Physics in Medicine and Biology

Predicting VMAT patient-specific QA results using a support vector classifier trained on treatment plan characteristics and linac QC metrics

To cite this article before publication: Dal Alexander Granville *et al* 2019 *Phys. Med. Biol.* in press https://doi.org/10.1088/1361-6560/ab142e

Predicting VMAT patient-specific QA results using a support vector classifier trained on treatment plan characteristics and linac QC metrics

Dal A. Granville[1], Justin G. Sutherland[1,2,3], Jason G. Belec[1,2], and Daniel J. La Russa[1,2,3]

[1]Radiation Medicine Program, The Ottawa Hospital, Ottawa, Canada

[2]Department of Radiology, Division of Radiation Oncology, The University of Ottawa, Ottawa, Canada

[3]The Ottawa Hospital Research Institute, Ottawa, Canada

Corresponding author: Dal A. Granville, Radiation Medicine Program, The Ottawa Hospital, 501 Smyth Rd, Box 927 Ottawa, Canada, K1H 8L6, Email: dgranville@toh.ca

Short title: Predicting VMAT QA results using machine learning

Keywords: VMAT, patient-specific QA, machine learning, linac QC

Abstract

The use of treatment plan characteristics to predict patient-specific quality assurance (QA) measurement results has recently been reported as a strategy to help facilitate automated pre-treatment verification workflows or to provide a virtual assessment of delivery quality. The goal of this work is to investigate the potential of using treatment plan characteristics and linac performance metrics (i.e., quality control test results) in combination with machine learning techniques to predict the results of VMAT patient-specific QA measurements.

Using features that describe treatment plan complexity and linac performance metrics, we trained a linear support vector classifier (SVC) to classify the results of VMAT patient-specific QA measurements. The 'targets' in this model were simple classes representing median dose difference between measured and expected dose distributions – 'hot' if the median dose deviation was >1%, 'cold' if it was <-1%, and 'normal' if it was within ±1%. A total of 1620 unique patient-specific QA measurements were available for model development and testing. 75% of the data were used to develop and cross-validate the model, and the remaining 25% were used for an independent assessment of model performance. For the model development phase, a recursive feature elimination cross-validation technique was used to eliminate unimportant features. Model performance was assessed using receiver operator characteristic (ROC) curve metrics.

Of the ten features found to be most predictive of patient-specific QA measurement results, half were derived from treatment plan characteristics and half from quality control (QC) metrics characterizing linac performance. The model achieved a micro-averaged area under the ROC curve of 0.93, and a macro-averaged area under the ROC curve of 0.88.

This work demonstrates the potential of using both treatment plan characteristics and routine linac QC results in the development of machine learning models for VMAT patient-specific QA measurements.

1 Introduction

Patient-specific quality assurance (QA) measurements are commonly employed to ensure the delivery accuracy of intensity modulated radiotherapy (IMRT) treatments (Ezzell *et al* 2003, Miften *et al* 2018). These measurements serve several important purposes, including the verification of data transfer, machine deliverability, and dosimetric calculations. However, these measurements often require considerable clinical resources and there is debate as to their necessity for all IMRT treatments (Siochi and Molineu 2013, Smith *et al* 2011, Neal *et al* 2016). Additionally, for measurement results that fail to meet pre-defined 'pass' criteria, it is often difficult to determine and correct the exact cause(s) of failure (Nelms *et al* 2013). Further, patient-specific QA measurements are typically performed late in the pre-treatment verification process, often outside of clinical operating hours when it may be too late to react without resulting in treatment delays. Radiotherapy clinics thus stand to benefit from a more streamlined, less resource-intensive approach to patient-specific QA, and the ability to proactively detect failing measurement results.

Typically, patient-specific QA involves the delivery of a patient treatment plan to a detector array that is used to measure the delivered dose or fluence. Comparisons between measured and planned distributions are usually quantified by means of a gamma analysis to assess delivery accuracy (Low and Dempsey 2003). This process compounds numerous potential sources of error, including dose calculation, data transfer, linac performance, device setup, and dosimeter response, among others (Miften *et al* 2018). As a result, it is often non-trivial for clinical physicists to identify and rectify the cause(s) of failing measurements. In addition, several reports have found common pre-treatment verification measurements to be insensitive to delivery errors (Nelms *et al* 2013, Stojadinovic *et al* 2015) and unable to predict the acceptability of plan delivery (Kry *et al* 2014). Despite its limitations, measurement-based patient-specific QA allows for some confidence in data transfer integrity, linac performance,

and dosimetric accuracy. Thus, it remains an integral part of the IMRT treatment process in many clinics (Miften *et al* 2018).

Several alternatives to measurement-based patient-specific QA have been proposed and evaluated. Notable and common examples include linac log-file analysis and secondary dose calculations (Kerns *et al* 2014, Childress *et al* 2015, Agnew *et al* 2016, Teke *et al* 2010, Siochi *et al* 2009). Log-file analysis provides an assessment of data transfer integrity and machine delivery based on parameters recorded by the linac control system. In addition, information contained within the files can be used to reconstruct expected dose distributions for comparison with treatment planning system (TPS) calculations; however, log-file analysis lacks a thorough measurement-based dose verification and discrepancies between actual MLC positions and those recorded in log-files have been noted (Neal *et al* 2016). Secondary dose calculations performed with an independent dose calculation engine are also used to validate TPS-generated dose distributions (Pisaturo *et al* 2009, Chen *et al* 2015) but are unable to provide verification of data transfer and deliverability. Log-file analysis combined with secondary dose calculations can be used to verify data transfer, deliverability, and TPS dose distributions. This approach, however, does not account for inter- and intra-linac performance variability.

More recent approaches to patient-specific QA involve the use of machine learning models to predict the results of measurement-based methods. Valdes *et al* have demonstrated the potential of training a machine learning algorithm to predict gamma passing rates for static gantry, sliding window IMRT plans delivered to 2D diode arrays using features that describe treatment plan complexity (Valdes *et al* 2016, 2017). Other studies have investigated the use of deep learning algorithms trained on fluence maps to make similar predictions (Tomori *et al* 2018, Interian *et al* 2018). While these techniques do not address issues associated with data transfer integrity and machine delivery, such ‘virtual’ patient-specific

QA processes are potentially useful in alerting clinicians to treatment plans that have a high likelihood of failing to meet clinical passing criteria. This would allow for a more proactive approach to dealing with potential failures. Additionally, it would allow for prioritization of treatment plans that would most benefit from measurement-based QA versus alternative methods. Although statistical process control charts can be used to alert physicists to systematic changes that may be predictive of patient-specific QA failures (Sanghangthum *et al* 2013) the machine learning approach is advantageous in that it accounts for the unique features of each plan and allows for failure predictions on an individualized basis.

Existing studies have clearly demonstrated the potential of using machine learning techniques to predict the results of patient-specific QA measurements (Valdes *et al* 2016, 2017). Thus far, these studies have included static gantry, sliding window IMRT treatment plans generated using the Eclipse TPS and delivered on Varian linacs to 2D arrays. Additionally, the machine learning models were trained on features related to treatment plan complexity, and variations in linac performance were not explicitly considered in the feature selection. In this work, we investigate the potential of using similar machine learning techniques trained on features that describe both treatment plan complexity and linac performance to predict patient-specific QA results for VMAT treatments developed and delivered using Elekta treatment planning and delivery infrastructure. A model capable of accurately predicting patient-specific QA results represents a useful clinical tool that could allow physicists to flag potentially problematic treatment plans and machine performance issues in a proactive manner, and reduce patient delays associated with substandard measurement results. Additionally, such tools could be used to support the reduction, or even elimination, of patient-specific QA measurements if used in combination with other reliable methods to verify data transfer and delivery accuracy.

2 Methods

2.1 Data sources

In this work, we trained a support vector classifier (SVC) to predict the results of VMAT patient-specific QA measurements. Two categories of data sources were used to calculate model features: those describing treatment plan characteristics (i.e. complexity metrics), and those that describe linac performance (i.e. linac QC metrics). Data sources were accrued through routine clinical operations as described in the following sections. The SVC model was developed and evaluated using 1620 patient-specific QA measurements, each corresponding to a unique VMAT treatment plan and delivery. These data were split into a training set (75%) used for model development and cross-validation, and a testing set (25%) used for evaluation. The Python library scikit-learn version 0.18 was used for all data analysis and model development (Pedregosa *et al* 2011). Prior to model training, the feature data were standardized by subtracting the mean and scaling to unit variance.

2.1.1 Model targets: Patient-specific QA results

At our institution, we perform patient-specific QA measurements prior to the first fraction of every VMAT treatment. These routine measurements have resulted in a database of 1620 patient-specific QA measurement results that was used to extract the targets for this study. All measurements were performed using Delta^4 biplanar diode arrays (ScandiDos, Sweden). Each measurement was performed using one of three clinical Delta^4 devices and measurements were performed on one of seven dosimetrically matched Elekta linacs equipped with Agility MLCs (Elekta, Sweden). No corrections for linac output variation were applied at the time of measurement. All plans were delivered with a nominal beam energy of 6 MV and were developed using the same beam model in the Monaco 5.11 treatment planning system (Elekta, Sweden). All measurements were stored in, and retrieved from, a database via the QATrack+ application (v0.29) (Studinski *et al* 2014). Each patient-specific QA measurement was

associated with the treatment plan, the treatment unit used to perform the measurement, and the date the measurement was performed.

2.1.2 Model features: Treatment plan characteristics

Data extracted from DICOM-RT Plan files associated with each patient-specific QA measurement were used to calculate a series of features describing treatment plan characteristics, complexity metrics, and delivery patterns. Data were extracted from the DICOM files using pydicom (Mason 2011), and features were calculated using scripts developed in-house. A number of studies have demonstrated that such complexity metrics are predictive of delivery accuracy (Du *et al* 2014, McNiven *et al* 2010, Younge *et al* 2016, Crowe *et al* 2014, Agnew *et al* 2016). The metrics that we used are largely similar to those used by Valdes *et al* to predict patient-specific QA results for static-gantry IMRT plans (Valdes *et al* 2016). The features included MU-averaged quantities such as the aperture area, aperture perimeter, aperture irregularity, centroid location, small aperture scores, leaf gaps, and jaw positions. Other metrics describing the overall plan included the maximum aperture dimensions, MU factor, total number of MUs, control points per dose/fraction, and the average motion per MU of the jaws and MLC leaves. Additionally, we included three VMAT-specific features to describe gantry motion; total angular arc range, the number of individual arcs, and the average gantry angle variation per control point. A complete list of features investigated is available in the supplementary material.

2.1.3 Model features: Linac performance features

Our clinic uses a QATrack+ (Studinski *et al* 2014) database to archive the results of all routine linac QC measurements. For each patient-specific QA measurement result, we extracted the most recent preceding routine QC results for the linac on which the measurement was performed. These features include commonly recommended dosimetric QC metrics describing linac output (dose/MU), profile flatness, and profile symmetry (Klein *et al*

2009). Beam profile symmetry and flatness measurements were acquired using a gantry-mounted MatrixX ion chamber array (IBA Dosimetry, Belgium). Additionally, we extracted the results of non-standard measurements that we perform, including a sliding window output (dose/MU) test involving a 2 cm-wide sliding MLC window, and a Delta^4 calibration check that compares an ion chamber dose measurement with the Delta^4 diode measurement to assess the drift in the absolute dose reading from the Delta^4. A complete list of the linac performance and QC features that we investigated is available in the supplementary material.

2.2 Feature selection

In an effort to eliminate unimportant features and reduce the complexity of our SVC model, we took an automated approach to feature selection using the training dataset. This was accomplished using a recursive feature elimination (RFE) cross-validation (RFECV) algorithm (Guyon *et al* 2002). The RFE process involves fitting a model using all available features, evaluating the model performance, and then eliminating the least important feature. Relative importance of features is determined using the feature weights. This process is repeated in an iterative manner, and allows for an assessment of performance with varying numbers of features. To evaluate performance, we used the average area under the receiver-operator characteristic (ROC) curves. A simple SVC with a linear kernel was used during RFECV, the same type of algorithm used to train the final model (section 2.4). We repeated the RFE process using a 10×5-fold cross validation. A total of 60 features were initially investigated in the RFECV process.

2.3 Target selection

The target metric used to describe the results of patient-specific QA measurements was defined as the median deviation between the dose calculated by the treatment planning system (Monaco5) and the dose measured by the Delta^4 diodes receiving >80% of the maximum planned dose. This differs from the more common gamma pass rate metric with dose

deviation/distance-to-agreement criteria of 3%/3mm or similar. We chose the median dose deviation because it is less sensitive to small errors in our measurement setup and provides more intuitive information than gamma pass rates. The median dose deviation is also advantageous in that it distinguishes between treatment plans that measure hotter or colder than expected based on the calculated dose distribution.

We binned the median dose deviation for each plan into three classes; 'hot' if the measured median dose deviation was >1%, 'normal' if it was within ±1%, and 'cold' if it was <-1%. The 1% threshold was selected as a conservatively small value that approximately corresponds to the range of fluctuations that we routinely see in repeated deliveries of the same treatment plans over time.

2.4 Model training and evaluation

An SVC model with a linear kernel was used throughout this study (Cortes and Vapnik 1995). As the primary purpose of this work was to demonstrate the feasibility of predicting VMAT patient-specific QA results rather than to rigorously optimize algorithm selection, we elected to choose a common, simple model. Linear kernels are advantageous in their simplicity – there is not a large number of hyper-parameters that require tuning – and have a reduced risk of overfitting compared to some more complex models. A further advantage of the linear kernel selection is that it provides a means of evaluating feature relevance through the feature weights (Guyon *et al* 2002). Thus, some insight into the relative importance of the features chosen is made possible when using a linear kernel. We emphasize, however, that this algorithm selection is somewhat arbitrary – it is possible, and indeed likely, that some performance gains may result from optimizing a different algorithm.

We used a 10×5-fold cross validation on the training data set to tune the *C* value of the linear SVC. The *C* value is a regularization parameter that controls the tradeoff between hyperplanes that maximize the distance between classes and hyperplanes that minimize the

misclassification rate in the training set. We evaluated the cross-validation performance using average ROC area under the curve (AUC) metrics. In addition to the cross validation scores described above, an independent test set of 405 plans was withheld for assessment of the final model's performance. Model performance was assessed using multiclass ROC analyses on the test set.

3 Results

3.1 Summary of the patient-specific QA dataset

The patient-specific QA dataset used in this study included all sites treated with VMAT at our centre. The breakdown of sites treated was approximately as follows: 25% palliative treatments of all sites; 25% genitourinary; 25% lung and mediastinum; 10% gastrointestinal; 5% dermatological; 5% gynecological. The remaining 5% was made up of sarcomas, lymphomas, and head & neck treatments. Note that the relatively small number of head and neck VMAT treatments is a result of these sites primarily being treated using TomoTherapy (Accuray Inc, CA) at our centre.

A histogram showing the distribution of median dose deviation values resulting from the patient-specific QA tests included in the training set is shown in figure 1. The breakdown of classes was as follows in the training set: 18% cold; 68.5% normal; 13.5% hot. In the test set, the breakdown was: 18.8% cold; 68.1% normal; 13.1% hot.

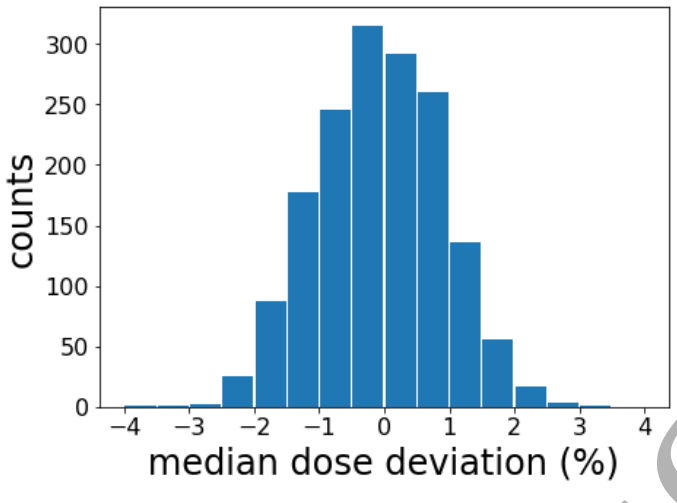

Figure 1: Histogram of the median dose deviations between TPS calculation and Delta[4] measurements in the training set.

3.2 Feature selection

To provide visualization of the feature space and a preliminary assessment of the potential of the features investigated to predict the targets, we performed a linear discriminant analysis (LDA) to reduce the feature dimensionality from sixty to two. Figure 2 shows a plot of the second linear discriminant versus the first, with the three classes of measurement results labelled (hot, cold, and normal). Clustering of hot, cold, and normal plans can be observed. This clustering provides qualitative evidence that the features we selected are predictive of the median dose deviation in patient-specific QA measurements.

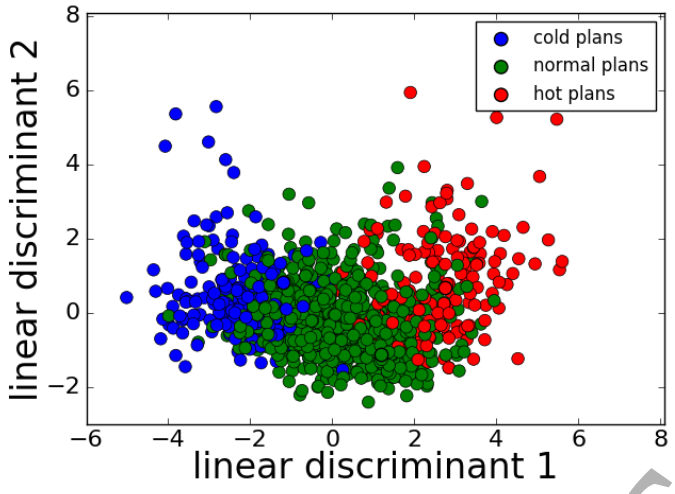

Figure 2: Results of a two-dimensional linear discriminant analysis. Hot, cold, and normal patient-specific QA measurements are shown in red, blue, and green respectively. Hot plans correspond to measured median dose differences > 1% of those calculated by the treatment planning system, normal between ±1%, and cold < -1%.

Figure 3 shows the results of the RFECV study used to eliminate unimportant features from the model. The model performance, as specified by ROC AUC values (both micro- and macro-averaged across the three classes of patient-specific QA results), is shown as a function of the number of features used during training. Performance metrics shown are the mean values obtained during a 10×5-fold cross validation, and error bars represent the standard deviation of the mean values obtained in each of the 10 repeated cross-validations. The ROC AUC metrics show an intial increase in performance up until approximately seven features are included in the SVC model, followed by more modest improvements with additional features, until reaching a plateau in performance with around 30 features. Based on this analysis, we selected the 30 most predictive features (out of the 60 investigated) to train the final SVC model.

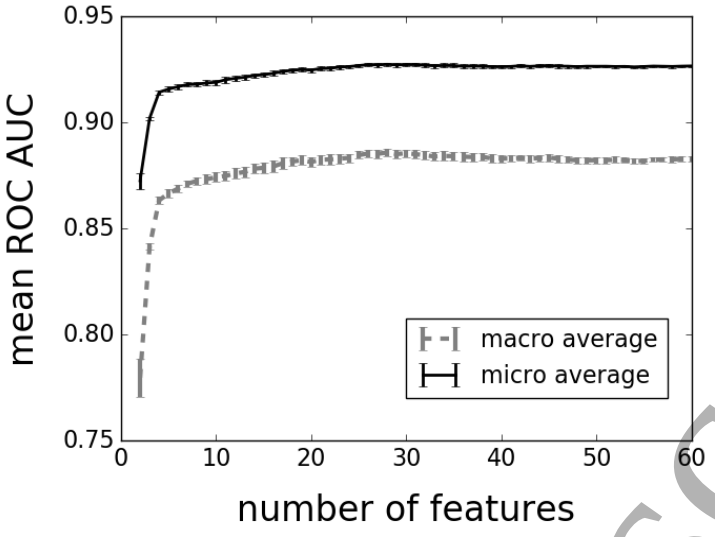

Figure 3: Mean macro- and micro-averaged ROC AUC values as a function of the number of features used to develop a linear SVC during RFECV (described in section 2.2).

3.3 Model training and performance

Model performance was not found to be strongly dependent on the *C* parameter value (section 2.4) of the linear SVC, as assessed by 10×5-fold cross-validation scoring of the micro- and macro- averaged ROC AUC values. For *C* parameter values ranging from 2^{-3} to 2^{5}, the micro- and macro-averaged ROC AUC varied by 0.004 ± 0.003 and 0.002 ± 0.001, respectively. The *C* value that maximized the micro-averaged ROC AUC (0.5) was selected.

The entire training dataset (1215 samples) was used to train a final model using the 30 features selected in section 3.2. The model's performance was evaluated by applying it to the independent testing set, consisting of 405 patient-specific QA measurements that were not used during model development. The resulting ROC curves are shown in figure 4. For the individual classes of patient-specific QA results, ROC AUC values were 0.90, 0.80, and 0.92 for cold, normal, and hot plans, respectively. The micro- and macro-averaged ROC values were 0.93 and 0.88.

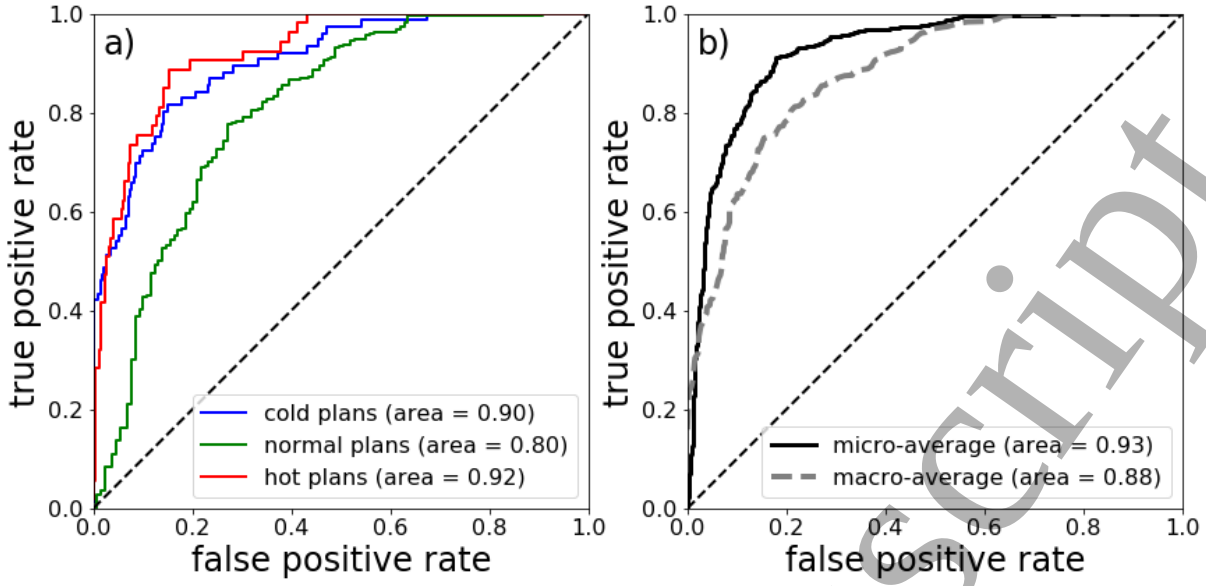

Figure 4: ROC curves for the final SVC model. Individual ROC curves for each class (a) are shown, as well as micro- and macro-averaged ROC curves (b).

3.4 Feature importance

As discussed in section 2.4, an advantage of using an SVC with a linear kernel is the ability to extract feature weights that provide some insight into the relative importance of each feature to the model. The weights associated with each of the 30 features used in the final model are shown in figure 5. Perhaps obviously, the feature rankings suggest that the daily measured linac output (dose/MU) fluctuations have the most significant impact on the measured median dose deviation. In addition to dose/MU, linac QC metrics that were amongst the 15 most important features were beam profile symmetry (at multiple gantry angles, both inplane and crossplane), beam profile flatness, and the dose/MU from the sliding window field.

The most important VMAT treatment plan characteristic was the MU factor, which is consistent with the results in work by Valdes *et al*. This was followed by the maximum aperture width, the average gantry angle rotation per control point, range of gantry angles (total arc length), dose per fraction, total number of control points, and the average aperture area outside of a 2.5 cm radius.

In addition to the linac QC metrics and VMAT treatment plan characteristics, and as might be expected, the results of our calibration check of the Delta^4 devices were also found to be of importance, ranking 15th.

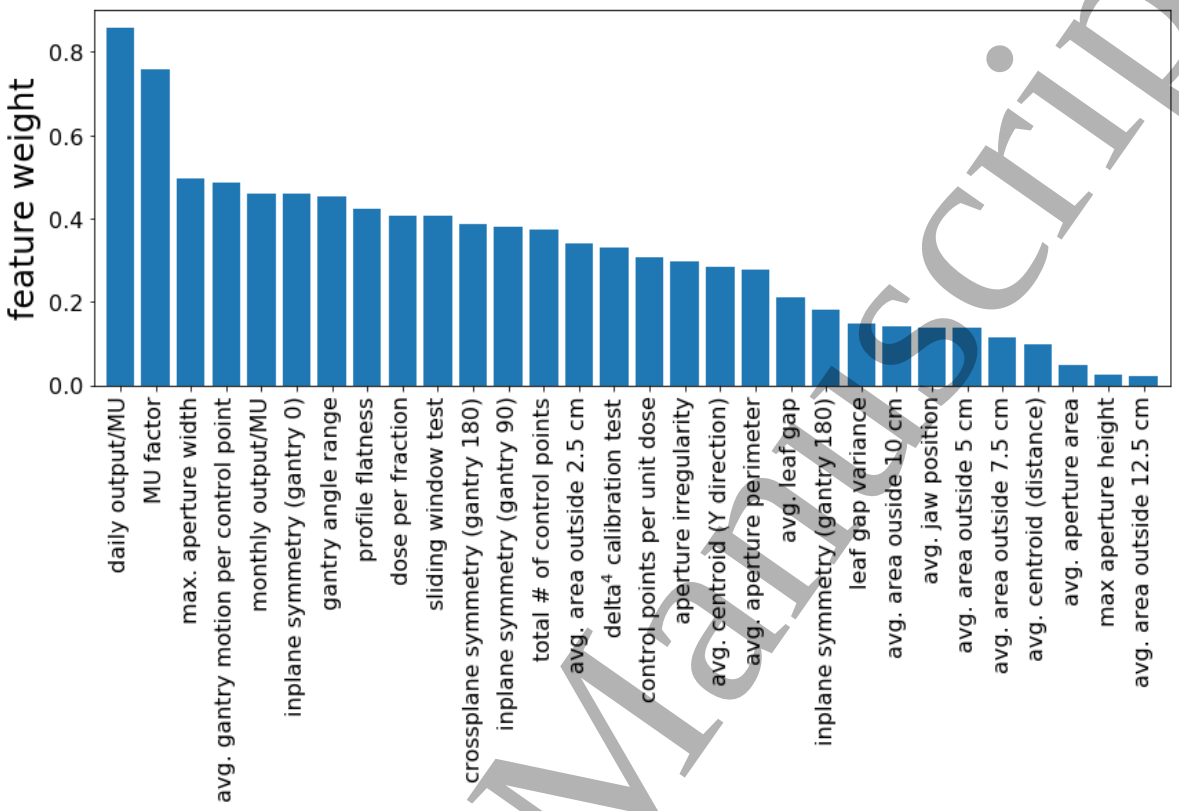

Figure 5: Feature importance, as determined by the feature weights of the support vectors between hot and cold classes of patient-specific QA measurement results.

4 Discussion

The model developed in this work to classify outcomes of VMAT patient-specific QA measurements based on median dose deviation achieved a macro-average ROC AUC value of 0.88 and a micro-average was 0.93. These results demonstrate that it is feasible to use machine learning techniques to classify the median dose deviation that would be measured for a particular VMAT patient-specific QA measurement as either hot (> 1%), cold (< -1%), or within a normal range (i.e., within ±1%).

Although we focused on VMAT treatments created and delivered using Elekta treatment planning and delivery systems, other studies have demonstrated machine learning approaches for static-gantry IMRT treatments using Varian linacs and 2D diode arrays. Importantly, our work used a different delivery technique (VMAT), TPS (Monaco) and linac (Elekta), as well as a unique measurement device (Delta^4). Further, the analysis in this work takes into account treatment delivery performance without collapsing the delivery control points into a single gantry angle. As our model was trained on data that were generated by products and techniques that vary institutionally, the specific model that was evaluated in this work may not achieve similar performance at other institutions without re-training on local data. However, our results, combined with those already in the literature, suggest that similar machine learning methods may be generally applicable to other clinics, regardless of the treatment planning systems, linacs, and measurement devices employed. While promising in terms of clinical utility, a key limitation is that replication of these methods in a radiotherapy center that is distinct from ours likely requires a model trained on an appropriately-sized sample of patient-specific QA measurements. Smaller clinics may have difficulty acquiring large enough datasets to train such a model.

The machine learning model-building process used in this work involved many decisions that could be considered somewhat arbitrary. One could be justified in selecting different features to investigate, different targets to predict, and different machine learning algorithms to train. We differed from previous works in our selection of features to investigate (though plan complexity features were similar to those used in previous investigations), target metric (median dose deviation versus gamma passing rate), and machine learning algorithm. We emphasize that the work that we have presented is not intended to represent the absolute optimal method by which to predict patient-specific QA results. Rather, we intended to investigate the feasibility of using machine learning techniques applied to a delivery

technique, TPS, and measurement device that have not been previously studied as well as the potential benefits of explicitly including linac QC results as features.

The features describing treatment plan characteristics primarily consisted of aperture-based features. Although other researchers have had success using complexity metrics based on fluence maps (including using deep learning approaches), we preferred to use aperture-based metrics due to their interpretability and direct relationship with linac delivery parameters. This allows for a better understanding of limitations in plan optimization parameters and provides a pathway for feedback on improvements in TPS beam modelling. Notably, the MU factor was found to be the most important plan complexity feature in both our work and that of Valdes *et al* (Valdes *et al* 2016). Also of note, two plan features that are specific to VMAT treatments, and therefore not previously investigated, were found to be among the most important features (total arc length and average gantry rotation per control point).

In contrast to previous studies, we explicitly included linac QC test results as features in our SVC model. A potential advantage to the explicit inclusion of such features is the ability to assess variations in expected results from linac-to-linac, and to identify problematic or beneficial linac-plan combinations. To our knowledge, previous studies did not consider linac performance features, and attempted to remove the linac-to-linac variations by correcting for linac calibration discrepancies in all patient-specific QA measurements, and renormalizing if discrepancies were large. In our local patient-specific QA process, we do not perform corrections for linac output (dose/MU) fluctuations, as this gives a more realistic representation of actual patient treatment deliveries. Our results show linac output (dose/MU) was by far the most important feature describing linac performance; however, other linac QC metrics, including beam flatness, symmetry, and the sliding window dose measurement, were also found to be among the most important features in the SVC model. While we elected to

use 30 features in our final model, the results presented in figure 3 suggest that comparable performance of the SVC model could be achieved with fewer features, which would simplify both the data collection and modelling processes.

The most common practice in the analysis of patient-specific QA measurements is to summarize results using a gamma passing rate. Despite this, we chose to train the SVC model in this work to predict median dose deviation since this metric is relatively insensitive to small errors in setup, better distinguishes between plans that are hotter and colder than expected, and is generally more interpretable than gamma pass rates. In addition, we elected to choose simple classes of targets – hot, cold, and normal – for our model. Since the median dose deviation is a continuous value, we could have chosen to perform a regression rather than classification with our model. Regression models have the advantage of providing more quantitative information. Classification, on the other hand, has the advantages of providing a quick, unambiguous, and actionable result, similar to routine linac QC results being binned into passing, tolerance, and action-level categories. Additionally, our preliminary investigations into regression analyses have typically underestimated large dose discrepancies, possibly due to the relative infrequency of such occurrences. Rather than risk overconfidence in a regression prediction, it may be considered preferable to set conservative limits (e.g., 1%) on plans that were considered ‘hot’ and ‘cold’ and treat all plans within a class in the same manner. In future work, we intend to explore the potential of performing a regression analysis in more detail.

The ability to accurately predict patient-specific QA results would allow clinical medical physicists to deal with likely failures on a more proactive basis, and reduce patient treatment delays. Additionally, as more efficient (but ultimately less comprehensive) alternatives to traditional patient-specific QA approaches become more prevalent, such predictive ability would allow clinicians to prioritize those treatment plans that are most likely

to be problematic for more rigorous, measurement-based QA. Models that explicitly include linac performance features could allow clinicians in a multi-linac setting to identify problematic linac-plan combinations, and pre-select those plans which maximize delivery accuracy, while also providing useful guidance on where to direct linac QC efforts for more robust beam matching.

5 Conclusion

Our work has demonstrated the feasibility of using data that is readily available in most radiotherapy clinics to train a machine learning model to predict the results of VMAT patient-specific QA measurements. Such a model has the potential to provide more timely failure detection for patient-specific QC measurements.

6 References

Agnew C E, Irvine D M and Mcgarry C K 2016 Correlation of phantom-based and log file patient-specific QA with complexity scores for VMAT *J. Appl. Clin. Med. Phys.* **15** 204–16

Chen X, Bush K, Ding A and Xing L 2015 Independent calculation of monitor units for VMAT and SPORT *Med. Phys.* **42** 918–24

Childress N, Chen Q and Rong Y 2015 Parallel/Opposed: IMRT QA using treatment log files is superior to conventional measurement-based method *J. Appl. Clin. Med. Phys.* **16** 5385

Cortes C and Vapnik V 1995 Support-Vector Networks *Mach. Learn.* **20** 273–97

Crowe S B, Kairn T, Kenny J, Knight R T, Hill B, Langton C M and Trapp J V. 2014 Treatment plan complexity metrics for predicting IMRT pre-treatment quality assurance results *Australas. Phys. Eng. Sci. Med.* **37** 475–82

Du W, Cho S H, Zhang X, Hoffman K E and Kudchadker R J 2014 Quantification of beam complexity in intensity-modulated radiation therapy treatment plans. *Med. Phys.* **41** 021716

Ezzell G A, Galvin J M, Low D, Palta J R, Rosen I, Sharpe M B, Xia P, Xiao Y, Xing L and Yu C X 2003 Guidance document on delivery, treatment planning, and clinical implementation of IMRT: Report of the IMRT subcommittee of the AAPM radiation therapy committee *Med. Phys.* **30** 2089–115

Guyon I, Weston J, Barnhill S and Vapnik V 2002 Gene selection for cancer classification using support vector machines *Mach. Learn.* **46** 389–422

Interian Y, Rideout V, Kearney V P, Gennatas E, Morin O, Cheung J, Solberg T and Valdes

G 2018 Deep nets vs expert designed features in medical physics: An IMRT QA case study *Med. Phys.* **45** 2672–80

Kerns J R, Childress N and Kry S F 2014 A multi-institution evaluation of MLC log files and performance in IMRT delivery *Radiat. Oncol.* **9** 176

Klein E E, Hanley J, Bayouth J, Yin F-F, Simon W, Dresser S, Serago C, Aguirre F, Ma L, Arjomandy B, Liu C, Sandin C and Holmes T 2009 Task Group 142 report: Quality assurance of medical accelerators *Med. Phys.* **36** 4197–212

Kry S F, Molineu A, Kerns J R, Faught A M, Huang J Y, Pulliam K B, Tonigan J, Alvarez P, Stingo F and Followill D S 2014 Institutional patient-specific IMRT QA does not predict unacceptable plan delivery *Int. J. Radiat. Oncol. Biol. Phys.* **90** 1195–201

Low D A and Dempsey J F 2003 Evaluation of the gamma dose distribution comparison method *Med. Phys.* **30** 2455–64

Mason D 2011 Pydicom: An open source DICOM library *Med. Phys.* **38** 3493

McNiven A L, Sharpe M B and Purdie T G 2010 A new metric for assessing IMRT modulation complexity and plan deliverability. *Med. Phys.* **37** 505–15

Miften M, Olch A, Mihailidis D, Moran J, Pawlicki T, Molineu A, Li H, Wijesooriya K, Shi J, Xia P, Papanikolaou N and Low D A 2018 Tolerance limits and methodologies for IMRT measurement-based verification QA: Recommendations of AAPM Task Group No . 218 *Med. Phys.* **45** e53–8

Neal B, Ahmed M, Kathuria K, Watkins T, Wijesooriya K, Siebers J, Neal B, Ahmed M, Kathuria K, Watkins T, Wijesooriya K and Siebers J 2016 A clinically observed discrepancy between image-based and log-based MLC positions *Med. Phys.* **43** 2933–5

Nelms B E, Chan M F, Jarry G, Lemire M, Lowden J, Hampton C and Feygelman V 2013 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.* **40** 111722

Pedregosa F, Weiss R and Brucher M 2011 Scikit-learn: Machine Learning in Python *J. Mach. Learn. Res.* **12** 2825–30

Pisaturo O, Moeckli R, Mirimanoff R-O and Bochud F O 2009 A Monte Carlo-based procedure for independent monitor unit calculation in IMRT treatment plans *Phys. Med. Biol.* **54** 4299–310

Sanghangthum T, Suriyapee S, Srisatit S and Pawlicki T 2013 Statistical process control analysis for patient-specific IMRT and VMAT QA *J. Radiat. Res.* **54** 546–52

Siochi R A C and Molineu A 2013 Point/Counterpoint: Patient-specific QA for IMRT should be performed using software rather than hardware methods *Med. Phys.* **40** 1–3

Siochi R A, Pennington E C, Waldron T J and Bayouth J E 2009 Radiation therapy plan checks in a paperless clinic *J. Appl. Clin. Med. Phys.* **10** 43–62

Smith J C, Dieterich S and Orton C G 2011 Point/Counterpoint: It is STILL necessary to validate each individual IMRT treatment plan with dosimetric measurements before delivery. *Med. Phys.* **38** 553–5

Stojadinovic S, Ouyang L, Gu X, Pompoš A, Bao Q and Solberg T D 2015 Breaking bad IMRT QA practice *J. Appl. Clin. Med. Phys.* **16** 5242

Studinski R, Taylor R, Angers C, La Russa D J and Clark B 2014 Development and implementation of a web based quality control software *Med. Phys.* **41** 246

Teke T, Bergman A M, Kwa W, Gill B, Duzenli C and Popescu I A 2010 Monte Carlo based, patient-specific RapidArc QA using linac log files *Med. Phys.* **37** 116–23

Tomori S, Kadoya N, Takayama Y, Kajikawa T, Shima K, Narazaki K and Jingu K 2018 A deep learning-based prediction model for gamma evaluation in patient-specific quality

assurance *Med. Phys.* **45** 4055–65

Valdes G, Chan M F, Lim B, Scheuermann R, Deasy J O, Solberg T D, Lim S B, Scheuermann R, Deasy J O and Solberg T D 2017 IMRT QA using machine learning: A multi-institutional validation *J. Appl. Clin. Med. Phys.* 279–84

Valdes G, Scheuermann R, Hung C Y, Olszanski A, Bellerive M and Solberg T D 2016 A mathematical framework for virtual IMRT QA using machine learning *Med. Phys.* **43** 4323–34

Younge K C, Roberts D, Janes L A, Anderson C, Moran J M and Matuszak M M 2016 Predicting deliverability of volumetric-modulated arc therapy (VMAT) plans using aperture complexity analysis *J. Appl. Clin. Med. Phys.* **17** 124–31