Radiation Oncology

RESEARCH

Open Access

Improving the performance of deep learning models in predicting and classifying gamma passing rates with discriminative features and a class balancing technique: a retrospective cohort study

Wei Song[1], Wen Shang[1], Chunying Li[1], Xinyu Bian[1], Hong Lu[1], Jun Ma[1*] and Dahai Yu[1*]

Abstract

Background The purpose of this study was to improve the deep learning (DL) model performance in predicting and classifying IMRT gamma passing rate (GPR) by using input features related to machine parameters and a class balancing technique.

Methods A total of 2348 fields from 204 IMRT plans for patients with nasopharyngeal carcinoma were retrospectively collected to form a dataset. Input feature maps, including fluence, leaf gap, leaf speed of both banks, and corresponding errors, were constructed from the dynamic log files. The SHAP framework was employed to compute the impact of each feature on the model output for recursive feature elimination. A series of UNet++ based models were trained on the obtained eight feature sets with three fine-tuning methods including the standard mean squared error (MSE) loss, a re-sampling technique, and a proposed weighted MSE loss (WMSE). Differences in mean absolute error, area under the receiver operating characteristic curve (AUC), sensitivity, and specificity were compared between the different models.

Results The models trained with feature sets including leaf speed and leaf gap features predicted GPR for failed fields more accurately than the other models ($F(7, 147) = 5.378$, $p < 0.001$). The WMSE loss had the highest accuracy in predicting GPR for failed fields among the three fine-tuning methods ($F(2, 42) = 14.149$, $p < 0.001$), while an opposite trend was observed in predicting GPR for passed fields ($F(2, 730) = 9.907$, $p < 0.001$). The WMSE_FS5 model achieved a superior AUC (0.92) and more balanced sensitivity (0.77) and specificity (0.89) compared to the other models.

Conclusions Machine parameters can provide discriminative input features for GPR prediction in DL. The novel weighted loss function demonstrates the ability to balance the prediction and classification accuracy between the passed and failed fields. The proposed approach is able to improve the DL model performance in predicting and classifying GPR, and can potentially be integrated into the plan optimization process to generate higher deliverability plans.

*Correspondence:
Jun Ma
bluemaple@sina.com
Dahai Yu
yudahaipumc@hotmail.com
Full list of author information is available at the end of the article

Trial registration: This clinical trial was registered in the Chinese Clinical Trial Registry on March 26th, 2020 (registration number: ChiCTR2000031276). https://clinicaltrials.gov/ct2/show/ChiCTR2000031276

Keywords Deep learning, Prediction, Classification, Quality assurance, Machine parameters, Class imbalance

Introduction

Intensity modulated radiation therapy (IMRT) is an advanced form of radiotherapy that can deliver highly conformal dose distributions to the tumor while minimizing dose to surrounding normal tissues [1]. Due to the increasing complexity of IMRT planning and delivery, patient-specific quality assurance (QA) is an essential process employed to verify the accuracy of IMRT plan dose calculations and to detect clinically relevant errors in radiation delivery, thereby ensuring the safety and efficacy of radiation treatment [2]. In clinical practice, patient-specific QA is commonly performed prior to the initiation of patient treatment with various measurement-based methods, including film dosimetry, electronic portal imaging device, two-dimensional ionization chamber array, and three-dimensional dosimetric systems, etc. [3]. To quantitatively evaluate the agreement between measured and calculated dose distributions, the gamma analysis method is commonly used to calculate the gamma passing rates for each measured plan or field [4]. However, as treatment planning becomes more efficient and the number of patients treated with advanced radiotherapy techniques steadily increases, this measurement-based IMRT plan verification procedure requires a substantial clinical workload and is often time-consuming and laborious [5].

In recent years, there has been growing interest in applying machine learning (ML) or deep learning (DL) algorithms to predict patient-specific QA outcomes for IMRT/VMAT plans. Several previous studies have applied ML methods to predict IMRT or VMAT delivery accuracy by mapping selected input features to patient-specific QA outcomes of interest. Valdes et al. [6] trained a Poisson regression model with Lasso regularization on a set of 78 aperture-based complexity metrics, which was found to be able to predict gamma passing rates (GPR) within 3% accuracy using a gamma criterion of 3%local/3 mm for static gantry IMRT plans. Wall et al. [7] observed that a support vector machine was the best model for predicting GPR for VMAT plans based on 100 treatment planning features compared to other ML algorithms. Other traditional ML models, including tree-based models [7–9], logistic regression [10], random forest [7, 11], Naïve-Bayes [11, 12], neural networks [13], and regression tree analysis [13], trained on handcrafted complexity features or radiomic texture features extracted from fluence/dose maps, have also been reported to achieve preliminary success in predicting IMRT/VMAT plan delivery accuracy. Unlike those models developed based on handcrafted features, DL algorithms demonstrate the ability to automatically learn representations from the input data without the need for human domain knowledge. Studies have shown that DL models can achieve higher prediction performance than traditional ML algorithms that use handcrafted features in virtual IMRT QA. However, only limited input feature maps, such as fluence or dose maps, have been fed into the DL networks to predict QA outcomes in previous studies [14–16]. As the main sources of error, spatial and dosimetric delivery system uncertainties also affect the accuracy of IMRT plan delivery [2]. By constructing input feature maps related to the delivery process, additional discriminative information can be provided to improve the prediction and classification performance of DL models.

Class imbalance is a common issue in DL studies. DL models trained on imbalanced datasets tend to perform significantly better on instance-rich classes than on instance-scarce classes [17, 18]. As the collected patient-specific QA training datasets are dominated by negative samples (passed fields or plans), previous studies have reported poorer prediction performance for positive samples (failed fields or plans) using DL models. Samples with lower measured GPR were found to have a higher mean absolute error (MAE) in GPR prediction under different gamma criteria [14, 19]. Methods for dealing with the class imbalance problem faced by studies applying DL algorithms to predict patient-specific QA accuracy have rarely been investigated so far.

In this study, a UNet++ architecture based DL neural network is proposed to predict IMRT field GPR. In addition to the commonly used fluence map, input feature maps that reflect machine delivery parameters are constructed to provide meaningful information for improving model performance. A Shapley-based framework is utilized to interpret the effect of each feature on the model prediction and to identify the best combination of feature maps. A novel re-weighting method is introduced and compared to a re-sampling method to evaluate its effectiveness in alleviating the class imbalanced issue.

Materials and methods

This study was approved by the Ethics Committee of Affiliated Hospital of Nanjing University of Chinese Medicine (Jiangsu Province Hospital of Chinese Medicine), and written informed consent was obtained from all the patients. The general method of the processes performed in this research is shown in Fig. 1.

Treatment planning and data collection

A total of 2348 fields from 204 static gantry IMRT plans for patients with nasopharyngeal carcinoma, treated at the Jiangsu Province Hospital of Chinese Medicine from August 2020 to December 2023, were retrospectively collected to form a dataset. All plans were generated in the Eclipse 15.6 TPS using the configured 6-MV photon beam data for a Clinac iX linear accelerator (Varian Medical Systems, Palo Alto, CA). This linac is equipped with a Millennium 120 multileaf collimator (MLC), which has 40 central leaf pairs with a projected width of 5 mm and 20 outer leaf pairs with a projected width of 10 mm at the isocenter. The sliding window technique was selected for treatment delivery at a dose rate of 500 MU/min. All dose distributions were calculated using the analytical anisotropic algorithm with a 2.5-mm grid size including corrections for tissue heterogeneity. Patient-specific quality assurance was performed with a two-dimensional ionization chamber array PTW OCTAVIUS Detector 729 (PTW, Freiburg, Germany), which consists of a matrix of 729 cubic vented ionization chambers with 5 mm × 5 mm cross section, covering an area of 27 cm × 27 cm. The distance between the centers of adjacent ionization chambers is 1 cm. The Detector array was positioned in the isocenter plane at a depth of 5 cm in a water-equivalent RW3 phantom with 5 cm of backscatter RW3 material. Each day, before starting the patient-specific QA delivery, a reference field (10 cm × 10 cm at the isocenter) was delivered and a cross-calibration factor was calculated as the ratio of the expected dose to the measured dose to the central chamber. All subsequent measurements were corrected by the calibration factor to eliminate daily linac output fluctuation. The verification plans were created by transferring each field of the treatment plans to the phantom CT. The gantry, collimator, and couch angles were all reset to 0°. The verification plans were calculated using the same algorithm and parameters as the treatment plans. After delivery of the verification plans, the gamma passing rates for each field were calculated with PTW VeriSoft 6.0 software (PTW, Freiburg, Germany) at a 2%global/2 mm criterion, and the measurement points below 10% of the maximum dose value were excluded. Previous studies have reported that the 2%/2 mm criterion is the most sensitive criterion for detecting clinically relevant errors [20, 21], and hence it was used in this study to develop the prediction models. Our institution-specific action limits (AL) for patient-specific QA were calculated based on the recommendations of the AAPM Task Group (TG)-218 report. For fixed gantry IMRT QA using the PTW OCTAVIUS Detector 729 array at our institution, the AL was determined to be 91% at the 2%global/2 mm criterion according to Eq. (3) in the TG-218 report [2]. Fields with GPR values below the AL are defined as failing QA, while those with GPR values above the AL are defined as passing QA.

Input feature construction

The dynamic log (Dynalog) files generated by the MLC controller during the verification plan delivery were collected for each delivered field. These files contain relevant information such as the dose fraction, expected and actual leaf positions, collimator jaw positions, and

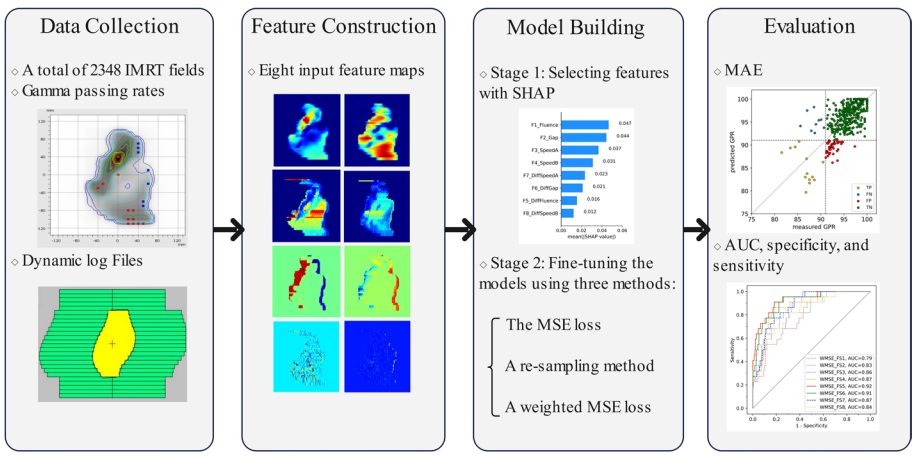

Fig. 1 The overall workflow for the study

the machine status (beam-on or beam hold-off) sampled every 50 ms. As shown in Table 1, eight input feature maps were constructed from the raw data extracted from the Dynalog files using an in-house Matlab script adapted from the Dynalog File Analyser [22]. Each map covers the 27 cm × 27 cm measurement region in the isocenter plane with a 256 × 256 matrix. The MLC leaf gap and leaf speed features were chosen based on their strong correlation with plan complexity, according to our previous experience [23]. The input features related to the difference between actual and planned parameters directly reflect the spatial and dosimetric uncertainties of the linac in the dynamic process of treatment delivery [2]. Most importantly, the simple form of feature definition can preserve the crucial discriminative information in the raw data as much as possible for model training. The detailed definition of the input feature maps can be found in the supplementary material.

Architecture of the proposed network

The UNet++ architecture has been widely used for feature extraction and classification in segmentation and detection tasks, achieving significant performance gain over U-Net proposed by Ronneberger et al. [24–26]. By introducing intermediate layers and redesigned dense skip connections, UNet++ can reduce the semantic gap between the encoder and decoder feature maps. The schematic diagram of the UNet++ based architecture used in this study is shown in Fig. 2a. The network consists of convolution blocks, down-sampling layers, up-sampling layers, and skip connections. Each node X*i,j* in the graph represents a basic convolution block, where *i* denotes the *i*th down-sampling layer along the encoder and *j* indexes the convolution layer along the skip pathway. Generally, nodes at level $j=0$ receive only one input from the previous layer of the encoder, while nodes at level $j>0$ receive and concatenate $j+1$ inputs, of which j inputs are the outputs of the previous j nodes of the dense block in the same skip pathway and the last input is the up-sampled output from the lower skip pathway. As can be seen in Fig. 2b, the basic convolution block consists of two stacked Squeeze-and-Excitation (SE) residual blocks, a batch normalization layer, and a ReLU activation layer. The SE block is integrated into the architecture and placed before the residual block, which allows the proposed network to focus on important features and suppress less useful ones by adaptively recalibrating channel-wise feature responses [27]. The residual block is constructed using stacked batch normalization, ReLU, and 3 × 3 convolution layers with a shortcut connection

Table 1 Description of the designed input feature maps

ID	Input feature map	Description
F1	*Fluence*	The planned fluence to be delivered to each pixel in the matrix, including that delivered through the leaf gap opening, transmitted through the MLC leaves, and transmitted through the jaw collimators
F2	*Gap*	The planned average width of the leaf gaps passing each pixel in the matrix
F3	*SpeedA*	The planned leaf speed for bank A leaves passing each pixel in the matrix
F4	*SpeedB*	The planned leaf speed for bank B leaves passing each pixel in the matrix
F5	*DiffFluence*	The difference between the actual and planned fluence delivered to each pixel in the matrix
F6	*DiffGap*	The difference between the actual and planned average width of the leaf gaps passing each pixel in the matrix
F7	*DiffSpeedA*	The difference between the actual and planned leaf speed for bank A leaves passing each pixel in the matrix
F8	*DiffSpeedB*	The difference between the actual and planned leaf speed for bank B leaves passing each pixel in the matrix

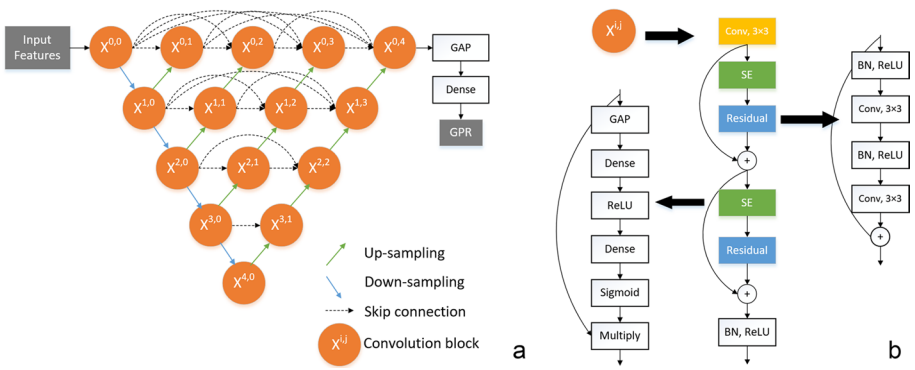

Fig. 2 **a** The proposed deep neural network based on the UNet++ architecture. **b** The structure of the basic convolution block

between the block input and the output of the second convolution layer. Residual learning has proven to be effective in overcoming the optimization difficulties and finding an optimal solution. The input feature maps are fed into the first node of UNet++ and the learned representations are finally passed through a global average pooling (GAP) layer and a dense layer to generate the predicted GPR.

Feature selection

As the number of input features increases, model training may be prone to over-fitting, so it is necessary to choose an appropriate combination of extracted input features to maximize model performance. To rank the importance of each input feature for model prediction, we used a unified Shapley-based framework called SHAP (SHapley Additive exPlanations) to assign an importance value to each input feature [28]. Shapley values come from the game theory literature and provide a theoretically justified method for assigning the impact of each input feature on the model prediction [29, 30]. Feature impact is defined as the change in the expected value of the model output when a feature is present or absent [28]. To identify the optimal set of input features, a model was trained using the full set of input features. Then, the least important feature was recursively excluded from the current feature set (FS), and the remaining features were used to train a new model from scratch. Finally, a total of eight models trained with a reduced number of input features were generated.

Handling class imbalance

According to the locally established AL for this IMRT procedure, only 124 fields out of the collected dataset (N = 2348) failed the gamma analysis. Therefore, the passed fields (N = 2224) dominated the dataset compared to the failed fields. This is a common problem in real-world applications of DL methods, known as the class imbalance problem [31, 32]. To address this issue, a two-stage training procedure was used in this study. The first stage is the feature selection process. All models were initially trained to convergence using the standard MSE loss. In the second stage, three fine-tuning methods, including the use of the mean squared error (MSE) loss without re-sampling, the MSE loss with re-sampling (SMSE), and a proposed weighted MSE loss without re-sampling (WMSE), were separately implemented to further fine-tune the models obtained in the first stage. For the SMSE method, we adopted the most commonly used over-sampling technique by randomly replicating the failed cases in the dataset so that the number of passed and failed cases was approximately equal. Furthermore, inspired by the focal loss proposed by Lin et al. [33], we proposed a WMSE loss as follows:

$$\text{loss} = \alpha_t \cdot \beta \cdot |t - p|^2 \tag{1}$$

$$\alpha_t = \left(\frac{1-\alpha}{\alpha}\right)^{-\left(t-\log_{\left(\frac{1-\alpha}{\alpha}\right)}(1-\alpha)\right)} \tag{2}$$

$$\beta = [(p-\theta)\cdot(\mathrm{I}(p \geq \theta) - \mathrm{I}(t \geq \theta)) + 1]^{\gamma} \tag{3}$$

where t is the true GPR value of a given IMRT field, scaled by min–max scaling to the range [0, 1], and p is the corresponding predicted GPR value. θ is the AL for evaluating QA results. α_t is a weighting factor that balances the importance of passed/failed fields. It is equal to $1-\alpha$ for the minimum GPR field ($t=0$) and decays exponentially to $\alpha \in [0, 0.5)$ for the maximum GPR field ($t=1$). β is a factor introduced to penalize the misclassification of the QA results. $\mathrm{I}(\cdot)$ is an indicator function that equals 1 if the event in the parentheses occurs, and 0 otherwise. For passed fields ($t \geq \theta$), if the predicted GPR is lower than the AL ($p \leq \theta$) (i.e. misclassified examples), the term $(p-\theta)\cdot(\mathrm{I}(p \geq \theta) - \mathrm{I}(t \geq \theta))$ in Eq. (3) equals $|p-\theta|$. Thus, β is proportional to the absolute deviation of the predicted GPR from the AL, and the effect of penalization is increased by increasing the tunable hyperparameter γ ($\gamma \geq 0$). However, if the passed fields are correctly classified ($p \geq \theta$), β is equal to 1 and thus the loss is unaffected. These properties also hold for failed fields. When α_t and β are set to 1 in Eq. (1), the WMSE loss becomes equivalent to the standard MSE loss.

Training

The collected dataset was split into a training/validation set (N = 1960) and a test set (N = 388) using a stratification technique to ensure that each set had approximately the same GPR distribution. The DL models were built using TensorFlow 2.4 and the Keras API and trained on an NVIDIA RTX 3080Ti GPU card with 12 GB of dedicated memory. The Adam algorithm was chosen as the optimizer to minimize the loss function. The learning rate was initially set at 10^{-4} and reduced by a factor of 0.1 when no improvement in the validation loss was observed after 10 epochs. The models were trained using a mini-batch method with a batch size of 4. An early stopping callback was used to stop training the models when the validation loss did not improve for 20 epochs. L2 regularization was also utilized to prevent over-fitting. All hyperparameters were tuned using grid search and fivefold cross validation. The L2 regularization coefficient was set to 10^{-4}. α and γ were set to 0.2 and 2, respectively. The model with the best performance on the validation

set over the five folds was selected for the final evaluation on the test set.

Statistical analysis

Statistical analysis was performed using SPSS Statistics26.0 software (IBM Corp., Armonk, NY, USA). As the Shapiro–Wilk test indicated that the absolute error of the predicted GPR was not normally distributed, differences in this variable between different models were analyzed by two-way repeated measures analysis of variance (ANOVA) after applying the aligned rank transform (ART) procedure to the data [34]. Post-hoc pairwise comparisons of group means were performed using the Bonferroni test. A p-value < 0.05 (two-tailed) was considered statistically significant.

Results

Feature selection

Figure 3a–h show the corresponding eight feature sets (FS1-8) obtained from the recursive feature elimination. It can be seen that the input features of F5_DiffFluence, F6_DiffGap, F7_DiffSpeedA, and F8_DiffSpeedB had a relatively lower mean SHAP value compared to the other four features, indicating a lower impact of machine delivery errors on the model's prediction, while the machine parameters themselves had a higher impact on the prediction result.

Prediction accuracy

Table 2 summarizes the MAE of the predicted GPR for different models in the test dataset. Figure 4 shows the discrepancies between the measured and predicted GPR in the test dataset for models trained with the full FS (FS1), the optimal FS (FS5), and the FS with only the *Fluence* map (FS8) using each of the three fine-tuning methods. As shown in Table 2, the results of the two-way repeated measures ANOVA revealed that there was no statistically significant interaction between the effects of input FS and fine-tuning method on model prediction accuracy for all fields ($F(14, 5418) = 0.110$, $p = 1.000$), passed fields ($F(14, 5110) = 0.123$, $p = 1.000$), and failed fields ($F(14, 294) = 0.102$, $p = 1.000$). Main effects analysis showed that neither the input FS ($F(7, 2709) = 1.925$, $p = 0.062$) nor the fine-tuning method ($F(2, 774) = 2.209$, $p = 0.111$) had a statistically significant effect on the MAE of the predicted GPR for all fields. For passed fields, the MAE of the predicted GPR values was not significantly different between the models trained with different input FS ($F(7, 2555) = 0.481$, $p = 0.849$), but was significantly different between the models trained with different fine-tuning methods ($F(2, 730) = 9.907$, $p < 0.001$). For failed fields, there was a significant difference in the MAE of predicted GPR values between the models trained with different input FS ($F(7, 147) = 5.378$, $p < 0.001$) and between the models trained with different fine-tuning methods ($F(2, 42) = 14.149$, $p < 0.001$). Post-hoc analyses showed that the models trained with FS4, FS5, FS6, or FS7 predicted the GPR for failed fields more accurately than the other models trained with more or fewer input features. Additionally, the models trained with the WMSE loss had the highest accuracy in predicting GPR for failed fields among the three fine-tuning methods and the models trained with MSE loss had the lowest accuracy, while an opposite

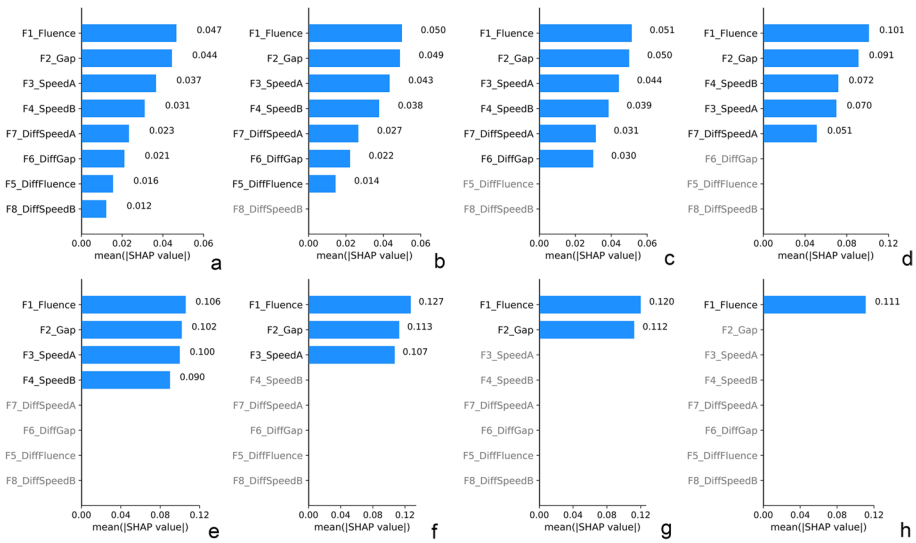

Fig. 3 Mean SHAP values computed on the validation set for each input feature in each of the eight models **a**–**h** trained during recursive feature elimination

Table 2 Comparison of the mean absolute error of the predicted GPR for different models in the test dataset. (mean ± standard deviation)

	All fields (N = 388)			Passed fields (N = 366)			Failed fields (N = 22)		
	MSE	SMSE	WMSE	MSE	SMSE	WMSE	MSE	SMSE	WMSE
Absolute error (μ ± σ) (%)									
FS1	2.17 ± 1.71	2.23 ± 1.66	2.28 ± 1.52	1.89 ± 1.22	1.97 ± 1.26	2.07 ± 1.23	6.91 ± 1.73	6.50 ± 1.78	5.76 ± 1.76
FS2	2.19 ± 1.74	2.20 ± 1.72	2.25 ± 1.64	1.90 ± 1.25	1.96 ± 1.38	2.05 ± 1.39	6.93 ± 1.84	6.26 ± 1.86	5.58 ± 1.99
FS3	2.17 ± 1.70	2.15 ± 1.63	2.19 ± 1.55	1.89 ± 1.22	1.92 ± 1.30	2.00 ± 1.32	6.76 ± 1.94	5.99 ± 1.84	5.34 ± 1.74
FS4	2.07 ± 1.65	2.14 ± 1.50	2.17 ± 1.48	1.84 ± 1.31	1.95 ± 1.24	2.01 ± 1.29	5.92 ± 1.90	5.31 ± 1.86	4.83 ± 1.89
FS5	2.02 ± 1.50	2.08 ± 1.41	2.11 ± 1.41	1.80 ± 1.16	1.91 ± 1.18	1.98 ± 1.27	5.64 ± 1.84	4.98 ± 1.77	4.32 ± 1.82
FS6	2.09 ± 1.48	2.08 ± 1.39	2.10 ± 1.44	1.88 ± 1.17	1.90 ± 1.12	1.95 ± 1.25	5.50 ± 1.89	5.01 ± 2.04	4.60 ± 1.98
FS7	2.13 ± 1.67	2.14 ± 1.55	2.19 ± 1.52	1.90 ± 1.35	1.94 ± 1.29	2.02 ± 1.31	5.96 ± 1.79	5.47 ± 1.76	5.03 ± 1.92
FS8	2.12 ± 1.73	2.19 ± 1.66	2.23 ± 1.64	1.85 ± 1.29	1.95 ± 1.31	2.02 ± 1.35	6.61 ± 1.92	6.15 ± 1.88	5.80 ± 1.97

	All fields (N = 388)	Passed fields (N = 366)	Failed fields (N = 22)
Two-way ANOVA			
FS	F(7, 2709) = 1.925, p = 0.062	F(7, 2555) = 0.481, p = 0.849	F(7, 147) = 5.378, p < 0.001
FM	F(2, 774) = 2.209, p = 0.111	F(2, 730) = 9.907, p < 0.001	F(2, 42) = 14.149, p < 0.001
Interaction	F(14, 5418) = 0.110, p = 1.000	F(14, 5110) = 0.123, p = 1.000	F(14, 294) = 0.102, p = 1.000

FS = feature set; FM = fine-tuning method; MSE = mean squared error; SMSE = MSE loss with re-sampling; WMSE = weighted MSE

trend was observed in predicting GPR for passed fields ($p < 0.05$).

Classification accuracy

The AUC, sensitivity and specificity for different models in the test dataset are summarized in Table 3. The corresponding ROC curves are shown in Fig. 5. The WMSE_FS5 model achieved a superior AUC (0.92) and more balanced sensitivity (0.77) and specificity (0.89) compared to the other models.

Directional impact of input features on model output

Figure 6 shows that F2_Gap had a positive effect on the GPR prediction of the two models (WMSE_FS1 and WMSE_FS5), whereas F3_SpeedA and F4_SpeedB had a negative effect on the model output. The effect of the other input features on the model output was less clear.

Discussion

In this study, a set of input feature maps characterizing the plan delivery process were constructed by extracting meaningful data from log files to complement the most commonly used fluence map in an attempt to improve the performance of GPR prediction. Previous studies applying traditional ML models to virtual patient-specific QA have often required the construction of tens or hundreds of hand-crafted features, which rely heavily on the expertise and experience of domain experts and may not necessarily provide the most meaningful representation of the raw data. In contrast, our defined DL feature maps, which represent the spatial variation of machine parameters, are more intuitive and easier to compute. Although DL models with sophisticated structures can learn high-quality representations on their own, a major concern is how to explain the relationship between input features and model output [15]. The SHAP framework proposed by Lundberg has proven useful in interpreting the prediction results of DL models [28]. Therefore, it was used in this study to interpret the DL models. The feature importance bar plots show that input features related to machine delivery errors, such as fluence error, MLC speed error, and leaf gap error, were excluded during the early stage of feature selection. This suggests that delivery system errors may have a lesser impact on the overall accuracy of IMRT plan delivery in our department. The models trained with input feature maps including fluence, leaf gap, and MLC speed demonstrated superiority over other models. They showed comparable performance in predicting GPR for passed fields, but significantly lower MAE for failed fields compared to the other models. As a result, the smaller discrepancy in GPR prediction accuracy between passed and failed fields led to higher accuracy in classifying IMRT fields as passing or failing patient-specific QA. In addition to feature selection, we also used SHAP to interpret the directional impact of input features on model prediction. The results show that a larger leaf gap width increased the predicted GPR, while a higher leaf speed of both banks had a negative effect on the prediction. These findings are consistent with the observations in our previous study [23]. The

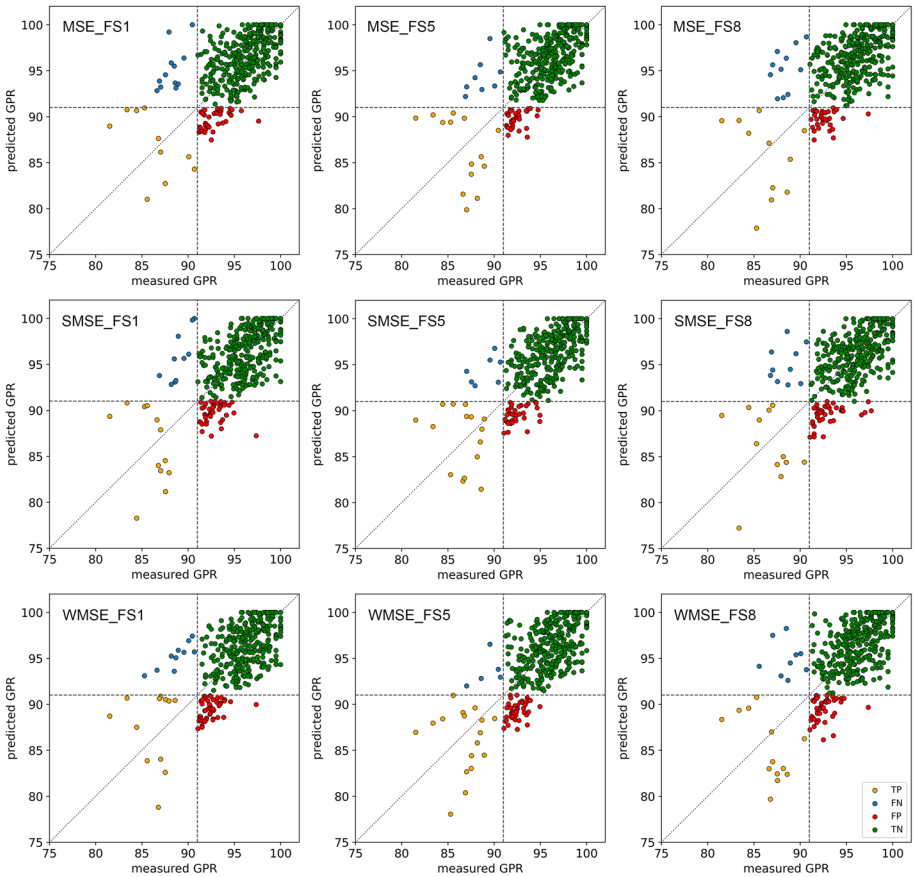

Fig. 4 Scatter plot of measured versus predicted gamma passing rate (GPR) for models trained with the full feature set (FS1), the optimal FS (FS5), and the FS with only the *Fluence* map (FS8) using each of the three fine-tuning methods. The diagonal dotted line represents the perfect prediction by an ideal model. The vertical and horizontal dashed lines represent the action limit. TP = true positive; TN = true negative; FP = false positive; FN = false negative

Table 3 Comparison of the classification performance for different models in the test dataset

	AUC			SEN			SPE		
	MSE	SMSE	WMSE	MSE	SMSE	WMSE	MSE	SMSE	WMSE
FS1	0.78	0.77	0.79	0.45 (10/22)	0.55 (12/22)	0.55 (12/22)	0.90 (330/366)	0.89 (325/366)	0.87 (320/366)
FS2	0.78	0.79	0.83	0.45 (10/22)	0.55 (12/22)	0.59 (13/22)	0.90 (329/366)	0.89 (327/366)	0.88 (321/366)
FS3	0.80	0.82	0.86	0.50 (11/22)	0.59 (13/22)	0.64 (14/22)	0.90 (330/366)	0.90 (329/366)	0.89 (325/366)
FS4	0.83	0.86	0.87	0.59 (13/22)	0.64 (14/22)	0.68 (15/22)	0.91 (332/366)	0.89 (327/366)	0.89 (324/366)
FS5	0.88	0.88	0.92	0.64 (14/22)	0.68 (15/22)	0.77 (17/22)	0.91 (334/366)	0.90 (329/366)	0.89 (326/366)
FS6	0.86	0.86	0.91	0.64 (14/22)	0.68 (15/22)	0.73 (16/22)	0.90 (331/366)	0.90 (329/366)	0.90 (329/366)
FS7	0.85	0.82	0.87	0.59 (13/22)	0.59 (13/22)	0.68 (15/22)	0.90 (328/366)	0.89 (326/366)	0.89 (324/366)
FS8	0.81	0.81	0.84	0.50 (11/22)	0.55 (12/22)	0.59 (13/22)	0.91 (332/366)	0.89 (325/366)	0.88 (323/366)

FS = feature set; AUC = the area under the receiver operating characteristic curve; SEN = sensitivity; SPE = specificity; MSE = mean squared error; SMSE = MSE loss with re-sampling; WMSE = weighted MSE

study revealed that the gap width increased as the plan complexity decreased and had a positive correlation with the measured GPR. Additionally, the leaf speed was found to be negatively correlated with the measured GPR. For low complexity IMRT fields, the leaf position errors increase due to higher leaf speed. However, the increased gap width reduces dose calculation uncertainties, which plays a dominant role in plan delivery accuracy. Several

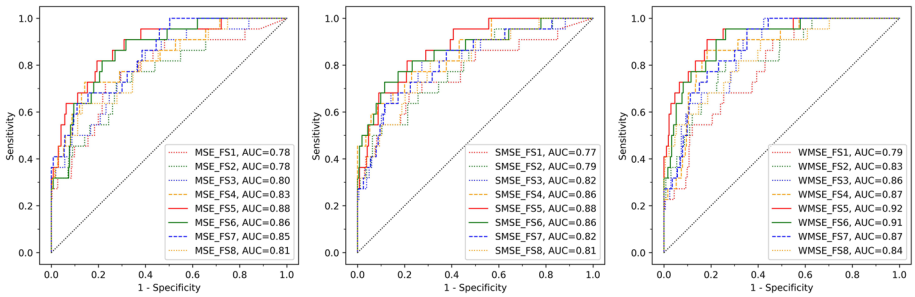

Fig. 5 Comparison of the area under the receiver operating characteristic curve (AUC) for different models in the test dataset

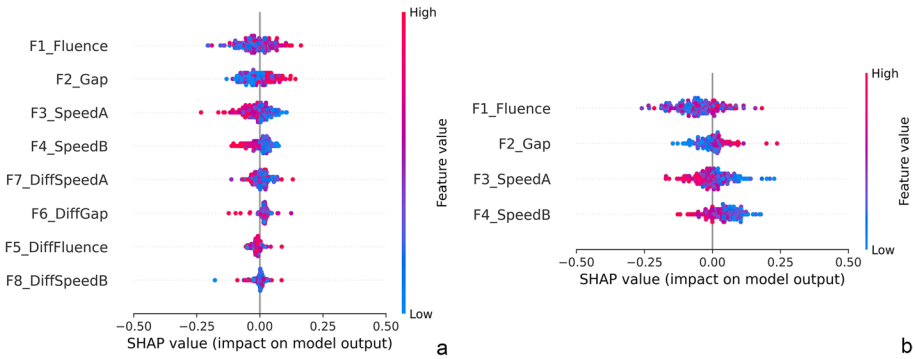

Fig. 6 The distribution of SHAP values computed on the test dataset for each feature in the WMSE_FS1 (**a**) and WMSE_FS5 (**b**) models

studies have reported that aperture-based complexity metrics can explain the variance in QA results. However, there are conflicting conclusions regarding the importance of leaf speed features. Valdes et al. [6] modeled the effect of MLC speed by averaging the leaf speed over all control points and beams for a given plan and concluded that MLC speed did not improve their model's prediction accuracy. In contrast, Braun et al. [35] found that the MLC movement variability score was the best performing single metric for their model. This score is defined as the number of leaf movements that exceed the standard deviation of leaf speeds. These findings suggest that the definition of input features is particularly important for traditional ML applications. The inappropriate methods used to aggregate raw data in ML can inadvertently result in the loss of discriminative information. In contrast, DL coupled with the SHAP framework represents a promising direction for overcoming this challenge due to its ability to automatically learn high-quality feature representations while maintaining model interpretability.

Machine models often predict GPR less accurately for failed fields than for passed fields due to the imbalanced distribution of examples in the training dataset. Previous studies have reported that the prediction accuracy of DL models decreases as the measured GPR decreases. Huang et al. [14] found that fields with measured GPR less than 80% had an average prediction error of up to 15%, compared to less than 3% accuracy for fields with measured GPR greater than 90%. Fondevila et al. [19] also noted that CNN models performed worse in the range of measured GPR below 95%. A variety of methods have been proposed in the literature to alleviate the problem of class imbalance in other ML application domains, such as loss re-weighting, re-sampling, data synthesis, and hybrid techniques [36–39]. However, there are few studies that compare these methods for predicting patient-specific QA results. Fondevila et al. [19] reported that they performed synthetic over-sampling of fields with measured GPR less than 95% through rotational and translational transformations, and under-sampling of fields with measured GPR greater than 95%. However, this approach was found to produce unfavorable results. Similar results were also observed when these basic data augmentation techniques were applied to our task, including image translation, rotation, and flipping. This may be attributed to the fact that the generated samples were less likely to be encountered in the real data distribution. Therefore, these techniques were not employed to augment the minority class samples in this study. In order to provide meaningful samples for model training, further research is warranted to identify a task-specific data augmentation technique. This may be accomplished by further

exploring more sophisticated data augmentation techniques, such as automated data augmentation, generative model-based augmentation, and feature space augmentation [40]. To address the class imbalance problem, we have developed a novel weighted MSE loss. First, we balanced the impact of passed and failed fields by weighting the loss value contributed by them with a factor inversely proportional to the measured GPR of each field in the dataset. Second, we introduced another factor to penalize the misclassification of the failed fields as passed ones, or vice versa, according to the magnitude of the deviation of the predicted GPR from the AL. The results show that the proposed re-weighting method can achieve a more balanced prediction and classification accuracy between passed and failed fields compared to the re-sampling method. Although the proposed re-weighting method has achieved promising results, a thorough investigation of the optimal combination of WMSE with various advanced class balancing and data augmentation techniques reported in the literature is required, as the joint use of certain techniques may degrade the overall performance [40].

Some limitations of this study should be mentioned. Although our results suggest that machine delivery errors play a less significant role in predicting the measured GPR, it should be noted that a comprehensive machine-specific QA program is routinely performed to ensure that the linac operates within tolerances. Other sources of error, including daily beam variation (beam profiles, percent depth doses, dose rate), MLC miscalibration, and poor beam modeling, may still have a potential impact on patient-specific QA results. The effect of the discrepancy between measured and calculated dose distributions caused by these error sources on the prediction and classification of QA results needs to be further investigated in the future. Therefore, the proposed approach is unlikely to completely replace measurement-based methods for patient-specific QA, but rather to enhance the existing QA program at different stages of IMRT planning and delivery. Given that mechanical parameter errors exert a relatively minor influence on the prediction of the measured GPR for a well-calibrated linac, it is possible to construct the most relevant input features, such as planned fluence, MLC leaf speed, and leaf gap, using the calculated control point sequence data instead of log files prior to irradiation. In addition to being a pre-treatment QA verification tool, it can also be integrated into the plan optimization phase to reduce unnecessary plan complexity, thereby reducing the proportion of plans that fail the measurement-based QA test [7]. It can also be used to monitor plan deliverability at each fraction during the course of treatment for each patient, if a large dataset is curated that covers all relevant sources of error in IMRT planning and delivery. Therefore, a public database of clinical patient-specific QA data collected from multiple institutions, treatment machines, and techniques, and freely available for research, will facilitate cross-comparison of the performance of different approaches to GPR prediction and benefit the deployment of prediction models in clinical settings. Additionally, it is important to note that our study was conducted using a dataset generated from a single IMRT delivery technique. However, the proposed approach and concept have the potential to be adapted to predict VMAT QA outcomes, provided that the variable gantry speed and dose rate are appropriately constructed as input feature maps. To achieve higher prediction accuracy, it is necessary to build separate prediction models for each combination of treatment sites, delivery machines, measurement devices, and delivery techniques. Further studies are warranted to determine the generalizability of these findings to other institutions using different combinations of planning and delivery techniques.

Conclusions

Machine parameters can provide discriminative input features for GPR prediction in DL. The novel weighted loss function demonstrates the ability to balance the prediction and classification accuracy between the passed and failed fields. The proposed approach is able to improve the DL model performance in predicting and classifying GPR, and can potentially be integrated into the plan optimization process to generate higher deliverability plans.

Abbreviations

DL	Deep learning
GPR	Gamma passing rate
FS	Feature set
MSE	Mean squared error
WMSE	Weighted MSE loss
AUC	Area under the receiver operating characteristic curve
IMRT	Intensity modulated radiation therapy
QA	Quality assurance
ML	Machine learning
MAE	Mean absolute error
MLC	Multileaf collimator
AL	Action limit
TG	Task Group
Dynalog	Dynamic log
SE	Squeeze-and-Excitation
GAP	Global average pooling
SHAP	SHapley Additive exPlanations
SMSE	MSE loss with re-sampling
ANOVA	Analysis of variance
FM	Fine-tuning method
SEN	Sensitivity
SPE	Specificity

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13014-024-02496-5.

Additional file1

Acknowledgements
Not applicable.

Author contributions
WS (Wei Song) contributed to the design of the study, data collection, data analysis, and manuscript drafting. WS (Wen Shang) and CL were responsible for data collection, data analysis, and interpretation of findings. XB and HL provided clinical expertise and helped to draft the manuscript. JM and DY participated in the design of the study and have critically revised the manuscript. All authors read and approved the final manuscript.

Funding
Financial support for this work was provided by the National Natural Science Foundation of China (81703758) and Innovation development fund of Jiangsu Province Hospital of Chinese Medicine (Y2019CX26).

Availability of data and materials
The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
This study was approved by the Ethics Committee of Affiliated Hospital of Nanjing University of Chinese Medicine (Jiangsu Province Hospital of Chinese Medicine), and written informed consent was obtained from all the patients.

Consent for publication
The authors give their permission for the submission.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Radiation Oncology, Jiangsu Province Hospital of Chinese Medicine, Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing 210029, China.

Received: 25 January 2024 Accepted: 24 July 2024
Published online: 31 July 2024

References

1. Chen D, Cai SB, Soon YY, Cheo T, Vellayappan B, Tan CW, Ho F. Dosimetric comparison between intensity modulated radiation therapy (IMRT) vs dual arc volumetric arc therapy (VMAT) for nasopharyngeal cancer (NPC): systematic review and meta-analysis. J Med Imaging Radiat Sci. 2023;54(1):167–77.
2. Miften M, Olch A, Mihailidis D, Moran J, Pawlicki T, Molineu A, et al. Tolerance limits and methodologies for IMRT measurement-based verification QA: recommendations of AAPM task group No 218. Med Phys. 2018;45(4):e53–83.
3. Chan LT, Tan YI, Tan PW, Leong YF, Khor JS, Teh MW, et al. Comparing log file to measurement-based patient-specific quality assurance. Phys Eng Sci Med. 2023;46(1):303–11.
4. Stasko JT, Ferris WS, Adam DP, Culberson WS, Frigo SP. IMRT QA result prediction via MLC transmission decomposition. J Appl Clin Med Phys. 2023;24(8):1–10.
5. Zhu TC, Stathakis S, Clark JR, Feng W, Georg D, Holmes SM, et al. Report of AAPM task group 219 on independent calculation-based dose/MU verification for IMRT. Med Phys. 2021;48(10):e808–29.
6. Valdes G, Scheuermann R, Hung CY, Olszanski A, Bellerive M, Solberg TD. A mathematical framework for virtual IMRT QA using machine learning. Med Phys. 2016;43(7):4323–34.
7. Wall PDH, Fontenot JD. Quality assurance-based optimization (QAO): Towards improving patient-specific quality assurance in volumetric modulated arc therapy plans using machine learning. Physica Med. 2021;87:136–43.
8. Hirashima H, Ono T, Nakamura M, Miyabe Y, Mukumoto N, Iramina H, Mizowaki T. Improvement of prediction and classification performance for gamma passing rate by using plan complexity and dosiomics features. Radiother Oncol. 2020;153:250–7.
9. Lam D, Zhang X, Li H, Deshan Y, Schott B, Zhao T, et al. Predicting gamma passing rates for portal dosimetry-based IMRT QA using machine learning. Med Phys. 2019;46(10):4666–75.
10. Viola P, Romano C, Craus M, Macchia G, Buwenge M, Indovina L, et al. Prediction of VMAT delivery accuracy using plan modulation complexity score and log-files analysis. Biomed Phys Eng Express. 2022;8(5):1–11.
11. Noblet C, Duthy M, Coste F, Saliou M, Samain B, Drouet F, et al. Implementation of volumetric-modulated arc therapy for locally advanced breast cancer patients: Dosimetric comparison with deliverability consideration of planning techniques and predictions of patient-specific QA results via supervised machine learning. Physica Med. 2022;96:18–31.
12. Cilla S, Viola P, Romano C, Craus M, Buwenge M, Macchia G, et al. Prediction and classification of VMAT dosimetric accuracy using plan complexity and log-files analysis. Physica Med. 2022;103:76–88.
13. Ono T, Hirashima H, Iramina H, Mukumoto N, Miyabe Y, Nakamura M, Mizowaki T. Prediction of dosimetric accuracy for VMAT plans using plan complexity parameters via machine learning. Med Phys. 2019;46(9):3823–32.
14. Huang Y, Pi Y, Ma K, Miao X, Fu S, Zhu Z, et al. Deep learning for patient-specific quality assurance: predicting gamma passing rates for IMRT based on delivery fluence informed by log files. Technol in Cancer Res Treat. 2022;21(1):1–9.
15. Osman AFI, Maalej NM. Applications of machine and deep learning to patient-specific IMRT/VMAT quality assurance. J Appl Clin Med Phys. 2021;22(9):20–36.
16. Tomori S, Kadoya N, Takayama Y, Kajikawa T, Shima K, Narazaki K, Jingu K. A deep learning-based prediction model for gamma evaluation in patient-specific quality assurance. Med Phys. 2018;45(9):4055–65.
17. Raeisi K, Khazaei M, Tamburro G, Croce P, Comani S, Zappasodi F. A class-imbalance aware and explainable spatio-temporal graph attention network for neonatal seizure detection. Int J of Neural Syst. 2023;33(9):1–16.
18. Wei M, Zhou Y, Li Z, Xu X. Class-imbalanced complementary-label learning via weighted loss. Neural Netw. 2023;166:555–65.
19. Fondevila DM, Rios PJ, Peñalva DMD, Arbiser S. Predicting gamma passing rates for portal dosimetry-based IMRT QA using deep learning. Int J Radiat Oncol Biol Phys. 2021;111(3S):e110–1.
20. Heilemann G, Poppe B, Laub W. On the sensitivity of common gamma-index evaluation methods to MLC misalignments in Rapidarc quality assurance. Med Phys. 2013;40(3):0317021–112.
21. Nelms BE, Chan MF, Jarry G, Lemire M, Lowden J, Hampton C, et al. Evaluating IMRT and VMAT dose accuracy: practical examples of failure to detect systematic errors when applying a commonly used metric and action levels. Med Phys. 2013;40(11):1117221–315.
22. Hughes M. Dynalog File Analyser. 2023.
23. Song W, Ma J, Lu H, Zhao D, Yu D. Determination of the optimal fluence smoothing parameters of IMRT plans for nasopharyngeal carcinoma based on log files. Chin J Cancer Prev Treat. 2022;29(2):147–52.
24. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. UNet++: redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans Med Imaging. 2020;39(6):1856–67.
25. Yin Y, Xu W, Chen L, Wu H. CoT-UNet++: A medical image segmentation method based on contextual transformer and dense connection. Math Biosci and Eng. 2023;20(5):8320–36.
26. Lin J, She Q, Chen Y. Pulmonary nodule detection based on IR-UNet + +. Med Biol Eng Comput. 2023;61(2):485–95.
27. Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-Excitation networks. IEEE Trans Pattern Anal Mach Intell. 2020;42(8):2011–23.
28. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng. 2018;2(10):749–60.

29. Tarabanis C, Kalampokis E, Khalil M, Alviar CL, Chinitz LA, Jankelson L. Explainable SHAP-XGBoost models for in-hospital mortality after myocardial infarction. Cardiovas Digital Health J. 2023;4(4):126–32.
30. Yi F, Yang H, Chen D, Qin Y, Han H, Cui J, et al. XGBoost-SHAP-based interpretable diagnostic framework for alzheimer's disease. BMC Med Inf Decis Making. 2023;23(1):137–50.
31. Almeida RL, Maltarollo VG, Coelho FGF. Overcoming class imbalance in drug discovery problems: Graph neural networks and balancing approaches. J Mol Graphics Modell. 2024;126:108627–34.
32. Srv S, Sivapuram AK, Ravi V, Senthil G, Gorthi RK. VISAL-A novel learning strategy to address class imbalance. Neural Netw. 2023;161:178–84.
33. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell. 2020;42(2):318–27.
34. Wobbrock JO. ARTool. 2023.
35. Braun J, Quirk S, Tchistiakova E. Machine learning-generated decision boundaries for prediction and exploration of patient-specific quality assurance failures in stereotactic radiosurgery plans. Med Phys. 2022;49(3):1955–63.
36. Walsh R, Tardy M. A comparison of techniques for class imbalance in deep learning classification of breast cancer. Diagnostics. 2023;13(1):67–85.
37. Rezvani S, Wang X. A broad review on class imbalance learning techniques. Appl Soft Comput. 2023;143:110415–43.
38. Saini M, Susan S. Tackling class imbalance in computer vision: a contemporary review. Artif Intell Rev. 2023;56(1):1279–335.
39. de Oliveira WDG, Berton L. A systematic review for class-imbalance in semi-supervised learning. Artif Intell Rev. 2023;56(2):2349–82.
40. Mumuni A, Mumuni F. Data augmentation: A comprehensive survey of modern approaches. Array. 2022;16:1–22.

Publisher's Note