XXV ISSSD International Symposium on Solid State Dosimetry
September 22 – 26, 2025
Belo Horizonte, Minas Gerais - Brazil
CDTN UFMG

Machine Learning in PSQA: Performance of a Random Forest Model on Radiotherapy Metrics

Nathália L.A. Alves[1]*, Mário Olímpio de Menezes[1]**, Hélio Yoriyaz[2], Lucas A. Radicchi[3], André L. S.Mattesco[3], André G. Prospero[3]

[1]Research Reactor Center – CERPQ,[2]Nuclear Engineering Center – CEENG, Nuclear and Energy Research Institute – IPEN-CNEN,[3]RTCON Radiotherapy Solutions, São Paulo, Brazil

*nathalia.aidar@usp.br, **momenezes@usp.br

Abstract

This study explores the application of supervised machine learning (ML) algorithms to predict outcomes of patient-specific quality assurance (PSQA) in intensity-modulated radiotherapy (IMRT) and volumetric-modulated arc therapy (VMAT). A cohort of 133 anonymized clinical radiotherapy plans covering treatment sites such as prostate, head and neck, pelvis, central nervous system (CNS), and stereotactic radiosurgery (SRS) was analyzed. From each DICOM-RT (Digital Imaging and Communications in Medicine – Radiation Therapy) plan, 60 quantitative features were extracted, encompassing multileaf collimator (MLC) dynamics, dose distribution characteristics, and complexity indices including Modulation Complexity Score (MCS), Average Leaf Speed (ALS), and Leaf Sequence Variability (LSV). To address class imbalance between QA pass/fail outcomes, a synthetic dataset of 66 perturbed plans simulating realistic delivery deviations was generated, complemented by the use of the Synthetic Minority Oversampling Technique (SMOTE) [8]. Feature importance analysis identified MCS, ALS, and dose irregularity metrics as key predictors of QA failure. The optimized Random Forest classifier achieved a weighted F1-score of 0.99 during cross-validation and 0.92 on the independent test set. The results corroborate findings in the literature that support the integration of ML models to enhance PSQA accuracy and sensitivity beyond conventional gamma-based analysis [1, 10, 7, 9]. By leveraging DICOM-RT metadata, this study highlights the feasibility of implementing predictive algorithms in radiotherapy QA pipelines, improving workflow efficiency and reducing dependence on manual dosimetric verification.

Keywords: radiotherapy, machine learning, DICOM-RT

1. Introduction

The integration of artificial intelligence (AI) techniques into quality assurance (QA) workflows in radiotherapy has garnered significant attention in recent years, with the primary objective

of enhancing treatment accuracy and optimizing clinical resources. A growing body of literature has demonstrated the potential of machine learning algorithms to identify treatment discrepancies and improve the reliability of patient-specific QA procedures [1, 2]. Predictive modeling approaches that leverage plan complexity metrics and historical QA data have shown considerable promise in anticipating QA outcomes, thereby minimizing the need for extensive physical measurements and contributing to increased operational efficiency [3, 4]. This study seeks to promote a deeper integration between academic investigation and clinical practice, fostering the development of intelligent computational tools to support radiotherapy QA processes. Concurrently, gamma analysis continues to represent a cornerstone in the verification of dose delivery accuracy in radiotherapy. Despite its widespread clinical adoption, conventional QA methodologies frequently fall short in detecting subtle deviations and long-term trends, which underscores the growing relevance of combining traditional techniques with data-driven approaches [5]. The transition toward AI-assisted QA paradigms reflects a broader shift in the field, emphasizing adaptability, automation, and evidence-based decision-making. The present work aims to investigate the application of machine learning in PSQA in radiotherapy, with a particular emphasis on the assessment of delivery complexity and the prediction of gamma passing rates. The proposed approach is grounded in the extraction of quantitative features from planning and delivery data, with the objective of training a predictive model capable of classifying treatment plans according to TG-218 gamma criteria [5].

2. Methodology

2.1 Dataset Construction and Feature Extraction

The implementation was carried out in Python, leveraging specialized libraries such as pydicom for parsing DICOM files, and pandas for organizing extracted data into relational structures. The pipeline was designed in a modular fashion, with dedicated scripts for handling each DICOM modality: RTPLAN, RTDOSE, and RTIMAGE.

The RTPLAN parser was responsible for extracting treatment beam configurations, delivery times, and prescribed doses. Iterative loops were employed to traverse patient datasets and store relevant metadata. For RTDOSE files, the parser converted raw voxel data into three-dimensional matrices, enabling the computation of aggregate dosimetric parameters such as mean and maximum dose within defined regions of interest. In turn, the RTIMAGE module focused on retrieving portal image metadata, including acquisition timestamps and pixel intensity values, ensuring their alignment with the corresponding treatment arcs.

A total of 133 radiotherapy treatment plans were evaluated, encompassing a variety of anatomical sites, including 70 prostate, 15 head and neck, 5 central nervous system (CNS), 3 stereotactic radiosurgery (SRS), 20 pelvis, and a general 20 as "other" category. For each plan, a structured dataset was built containing 60 distinct metrics extracted from DICOM-RT files. These metrics included parameters such as prescribed dose, maximum dose, number of fractions, and total monitor units (MU).

The data were organized using the pandas library, which enabled efficient data manipulation and supported exploratory data analysis. Visual analytics were developed using matplotlib and seaborn to aid in the identification of patterns and detection of potential outliers. For example, scatter plots were used to examine the relationship between monitor units and acquisition duration, while histograms helped visualize inter-patient variability in dose per fraction.

EPID-based measurements were incorporated for real-time delivery validation. A dedicated script segmented treatment plans by arc, guided by acquisition timestamps extracted from RTIMAGE metadata. Matching algorithms were employed to associate EPID frames with specific RTPLAN segments. Spatial filtering techniques were applied to reduce image noise, followed by dose profile extraction and comparison with RTDOSE data through metrics such as mean percentage difference and gamma index similarity.

The validation process encompassed both manual and automated cross-checks. Scripts compared extracted metadata against expected clinical protocol values to identify inconsistencies or outliers. For instance, plans where the total dose derived from RTDOSE deviated substantially from RTPLAN prescriptions were flagged and, in less than 1% of cases, excluded due to formatting anomalies. Temporal consistency of EPID acquisitions was also verified to ensure synchronization with documented treatment phases.

A validation module was integrated into the pipeline to assess the integrity of input files before processing. This routine detected and excluded corrupted or incomplete datasets. Performance optimization was achieved by enabling batch processing and parallel computation through Python’s multiprocessing library, significantly reducing execution time across the dataset comprising thousands of records from 133 patients. Additionally, a logging routine was incorporated to capture progress, runtime errors, and diagnostic information, enhancing traceability and reproducibility.

These efforts culminated in the construction of a robust, well-structured database, which serves as the foundation for future developments in predictive modeling and automated QA optimization in radiotherapy workflows.

2.2 Balancing QA Datasets with Simulated Delivery Errors

To improve the predictive performance of the machine learning models developed in this study, a synthetic cohort of 66 patient treatment plans was generated to simulate QA failures. These synthetic plans were derived by applying controlled perturbations to the original dataset, such as altering multileaf collimator (MLC) trajectories, beam-on times, and dose distributions. The objective was to emulate realistic delivery deviations that could result in failing gamma index criteria, thereby increasing the representation of the minority class (QA failures) in the training dataset.

This strategy aligns with the general recommendations found in recent reports on the implementation of online treatment monitoring (OTM) systems, which emphasize the importance of evaluating system robustness using diverse clinical scenarios and simulated errors [6, 7]. The original dataset, composed predominantly of QA-approved cases, presented a marked class imbalance that could negatively impact model training. Unbalanced datasets tend to bias classifiers toward the majority class, reducing sensitivity to clinically relevant errors. To mitigate this issue, the Synthetic Minority Oversampling Technique (SMOTE) was used in this case [8]. SMOTE generates new samples of the minority class (QA failures) by interpolating between existing observations in the feature space, thereby enhancing the diversity and representativeness of this class during model training [8]. The use of SMOTE in this context is justified by the need for a more balanced training set, which improves model generalizability and reduces the risk of false negatives—cases in which failing plans are incorrectly classified as acceptable. Such errors could lead to undetected deviations in clinical practice, a risk highlighted in QA guidelines for IMRT and VMAT [9]. Additionally, the balanced dataset facilitates robust evaluation metrics such as sensitivity, specificity, and area under the receiver operating characteristic (ROC) curve, as emphasized in recent studies applying machine learning to radiotherapy QA [1, 10, 11]. Through this data augmentation and balancing process, the final model was trained on a dataset with equal representation of passed and failed QA cases. The trained Random Forest classifier achieved high performance metrics in the test phase, including an F1-score of 0.92, which reflects a strong balance between precision (low false positives) and recall (low false negatives). These results highlight how crucial it is to address class imbalance when designing supervised learning approaches for radiotherapy quality assurance.

3. Results

A comprehensive dataset encompassing 133 patients was analyzed, generating an extensive collection of records with 60 features related to treatment planning and delivery per arc. Following meticulous data preprocessing and validation, supervised machine learning models were developed to predict quality assurance (QA) outcomes, classified as either pass or fail based on gamma analysis using the 3%/3 mm criteria with a 10% low-dose threshold, as stipulated by the AAPM TG-218 guidelines [5].

A total of 184 treatment arcs were assessed, of which 128 (69.6%) satisfied the gamma passing criteria, while 56 (30.4%) were identified as failing. A pronounced discrepancy was observed between the original (Modified = 0) and perturbed (Modified = 1) datasets: 91% of the failing arcs were associated with modified plans, underscoring the model's sensitivity to variations in plan complexity and dose delivery precision.

To address the inherent class imbalance and enhance model robustness, a synthetic oversampling technique was employed, resulting in a balanced dataset with 128 instances for each class (pass and fail). A Random Forest classifier was subsequently trained and evaluated using five-fold cross-validation, achieving a mean F1-weighted score of 0.992 ± 0.009, indicative of exceptional predictive performance across validation subsets.

On an independent test set comprising 77 instances, the classifier attained an overall accuracy of 92%. The confusion matrix, presented in Table 1, elucidated the model's classification performance, revealing only six misclassifications: three false positives (FP), where passing arcs were erroneously predicted as failing, and three false negatives (FN), where failing arcs were incorrectly predicted as passing. The true positives (TP, correctly predicted failing arcs) and true negatives (TN, correctly predicted passing arcs) dominated, affirming the model's capability to accurately differentiate between clinically acceptable and suboptimal treatment deliveries.

	Predicted Pass	Predicted Fail
Actual Pass	**45 (TN)**	**3 (FP)**
Actual Fail	**3 (FN)**	**26 (TP)**

Table 1. Confusion Matrix for the Random Forest Classifier on the Test Set

Note: TN = True Negative, FP = False Positive, FN = False Negative, TP = True Positive.

The model achieved robust performance across both QA outcome classes. For failing arcs, it yielded a precision of 0.97 and recall of 0.87, while passing arcs presented the inverse trend, with precision and recall of 0.88 and 0.97, respectively. These complementary patterns resulted in balanced F1-scores of 0.92 for both classes, indicating the model's ability to effectively distinguish between acceptable and unacceptable treatment deliveries without bias toward the majority class.

The feature importance analysis identified the Modulation Complexity Score (MCS) as the primary predictor of quality assurance (QA) outcomes, with an importance score of 0.292, followed by Average Leaf Speed (ALS) at 0.190 and the Irregularity Factor at 0.142. These results underscore the critical influence of dynamic multileaf collimator (MLC) parameters on QA failures. Other notable features include the number of Monitor Units (MU_1, 0.073), Maximum Dose (0.034), and geometric field parameters such as Field_Y_1 (0.018). These findings highlight that both dosimetric and geometric characteristics of the treatment plan play significant roles in determining QA performance, with MLC-related dynamics being particularly impactful, as shown in Figure 1.

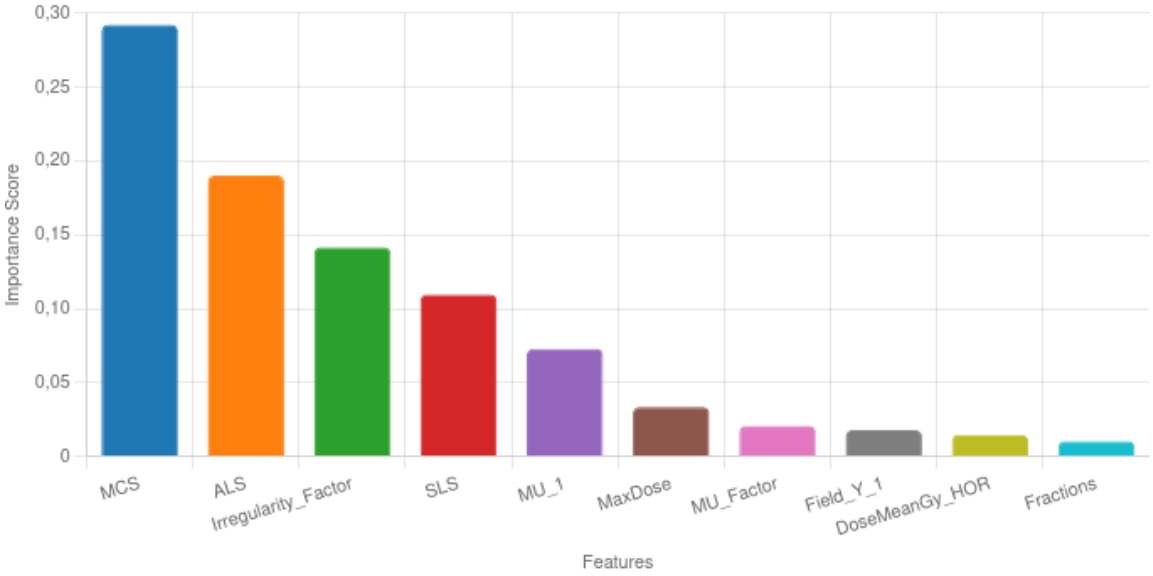

Figure 1. Feature Importance Scores for the Random Forest Classifier on the Test Set

The low incidence of false positives (3/77) minimizes the risk of unwarranted clinical interventions for acceptable plans, while the equally low rate of false negatives (3/77) ensures that most suboptimal plans are detected, thereby enhancing patient safety. These results validate the robustness of the Random Forest classifier and its potential for integration into clinical QA workflows for VMAT plans in radiotherapy.

4. Discussion

The high F1-score of 0.92 on the independent test set substantiates the robustness of the Random Forest classifier and its capacity for generalization to unseen data. Beyond overall accuracy, a detailed analysis of the confusion matrix reveals the critical implications for clinical implementation. The model's low misclassification rates—just three false positives and three false negatives on the test set—indicate its potential as a reliable screening tool to enhance the efficiency of patient-specific quality assurance (PSQA) workflows.

The minimal rate of false positives (FP) is particularly significant for clinical efficiency, as it suggests that the model would rarely flag an acceptable plan for unnecessary manual review. This minimizes the expenditure of clinical resources. Conversely, the rate of false negatives (FN) represents the most critical concern from a patient safety perspective, as these are clinically failing plans that the model incorrectly classified as acceptable. While this rate is low, it underscores that the model should function as a decision support tool rather than a complete replacement for established dosimetric verification procedures. Its primary role is to assist medical physicists in prioritizing which plans require the most rigorous review, thereby optimizing resource allocation without compromising safety.

The feature importance analysis offers valuable insight into the physical determinants of QA failure. The Modulation Complexity Score (MCS) and Average Leaf Speed (ALS) emerged as the most influential predictors, which aligns with the well-documented clinical observation that highly complex treatment plans requiring rapid multileaf collimator (MLC) movements are inherently more susceptible to delivery errors. This finding validates the model's physical foundation, as it effectively learns to identify the same risk factors that contribute to real-world delivery discrepancies. The successful classification of these complex cases was significantly enabled by the use of a synthetic dataset of realistic failure scenarios, a crucial step in mitigating the class imbalance inherent in clinical QA data.

Despite its promising performance, this study has several limitations. The dataset, while augmented, is derived from a single-institution cohort. Consequently, the model's generalizability across diverse clinical environments with varying equipment and protocols remains to be fully validated. Additionally, the simulated perturbations, although informed by clinical experience, may not encompass the full spectrum of potential delivery errors. Future research should focus on validating this predictive framework using multi-institutional data. Further improvements could be achieved by integrating more sophisticated machine learning architectures, such as deep learning, and by incorporating real-time data from treatment log files to enhance predictive accuracy and facilitate the development of an online monitoring system.

5. Conclusions

This study presents a comprehensive approach to predictive modeling of radiotherapy plan quality assurance (QA) through the integration of clinical data, complexity metrics, and machine learning techniques.

Descriptive and exploratory analyses revealed substantial variability across anatomical regions, particularly in parameters such as prescribed dose and maximum dose per fraction, emphasizing the clinical heterogeneity inherent to radiotherapy planning. Notably, pelvic treatments exhibited the highest variability in dose delivery, underscoring the need for site-specific QA strategies.

The inclusion of synthetically generated failure cases, SMOTE, led to a more balanced training distribution, allowing the model to better distinguish between passing and failing QA scenarios. This improved representation of both classes contributed directly to the robustness of the Random Forest classifier, which achieved a weighted F1-score of 0.99 during cross-validation and 0.92 when evaluated on an independent test set.

The analysis of feature importance highlighted the relevance of plan complexity metrics, with the Modulation Complexity Score (MCS), Average Leaf Speed (ALS), and Sequence Leaf Speed (SLS) emerging as the most influential predictors of gamma passing rate. These findings are consistent with the literature and reinforce the clinical interpretability of model outputs.

In summary, the proposed framework demonstrates that supervised machine learning models, informed by quantitative radiotherapy metrics, can effectively support QA decision-making. By reducing dependence on manual measurement procedures and enabling early identification of potentially suboptimal plans, this approach contributes to the development of more efficient, data-driven, and personalized QA processes in contemporary radiotherapy practice.

Acknowledgements

The authors would like to thank RTCON Solutions in Radiotherapy and IBCC Hospital for their partnership and support throughout the research. Special thanks to INCT INTERAS - PROCESSO CNPq 406761/2022-1 for supporting the participation in scientific events and funding the research activities.

References

1. Crowe, S. B. et al. (2015). J. Phys.: Conf. Ser., 573: 012042.

2. Li, Y. et al. (2015). Med. Phys., 42(6): 3456–3464.

3. Houssen, A. et al. (2017). J. Appl. Clin. Med. Phys., 18(3): 134–142.

4. Moreau, N. et al. Deep hybrid learning prediction of patient-specific quality assurance in radiotherapy: Implementation in clinical routine.

5. Miften, M. et al. (2018). Tolerance limits and methodologies for IMRT measurement-based verification QA: Recommendations of AAPM Task Group No. 218. Med. Phys., 45(4): e53–e83. https://doi.org/10.1002/mp.12810

6. Nelms, B. E. et al. (2012). Med. Phys., 39(2): 765–778.

7. Sadowski, B., Milewska, K., and Ginter, J. (2022). Machine Learning Based Prediction of Gamma Passing Rate for VMAT Radiotherapy Plans. J. Pers. Med., 12(2071). https://doi.org/10.3390/jpm12122071

8. Pradipta, G. A. et al. (n.d.). SMOTE for Handling Imbalanced Data Problem: A Review. Faculty of Social and Political Sciences, UIN Syarif Hidayatullah, Jakarta, Indonesia.

9. Stevens, S. et al. (2023). IPEM topical report: guidance for the clinical implementation of online treatment monitoring solutions for IMRT/VMAT. Phys. Med. Biol., 68: 18TR02.

10. Nguyen, T. et al. (2023). Applications of machine and deep learning in patient-specific IMRT/VMAT quality assurance. Med. Phys., 50(1): 24–36. https://doi.org/10.1002/mp.16000

11. Sakai, M. et al. (2021). Detecting MLC modeling errors using radiomics-based machine learning in patient-specific QA with an EPID for intensity-modulated radiation therapy. Med. Phys., published online January 27, 2021. https://doi.org/10.1002/mp.14820

12. Valdes, G. et al. (2016). Int. J. Radiat. Oncol. Biol. Phys., 94(5): 1080–1087