Implementation of volumetric-modulated arc therapy for locally advanced breast cancer patients: Dosimetric comparison with deliverability consideration of planning techniques and predictions of patient-specific QA results via supervised machine learning

Caroline Noblet, Marie Duthy, Frédéric Coste, Marie Saliou, Benoît Samain, Franck Drouet, Thomas Papazyan, Matthieu Moreau

Implementation of volumetric-modulated arc therapy for locally advanced breast cancer patients: Dosimetric comparison with deliverability consideration of planning techniques and predictions of patient-specific QA results via supervised machine learning.

Caroline NOBLET [a] *, Marie DUTHY [a], Frédéric COSTE [a], Marie SALIOU [b], Benoît SAMAIN [b], Franck DROUET [b], Thomas PAPAZYAN [b], Matthieu MOREAU [a]

[a] Department of Medical Physics, Clinique Mutualiste de l'Estuaire, Cité Sanitaire, Saint-Nazaire, France

[b] Department of Radiation Oncology, Clinique Mutualiste de l'Estuaire, Cité Sanitaire, Saint-Nazaire, France

*Corresponding author at : Service de Physique Médicale, 11 boulevard Georges Charpak - CS 20252 44606 Saint-Nazaire Cedex, France. E-mail address : caroline.noblet@hospigrandouest.fr

Abstract

Purpose: The aim of this study was to implement a clinically deliverable VMAT planning technique dedicated to advanced breast cancer, and to predict failed QA using a machine learning (ML) model to optimize the QA workload.

Methods: For three planning methods (2A: 2-partial arc, 2AS: 2-partial arc with splitting, 4A: 4-partial arc), dosimetric results were compared with patient-specific QA performed with the electronic portal imaging device of the linac. A dataset was built with the pass/fail status of the plans and complexity metrics. It was divided into training and testing sets. An ML metamodel combining predictions from six base classifiers was trained on the training set to predict plans as 'pass' or 'fail'. The predictive performances were evaluated using the unseen data of the testing set.

Results: The dosimetric comparison highlighted that 4A was the highest dosimetric performant method but also the poorest performant in the QA process. 2AS spared the best heart, but provided the highest dose to the contralateral breast and lowest node coverage. 2A provides a dosimetric compromise between organ at risk sparing and PTV coverage with satisfactory QA results. The metamodel had a median predictive sensitivity of 73% and a median specificity of 91%.

Conclusions: The 2A method was selected to calculate clinically deliverable VMAT plans; however, the 2AS method was maintained when the heart was of particular importance and breath-hold techniques were not applicable. The metamodel provides promising predictive performance, and it is intended to be improved as a larger dataset becomes available.

Key words : breast, VMAT, patient-specific QA, machine learning.

1. Introduction

Breast cancer is one of the most common cancers, but remains one of the most technically challenging to treat with radiation therapy. The shape and size diversity of the breast or chest wall, its superficial location, and its direct proximity to organs at risk (OARs) make this location particularly challenging. Conventional 3D conformal radiation therapy techniques (3D-CRT) can fail to provide appropriate target dose conformity and homogeneity while sparing OARs, in particular, for locally advanced breast cancers involving the internal mammary chain (IMC). In response to this limitation, over the past few years, intensity modulated radiation therapy (IMRT) has been increasingly popular for the treatment of breast cancer. Among the advanced advantages are better target coverage and homogeneity, sparing surrounding OARs from a high dose, and a lower dose to the heart and ipsilateral (IL) lung [1-8]. Several modulation techniques have been proposed, including static IMRT, tomotherapy, volumetric modulated arc therapy (VMAT), and hybrid techniques. This study focuses on VMAT techniques dedicated to lymph node-positive breast cancer with IMC. The literature offers a wide range of planning methods for VMAT techniques. After Popescu et al. [1]] published their two partial-arc methods in 2010, many published methods involved a similar two-arc ballistic [5-17]. Nonetheless, some published VMAT techniques also proposed three or four partial arcs [18-21] and more confidentially only one partial arc [22,23]. To the best of the author's knowledge, patient-specific quality assurance (PSQA) results are rarely presented with published VMAT planning techniques [10,22]. PSQA measurements are strongly recommended for verification before each modulated radiation therapy treatment delivery [24-25]. For example, in our institution, a PSQA measurement is performed for each VMAT plan after it is approved by a radiation oncologist. Breast cancer is the first female cancer in France, and the implementation of a modulated technique for breasts in a clinical routine may therefore imply a severe increase in quality control activity. As evidenced by the multiplicity of published techniques, VMAT planning dedicated to the breast remains challenging and can involve highly complex modulated plans, leading to PSQA failure. This may involve re-planning until PSQA is successful, which may overload the dosimetric/pre-treatment QA workflow. Consequently, with the implementation of such a technique for breasts in clinical routine, it would be interesting to predict the outcome of the PSQA to minimize the time lost in measuring treatment plans that are likely to fail QA. One way to predict the PSQA results is to use metrics that quantify the complexity of the plans and correlate them with the results of the PSQA already measured. Several studies have proposed plan complexity metrics (PCMs), as reviewed by Chiavassa et al. [26] and Antoine et al. [27]. As discussed in these two reviews, the correlation between PCMs and PSQA results depends on the local procedures, as it is impacted for example by anatomical localization, dose calculation algorithm, dose measurement technique, or analysis protocol. In addition, only weak or moderate correlations were found between PCM and PSQA results, leading to unreliable PSQA result predictions. Recently, the rise of machine learning (ML) techniques has paved the way for the further optimization of these predictions through the use of supervised learning methods based on PCMs. Such methods make it possible to combine several PCMs, and have led to more accurate predictions [28-37]. Most published ML methods for PSQA predictions involve ML algorithms, such as support vector machine models, decision trees, or linear regressions. In the context of technique implementation with a small dataset of PSQA results, combining multiple ML algorithms in an ensemble model is proposed, to enhance the robustness of the predictions.

This study aimed to select a clinically deliverable VMAT technique dedicated to advanced breast cancer, that meets our dosimetric objectives, while preventing the overload of our workflow by limiting the number of QA failures. In summary, the objectives were to 1) compare the dosimetric results of three different VMAT planning techniques, 2) compare the QA results of these techniques, and 3) propose a method for predicting the QA results.

2. Methods

2.1. Patient selection and target definition

Computed-tomography (CT) images of 28 lymph node-positive breast cancer patients, previously treated with 3D-CRT technique, were retrospectively selected for this study (14 left-sided and 14 right sided, named 'LS' group and 'RS' group, respectively). Eighteen patients underwent mastectomy (eight LS and ten RS), and ten underwent breast conservative surgery (six LS and four RS). All the patients were imaged with a 2.5 mm slice thickness

(Optima CT 580RT, GE Medical Systems) in free breathing. The clinical target volumes (CTVs) included the supraclavicular and axillary nodes (CTV_n), IMC (CTV_{IMC}), and breast/chest wall ($CTV_{B/CW}$). The planning target volumes (PTVs) were created from the CTVs with a 5 mm margin and cropped within 3 mm from the skin. The prescribed dose was 50 Gy in 25 fractions for all PTVs.

2.2. Treatment planning

All VMAT plans were optimized for a Varian iX linear accelerator (Varian Medical Systems) with a Millennium 120 multileaf collimator (MLC). The Eclipse treatment planning system (TPS) was used to perform plan optimization (photon optimizer algorithm version 13.7.16) and dose calculation (anisotropic analytical algorithm version 13.7.16) with a 2.5 mm dose matrix resolution. An 8 mm virtual bolus (set to -100 HU) was added to the surface of the breast/chest wall. An optimized PTV was created by expanding the original $PTV_{B/CW}$ 5 mm outside the skin in the 8 mm bolus. Bolus and PTV_{opt} were used for the plan optimization. The virtual bolus method forces the MLC leaves to be positioned beyond the patient's outer contour to improve the robustness of the treatment delivery. The efficiency of this method has been demonstrated in several studies [13,14,38,39]. For each plan, the final dose was calculated with and without a virtual bolus with the same number of monitor units (MU). The plan had to be acceptable in both situations regarding the dosimetric objectives listed in Appendix A.

As previously mentioned, the most published VMAT ballistics for treating advanced breast cancer involve two partial arcs from 280-340° to 130-180° (for LS patients) with or without collimator rotation. Based on this typical ballistic, Boman et al. [9] proposed in 2016, to split the two arcs into two, towards the middle of their trajectory, to optimize collimator rotations to patient anatomy and better spare the heart and ipsilateral (IL) lung. In addition, based on the classical two-partial-arc ballistic, Lang et al. [20] proposed in 2019, to add two partial arcs with a 90° collimator rotation to improve dose conformity.

Three VMAT ballistics were compared in this study:

- 2A: two partial arcs with a 10/350° collimator angle based on the two-arc classic method.
- 2AS: two partial arcs based on the method of Boman et al. [9]. The collimator angle of the subarcs was set to ±10°/±30° for the two upper arcs and ±330°/±350° for the two lower arcs. The internal X-jaw (to the patient) of each subarc was set closer to the isocenter (about 2 cm depending on patient anatomy) to reduce the heart and IL lung irradiation.
- 4A: four partial arcs inspired by the Lang et al. method [20], with 10/350° and 80/280° collimator angles. The 80/280° collimator angle arcs overlap by at least 4 cm at the isocenter.

For all ballistics, the arcs ranged from 300° to 179.9°, or 60° to 180.1°, respectively, for patients with LS and RS breast cancer. The jaw opening in the X-direction was limited to 15-17 cm to avoid modulation deterioration owing to the maximum 15 cm leaf span of the MLC. The three VMAT plans were calculated for each of the 28 selected CT scans. The dose-volume histogram (DVH) data were extracted from the 84 plans calculated without the virtual bolus and statistically compared using Wilcoxon signed-rank tests. Statistical significance was set at $p < 0.05$. The 2A ballistic dosimetric results were taken as a reference, since 2A is the most published VMAT technique.

2.3. Pre-treatment quality assurance

Pre-treatment quality assurance measurements were carried out for all plans using portal dosimetry software (Varian Medical System, PDIP v13.7.16). The TPS dose calculation was compared to VMAT integrated images acquired using an electronic portal imaging device (Varian EPID aSi 500). Comparison was made in terms of gamma passing rate (GPR) based on an absolute dose global gamma analysis with a lower dose threshold of 10%. For a plan to be validated by a medical physicist, all arcs in that plan must pass QA, that is, they must have a GPR greater than 95%. A plan was classified as failed if any of its arcs failed quality control. Gamma criteria for dose difference and distance to agreement were set at 2% and 2 mm, respectively. In the context of technique implementation, these criteria were in agreement with those of Heilemann et al. [40] and with the AAPM Task Group No. 218 [25] recommendation to use tighter criteria than their recommended 3%/2mm criteria to emphasize

subtle regional errors and systematic errors for a specific treatment site. Judging in a sufficiently strict manner the quality of the plans optimized with a newly implemented technique was of particular importance, especially since a particular modulation effort was required of the treatment machines, with asymmetric fields and several OARs close to the target.

2.4. Supervised machine learning classifier

2.4.1. Dataset implementation

Twenty-two PCMs [41-46] were selected to characterize the VMAT plan complexity and were calculated from the DICOM RT plan files using an in-house Python program. PCMs were averaged over the control points (every 2°) and over the beams according to their respective weights. The $PTV_{B/CW}$ volume in cc was added to the features, in addition to the PCMs. These are listed in Appendix B.

A dataset was created with the minimum GPR for each plan and corresponding features. The GPR was transformed into a binary class of PSQA status for each plan ("pass" or "fail") based on a predefined GPR threshold. This threshold was defined as the 95% validation threshold plus the margin estimated from the EPID delivery uncertainty. To evaluate the EPID response variation over time, 18 randomly selected plans (equally distributed among the planning methods) and 60 partial arcs in total were acquired three times with the portal imager, three months apart.

2.4.2. Data pre-processing

The dataset was divided into training and testing sets (70% and 30% of the original dataset, respectively) using stratification to ensure the same distribution of the plan QA classes over both sets. Training data was used to train the ML model and testing data was unseen data only used to evaluate the performances of the validated ML model.

Training set features were scaled via a standardization process so that the features had a mean of 0 and a standard deviation of 1. The testing set features were scaled based on the mean and standard deviation of the training set features.

2.4.3. Feature relative importance

The number and nature of the features used in model training have an impact on its accuracy. The use of irrelevant features can reduce the predictive performance of ML models and unnecessarily increase the cost of acquiring data. An automated feature selection strategy was applied to the training set to estimate the relative importance of each feature contributing to the model performance. They were ranked using a recursive feature elimination (RFE) method based on six different algorithms that were used to build the metamodel:

- K nearest neighbors (KNN)
- Linear discriminant analysis (LDA)
- Regularised logistic regression (LR)
- Naïve-Bayes (NB)
- Support vector machine with a radial kernel (SVM)
- Random forest (RF)

The process was performed with R software version 3.6.1. and the 'caret' library (Classification And REgression Training, from the Comprehensive R Archive Network).

The RFE algorithm starts by building a classifier (based on the chosen algorithm) with all features and calculates its performance. Thereafter, for each feature F_i in the feature set F_{all}, it trains a classifier with F_{all}-F_i features and calculates its performance and the difference between the performance of the classifier trained with F_{all}. Finally, it computes the feature rank from the performance-loss profile. Performance was evaluated using the average area under the receiver-operator characteristic curve (AUROC). The RFE process was performed using 5-fold cross-validation (CV). During a 5-fold CV, the input dataset was randomly divided into five subsets, four training subsets,

and one validation subset. Five sub-models were iteratively fitted on the training subsets and evaluated on the remaining validation subset. The mean performance was computed for the five sub-models (Fig 1.).

To evaluate the variation in the final RFE feature rank, the process was performed for each ML algorithm using 100 random splits of the original dataset into training and testing sets. The feature rank was averaged over all splits and algorithms to limit its dependence on the data split and ML algorithm.

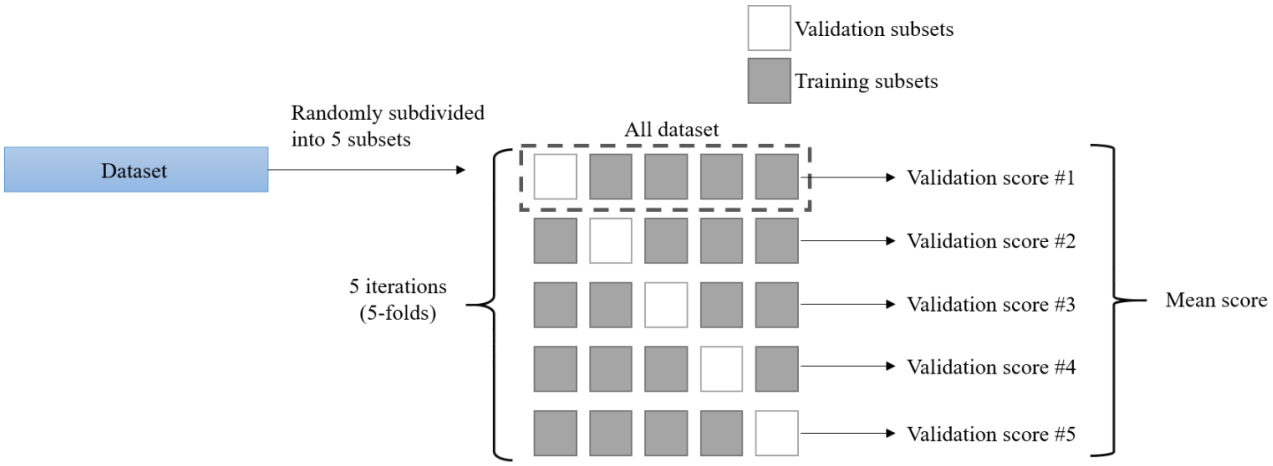


Fig 1. Five-fold cross-validation process.

2.4.4. Model building

An ensemble approach was implemented to generate QA outcome predictions. This approach is based on the principle that combining predictions from several base classifiers allows the construction of a more robust model. Base classifiers were trained using the same training data but with different families of algorithms (Appendix C.) to generate the diversity needed to obtain a complementary set of predictions. The aim is to combine models with sufficiently dissimilar characteristics that are not constantly unanimous in predictions. Simple algorithms with no more than two hyperparameters to tune were chosen to limit the complexity of the models and reduce the risk of overfitting. Hyperparameters are used to train the model for a specific problem. For example, they can control the strength of the regularization process, which modifies the cost function to penalize the complexity of the model. These parameters were tuned during the training process.

For each base learner, training was performed on the training set using a 100-repeated 5-fold CV that repeats 100 times the random split of the input dataset into five subsets, resulting in 500 sub-models. Global performance was estimated by averaging the performances resulting from each 5-fold CV. The 100-repeated 5-fold CV was used to select the best hyperparameter combination using a random search process. The hyperparameters were chosen to maximize the average CV score, defined as the AUROC. In addition to hyperparameter tuning, the CV process provides an early estimate of the skill of the validated model over the validation subsets (validation AUROC). Once the hyperparameter combination was validated, each base model was trained using the entire training set. Prediction probabilities from each validated base model were used as inputs to build the ensemble metamodel using a 100-repeated 5-fold CV. The performances of each validated model were evaluated on the unseen data of the testing set. To measure the variation of the predictive performances and obtain a robust estimate of the performances that was less impacted by the original data split, the whole process was repeated 100 times, and the overall performances were averaged over the 100 resulting validated models. The entire process is summarized in Fig 2.

The final metamodel used to predict PSQA outcomes of future plans was trained with exactly the same process but with the full dataset in order to enhance its predictive performances. Consequently, they were only estimated on the validation sets of the CV process and compared to the median validation performance obtained over the 100 data splits.

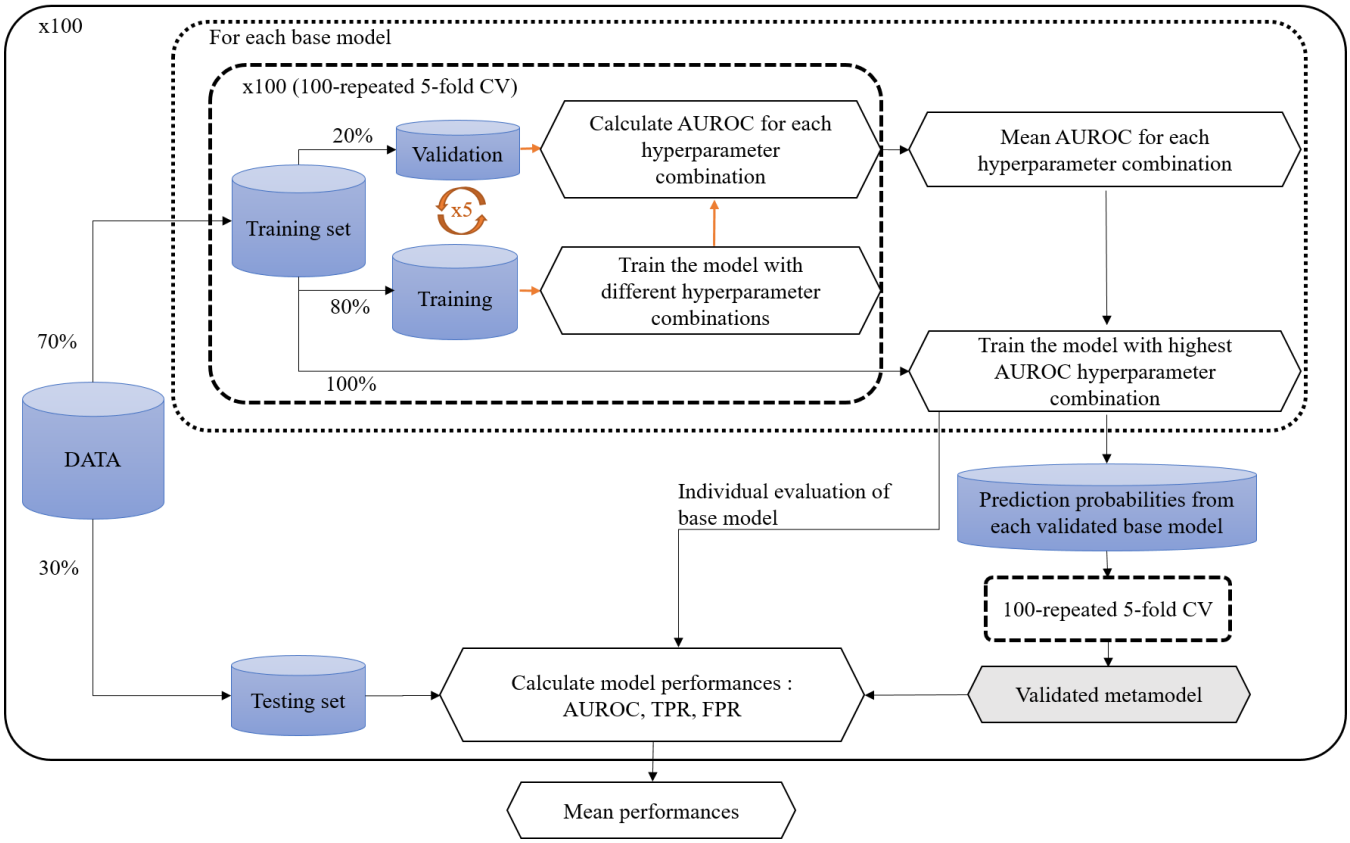


Fig 2. Metamodel construction scheme. Pre-processed original dataset was randomly split into training (70%) and testing set (30%) using stratification to ensure the original class distribution. A 100-repeated 5-fold cross-validation (CV) was performed to tune hyperparameters of each model with a random search and train the validated model. The CV performance was evaluated with the area under the receiver-operator characteristic curve (AUROC). Prediction probabilities from each base model was used as input to train the metamodel via the same 100-repeated 5-fold cross validation process. Final performances were evaluated with the AUROC, true positive rate (TPR) and false positive rate (FPR) and averaged over 100 random splits of the original dataset.

A naïve Bayes algorithm was used to train the metamodel, which assumes the independence of the features from each other (naïve aspect) so it can quickly learn to use a high number of features relative to the number of observations compared to more sophisticated methods, and requires less training data to converge. In many algorithms, features are not considered independently which is more representative of real life. However, this implies the need for a covariance matrix to estimate predicted probabilities. A small training set can then lead to a highly variable covariance matrix which can decrease the performance of the maximum likelihood estimator (defined by the cost function that maximizes the likelihood of obtaining the desired output data). A naïve Bayesian classifier does not require the covariance matrix, as its independence assumption only implies the calculation of one-dimensional variances for each predictor; thus, the maximum likelihood estimator is less affected by the problem of a small training set.

2.4.5. Model performance evaluation

The classification performance of the validated models was evaluated on the validation and testing sets using four metrics:

- AUROC (a value of 0.5 indicates that model predictions are equivalent to random predictions; a value of 1 indicates a perfect predictive model).
- True positive rate (TPR): rate of failed plans predicted failed
- False positive rate (FPR): rate of failed plans predicted passed
- AUROC difference: the difference between validation and testing AUROCs.

The AUROC was chosen as the performance score for training ML models, as it characterizes a trade between TPR (also known as "sensitivity") and FPR (also known as "1-specificity").

The AUROC difference allows for the evaluation of the generalizability of models, that is, their ability to generate predictions on unseen data as accurately as on training data. A large difference between the validation and testing AUROC indicates that the models were overfitted to the training data.

For the final classification, the default 0.5 decision threshold on the output probabilities of belonging to a class was lowered to 0.2 with the aim of reducing the FPR ratio as much as possible by maintaining a satisfactory TPR. Indeed, it is important to not re-optimize too many plans that will have passed the QA; the risk will be to overload the dosimetric workflow by optimizing the QA workflow.

An appropriate set of features was selected by building the ML models with four subsets of 5, 10, 15, and 22 (whole set) features and evaluating the performance of each ML model. The features were selected in each subset according to the RFE feature rankings.

3. Results

3.1. Dosimetric comparison

Dosimetric results from the investigated VMAT techniques are detailed in Tables 1 and 2 for RS and LS breast cancer patients, respectively. Results were averaged over the 14 patients in each group, except for the liver; only seven CT scans from the LS group and eight from the RS group included the whole liver, therefore the absorbed dose to the liver was only reported for these cases.

All VMAT techniques resulted in similar dose coverage of PTV_n and $PTV_{B/CW}$ ($D_{95\%}$, $D_{2\%}$, and D_{max}). For PTV_{IMC}, the 2AS method resulted in a significantly lower dose delivered to 95% of the volume in both patient sets.

In the LS group, a significant difference in heart sparing was observed between methods 2AS and 2A, with a decrease in D_{mean} and V_{5Gy} of 1.8 Gy and 24.3%, respectively, with 2AS. The same trend was observed for V_{5Gy} in the IL lung, with a 6% decrease in 2AS compared with 2A. Esophagus mean dose was also significantly lower with 2AS (-2 Gy compared to 2A). However, the 2AS method significantly increased the mean dose and V_{5Gy} to the contralateral (CL) breast, respectively, from 3.1 Gy and 12.8% with 2A to 4.3 Gy and 28.4% with 2AS. In the same group, the 4A method showed similar results to 2A, except in the IL lung, where significantly lower V_{20Gy}, V_{30Gy}, and D_{mean} values were obtained.

For the RS group, similar results were obtained between the VMAT techniques, except in the IL lung: V_{5Gy} increased from 68.5% with the 2A method to 75.8% with the 4A while it decreased to 62.6 % with 2AS.

Structure	Parameter	2A mean (min - max)	4A mean (min - max)	p-value 4A vs 2A	2AS mean (min - max)	p-value 2AS vs 2A
PTV_{Total}	$D_{95\%}$ (Gy)	47.3 (45.9-48.1)	47.4 (46.5-48.1)	0.612	47.3 (45.4-48.0)	0.645
	$D_{2\%}$ (Gy)	52.4 (51.4-53.1)	52.2 (51.3-53.4)	0.167	52.4 (51.2-53.2)	0.963
	D_{max} (Gy)	53.9 (52.7-54.7)	53.5 (52.7-54.4)	0.093	54.0 (52.8-55.0)	0.765
$PTV_{Breast/Wall}$	$D_{95\%}$ (Gy)	47.7 (46.4-48.5)	47.4 (46.2-48.2)	0.300	47.6 (45.1-48.9)	0.890
	$D_{2\%}$ (Gy)	52.5 (51.5-53.2)	52.2 (51.4-53.4)	0.167	52.5 (51.2-53.4)	0.982
	D_{max} (Gy)	53.7 (52.4-54.6)	53.4 (52.5-54.4)	0.154	53.9 (52.3-55.0)	0.519
PTV_{n}	$D_{95\%}$ (Gy)	47.1 (44.8-48.8)	47.7 (47.0-48.4)	0.073	47.4 (46.0-48.6)	0.519
	$D_{2\%}$ (Gy)	51.6 (50.4-52.7)	51.4 (49.8-53.2)	0.394	51.6 (50.4-52.1)	0.782
	D_{max} (Gy)	53.2 (52.1-54.6)	52.7 (50.9-54.3)	0.118	53.1 (52.4-54.4)	0.747
PTV_{IMC}	$D_{95\%}$ (Gy)	46.0 (44.4-47.0)	46.5 (45.0-48.3)	0.147	44.8 (41.5-46.5)	0.038 *
	$D_{2\%}$ (Gy)	52.6 (51.8-53.7)	52.2 (51.5-52.9)	0.038 *	52.6 (51.6-53.4)	1.000
	D_{max} (Gy)	53.5 (52.4-54.7)	53.1 (52.3-54.1)	0.069	53.5 (52.4-54.5)	0.890
Heart	D_{max} (Gy)	30.1 (15.2-38.9)	27.7 (13.7-36.2)	0.352	31.1 (13.5-40.8)	0.597
	D_{mean} (Gy)	5.4 (3.4-8.5)	5.3 (3.8-8.3)	0.854	5.1 (3.2-7.0)	0.730
	V_{5Gy} (%)	42.7 (21.0-76.2)	42.7 (25.2-76.0)	0.982	35.4 (14.9-51.4)	0.301
	V_{10Gy} (%)	10.1 (1.9-26.4)	9.6 (1.5-26.2)	0.874	11.3 (0.7-23.4)	0.535
	V_{20Gy} (%)	1.1 (0.0-4.7)	0.7 (0.0-2.6)	0.592	1.7 (0.0-4.6)	0.379
IL Lung	V_{30Gy} (%)	9.1 (6.7-14.1)	8.1 (5.4-13.1)	0.103	8.6 (6.4-14.3)	0.447
	V_{20Gy} (%)	18.2 (14.9-25.7)	16.8 (13.3-24.3)	0.118	17.4 (13.1-24.1)	0.454
	V_{10Gy} (%)	38.2 (31.4-47.6)	37.8 (30.0-51.0)	0.730	36.0 (26.0-46.1)	0.421
	V_{5Gy} (%)	68.5 (61.8-76.7)	75.8 (60.1-88.0)	0.024 *	62.6 (52.4-70.8)	0.002 *
	D_{mean} (Gy)	11.8 (10.7-14.3)	11.7 (10.3-14.6)	0.747	11.0 (9.5-13.1)	0.084
CL Lung	V_{5Gy} (%)	27.6 (13.5-53.1)	27.2 (17.2-51.4)	0.890	35.7 (20.0-55.5)	0.085
	D_{mean} (Gy)	4.2 (3.0-6.1)	4.1 (3.3-6.0)	0.747	4.6 (3.5-6.2)	0.323
Total Lung	V_{5Gy} (%)	50.7 (45.3-60.7)	54.6 (43.9-67.4)	0.069	51.0 (42.9-61.5)	0.818
	V_{20Gy} (%)	10.2 (8.5-14.3)	9.4 (7.8-13.5)	0.051	9.9 (7.8-13.5)	0.565
	V_{30Gy} (%)	5.0 (3.8-7.9)	4.5 (3.1-7.3)	0.107	4.8 (3.7-8.0)	0.205
Spinal Cord	D_{max} (Gy)	18.1 (11.2-22.6)	17.5 (12.0-21.2)	0.629	18.5 (11.1-24.6)	0.910
	D_{2cc} (Gy)	13.3 (5.8-18.2)	14.0 (9.5-16.7)	0.581	14.6 (7.8-19.3)	0.323
CL Breast	D_{max} (Gy)	14.0 (9.9-31.5)	14.0 (9.5-30.0)	0.783	16.0 (10.7-35.0)	0.135
	D_{mean} (Gy)	3.8 (2.2-5.8)	3.5 (2.0-6.0)	0.475	4.2 (2.3-7.4)	0.448
	V_{5Gy} (%)	24.6 (4.5-56.7)	19.8 (3.7-58.7)	0.421	29.1 (6.3-75.8)	0.667
Esophagus	V_{20Gy} (%)	16.5 (0.0-27.7)	13.7 (0.0-22.7)	0.250	16.0 (3.1-30.2)	0.730
	D_{mean} (Gy)	9.2 (5.9-11.9)	9.6 (6.7-11.7)	0.566	7.8 (4.9-10.8)	0.081
Thyroid	D_{mean} (Gy)	16.3 (5.3-32.5)	16.2 (5.3-35.1)	0.927	16.5 (4.7-34.2)	0.927
Liver	V_{5Gy} (%)	19.1 (10.1-28.9)	20.1 (12.5-37.8)	0.878	16.3 (10.0-20.2)	0.103

Table 1. Dosimetric comparison between 2A, 4A, and 2AS plans of 14 right-sided breast cancer patients in free breathing for 25 x 2 Gy fractionation scheme. P-value comes from a Wilcoxon rank sum test. * p-value <0.05: sample means are significantly different. Dmean is the mean dose to the volume. Dmax is the maximum dose delivered to the volume. Dx% is the dose delivered to x% of the volume. VxGy is the volume receiving at least xGy. IL: ipsilateral, CL: contralateral

Structure	Parameter	2A mean (min - max)	4A mean (min - max)	p-value 4A vs 2A	2AS mean (min - max)	p-value 2AS vs 2A
PTV_{Total}	$D_{95\%}$ (Gy)	47.7 (46.7-48.5)	47.7 (46.4-48.2)	0.982	47.5 (46.7-48.4)	0.333
	$D_{2\%}$ (Gy)	52.6 (51.8-53.0)	52.5 (51.6-53.0)	0.694	52.6 (51.8-53.9)	0.800
	D_{max} (Gy)	54.3 (53.3-54.9)	54.0 (53.4-54.8)	0.180	54.4 (53.6-56.1)	0.532
$PTV_{Breast/Wall}$	$D_{95\%}$ (Gy)	47.6 (46.4-48.8)	47.5 (45.6-48.3)	0.596	47.4 (46.0-48.3)	0.357
	$D_{2\%}$ (Gy)	52.6 (51.8-53.1)	52.6 (51.7-53.1)	0.661	52.6 (51.8-53.7)	0.982
	D_{max} (Gy)	54.0 (53.0-54.9)	54.0 (53.4-54.8)	0.890	54.2 (53.1-55.3)	0.405
PTV_{n}	$D_{95\%}$ (Gy)	48.0 (47.3-48.8)	48.2 (47.5-48.9)	0.288	48.1 (47.3-48.6)	0.661
	$D_{2\%}$ (Gy)	52.0 (51.1-52.9)	52.0 (51.0-52.7)	0.818	52.0 (50.6-53.3)	0.890
	D_{max} (Gy)	53.5 (52.3-54.5)	53.3 (52.7-54.1)	0.240	53.6 (52.1-55.1)	0.872
PTV_{IMC}	$D_{95\%}$ (Gy)	47.6 (46.1-49.0)	47.6 (46.7-49.7)	0.890	46.3 (43.5-50.9)	0.019 *
	$D_{2\%}$ (Gy)	52.8 (52.2-53.7)	52.5 (51.6-53.2)	0.116	52.8 (51.8-55.3)	0.564
	D_{max} (Gy)	53.9 (53.0-54.9)	53.4 (52.4-54.3)	0.027 *	53.9 (53.2-56.1)	0.501
Heart	D_{max} (Gy)	43.6 (31.4-52.2)	42.4 (29.9-53.9)	0.713	41.6 (23.6-54.7)	0.535
	D_{mean} (Gy)	7.0 (5.1-9.3)	6.5 (4.4-9.3)	0.434	5.2 (3.2-8.7)	0.009 *
	V_{5Gy} (%)	56.1 (33.6-80.3)	50.5 (26.0-84.8)	0.401	31.8 (15.7-52.5)	0.000 *
	V_{10Gy} (%)	17.4 (7.4-29.5)	14.0 (3.8-28.3)	0.312	13.2 (3.2-28.3)	0.135
	V_{20Gy} (%)	3.2 (0.5-7.3)	2.4 (0.2-6.1)	0.334	3.3 (0.0-11.2)	0.836
IL Lung	V_{30Gy} (%)	8.7 (4.8-12.1)	6.5 (4.2-9.2)	0.004 *	8.4 (5.9-10.2)	0.730
	V_{20Gy} (%)	18.2 (11.0-23.1)	15.6 (12.2-19.3)	0.006 *	18.2 (14.1-22.8)	0.818
	V_{10Gy} (%)	39.5 (30.1-48.1)	36.8 (31.9-40.7)	0.069	40.0 (32.7-46.7)	0.550
	V_{5Gy} (%)	70.2 (57.7-84.5)	71.7 (60.9-85.4)	0.713	64.0 (55.5-78.3)	0.006 *
	D_{mean} (Gy)	11.8 (9.7-13.2)	11.1 (10.3-12.3)	0.007 *	11.3 (10.1-12.5)	0.174
CL Lung	V_{5Gy} (%)	30.9 (18.4-50.4)	29.5 (17.5-53.9)	0.748	36.2 (23.7-60.8)	0.246
	D_{mean} (Gy)	4.5 (3.5-5.8)	4.3 (3.6-5.7)	0.320	4.6 (3.4-6.2)	0.782
Total Lung	V_{5Gy} (%)	48.3 (38.7-59.3)	48.2 (38.3-65.2)	0.982	48.6 (39.5-62.5)	0.734
	V_{20Gy} (%)	8.2 (4.5-10.7)	7.0 (5.0-8.9)	0.017 *	8.5 (5.8-10.6)	0.581
	V_{30Gy} (%)	3.9 (2.0-6.3)	2.9 (1.9-4.5)	0.006 *	3.8 (2.4-5.3)	0.765
Spinal Cord	D_{max} (Gy)	17.3 (11.6-24.2)	18.1 (14.3-22.1)	0.646	17.9 (13.7-21.7)	0.696
	D_{2cc} (Gy)	12.8 (8.5-16.6)	13.6 (10.9-16.9)	0.535	12.9 (9.6-17.2)	0.872
CL Breast	D_{max} (Gy)	13.7 (8.9-35.9)	13.2 (8.7-33.4)	0.565	16.6 (10.6-41.7)	0.190
	D_{mean} (Gy)	3.1 (2.3-4.4)	3.0 (2.0-4.9)	0.490	4.3 (2.5-10.4)	0.040 *
	V_{5Gy} (%)	12.8 (3.2-28.5)	12.7 (2.2-35.0)	0.945	28.4 (8.8-88.7)	0.016 *
Esophagus	V_{20Gy} (%)	18.5 (8.9-33.4)	19.9 (9.0-36.8)	0.581	17.8 (9.7-32.5)	0.662
	D_{mean} (Gy)	10.9 (7.3-17.2)	11.7 (7.5-18.4)	0.408	8.8 (5.8-15.8)	0.005 *
Thyroid	D_{mean} (Gy)	20.1 (0.0-32.9)	19.7 (0.0-33.4)	0.748	20.6 (0.0-33.0)	0.730
Liver	V_{5Gy} (%)	8.2 (1.9-22.3)	6.1 (1.2-16.4)	0.383	13.7 (0.0-36.6)	0.318

Table 2. Dosimetric comparison between 2A, 4A, and 2AS plans of 14 left-sided patients in free breathing for 25 x 2 Gy fractionation scheme. P-value comes from a Wilcoxon rank sum test. * p-value <0.05: sample means are significantly different. Dmean is the mean dose to the volume. Dmax is the maximum dose delivered to the volume. Dx% is the dose delivered to x% of the volume. VxGy is the volume receiving at least xGy. IL: ipsilateral, CL: contralateral

The dosimetric results published in the literature and derived from equivalent methods are listed in Table 3. When available, only results from LS breast cancer patients were extracted since LS irradiation involves higher dosimetric constraints on the heart. The 2AS results were extracted from the study by Boman et al. [9] in 11 LS patients, including three patients with deep inspiration breath hold (DIBH). The 4A results were taken from the study by Lang et al. [20] in 11 patients: four RS and seven LS DIBH patients- except for the heart for which results exclusively from LS patients were available in the study. 2A results were obtained from seven studies [1, 6, 9, 12-14, 17] that performed two-partial arc VMAT for advanced LS patients in free-breathing. Two studies [6, 13] also included RS patients in their results, and one study [9] partially used DIBH. For 2A method, published results were consistent with Table 2, except for the IL lung (+10% for V_{5Gy} and +4 Gy for Dmean) and the CL breast (+20% for V_{5Gy}). These differences can be explained by the reported greater tolerance to the dose delivered to the IL lung and in most cases the only constraint on the mean dose for the contralateral breast. Published results for 2AS were

significantly lower for the heart (-13% for V_{5Gy}) owing to the use of DIBH and a greater tolerance to the dose delivered to the CL breast (+13.8% for V_{5Gy} and +1.7 Gy for D_{mean}) allowing an easier spare of the heart. Published results for the 4A method showed a higher dose for IL lung (+12% for V_{5Gy} and +3 Gy for D_{mean}) and CL breast (+36.2% for V_{5Gy} and +3 Gy for D_{mean}), that probably led to the lower dose to the heart (-2.8 Gy for D_{mean}) associated with the use of DIBH.

		Reported results in literature		
		2A methods [1][6][9][12-14][17]	2AS method [9]	4A method [20]
Structure	Parameter	mean (min-max)	mean ± SD	mean (min-max)
IL Lung	V_{20Gy} (%)	26.2 (15.4-35.9)	28.0 ± 4.0	24.6 (22-27)
	V_{5Gy} (%)	80.3 (70.2-89.3)	65.9 ± 5.5	84.3 (71-95.6)
	D_{mean} (Gy)	14.9 (11.4-18.2)	14.4 ± 1.4	15.1 (13.4-17.1)
Heart	V_{5Gy} (%)	54.4 (39.0-83.0)	18.9 ± 11.7	-
	V_{10Gy} (%)	26.0 (20.2-35.7)	6.3 ± 5.6	-
	D_{mean} (Gy)	7.8 (4.6-10.9)	3.9 ± 1.3	3.7 (3.3-5.6)
CL Lung	V_{5Gy} (%)	27.6 (8.1-50.7)	27.7 ± 17.5	36.5 (27.3-50.7)
	D_{mean} (Gy)	3.7 (2.5-6.0)	4.1 ± 2.1	4.8 (2.7-5.8)
CL Breast	V_{5Gy} (%)	32.6 (24.4-43.5)	42.2 ± 22.9	48.9 (max 67.7)
	D_{mean} (Gy)	4.6 (2.0-8.7)	6.0 ± 3.6	6.0 (max 7.2)

Table 3: Dosimetric results extracted from published studies with equivalent method. For 2A method, studies included between 6 and 19 patients. Results were calculated from free-breathing LS breast cancer treatments except for two studies [6, 13] including LS and RS patients in their results and one study [9] partially using deep inspiration breath hold for LS treatments. For 2AS method, Boman et al. [9] results in IL lung, heart and CL lung included 8 free-breathing and 3 DIBH LS breast cancer patients. Results from 8 RS breast cancer patients were included in the calculation of the mean for the CL breast. For 4A method, Lang et al [20] included 4 RS and 7 DIBH LS cancer patients in their results except for the heart for which only LS cancer patients were included.

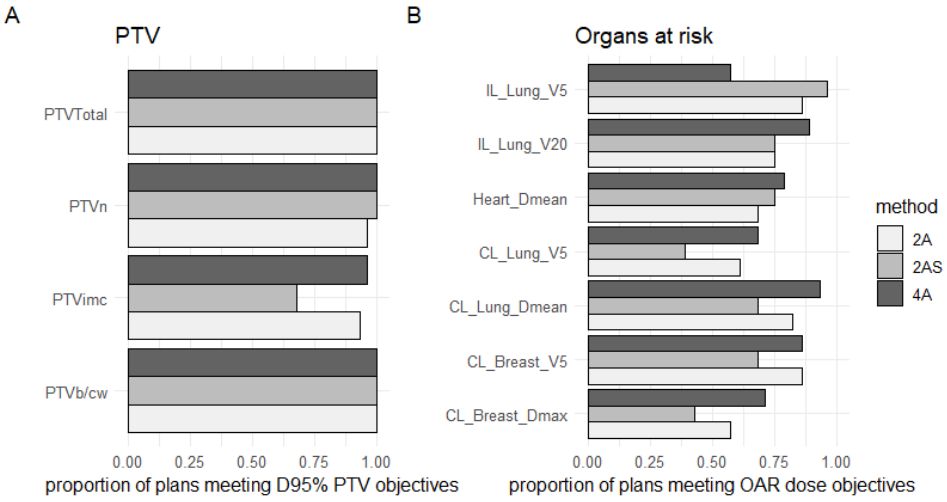


Fig 3. Proportion of plans meeting dosimetric objectives for the 28 patients for A. PTV and B. OARs.

Fig 3.A shows the proportion of plans meeting $D_{95\%}$ PTV objective >45 Gy. This objective was met by all the planning methods except for the IMC: only 68% of the 2AS plans reached the objective against more than 90% of the 2A and 4A.

Fig 3.B shows proportion of plans meeting dose limit objectives for most important OARs. The 2AS method achieved better agreement to the dose objectives for the volume of IL lung receiving at least 5 Gy, compared to

the 2A and 4A methods. 4A plans fulfilled most of the other objectives (IL lung V_{20Gy}, Heart D_{mean}, CL lung V_{5Gy} and D_{mean}, CL breast V_{5Gy} and D_{max}).

3.2. Deliverability considerations

Approximately 18% of the 2A, 12% of the 2AS and 29% of the 4A arcs did not pass the action level of >95% (Fig 4.A). Assuming that a plan is not valid if one of its arcs is out of tolerance, Fig 4.B shows the distribution of the minimum GPR obtained for each plan. Whereas 71% of the 4A arcs were within the acceptable GPR limits, this proportion fell to 21% when considering 4A plan validity (6 plans out of 28, Fig 4.C). To a lesser extent, the same trend was observed for 2A and 2AS plans with 68% of plans within the GPR tolerances (19 plans out of 28, Fig 4.C).

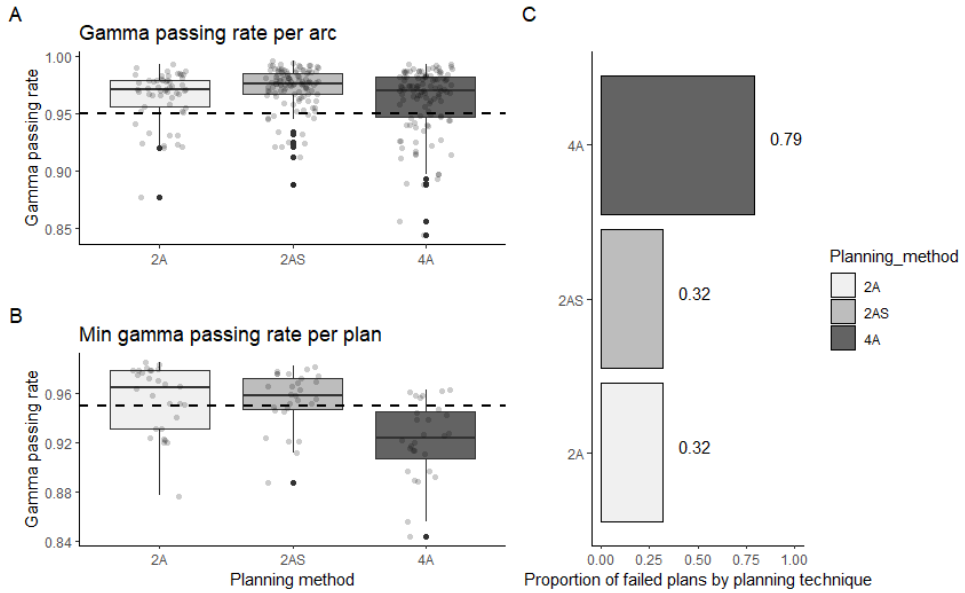


Fig 4. GPR distribution for each VMAT method. A. GPR for each arc of VMAT plans. B. Minimum GPR by plan. C. Proportion of plans that failed PSQA for each VMAT technique. Dashed line represents the 95% GPR action level.

3.3. Supervised ML classifier

3.3.1. Data pre-processing: GPR conversion to pass/fail binary class

For the 60 EPID images acquired three months apart, 75% of the GPR amplitudes obtained were below 1% (Fig 5). As a result, a GPR threshold of 96% was considered to convert plans into a pass/fail binary class.

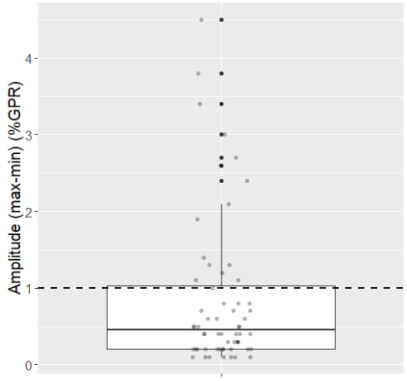
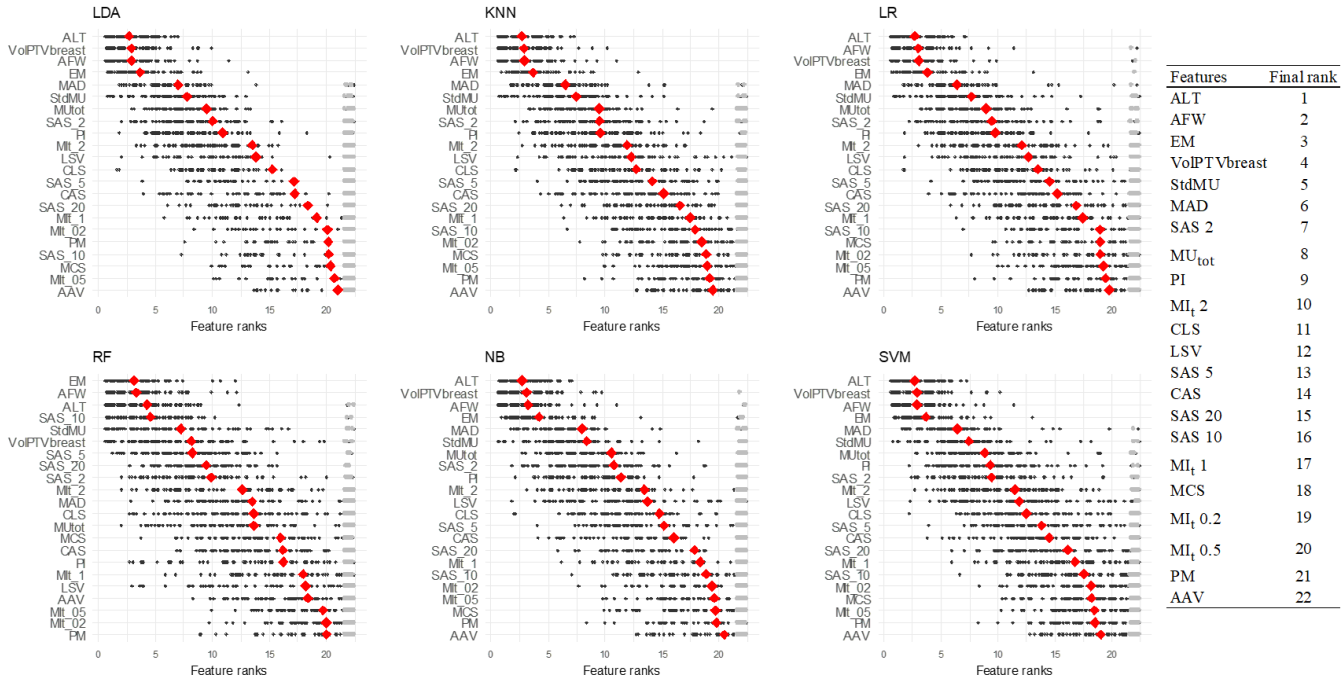


Fig 5. Amplitude of EPID response variation for 60 images acquired three months apart.

3.3.2. Feature relative importance



Features	Final rank
ALT	1
AFW	2
EM	3
VolPTVbreast	4
StdMU	5
MAD	6
SAS 2	7
MU_{tot}	8
PI	9
MI_t 2	10
CLS	11
LSV	12
SAS 5	13
CAS	14
SAS 20	15
SAS 10	16
MI_t 1	17
MCS	18
MI_t 0.2	19
MI_t 0.5	20
PM	21
AAV	22

Fig 6. Ranks of features of SVM, NB, RF, LR, KNN and LDA algorithms, from the highest, 1, to the lowest, 22, for all data splits. If for a particular split, a feature was not selected at the end of the RFE process and not ranked, then it was manually ranked 22 to account its irrelevance determined by the RFE process. Grey points represents these manually ranked features. Red points are the mean of the ranks over all splits for each feature. Final rankings of the relative importance of features were the mean of rankings over all the splits and algorithms

Fig 6 shows that the rankings of the relative importance of the features were globally in accordance with the different algorithms. The main differences were found in the RF algorithm. The RF mean ranks of the MAD and MU_{tot} metrics were 11 and 13, respectively, whereas they were 5 and 7, respectively, for all other algorithms. In addition, RF ranked SAS 10 fourth, contrary to other algorithms that ranked it sixteenth. In the final ranking, the first ranked metric characterizes the MLC modulation complexity, and the next two metrics are representative of the overall field complexity. The volume of $PTV_{b/cw}$ and standard deviation of the MU among the control points were also in the top five ranks. The last five ranked were MCS, PM, AAV, and MIt(f=0.5 and f=0.2). MCS, PM, and AAV are relative to leaf sequence and aperture area variability, and MIt is relative to the speed and acceleration of the MLC, gantry acceleration, and dose rate variation.

3.3.3. Model performance evaluation

The AUROC scores for the metamodel trained with different subsets of features reached a plateau for the subset of the 15 first-ranked features (Fig 7.A). In addition, the average AUROC difference was minimal for the 15-feature subset (Fig 7.B). It can also be noticed that standard deviation of the results was the smallest for the 15-feature subset. Consequently, the final metamodel was built with this subset.

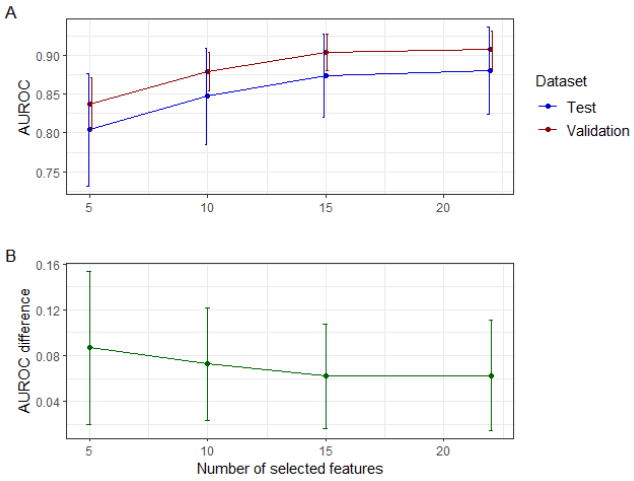


Fig 7. Validation and testing AUROC (A.) and AUROC difference (B.) for the validated metamodel trained with 5, 10, 15 or 22 features averaged over 100 data splits. Error bars represents standard deviation.

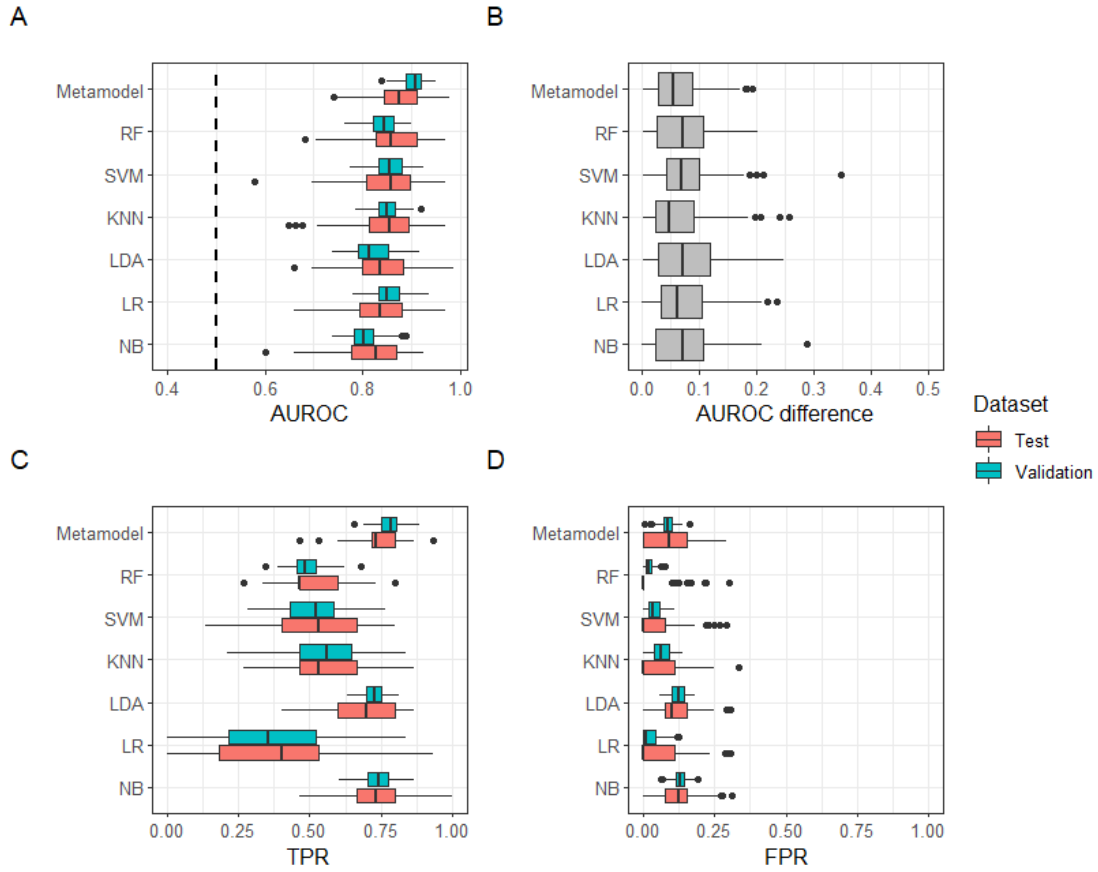


Fig 8. Classification performances of each of the validated models trained with 15 features on 100 data splits. A. Validation and testing AUROC. Dashed line is the 0.5 AUROC threshold below which classification performance is worse than random classification. B. AUROC differences between validation and testing AUROCs calculated for each split. C. True positive and D. false positive rates calculated on validation and testing set. Average ROC curve for each model are available in Appendix D.

The largest median difference between the testing and validation AUROCs (Fig 8.B) was 0.07 for the SVM, RF, NB and LDA models. The meta- and KNN models exhibited the smallest median AUROC difference of 0.05. AUROC values for all the ML models over the 100 splits were above the 0.5 threshold representative of a random guess, indicating satisfactory predictive performance (Fig 8.A). The median AUROC was above 0.8 for both the validation and testing sets for all the models. The scores on the testing sets were representative of model predictive

performance with new, unseen data. On the testing sets, metamodel showed the highest median AUROC of 0.87. The lowest value was obtained for the NB model at 0.83. The meta- and NB models exhibited the highest median TPR at 73%. The lowest median TPR scores were for the LR model at 40%. The metamodel median FPR shown in Fig 8.D, was at 9%, below NB and LDA models, respectively, at 13% and 10%, but above the four other models at 0%. The range among the testing values within each model was higher than among the validation values since validation data were already seen by the model during the CV training.

The range of interquartile variation of the performance scores over the 100 data splits was around 0.1 and 0.05 among testing and validation AUROCs, between 5% and 30% among TPR values, and between 0 and 15% among FPR values. These variations indicated that the results varied according to the data split. Consequently, only one data split will not be sufficient to fairly estimate the classification performances of ML models, as it can lead to an overly optimistic or pessimistic estimate.

	Metamodel	
	trained on 70% of the data	trained on 100% of the data
AUROC	0.908	0.910
FPR	0.783	0.785
TPR	0.086	0.074

Table 4. Median performance scores of metamodel trained on 70% of the dataset over the 100 data splits and performance scores of the metamodel trained on 100% of the dataset.

The validation AUROC and TPR of the final metamodel (trained with the full dataset) were equivalent to the median scores of the 100 metamodels trained on 70% of the dataset (Table 4). The TPR was smaller, at 7.4% versus 8.6%. These results could indicate that predictive performances of the final metamodel on new data is likely to be equivalent or better (for FPR) than observed median scores obtained on the testing sets over the 100 data splits.

4. Discussion

The aim of this study was to implement in clinical routine, the VMAT technique for advanced breast cancers, by selecting a relevant planning method and preventing QA workflow perturbation owing to the introduction of a technique, potentially affecting 30% of the annual pool of breast patients. Three VMAT treatment methods for advanced breast cancer with IMC involvement were evaluated. A cohort of 28 patients (14 LS and 14 RS patients) was planned for free-breathing using each method. Several studies have demonstrated the benefit of including IMC in target volumes to improve the overall survival of patients with node-positive breast cancer [47-49]. Compared with 3D-CRT and static IMRT techniques, VMAT techniques have been shown to achieve better target coverage with greater cardiac and pulmonary sparing, especially when IMC is included [1,4,22,50,51]. Nevertheless, cardiac and pulmonary toxicities in patients with breast cancer remain an important issue because they have been shown to be related to heart and lung exposure to radiotherapy [52-57]. In this study, the 2AS method achieved the lowest dose on the heart, but at the cost of an increased dose delivered to the CL breast and a significantly lower IMC coverage. This result was in agreement with the study by Boman et al. [9] for OARs but not for the IMC coverage because the IMC was not distinguished from other targets in their study. The advantage of the 4AS method was the highest target dose coverage while reasonably sparing the OARs compared to the other two methods. The 2A method led to acceptable target coverage and best achieved OAR dose objectives after the 4A method. The dosimetric comparison was conducted in free-breathing. Deep inspiration breath-hold techniques would improve heart sparing, in particular, for patients with LS cancer. However, for a proportion of patients, holding a blocked inspiration is not an option, and for these patients, it remains interesting to propose alternative techniques that make it possible to best achieve individual dosimetric constraints.

In addition to the dosimetric results, this study focused on the clinical deliverability of the plans. The PSQA investigation showed that the 4A plans exhibited the poorest QA results despite their promising dosimetric results, indicating that caution should be taken when selecting the 4A method. The increased proportion fails in the 4A plans compared to 2A and 2AS can be explained by a significantly higher value of modulation units delivered per plan, and a higher value of SAS, MAD, and PI metrics (Fig 9), representative of aperture complexity. In addition, CAS and AFW values were significantly lower for the 4A method (Fig 9), synonymous with greater asymmetry

of the MLC. These observations are in accordance with the published results of Li et al. [46] and Crowe et al. [41] based on the same complexity metrics.

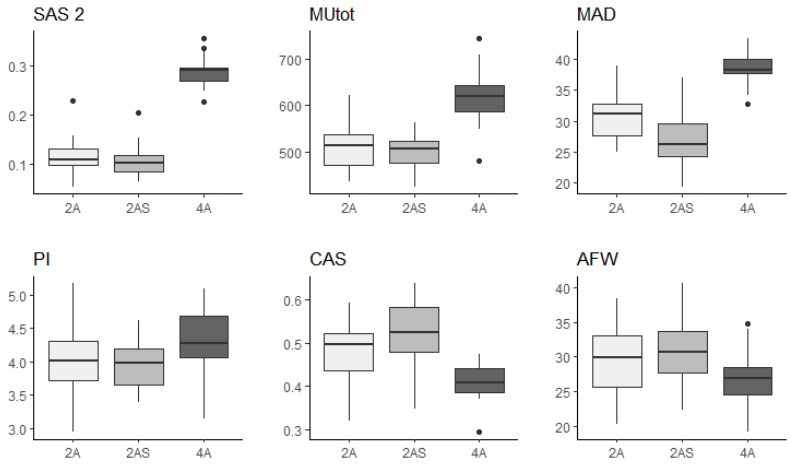


Fig 9. Distribution by planning methods of PCMs for which a significant difference (Wilcoxon test) was found between the 4A planning method and the other two methods (2A and 2AS).

The mathematical introduction of the VMAT technique for breast cancer involves an increase in the QA workload. In addition, VMAT techniques are also more sensitive than 3D-CRT techniques to anatomical changes in patients during treatment owing to beam modulation [39, 58]. During the course of breast cancer radiotherapy, significant changes in breast anatomy can be observed, in particular, patients who have undergone breast-conserving surgery. The dosimetric margin added outside the breast surface could not be sufficient to compensate for the anatomical change, and a re-evaluation of the VMAT plan followed by a new QA measurement could be required. Therefore, with the implementation of the VMAT technique in routine clinical practice for breast cancers, there is an additional risk of increased QA workload. In this context, a lever of action is to prevent QA process overload by predicting plans that are likely to fail. Building a ML model was a necessary step to reach robust predictions of QA outcomes, since a direct relationship between PCMs and QA outcomes would have led to a significantly lower prediction accuracy, as shown in Appendix E. ML techniques have been increasingly developed in recent years, in the field of medical physics in radiation therapy, to optimize the dosimetric workflow [59-61]. Valdes et al. [33] were among the first to publish a study on PSQA prediction based on an ML algorithm in 2017. They used a Poisson regression with Lasso regularization to predict GPR values from diode array detector and EPID measurements, with 3% and 3.5% accuracy, respectively. After their study, several ML methods were proposed to predict the IMRT and VMAT plan QA outcomes. They are based on convolutional networks [34,35], SVM [28,36], decision trees [30, 31,35], linear regression [35,30] or an association of two algorithms to combine the regression and classification methods [37]. Ensemble strategies were applied for example by Interian et al [62], Tomori et al. [34] or Li et al. [30] to improve the prediction accuracy by decreasing the variance of the models. These strategies relied on ensembled ML models, based on the same algorithm, but tuned with different hyperparameters. In this study, the originality of the ML classifier of VMAT plans as ‘pass’ or ‘fail’ was to rely on the predictive performance diversity from the different ML algorithms. The metamodel was built on a stack of ML models trained with different algorithms, but tuned with no more than two hyperparameters to limit their complexity, and reduce the generalization error. The relative importance of PCMs was ranked prior to ML training to limit the complexity of the model and improve its generalization performance. It highlighted that the five metrics that came first had a low degree of complexity, characterizing the variation in MLC aperture or MU and PTV volume. More complex PCMs, such as MIt and MCS, were ranked near the end. Metamodel performance evaluation over different subsets of these features favored the 15 first-ranked metrics. The predictive performances evaluated over the 100 testing sets showed that the metamodel significantly outperformed the base classifiers’ individual performances with the highest median AUROC and sensitivity (respectively at 87% and 73%), by keeping the specificity in a reasonably high level (91%). The base classifiers for which sensitivity was the highest

were those with the lowest sensibility, and vice versa. Their combination in the metamodel improved the overall predictions based on a complementary set of individual predictions. Predictive performances were expected to be better on the validation data as ML models already seen this data during the CV process whereas they never seen the testing data, but the difference between validation and testing performances should be kept reasonable. For the base and meta- models, testing AUROCs were sufficiently consistent with validation AUROCs to estimate that the generalization performances from training to testing were satisfactory. The AUROC differences were the smallest for the metamodel, as well as the variation of the testing TPR and AUROC values over the 100 splits. These results demonstrate that the metamodel reduced the variance and achieved more robust classification results by improving generalization performance. The final metamodel that will be used for the PSQA prediction was trained with the full dataset. The estimated predictive performances on the cross-validation sets were similar to that estimated on the 100 training splits. These results indicate that the estimated predictive performance over the 100 testing splits could be a fair estimate of the final predictive performance of the model on future unseen data. Despite these encouraging results, the proposed metamodel has limitations. One limitation is directly related to the small sample size of plans with which it was trained in the context of VMAT implementation for breast localization. Even though efforts have been devoted to reducing generalization error through feature selection, use of simple ML models, repeated cross-validation and stacking of ML models via a naïve Bayesian algorithm, generalization performance is expected to improve as a larger sample of training becomes available; that is, additional QA was carried out. The second limitation is the performance dependence of the selected features designed by domain experts. To overcome this limitation, some studies have recently paved the way for a fully automated process using radiomics features automatically extracted from gamma images [63, 64] or fluence maps [62] without human expert supervision.

5. Conclusion

Prior to the clinical implementation of VMAT for advanced breast cancer in our institution, this study provided a comprehensive process for obtaining VMAT plans meeting the dosimetric objectives, clinically deliverable by the treatment machines, and proposed a way to anticipate the increase in QA workload. The three investigated VMAT methods led to significant differences in the dose distribution. The 2A method could be suggested as a dosimetric basis because it allows a reasonable sparing of OARs while covering the targets in an acceptable manner. In the event that the heart is of particular importance for patients with no DIBH option, the 2AS method leaves open the possibility of achieving better heart-sparing, bearing in mind that this comes at the cost of a lower target coverage and the highest contralateral breast irradiation. 4A method should be selected with caution since 4A plans exhibited higher MLC modulation and field asymmetry, making them poorly reproducible by the treatment machine in the context of this study. The proposed ML metamodel, with 73% sensitivity and 91% specificity, proved to be a promising tool for classifying PSQA results to prevent the measurement of treatment plans that are likely to fail QA. As a larger dataset becomes available, the performance of the metamodel can be improved. Future investigations will focus on automated radiomic characteristic extraction as features for the ML model to achieve a fully automated model and overcome the dependence of the model on the selected PCMs.

Declaration of Competing Interest

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

Appendix

A*: Dosimetric objectives used for each plan optimization. IL = ipsilateral. CL = contralateral.*

Volume	Dosimetric objective
PTVs	$D_{95\%} > 45Gy$
	$D_{2\%} < 53.5Gy$
	$D_{max} < 55Gy$
Heart	$D_{mean} < 8Gy$ (L)
	$D_{mean} < 6Gy$ (R)
IL Lung	$V_{20Gy} < 20\%$
	$V_{10Gy} < 50\%$
	$V_{30Gy} < 10\%$
	$V_{5Gy} < 75\%$
	$D_{mean} < 13Gy$
CL Lung	$D_{mean} < 5Gy$
	$V_{5Gy} < 30\%$
Total Lung	$V_{5Gy} < 50\%$
	$V_{20Gy} < 10\%$
	$V_{30Gy} < 5\%$
Spinal Cord	$D_{2cc} < 20Gy$
CL Breast	$D_{mean} < 4Gy$
	$D_{max} < 13Gy$
	$V_{5Gy} < 30\%$
Esophagus	$D_{mean} < 10Gy$
	$V_{20Gy} < 20\%$
Thyroid	$D_{mean} < 30Gy$
Liver	$V_{5Gy} < 20\%$

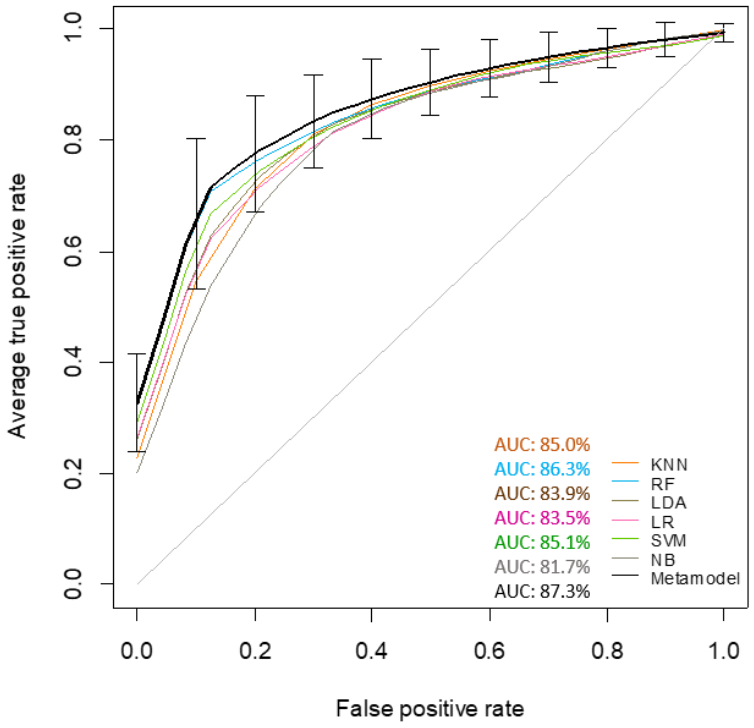
B: Summary of the calculated plan complexity metrics (PCMs) . Metrics were averaged over the control points (every 2°) and over the beams according to their respective weights.

Abbrev	Metric	Ref
MU_{tot}	Total number of monitor units of the plan	-
StdMU	Standard deviation of monitor units delivered by control point	-
Vol_{PTV}	Volume of the $PTV_{B/CW}$	-
MI_t 0.2	Total modulation index with f = 0.2	44
MI_t 0.5	Total modulation index with f = 0.5	44
MI_t 1	Total modulation index with f = 1	44
MI_t 2	Total modulation index with f = 2	44
MCS	Modulation complexity score	42
EM	Edge metric	43
SAS 2	Small aperture score 2 mm	41
SAS 5	Small aperture score 5 mm	41
SAS 10	Small aperture score 10 mm	41
SAS 20	Small aperture score 20 mm	41
LSV	Leaf sequence variability	42
MAD	Mean asymmetry distance	41
PI	Plan averaged beam irregularity	45
PM	Plan averaged beam modulation	45
CAS	Cross-axis score	41
AAV	Aperture area variability	42
CLS	Closed leaf score	41
AFW	Average field width	46
ALT	Average leaf travel	46

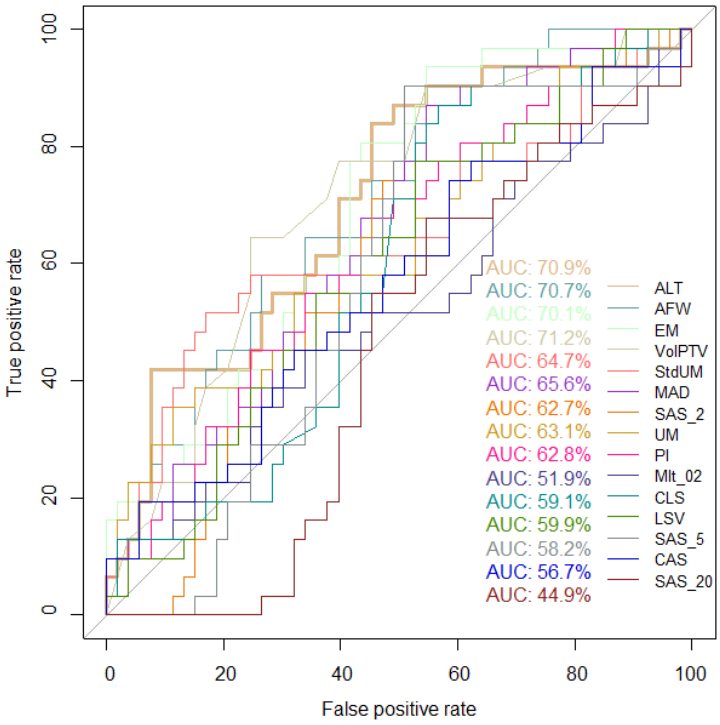
C: Algorithms used in base models and tuned hyperparameters during the cross-validation process. Abbreviations: KNN: K nearest neighbors, LDA : linear discriminant analysis, LR: regularized logistic regression, NB: naïve Bayes, SVM: support vector machines, RF: random forest.

Algorithm	Description	Tuned hyperparameters
KNN	Find k samples that have similar features and assign the observation to the class to which the majority of its neighbors belong.	number of neighbors considered in the nearest neighbor calculation
LDA	Calculates a linear discriminant function using a covariance matrix to maximize the scatter between class and minimize it within class.	none
LR	Binary regression based on sigmoid function with regularization.	cost penalty for misclassifications (regularization strength) and regularization type (Ridge regression or Lasso)
NB	Uses the prior probability of the classes (based on class distribution) to estimate the posterior probability of an observation to belong to the class given the predictor information.	kernel used to calculate the conditional probability (a kernel density estimation or a gaussian distribution)
SVM	Identifies the optimal non-linear hyperplane to separate classes by mapping the original space into an higher dimensionnal space to facilitate the class separation using a kernel function.	sigma (parameter that controls the SVM decision boundary) and cost penalty for misclassifications (regularization strength)
RF	Fits many decision trees to random subsamples of the data and features and averages the trees.	number of randomly selected features at each cut in the tree

D: ROC curves for each base model and the metamodel, averaged over all testing splits. Error bars represent standard deviation of the true positive rate, only displayed for the metamodel. Abbreviations: AUC: area under the curve, KNN: K nearest neighbors, LDA : linear discriminant analysis, LR: regularized logistic regression, NB: naïve Bayes, SVM: support vector machines, RF: random forest.



E: ROC curves generated for each of the 15 first-ranked complexity metrics. The area under the curve (AUC) quantifies the ability of the metric to distinguish between failed and successful QA (defined with the GPR threshold at 96%). The AUC for each metric taken independently was more than 10% lower than the AUC obtained with the ML models.



References

[1] Popescu CC, Olivotto IA, Beckham WA, et al. Volumetric modulated arc therapy improves dosimetry and reduces treatment time compared to conventional intensity-modulated radiotherapy for locoregional radiotherapy of left-sided breast cancer and internal mammary nodes. Int. J. Radiation Oncology Biol. Phys. 2010;76(1):287-295. https://doi.org/10.1016/j.ijrobp.2009.05.038

[2] Jo IY, Kim ES, Kim WC, Min, Kee C, Yeo SG. Dosimetric comparison of incidental axillary irradiation between three-dimensional conformal and volumetric modulated arc techniques for breast cancer. Molecular and Clinical Oncology 2020;12:551-556.

[3] Johansen S, Cozzi L, Olsen DR. A planning comparison of dose patterns in organs at risk and predicted risk for radiation induced malignancy in the contralateral breast following radiation therapy of primary breast using conventional IMRT and volumetric modulated arc treatment technique. Acta Oncologica 2009;48(4):495-503. https://doi.org/10.1080/02841860802657227

[4] Koivumäki T, Fogliata A, Zeverino M, et al. Dosimetric evaluation of modern radiation therapy techniques for left breast in deep-inspiration breath-hold. Physica Medica 2018;45:82-87. https://doi.org/10.1016/j.ejmp.2017.12.009

[5] Tyran M, Mailleux H, Tallet A, et al. Volumetric-modulated arc therapy for left-sided breast cancer and all regional nodes improves target volumes coverage and reduces treatment time and doses to the heart and left coronary artery, compared with a field-in-field technique. Journal of Radiation Research 2015;56(6):927-937. https://doi.org/10.1093/jrr/rrv052

[6] Dumane VA, Bakst R, Green S. Dose to organs in the supraclavicular region when covering the Internal Mammary Nodes (IMNs) in breast cancer patients: A comparison of Volumetric Modulated Arcs Therapy (VMAT) versus 3D and VMAT. PLoS ONE 2018;13(10). https://doi.org/10.1371/journal.pone.0205770

[7] Jensen CA, Acosta Roa AM, Johansen M, Lund JA, Frengen J. Robustness of VMAT and 3DCRT plans toward setup errors in radiation therapy of locally advanced left-sided breast cancer with DIBH. Physica Medica 2018;45:12-18. https://doi.org/10.1016/j.ejmp.2017.11.019
[8] Osman SOS, Hol S, Poortmans PM, Essers M. Volumetric modulated arc therapy and breath-hold in image-guided locoregional left-sided breast irradiation. Radiotherapy and Oncology 2014;112:17-22. https://doi.org/10.1016/j.radonc.2014.04.004
[9] Boman E, Rossi M, Haltamo M, Skytta T, Kapanen M. A new split arc VMAT technique for lymph node positive breast cancer. Physica Medica 2016;32:1428-1436. https://doi.org/10.1016/j.ejmp.2016.10.012
[10] Fogliata A, Seppala J, Reggiori GLF, et al. Dosimetric trade-offs in breast treatment with VMAT technique. BR J Radiol 2017;90. https://doi.org/10.1259/bjr.20160701
[11] Dunlop A, Colgan R, Kirby A, Ranger A, Blasiak-Wal I. Evaluation of organ motion-based robust optimisation for VMAT planning for breast and internal mammary chain radiotherapy. Clinical and Translational Radiation Oncology 2019;16:60-66. https://doi.org/10.1016/j.ctro.2019.04.004
[12] Pham TT, Ward R, Latty D, et al. Left-sided breast cancer loco-regional radiotherapy with deep inspiration breath-hold: Does volumetric-modulated arc radiotherapy reduce heart dose further compared with tangential intensity-modulated radiotherapy ? Journal of Medical Imaging and Radiation Oncology 2016;60:545-553. https://doi.org/10.1111/1754-9485.12459
[13] Nicolini G, Fogliata A, Clivio A, Vanetti E, Cozzi L. Planning strategies in volumetric modulated arc therapy. Medical Physics 2011;38 (7):4025-4030. https://doi.org/10.1118/1.3598442
[14] Tyran M, Tallet A, Resbeut M, et al. Safety and benefit of using a virtual bolus during treatment planning for breast cancer treated with arc therapy. Radiation Oncology Physics 2018;19:463-472. https://doi.org/10.1002/acm2.12398
[15] Liao X, Wu F, Wu J, et al. Impact of positioning errors in the dosimetry of VMAT left-sided post mastectomy irradiation. Radiation Oncology 2020;15(103).
[16] van der Veen GJ, Janssen T, Duijn A, et al. A robust volumetric arc therapy planning approach for breast cancer involving the axillary nodes. Medical Dosimetry 2019;44(2):183-189. https://doi.org/10.1016/j.meddos.2018.06.001
[17] Zhang R, Heins D, Sanders M, Guo B, Hogstrom K. Evaluation of a mixed beam therapy for postmastectomy breast cancer patients: Bolus electron conformal therapy combined with intensity modulated photon radiotherapy and volumetric modulated photon arc therapy. Medical Physics 2018;45(7):2912-2924. https://doi.org/10.1002/mp.12958
[18] Zhang W, Ruisheng L, You D, Su Y, Dong W, Ma Z. Dosimetry and Feasibility Studies of Volumetric Modulated Arc Therapy With Deep Inspiration Breath-Hold Using Optical Surface Management System for Left-Sided Breast Cancer Patients. Front. Oncol. 2020. https://doi.org/10.3389/fonc.2020.01711
[19] Kuo L, Ballangrud AM, Ho AY, Mechalakos JG, Li G, Hong L. A VMAT planning technique for locally advanced breast cancer patients with expander or implant reconstructions requiring comprehensive postmastectomy radiation therapy. Medical Dosimetry 2019;44(2):150-154. https://doi.org/10.1016/j.meddos.2018.04.006
[20] Lang K, Loritz BSA, Hunzeker A, Lenards N, Culp L, Finley R. Dosimetric comparison between volumetric-modulated arc therapy and a hybrid volumetric-modulated arc therapy and segmented field-in-field technique for postmastectomy chest wall and regional lymph node irradiation. Medical Dosimetry 2020;45(2):121-127. https://doi.org/10.1016/j.meddos.2019.08.001
[21] De Rose F, Fogliata A, Franceschini D, et al. Postmastectomy radiation therapy using VMAT technique for breast cancer patients with expander reconstruction. Medical Oncology 2019;36(48). https://doi.org/10.1007/s12032-019-1275-z
[22] Pasler M, Georg D, Bartelt S, Lutterbach J. Node-positive left-sided breast cancer: does VMAT improve treatment plan quality with respect to IMRT ? Strahlenther Onkol 2013;5:380-386.
[23] Zhao H, He M, Cheng G, et al. A comparative dosimetric study of left sided breast cancer after breast-conserving surgery treated with VMAT and IMRT. Radiation Oncology 2015;10(231). https://doi.org/10.1186/s13014-015-0531-4
[24] Moran JM, Dempsey M, Eisbruch A, et al. Safety considerations for IMRT: Executive summary. Practical Radiation Oncology 2011;1:190-195. https://doi.org/10.1016/j.prro.2011.04.008
[25] Miften M, Olch A, Mihailidis D, et al. Tolerance limits and methodologies for IMRT measurement-based verification QA: Recommendations of AAPM Task Group No. 218 2018;45(4).

[26] 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 2019;92(20190270). https://doi.org/10.1259/bjr.20190270
[27] Antoine M, Ralite F, Soustiel C, et al. Use of metrics to quantify IMRT and VMAT treatment plan complexity: A systematic review and perspectives. Physica Medica 2019;64:98-108. https://doi.org/10.1016/j.ejmp.2019.05.024
[28] 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 2021;87:136-143. https://doi.org/10.1016/j.ejmp.2021.03.017
[29] Chan MF, Witztum A, Valder G. Integration of AI and Machine Learning in Radiotherapy QA. Front. Artif. Intell. 2020;3:577620. https://doi.org/10.3389/frai.2020.577620
[30] Li J, Wang L, Zhang X, et al. Machine Learning for Patient-Specific Quality Assurance of VMAT: Prediction and Classification Accuracy. Int J Radiat Oncol Biol Phys 2019;105(4):893-902. https://doi.org/10.1016/j.ijrobp.2019.07.049
[31] Lam D, Zhang X, Li H, et al. Predicting gamma passing rates for portal dosimetry-based IMRT QA using machine learning. Medical Physics 2019;46(10):4666-4675. https://doi.org/10.1002/mp.13752
[32] El Naqa I, Irrer J, Ritter TA, et al. Machine learning for automated quality assurance in radiotherapy: A proof of principle using EPID data description. Medical Physics 2019;46(4):1914-1921. https://doi.org/10.1002/mp.13433
[33] Valdes G, Scheuermann R, Hung CY, Olszanski A, Bellerive M, Solberg TD. A mathematical framework for virtual IMRT QA using machine learning. Medical Physics 2016;43(7):4323-4334. https://doi.org/10.1118/1.4953835
[34] Tomori S, Kadoya N, Takayama Y, et al. A deep learning-based prediction model for gamma evaluation in patient-specific quality assurance. Medical Physics 2018;45(9):4055-4065. https://doi.org/10.1002/mp.13112
[35] Ono T, Hirashima H, Iramina H, et al. Prediction of dosimetric accuracy for VMAT plans using plan complexity parameters via machine learning. Med Phys 2019;46(9). https://doi.org/10.1002/mp.13669
[36] 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 2019;64(9). https://doi.org/10.1088/1361-6560/ab142e
[37] Yang R, Yang X, Wang L, et al. Commissioning and clinical implementation of an Autoencoder based Classification-Regression model for VMAT patient-specific QA in a multi-institution scenario. Radiotherapy and Oncology 2021;161:230-240. https://doi.org/10.1016/j.radonc.2021.06.024
[38] Lizondo M, Latorre-Musoll ARM, Carrasco P, Espinosa N, Coral A, Jornet N. Pseudo skin flash on VMAT in breast radiotherapy: Optimization of virtual bolus thickness and HU values. Physica Medica 2018;63:56-62. https://doi.org/10.1016/j.ejmp.2019.05.010
[39] Rossi M, Boman A, Kapanen M. Optimal selection of optimization bolus thickness in planning of VMAT breast radiotherapy treatments. Medical Dosimetry 2019;44(3):266-273. https://doi.org/10.1016/j.meddos.2018.10.001
[40] Heilemann G, Poppe B, Laub W. On the sensitivity of common gamma-index evaluation methods to MLC misalignments in Rapidarc quality assurance. Medical Physics 2013;40(3). 10.1118/1.4789580.
[41] Crowe SB, Kairn T, Kenny J, et al. Treatment plan complexity metrics for predicting IMRT pre-treatment quality assurance results. Australasian Physical & Engineering Sciences in Medicine 2014;37:475-482. https://doi.org/10.1007/s13246-014-0274-9
[42] McNiven AL, Sharpe MB, Purdie TG. A new metric for assessing IMRT modulation complexity and plan deliverability. Medical Physics 2010;37(2):505-515. https://doi.org/10.1118/1.3276775
[43] Younge KC, Matuszak MM, Moran JM, McShan DL. Penalization of aperture complexity in inversely planned volumetric modulated arc therapy. Medical Physics 2012;39(11):7160-7170. https://doi.org/10.1118/1.4762566
[44] Park JM, Park SY, Kim H, Kim JJ, Carlson J, Ye SJ. Modulation indices for volumetric modulated arc therapy. Phys. Med. Biol. 2014;59:7315-7340. https://doi.org/10.1088/0031-9155/59/23/7315
[45] Du W, CHo HS, Zhang X, Hoffman KE, Rajat K. Quantification of beam complexity in intensity-modulated radiation therapy treatment plans. Med Phys 2014;41(2). . https://doi.org/10.1118/1.4861821

[46] Li G, Wu K, Peng G, Zhang Y, Bai S. A retrospective analysis for patient-specific quality assurance of volumetric-modulated arctherapy plans. Medical Dosimetry 2014;39:309-313. https://doi.org/10.1016/j.meddos.2014.05.003
[47] Thorsen LBJ, Offersen BV, Danø H, et al. DBCG-IMN: A Population-Based Cohort Study on the Effect of Internal Mammary Node Irradiation in Early Node-Positive Breast Cancer. Journal of Clinical Oncology 2016;34(4):314-320. https://doi.org/10.1200/JCO.2015.63.6456
[48] Poortmans PM, Weltens C, Fortpied C, et al. Internal mammary and medial supraclavicular lymph node chain irradiation in stage I–III breast cancer (EORTC 22922/10925): 15-year results of a randomised, phase 3 trial. The Lancet Oncology 2020;21(12):1602-1610. https://doi.org/10.1016/S1470-2045(20)30472-1
[49] Borm KJ, Simonetto C, Kundrat P, et al. Toxicity of internal mammary irradiation in breast cancer. Are concerns still justified in times of modern treatment techniques? Acta Oncologica 2020;59(10):1201-1209. https://doi.org/10.1080/0284186X.2020.1787509
[50] Ranger A, Dunlop A, Hutchinson K, et al. A Dosimetric Comparison of Breast Radiotherapy Techniques to Treat Locoregional Lymph Nodes Including the Internal Mammary Chain. Clinical Oncology 2018;30(6):346-353. https://doi.org/10.1016/j.clon.2018.01.017
[51] Zhang Q, Yu XL, Hu WG, et al. Dosimetric comparison for volumetric modulated arc therapy and intensity-modulated radiotherapy on the left-sided chest wall and internal mammary nodes irradiation in treating post-mastectomy breast cancer. 2015;49(1):91-98. https://doi.org/10.2478/raon-2014-0033
[52] Darby SC, Ewertz M, McGale P, et al. Risk of Ischemic Heart Disease in Women after Radiotherapy for Breast Cancer. The New England Journal of Medicine 2013;368(11). https://doi.org/10.1056/NEJMoa1209825
[53] Schneider U, Ernst M, Hartmann M. The dose-response relationship for cardiovascular disease is not necessarily linear. Radiation Oncology 2017;12(74). https://doi.org/10.1186/s13014-017-0811-2
[54] van den Bogaard VAB, van Luijk P, Hummel YM, et al. Cardiac function after radiotherapy for breast cancer. Int J Radiation Oncol Biol Phys 2019;104(2):392-400. https://doi.org/10.1016/j.ijrobp.2019.02.003
[55] Skytta T, Tuohinen S, Luukkaala T, Virtanen V, Raatikainen P, Kellokumpu-Lehtinen PL. Adjuvant radiotherapy-induced cardiac changes among patients with early breast cancer: a three-year follow-up study. Acta Oncol 2019;58(9):1250-1258. https://doi.org/10.1080/0284186X.2019.1630751
[56] Goldman UB, Svane G, Anderson M, Wennberg B, Lind P. Long-term functional and radiological pulmonary changes after radiation therapy for breast cancer. Acta Oncol 2014;53(10):1373-1379. https://doi.org/10.3109/0284186X.2014.934967
[57] Erven K, Weltens C, Nackaerts K, Fieuws S, Decramer M, Lievens Y. Changes in pulmonary function up to 10 years after locoregional breast irradiation. Int J Radiat Oncol Biol Phys 2012;82(2):701-707. https://doi.org/10.1016/j.ijrobp.2010.12.058
[58] Seppälä J, Vuolukka K, Viren T, et al. Breast deformation during the course of radiotherapy: The need for an additional outer margin. Physica Medica 2019;65:1-5. https://doi.org/10.1016/j.ejmp.2019.07.021
[59] Bosmans H, Zanca F, Gelaude F. Procurement, commissioning and QA of AI based solutions: An MPE's perspective on introducing AI in clinical practice. Physica Medica 2021;83:257-263. https://doi.org/10.1016/j.ejmp.2021.04.006.
[60] Harrer C, Ullrich W, Wilkens JJ. Prediction of multi-criteria optimization (MCO) parameter efficiency in volumetric modulated arc therapy (VMAT) treatment planning using machine learning (ML). Physica Medica 2021;81:102-113. https://doi.org/10.1016/j.ejmp.2020.12.004.
[61] Olaciregui-Ruiz I, Torres-Xirau I, Teuwen J, van der Heide UA. A Deep Learning-based correction to EPID dosimetry for attenuation and scatter in the Unity MR-Linac system. Physica Medica 2020;71:124-131. https://doi.org/10.1016/j.ejmp.2020.02.020.
[62] Interian Y, Rideout V, Kearney VP, et al. Deep Nets vs Expert Designed Features in Medical Physics: An IMRT QA case study. Medical Physics 2018;45(6):2672-2680. https://doi.org/10.1002/mp.12890

[63] Lizar JC, Yaly CC, Bruno AC, Viani GA, Pavoni JF. Patient-specific IMRT QA verification using machine learning and gamma radiomics. Physica Medica 2021;82:100-108. https://doi.org/10.1016/j.ejmp.2021.01.071
[64] Wootton LS, Nyflot MJ, Chaovalitwongse WA, Ford E. Error Detection in Intensity-Modulated Radiation Therapy Quality Assurance Using Radiomic Analysis of Gamma Distributions. Radiation Oncology Biology Physics 2018;102(1):219-228. https://doi.org/10.1016/j.ijrobp.2018.05.033