Patient-specific Quality Assurance Failure Prediction with Deep Tabular Models

R. Levin *, A. Y. Aravkin, M. Kim
University of Washington, Seattle WA

Abstract

Background: Patient-specific quality assurance (PSQA) is part of the standard practice to ensure that a patient receives the dose from intensity-modulated radiotherapy (IMRT) beams as planned in the treatment planning system (TPS). PSQA failures can cause a delay in patient care and increase workload and stress of staff members. A large body of previous work for PSQA failure prediction focuses on non-learned plan complexity measures. Another prominent line of work uses machine learning methods, often in conjunction with feature engineering. Currently, there are no machine learning solutions which work directly with multi-leaf collimator (MLC) leaf positions, providing an opportunity to improve leaf sequencing algorithms using these techniques.
Purpose: To improve patient safety and work efficiency, we develop a tabular transformer model based directly on the MLC leaf positions (without any feature engineering) to predict IMRT PSQA failure. This neural model provides an end-to-end differentiable map from MLC leaf positions to the probability of PSQA plan failure, which could be useful for regularizing gradient-based leaf sequencing optimization algorithms and generating a plan that is more likely to pass PSQA.
Method: We retrospectively collected DICOM RT PLAN files of 968 patient plans treated with volumetric arc therapy. We construct a beam-level tabular dataset with 1873 beams as samples and MLC leaf positions as features. We train an attention-based neural network FT-Transformer to predict the ArcCheck-based PSQA gamma pass rates. In addition to the regression task, we evaluate the model in the binary classification context predicting the pass or fail of PSQA. The performance was compared to the results of the two leading tree ensemble methods (CatBoost and XGBoost) and a non-learned method based on mean MLC gap.
Results: The FT-Transformer model achieves 1.44% Mean Absolute Error (MAE) in the regression task of the gamma pass rate prediction and performs on par with XGBoost (1.53 % MAE) and CatBoost (1.40 % MAE). In the binary classification task of PSQA failure prediction, FT-Transformer achieves 0.85 ROC AUC (with CatBoost and XGBoost achieving 0.87 ROC AUC and the mean-MLC-gap complexity metric achieving 0.72 ROC AUC). Moreover, FT-Transformer, CatBoost, and XGBoost all achieve 80% true positive rate while keeping the false positive rate under 20%.
Conclusions: We demonstrate that reliable PSQA failure predictors can be successfully developed based solely on MLC leaf positions. Our FT-Transformer neural network can reduce the need for patient rescheduling due to PSQA failures by 80% while sending only 20% of plans that would not have failed the PSQA for replanning. FT-Transformer achieves comparable performance with the leading tree ensemble methods

*This paper was written prior to the author joining Amazon

while having an additional benefit of providing an end-to-end differentiable map from MLC leaf positions to the probability of PSQA failure.

Contents

I. Introduction 1

II. Methods 3

II.A. Data Description 3

II.B. Transformer-based tabular deep learning model 4

III. Results 5

IV. Discussion 7

V. Conclusion 8

VI. Conflict of Interest Statement 8

References 9

A Hyperparameter search spaces 16

A.1. FT-Transformer 16

A.2. Catboost 16

A.3. XGBoost 16

I. Introduction

Intensity-modulated radiation therapy (IMRT)[1] achieves a dose distribution that is highly conformal to the target while minimizing the dose to normal tissue by modulating beam intensities within the radiation fields, often termed fluence maps. The beam modulation is performed using multi-leaf collimators (MLC) located within the gantry of a linear accelerator by varying the speed and position of each leaf and gantry angle.

Leaf sequencing algorithms[2,3,4,5,6,7,8] in the treatment planning system (TPS) optimize the MLC movements to deliver a desirable dose distribution as a treatment planer specifies. Ultimately, final dose distributions to patients are computed using the optimal leaf sequences.

IMRT delivery is a complex, multi-step process with a number of possible sources of noise ranging from computational approximations in the underlying algorithms to physical effects in the linear accelerator components. Therefore, an extensive quality assurance (QA) process is required to prevent any unintended error from reaching the patient and affecting the patient's clinical outcome. It is current practice in many clinics to perform a patient-specific QA (PSQA) for each patient's radiation treatment plan[9,10,11] to ensure that the linear accelerator delivers the correct dose distributions as designed and shown by TPS.

One of the prevalent ways to perform PSQA is using a 3D phantom with an embedded array of detectors to measure the dose delivered using the patient's treatment beams. Then the computed dose distribution in the TPS is compared with the measured dose distribution, and a gamma analysis is performed to quantify the agreement between the two[12,13]. Sometimes, PSQA fails due to a poor agreement between the computed and measured dose distributions requiring a replanning process and another PSQA, which is often done outside clinic hours. PSQA failure can cause increased workloads and stress for hospital staff members, delay patient treatment, or compromise patient safety if the work has to be rushed to preserve the patient's original treatment schedule.

To mitigate those issues and improve patient safety, many studies explored PSQA failure prediction. An extensive line of research focused on developing non-learned treatment plan complexity metrics such as modulation complexity score, mean aperture displacement, or small aperture score and investigating their correlation with PSQA failure[14,15,16,17,18,19,20]. A large number of papers further extended these approaches by developing classical machine

learning and deep learning models to predict the PSQA failure based on a vast array of the plan complexity metrics as well as other heuristic features[21,22,23,24,25,26,27,28]. Thongsawad et al. used MLC texture analysis and boosting algorithms for predicting gamma evaluation results[29]. Kimura *et al.* and Huang *et al.* used target metrics alternative to gamma pass rates, such as dose difference[30,31]. Other works leveraged convolutional neural networks to predict the PSQA failure directly from fluence maps[32] or dose distributions[33,34] obtained from TPS. Since these previous efforts leveraged heuristic feature engineering, their models are not differentiable and are unable to provide a differentiable map from MLC leaf positions to the probability of PSQA plan failure. This means that their models are not applicable to be directly used in the leaf sequencing algorithms to produce MLC positions that are likely to pass PSQA.

In this study, we develop a tabular transformer neural network model FT-Transformer[35] based directly on MLC leaf positions to predict volumetric arc therapy (VMAT) PSQA failure. Using 968 patient plans previously treated with 2–4 VMAT arcs, we trained a regression model to predict the ArcCheck-based PSQA gamma pass rates. We evaluated our model in both the regression context and additionally in the classification context of predicting the pass or fail of PSQA by directly computing receiver operating characteristic (ROC) area under the curve (AUC) on the regression predictions.

We compared the performance of our model with the results from two leading gradient boosted decision tree models in their CatBoost and XGBoost implementations[36,37] widely used for tabular data as well as to a non-learned complexity metric, mean MLC gap.

Neither FT-Transformer nor CatBoost have been used in the context of PSQA failure prediction. Our proposed approach is distinguished from the previous efforts in that we predict PSQA failure directly from MLC leaf positions and the FT-Transformer model we applied is end-to-end differentiable with no heuristic feature engineering. As the MLC leaf positions are the output of leaf sequencing optimization algorithms, our model could be directly leveraged as a differentiable regularizer to improve the leaf sequencing algorithms to produce deliverable treatment plans (i.e., plans with a lower chance of PSQA failure). This is especially useful for the algorithms that employ gradient-based optimization, some of which are implemented in commercial TPS[4,8].

II. Methods

In this section, we describe the pipeline of our study including the description of data collection and processing as well as the models, evaluation metrics and hyperparameter tuning approaches we use. This study was approved by the institutional review board of the University of Washington (STUDY00015736).

II.A. Data Description

We retrospectively collected DICOM-RT PLAN[38] files of 968 patients previously treated with 2 – 4 VMAT arcs using Elekta linear accelerators with Agility collimators between January 2019 and August 2021. All plans were designed in Raystation TPS*. PSQA of each plan was done using ArcCHECK† and the gamma analysis of each PSQA used the criteria of 3% dose difference and 3 mm distance-to-agreement (3%/3mm). We excluded stereotactic body radiotherapy (SBRT) patients since our clinic applies different criteria for the gamma analysis with SBRT patients. We constructed a tabular dataset on beam level leveraging the DICOM-RT PLAN[38] files of the treatment plans to form the samples: for each arc in a treatment plan, we used the leaf and jaw positions of the MLC collimators at each gantry angle.

We aggregated the MLC positions by computing the MLC gap for each leaf-jaw pair at every gantry angle and averaging every 10 neighboring MLC pairs. Additionaly, we averaged the gantry angles over every 8-degree sector. For the labels, we used the ArcCheck-based percentage gamma pass rate of each arc obtained as part of the standard PSQA process in our clinic. To obtain the gamma pass rates, we parsed the ArcCheck-generated PDF reports corresponding to each patient using the `PyPDF2`‡ Python package. As the result, we obtained a tabular regression dataset with 360 purely numerical features and 1873 samples.

For our ultimate goal of PSQA failure prediction, we consider the same data in the classification context by thresholding the regression labels and converting them into binary classification labels. We defined the action threshold level in the gamma analysis to be at 95 % as is common in clinical practice[39,40,41] and obtained binary classification labels (pass

*RaySearch Laboratories
†Sun Nuclear corporation
‡https://pypi.org/project/PyPDF2/

or fail) based on this threshold. We reserved 65% of the samples for the training set, 15% for the validation set and 20% for the test set. To pre-process the data, we normalized the features and regression targets by subtracting their mean over the training set and dividing by their standard deviation over the training set.

II.B. Transformer-based tabular deep learning model

Background of machine learning models for tabular data. Gradient boosted decision trees (GBDT)[36,37,42,43] are the traditionally dominant machine learning approaches for tabular data. These models are commonly used in practice and widely deployed in industry in various domains[44]. Although numerous models have been proposed based on using differentiable ensembles[45,46,47,48,49], leveraging attention-based transformer neural networks[35,50,51,52,53,54], as well as other approaches[55,56,57,58,59,60], recent work on systematic evaluation of deep tabular models[35,44] shows that there is no universally best model capable of consistently outperforming GBDT. Transformer-based models have been shown to be the strongest competitor of GBDT[35,50,54,61,62], especially when coupled with a powerful hyperparameter tuning toolkit[35,63].

Tabular transformer model. We employ the recent transformer-based tabular deep learning method FT-Transformer proposed by Gorishniy *et al.*[35] which has been shown to be the strongest neural network approach in the tabular data domain[35,61]. Additionally, we compare the performance of our model with the gradient boosted decision trees, and we use the popular CatBoost[36] and XGBoost[37] packages.

Evaluation of model performance. We evaluate the models in the regression context of predicting the gamma pass rates as well as in the classification context of predicting the PSQA plan failures. In the regression context, we use mean absolute error (MAE) and root mean squared error (RMSE) metrics as well as Pearson's and Spearman's correlation coefficients between the predictions and the ground truth gamma pass rate values. In the classification context, we use the receiver operating characteristic (ROC) area under the curve (AUC) to evaluate the model performance. We report the beam-level ROC AUC and patient-level ROC AUC. The patient-level predictions and labels are obtained by converting

the beam-level predictions and labels such that a plan is labeled as fail if at least one beam in the plan failed QA. In the classification context we also evaluate the performance of a non-learned baseline approach based on the average MLC gap[15] for comparison.

Hyperparameter tuning. We use the Optuna Bayesian optimization toolkit[63] for hyperparameter tuning. The hyperparameter search spaces for each model are reported in Appendix A. To avoid overfitting, we use early stopping with patience for each model, i.e., we stop training the models if no improvement in the validation score is observed for 30 epochs with FT-Transformer or for 50 boosting rounds with CatBoost and XGBoost.

III. Results

In this section we present the performance of the FT-Transformer model and compare it to the gradient boosted decision trees as well as to the non-learned mean-MLC-gap complexity metric baseline. We investigate the model performance both on the regression task of predicting the ArcCHECK gamma pass rates and the classification task of predicting the QA failure.

Regression results. We first present the performance of all models in predicting the gamma pass rates in Table 1. For each model we present four regression performance metrics: mean absolute error (MAE), root mean squared error (RMSE), Pearson's r and Spearman's r correlation coefficients. FT-Transformer offers competitive performance with CatBoost and XGBoost and all models achieve good results, with e.g. MAE of the gamma rate predictions between 1.4% and 1.53%. The MAE, RMSE, Pearson's r and Spearman's r values are consistent and are on the same order with the results of other studies in the literature[21,22,23,28,32] even though they are not directly comparable given the differences in the experimental setups due to the varying hospital equipment and PSQA processes.

Classification results. The ultimate clinical utility of our models is predicting the PSQA failures to reduce the patient treatment delays and the load on the hospital resources. This practical setup is best emulated by considering our models in the classification context. However, training the models using the regression labels instead of the classification labels

Table 1: **Regression results.** Rows correspond to models and columns correspond to regression metrics.

	MAE (%)	RMSE (%)	Pearson's r	Spearman's r
FT-Transformer	1.44	1.95	0.51	0.51
XGBoost	1.53	1.89	0.58	0.59
CatBoost	1.40	1.84	0.6	0.59

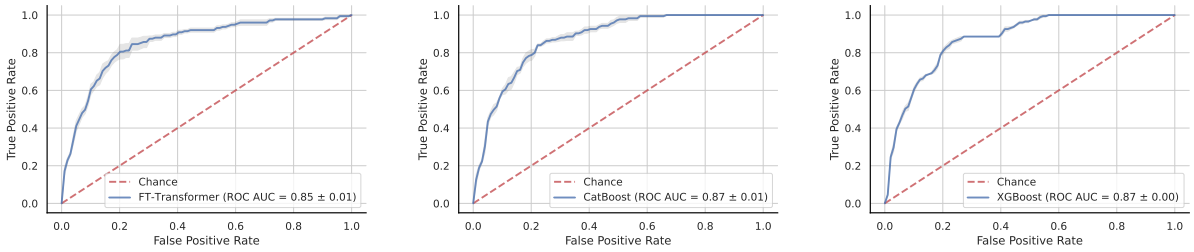
directly allows us to leverage more fine-grained target information and avoid the challenges of severe class imbalance in the classification labels. Nonetheless, the predictions of our regression models could be evaluated in the classification context and we present these results in Table 2. We highlight that Table 2 shows two types of ROC AUC metrics: beam-level and patient-level. As mentioned in section II.B., the patient-level predictions are formed from the beam-level predictions by considering a patient plan to be failed if at least one of the beams in the plan is failed.

Table 2: **Classification results.** Rows correspond to models and columns correspond to classification metrics.

	Beam-level ROC AUC	Patient-level ROC AUC
FT-Transformer	0.82	0.85
XGBoost	0.87	0.87
CatBoost	0.86	0.87
Mean MLC Gap Baseline	0.71	0.72

As the main takeaways of Table 2, we observe that the patient-level ROC AUC classification performance of FT-Transformer is very close to that of CatBoost and XGBoost and that all of the machine learning approaches significantly outperform the Mean-MLC-Gap baseline.

While ROC AUC summarizes the classification performance for all of the prediction thresholds, a particular threshold has to be selected in practice. To investigate this, we further report the patient-level ROC curves for each of the machine learning models in Figure 1. Since missing a failed plan results in patient rescheduling, it is more costly than sending a successful plan for replanning. Therefore, in our clinical scenario it is beneficial to maximize the true positive rate of PSQA failure identification while keeping the false positive rate at a reasonable value. From the shape of the ROC curves in Figure 1, we observe that



(a) FT-Transformer (b) CatBoost (c) XGBoost

Figure 1: **Patient-level ROC curves.** (a) FT-Transformer (b) CatBoost (c) XGBoost. The error bars represent the standard error across 5 seeds. The positive label corresponds to plan failure.

FT-Transformer, CatBoost, and XGBoost serve this purpose well and all allow to achieve 80% true positive rate while keeping the false positive rate under 20%.

IV. Discussion

We demonstrated that PSQA failure prediction is feasible using just the MLC leaf position data without feature engineering. We evaluated the FT transformer model in both regression and classification contexts and found that it outperforms the non-learned model with a mean MLC gap complexity metric, and performs similarly with the two leading gradient boosted decision tree models, CatBoost and XGBoost. The FT-Transformer neural network model, CatBoost, and XGBoost all provide a substantial improvement over the complexity-metric-based baseline. However, the FT-Transformer model comes with a benefit of being end-to-end differentiable, providing a differentiable map from MLC positions to the probability of PSQA failure. Therefore, this model could be leveraged as a differentiable regularizer that allows gradient-based leaf sequencing optimization algorithms to produce a deliverable treatment plan that is likely to pass PSQA.

It is challenging to directly compare models across different studies due to the lack of existing benchmark datasets and there being numerous combinations of TPS, beam models, linear accelerators, MLC designs, and PSQA procedures, all of which can affect the performance, making apple-to-apple comparison difficult. However, we note that our results are consistent with the performance published in the literature[21,22,23,28,32]. Our models achieve classification performance of 0.85-0.87 ROC AUC and are able to identify 80% of treatment

plans that would have failed the PSQA while sending for replanning only up to 20% of successful plans. Using these models in clinical practice can substantially reduce the need for replanning and possibly rescheduling patient due to PSQA failure, which imposes extra workload and stress, and can ultimately compromise patient safety.

Our work was motivated by recognizing the correlation between MLC related complexity metrics and PSQA failures. This leads to the idea of improving leaf sequencing algorithms to produce MLC movements that are more likely to pass PSQA to begin with, which we believe is an improvement from the previous efforts to reduce the frequency of replanning and redoing PSQA by identifying a treatment plan that is likely to fail in the upstream of the workflow, i.e., prior to doing PSQA. We successfully built a model to predict PSQA failure solely based on MLC and jaw positions exploiting recent advances in tabular machine learning models. Incorporating FT-Transformer model in the leaf sequencing algorithms to estimate the potential reduction in the PSQA failure probability of the resulting plans is left for future work.

V. Conclusion

In this work we applied the leading tabular machine learning approaches to the problem of PSQA failure prediction based solely on MLC leaf positions, and obtained effective models which have both direct clinical practice impact to reduce the PSQA failure as well as potential to improve MLC leaf sequencing algorithms to produce treatment plans that are more likely to pass PSQA.

VI. Conflict of Interest Statement

The authors have no relevant conflicts of interest to disclose.

References

[1] J. R. Palta, T. R. Mackie, and R. Lee, Intensity-modulated radiation therapy state of the art, in *Proceedings of the Korean Society of Medical Physics Conference*, pages 4–4, Korean Society of Medical Physics, 2006.

[2] C. Yu, D. Yan, M. Du, S. Zhou, and L. Verhey, Optimization of leaf positions when shaping a radiation field with a multileaf collimator, Physics in Medicine & Biology **40**, 305 (1995).

[3] T. Long, M. Chen, S. Jiang, and W. Lu, Continuous leaf optimization for IMRT leaf sequencing, Medical Physics **43**, 5403–5411 (2016).

[4] A. Cassioli and J. Unkelbach, Aperture shape optimization for IMRT treatment planning, Physics in Medicine & Biology **58**, 301 (2012).

[5] D. M. Shepard, M. A. Earl, X. A. Li, S. Naqvi, and C. Yu, Direct aperture optimization: a turnkey solution for step-and-shoot IMRT, Medical physics **29**, 1007–1018 (2002).

[6] D. A. Granville, J. G. Sutherland, J. G. Belec, and D. J. La Russa, Predicting VMAT patient-specific QA results using a support vector classifier trained on treatment plan characteristics and linac QC metrics, Physics in Medicine & Biology **64**, 095017 (2019).

[7] M. Earl, M. Afghan, C. Yu, Z. Jiang, and D. Shepard, Jaws-only IMRT using direct aperture optimization, Medical physics **34**, 307–314 (2007).

[8] B. Hardemark, A. Liander, H. Rehbinder, and J. Löf, Direct machine parameter optimization with RayMachine in Pinnacle, Ray-Search White Paper (2003).

[9] T. LoSasso, C.-S. Chui, and C. C. Ling, Comprehensive quality assurance for the delivery of intensity modulated radiotherapy with a multileaf collimator used in the dynamic mode, Medical physics **28**, 2209–2219 (2001).

[10] G. A. Ezzell, J. M. Galvin, D. Low, J. R. Palta, I. Rosen, M. B. Sharpe, P. Xia, Y. Xiao, L. Xing, and C. X. Yu, Guidance document on delivery, treatment planning, and clinical

implementation of IMRT: report of the IMRT Subcommittee of the AAPM Radiation Therapy Committee, Medical physics **30**, 2089–2115 (2003).

[11] D. A. Low, J. M. Moran, J. F. Dempsey, L. Dong, and M. Oldham, Dosimetry tools and techniques for IMRT, Medical physics **38**, 1313–1338 (2011).

[12] D. A. Low, W. B. Harms, S. Mutic, and J. A. Purdy, A technique for the quantitative evaluation of dose distributions, Medical physics **25**, 656–661 (1998).

[13] D. A. Low and J. F. Dempsey, Evaluation of the gamma dose distribution comparison method, Medical physics **30**, 2455–2464 (2003).

[14] K. C. Younge, D. Roberts, L. A. Janes, C. Anderson, J. M. Moran, and M. M. Matuszak, Predicting deliverability of volumetric-modulated arc therapy (VMAT) plans using aperture complexity analysis, Journal of applied clinical medical physics **17**, 124–131 (2016).

[15] S. Crowe, T. Kairn, N. Middlebrook, B. Sutherland, B. Hill, J. Kenny, C. M. Langton, and J. Trapp, Examination of the properties of IMRT and VMAT beams and evaluation against pre-treatment quality assurance results, Physics in Medicine & Biology **60**, 2587 (2015).

[16] J. M. Park, S.-Y. Park, H. Kim, J. H. Kim, J. Carlson, and S.-J. Ye, Modulation indices for volumetric modulated arc therapy, Physics in Medicine & Biology **59**, 7315 (2014).

[17] S. Crowe, T. Kairn, J. Kenny, R. Knight, B. Hill, C. M. Langton, and J. Trapp, Treatment plan complexity metrics for predicting IMRT pre-treatment quality assurance results, Australasian physical & engineering sciences in medicine **37**, 475–482 (2014).

[18] L. Masi, R. Doro, V. Favuzza, S. Cipressi, and L. Livi, Impact of plan parameters on the dosimetric accuracy of volumetric modulated arc therapy, Medical physics **40**, 071718 (2013).

[19] J. Park, H. Wu, J. Kim, J. Carlson, and K. Kim, The effect of MLC speed and acceleration on the plan delivery accuracy of VMAT, The British journal of radiology **88**, 20140698 (2015).

[20] M. Antoine, F. Ralite, C. Soustiel, T. Marsac, P. Sargos, A. Cugny, and J. Caron, Use of metrics to quantify IMRT and VMAT treatment plan complexity: A systematic review and perspectives, Physica Medica **64**, 98–108 (2019).

[21] J. Li, L. Wang, X. Zhang, L. Liu, J. Li, M. F. Chan, J. Sui, and R. Yang, Machine learning for patient-specific quality assurance of VMAT: prediction and classification accuracy, International Journal of Radiation Oncology* Biology* Physics **105**, 893–902 (2019).

[22] L. Wang, J. Li, S. Zhang, X. Zhang, Q. Zhang, M. F. Chan, R. Yang, and J. Sui, Multi-task autoencoder based classification-regression model for patient-specific VMAT QA, Physics in Medicine & Biology **65**, 235023 (2020).

[23] H. Hirashima, T. Ono, M. Nakamura, Y. Miyabe, N. Mukumoto, H. Iramina, and T. Mizowaki, Improvement of prediction and classification performance for gamma passing rate by using plan complexity and dosiomics features, Radiotherapy and Oncology **153**, 250–257 (2020).

[24] R. Yang 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 **161**, 230–240 (2021).

[25] J. C. Lizar, C. C. Yaly, A. C. Bruno, G. A. Viani, and J. F. Pavoni, Patient-specific IMRT QA verification using machine learning and gamma radiomics, Physica Medica **82**, 100–108 (2021).

[26] T. Kairn, S. Crowe, J. Kenny, R. Knight, and J. Trapp, Predicting the likelihood of QA failure using treatment plan accuracy metrics, in *Journal of Physics: Conference Series*, volume 489, page 012051, IOP Publishing, 2014.

[27] T. Kusunoki, S. Hatanaka, M. Hariu, Y. Kusano, D. Yoshida, H. Katoh, M. Shimbo, and T. Takahashi, Evaluation of prediction and classification performances in different machine learning models for patient-specific quality assurance of head-and-neck VMAT plans, Medical physics **49**, 727–741 (2022).

[28] D. Lam, X. Zhang, H. Li, Y. Deshan, B. Schott, T. Zhao, W. Zhang, S. Mutic, and B. Sun, Predicting gamma passing rates for portal dosimetry-based IMRT QA using machine learning, Medical physics **46**, 4666–4675 (2019).

[29] S. Thongsawad, S. Srisatit, and T. Fuangrod, Predicting gamma evaluation results of patient-specific head and neck volumetric-modulated arc therapy quality assurance based on multileaf collimator patterns and fluence map features: A feasibility study, Journal of Applied Clinical Medical Physics , e13622 (2022).

[30] Y. Kimura, N. Kadoya, Y. Oku, T. Kajikawa, S. Tomori, and K. Jingu, Error detection model developed using a multi-task convolutional neural network in patient-specific quality assurance for volumetric-modulated arc therapy, Medical Physics **48**, 4769–4783 (2021).

[31] Y. Huang et al., Virtual Patient-Specific Quality Assurance of IMRT Using UNet++: Classification, Gamma Passing Rates Prediction, and Dose Difference Prediction, Frontiers in Oncology , 2798 (2021).

[32] S. Tomori, N. Kadoya, T. Kajikawa, Y. Kimura, K. Narazaki, T. Ochi, and K. Jingu, Systematic method for a deep learning-based prediction model for gamma evaluation in patient-specific quality assurance of volumetric modulated arc therapy, Medical Physics **48**, 1003–1018 (2021).

[33] T. Matsuura, D. Kawahara, A. Saito, H. Miura, K. Yamada, S. Ozawa, and Y. Nagata, Predictive gamma passing rate of 3D detector array-based volumetric modulated arc therapy quality assurance for prostate cancer via deep learning, (2022).

[34] S. Tomori, N. Kadoya, Y. Takayama, T. Kajikawa, K. Shima, K. Narazaki, and K. Jingu, A deep learning-based prediction model for gamma evaluation in patient-specific quality assurance, Medical physics **45**, 4055–4065 (2018).

[35] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, Revisiting Deep Learning Models for Tabular Data, arXiv preprint arXiv:2106.11959 (2021).

[36] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, CatBoost: unbiased boosting with categorical features, Advances in neural information processing systems **31** (2018).

[37] T. Chen and C. Guestrin, Xgboost: A scalable tree boosting system, in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[38] M. Y. Law and B. Liu, DICOM-RT and its utilization in radiation therapy, Radiographics **29**, 655–667 (2009).

[39] G. H. Chan, L. C. Chin, A. Abdellatif, J.-P. Bissonnette, L. Buckley, D. Comsa, D. Granville, J. King, P. L. Rapley, and A. Vandermeer, Survey of patient-specific quality assurance practice for IMRT and VMAT, Journal of Applied Clinical Medical Physics **22**, 155–164 (2021).

[40] Y. Pan, R. Yang, S. Zhang, J. Li, J. Dai, J. Wang, and J. Cai, National survey of patient specific IMRT quality assurance in China, Radiation Oncology **14**, 1–10 (2019).

[41] H. Mehrens, P. Taylor, D. S. Followill, and S. F. Kry, Survey results of 3D-CRT and IMRT quality assurance practice, Journal of applied clinical medical physics **21**, 70–76 (2020).

[42] J. H. Friedman, Greedy function approximation: a gradient boosting machine, Annals of statistics , 1189–1232 (2001).

[43] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, Advances in neural information processing systems **30** (2017).

[44] R. Shwartz-Ziv and A. Armon, Tabular data: Deep learning is not all you need, Information Fusion **81**, 84–90 (2022).

[45] S. Popov, S. Morozov, and A. Babenko, Neural oblivious decision ensembles for deep learning on tabular data, arXiv preprint arXiv:1909.06312 (2019).

[46] H. Hazimeh, N. Ponomareva, P. Mol, Z. Tan, and R. Mazumder, The tree ensemble layer: Differentiability meets conditional computation, in *International Conference on Machine Learning*, pages 4138–4148, PMLR, 2020.

[47] Y. Yang, I. G. Morillo, and T. M. Hospedales, Deep neural decision trees, arXiv preprint arXiv:1806.06988 (2018).

[48] P. Kontschieder, M. Fiterau, A. Criminisi, and S. R. Bulo, Deep neural decision forests, in *Proceedings of the IEEE international conference on computer vision*, pages 1467–1475, 2015.

[49] S. Badirli, X. Liu, Z. Xing, A. Bhowmik, K. Doan, and S. S. Keerthi, Gradient boosting neural networks: Grownet, arXiv preprint arXiv:2002.07971 (2020).

[50] G. Somepalli, M. Goldblum, A. Schwarzschild, C. B. Bruss, and T. Goldstein, SAINT: Improved Neural Networks for Tabular Data via Row Attention and Contrastive Pre-Training, arXiv preprint arXiv:2106.01342 (2021).

[51] S. O. Arık and T. Pfister, Tabnet: Attentive interpretable tabular learning, in *AAAI*, volume 35, pages 6679–6687, 2021.

[52] X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin, Tabtransformer: Tabular data modeling using contextual embeddings, arXiv preprint arXiv:2012.06678 (2020).

[53] W. Song, C. Shi, Z. Xiao, Z. Duan, Y. Xu, M. Zhang, and J. Tang, Autoint: Automatic feature interaction learning via self-attentive neural networks, in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1161–1170, 2019.

[54] J. Kossen, N. Band, C. Lyle, A. N. Gomez, T. Rainforth, and Y. Gal, Self-attention between datapoints: Going beyond individual input-output pairs in deep learning, Advances in Neural Information Processing Systems **34** (2021).

[55] R. Wang, B. Fu, G. Fu, and M. Wang, Deep & cross network for ad click predictions, in *Proceedings of the ADKDD'17*, pages 1–7, 2017.

[56] R. Wang, R. Shivanna, D. Cheng, S. Jain, D. Lin, L. Hong, and E. Chi, DCN V2: Improved deep & cross network and practical lessons for web-scale learning to rank systems, in *Proceedings of the Web Conference 2021*, pages 1785–1797, 2021.

[57] A. Beutel, P. Covington, S. Jain, C. Xu, J. Li, V. Gatto, and E. H. Chi, Latent cross: Making use of context in recurrent recommender systems, in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 46–54, 2018.

[58] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, Self-normalizing neural networks, Advances in neural information processing systems **30** (2017).

[59] J. Fiedler, Simple modifications to improve tabular neural networks, arXiv preprint arXiv:2108.03214 (2021).

[60] B. Schäfl, L. Gruber, A. Bitto-Nemling, and S. Hochreiter, Hopular: Modern Hopfield Networks for Tabular Data, (2021).

[61] Y. Gorishniy, I. Rubachev, and A. Babenko, On Embeddings for Numerical Features in Tabular Deep Learning, arXiv preprint arXiv:2203.05556 (2022).

[62] R. Levin, V. Cherepanova, A. Schwarzschild, A. Bansal, C. B. Bruss, T. Goldstein, A. G. Wilson, and M. Goldblum, Transfer Learning with Deep Tabular Models, arXiv preprint arXiv:2206.15306 (2022).

[63] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.

A Hyperparameter search spaces

A.1. FT-Transformer

The number of attention heads is always set to 8.

Table 3: **Optuna hyperparameter search space for FT-Transformer**

Parameter	Search Space
Number of layers	UniformInt$[1, 4]$
Feature embedding size	UniformInt$[64, 512]$
Residual dropout	$\{0,$ Uniform$[0, 0.2]\}$
Attention dropout	Uniform$[0, 0.5]$
FFN dropout	Uniform$[0, 0.5]$
FFN factor	Uniform$[2/3, 8/3]$
Learning rate	LogUniform$[1e-5, 1e-3]$
Weight decay	LogUniform$[1e-6, 1e-3]$

A.2. Catboost

The hyperparameter search space and distributions are presented in Table 4.

Table 4: **Optuna hyperparameter search space for Catboost**

Parameter	Search Space
Max depth	UniformInt$[3, 10]$
Learning rate	LogUniform$[1e-5, 1]$
Bagging temperature	Uniform$[0, 1]$
L2 leaf reg	LogUniform$[1, 10]$
Leaf estimation iterations	UniformInt$[1, 10]$

A.3. XGBoost

The hyperparameter search space and distributions are presented in Table 5.

Table 5: **Optuna hyperparameter search space for XGBoost**

Parameter	Search Space
Max depth	UniformInt$[3, 10]$
Min child weight	LogUniform$[1e-8, 1e5]$
Subsample	Uniform$[0.5, 1]$
Learning rate	LogUniform$[1e-5, 1]$
Col sample by level	Uniform$[0.5, 1]$
Col sample by tree	Uniform$[0.5, 1]$
Gamma	$\{0$, LogUniform$[1e-8, 1e2]\}$
Lambda	$\{0$, LogUniform$[1e-8, 1e2]\}$
Alpha	$\{0$, LogUniform$[1e-8, 1e2]\}$