dataset, Interian et al. [78] proposed a convolutional neural network (CNN) using transfer learning and a VGG-16 architecture [60] to predict the GPR values using the dose fluence maps from each field, reporting a mean absolute error (MAE) of 0.70 ± 0.05. As a result of all these previous works, *they noted that the model performance might be compromised by factors such as the dataset size, potential detector misalignments, the use of different dose detectors, and, more critically, the imbalanced dataset* (*i.e.,* most of the plans from the training dataset had GPR values close to 100%, inducing a poor model performance when plans with lower GPR values are predicted).

Exploring new GPR prediction strategies, Tomori et al.[81] implemented a 15-layer CNN architecture with sagittal planar dose distributions (Gafchromic EBT3 films) from 60 IMRT prostate plans and the volume data from the planning target volume (PTV), rectum, and their corresponding overlapping regions. Despite having a small dataset, Tomori et al. found moderated and strong correlations between the predicted and calculated GPR values. In addition, they suggested that exploring other GPR criteria with new threshold reference levels might provide better information about potential clinical errors [106] dealing with imbalanced datasets simultaneously. This rationale was further explored by Li et al. [135], who presented GPR regression and classification implementing Poisson Lasso and random forest (RF) models,

[2] Breast (N=110), Central Nervous System (N=58), Gastrointestinal (N=78), Genitourinary (N=64), Gynaecologic (N=19), H&N (N= 5), Lung (N=134), Paediatrics (N=30).

respectively, evaluating the GPR criteria of 3%/3 mm, 3%/2 mm, and 2%/2 mm. They calculated 54 complexity metrics and used a dataset with 303 VMAT plans with two different anatomical regions (head and neck - H&N, and gynaecologic) and various dose per fraction schemes, achieving a prediction error of up to 4.2% and classification sensitivity of 100% for 2%/2 mm criteria, which was the GPR criteria with less imbalanced results. Moreover, Li et al. suggested that classification would be a more convenient method than regression for virtual specific-plan verification, following the American Association of Physicists in Medicine (AAPM) task group 218 [4] recommendations regarding clinical feasibility and action limits to classify one plan as fail or pass. Indeed, as it was studied later by Nguyen and Chan [107], after choosing one GPR criteria and one tolerance reference level able to detect potential clinical errors in practice, as more interpretable to identify if one plan will pass or fail rather than predict one single GPR value.

The features variations and their weighted importance for GPR predictions were investigated by Lam et al.[136] with AdaBoost, RF, and extreme gradient boosting (XG-Boost) algorithms. Despite the differences in their operational basis[57,137], the three models identified the same nine most important complexity features (all related with modulation complexity and treatment unit model). Lam et al. used a dataset with 189 IMRT plans consisting, heterogeneously, of treatments from 10 different anatomical regions[3], 31 complexity metrics, four dosimetrically matched treatment units, and EPID dose measurements. Although their dataset was highly unbalanced, they evidenced that GPR prediction is feasible with EPID measurements achieving an accuracy of up to 3%. Also, Lam et al. emphasized *the need to explore potential variations in complexity metrics when considering different anatomic regions and treatment units*. Similarly, Ono et al.[80] investigated the variations of the features importance for GPR predictions using one regression tree analysis (RTA), multiple regression analysis (MRA), and one neural network (NN) models adopting a dataset with 600 VMAT plans with 28 complexity metrics and helical diode-array measurements. Even with an unbalanced dataset consisting of different anatomic regions[4] (heterogeneously represented), they noticed that regression tree models are not always suitable for continuous values predictions due to their implicit accuracy dependency on the number of nodes and the values range of the predicted feature. On the other hand, the NN

[3] head-and-neck (N=38), abdomen (N=5), bladder and rectum (N=6), brain (N=36), lung and oesophagus (N=20), breast and chest-wall (N=12), pelvis(N=27), extremity (N=2), and prostate (N=36)
[4] Brain(N=184), head and neck (N=89), lungs (N=36), oesophagus (N=39), abdomen (N=15), pancreas (N=82), prostate (N=106), pelvis (N=46), and spine (N=3)

model presented the best prediction performance with an error of up to 0.70% ± 0.05%, evaluating the GPR 3%/3 mm criteria.

As an alternative method to predict dose deliverability, Granville et al.[138] demonstrated that complexity metrics, plan parameters, and data from daily quality assurance (QA) measurements applied to the treatment units were adequate predictor features of 'dose differences' between measured and calculated dose distributions. They used a support vector (SV) classifier with a linear kernel to identify three potential conditions: hot, normal, and cold plans based on dose differences higher, in between, or lower than 1%, respectively. Simultaneously, they showed the model performance advantages associated with recursive feature elimination, reducing the model complexity to increase the results interpretability. More recently, Wall and Fenot [87] also investigated the positive impact of feature selection on GPR predictions, implementing three feature selection methods (extra-trees, mutual information, and linear regression), three machine learning algorithms (SV, tree-based model, and NN), and using a 500 VMAT plans dataset with different anatomical sites[5]. Besides confirming a considerable improvement in prediction performance with a reduced number of features, they also suggested the potential benefit of bringing dedicated datasets for each anatomical region due to their different variations in GPR values. Subsequently, Well and Fenot [139] used the same dataset and the SV machine to improve the inverse optimization process for VMAT plans, detecting potential lower GPRs and changing specific treatment unit parameters (e.g., field aperture) and benefit the final treatment prediction. Aside from proposing the first application of GPR prediction directly in VMAT optimization, they recommended exploring the benefits of dedicated machine-specific models for GPR predictions, supported by previously reported suggestions[63].

New approaches to improve ML-based GPR predictions have been recently studied, proposing alternative features or datasets, and changing how dose deliverability could be inferred from the calculated and measured dose data. That is the case of Hirashima et al. [77], who reported model performance improvements using radiomic features [83,84] as additional predictors for GPR prediction. They used a dataset containing 1225 VMAT plans[6] with 24 complexity metrics and 851 radiomic features extracted from 3D dose distributions of each plan (3D dosiomics). In addition, the regression and classification of GPR values were performed with the XG-Boost model, showing improvements in sensitivity and specificity using a hybrid dataset with both

[5] Abdomen (N=29), reast (N=13), Chest (36), chest wall (13), H&N (148), lung (127), prostate (61), prostatic fossa (30), pelvis (32), and miscellaneous (11).
[6] Brain (N=480), H&N (N=171), Oesophagus (N=70), Lung (109), Pancreas (115), Abdomen (38), Pelvis (N=119), Prostate (N=153)

types of features. Also, *Hirashima et al. found that the classification performance might be affected if the dataset includes many treatment sites due to the correlation between modulation complexity and the anatomical region.* In contrast, Tomori et al.[79] considered a new synthetic dataset for model training using 96 dummy plans based on virtual spheric phantoms with different dose restrictions and OARs to emulate various dose delivery scenarios. They tested the model with a 51 clinical plans dataset, predicting 36 pairs of GPR criteria simultaneously. Although they found moderate statistical significance in testing dataset results, even in extreme GPR criteria like 0.5%/1 mm, they proposed new ways to understand and extract potential predictor features for GPR values. Nevertheless, the results might be limited by the training dataset size and its variations within the synthetic dataset in the target volumes, number of OARs, and dose per fraction.

Considering the previously mentioned ML-based methods and their potential options for virtualspecific-plan verification, their implementations in clinical practice can represent benefits regarding time and new safety filter protocols. However, two limitations might influence the models' interpretability and the quality of their predictions. First, previous publications had imbalanced datasets in GPR terms (most plans have the 'passing' criteria label or high GPR values), leading to a limited prediction performance in plans with low GPR values when it is more critical to act. In addition, the classification performances were measured with the area under the curve (AUC) from the receiver operating characteristic (ROC) analysis (ROC-AUC), which is a metric suited for balanced datasets. Second, the datasets referenced before were assembled, *not proportionally*, by treatments for different anatomic regions having, in turn, unbalanced differences in the numbers of beams, the dose per fraction scheme, the beam energy, and the treatment units. These heterogeneity factors might influence the model performance due to the demonstrated correlation between GPR values and treatment parameters such as modulation complexity metrics, anatomic region, and dose per fraction [76,92,103,140–142].

Moreover, the dataset heterogeneities could impact the weighted importance of the predictors used by the model [87,138], missing the ‘real’ physical aspects involved in the prediction process that correspondingly might explain the GPR prediction causes of one unrepresented plan. For this reason, the main goal of this study was to understand how dataset heterogeneities might impact the model performance in GPR predictions. Consequently, it was aimed to explore enhanced dataset conditions to create dedicated datasets that will be explored in the Chapter 6.

5.1 Specific materials and methods

Anonymized VMAT plans for 945 cases treated in one institution[7] were retrospectively extracted to evaluate the potential effects of dataset heterogeneities on model prediction performance for GPR binary classification (pass/fail). Accordingly, 25 datasets were designed controlling the number of treatments based on four heterogeneity factors: *anatomical region*, *dose per fraction*, *treatment unit*, and the *number of arcs* (VMAT-arcs). From these plans, 309 predictor features were extracted, 20 of them were plan parameters (dose and volume data), 14 were complexity metrics, and 285 were radiomic features. Finally, the classification performance of three AI algorithms were evaluated for each dataset to verify the most favourable dataset assembling conditions.

The 945 VMAT plans contained a total of 1150 VMAT-arcs and represented prostate (N= 840), H&N (N= 49), and brain (N= 56) treatment sites. The plans were optimized with Eclipse 15.6 (Varian Medical Systems, CA) using the anisotropic analytical algorithm (AAA), setting a 0.25 mm grid size calculation resolution and 2^{o} spacing per control point. The treatment unit models were three Varian linear accelerators calibrated at the same reference conditions, two dosimetrically matched TrueBeams (TB) and one Halcyon-v2 (HL), with the same nominal resolution at isocentre (5 mm) and same EPID model. Both TBs and HL hardware specifications were described in Section 3.2.1.

5.1.1 The datasets

Four groups with six datasets each (24 datasets) were assembled to study the individual impact of their corresponding four heterogeneity factors. For each datasets group, the six datasets were designed with systematic variations of the number of plans with one of the two contrasting treatment characteristics ({*a*}, {*b*}), representing their specific heterogeneity (Table 1). For instance, for the *dose per fraction* heterogeneity, the six datasets were assembled with variations of the number of prostate plans having {*a*}= 2 Gy or {*b*}= 3 Gy per fraction, keeping constant the dataset size (n= 210) and the other heterogeneity factors, such as *treatment unit*, *number of arcs*, and *anatomical region*. Therefore, the first of the six datasets was constituted by 100% of plans having the characteristic {a} plus 0% of plans with the characteristic {b} ({*a%/b%*} = 100%/0%), and subsequently the other five datasets had 80%/20%, 60%/40%, 40%/60%, 20%/80%, and 0%/100% proportions. The makeup of the datasets groups is defined in Table 5.1.

[7] Queen's Centre for Oncology and Haematology - Castle Hill Hospital. Hull University Teaching Hospitals NHS Trust.

The heterogeneity factor of *anatomical region* considered prostate plans as characteristic {*a*}, and non-prostate plans as characteristic {*b*}. Due to the limited dataset, H&N and brain plans were combined as the non-prostate characteristic. For the *number of arcs* heterogeneity, {*a*} and {*b*] were prostate plans delivered by one single arc and two arcs (first and second arc), respectively. As it was mentioned before, the *dose per fraction* heterogeneity considered 2 Gy as characteristic {*a*} and 3 Gy as characteristic {*b*}. Finally, for *treatment unit* heterogeneity, the TB and HL were the characteristic {*a*} and {*b*}, respectively.

Additionally, one reference dataset was randomly assembled without any control of the characteristics within the heterogeneity factors (Table 1). This reference dataset was used to compare the models' performances against those observed with the datasets with controlled heterogeneity factors. Thus, to ensure that across the created datasets, the inputs to each model were the same, 210 inputs were provided for each dataset. Given the plans had one or two arcs, each arc was taken as a single input to the model, therefore the number of plans contributing to the heterogeneity characteristics within the datasets reflected the relative combinations as well as the number of arcs.

Table 5.1 Datasets classification for each principal heterogeneity

Heterogeneity evaluated	Total inputs	Number Datasets	Anatomic region	Type of arc	Dose per fraction [Gy]	Treatment Unit
Reference	210	1	P (n= 94) B (n=71) H&N (n=45)	Single (n=35) Two (n =175)	2.0 (n= 138) 3.0 (n= 72)	HL (n=57) TB (n=153)
Anatomical region	210	6	{*a*} P (n= 105) ---------- {*b*} B (n= 53), H&N (n=52)	Two	2	TB
Number of arcs	210	6	P	{*a*} Single ---------- {*b*} Two	3.0	HL
Dose per fraction	210	6	P	Single	{*a*} 2.0 ---------- {*b*} 3.0	HL
Treatment unit	210	6	P	Single	3.0	{*a*} TB ---------- {*b*} HL
Abbreviations: Treatment units, TU; flattening-filter-free, FFF; flattened-filter, FF; Halcyon-v2, HL; matched treatment unit, ML; prostate, P; brain, B; head and neck, H&N. {*a*}: Initial plan characteristic, {*b*}: contrasting plan characteristic.						

The specific-plan verification was performed with EPID measurements and their respective image analysis by Eclipse's portal dosimetry tools, applying the absolute dose correction and the improved gamma evaluation mode. The TBs and HL machines had the detector model aS1200, specified in Section 3.2.3. These GPR values were used to identify a GPR threshold (percentage of evaluated points passing the comparison criteria) that grants balanced distributions within the dataset (i.e., ideally 50% pass and 50% fail). Simultaneously, the chosen GPR criteria and threshold had to be suitable for detecting potential clinical errors[106], avoiding unbalanced datasets[81,104,107], and excluding this potential bias within the model performance.

5.1.2 Feature selection

For each of the 945 VMAT plans, the anonymized files of the Digital Imaging and Communications in Medicine Radio Therapy (DICOM-RT) [118] plan, DICOM-RT dose, and DICOM-RT image were extracted. These files contain the plan parameters (such as the MLC trajectories and field geometry), the dose distribution, and the composite dose image (CDI) of each individual VMAT-arc used for portal dosimetry analysis, respectively. Based on previous studies[80,143], the predictor features associated with complexity metrics and plan parameters were calculated using Python [98] and ESAPI [100] (Eclipse Scripting Application Interface) scripting. Considering the different designs, appropriate complexity metrics were used respectively for single (TB) and dual (HL) layered MLCs. For HL, the complexity metrics were weighted by the respective MLC-layer contribution to the final field conformation [143], following the recommendations of Tamura et al.,[97]. Additionally, and considering the demonstrated improvements in model performance due to the inclusion of texture analysis features from dose distributions [77,83,84], radiomic features were calculated with Pyradiomics [131] using the 3D dose distribution (*i.e.,* dosiomics). Also, it was proposed two new radiomic features sets. First, using the 2D image created with all the MLC movements through each control point per arc (modulation maps, MM), extracting additional features related to modulation complexity that might not be calculated using conventional equations. And secondly, it was extracted radiomic features using the CDI per arc used for portal dosimetry evaluations to consider the final composite dose distribution of each beam used for gamma index analysis. All features are listed in Table 2.

To evaluate the impact of all these radiomic features on the classification performance, the AI algorithms were tested using the reference dataset evaluating four conditions, (1) dataset with only plan (volume and dose) parameters and complexity metrics as predictor features, (2) previous condition plus radiomic features from MMs, (3) previous conditions plus radiomic

features from 3D dose distributions, and (4) previous conditions plus radiomic features from CDIs.

Table 5.2 Classification and summary of predictor features.

Feature Classification	Description	N	Features
Volume/Dose Plan parameters	Planning parameters specific dose values received in PTV and OARs Volume of overlapping structures Treatment unit	20	Dose per fraction, dose calculation algorithm, treatment unit, anatomic region, number of arc, PTV volume, volume of OARs in contact with PTV, volume of the overlapping region between PTV and OARs, beam mode, PTV-D98%, PTV-D95%, PTV-D50%, PTV-MD, OAR1-D50%, OAR1-MD, OAR1-D2%, OAR2-D50%, OAR2-MD, OAR2-D2% (Section 4.2.1.4)
Modulation Complexity	Beam modulation, Field variability Gantry speed and dose rate variation	14	LSV, AAV, MCSv, MCSw, LTMCS, NP, MU, MUcp, AA, AI, AM, PA, PI, PM (Section 4.1)
Radiomic metrics	pixel size, / voxel volume, skewness, etc.[144] MM (Radiomics1) 3D dose distribution (Radiomics2) CDI (Radiomics3)	285	GLCM (N= 24), GLDM (N= 14), GLRLM (N= 16), GLSZM (N= 16), NGTDM (N= 5), first order (N= 10), shape (N= 10)

Abbreviations: leaf sequence variability, LSV; aperture area, AAV; modulation complexity score for volumetric modulated arc therapy, MCSv; MCSv weighted by dual-layer multi-leaf collimator, MCSw; average leaf travel for MCSv, LTMCS; number of peaks of leaf trajectory, NP; monitor units, MU; averaged MU increment per control point, MUcp; aperture area, AA; area irregularity, AI; aperture modulation, AM; plan-averaged beam area, PA; plan-averaged beam irregularity, PI; plan-averaged beam modulation, PM; planning target volume, PTV; organ at risk, OAR; dose received by y% of the X structure's volume, X-Dy%, grey level co-occurrence matrix, GLCM; Gray level dependence matrix, GLDM; grey level run length matrix, GLRLM; Gray level size zone matrix, GLSZM; neighbouring grey tone difference matrix, NGTDM.

5.1.3 Models

Three machine learning (ML) models, RF, XG-Boost, and NN (Section 2.2) were implemented to perform GPR binary classification, using standardized datasets with 80% and 20% stratified train-test split. Five-fold-cross-validation (5-CV) was applied to reduce potential overfitting effects, and the hyperparameters tuning was performed by the grid-search method provided by Scikit-learn [145] package. The three models were optimized using the reference dataset to set the fixed model parameters for the remaining datasets and analyse the models' capabilities considering only the different datasets compositions. In addition, for each modelled dataset for RF, the features importance was calculated based on Gini importance [146] method using the Scikit-learn function '*feature_importances_*' due to its features managing consideration.

5.1.4 Model evaluation

The model performance was evaluated using ROC-AUC [77,104], which is defined as the area score from the curve calculated with the true positive rate (TPR) and the false positive rate

(FPR)[147]. The TPR (also known as Sensitivity) is defined as the ratio between the number of correctly positive classified cases, true positives (TP), and the actual total number of positive cases, which are the TP plus the classified false negative (FN) events (TPR = TP/(TP+FN)). In this study, sensitivity measures the model's capability to detect pass-plans that will pass the PPSTV when actually performed. Similarly, the FPR is defined as the inverted Specificity (FPR = 1-Specificity), or the number of false positives (FP) divided by the sum of FP and the true negative (TN) classifications (FPR = FP/(FP+TN)). This metric summarizes how often a plan is classified as pass when, in fact, it will fail in practice. When used as a 'screening-tool' to decide which plans may require actual measurements, a model with low TPR and high FPR would over-estimate those predicted to fail and result in the irradiation of more plans than necessary.

The Spearman's rank test (r) was performed to investigate whether the predictors features correlated with GPR values for a specific dataset. Low, moderate, and high correlations were considered for values of $|r| < 0.4$, $0.4 \leq |r| \leq 0.7$, and $|r| > 0.7$ respectively.

5.2 Results

5.2.1 Analysis of datasets

The clinically measured/calculated GPR values for all HL plans with 3%/3 mm, 3%/2 mm, 2%/3 mm, and 2%/ 2mm criteria were 100%. Thus, to ensure balanced datasets, the GPR classifications were performed adopting a GPR threshold of 95% as a plan pass/fail indicator for 2%/1 mm GPR criteria. Their respective distributions of the GPR values in all six datasets for each of the reference and the four heterogeneity factors are displayed in Figure 5.1.

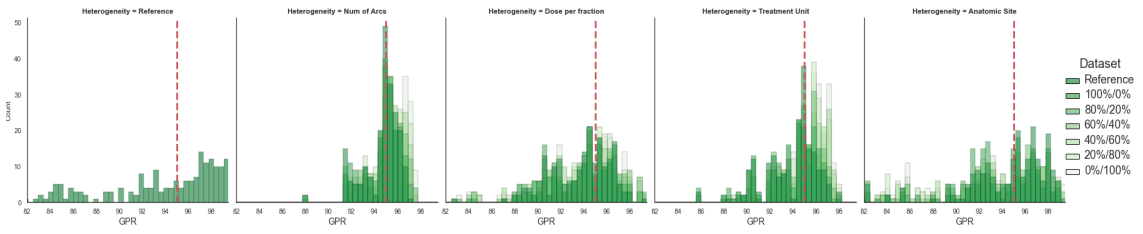


Figure 5.1 Distribution of Gamma Passing Rate (GPR) values for all dataset groups considering the reference dataset and each heterogeneity factor: Number of arcs, Dose per fraction, Treatment Unit, and Anatomic site. Each distribution group accounts their respective six datasets. The red dashed line marks the GPR threshold of 95% considered for the model classification pass/fail. Plans with GPR < 95% might be considered as plans that need to be investigated and will potentially fail.

Statistically moderate correlations to GPR with 2%/1 mm criteria were found for the predictors MCSv and the GLRLM radiomic feature extracted from MMs (Table 5.1). The scatterplot of these features against GPR values are included in supplementary material 1.1.

Table 5.3 Predictor features with statistically moderate correlation (Spearman's rank test) with GPR measurements based on 2%/1 mm criteria.

Dataset [Heterogeneity (a%/b%)]	**Feature**	**|r|**
Number of arcs (20%/80%) *	MCSv	0.47
Number of arcs (0%/100%)	MCSv	0.62
Treatment Unit (60%/40%) **	original_GLRLM_RunLengthNonUniformity	0.45
	original_ GLRLM _RunVariance	0.43
	original_GLRLM_RunLengthNonUniformity Normalized	0.42
	original_ GLRLM _RunEntropy	0.45
Treatment Unit (40%/60%)	original_ GLRLM _RunLengthNonUniformity	0.44
	original_ GLRLM _RunVariance	0.41
Treatment Unit (20%/80%)	original_ GLRLM _RunVariance	0.41
Abbreviations: GLRLM, grey level run length matrix from MM (Modulation Map); MCS, modulation complexity score. *Number of arcs: {a} Single arc, {b} Two arcs, ** Treatment Unit: {a}TrueBeam {b}: Halcyon		

The split between plans passing and failing the specific-plan verification is displayed in Figure 5.2 for each dataset. These two plots represent the plans reserved for the training and testing data sets for the modelling, respectively. The uncertainty bars represent the variations between the six datasets (with characteristic composition, *a%/b%*) for each heterogeneity group and for the reference group.

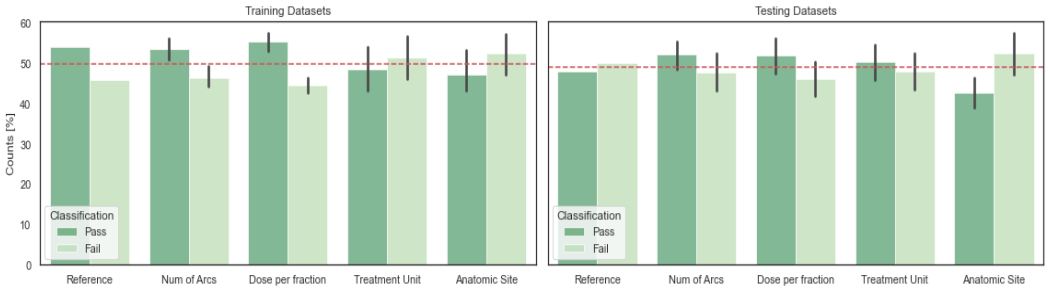


Figure 5.2 Distribution of dataset split between pass and fail plans for each heterogeneity factor, considering the training (left) and testing (right) datasets. The deviation of each column is calculated with the variation of the dataset split for each of the six datasets in every heterogeneity dataset group.

5.2.2 Analysis of modelling

As a result of the grid-search optimization, the hyperparameters for RF were 100 trees and 3 maximum depths of trees; for XG-Boost, the hyperparameters were 170 trees, 2 maximum depths of trees, and a 0.01 learning rate; and for NN, three layers (121, 60, and one neuron for each layer, respectively) and binary-cross entropy loss function were selected.

5.2.2.1 Reference dataset and evaluation of radiomic features

The reference dataset was randomly created with prostate (45.2%), H&N (33.8%), and brain (21%) treatment sites; dose per fraction of 1.8 (8.1%), 2 (57.7%), and 3 Gy (34.2%); plans with single (16.7%), first or clockwise (42.9%), and second or counter clockwise (40.4%) VMAT-arcs; and TB (72.8%) and HL (27.2%) treatment units (Table 5.1). The effect of radiomic features inclusion in the reference dataset were evaluated in all three ML models using the ROC-AUC metric (Table 5.4).

Table 5.4 Mean and standard deviation of AUC values for RF, XG-Boost and NN binary classification using the reference dataset with four sets of predictor features.

Model \ Condition	1	2	3	4
	V/D + C	V/D + C + R1	V/D + C + R1+ R2	V/D + C + R1 + R2 + R3
RF	0.58 ± 0.16	0.67 ± 0.25	0.75 ± 0.14	0.78 ± 0.15
XG-Boost	0.60 ± 0.14	0.59 ± 0.12	0.61 ± 0.16	0.65 ± 0.13
NN	0.82 ± 0.08	0.85 ± 0.05	0.87 ± 0.04	0.87 ± 0.03

Abbreviations: area under the curve, AUC; random forest, RF; neural network, NN; feature predictors based on volume and dosimetric plan parameters, V/D; complexity metrics, C; radiomic features extracted from the leaves-trajectories maps, R1, radiomic features extracted from dose distribution, R2; radiomics features extracted from calculated CDI, R3.

5.2.2.2 Features importance

To evaluate model's most important features, the predictor features were grouped and labelled as Volume/Dose, Complexity, Radiomics1, Radiomics2, and Radiomics3, corresponding to plan parameters, complexity metrics, the MMs' radiomics, the 3D dose distributions' radiomics, and the radiomics from CDI, respectively. The ten most important features distribution for the reference dataset were: one feature from Volume/Dose ('Dose_per_fraction'), two from Complexity ('MUcp' and 'Number_of_arcs'), four from Radiomics1 ('first_order_*x*'), and 3 from Radiomics2 ('GLDM', 'GLRLM', 'first_order_*x*'). Furthermore, the distribution of the 10 most

important features for each heterogeneity factor with their respective datasets is summarized in Figure 5.3.

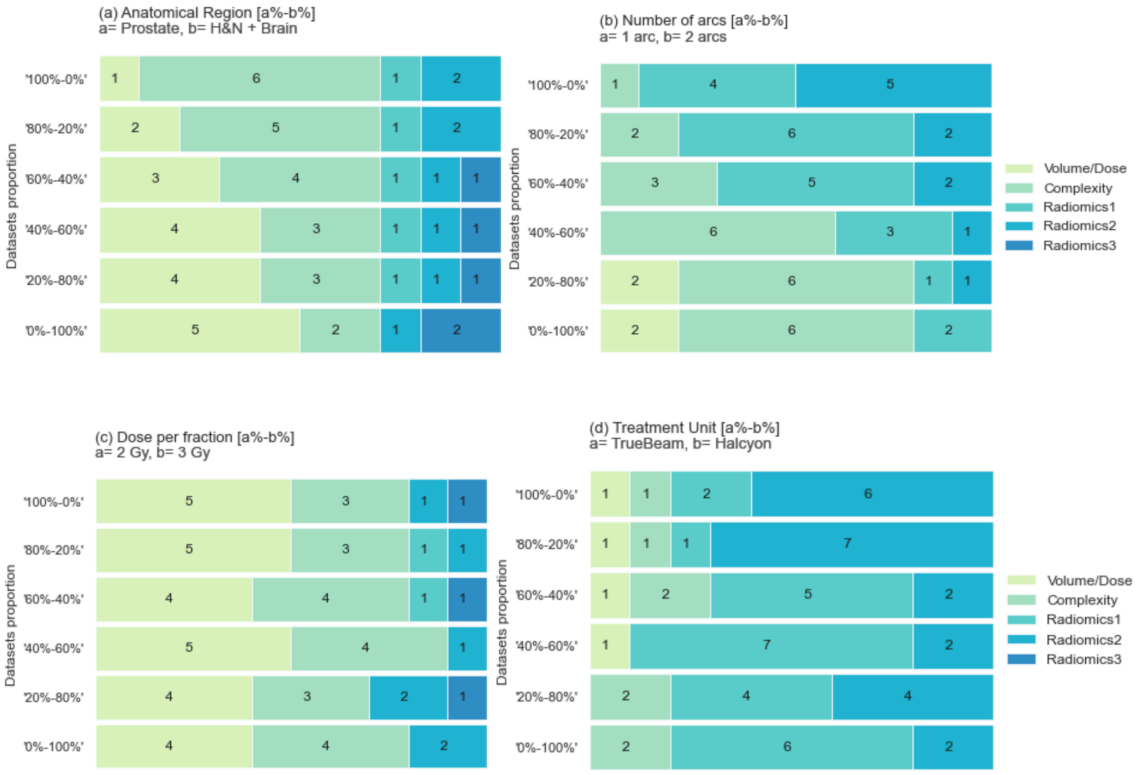


Figure 5.3 Ten most important feature classes distribution for each dataset scenario considering the heterogeneities of (a) anatomic region, (b) number of arcs, (c) dose per fraction, and (d) treatment unit.

5.2.2.3 Model performance

Figure 5.4 shows the plots of ROC-AUC values for RF, XG-Boost, and NN based on the reference and the 24 controlled heterogeneities datasets. The standard deviation is represented by a dotted red line for the reference models, and green shades for the other models. The individual lines correspond to the heterogeneity conditions. The additional data of ROC curves and TPR-FPR values are included in supplementary material 1.3.

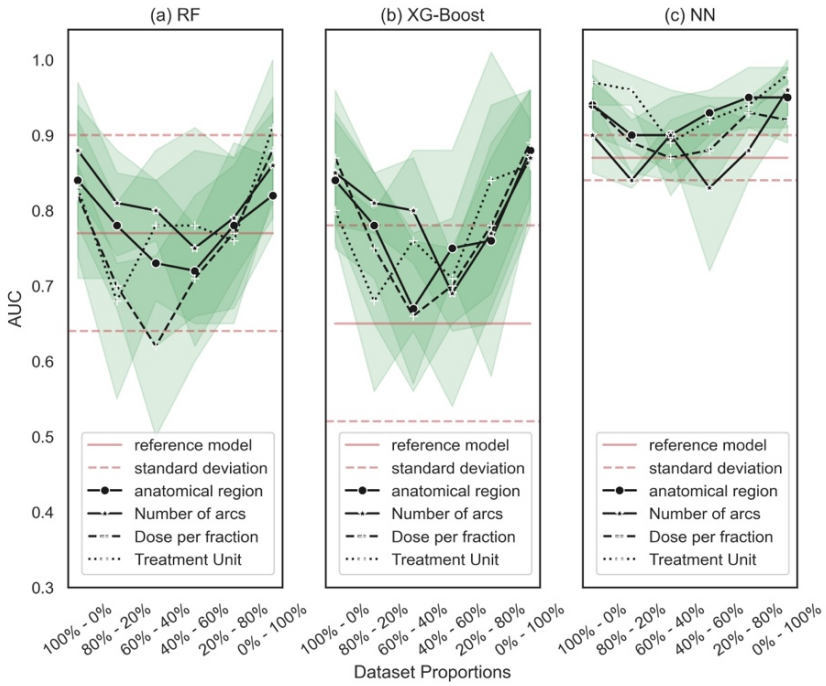


Figure 5.4 AUC results and its standard deviations values for (a) RF, (b) XG-Boost, and (c) NN models, considering the reference dataset, each heterogeneity source, and its different proportions.

5.3 Discussion

Having considered the reported feasibility of GPR predictions in the literature using ML methods [77,80,88,148], this study was designed to investigate the impact of the dataset composition on classification performance from models dedicated to virtual specific-plan verification. Rather than proposing new GPR prediction model in this chapter, it was intended to investigate strategies to increase models' reliability from the dataset quality and their interpretability from a physical perspective. For this reason, it was evaluated the ROC-AUC of three different models (RF, XG-Boost, and NN) by applying them to one reference dataset (constructed from random cases to be unbalanced by design) and 24 datasets with controlled heterogeneities variations (anatomical region, number of arcs, dose per fraction, and treatment units). These heterogeneity factors were implemented based on the demonstrated correlation of GPR values with the PTV size or anatomical region [5,149], the MU or dose delivered by each field (beam or VMAT-arc) [6], and the treatment unit or beam modelling [5,6,99]. Thus, it was hypothesized that an ML model predicting GPR values would not have a reliable performance if the plan to be predicted is underrepresented in the training dataset (*i.e.*, the model dataset would have a few or any plans comparable with the predicted treatment conditions).

The fail/pass plan classification selected in this study was resulted from (I) the low extrapolation capabilities of tree-based models beyond the training dataset's range values [57,146] and (II)

the reported advantages of implementing a decision-making tool within a specific-plan verification program [4,77]. Duly, it was considered this classification approach to be the more reasonable strategy for GPR predictions using RF or XG-Boost models, having the benefit of retrieving the weighted features' importance to understand the main variables involved in the predictions. Consequently, it was implemented the ROC-AUC metric to measure the model performance, ensuring that all datasets had comparable balanced cases of 'pass' and 'fail' plans (Figure 5.2). For this reason, the 2%/ 1mm GPR criteria and a 95% cut-off value (Figure 5.1) was chosen, avoiding any unbalanced effects within the model performance, and excluding this error factor from the results. Otherwise, unlike most reported datasets [77,80,147], it would suggest that the PR-AUC (AUC from the precision-recall curve) might be a more relevant metric (than ROC-AUC) if the datasets are unbalanced with a bias towards passing plans [147]. Finally, based on the differences in the structural basis between RF, XG-Boost, and NN algorithms [1,55], and because in this section it was intended to understand the dataset's effects rather than achieve the optimal classification model, the classification performance comparisons between these models were not intended. Nevertheless, the lower variability and higher reproducibility of NN models were observed (Figure 5.4).

Besides this analysis of dataset composition effects on the model performance, it was also proposed new features to be included based on radiomics. It was included the MM texture analysis, as it represents the whole arc modulation behaviour in each control point [150] and contains the information needed to calculate most of the reported complexity metrics [91,94]. In addition, the conventional modulation complexity metrics have shown to be not always highly correlated to GPR values[94]. Hence, this automatic texture analysis might bring additional information beyond the 'hand-extracted' features by the complexity equations. In the same way, the CDI were included for radiomic features extraction because the 'blended' image represents the final dose distribution that will be compared with the EPID dose measurements for gamma analysis. Thus, shape and texture variations from the calculated blended dose image could be considered as an indirect way to measure dose fluence complexity.

Similarly, as reported by Hirashima et al., [77] it was confirmed that radiomic features extracted from 3D dose distributions improved the three models' classification performance, compared to the models using just the plan parameters and complexity metric features (Table 5.4). Furthermore, it was showed that combining all the radiomic features, including the extracted from the MMs and the CDI, might improve the classification performances. However, it was noticed that XG-Boost was less susceptible to improvement until the three radiomic features were used together. Also, the NN model was not sensitive to the CDI radiomic features. These behaviours could be attributed to the boosting method implemented by XG-Boost [56] and the

data generalization power of NNs [60,83]. As it has been described [56,57], the model classification performance might not increase if the newly added features represent redundant information about the dataset system. Nevertheless, further studies must be performed to analyse each radiomic group individually with higher datasets.

This study confirms that modulation complexity metrics were important classification features for datasets having mostly plans with two arcs ({*a%/b%*} = 20%/80%, 0%/100%) and that features based on GLRLM analysis from MMs radiomics (*Radiomics1*) were relevant predictors in datasets having plans with different treatment units ({*a%/b%*} = 60%/40%, 40%/60%, 60%/80%) (Table 5.3). The same behaviours were corroborated by Figure 5.3.a, Figure 5.3.b and Figure 5.3.d, respectively. These dependencies were expected due to the prominent difference in modulation complexity values, between the first and second arc, due to their respective MU range values differences (first arcs have higher MU values and modulation than second arcs in prostate plans).

This study demonstrates the dataset heterogeneities effects in two main related aspects. First, the type of features the model relies on to make its predictions (Figure 5.3), and second, the classification model performance (Figure 5.4). Indeed, the progressive variations in features importance for each heterogeneity factor (Figure 5.3) and dataset variations confirm that the model does not account for the same kind of plan parameters to perform its GPR predictions, except the datasets from the dose per fraction heterogeneity (Figure 3.c). Although further investigations are needed, these results suggest that potential ML model implementations in practice for GPR classification should be considered more stable (in terms of explainable predictors) if the dataset account plans with similar treatment conditions excluding the differences in dose per fraction. Consequently, it is considered that analysing the variation of features' importance might explain the predictors' stability in differentiating and predicting specific GPR values (or categories). Therefore, the presented feature analysis could be considered as a potential strategy to evaluate and control the model stability of long-term or dedicated virtual specific-plan verification programs implemented in radiotherapy facilities after feeding the model dataset with additional inputs.

The present study found that models based on datasets with fewer heterogeneities have higher classification performance (higher AUC values) and lower variability than models with highly heterogeneous datasets (Figure 4). This trend might represent the potential effect of dataset composition on the model efficiency to find solid predictor factors for GPR values. Indeed, models with one predominant treatment condition (or just one), presented higher AUC values compared with those more heterogeneous models ({50%-50%}), which can intricate an

appropriate model optimization. Specifically, a poor model data generalization due to the mentioned heterogeneous datasets could promote model predictions less related to physical treatment characteristics and dose deliverability parameters, increasing random feature associations that might fit or describe the training dataset, but it might not have an optimal performance predicting new underrepresented plans. Otherwise, well-classified plans from these kinds of unbalanced heterogeneous models could be expected just because of random correlations instead of dosimetric or physical properties that reflect the actual plan deliverability, reducing the model reliability.

Considering the RF and XG-Boost results from Figure 5.4, it is also important to note that the heterogeneity factor with a lower impact on prediction performance was the treatment unit, which suggests that the data generalization and model training are more affected by plan parameters such as PTV volume (anatomic region), number of arcs, or dose per fraction. These results might be considered in the dataset design when there is a limited number of treatment plans. However, these variations between treatment conditions and their implicit effects on modulation complexity scores, and consequently, on dose deliverability, ease the GPR models *based on numeric features* to rely on the differences between the dataset plan categories rather than individual patient-specific properties, which compromise the prediction reliability and further analysis. For this reason, models using CNN architectures (automatic extraction features) based on high dimensional plan parameters might be more reliable in including specific physical aspects of the treatment (Chapter 6, Chapter 7).

This work is among the first to analyse the implications of the dataset heterogeneities in the model performance considering its potential applications for virtual plan verification protocols. However, there are some limitations in this study. First, the results obtained in this study were based in plans from one institution, and further studies are required using hybrid datasets from other radiotherapy centres to achieve broad conclusions. Additionally, it is important to highlight the potential benefits of larger datasets in further investigations in terms of reproducibility and feature predictor definitions.

Finally, confidence in any virtual specific-plan evaluation requires a deep understanding of its results and the selection process of the features used in these predictions, rather than just demonstrating a high accuracy model without practical interpretability and using datasets that do not represent the treatment conditions related to the GPR value intended to predict. Based on these results, it is recommend performing GPR predictions based on datasets with similar treatment characteristics. These considerations suggest a better understanding of the features needed in the prediction process and their physical impact.

5.4 Conclusion

Evaluating ML-based models applied to virtual specific-plan verification needs to consider strategies to measure their prediction reliability and interpretability to ease their implementation in practice. Therefore, assessing the impact of the dataset components on the model data-generalization (given by the treatment plan characteristics in the dataset) must be an essential strategy to understand the physical aspects involved in the prediction process. Additionally, radiomic features from MM, dose distribution, and CDI were associated with improvements in model prediction performance. Finally, the plan parameter *treatment unit* represented the heterogeneity factor with less adverse impact on model performance.

Contrastingly, with these results, the most relevant question now relies on the actual utility of the 'more important features' used by the model to predict certain GPR values. Once the model assembling was analysed, the reliability of their predictions is still questionable because their predictor features cannot retrieve specific plan parameters associated with dose deliverability. Thus, in Chapter 6, high-dimensional features associated to dynamic treatment unit conditions (MM and MUcp profile) were explored to predict GPR, attempting to locate specific activated features that might improve the model reliability (Chapter 7).

Chapter 6 GPR modelling

As more widely mentioned in Chapter 5, ML models dedicated to QA have often been implemented to predict GPR values [2,29], exploring calculated modulation complexity metrics and implementing models such as Poisson regression [76], decision trees [77], support vector machine (SVM) [63], and artificial neural networks (ANN). Also, CNN-based models [78–81] have been reported using dose distributions or static beam fluence maps [2]. However, neither the empirical complexity metrics nor the dose distributions directly account for the modulation complexity. For this reason, it is necessary to contemplate a more comprehensive prediction method considering high dimensional information, such as the MM and the MUcp profile variations (Section 4.4) as potential GPR predictors, implementing automatic-feature extraction methods, and avoiding the use of conventional complexity formulas [74,93,97] that might limit the amount of information extracted.

Considering the above mentioned, this chapter 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 design 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 profiles as input features for GPR predictions. Consequently, since the GPR values were calculated using EPID measurements, it was decided to include the calculated CDI as a third evaluated input feature (i.e., dosimetric input feature). The second objective was to verify whether concatenated models presented an improved GPR prediction performance or not. Furthermore, it was aimed to evaluate the model stability in terms of the quality of the learned features extracted by each model.

In this chapter a workflow followed to develop this study, a descriptions of the dataset assembled based on the recommendations and conclussions from Chapter 5, the designed CNN-models' architectures, and the predicting performance evaluation of those models. Finally, a discussion and conclusions about the present findings implications in RT virtual QA is presented.

6.1 Method

6.1.1 Workflow

The four-step workflow followed in this chapter is illustrated in Figure 6.1. (I) From 1024 DICOM-RT files, the MM, MUcp, 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 tuning, and stability tests were performed with TensorFlow [105]. (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) Finally, 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.

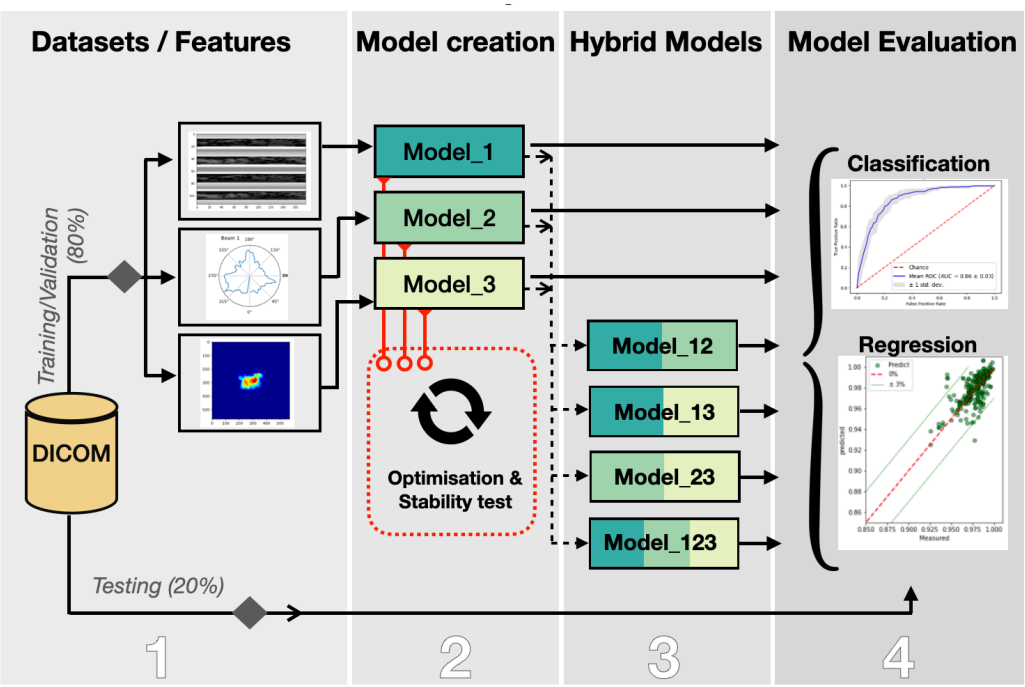


Figure 6.1 Workflow of the present study, including the (1) dataset creation, (2) the corresponding designed main models (M_1, M_2, and M_3) plus their optimization and stability evaluation, (3) the design of the assembled hybrid models, and (4) the prediction performance evaluation, for the training and testing sub-datasets.

6.1.2 Dataset

A total of 1024 anonymized DICOM-RT files from 746 prostate plans were retrieved to extract the MM, the MUcp profiles, and the CDI features by Python scripting [98]. 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. Since the dataset was limited, it was decided to create a dataset with just variations of the treatment unit based on Chapter 5 findings. 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 and 50 mm/s, 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 6.2. The treatment plan conditions are summarised in Table 6.1.

Table 6.1 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.

The GPRs were calculated from gamma analysis evaluation [29] based on EPID measurements and a global 2% dose and 1 mm distance differences criteria (2%/1 mm) because of the same rationale explained in Section 5.2.1. For classification models, the VMAT dose distributions with a GPR ≥ 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 [4]. However, this value also promoted the best-balanced conditions in GPR terms when the datasets were divided into sub-datasets (Figure 6.2.b), avoiding unreliable classification modelling and overfitting effects [151]. As it is registered in the Supplementary Material 2.1, most measured plans evaluated with 3%/3mm, 3%/2 mm, 2%/ 3mm, and 2%/2 mm criteria presented GPR values of 100%, generating highly unbalanced datasets.

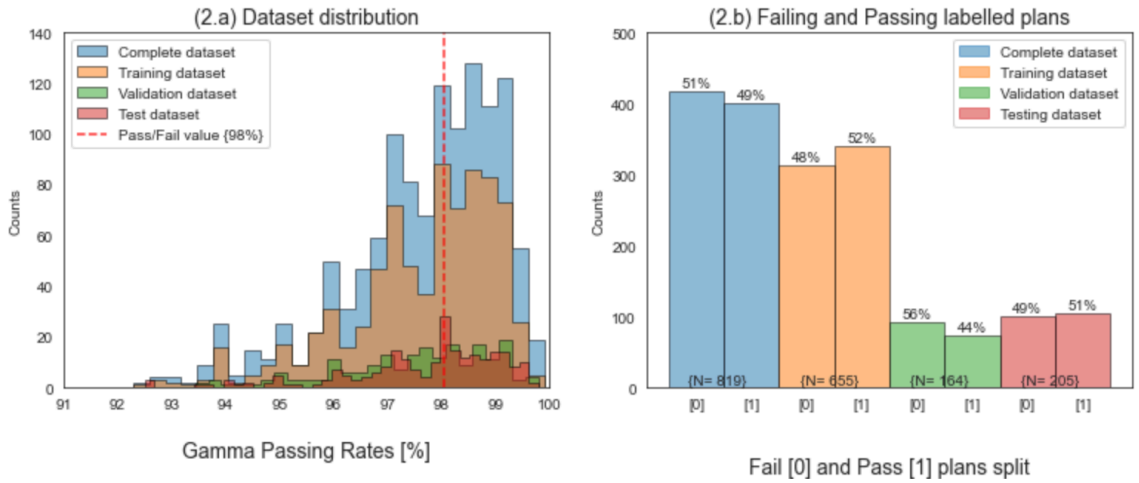


Figure 6.2 Distribution of the number of plans (counts) for (a) all GPR values with the GPR criteria of 98% (dotted red line), and (b) the representation of all sub-dataset splits. Plans labelled as 'fail' were represented with [0] and plans labelled as 'pass' were represented with [1].

6.1.3 Input features

The input features used in this chapter were described in Section 4.4. They are MM, MUcp profile, and CDI (Figure 6.3).

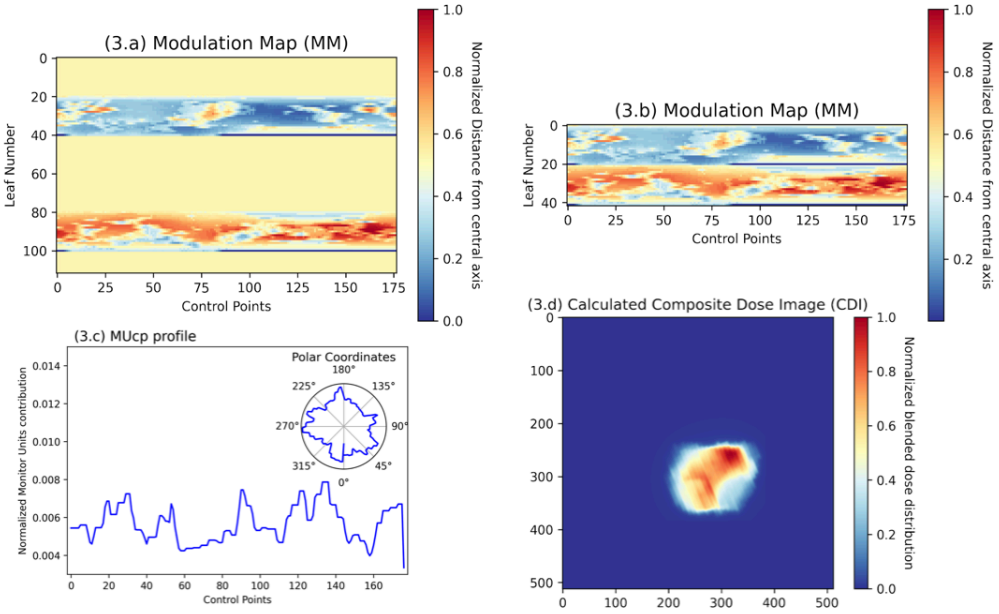


Figure 6.3 Representation of the three features used in this study. (a) The full modulation map (MM) and (b) the edited MM removing the static leaves. (c) The monitor units per control point (MUcp) profile and its representation in polar coordinates. (d) Composite dose image (CDI) calculated by the portal dosimetry tools in the treatment planning system.

6.1.4 Models

The designed models for MM, MUcp profile, 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 (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 [61,151,152].

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.

6.1.5 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 $r<0.4$, $0.4\leq r\leq 0.7$, 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 3.1)

6.2 Results

6.2.1 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 6.4 and a detailed description is available in the Supplementary Material 2.2 [105].

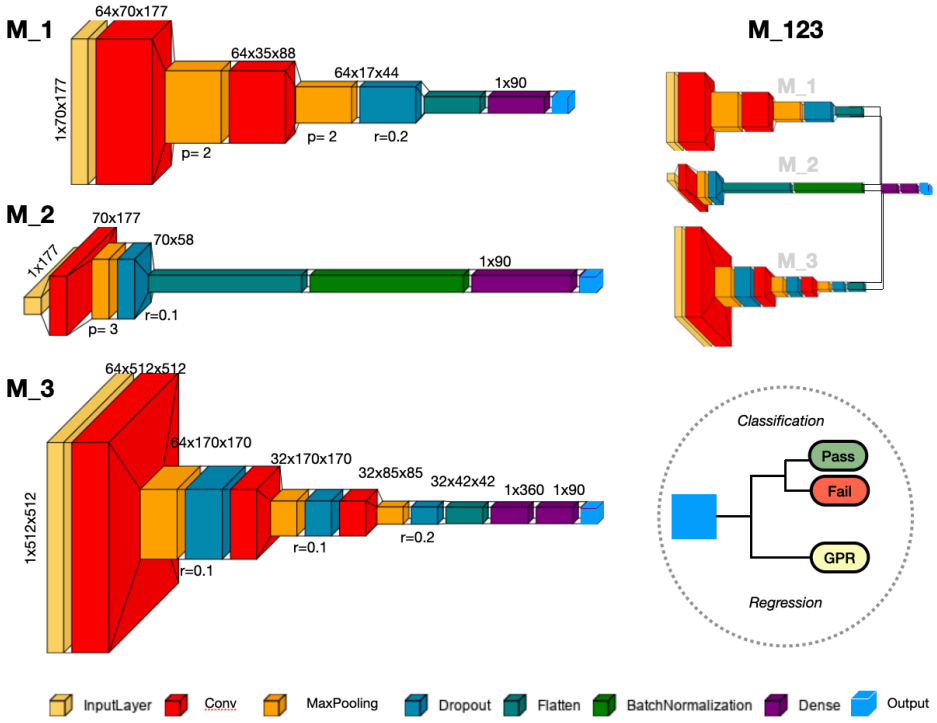


Figure 6.4 Convolutional Neural network architectures corresponding to the models M_1, M_2, M_3, and M_123. The output is also represented as a dual output for classification (*pass-fail*) and a single output for regression.

6.2.2 Architecture stability

The results for the model stability test are represented in Figure 6.5. 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.

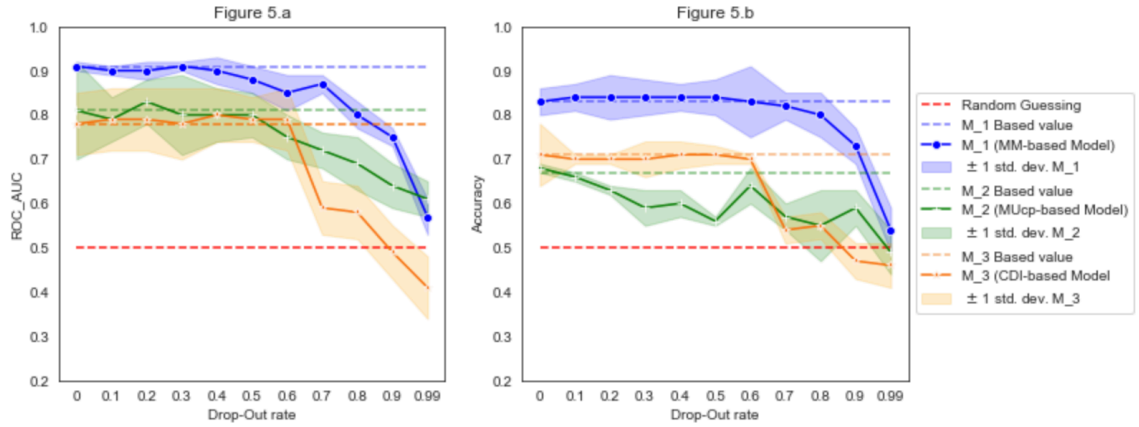


Figure 6.5 Model stability test of ROC_AUC and accuracy for models M_1c, M_2c, and M_3c.

6.2.3 Modelling performance

The modelling classification and regression performances for all models were summarised in Table 6.2, and in Figure 6.6 and Figure 6.7.

Table 6.2 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 spear corr.	*Val.*	0.62	0.46	0.65	0.66	0.53	0.58	0.68
		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.

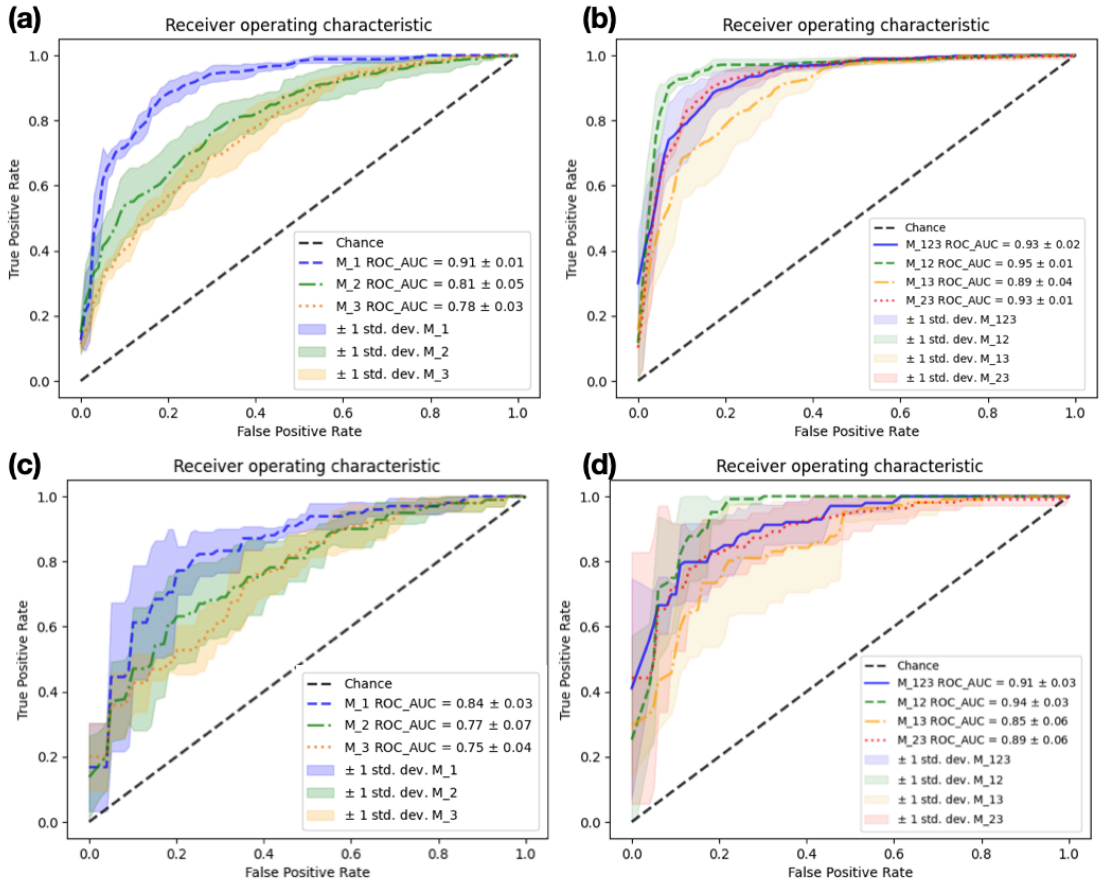


Figure 6.6 ROC plots and ROC_AUC values of the main models (M_1c, M_2c, and M_3), and the hybrid models (M_12c, M_13c, M_23c, M_123c) for validation (Fig. 6.a, Fig. 6.b) and training sub-datasets (Fig. 6.c, Fig. 6.d).

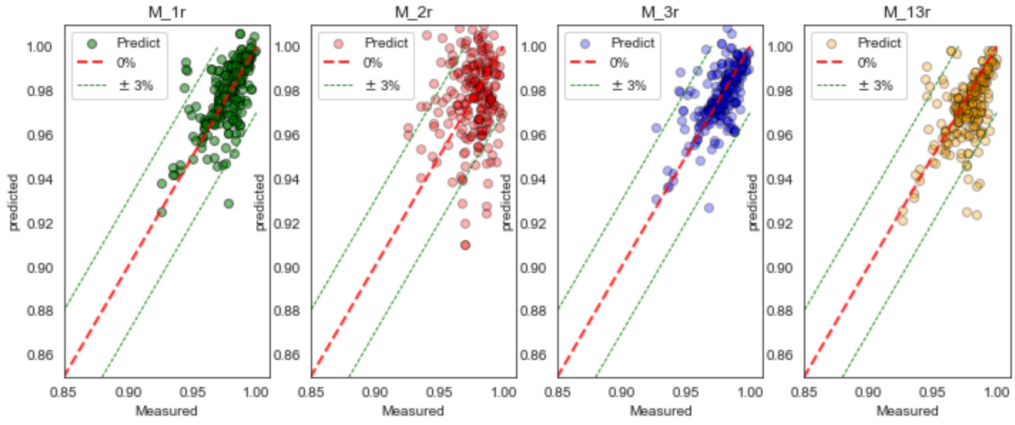


Figure 6.7 Regression results for the models M_1r, M_2r, M _3r, and M_13r with a 3% deviation (dotted green lines) from the ideal GPR distribution represented by the red line.

6.3 Discussion

This chapter investigated the suitability of MM, MUcp profiles, and CDI for GPR predictions implementing ML models. These three input features were used to explore new treatment-plan

information apart from the already studied dose distributions and reported complexity metrics [74,93,94,97]. Indeed, the MM and MUcp profiles can be considered high-dimensional modulation complexity features directly related to the treatment unit performance, which correlates to the dose deliverability [91,92,94]. Hence, it was intended to predict GPRs based on practical physical aspects of the treatment delivery, avoiding calculating limited complexity metrics from empirical equations. Furthermore, the CDI was evaluated 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 [89,120,153]. In addition to the exploratory study, it evaluated and confirmed the potential benefit of including more than one treatment feature within the GPR prediction process (Figure 6.6). Indeed, 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 profiles, and CDI features in GPR predictions, (2) verify if concatenated models increase the GPR prediction performance, and (3) assess the quality of the learned features extracted by each model in GPR predictions.

This study is the first reported evaluation of the MM, MUcp profiles, and CDI as potential GPR predictors using ML methods [2,3]. Previous works have implemented regression models based on modulation complexity metrics and dosimetric parameters, reporting mean prediction errors between 2.2% and 4.5% [88,89,104,152]. Similarly, MAE values between 0.74-4.2, 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 [78]; one CNN-based hybrid model based on planar (sagittal) dose images, volumes data, and MU values [81]; one gradient-boosting model based on radiomic features, clinical parameters, and modulation complexity metrics [77]; and one support vector machine based on complexity metrics and plan parameters [87]. Likewise, using similar input features, reported classification models presented ROC_AUC values between 0.7-0.88 [77,138]. In contrast, this study's MAE, RMSE, r, and ROC_AUC values presented comparable results for all models (Table 6.2), 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 [2,3], 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 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 [77,81] 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[82,154]. Furthermore, although the benefits of concatenating various multi-scale features have been reported, even in radiotherapy [28,31], concatenating too many features might compromise the model's performance and the training model[154]. However, using concatenated models and controlling the different types of inputs might represent a technical advantage in mitigating premature or suboptimal gradient optimization[81], 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 [87,99,139]. 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[151,155], heading suboptimal predictions. Therefore, it is deemed 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 6.1) in accordance to Chapter 5, 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, it was 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, it was only explored 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 (as it is described in Chapter 7), tools like activation maps [155] 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, it is 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, it is 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.

The GPR evaluation is widely used as a deliverability metric and is one of the worldwide standard tests for specific treatment verification [4]. However, it has been thoroughly questioned because of its arguable sensitivity to reflect or discriminate plan errors with potential clinical implications[5]. Nevertheless, this study, rather than predicting just one metric, shows the promising opportunity to explore more treatment-associated parameters that can be part of an integral evaluation method of dose deliverability evaluation. It is considered 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, it is noted 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, as it is discussed in Chapter 7.

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 [79,81]. 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, it is acknowledged that the extracted datasets were based on treatment plan information from one institution, and external verifications will be necessary to perform further validations. Finally, it is also acknowledged 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 [1,51,83]. It might be considered 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, this research was 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 traceable physical factors. 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.

6.4 Conclusions

The MP, MUcp profiles, 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, as it is explored in Chapter 7.

Chapter 7 Decision Suppor applications

Several contributions of ML models applied to patient-specific QA have been discussed previously (Chapter 5 and Chapter 6), highlighting their *acceptable* performance in predicting GPRs within 3% error (accuracy, MAE or RMSE) [2]. Among these reported models, Poisson regression [76], AdaBoost [89], and Random Forest [89] presented a 3% error, and DNN [156], CNN [81], and ANN [80] models reported up to 1.8%, 1.1%, and <1% errors, respectively. However, as mentioned before (Section 5.3), all the datasets used for model training were unbalanced in terms of GPR values, and they were created heterogeneously with plans from different anatomic regions treated with different energies and plan modalities (VMAT or IMRT). These factors are the leading causes of model overfitting and lack of model generalizability, which might lead to incorrect predictions due to the known dependence of modulation complexity [94], anatomic region [138], and beam energy [88] on GPRs.

From the model interpretability point of view, the reported CNN-based models dedicated to GPR predictions [79,81] do not offer straightforward ways to retrieve or identify the features associated with the predictions [65,157], limiting the understanding and evaluation of the model quality because they were developed using dose distribution regions as predictors [2]. These inputs do not provide enough explanatory parameters for plan deliverability analysis; hence, DL 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, CP). In general terms, the activation map of one input image is generated by applying the model filters from one layer to the original input, identifying the regions or features considered to compute the resulting prediction. As it was defined initially by Bolei et al. [158], this activated filter map can be extracted using a global average pooling (GAP) layer or a global max pooling (GMP) layer method, which respectively, computes all the different input regions activated by the filters or calculates one single discriminative region maximizing all these activated regions.

Besides the previously considered model's performance aspects (Chapter 5), it is noticeable that their actual application in practice has been a poorly studied topic, generating some technical gaps about their relevance within a QA protocol in RT departments. Indeed, gaps regarding indicators of model reliability to implement these algorithms in practice are poorly defined. Accordingly, the reported reviews of ML contributions in patient-specific QA protocols for RT [2,157] only agreed on the need to establish precise ML support-decision tools based on trained

models with at least more extended datasets (according to Valdes et al. [88]) and dedicated-anatomy conditions. In the same way, Kalet et al. [159,160] briefly mentioned the need to consider the data quality, adaptability, and limitations of each model applied in RT. Nevertheless, specific parameters to contemplate "before, during, and after" ML model implementations for virtual patient QA verification were not widely discussed. For these reasons, the designed models in Chapter 6 were considered to design a potential QA protocol dedicated to prostate treatment verification, extracting the activation maps of MM and MUcp profiles to retrieve and understand the main features or regions of interest considered by the models to perform their GPR prediction. The designed workflow using the activated features are considered to open a discussion about the utility of ML model applications and their stability in evaluating dose deliverability.

Finally, it is paramount to mention that all previously reported ML models were trained with datasets having treatment information from plans created in independent institutions following their planning protocol and technology availability. Thus, the models' capabilities to transfer the 'learned features' to new scenarios have not been evaluated yet, which might be the final verification of the model's utility. For this reason, the models developed in Chapter 6 were applied to predict GPR values to one dataset with plans designed and delivered in an external institution with similar dose prescription and technology conditions to those that were contemplated in the original training, testing, and validation datasets.

This chapter gives the activation maps from models M_1, M_2, and M_3 from Chapter 6. Next, it is proposed a designed treatment verification QA workflow to assist the dose deliverability evaluation of RT prostate plans, including the virtual patient-specific treatment verification. In this workflow, the dose deliverability analysis will be supported by the activation maps of MM and MUcp profiles to identify potential error causes. Additionally, the models' generalization is analysed using an external dataset and their respective activation maps.

7.1 Methods

7.1.1 Activation maps and workflow design

The activation maps of six plans from the testing datasets used in Chapter 6 were generated using the GAP method to demonstrate the potential applicability of virtual plan verification identifying regions of interest linked to the predictions. Three cases were randomly selected from the correctly classified plans labelled as 'Pass', and three plans correctly labelled as 'fail'. The regions were associated to plan conditions that might be considered to change in future re-planning cases. Furthermore, one workflow for dose deliverability evaluation protocol was designed by incorporating a virtual patient-specific plan verification section, which might be able

to retrieve specific plan parameters or physical aspects associated with dose deliverability that could be included in new plans or re-planning scenarios.

7.1.2 Model validation with external dataset

To externally validate the models M_1, M_2, and M_3, and the hybrid models M_12, M_13, M_23, and M_123 (Chapter 6), the information from 32 anonymized prostate plans from an external institution[8] was extracted. The treatment plan conditions were the same as the TrueBeam parameters specified in Section 6.1,with the same detector model and calibration conditions. The prediction performance was evaluated with the same metrics used in Section 6.1.5. Finally, the activation maps were generated for 5 external plans to verify the pertinence of the activated features.

7.2 Results

7.2.1 Activation maps and workflow design

The activation maps from 'failing' plans (Plan_181, Plan_200, and Plan_197) are displayed in Figure 7.1, Figure 7.2, and Figure 7.3, respectively. Contrastingly, the activation maps from the 'passing' plans (Plan_3, Plan_119, and Plan_2) are displayed in Figure 7.4, Figure 7.5, and Figure 7.6, respectively. These figures show a distinctive and consistent difference for activation maps in MM. The failing plans presented specific activated leaf movement regions, which might be associated to challenging leaf trajectories (further studies and robust research implementing dosimetric tests are needed to confirm that these group of leaf movements are associated to lower dose deliverability conditions). On the contrary, In the case of passing plans, longer and narrow activated regions corresponding to static leaf trajectories were identified. Furthermore, for MUcp profiles and CDIs, the activation maps do not provide differentiated regions between passing and failing plans.

[8] Centro de Control de Cancer, Bogotá-Colombia

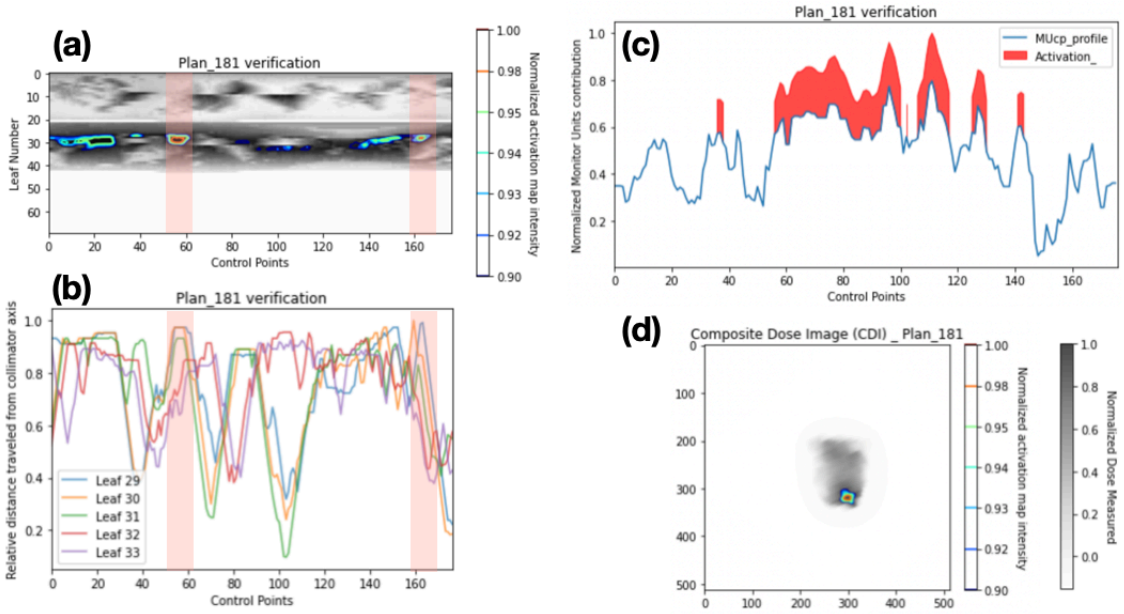


Figure 7.1 The activation maps of model M_1, M_2, and M_3 applied to features extracted from the 'failing' plan Plan_181. (a) Activation map from model M_1 applied to the modulation map. (b) Leaf trajectories corresponding to the activated regions, highlighting in red the control points. (c) Activated regions, in red, from model M_2 applied to the respective MUcp profile. (d) Activation map from model M_3 applied to the CDI.

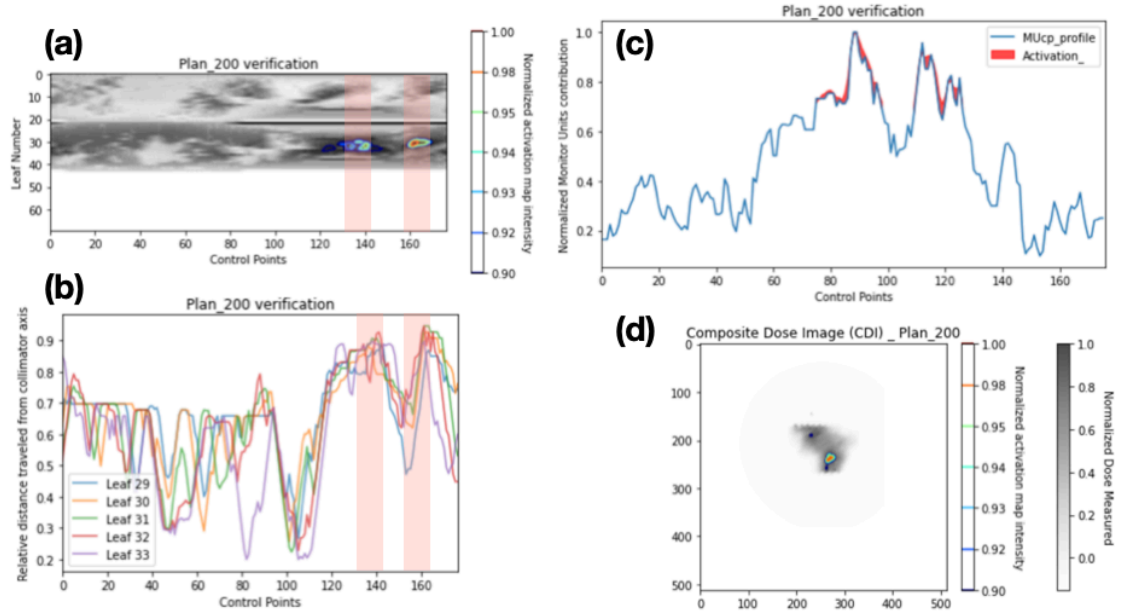


Figure 7.2 The activation maps of model M_1, M_2, and M_3 applied to features extracted from the 'failing' plan Plan_200. (a) Activation map from model M_1 applied to the modulation map. (b) Leaf trajectories corresponding to the activated regions, highlighting in red the control points. (c) Activated regions, in red, from model M_2 applied to the respective MUcp profile. (d) Activation map from model M_3 applied to the CDI.

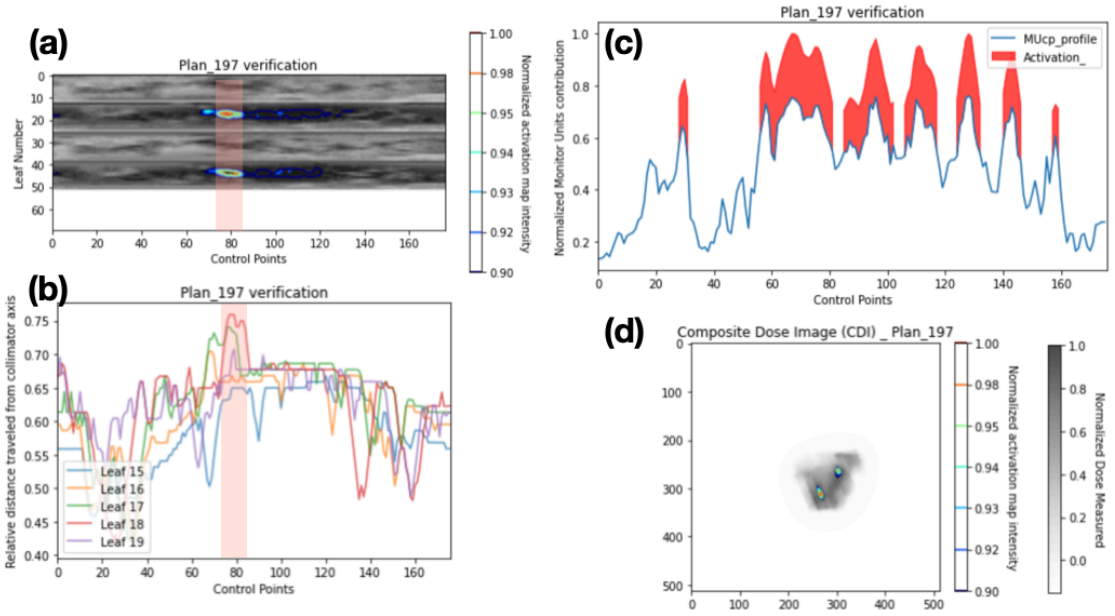


Figure 7.3 The activation maps of model M_1, M_2, and M_3 applied to features extracted from the 'failing' plan Plan_197. (a) Activation map from model M_1 applied to the modulation map. (b) Leaf trajectories corresponding to the activated regions, highlighting in red the control points. (c) Activated regions, in red, from model M_2 applied to the respective MUcp profile. (d) Activation map from model M_3 applied to the CDI.

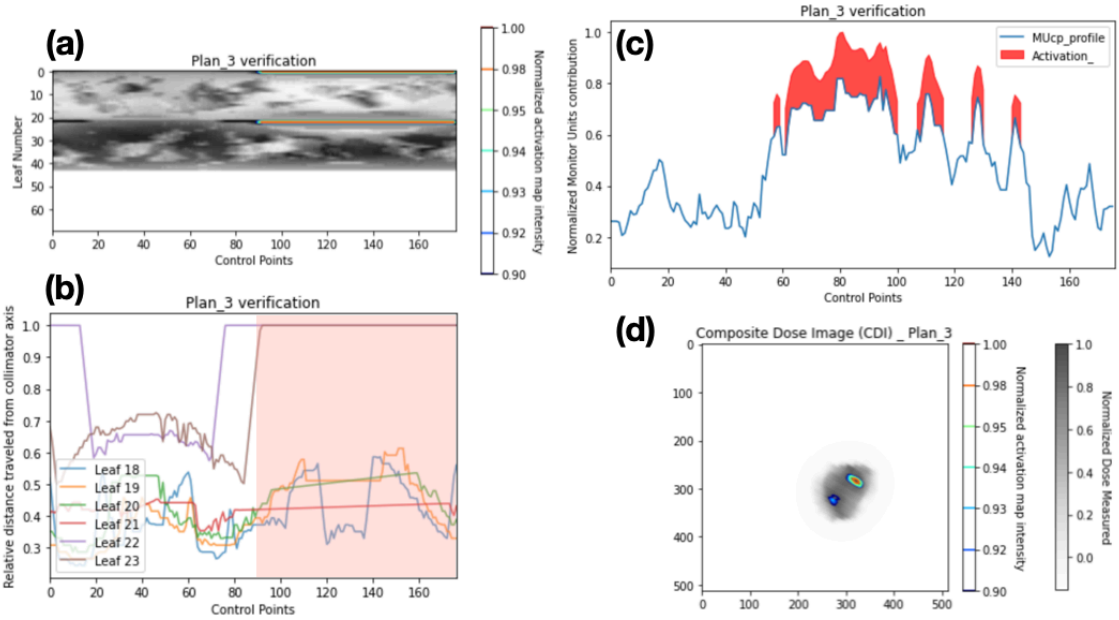


Figure 7.4 The activation maps of model M_1, M_2, and M_3 applied to features extracted from the 'passing' plan Plan_3. (a) Activation map from model M_1 applied to the modulation map. (b) Leaf trajectories corresponding to the activated regions, highlighting in red the control points. (c) Activated regions, in red, from model M_2 applied to the respective MUcp profile. (d) Activation map from model M_3 applied to the CDI.

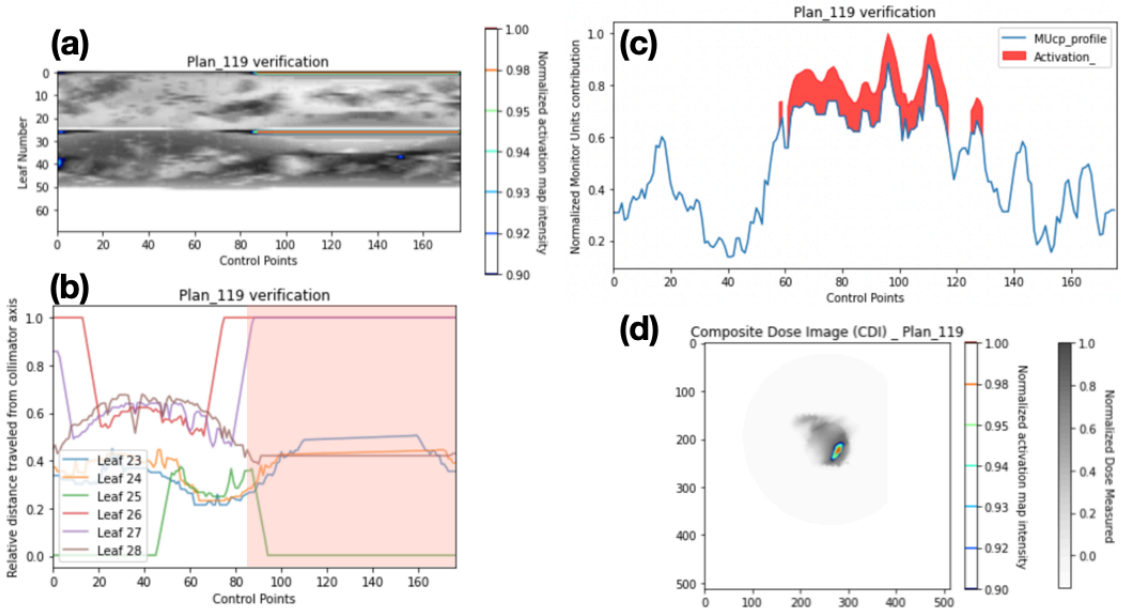


Figure 7.5 The activation maps of model M_1, M_2, and M_3 applied to features extracted from the 'passing' plan Plan_119. (a) Activation map from model M_1 applied to the modulation map. (b) Leaf trajectories corresponding to the activated regions, highlighting in red the control points. (c) Activated regions, in red, from model M_2 applied to the respective MUcp profile. (d) Activation map from model M_3 applied to the CDI.

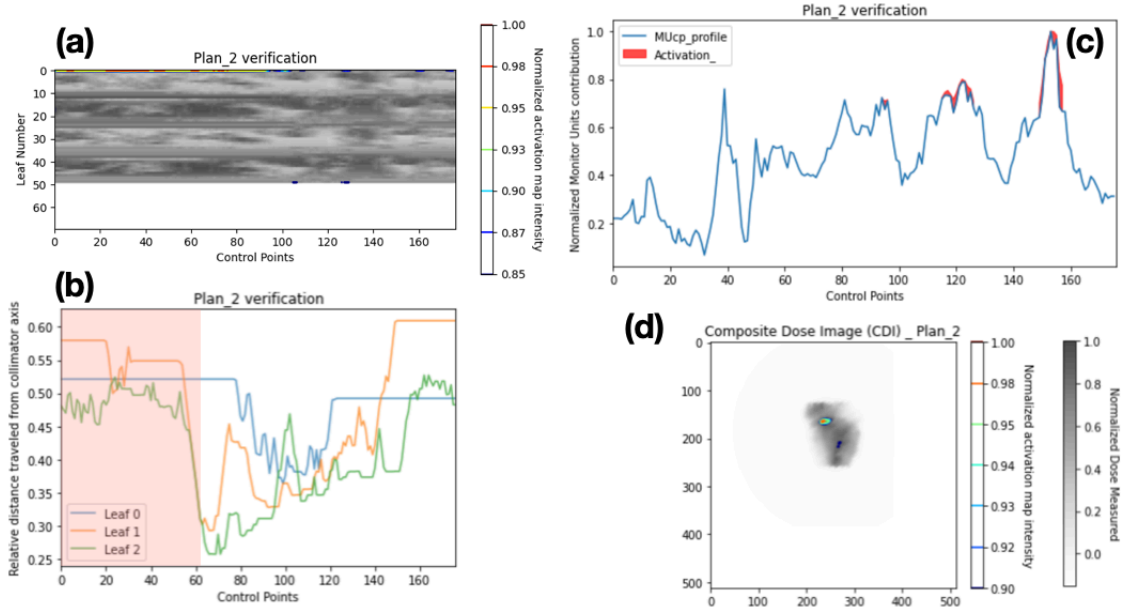


Figure 7.6 The e activation maps of model M_1, M_2, and M_3 applied to features extracted from the 'passing' plan Plan_2. (a) Activation map from model M_1 applied to the modulation map. (b) Leaf trajectories corresponding to the activated regions, highlighting in red the control points. (c) Activated regions, in red, from model M_2 applied to the respective MUcp profile. (d) Activation map from model M_3 applied to the CDI.

The workflow for treatment QA was designed including a virtual patient-specific verification, as it is displayed in Figure 7.7. The present workflow, similar to standard adaptative RT workflows, is a more comprehensive way to use and evaluate ML tools predicting GPRs, since the model has to be robust enough to predict a GPR value and also retrieve potential causes that might explain the prediction. Besides, the information retrieved should be used to improve and reconsider change certain parameters of the actual or further treatment plans. The workflow will be widely commented in Section 7.3.

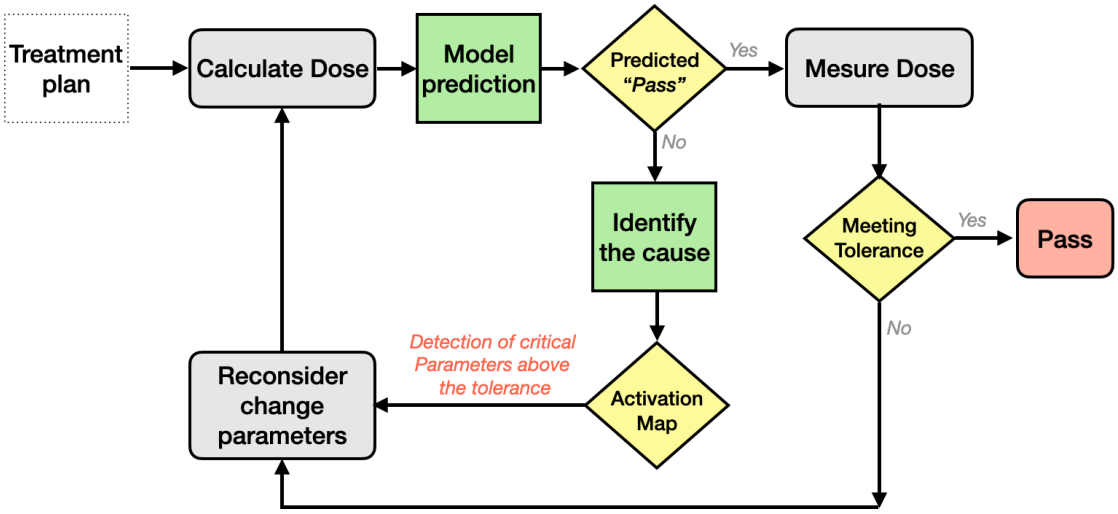


Figure 7.7 Workflow dedicated to patient-specific treatment verification including a section for virtual plan verification with the opportunity to retrieve specific plan parameters associated with the prediction.

7.2.2 Model validation with external dataset

The classification results for the modes' prediction performance using an external dataset are summarised in Table 7.1. Additionally, the ROCs with their respective calculated AUC are displayed in Figure 7.8 for models M_1, M_2, and M_3, and in Figure 7.9 for the hybrid models.

Table 7.1 Prediction performance of all models applied to the external dataset.

	Metric	M_1	M_2	M_3	M_12	M_13	M_23	M_123
Classification	ROC_AUC	0.70	0.49	0.47	0.51	0.50	0.47	0.57
	Accuracy	0.66	0.44	0.53	0.47	0.44	0.63	0.47
Regression	MAE [%]	7.3	8.6	10.6	11.8	13.6	17.6	14.2
	RMSE [%]	7.5	9.2	13.4	13.6	14.3	18.3	14.9
	r spear corr.	0.5	0.2	0.1	0.2	0.1	0.1	0.4

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

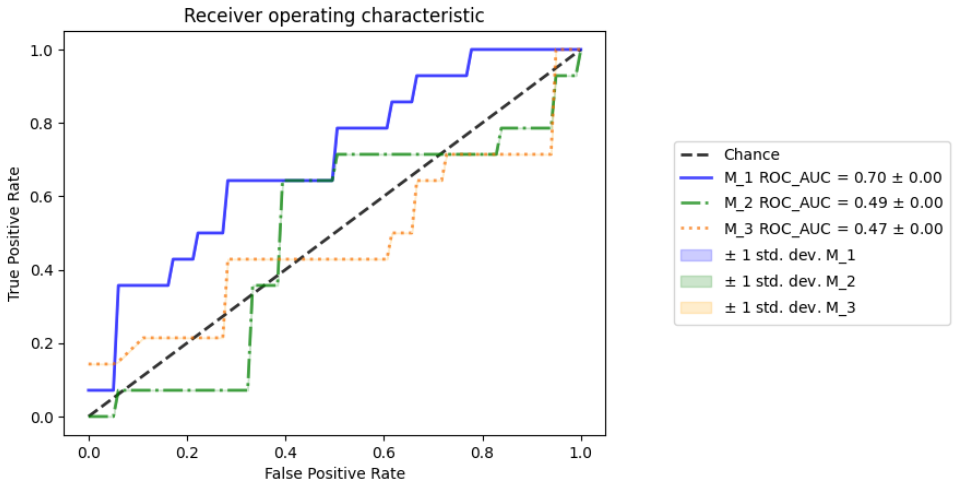


Figure 7.8 ROC plot and ROC_AUC value of the model classification performance for models M_1, M_2, and M3, predicting the 'passing' or 'failing' dose deliverability evaluation based on one external dataset.

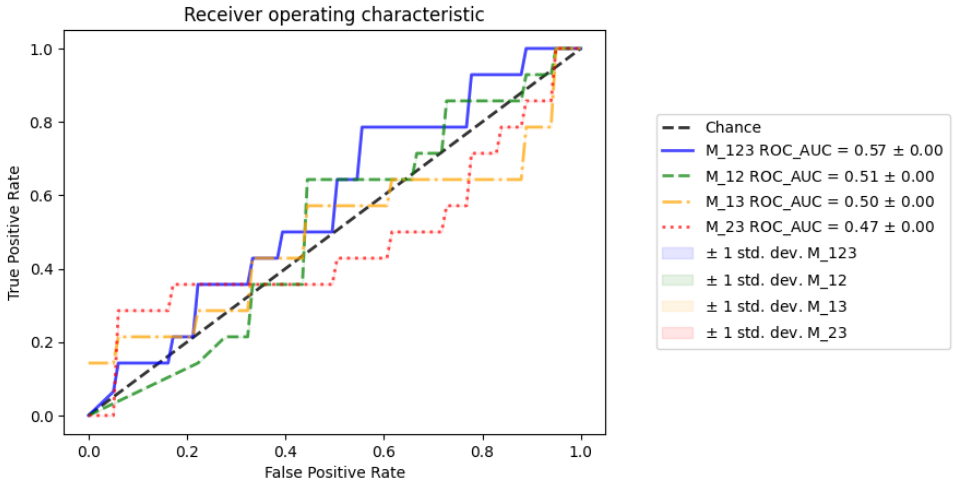


Figure 7.9 ROC plot and ROC_AUC value of the model classification performance for hybrid models M_12, M_13, M_23, and M_123, predicting the 'passing' or 'failing' dose deliverability evaluation based on one external dataset.

Given the low performance of all models except M_1, the activation maps from this model applied to the external dataset are displayed in Figure 7.10 and Figure 7.11 for correctly predicted “failing” and “passing” plans, respectively.

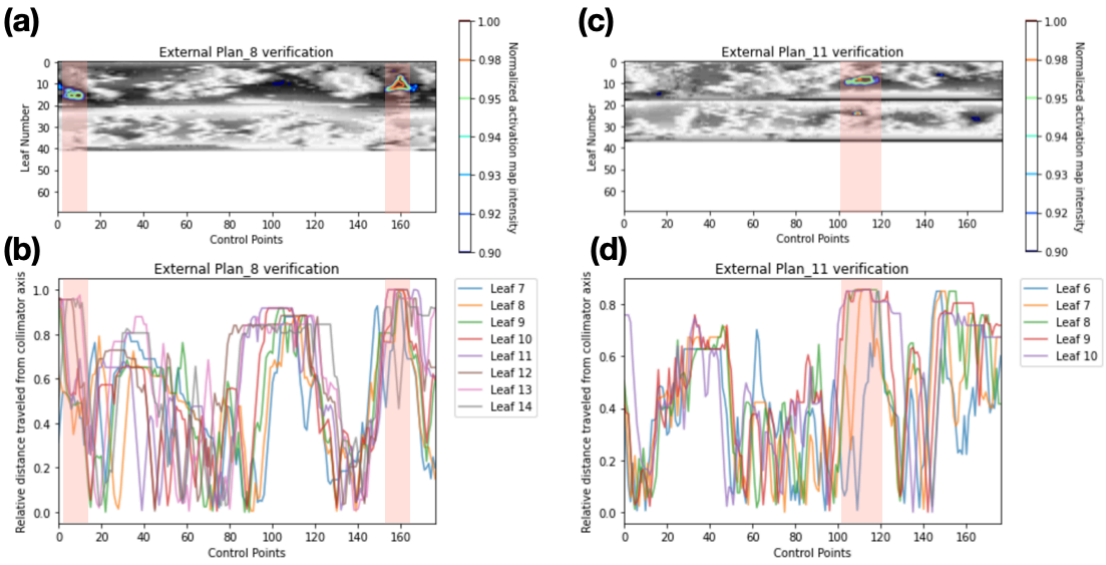


Figure 7.10 The activation maps of model M_1 applied to features extracted from the ‘failing’ plans (a) Plan_8 and (c) Plan_11. Leaf trajectories corresponding to the activated regions, highlighting in red the control points of interest for (b) Plan_8 and (d) Plan_11.

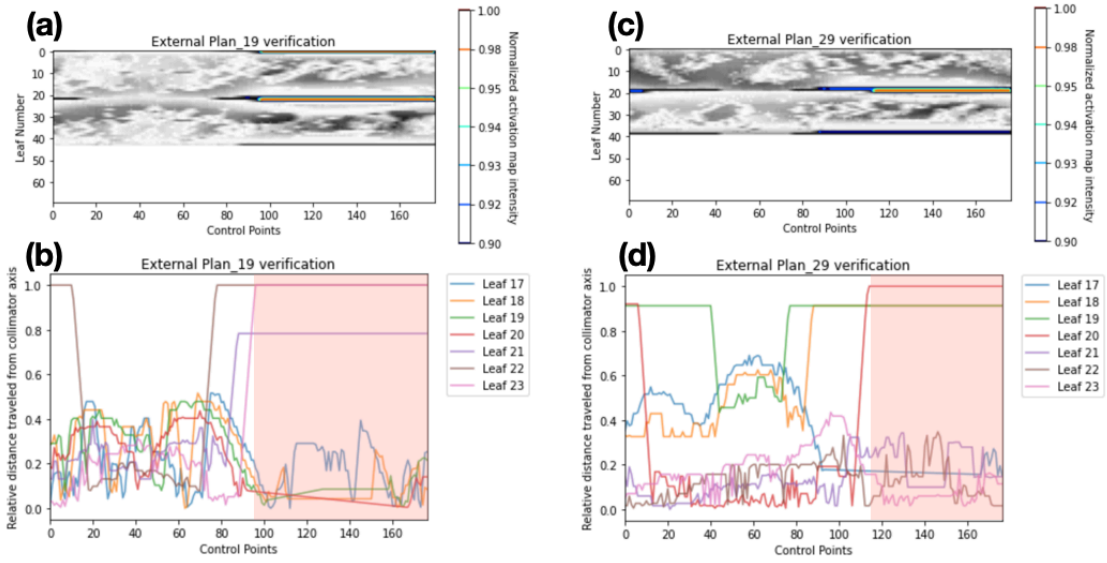


Figure 7.11 The activation maps of model M_1 applied to features extracted from the ‘passing’ plans (a) Plan_19 and (c) Plan_29. Leaf trajectories corresponding to the activated regions, highlighting in red the control points of interest for (b) Plan_19and (d) Plan_29.

7.3 Discussion

To support the implementation of one virtual patient-specific treatment verification protocol using ML tools, the designed models developed in Chapter 6 were implemented in this section opening the discussion about their potential applications, limitations, and additional required developments when it is included in a general treatment QA verification workflow. Furthermore, the same models were tested with an external dataset, evaluating the level of data generalization acquired with the original training dataset. For these reasons, this section will address the main issues regarding the quality of the dataset, the model interpretability, and the model stability.

The proposed workflow in Figure 7.7 starts with the optimized treatment plan and dose calculation, which will be verified to determine if this 'theoretical' dose distribution agrees with the actual dose delivered by the treatment unit. However, since the dose verification procedure is time-consuming because of the detector/phantom setup and the irradiation/evaluation time, an ML model can be implemented to determine if a specific plan has a high probability of 'passing' the evaluation criteria (e.g., gamma passing rate) and proceed to the dose measurements. Otherwise, if the plan prediction is 'fail,' it is recommended not to be irradiated and to analyse the potential causes. Nevertheless, there is no reported any direct way to establish the potential treatment causes of such 'failing' prediction [2,3], and this is one of the main issues that this research addresses, the interpretability of each prediction, which is necessary for medical physicists to understand and possibly correct error causes. Indeed, the previously reported models were trained with empirical functions based on the MLC movements or automatically extracted feature methods from dose fluencies[2], and these predictors do not ease the tracking or identification of real plan parameters involved in the prediction performance.

Accordingly, it is essential to highlight the importance of the present study, using plan features that describe the treatment unit and the dosimetric performance (MM and MUcp profile). Thus, the activation maps can be considered as one essential tool to be included in a virtual verification protocol for two main aspects, (I) it might help to verify if the model is 'learning' meaningful features from plan parameters linked to physical aspects, and (II) it might help finding challenging treatment conditions in new plans and aid re-planning scenarios. The latter is the case of the failing plans from Figure 7.1, Figure 7.2, and Figure 7.3, where it was possible to retrieve the leaf positions (Figures 7.1a, 7.1b, 7.2a, 7.2b, 7.3a, 7.3b) potentially related to a demanding hardware performance and the MU output variations also associated with challenging gantry speed variations (MUcp profiles in Figures 7.1c, 7.2c, 7.3c). These parameters

can be verified with parallel and additional dosimetric and mechanical tests, setting tolerance limits in future designed treatments. Contrastingly, the activation maps from model M_3 (Figures 7.1.d, 7.2.d, 7.3.d) have the same issue as the previously reported ML models in terms of interpretability, and it corresponds to the problematic applicability of these dose region features found by the model in a re-planning scenario or to set tolerance limits based on the mechanical performance tests.

Despite the activation maps between failing and passing plans localized distinctive regions, mainly for MMs, further studies are needed to verify that these changes in MLC position represent demanding hardware scenarios that might compromise the dose deliverability, setting tolerance limits for MLC trajectories or configuring TPS tools associated to the MLC sequencing algorithms [31]. These critical features are the key to proposing a patient verification workflow with the possibility to reconsider treatment parameters and reduce the number of measurements from plans with unacceptable dose deliverability. However, these retrieved features might become obsolete with time if the hardware or plan conditions change since the new planning conditions might no longer represent the conditions that the training dataset represents. Consequently, it is necessary to ensure that the new evaluated plan conditions remain the same as those represented by the training dataset; otherwise, the physical aspects extracted from the model will not represent reliable predictors. Therefore, it is recommended to control at least the anatomic region, beam energy, TPS configuration (optimization and dose calculation algorithms), and hardware (model and performance) as the minimum constant planning parameters.

From the clinical evaluation point of view, it is important to note that the GPR predictions implemented in this study, like the real GPR calculations, are just one metric associated with dose deliverability. Thus, additional tests should be performed to analyse the impact of a specific GPR value (predicted or calculated) on clinical endpoints or changes in the dose coverage or dose sparing. For this reason, it is essential to clarify that these ML models are intended to aid decision-support tools in deciding if it is necessary to recalculate or adjust plan parameters before one treatment is verified by dose measurements, and additional models are needed to predict/evaluate their clinical impact. Moreover, this is one of the main clarifications needed to be considered before implementing these models in practice. However, besides this limitation, the aggregated value of the present ML applications in this matter is the extraction and identification of potential problematic hardware parameters that compromise the dose deliverability.

Considering the results from the model evaluation using an external dataset, the model with better prediction performance was M_1 based on MM inputs, suggesting that the MLC modulation images given by the leaf trajectories might represent more transferable features; still, it needs to be explored in further applications. However, it is noticed that mechanical and dosimetric aspects, such as leaf speed or DLG (dosimetric leaf gap) variations, are needed to be considered following the minimum requirements to maintain homogeneous planning conditions to generate models capable of identifying actual physical parameters linked to appropriate dose deliverability conditions. On the other hand, the other models' performance confirms the need to develop dedicated ML applications (predicting GPRs) trained with the datasets from each institution to learn and represent their own mechanical and dosimetric conditions. Consequently, further studies with more extensive datasets will be required to generate more transferable ML models while keeping the same physical considerations.

7.4 Conclusions

A reliable ML application incorporated within a plan verification QA workflow must retrieve and detect relevant physical aspects linked to dose deliverability to understand and control the pertinence of their predictions. Additionally, this information should be included in re-planning scenarios.

ML models dedicated to GPR predictions are based on specific dosimetric and mechanical conditions; thus, designing transferable or general models to be applied in different institutions might require strictly controlled conditions to maintain their reliability and physical interpretability.

Chapter 8 Conclusions

8.1 Conclusions

The outstanding contribution of the present thesis is the better understanding of the minimum conditions required to propose and implement an ML tool to support virtual patient-specific plan verification protocols in RT. Considering the findings from all the studies followed to address the research questions, it is essential to highlight the relevance of the conclusions obtained chronologically.

First, Chapter 3 opened the discussion about the predictors' quality, how they are calculated, and which metrics represent physical aspects within the treatment plan. Although this section recorded the retrieved and calculated metrics, it also gives the proposed and verified new complexity metrics for Halcyon-v2, considering the potential leaf gaps that the MLC sequencer in the TPS generated during the MLC movements. In this study, there are two main aspects of notice. First, the variations of hardware and planning optimization parameters influence the range of possible modulation complexity scores, even for similar clinical conditions. And secondly, the conventional modulation complexity scores do not provide exact information about the hardware planning parameters linked to a specific predicted GPR value. Therefore, although these complexity scores were dedicated to train ML models predicting GPR values with acceptable results, in this context, the predictors' quality is not optimal and might reduce the model interpretability due to the difficulty of retrieving specific treatment aspects related to the prediction. This analysis, and the consideration of the other predictor metrics and high-dimensional features, promote the discussion regarding which predictor features might be more suitable to design a reliable model and how all these metrics are susceptible to change between different treatment conditions. In this section, and in accordance with the initial research aiming to predict GPR values, it was suggested that ML models should consider the nature of the predictors (which specific treatment unit parameters and clinical conditions represents) to understand better the physical meaning of the predictions.

Consequently, in chapter 4, following the rationale of finding the physical sense of ML model predictions to generate reliable decision support tools in RT, a study was performed to demonstrate how heterogeneous datasets with no balanced representation of defined treatment plan conditions might present suboptimal performance, training models based on random feature associations rather than physical predictors related directly to dose deliverability. This study confirmed that models must be trained by predictor features corresponding to similar treatment conditions to obtain a more robust performance. However, in an indirect way and echoing those mentioned above, it was also noted that the most critical

numeric feature predictors based on modulation complexity scores do not retrieve specific physical conditions that are needed in the verification of potentially challenging treatment unit conditions, which opens the need to use high-dimensional features, as discussed in Chapter 5

Chapter 6 demonstrates and confirms the suitability of high-dimensional features to predict GPRS, especially MM and MUcp profiles are convenient features that retrieve specific treatment plan parameters within a particular treatment time (specific control points) related to the final prediction performance. As well as confirming the benefits of hybrid models, this section promotes the discussion about accurate and more reliable ML models applied to predicting dose deliverability indicators, as discussed in Chapter 7. In this section, a practical application workflow of these models suggested the benefits of retrieving the treatment unit and plan parameters associated with challenging dose deliverability scenarios. Finally, the model verification with an external dataset confirms the technological and clinical configuration's dependence on the treatment plans and their impact on the model generalization. However, this study also found that the MM might be a predictor with potential transferable attributes to be explored in interinstitutional studies.

This thesis using a prostate dataset demonstrates the suitability and potential advantages of virtual patient-specific treatment verification workflow, supported by the activation of detected physical aspects associated with dose deliverability. Summarizing this thesis' findings from a practical application point of view: ML methods trained with high-dimensional features (corresponding to similar treatment approaches) are more convenient to assist the dose deliverability evaluation of treatment plans because their application goes further than just predicting or classifying one GPR value. These applications are essential to open the discussion regarding the main physical causes related to dose deliverability predictions, which might represent a potential benefit in re-planning and adapting treatment plans, or in exploring mechanical-dosimetric tolerance limits to be considered in each institution.

In a more detailed pathway to the routine use of these ML applications, it is necessary to note the need to establish the reference parameters of a specific pathology or treatment modality. Therefore, before creating the dataset to start a customized model training, it is necessary to retrieve and control any change in the TPS version for plan optimization or dose calculation, register any change or upgrade in the hardware parameters, and manage any modifications in the portal dosimetry or radiation dose detector. Once this is controlled and model training is implemented, a constancy test must be incorporated into the whole QA process, determining if the model can still identify or determine the same plan parameters linked to the predicted GPR

value. In the same way, identifying any technical or planning process related to changes in the physical conditions used to train the model is mandatory to ensure the model's interpretability.

Figure 8.1 displays a detailed workflow with the main steps needed to implement an ML method that predicts GPR values in practice. First, deciding which treatment modality and protocol is an initial step to analyse the consistency of the irradiated volume. This step is essential to ensure the homogeneity of the physical parameters linked to the delivered treatments within the training dataset. Second, identify and set the inclusion criteria for those treatments in the training dataset. For instance, plans with prosthetic implants close to the target volume or having high-density changes might not contribute to the model generalization. Considering these previous steps, training and validating the model is the next task to test and evaluate the prediction power of the model. Once the model is evaluated, a pilot test must be implemented with specific plans, verifying its efficacy. After showing acceptable results, it is ready to be implemented in practice. Furthermore, it is necessary to constantly evaluate any change in the critical parameter of the planning process considered in the training dataset to assess the model obsolescence using customized constancy tests based on reference plans.

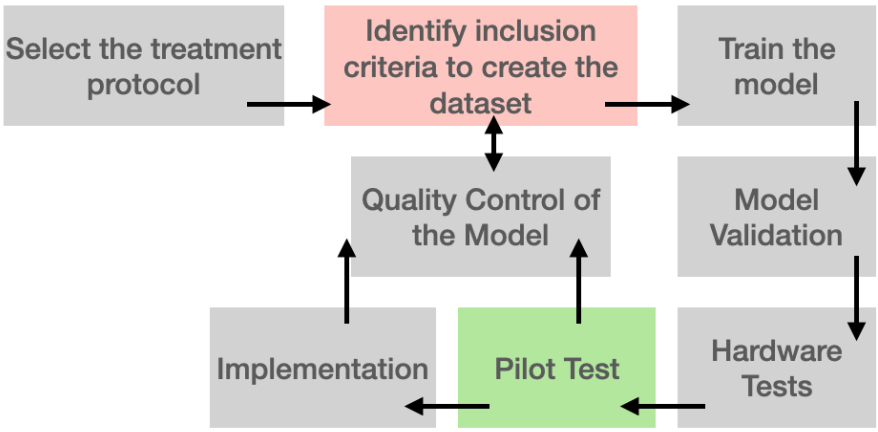


Figure 8.1 Workflow for clinical implementation of ML applications in radiotherapy facilities predicting GPR values

8.2 Future work

The research developed in this thesis opens the discussion regarding reliable and interpretable ML applications in dose deliverability evaluation, assisting virtual patient-specific QA. Therefore, the future work to support the thesis hypothesis is to explore and implement additional high-dimensional input features to keep a critical dose deliverability evaluation. A more comprehensive and integrated evaluation workflow should consider the ML model prediction of hardware position misalignments (e.g., gantry or MLC), the potential clinical differences predicted based on the dose-volume histogram variations, or dose deliverability changes by

anatomic deformations. Indeed, all these considerations might be highly beneficial for onboard plan verification workflows when automatic re-plan protocols are implemented.

References

1. el Naqa I, Das S. The role of machine and deep learning in modern medical physics. Med Phys. John Wiley and Sons Ltd.; 2020. p. e125–6.

2. Osman AFI, Maalej NM. Applications of machine and deep learning to patient-specific IMRT/VMAT quality assurance. J Appl Clin Med Phys. John Wiley & Sons, Ltd; 2021;22:20–36.

3. Chan MF, Witztum A, Valdes G. Integration of AI and Machine Learning in Radiotherapy QA. Front Artif Intell. Frontiers Media S.A.; 2020;3:76.

4. Miften M, Olch A, Mihailidis D, Moran J, Pawlicki T, Molineu A, et al. Tolerance limits and methodologies for IMRT measurement-based verification QA: Recommendations of AAPM Task Group No. 218. Med Phys [Internet]. John Wiley and Sons Ltd.; 2018 [cited 2020 Sep 21];45:e53–83. Available from: https://aapm.onlinelibrary.wiley.com/doi/full/10.1002/mp.12810

5. Hussein M, Clark CH, Nisbet A. Challenges in calculation of the gamma index in radiotherapy – Towards good practice. Physica Medica. Elsevier; 2017;36:1–11.

6. Hussein M, Rowshanfarzad P, Ebert MA, Nisbet A, Clark CH. A comparison of the gamma index analysis in various commercial IMRT/VMAT QA systems. Radiotherapy and Oncology [Internet]. Elsevier; 2013 [cited 2020 Sep 21];109:370–6. Available from: http://www.thegreenjournal.com/article/S0167814013004593/fulltext

7. Hanna TP, Shafiq J, Delaney GP, Vinod SK, Thompson SR, Barton MB. The population benefit of evidence-based radiotherapy: 5-Year local control and overall survival benefits. Radiother Oncol [Internet]. Radiother Oncol; 2018 [cited 2022 Dec 24];126:191–7. Available from: https://pubmed.ncbi.nlm.nih.gov/29229506/

8. Abdel-Wahab M, Zubizarreta E, Polo A, Meghzifene A. Improving Quality and Access to Radiation Therapy—An IAEA Perspective. Semin Radiat Oncol [Internet]. W.B. Saunders; 2017 [cited 2022 Dec 24];27:109–17. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1053429616300601

9. Podgorsak EB. Radiation Physics for Medical Physicists [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2010 [cited 2020 Sep 7]. Available from: http://link.springer.com/10.1007/978-3-642-00875-7

10. Deng J, Pawlicki T, Chen Y, Li J, Jiang SB, Ma C-M. The MLC tongue-and-groove effect on IMRT dose distributions. Phys Med Biol [Internet]. IOP Publishing; 2001 [cited 2019 Nov 22];46:1039–60. Available from: http://stacks.iop.org/0031-9155/46/i=4/a=310?key=crossref.37000f9997827673fa355e28dbca9ce5

11. Arnfield MR, Wu Q, Tong S, Mohan R. Dosimetric validation for multileaf collimator-based intensity-modulated radiotherapy: a review. Medical Dosimetry [Internet]. Pergamon; 2001 [cited 2019 Nov 22];26:179–88. Available from: https://www.sciencedirect.com/science/article/pii/S095839470100058 9?via%3Dihub

12. Arnfield MR, Otto K, Aroumougame VR, Alkins RD. The use of film dosimetry of the penumbra region to improve the accuracy of intensity modulated radiotherapy. Med Phys [Internet]. John Wiley & Sons, Ltd; 2004 [cited 2019 Nov 22];32:12–8. Available from: http://doi.wiley.com/10.1118/1.1829246

13. Kielar KN, Mok E, Hsu A, Wang L, Luxton G. Verification of dosimetric accuracy on the TrueBeam STx: Rounded leaf effect of the high definition MLC. Med Phys [Internet]. 2012 [cited 2019 Nov 22];39:6360–71. Available from: http://doi.wiley.com/10.1118/1.4752444

14. Shende R, Patel G. Validation of Dosimetric Leaf Gap (DLG) prior to its implementation in Treatment Planning System (TPS): TrueBeam™ millennium 120 leaf MLC. Reports of Practical Oncology & Radiotherapy [Internet]. Elsevier; 2017 [cited 2019 Nov 22];22:485–94. Available from: https://www.sciencedirect.com/science/article/pii/S150713671730079 2?via%3Dihub

15. Chae S-M, Lee G, Son S. The effect of multileaf collimator leaf width on the radiosurgery planning for spine lesion treatment in terms of the modulated techniques and target complexity. Radiation Oncology [Internet]. 2014 [cited 2019 Nov 19];9:72. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24606890

16. Tanyi JA, Summers PA, McCracken CL, Chen Y, Ku L-C, Fuss M. Implications of a high-definition multileaf collimator (HD-MLC) on treatment planning techniques for stereotactic body radiation therapy (SBRT): a planning study. Radiation Oncology [Internet]. 2009 [cited 2019 Nov 19];4:22. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19591687

17. Dhabaan A, Elder E, Schreibmann E, Crocker I, Curran WJ, Oyesiku NM, et al. Dosimetric performance of the new high-definition multileaf collimator for intracranial stereotactic radiosurgery. J Appl Clin Med Phys [Internet]. 2010 [cited 2019 Nov 19];11:197–211. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20717077

18. Cozzi L, Fogliata A, Thompson S, Franzese C, Franceschini D, de Rose F, et al. Critical Appraisal of the Treatment Planning Performance of Volumetric Modulated Arc Therapy by Means of a Dual Layer Stacked Multileaf Collimator for Head and Neck, Breast, and Prostate. Technol Cancer Res Treat [Internet]. 2018 [cited 2020 Jan 10];17:1533033818803882. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30295172

19. Lim TY, Dragojevic I, Hoffman D, Flores-Martinez E, Kim G. Characterization of the Halcyon TM multileaf collimator system. J Appl Clin Med Phys [Internet]. John Wiley & Sons, Ltd; 2019 [cited 2019 Nov 19];20:106–14. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/acm2.12568

20. Petroccia HM, Malajovich I, Barsky AR, Ghiam AF, Jones J, Wang C, et al. Spine SBRT With Halcyon Plan Quality, Modulation Complexity, Delivery Accuracy, and Speed. Front Oncol [Internet]. Frontiers Media SA; 2019 [cited 2019 Nov 15];9:319. Available from: http://www.ncbi.nlm.nih.gov/pubmed/31106151

21. Li T, Irmen P, Liu H, Shi W, Alonso-Basanta M, Zou W, et al. Dosimetric Performance and Planning/Delivery Efficiency of a Dual-Layer Stacked and Staggered MLC on Treating Multiple Small Targets: A Planning Study Based on Single-Isocenter Multi-Target Stereotactic Radiosurgery (SRS) to Brain Metastases. Front Oncol [Internet]. Frontiers; 2019 [cited 2019 Nov 21];9:7. Available from: https://www.frontiersin.org/article/10.3389/fonc.2019.00007/full

22. Lloyd SAM, Lim TY, Fave X, Flores-Martinez E, Atwood TF, Moiseenko V. TG-51 reference dosimetry for the Halcyon: A clinical experience. J Appl Clin Med Phys [Internet]. John Wiley & Sons, Ltd; 2018 [cited 2019 Nov 19];19:98–102. Available from: http://doi.wiley.com/10.1002/acm2.12349

23. Gay SS, Netherton TJ, Cardenas CE, Ger RB, Balter PA, Dong L, et al. Dosimetric impact and detectability of multi-leaf collimator positioning errors on Varian Halcyon. J Appl Clin Med Phys [Internet]. John Wiley & Sons, Ltd; 2019 [cited 2019 Nov 15];20:47–55. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/acm2.12677

24. Taylor & Francis. Handbook of radiotherapy physics. 2007.

25. Hunte SO, Clark CH, Zyuzikov N, Nisbet A. Volumetric modulated arc therapy (VMAT): a review of clinical outcomes-what is the clinical evidence for the most effective implementation? Br J Radiol [Internet]. Br J Radiol; 2022 [cited 2022 Dec 24];95. Available from: https://pubmed.ncbi.nlm.nih.gov/35616646/

26. Teoh M, Clark CH, Wood K, Whitaker S, Nisbet A. Volumetric modulated arc therapy: a review of current literature and clinical use in practice. Br J Radiol [Internet]. Br J Radiol; 2011 [cited 2022 Dec 24];84:967–96. Available from: https://pubmed.ncbi.nlm.nih.gov/22011829/

27. Betzel GT, Yi BY, Niu Y, Yu CX. Is RapidArc more susceptible to delivery uncertainties than dynamic IMRT? Med Phys. John Wiley and Sons Ltd; 2012;39:5882–90.

28. Antoine M, Ralite F, Soustiel C, Marsac T, Sargos P, Cugny A, et al. Use of metrics to quantify IMRT and VMAT treatment plan complexity: A systematic review and perspectives. Physica Medica [Internet]. Associazione Italiana di Fisica Medica; 2019 [cited 2020 Jul 29];64:98–108. Available from: https://doi.org/10.1016/j.ejmp.2019.05.024

29. Low DA, Harms WB, Mutic S, Purdy JA. A technique for the quantitative evaluation of dose distributions. Med Phys [Internet]. John Wiley and Sons Ltd; 1998 [cited 2020 Jul 27];25:656–61. Available from: http://doi.wiley.com/10.1118/1.598248

30. Varian Medical Systems. Eclipse Photon and Electron Reference Guide. 2017.

31. Varian Medical Systems. TPS New Features Workbook v15.6. 2018.

32. Otto K. Volumetric modulated arc therapy: IMRT in a single gantry arc. Med Phys [Internet]. John Wiley and Sons Ltd; 2008 [cited 2020 Oct 13];35:310–7. Available from: https://pubmed.ncbi.nlm.nih.gov/18293586/

33. Varian Medical Systems. Eclipse Photon and Electron Reference Guide. 2014;263–348.

34. Liu H, Sintay B, Pearman K, Shang Q, Hayes L, Maurer J, et al. Comparison of the progressive resolution optimizer and photon optimizer in VMAT optimization for stereotactic treatments. J Appl Clin Med Phys [Internet]. John Wiley and Sons Ltd; 2018 [cited 2020 Apr 18];19:155–62. Available from: http://doi.wiley.com/10.1002/acm2.12355

35. Tol JP, Dahele M, Peltola J, Nord J, Slotman BJ, Verbakel WF. Automatic interactive optimization for volumetric modulated arc therapy planning. Radiation Oncology [Internet]. BioMed Central; 2015 [cited 2019 Nov 18];10:75. Available from: https://ro-journal.biomedcentral.com/articles/10.1186/s13014-015-0388-6

36. Shende R, Gupta G, Patel G, Kumar S. Assessment and performance evaluation of photon optimizer (PO) vs. dose volume optimizer (DVO) for IMRT and progressive resolution optimizer (PRO) for RapidArc planning using a virtual phantom. International Journal of Cancer Therapy

and Oncology [Internet]. 2016 [cited 2019 Nov 18];4. Available from: http://www.ijcto.org/index.php/IJCTO/article/view/ijcto.43.7

37. Binny D, Kairn T, Lancaster CM, Trapp J v., Crowe SB. Photon optimizer (PO) vs progressive resolution optimizer (PRO): a conformality- and complexity-based comparison for intensity-modulated arc therapy plans. Medical Dosimetry [Internet]. Pergamon; 2018 [cited 2019 Nov 15];43:267–75. Available from: https://www.sciencedirect.com/science/article/abs/pii/S0958394717301140

38. Sanford L, Pokhrel D. Improving treatment efficiency via photon optimizer (PO) MLC algorithm for synchronous single-isocenter/multiple-lesions VMAT lung SBRT. J Appl Clin Med Phys [Internet]. John Wiley & Sons, Ltd; 2019 [cited 2019 Nov 19];20:201–7. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/acm2.12721

39. Hodapp N. Der ICRU-Report 83: Verordnung, Dokumentation und Kommunikation der fluenzmodulierten Photonenstrahlentherapie (IMRT). Strahlentherapie und Onkologie [Internet]. Urban and Vogel; 2012 [cited 2019 Nov 20];188:97–100. Available from: http://link.springer.com/10.1007/s00066-011-0015-x

40. Grégoire V, Mackie TR. State of the art on dose prescription, reporting and recording in Intensity-Modulated Radiation Therapy (ICRU report No. 83). Cancer Radiother [Internet]. 2011 [cited 2020 Mar 29];15:555–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21802333

41. Brock KK. Adaptive Radiotherapy: Moving Into the Future. Semin Radiat Oncol [Internet]. W.B. Saunders; 2019 [cited 2019 Nov 14];29:181–4. Available from: https://www.sciencedirect.com/science/article/pii/S1053429619300207#fig0001

42. 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. Physica Medica [Internet]. Elsevier; 2016 [cited 2019 Oct 9];32:1659–66. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1120179716309450

43. Sonke J-J, Aznar M, Rasch C. Adaptive Radiotherapy for Anatomical Changes. Semin Radiat Oncol [Internet]. 2019 [cited 2019 Nov 14];29:245–57. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1053429619300165

44. LeCun Y, Boser BE, information processing … JSD-, undefined 1990, … JD-… information processing, 1990 undefined. Handwritten digit recognition with a back-propagation network.

papers.nips.cc [Internet]. [cited 2019 Oct 20]; Available from: http://papers.nips.cc/paper/293-handwritten-digit-recognition-with-a-back-propagation-network.pdf

45. McKevitt P. Daniel Crevier, **AI: The Tumultuous History of the Search for Artificial Intelligence** . London and New York: Basic Books, 1993. Pp. xiv+386. ISBN 0-465-02997-3. £17.99, $27.50. The British Journal for the History of Science [Internet]. Cambridge University Press; 1997 [cited 2019 Oct 20];30:101–21. Available from: https://www.cambridge.org/core/product/identifier/S0007087496302963/type/journal_article

46. Meyer P, Noblet V, Mazzara C, Lallement A. Survey on deep learning for radiotherapy. Comput Biol Med [Internet]. 2018 [cited 2019 Oct 8];98:126–46. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0010482518301318

47. McCarthy J, intelligence PH-R in artificial, 1981 undefined. Some philosophical problems from the standpoint of artificial intelligence. Elsevier [Internet]. [cited 2019 Oct 20]; Available from: https://www.sciencedirect.com/science/article/pii/B9780934613033500337

48. Boden M. Artificial intelligence and natural man. 1980 [cited 2019 Oct 20]; Available from: https://philpapers.org/rec/BODAIA-6

49. Samuel AL. Some Studies in Machine Learning Using the Game of Checkers. IBM J Res Dev [Internet]. 1959 [cited 2019 Oct 11];3:210–29. Available from: http://ieeexplore.ieee.org/document/5392560/

50. Sahiner B, Pezeshk A, Hadjiiski LM, Wang X, Drukker K, Cha KH, et al. Deep learning in medical imaging and radiation therapy. Med Phys [Internet]. John Wiley & Sons, Ltd; 2019 [cited 2019 Oct 2];46:e1–36. Available from: http://doi.wiley.com/10.1002/mp.13264

51. el Naqa I, Murphy MJ. What Is Machine Learning? Machine Learning in Radiation Oncology [Internet]. Cham: Springer International Publishing; 2015 [cited 2019 Oct 11]. p. 3–11. Available from: http://link.springer.com/10.1007/978-3-319-18305-3_1

52. Alpaydin E. Introduction to machine learning [Internet]. 2014 [cited 2019 Nov 2]. Available from: https://books.google.com/books?hl=en&lr=&id=7f5bBAAAQBAJ&oi=fnd&pg=PR5&dq=Alpaydin+E.+Introduction+to+machine+learning.+3rd+ed.+Cambridge,+MA:+The+MIT+Press%3B+2014.&ots=C4aD3q2aKk&sig=MQ2_drQdEfiwnxylKfAu9yuQs1o

53. Mitchel TM. Machine Learning [Internet]. Michalski RS, Carbonell JG, Mitchell TM, editors. Berlin, Heidelberg: Springer Berlin Heidelberg; 1983 [cited 2019 Oct 11]. Available from: http://link.springer.com/10.1007/978-3-662-12405-5

54. el Naqa I, Li R, Murphy MJ. Machine Learning in Radiation Oncology [Internet]. el Naqa I, Li R, Murphy MJ, editors. Cham: Springer International Publishing; 2015 [cited 2019 Nov 2]. Available from: http://link.springer.com/10.1007/978-3-319-18305-3

55. Cui S, Tseng H, Pakela J, ten Haken RK, el Naqa I. Introduction to machine and deep learning for medical physicists. Med Phys. Wiley; 2020;47.

56. González S, García S, del Ser J, Rokach L, Herrera F. A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. Information Fusion. Elsevier B.V.; 2020;64:205–37.

57. Breiman L. Random Forests. Mach Learn [Internet]. Kluwer Academic Publishers; 2001 [cited 2019 Nov 25];45:5–32. Available from: http://link.springer.com/10.1023/A:1010933404324

58. Ibrahem Ahmed Osman A, Najah Ahmed A, Chow MF, Feng Huang Y, El-Shafie A. Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. Ain Shams Engineering Journal. Elsevier; 2021;12:1545–56.

59. Boldrini L, Bibault J-E, Masciocchi C, Shen Y, Bittner M-I. Deep Learning: A Review for the Radiation Oncologist. Front Oncol [Internet]. Frontiers; 2019 [cited 2019 Oct 30];9:977. Available from: https://www.frontiersin.org/article/10.3389/fonc.2019.00977/full

60. Lecun Y, Bengio Y, Hinton G. Deep learning. Nature [Internet]. 2015;521:436–44. Available from: https://www.nature.com/articles/nature14539.pdf

61. Chauhan R, Ghanshala KK, Joshi RC. Convolutional Neural Network (CNN) for Image Detection and Recognition. ICSCCC 2018 - 1st International Conference on Secure Cyber Computing and Communications. Institute of Electrical and Electronics Engineers Inc.; 2018;278–82.

62. Stewart J, Myrehaug SD, Lee Y, Tseng CL, Soliman H, Xie J, et al. Machine Learning-Based Tumor Contour Propagation for MRI Adaptive Radiotherapy of Gliobastoma. International Journal of Radiation Oncology*Biology*Physics [Internet]. Elsevier; 2019 [cited 2019 Nov

1];105:E784. Available from: https://www.sciencedirect.com/science/article/pii/S0360301619315822

63. Valdes G, Chan MF, Lim SB, Scheuermann R, Deasy JO, Solberg TD. <scp>IMRT QA</scp> using machine learning: A multi-institutional validation. J Appl Clin Med Phys. John Wiley and Sons Ltd; 2017;18:279–84.

64. ACM TM-C of the, 1999 undefined, of the ACM TMM-C, undefined 1999. Machine learning and data mining. ri.cmu.edu [Internet]. [cited 2019 Oct 11]; Available from: http://www.ri.cmu.edu/pub_files/pub1/mitchell_tom_1999_1/mitchell_tom_1999_1.pdf

65. Chan MF, Witztum A, Valdes G. Integration of AI and Machine Learning in Radiotherapy QA. Front Artif Intell. Frontiers Media S.A.; 2020;3:76.

66. Chen S, Qin J, Ji X, Lei B, Wang T, Ni D, et al. Automatic Scoring of Multiple Semantic Attributes With Multi-Task Feature Leverage: A Study on Pulmonary Nodules in CT Images. IEEE Trans Med Imaging [Internet]. 2017 [cited 2019 Oct 16];36:802–14. Available from: http://ieeexplore.ieee.org/document/7745891/

67. el Naqa I, Bradley JD, Lindsay PE, Hope AJ, Deasy JO. Predicting radiotherapy outcomes using statistical learning techniques. Phys Med Biol [Internet]. IOP Publishing; 2009 [cited 2019 Nov 1];54:S9--S30. Available from: http://stacks.iop.org/0031-9155/54/i=18/a=S02?key=crossref.42485df4b97868ff644413e820598dc1

68. el Naqa I, Bradley J, Blanco AI, Lindsay PE, Vicic M, Hope A, et al. Multivariable modeling of radiotherapy outcomes, including dose–volume and clinical factors. International Journal of Radiation Oncology*Biology*Physics [Internet]. Elsevier; 2006 [cited 2019 Nov 2];64:1275–86. Available from: https://www.sciencedirect.com/science/article/pii/S0360301605029718

69. Jayasurya K, Fung G, Yu S, Dehing-Oberije C, de Ruysscher D, Hope A, et al. Comparison of Bayesian network and support vector machine models for two-year survival prediction in lung cancer patients treated with radiotherapy. Med Phys [Internet]. John Wiley & Sons, Ltd; 2010 [cited 2019 Oct 31];37:1401–7. Available from: http://doi.wiley.com/10.1118/1.3352709

70. Pella A, Cambria R, Riboldi M, Jereczek-Fossa BA, Fodor C, Zerini D, et al. Use of machine learning methods for prediction of acute toxicity in organs at risk following prostate radiotherapy. Med Phys [Internet]. John Wiley & Sons, Ltd; 2011 [cited 2019 Nov 2];38:2859–67. Available from: http://doi.wiley.com/10.1118/1.3582947

71. Delaney G, Barton M, Jacob S. Estimation of an optimal radiotherapy utilization rate for melanoma. Cancer [Internet]. John Wiley & Sons, Ltd; 2004 [cited 2019 Nov 2];100:1293–301. Available from: http://doi.wiley.com/10.1002/cncr.20092

72. Oh JH, Craft JM, Townsend R, Deasy JO, Bradley JD, el Naqa I. A Bioinformatics Approach for Biomarker Identification in Radiation-Induced Lung Inflammation from Limited Proteomics Data. J Proteome Res [Internet]. American Chemical Society; 2011 [cited 2019 Nov 2];10:1406–15. Available from: https://pubs.acs.org/doi/abs/10.1021/pr101226q

73. Saligan LN, Fernández-Martínez JL, deAndrés-Galiana EJ, Sonis S. Supervised Classification by Filter Methods and Recursive Feature Elimination Predicts Risk of Radiotherapy-Related Fatigue in Patients with Prostate Cancer. Cancer Inform [Internet]. SAGE PublicationsSage UK: London, England; 2014 [cited 2019 Nov 3];13:CIN.S19745. Available from: http://journals.sagepub.com/doi/10.4137/CIN.S19745

74. Masi L, Doro R, Favuzza V, Cipressi S, Livi L. Impact of plan parameters on the dosimetric accuracy of volumetric modulated arc therapy. Med Phys [Internet]. 2013 [cited 2019 Nov 15];40:071718. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23822422

75. 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 [Internet]. John Wiley and Sons Ltd; 2009 [cited 2020 Sep 21];36:5359–73. Available from: https://aapm.onlinelibrary.wiley.com/doi/full/10.1118/1.3238104

76. Valdes G, Solberg TD, Heskel M, Ungar L, Simone CB. Using machine learning to predict radiation pneumonitis in patients with stage I non-small cell lung cancer treated with stereotactic body radiation therapy. Phys Med Biol [Internet]. IOP Publishing; 2016 [cited 2019 Nov 5];61:6105–20. Available from: http://stacks.iop.org/0031-9155/61/i=16/a=6105?key=crossref.ac429b395960e79f413858fa12df82c9

77. 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. Radiotherapy and Oncology [Internet]. Elsevier Ireland Ltd; 2020 [cited 2020 Sep 20]; Available from: https://doi.org/10.1016/j.radonc.2020.07.031

78. Interian Y, Rideout V, Kearney VP, Gennatas E, Morin O, Cheung J, et al. Deep nets vs expert designed features in medical physics: An IMRT QA case study. Med Phys [Internet]. John

Wiley and Sons Ltd.; 2018 [cited 2020 Sep 21];45:2672–80. Available from: http://doi.wiley.com/10.1002/mp.12890

79. Tomori S, Kadoya N, Kajikawa T, Kimura Y, Narazaki K, Ochi T, et al. Systematic method for a deep learning-based prediction model for gamma evaluation in patient-specific quality assurance of volumetric modulated arc therapy. Med Phys [Internet]. Wiley; 2020 [cited 2021 Jan 15];mp.14682. Available from: https://onlinelibrary.wiley.com/doi/10.1002/mp.14682

80. Ono T, Hirashima H, Iramina H, Mukumoto N, Miyabe Y, Nakamura M, et al. Prediction of dosimetric accuracy for VMAT plans using plan complexity parameters via machine learning. Med Phys [Internet]. John Wiley and Sons Ltd.; 2019 [cited 2020 Sep 21];46:3823–32. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/mp.13669

81. 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 [Internet]. John Wiley & Sons, Ltd; 2018 [cited 2019 Dec 16];45:4055–65. Available from: http://doi.wiley.com/10.1002/mp.13112

82. Shin H-C, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. IEEE Trans Med Imaging [Internet]. 2016 [cited 2019 Oct 16];35:1285–98. Available from: http://ieeexplore.ieee.org/document/7404017/

83. Avanzo M, Wei L, Stancanello J, Vallières M, Rao A, Morin O, et al. Machine and deep learning methods for radiomics. Med Phys. Wiley; 2020;47.

84. 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 [Internet]. 2018 [cited 2019 Dec 16];mp.13338. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/mp.13338

85. 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 [Internet]. John Wiley and Sons Ltd.; 2020 [cited 2020 Sep 21];47:1421–30. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/mp.14014

86. Carlson JNK, Park JMI, Park S-Y, Park JMI, Choi Y, Ye S-J. A machine learning approach to the accurate prediction of multi-leaf collimator positional errors. Phys Med Biol [Internet]. IOP

Publishing; 2016 [cited 2019 Nov 14];61:2514–31. Available from: http://stacks.iop.org/0031-9155/61/i=6/a=2514?key=crossref.d7d9ad6bf1750670d6d4a5cf840c2912

87. Wall PDH, Fontenot JD. Application and comparison of machine learning models for predicting quality assurance outcomes in radiation therapy treatment planning. Inform Med Unlocked. Elsevier Ltd; 2020;18:100292.

88. Valdes G, Scheuermann R, Hung CY, Olszanski A, Bellerive M, Solberg TD. A mathematical framework for virtual IMRT QA using machine learning. Med Phys. AAPM - American Association of Physicists in Medicine; 2016;43:4323–34.

89. Lam D, Zhang X, Li H, Deshan Y, Schott B, Zhao T, et al. Predicting gamma passing rates for portal dosimetry-based IMRT QA using machine learning. Med Phys. John Wiley and Sons Ltd.; 2019;46:4666–75.

90. Park JM, Kim JI, Park SY, Oh DH, Kim ST. Reliability of the gamma index analysis as a verification method of volumetric modulated arc therapy plans. Radiation Oncology [Internet]. BioMed Central Ltd.; 2018 [cited 2022 Aug 31];13:1–14. Available from: https://ro-journal.biomedcentral.com/articles/10.1186/s13014-018-1123-x

91. Park JM, Park S-Y, Kim H. Modulation index for VMAT considering both mechanical and dose calculation uncertainties. Phys Med Biol [Internet]. IOP Publishing; 2015 [cited 2019 Dec 12];60:7101–25. Available from: http://stacks.iop.org/0031-9155/60/i=18/a=7101?key=crossref.3c92e4247b0b0039a973b8d33b3c7d87

92. Park JM, Kim J, Park S. Modulation indices and plan delivery accuracy of volumetric modulated arc therapy. J Appl Clin Med Phys [Internet]. John Wiley & Sons, Ltd; 2019 [cited 2019 Dec 11];20:acm2.12589. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/acm2.12589

93. McNiven AL, Sharpe MB, Purdie TG. A new metric for assessing IMRT modulation complexity and plan deliverability. Med Phys [Internet]. 2010 [cited 2019 Nov 15];37:505–15. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20229859

94. Chiavassa S, Bessieres I, Edouard M, Mathot M, Moignier A. Complexity metrics for IMRT and VMAT plans: a review of current literature and applications. Br J Radiol [Internet]. The British Institute of Radiology.; 2019 [cited 2019 Dec 11];92:20190270. Available from: https://www.birpublications.org/doi/10.1259/bjr.20190270

95. Tambasco M, Nygren I, Yorke-Slader E, Villarreal-Barajas JE. FracMod: A computational tool for assessing IMRT field modulation. Physica Medica [Internet]. Elsevier; 2013 [cited 2019 Dec 12];29:537–44. Available from: https://www.sciencedirect.com/science/article/pii/S1120179712002037?via%3Dihub

96. S W. Use of a quantitative index of beam modulation tocharacterize dose conformality: illustration by acomparison of full beamlet IMRT, few-segment IMRT(fsIMRT) and conformal unmodulated radiotherapy. Phys Med Biol [Internet]. IOP Publishing; 2003 [cited 2020 Mar 26];48:2051–62. Available from: https://www.researchgate.net/publication/280869940

97. Tamura M, Matsumoto K, Otsuka M, Monzen H. Plan complexity quantification of dual-layer multi-leaf collimator for volumetric modulated arc therapy with Halcyon linac. Phys Eng Sci Med [Internet]. Springer; 2020 [cited 2020 Sep 1]; Available from: http://link.springer.com/10.1007/s13246-020-00891-2

98. Quintero P. pquinterome/MCS-calculation: Calculating the MCS for VMAT based on:" Masiet al. : Plan parameters and VMAT dosimetric accuracy - 2013" [Internet]. Github. 2020 [cited 2020 Jul 27]. Available from: https://github.com/pquinterome/MCS-calculation

99. 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 [Internet]. British Institute of Radiology; 2015 [cited 2021 Apr 7];88:20140577. Available from: http://www.birpublications.org/doi/10.1259/bjr.20140577

100. Pyyry EJ, Keranen W. Varian APIs A handbook for programming in the Varian oncology software ecosystem [Internet]. Available from: https://www.overleaf.

101. DICOM Processing and Segmentation in Python – Radiology Data Quest [Internet]. [cited 2020 Feb 5]. Available from: https://www.raddq.com/dicom-processing-segmentation-visualization-in-python/

102. Mason D, Guillaume L. GitHub - pydicom/pydicom: Read, modify and write DICOM files with python code [Internet]. [cited 2020 Feb 4]. Available from: https://github.com/pydicom/pydicom

103. 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. Institute of Physics Publishing; 2018;63.

104. 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. Elsevier Inc.; 2019;105:893–902.

105. Dillon J v., Langmore I, Tran D, Brevdo E, Vasudevan S, Moore D, et al. TensorFlow Distributions. 2017 [cited 2022 Sep 23]; Available from: https://arxiv.org/abs/1711.10604v1

106. 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.

107. Nguyen M, Chan GH. Quantified VMAT plan complexity in relation to measurement-based quality assurance results. J Appl Clin Med Phys. John Wiley and Sons Ltd; 2020;21:132–40.

108. Kim H, Huq MS, Lalonde R, Houser CJ, Beriwal S, Heron DE. Early clinical experience with varian halcyon V2 linear accelerator: Dual-isocenter IMRT planning and delivery with portal dosimetry for gynecological cancer treatments. J Appl Clin Med Phys [Internet]. John Wiley & Sons, Ltd; 2019 [cited 2019 Dec 10];20:111–20. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/acm2.12747

109. Semenenko VA, Li XA. Lyman–Kutcher–Burman NTCP model parameters for radiation pneumonitis and xerostomia based on combined analysis of published clinical data. Phys Med Biol [Internet]. IOP Publishing; 2008 [cited 2019 Nov 7];53:737–55. Available from: http://stacks.iop.org/0031-9155/53/i=3/a=014?key=crossref.21b843885a46d494ab5621e962108314

110. Wu Z, Xie C, Hu M, Han C, Yi J, Zhou Y, et al. Dosimetric benefits of IMRT and VMAT in the treatment of middle thoracic esophageal cancer: is the conformal radiotherapy still an alternative option? J Appl Clin Med Phys [Internet]. John Wiley and Sons Ltd; 2014 [cited 2020 Apr 18];15:93–101. Available from: http://doi.wiley.com/10.1120/jacmp.v15i3.4641

111. Park S-Y, Kim IH, Ye S-J, Carlson J, Park JM. Texture analysis on the fluence map to evaluate the degree of modulation for volumetric modulated arc therapy. Med Phys [Internet]. 2014 [cited 2019 Dec 12];41:111718. Available from: http://doi.wiley.com/10.1118/1.4897388

112. Du W, Cho SH, Zhang X, Hoffman KE, Kudchadker RJ. Quantification of beam complexity in intensity-modulated radiation therapy treatment plans. Med Phys [Internet]. John Wiley and Sons Ltd; 2014 [cited 2020 Sep 21];41:021716. Available from: http://doi.wiley.com/10.1118/1.4861821

113. Agnew CE, King RB, Hounsell AR, McGarry CK. Implementation of phantom-less IMRT delivery verification using Varian DynaLog files and R/V output. Phys Med Biol. IOP Publishing; 2012;57:6761–77.

114. Luo Y, Chen S, Valdes G. Machine learning for radiation outcome modeling and prediction. Med Phys. Wiley; 2020;47.

115. Lam D, Zhang X, Li H, Deshan Y, Schott B, Zhao T, et al. Predicting gamma passing rates for portal dosimetry-based IMRT QA using machine learning. Med Phys [Internet]. John Wiley and Sons Ltd.; 2019 [cited 2020 Sep 18];46:4666–75. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/mp.13752

116. Valdes G, Scheuermann R, Hung CY, Olszanski A, Bellerive M, Solberg TD. A mathematical framework for virtual IMRT QA using machine learning. Med Phys [Internet]. AAPM - American Association of Physicists in Medicine; 2016 [cited 2021 Jan 15];43:4323–34. Available from: http://doi.wiley.com/10.1118/1.4953835

117. Luo W, Li J, Price RA, Chen L, Yang J, Fan J, et al. Monte Carlo based IMRT dose verification using MLC log files and R/V outputs. Med Phys. John Wiley and Sons Ltd; 2006;33:2557–64.

118. Law MYY, Liu B. DICOM-RT and Its Utilization in Radiation Therapy1. https://doi.org/101148/rg293075172 [Internet]. Radiological Society of North America ; 2009 [cited 2021 Sep 20];29:655–67. Available from: https://pubs.rsna.org/doi/abs/10.1148/rg.293075172

119. NEMA. dicom.nema.org - /dicom/2004/ [Internet]. [cited 2020 Jan 27]. Available from: http://dicom.nema.org/dicom/2004/

120. Agnew A, Agnew CE, Grattan MWD, Hounsell AR, McGarry CK. Monitoring daily MLC positional errors using trajectory log files and EPID measurements for IMRT and VMAT deliveries. Phys Med Biol. Institute of Physics Publishing; 2014;59.

121. Thwaites DI, DuSautoy AR, Jordan T, McEwen MR, Nisbet A, Nahum AE, et al. The IPEM code of practice for electron dosimetry for radiotherapy beams of initial energy from 4 to 25 MeV based on an absorbed dose to water calibration. Phys Med Biol [Internet]. IOP Publishing; 2003 [cited 2022 Dec 24];48:2929. Available from: https://iopscience.iop.org/article/10.1088/0031-9155/48/18/301

122. Binny D, Spalding M, Crowe SB, Jolly D, Kairn T, Trapp J v., et al. Investigating the use of aperture shape controller in VMAT treatment deliveries. Medical Dosimetry. Elsevier Inc.; 2020;

123. Niemierko A. A generalized concept of equivalent uniform dose (EUD). Med Phys. 1999;26:1100.

124. Fogliata A, Thompson S, Stravato A, Tomatis S, Scorsetti M, Cozzi L. On the gEUD biological optimization objective for organs at risk in Photon Optimizer of Eclipse treatment planning system. J Appl Clin Med Phys [Internet]. Wiley-Blackwell; 2018 [cited 2019 Nov 21];19:106–14. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29152846

125. Tsougos I, Mavroidis P, Theodorou K, Rajala J, Pitkänen MA, Holli K, et al. Clinical validation of the LKB model and parameter sets for predicting radiation-induced pneumonitis from breast cancer radiotherapy. Phys Med Biol [Internet]. 2006;51:L1–9. Available from: http://stacks.iop.org/0031-9155/51/i=3/a=L01?key=crossref.fb43d1eb75266dcb6acb4afbd92add55

126. Luxton G, Keall PJ, King CR. A new formula for normal tissue complication probability (NTCP) as a function of equivalent uniform dose (EUD). Phys Med Biol [Internet]. 2008 [cited 2020 Mar 31];53:23–36. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18182685

127. McGarry CK, Agnew CE, Hussein M, Tsang Y, McWilliam A, Hounsell AR, et al. The role of complexity metrics in a multi-institutional dosimetry audit of VMAT. Br J Radiol [Internet]. British Institute of Radiology; 2016 [cited 2020 Mar 21];89:20150445. Available from: http://www.birpublications.org/doi/10.1259/bjr.20150445

128. Mukaka MM. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. Malawi Medical Journal. Medical Association of Malawi; 2012;24:69–71.

129. Flores-Martinez E, Kim G, Yashar CM, Cerviño LI. Dosimetric study of the plan quality and dose to organs at risk on tangential breast treatments using the Halcyon linac. J Appl Clin Med Phys [Internet]. John Wiley & Sons, Ltd; 2019 [cited 2019 Dec 10];20:acm2.12655. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/acm2.12655

130. van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging—"how-to" guide and critical reflection. Insights Imaging [Internet]. Springer; 2020 [cited 2022 Sep 23];11:1–16. Available from: https://insightsimaging.springeropen.com/articles/10.1186/s13244-020-00887-2

131. 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. American Association for Cancer Research Inc.; 2017;77:e104–7.

132. Hegi F, Atwood T, Keall P, Loo BW. Technical Requirements for Lung Cancer Radiotherapy. IASLC Thoracic Oncology [Internet]. Content Repository Only!; 2018 [cited 2019 Oct 17];318-329.e2. Available from: https://www.sciencedirect.com/science/article/pii/B9780323523578000342

133. Palma G, Monti S, D'Avino V, Conson M, Liuzzi R, Pressello MC, et al. A Voxel-Based Approach to Explore Local Dose Differences Associated With Radiation-Induced Lung Damage. Int J Radiat Oncol Biol Phys [Internet]. Elsevier Inc.; 2016 [cited 2020 Aug 3];96:127–33. Available from: https://pubmed.ncbi.nlm.nih.gov/27511851/

134. Berger L, François P, Gaboriaud G, Rosenwald J-C. Performance optimization of the Varian aS500 EPID system. J Appl Clin Med Phys [Internet]. John Wiley & Sons, Ltd; 2006 [cited 2022 Sep 12];7:105–14. Available from: https://onlinelibrary.wiley.com/doi/full/10.1120/jacmp.v7i1.2158

135. 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 [Internet]. Elsevier Inc.; 2019 [cited 2020 Sep 19];105:893–902. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0360301619335539

136. Lam D, Zhang X, Li H, Deshan Y, Schott B, Zhao T, et al. Predicting gamma passing rates for portal dosimetry-based IMRT QA using machine learning. Med Phys. John Wiley and Sons Ltd.; 2019;46:4666–75.

137. González S, García S, del Ser J, Rokach L, Herrera F. A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. Information Fusion. Elsevier; 2020;64:205–37.

138. Granville DA, Sutherland JG, Belec JG, la Russa DJ. Predicting VMAT patient-specific QA results using a support vector classifier trained on treatment plan characteristics and linac QC metrics. Phys Med Biol [Internet]. Institute of Physics Publishing; 2019 [cited 2020 Sep 21];64:095017. Available from: https://iopscience.iop.org/article/10.1088/1361-6560/ab142e

139. Wall PDH, Fontenot JD. Quality assurance-based optimization (QAO): Towards improving patient-specific quality assurance in volumetric modulated arc therapy plans using machine

learning. Physica Medica [Internet]. 2021;87:136–43. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1120179721001332

140. Lam D, Zhang X, Li H, Deshan Y, Schott B, Zhao T, et al. Predicting gamma passing rates for portal dosimetry-based IMRT QA using machine learning. Med Phys. John Wiley and Sons Ltd.; 2019;46:4666–75.

141. Hernandez V, Saez J, Pasler M, Jurado-Bruggeman D, Jornet N. Comparison of complexity metrics for multi-institutional evaluations of treatment plans in radiotherapy. Phys Imaging Radiat Oncol. Elsevier Ireland Ltd; 2018;5:37–43.

142. Li C, Chen J, Zhu J, Gong G, Tao C, Li Z, et al. Plan quality comparison for cervical carcinoma treated with Halcyon and Trilogy intensity-modulated radiotherapy. J Cancer [Internet]. Ivyspring International Publisher; 2019 [cited 2019 Dec 11];10:6135–41. Available from: http://www.ncbi.nlm.nih.gov/pubmed/31762823

143. Quintero P, Cheng Y, Benoit D, Moore C, Beavis A. Effect of treatment planning system parameters on beam modulation complexity for treatment plans with single-layer multi-leaf collimator and dual-layer stacked multi-leaf collimator. https://doi.org/101259/bjr20201011 [Internet]. The British Institute of Radiology.; 2021 [cited 2021 Sep 20];94. Available from: https://www.birpublications.org/doi/abs/10.1259/bjr.20201011

144. Joost van Griethuysen. pyradiomics/index.rst at master · AIM-Harvard/pyradiomics · GitHub [Internet]. [cited 2021 Sep 20]. Available from: https://github.com/AIM-Harvard/pyradiomics/blob/master/docs/index.rst

145. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research [Internet]. 2011 [cited 2021 Sep 20];12:2825–30. Available from: http://jmlr.org/papers/v12/pedregosa11a.html

146. Menze BH, Kelm BM, Masuch R, Himmelreich U, Bachert P, Petrich W, et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics 2009 10:1 [Internet]. BioMed Central; 2009 [cited 2021 Sep 21];10:1–16. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-213

147. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. ACM International Conference Proceeding Series. 2006;148:233–40.

148. Lam D, Zhang X, Li H, Deshan Y, Schott B, Zhao T, et al. Predicting gamma passing rates for portal dosimetry-based IMRT QA using machine learning. Med Phys. John Wiley and Sons Ltd.; 2019;46:4666–75.

149. Zhen H, Nelms BE, Tomé WA. Moving from gamma passing rates to patient DVH-based QA metrics in pretreatment dose QA. Med Phys [Internet]. John Wiley and Sons Ltd; 2011 [cited 2019 Dec 11];38:5477–89. Available from: http://doi.wiley.com/10.1118/1.3633904

150. Woon WA, Ravindran PB, Ekanayake P, Vikraman S, Amirah S, Lim YYF, et al. Trajectory log file sensitivity: A critical analysis using DVH and EPID. Reports of Practical Oncology and Radiotherapy. Urban and Partner; 2018;23:346–59.

151. Chen RC, Dewi C, Huang SW, Caraka RE. Selecting critical features for data classification based on machine learning methods. J Big Data [Internet]. Springer; 2020 [cited 2022 Sep 24];7:1–26. Available from: https://link.springer.com/articles/10.1186/s40537-020-00327-4

152. Kimura Y, Kadoya N, Tomori S, Oku Y, Jingu K. Error detection using a convolutional neural network with dose difference maps in patient-specific quality assurance for volumetric modulated arc therapy. Phys Med [Internet]. Phys Med; 2020 [cited 2022 Sep 24];73:57–64. Available from: https://pubmed.ncbi.nlm.nih.gov/32330812/

153. Miri N, Keller P, Zwan BJ, Greer P. EPID-based dosimetry to verify IMRT planar dose distribution for the aS1200 EPID and FFF beams. J Appl Clin Med Phys [Internet]. Wiley-Blackwell; 2016 [cited 2022 Sep 13];17:292. Available from: /pmc/articles/PMC5690494/

154. Li Y, Zhang T, Liu Z, Hu H. A CONCATENATING FRAMEWORK OF SHORTCUT CONVOLUTIONAL NEURAL NETWORKS.

155. Payer C, Štern D, Bischof H, Urschler M. Regressing Heatmaps for Multiple Landmark Localization Using CNNs. Springer, Cham; 2016 [cited 2019 Oct 16]. p. 230–8. Available from: http://link.springer.com/10.1007/978-3-319-46723-8_27

156. Wang L, Li J, Zhang S, Zhang X, Zhang Q, Chan MF, et al. Multi-task autoencoder based classification-regression model for patient-specific VMAT QA. Phys Med Biol [Internet]. Phys Med Biol; 2020 [cited 2022 Oct 6];65. Available from: https://pubmed.ncbi.nlm.nih.gov/33245054/

157. Feng M, Valdes G, Dixit N, Solberg TD. Machine learning in radiation oncology: Opportunities, requirements, and needs. Front Oncol. 2018;8:1–7.

158. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning Deep Features for Discriminative Localization. [cited 2022 Oct 5]; Available from: http://cnnlocalization.csail.mit.edu

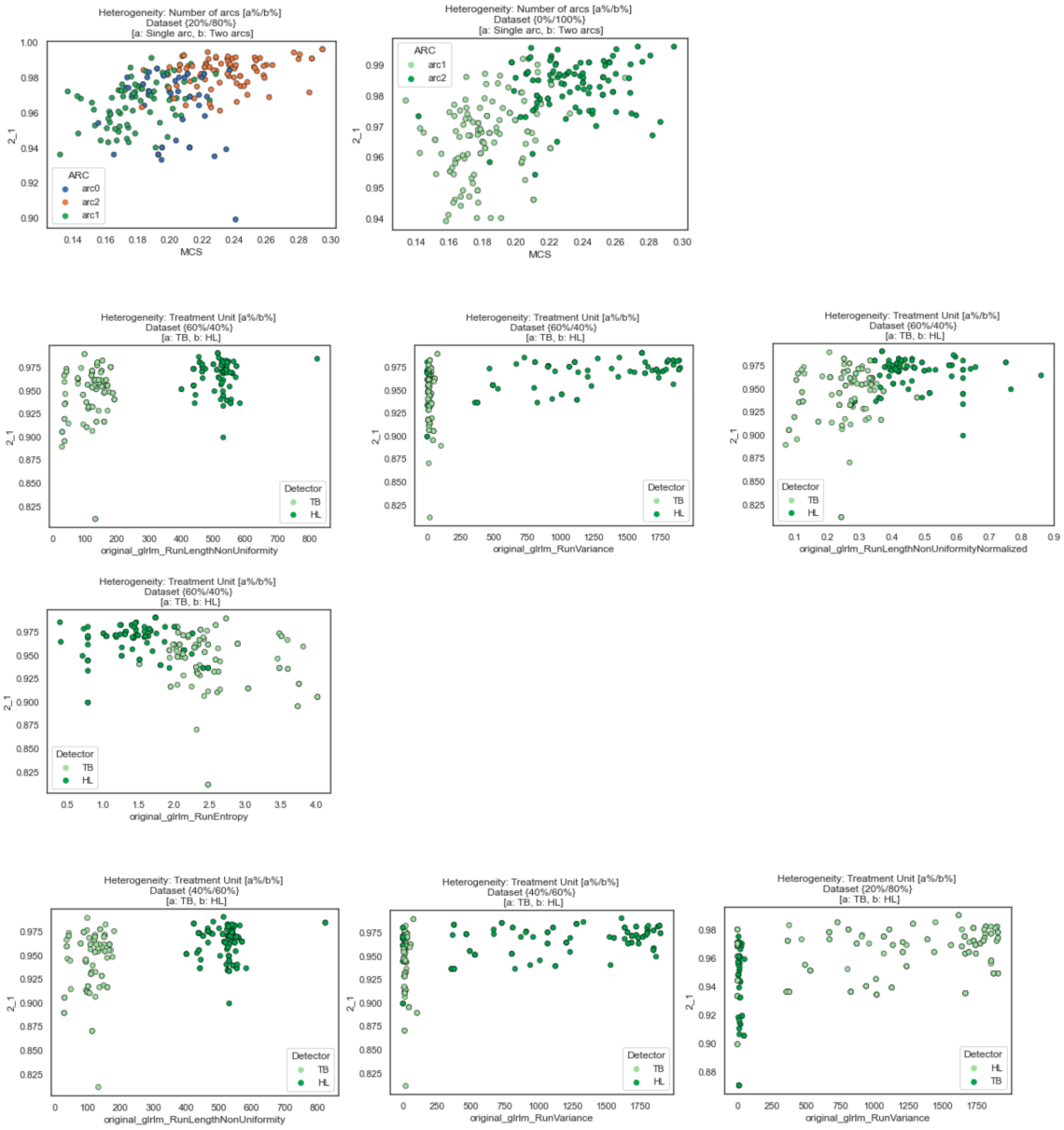
159. Kalet AM, Luk SMH, Phillips MH. Radiation Therapy Quality Assurance Tasks and Tools: The Many Roles of Machine Learning. Med Phys [Internet]. John Wiley and Sons Ltd; 2020 [cited 2021 Jan 21];47:e168–77. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/mp.13445

160. Kalet AM, Luk SMH, Phillips MH. Quality assurance tasks and tools: The many roles of machine learning. Med Phys. John Wiley and Sons Ltd.; 2019.

Appendix 1

Supplementary material 1.1

Scatterplot of moderate correlated features and GPR:2%/1 mm



Supplementary material 1.2

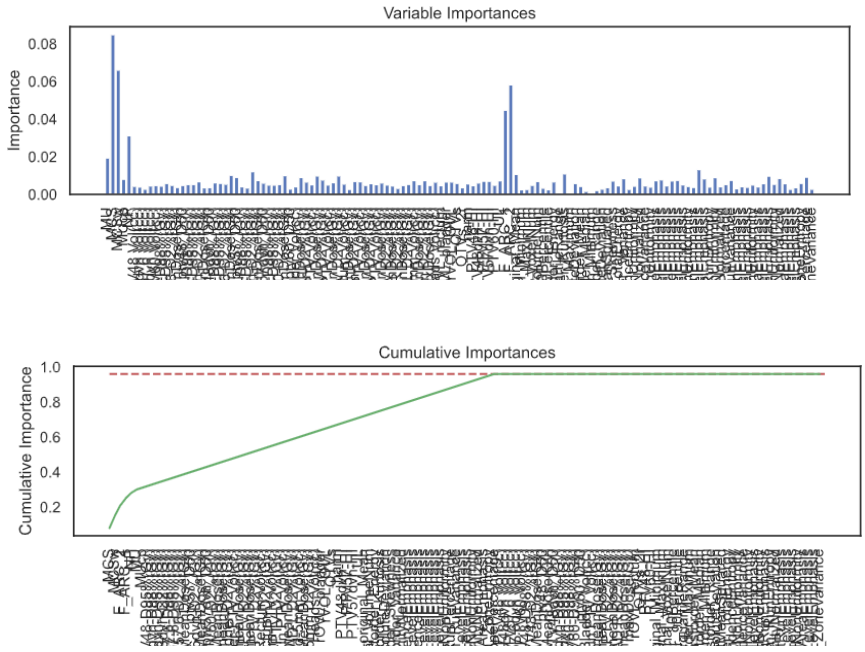
Grid search optimization for each model using the reference dataset.

Random Forest (RF)

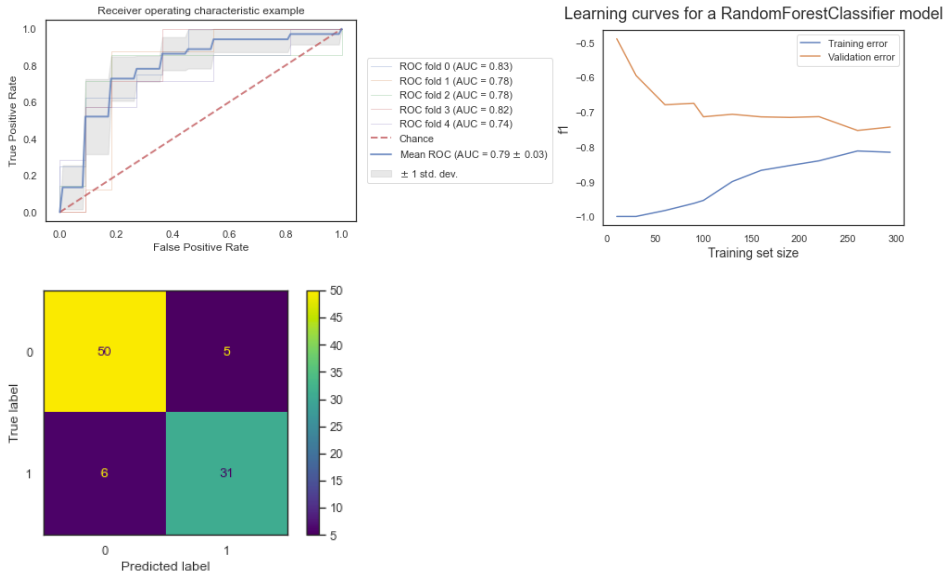
Grid search method (scoring= “roc_auc”) was performed in RF to find the best hyperparameters. The number of trees (n_estimators= 100) and the number of nodes

(max_deepth= 3) were calculated three times shuffling the dataset split. The roc curves (cv=5) for the model with their respective learning curve are displayed below.

Different modifications of the number of trees and nodes were performed and measured. It was concluded that 100 and 3 were de best parameters.

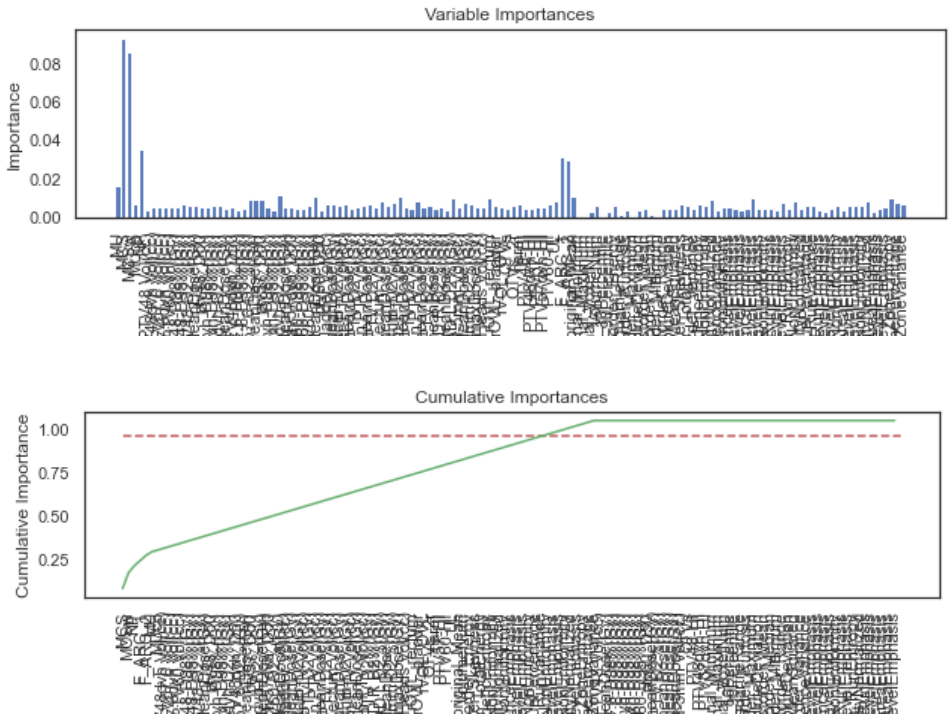


max_depth	n_estimators	Accuracy	ROC_AUC_CV_5	Precision	Recall	F1
3	100	0.73	0.79+-0.03	0.71	0.59	0.65

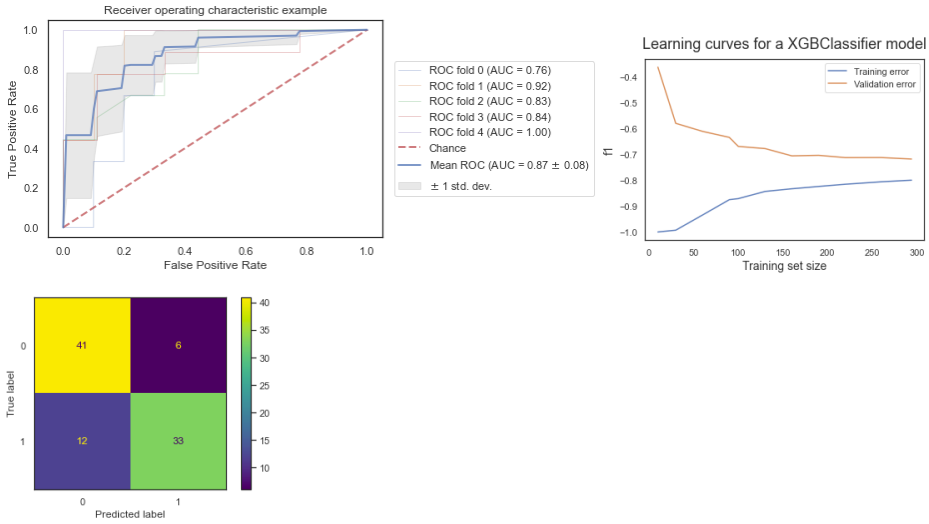


XG-Boost

To achieve up to 95% maximum prediction accuracy, a simple XGBoost identified 72 features as the most important. The hyperparameters resulting from Grid search for XGBoost with all features were learning_rate=0.01, max_depth=2, n_estimators=170.



max_depth	n_estimators	Learning Rate	Accuracy	ROC_AUC_CV_5	Precision	Recall	F1
2	170	0.01	0.80	0.87+-0.08	0.78	0.84	0.81



Neural Network

A simple NN was trained to predict GPRs, changing the number of layers and nodes intuitively keeping the simplest architecture. Using all features the model performance with CV= 5 is shown below in terms of AUC and confusion matrix.

Model: "sequential_157"

_____

Layer (type) Output Shape Param #

==========

layer_1 (Dense) (None, 210) 12474

_____

layer_2 (Dense) (None, 60) 2211

_____

layer_3 (Dense) (None, 1) 34

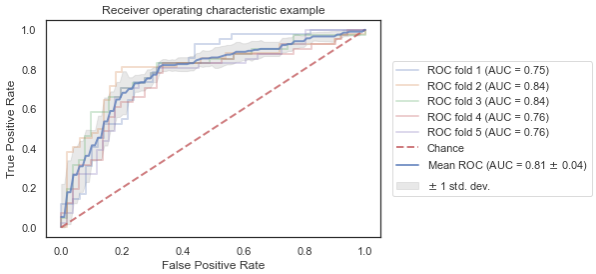
==========

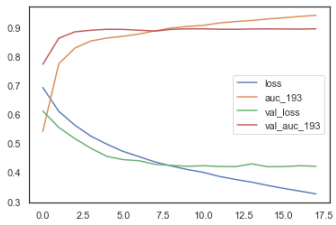
Total params: 14,719

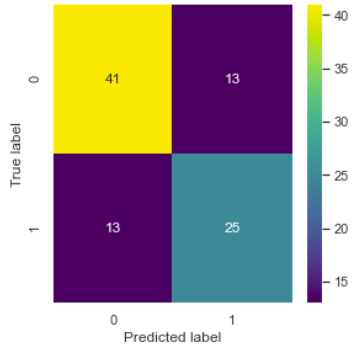
Trainable params: 14,719

Non-trainable params: 0

_____







Accuracy	ROC_AUC_CV_5	Precision	Recall	F1
0.72	0.81+-0.04	0.66	0.66	0.66

Supplementary Material 1.3

Classification performance metrics for each heterogeneity

Random Forest

Heterogeneity	Metric	100% - 0%		80% - 20%		60% - 40%		40% - 60%		20% - 80%		0% - 100%	
		mv	sd	mv	sd	mv	sd	mv	sd	mv	sd	mv	sd
Dose per fraction	Sensitivity	0.75	0.04	0.70	0.16	0.84	0.03	0.55	0.05	0.74	0.03	0.84	0.02
	Specificity	0.75	0.01	0.73	0.03	0.79	0.06	0.78	0.00	0.67	0.04	0.58	0.04
	Precision	0.77	0.01	0.65	0.05	0.61	0.03	0.68	0.02	0.69	0.03	0.73	0.02
	F1	0.76	0.03	0.70	0.09	0.70	0.02	0.61	0.04	0.71	0.02	0.78	0.01
	AUC	0.82	0.04	0.70	0.03	0.62	0.12	0.71	0.11	0.77	0.09	0.88	0.07
Number of arcs	Sensitivity	0.80	0.01	0.72	0.01	0.78	0.01	0.68	0.01	0.71	0.03	0.72	0.07
	Specificity	0.81	0.03	0.81	0.01	0.63	0.01	0.67	0.01	0.75	0.03	0.80	0.02
	Precision	0.83	0.02	0.80	0.01	0.73	0.01	0.63	0.03	0.73	0.02	0.72	0.02
	F1	0.81	0.01	0.75	0.01	0.75	0.00	0.65	0.02	0.72	0.02	0.72	0.04
	AUC	0.88	0.06	0.81	0.07	0.80	0.04	0.75	0.13	0.79	0.08	0.86	0.07
Treatment Unit	Sensitivity	0.80	0.10	0.84	0.08	0.81	0.11	0.80	0.10	0.78	0.12	0.81	0.04
	Specificity	0.65	0.02	0.60	0.08	0.56	0.13	0.48	0.06	0.65	0.13	0.83	0.05
	Precision	0.75	0.07	0.74	0.05	0.71	0.02	0.71	0.08	0.67	0.10	0.81	0.09
	F1	0.77	0.09	0.78	0.06	0.75	0.04	0.75	0.08	0.69	0.08	0.81	0.06
	AUC	0.83	0.09	0.68	0.13	0.78	0.10	0.78	0.13	0.76	0.11	0.91	0.09
Anatomic Region	Sensitivity	0.77	0.04	0.67	0.09	0.61	0.11	0.44	0.03	0.46	0.10	0.48	0.06
	Specificity	0.65	0.06	0.66	0.04	0.62	0.19	0.83	0.11	0.87	0.05	0.93	0.05
	Precision	0.73	0.03	0.66	0.06	0.65	0.06	0.68	0.16	0.67	0.10	0.79	0.09
	F1	0.75	0.03	0.66	0.08	0.62	0.06	0.53	0.03	0.52	0.10	0.64	0.03
	AUC	0.84	0.13	0.78	0.07	0.73	0.11	0.72	0.06	0.78	0.11	0.82	0.05

XG-Boost

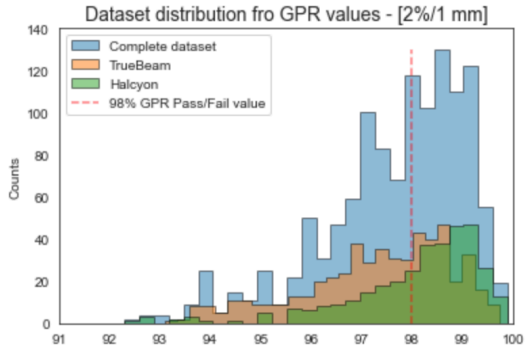
Heterogeneity	Metric	100% - 0%		80% - 20%		60% - 40%		40% - 60%		20% - 80%		0% - 100%	
		mv	sd	mv	sd	mv	sd	mv	sd	mv	sd	mv	sd
Dose per fraction	Sensitivity	0.69	0.02	0.72	0.09	0.80	0.04	0.56	0.06	0.73	0.03	0.84	0.02
	Specificity	0.64	0.06	0.65	0.08	0.44	0.10	0.71	0.06	0.67	0.16	0.58	0.04
	Precision	0.73	0.02	0.67	0.06	0.66	0.02	0.64	0.02	0.70	0.08	0.73	0.02
	F1	0.70	0.02	0.70	0.07	0.72	0.01	0.60	0.03	0.72	0.03	0.78	0.01
	AUC	0.87	0.09	0.75	0.1	0.66	0.1	0.70	0.05	0.78	0.09	0.89	0.07
Number of arcs	Sensitivity	0.82	0.02	0.71	0.04	0.77	0.02	0.56	0.07	0.61	0.13	0.75	0.15
	Specificity	0.75	0.10	0.84	0.05	0.66	0.03	0.70	0.01	0.68	0.04	0.89	0.06
	Precision	0.81	0.05	0.80	0.01	0.75	0.02	0.62	0.04	0.62	0.08	0.83	0.01
	F1	0.80	0.01	0.75	0.02	0.77	0.01	0.59	0.06	0.62	0.11	0.78	0.07
	AUC	0.85	0.07	0.81	0.04	0.80	0.07	0.69	0.05	0.77	0.12	0.87	0.09
Treatment Unit	Sensitivity	0.80	0.03	0.74	0.06	0.76	0.09	0.78	0.12	0.68	0.13	0.77	0.07
	Specificity	0.58	0.11	0.61	0.14	0.64	0.16	0.55	0.04	0.65	0.12	0.77	0.06
	Precision	0.70	0.07	0.74	0.04	0.75	0.09	0.72	0.10	0.65	0.11	0.75	0.08
	F1	0.75	0.04	0.73	0.05	0.75	0.08	0.75	0.11	0.66	0.12	0.76	0.07
	AUC	0.80	0.04	0.68	0.12	0.76	0.12	0.71	0.17	0.84	0.17	0.86	0.06
Anatomic Region	Sensitivity	0.81	0.02	0.65	0.11	0.60	0.05	0.46	0.14	0.63	0.24	0.65	0.10
	Specificity	0.63	0.10	0.68	0.02	0.67	0.08	0.78	0.12	0.73	0.18	0.91	0.02
	Precision	0.74	0.05	0.66	0.05	0.66	0.01	0.60	0.10	0.67	0.13	0.78	0.05
	F1	0.77	0.02	0.66	0.08	0.62	0.02	0.50	0.14	0.59	0.13	0.61	0.08
	AUC	0.84	0.09	0.78	0.07	0.67	0.1	0.75	0.04	0.76	0.18	0.88	0.08

Neural Network

Heterogeneity	Metric	100% - 0%		80% - 20%		60% - 40%		40% - 60%		20% - 80%		0% - 100%	
		mv	sd	mv	sd	mv	sd	mv	sd	mv	sd	mv	sd
Dose per fraction	Sensitivity	0.93	0.06	0.91	0.03	0.91	0.06	0.80	0.05	0.78	0.11	0.92	0.01
	Specificity	0.90	0.02	0.92	0.04	0.87	0.06	0.86	0.10	0.91	0.04	0.89	0.06
	Precision	0.90	0.02	0.93	0.03	0.91	0.03	0.84	0.10	0.90	0.05	0.88	0.06
	F1	0.90	0.02	0.92	0.01	0.91	0.04	0.81	0.04	0.83	0.08	0.90	0.03
	AUC	0.94	0.03	0.89	0.04	0.87	0.03	0.88	0.05	0.93	0.02	0.92	0.03
Number of arcs	Sensitivity	0.87	0.07	0.80	0.08	0.88	0.01	0.90	0.06	0.90	0.05	0.94	0.06
	Specificity	0.80	0.03	0.84	0.05	0.67	0.04	0.68	0.01	0.70	0.04	0.84	0.02
	Precision	0.83	0.04	0.82	0.03	0.89	0.05	0.87	0.02	0.89	0.02	0.92	0.03
	F1	0.84	0.01	0.80	0.01	0.90	0.03	0.88	0.01	0.91	0.01	0.91	0.01
	AUC	0.90	0.05	0.84	0.01	0.90	0.02	0.83	0.11	0.88	0.04	0.96	0.04
Treatment Unit	Sensitivity	0.98	0.04	0.98	0.02	0.89	0.01	0.95	0.05	0.86	0.10	0.93	0.08
	Specificity	0.92	0.08	0.90	0.10	0.91	0.02	0.86	0.07	0.89	0.02	0.95	0.08
	Precision	0.94	0.06	0.92	0.07	0.93	0.02	0.91	0.06	0.88	0.01	0.96	0.07
	F1	0.96	0.02	0.95	0.05	0.90	0.01	0.92	0.04	0.87	0.07	0.95	0.05
	AUC	0.97	0.03	0.96	0.02	0.89	0.07	0.92	0.03	0.94	0.01	0.98	0.01
Anatomic Region	Sensitivity	0.90	0.10	0.92	0.01	0.88	0.05	0.91	0.04	0.83	0.13	0.93	0.07
	Specificity	0.79	0.07	0.87	0.05	0.93	0.06	0.92	0.07	0.91	0.04	0.98	0.03
	Precision	0.83	0.06	0.88	0.05	0.92	0.08	0.88	0.10	0.86	0.06	0.97	0.04
	F1	0.86	0.03	0.89	0.03	0.90	0.06	0.89	0.07	0.85	0.09	0.95	0.03
	AUC	0.94	0.04	0.90	0.02	0.90	0.05	0.93	0.03	0.95	0.04	0.95	0.04

Appendix 2

Supplementary Material 2.1



Supplementary Figure 1: Distribution of GPR values from plans evaluated with 2%/1 mm criteria.

Supplementary Table 1: Mean and standard deviation values (mv, sdv) of gamma passing rate (GPR) evaluation of 547 prostate treatment plans for Halcyon-v2 and TrueBeam, using different criteria values of dose difference (DD) and distance to agreement (DTA).

DD [%]/DTA [mm]	Halcyon		TrueBeam	
	mv	sdv	mv	sdv
3/3	100	0	99.6	0.3
3/2	100	0	99.2	0.7
2/3	100	0	99.0	0.8
2/2	100	0	98.1	1.5
2/1	98	1.6	97.4	2.2

Supplementary Material 2.2

Model_1 architecture summary

```
Model: "model_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_2 (InputLayer)         [(None, 70, 177, 1)]      0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 70, 177, 64)       640
_________________________________________________________________
max_pooling2d_2              (MaxPooling2 (None, 35, 88, 64)   0
_________________________________________________________________
dropout_2 (Dropout)          (None, 35, 88, 64)        0
_________________________________________________________________
```

```
conv2d_3 (Conv2D)           (None, 35, 88, 64)          36928
_________________________________________________________________
max_pooling2d_3             (MaxPooling2 (None, 17, 44, 64)     0
_________________________________________________________________
dropout_3 (Dropout)         (None, 17, 44, 64)          0
_________________________________________________________________
flatten_1 (Flatten)         (None, 47872)               0
_________________________________________________________________
dense_2 (Dense) (None, 90) 4308570
_________________________________________________________________
dense_3 (Dense) (None, 1) 91
=================================================================
Total params: 4,346,229
Trainable params: 4,346,229
Non-trainable params: 0
```

Model_2 architecture summary

```
Model: "model_2"
_________________________________________________________________
Layer (type) Output Shape Param #
=================================================================
input_1 (InputLayer) [(None, 176, 1)] 0
_________________________________________________________________
conv1d (Conv1D) (None, 176, 70) 420
_________________________________________________________________
max_pooling1d (MaxPooling1D) (None, 58, 70) 0
_________________________________________________________________
dropout (Dropout) (None, 58, 70) 0
_________________________________________________________________
flatten (Flatten) (None, 4060) 0
_________________________________________________________________
batch_normalization (BatchNo (None, 4060) 16240
_________________________________________________________________
dense (Dense) (None, 90) 365490
_________________________________________________________________
dense_1 (Dense) (None, 1) 91
=================================================================
Total params: 382,241

Trainable params: 374,121

Non-trainable params: 8,120
```

Model_3 architecture summary

```
Model: "model_3"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_10 (InputLayer)        [(None, 512, 512, 1)]     0
_________________________________________________________________
conv2d_25 (Conv2D)           (None, 512, 512, 64)      1664
```

```
_________________________________________________________________
max_pooling2d_27 (MaxPooling (None, 170, 170, 64)      0
_________________________________________________________________
dropout_19 (Dropout)         (None, 170, 170, 64)      0
_________________________________________________________________
conv2d_26 (Conv2D)           (None, 170, 170, 32)      18464
_________________________________________________________________
max_pooling2d_28 (MaxPooling (None, 85, 85, 32)        0
_________________________________________________________________
dropout_20 (Dropout)         (None, 85, 85, 32)        0
_________________________________________________________________
conv2d_27 (Conv2D)           (None, 85, 85, 32)        9248
_________________________________________________________________
max_pooling2d_29 (MaxPooling (None, 42, 42, 32)        0
_________________________________________________________________
dropout_21 (Dropout)         (None, 42, 42, 32)        0
_________________________________________________________________
flatten_8 (Flatten)          (None, 56448)             0
_________________________________________________________________
dense_24 (Dense)             (None, 360)               20321640
_________________________________________________________________
dense_25 (Dense)             (None, 90)                32490
_________________________________________________________________
dense_26 (Dense)             (None, 1)                 91
=================================================================
Total params: 20,383,597
Trainable params: 20,383,597
Non-trainable params: 0
```