Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Delta-Dosiomics-Based Multilevel Precision Identification of Radiotherapy Treatment Errors in Cervical Cancer

Junyou Shen
First Affiliated Hospital of Chongqing Medical University
Xin Yi
Xi'an Jiaotong University
Wenli Lu
First Affiliated Hospital of Chongqing Medical University
Hanyin Zhang
First Affiliated Hospital of Chongqing Medical University
Haoling Chen
First Affiliated Hospital of Chongqing Medical University
Wenxin Chen
First Affiliated Hospital of Chongqing Medical University
Lifang Tong
First Affiliated Hospital of Chongqing Medical University
Yanbo Song
songyanbo@cqmu.edu.cn
First Affiliated Hospital of Chongqing Medical University

Research Article

Keywords: Cervical cancer radiotherapy, Delta-dosiomics, Treatment error, Machine learning, 3D dose-difference maps

Posted Date: June 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9822292/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background:

Conventional gamma analysis for dose verification has inherent limitations. It cannot accurately identify the types of treatment error or quantify their magnitudes. This study aimed to develop a multilevel error classification model based on delta-dosiomics features extracted from three-dimensional (3D) dose-difference maps for the classification treatment errors in cervical cancer radiotherapy, thereby improving treatment safety and precision.

Materials and methods:

A total of 52 cervical cancer patients who underwent volumetric modulated arc therapy (VMAT) were retrospectively included. Based on the original treatment plans, eight types of treatment errors with different magnitudes were introduced, yielding 1,984 delta-dosiomics datasets. Delta-dosiomics features were extracted from four evaluation volumes: the planning target volume (PTV), the external body contour (BODY), and the 10% and 50% prescription isodose volumes. Feature selection was performed using Spearman correlation analysis, random forest feature importance ranking and recursive feature elimination (RFE). Four machine learning algorithms, support vector machine (SVM), random forest (RF), artificial neural network (ANN), and extreme gradient boosting (XGBoost), were used to develop classification models. Model training was performed using 5-fold cross-validation across three levels: Level 1 (main error type), Level 2 (error direction), and Level 3 (error magnitude). Model performance was evaluated on the independent test set and clinical validation set.

Results:

First-order features were the dominant predictors, followed by texture features derived from the gray level co-occurrence matrix (GLCM) and gray level run length matrix (GLRLM). The XGBoost model, based on delta-dosiomics features extracted from the 50% prescription isodose volume, showed the most robust performance across all classification levels. On the independent test set, the model achieved accuracies of 94.19% (AUC = 0.998) for Level 1, 83.87% (AUC = 0.993) for Level 2, and 74.84% (AUC = 0.985) for Level 3. On the clinical validation set, the corresponding accuracies were 72.81%, 64.75%, and 51.61% for Levels 1, 2, and 3, respectively, with AUC values of 0.945, 0.935, and 0.911.

Conclusion:

3D dose-difference maps in dose verification may serve as a unique “fingerprint” of radiotherapy treatment errors. Combined with machine learning, delta-dosiomics may provide an effective complement to conventional gamma analysis, helping medical physicists rapidly localize and quantify treatment errors and potentially improving the safety and quality of radiotherapy.

1 Background

Cervical cancer remains one of the most common malignancies in women worldwide and ranks as the fourth most frequently diagnosed cancer among females [1]. It continues to pose a major public health burden, particularly in developing countries. Radiotherapy, especially intensity-modulated radiation therapy (IMRT) and volumetric modulated arc therapy (VMAT), plays a crucial role in the management of locally advanced cervical cancer [2]. Owing to their advanced beam modulation capabilities, IMRT and VMAT enable more precise dose sculpting, achieving superior target conformity while minimizing radiation exposure to surrounding normal tissues [3, 4]. However, the increased complexity of IMRT and VMAT planning and delivery imposes more stringent requirements on patient-specific quality assurance (PSQA) [5, 6].

Traditionally, PSQA has relied on gamma analysis, first proposed by Low et al. [7] in 1998. This method summarizes the agreement between calculated and measured dose distributions using gamma passing rates (GPRs), thereby providing a quantitative indicator of treatment planning calculation accuracy and linear accelerator delivery precision. Although gamma analysis is simple and widely adopted, numerous studies have highlighted its inherent limitations [8–14]. Firstly, gamma analysis is criticized for its weak correlation with clinically relevant metrics [9–11], such as estimated deviations in dose-volume histograms. Secondly, extensive

research [15–18] has demonstrated that gamma analysis is less sensitive to radiotherapy treatment errors. Even though gamma analysis can detect treatment errors, it cannot provide information on their source or magnitude. Thirdly, the evaluation volume recommended in the AAPM TG 218 report (10% prescription isodose volume) is also questioned for being too lenient [12, 13].

A major reason for these limitations is that gamma analysis condenses dose-comparison information into a single value, the GPRs, thereby losing potentially important spatial and textural information. To overcome this loss of information, recent studies have increasingly applied artificial intelligence (AI) to directly extract quantitative features from different types of dose-comparison information, including gamma maps, dose-gradient maps, structural similarity index maps (SSIMs), and dose-difference maps. Inspired by radiomics, Wootton et al. [19] and Lizar et al. [3] reported that combined with machine learning and deep learning, radiomic features extracted from gamma maps can improve treatment delivery error detection. Ma et al. [20] further proposed that radiomic analysis on SSIMs has great potential for error identification in PSQA. Furthermore, several studies have applied machine learning and deep learning methods to directly extract features from dose-difference maps for treatment delivery error classification or GPRs prediction, achieving good model performance. Wolfs et al. [21] systematically compared different types of dose-comparison information for deep learning-based error identification and suggested that simple dose-comparison information, such as dose-difference maps, may be beneficial for deep learning models. Because gamma maps, SSIMs, and dose-gradient maps are all ultimately derived from dose-difference maps, the latter contains too many original details for human interpreters. Therefore, direct extraction of high-dimensional features from dose-difference maps may be a more effective strategy for identifying treatment error.

However, current studies still have several important limitations. On the one hand, owing to computational complexity, most investigations have been limited to two-dimensional (2D) dose-difference maps. On the other hand, the optimal region for feature extraction remains unclear. Feature extraction from the entire region may introduce image noise and limit model performance. Therefore, incorporating three-dimensional (3D) dose-comparison information and exploring the optimal feature-extraction region may be essential for developing error classification models with improved performance and greater clinical interpretability [22].

In this study, we performed radiomics analysis of 3D dose-difference maps, termed *delta-dosiomics*. Delta-dosiomics features were extracted from multiple evaluation volumes, including the planning target volume (PTV), the external body contour (BODY), the 10% prescription isodose volume ($V_{10\%}$), and the 50% prescription isodose volume ($V_{50\%}$). Based on these delta-dosiomics features, several machine learning models, including support vector machine (SVM), random forest (RF), artificial neural network (ANN), and extreme gradient boosting (XGBoost), were developed and compared for the identification of radiotherapy treatment errors in cervical cancer VMAT plans. The aims of this study were: (1) to develop a delta-dosiomics-based approach for detecting treatment error types and quantifying error magnitudes; and (2) to evaluate the performance of machine learning models based on features from different evaluation volumes in order to explore the optimal evaluation volume. Ultimately, this study aimed to provide a novel, quantitative, and clinically meaningful framework for detecting radiotherapy treatment errors in dose-guided radiotherapy.

2 Methods

2.1 Workflow

Figure 1 summarizes the overall workflow of this study. Firstly, VMAT treatment plans from 52 cervical cancer patients were collected, and treatment errors with different types and magnitudes were simulated. Subsequently, 3D dose-difference maps were generated by comparing the error-induced plans with the original plans, and delta-dosiomics features were extracted using 3D Slicer (version 5.6.2). Next, feature selection was performed in three consecutive steps: correlation analysis, feature importance ranking, and recursive feature elimination (RFE). Finally, the selected features were used to develop machine learning models, which were further evaluated on an independent test set and a clinical validation set to assess their performance for treatment error detection.

2.2 Data Collection

A retrospective analysis was conducted on 52 VMAT plans for cervical cancer. The prescription dose to the planning target volume was either 50.4 Gy in 28 fractions (1.8 Gy per fraction) or 50 Gy in 25 fractions (2.0 Gy per fraction). These plans with dual-arc were calculated on a 2.5 mm isotropic dose grid using the Acuros XB (AXB) algorithm in Eclipse version 15.6 and delivered by the Varian VitalBeam machine (Varian Medical Systems, Palo Alto, CA).

PSQA measurements were performed using a gantry-mounted Dolphin detector (IBA Dosimetry, Germany). The Dolphin detector consists of 1,514 sensors (1,513 air-vented plate ionization chambers and 1 diode) arranged in a grid. The center-to-center distances of the sensors range from 5 mm to 10 mm. The COMPASS system (version 4.1, IBA Dosimetry, Germany) reconstructs the 3D dose distribution in the patients' body based on the fluence measured by the Dolphin detector. It performs 3D dose reconstruction on the same dose grid as the treatment planning system (TPS) using the collapsed cone algorithm and provides a 3D comparison between the planned and reconstructed dose distributions.

All plans included in this study were required to meet the following two criteria: (1) GPRs (3%/2 mm) ≥ 95% in the 10% prescription isodose volume [5], and (2) a mean gamma index ≤ 0.6 in the 50% prescription isodose volume [23]. During the data acquisition period, all treatment equipment, including the linear accelerator and dose verification system, underwent regular performance checks and dose calibrations to ensure system stability and data reliability.

This study was approved by the Research Ethics Committee of the First Affiliated Hospital of Chongqing Medical University (No. K2024-224-01). The requirement for patient-informed consent was waived due to the study's retrospective nature and the collection of existing, de-identified medical data.

2.3 Treatment Error Simulation

Machine-, plan-, and patient-related factors are the three principal sources of radiotherapy treatment errors. Accordingly, eight categories of treatment errors with different types and magnitudes were deliberately introduced into the original treatment plans, including setup errors, random and systematic multileaf collimator (MLC) errors, monitor unit (MU) errors, gantry position errors, bladder filling variations, patient weight loss, and differences in dose calculation algorithms. A multilevel error simulation framework was established to systematically model these factors: (1) Machine delivery errors were simulated by modifying DICOM RT files using an in-house Python program based on Pydicom (version 2.3.1). (2) In anatomical changes, the CT information was artificially modified to simulate variations in bladder filling and patient weight loss. (3) Setup errors were introduced by changing the field isocenter coordinates of the original plans in the lateral, longitudinal, and vertical directions.

As listed in Table 1, multiple treatment error types and magnitudes were introduced into the 52 patient treatment plans. By simulating errors based on both original treatment plans and clinically delivered dose distributions, a total of 1,984 error-induced plans were generated. Subsequently, 3D dose-difference maps were generated by voxel-wise subtraction of the original dose distribution from each corresponding error-induced dose distribution.

Table 1
Types and magnitudes of introduced radiotherapy treatment errors.

Error source	Main error type (Level 1)	Error direction (Level 2)	Error magnitude (Level 3)
Plan-related	Dose calculation algorithm differences	Dose calculation algorithm differences	AAA-AXBW, AAA-AXBM, AXBW-AXBM
	MU errors	MU increase	+ 3%, + 5%
		MU decrease	-3%, -5%
Machine-related	MLC random errors	MLC random error	σ = 0.5 mm, 1 mm, 2 mm
	MLC systematic errors	MLC outward	+ 0.5 mm, + 1 mm
		MLC inward	-0.5 mm, -1 mm
	Gantry position errors	Clockwise	+ 0.2°, + 0.4°
		Counterclockwise	-0.2°, -0.4°
Patient-related	Setup errors	Lateral	± 1 mm
		Longitudinal	± 1 mm
		Vertical	± 1 mm
	Bladder filling variation	Increased filling	+ 0.5 mm, + 1 mm
		Decreased filling	-0.5 mm, -1 mm
	Weight loss	Weight loss	-0.5 mm, -1 mm, -1.5 mm

For setup errors, "±" indicates the direction of the error, where "+" indicates right/anterior/superior shifts, and "–" indicates left/posterior/inferior shifts. AAA: Analytical Anisotropic Algorithm; AXBW: Acuros XB dose-to-water algorithm; AXBM: Acuros XB dose-to-medium algorithm; MLC: multileaf collimator; MU: monitor unit; σ: standard deviation of the Gaussian distribution.

2.4 Feature Extraction

The 3D dose-difference maps were resampled to 1-mm isotropic voxels using B-spline interpolation before feature extraction. Delta-dosiomics features were extracted using 3D Slicer (version 5.6.2) from each dose-difference map within four evaluation volumes: PTV, BODY, $V_{10\%}$, and $V_{50\%}$. For each evaluation volume, 107 delta-dosiomics features were obtained, comprising 14 shape features, 18 first-order features, 24 gray level co-occurrence matrix (GLCM) features, 16 gray level size zone matrix (GLSZM) features, 16 gray level run length matrix (GLRLM) features, 5 neighboring gray tone difference matrix (NGTDM) features, and 14 gray level dependence matrix (GLDM) features.

2.5 Feature Selection

Delta-dosiomics feature selection was performed separately for each classification level (error type, error direction, and error magnitude) and for each of the four evaluation volumes. This procedure yielded 12 feature sets, corresponding to the four evaluation volumes across the three classification levels. Feature selection was conducted in three steps: (1) Highly correlated features were removed based on Spearman's correlation analysis ($|\rho| > 0.8$). (2) An RF model was used to preliminarily assess and rank feature importance. (3) Using RF as the base learner, RFE was applied to iteratively remove the least important feature. The final feature subset was defined as the subset that achieved the highest classification accuracy.

2.6 Model Training and Evaluation

The development dataset (n = 1,550) was randomly divided into a training set (n = 1,240) and an independent test set (n = 310) at an 8:2 ratio. For each classification level, four machine learning models were trained using SVM, RF, ANN, and XGBoost based on selected features from four evaluation volumes. In total, 12 machine learning models were developed to detect treatment errors

across three classification levels. Hyperparameter optimization was performed on the training set using 5-fold cross-validation. Model performance was first assessed using the independent test set. To further evaluate clinical applicability, an additional Dolphin-based measurement dataset (n = 434) was used as the clinical validation set. The receiver operating characteristic (ROC) curve was generated, and performance metrics, including the area under the ROC curve (AUC), accuracy, precision, recall, and F1-score, were calculated to ensure a comprehensive evaluation.

3 Results

3.1 Selection of delta-dosiomics features

An RF model was employed to assess the relative importance of delta-dosiomics features. Figure 2 illustrates the cumulative feature importance scores for the four evaluation volumes, summed across the three classification levels. First-order features made the largest contribution to the model performance across all evaluation volumes. Key features, including 90th percentile, 10th percentile, maximum and energy, consistently ranked among the top predictors. Texture features, particularly those derived from the GLCM and GLRLM, also showed high importance. In contrast, features derived from the GLSZM, NGTDM, and GLDM, as well as all shape features, contributed less to model performance.

3.2 Model performance in different classification levels

To identify the optimal machine learning algorithm and evaluation volume for treatment error identification, model performance was compared across four algorithms and four evaluation volumes at three classification levels.

3.2.1 Performance comparison of machine learning algorithms

Table 2 presents the mean performance metrics of the four machine learning algorithms, averaged across all evaluation volumes at the three classification levels. The results showed distinct performance differences among the algorithms at varying levels of task complexity.

Table 2
Mean test accuracy of four machine-learning algorithms across three classification levels, averaged over all evaluation volumes.

Level	Performance	RF		SVM		ANN		XGBoost	
		Independent	Clinical	Independent	Clinical	Independent	Clinical	Independent	Clinical
Level 1	Accuracy	0.918	**0.689**	**0.931**	0.687	0.925	0.679	0.929	0.680
	Precision	0.922	**0.787**	**0.936**	0.729	0.928	0.731	0.931	0.751
	Recall	0.921	0.687	**0.937**	**0.688**	0.932	0.678	0.932	0.679
	F1-score	0.921	**0.694**	**0.936**	0.686	0.929	0.683	0.931	0.686
Level 2	Accuracy	0.837	**0.616**	0.836	0.608	0.827	**0.616**	**0.847**	0.607
	Precision	0.831	**0.691**	0.831	0.676	0.825	0.672	**0.841**	0.680
	Recall	0.831	**0.602**	0.828	0.593	0.819	0.600	**0.839**	0.590
	F1-score	0.827	**0.613**	0.826	0.602	0.819	0.605	**0.838**	0.603
Level 3	Accuracy	**0.774**	**0.474**	0.741	0.471	0.747	0.456	0.762	0.464
	Precision	**0.791**	**0.583**	0.752	0.552	0.755	0.536	0.779	0.553
	Recall	**0.775**	**0.474**	0.743	0.471	0.751	0.456	0.765	0.464
	F1-score	**0.770**	0.474	0.735	**0.476**	0.740	0.460	0.761	0.470

On the independent test set, the models demonstrated high discriminative ability. For identifying error types (Level 1), SVM and XGBoost achieved the highest mean accuracies of 0.931 and 0.929, respectively. For determining error directions (Level 2), XGBoost achieved the highest accuracy of 0.847, followed by RF, with an accuracy of 0.837. For the most challenging task of quantifying error magnitudes (Level 3), RF achieved the highest accuracy (0.774), followed by XGBoost and ANN (0.762 and 0.747).

Although model performance generally decreased across all algorithms on the clinical validation set, XGBoost and RF showed relatively stable performance. This robustness was further supported by representative ROC curves for Level 1 classification on the clinical validation set (Fig. 3). XGBoost consistently achieved the highest AUC values across all volumes, followed closely by RF.

3.2.2 Performance comparison of evaluation volumes

Table 3
Performance of the XGBoost model across three classification levels for four evaluation volumes.

Evaluation Volume	performance	Level1		Level2		Level3	
		Independent	Clinical	Independent	Clinical	Independent	Clinical
PTV	Accuracy	**0.955**	0.645	**0.894**	0.576	**0.836**	0.447
	AUC	**0.999**	0.915	**0.996**	0.912	**0.992**	0.903
BODY	Accuracy	0.867	0.652	0.835	0.590	0.712	0.412
	AUC	0.993	0.924	0.992	0.926	0.979	0.901
$V_{10\%}$	Accuracy	0.952	0.694	0.819	0.613	0.752	0.479
	AUC	**0.999**	0.941	0.992	0.927	0.985	**0.918**
$V_{50\%}$	Accuracy	0.942	**0.728**	0.839	**0.648**	0.748	**0.516**
	AUC	0.998	**0.945**	0.993	**0.935**	0.985	0.911

Based on the comparison results in Table 2, XGBoost was selected as the optimal model for treatment error detection. Table 3 presents the performance of XGBoost models using delta-dosiomics features extracted from the four evaluation volumes at the three classification levels. On the independent test set, the PTV-based model achieved accuracies of 0.955 (AUC = 0.999) at Level 1, 0.894 (AUC = 0.996) at Level 2, and 0.836 (AUC = 0.992) at Level 3. The $V_{50\%}$-based model also demonstrated highly competitive performance (Level 1: 0.942, Level 2: 0.839, Level 3: 0.748), followed by the $V_{10\%}$-based model. The BODY-based model consistently yielded the lowest classification accuracy. However, external validation on the clinical validation set showed a marked decrease in model performance. While the PTV-based model exhibited a severe performance drop, the $V_{50\%}$-based model consistently demonstrated the most stable predictive performance across all classification levels.

3.3 Confusion matrix analysis and error pattern recognition

To further analyze error-specific performance, Fig. 4 presents the normalized confusion matrices of the optimal $V_{50\%}$-based XGBoost model. For Level 1 classification (Fig. 4a), the model showed high classification performance. The classification accuracy exceeded 0.90 for most treatment error types on the independent test set, with relatively lower accuracy observed for random MLC errors (0.82) and gantry position errors (0.89). However, on the clinical validation set, the classification accuracies for bladder filling variations, systematic MLC errors, random MLC errors, and gantry position errors decreased markedly to 0.57, 0.59, 0.55, and 0.52, respectively. These cases were most frequently misclassified as either setup errors or error-free class.

For Level 2 classification (Fig. 4b), the model achieved class-wise accuracy values greater than 0.9 for 7 of the 14 error subtypes on the independent test set and for 4 of the 14 subtypes on the clinical validation set. Misclassifications were concentrated mainly in subtypes related to bladder filling variations, random and systematic MLC errors, gantry position errors, and setup errors, particularly in the clinical validation set.

For Level 3 classification (Fig. 4c), the simulated treatment errors were stratified into 31 subtypes, enabling a more refined, semi-quantitative assessment of error magnitude. Model performance decreased further on the clinical validation set, with a mean

accuracy of 0.516. Nevertheless, class-wise accuracy exceeded 0.8 for 17 of the 31 error subtypes on the independent test set and for 7 of the 31 error subtypes on the clinical validation set. Misclassifications were more frequent on the clinical validation set, and were mainly associated with bladder filling variations, random and systematic MLC errors, gantry position errors and setup errors.

4 Discussion

In this study, we extracted delta-dosiomics features from 3D dose-difference maps and developed machine learning models to detect potential treatment errors and quantify error magnitudes in PSQA for cervical cancer VMAT. The results showed that delta-dosiomics features extracted from four evaluation volumes, when combined with machine learning models, were capable of identifying types and magnitudes of treatment errors across different classification levels. In contrast to the information compression in gamma analysis, delta-dosiomics features preserve high-dimensional spatial and textural information from dose-difference maps. This may offer a novel method to qualitatively and even semi-quantitatively identify radiotherapy treatment errors.

Because delta-dosiomics analysis involves high-dimensional feature extraction from dose-difference maps, feature selection is required to reduce overfitting and improve model interpretability. In this study, feature importance analysis further clarified the mechanisms underlying delta-dosiomics-based error classification. Across the different evaluation volumes and classification levels, first-order features were the main contributors to model performance, including the 90th percentile, the 10th percentile, energy, interquartile range, maximum, and kurtosis. These features describe the magnitude and distribution of voxel-wise dose differences, suggesting that treatment-related errors are mainly reflected by changes in the statistical distribution of dose-difference maps. This may partly explain the relatively robust performance for plan-related and patient-related errors, which often produce more systematic dose-deviation patterns. These findings were generally consistent with those of Wootton et al. [15] and Lizar et al. [3]. They also found that first-order features, including the 10th percentile, the 90th percentile, and kurtosis, were associated with systematic MLC errors or gamma-based classification.

In addition to first-order features, texture features derived from the GLCM and GLRLM also contributed to model performance, capturing high-dimensional spatial information typically obscured in conventional gamma analysis. These features characterize the spatial arrangement and heterogeneity of dose deviation, which may be particularly relevant for treatment errors that produce non-uniform or localized dose changes, such as MLC-related errors and gantry position errors. Collectively, these spatial and textural features extract high-dimensional information that is typically ignored in conventional gamma analysis and may improve understanding of error mechanisms and guide targeted corrective actions. In contrast, shape features contributed little to model performance because the contours of the evaluation volumes were held constant during error simulation. As shape features mainly describe the geometric properties of ROIs [24], they may be less sensitive to dose variations when the ROI contours remain unchanged. This finding suggests that the models were mainly driven by dose variations rather than by contour-related information.

The selection of machine learning algorithms also contributed to differences in classification performance. In this study, XGBoost and RF generally showed relatively favorable and stable performance, consistent with previous reports that tree-based ensemble models can effectively predict GPRs in IMRT/VMAT plans [25, 26]. This may reflect their ability to capture nonlinear relationships and complex interactions among delta-dosiomics features while reducing overfitting through ensemble learning and regularization. In addition, RF has been frequently used for feature selection and feature importance analysis in radiomics-based PSQA studies [3]. Nevertheless, previous studies have also reported promising performance of SVM- and ANN-based models in MLC-related prediction and error-detection tasks [27, 28]. Therefore, although XGBoost showed better generalization in the present study, the optimal algorithm may depend on the used data sizes, the predictive input features, and the specific prediction target, as highlighted in previous reviews of ML- and DL-based approaches for PSQA [4].

In this study, classification sensitivity varied across different evaluation volumes, consistent with previous gamma-analysis studies [29, 30]. Notably, the $V_{50\%}$-based models showed the most stable and discriminative performance across all classification levels. These results suggest that both insufficient and excessive spatial coverage may reduce the robustness of delta-dosiomics-based error classification.

The relatively inferior performance of the BODY- and $V_{10\%}$-based models may be attributable to the inclusion of extensive low-dose regions. Low-dose evaluation regions may dilute error-related spatial and textural alterations in dose-difference maps [31, 32]. Small dosimetric changes in low-dose regions may be indistinguishable from random noise, potentially obscuring treatment-error-related dose deviations and compromising the robustness of delta-dosiomics features. This effect may be further amplified in the BODY-

based model, as the BODY volume includes broad low-dose regions and skin–air interfaces. Sharp density changes at the skin–air boundary may introduce dose deviations unrelated to actual treatment delivery errors [33], thereby adding non-informative variations to the dose-difference maps. As a high-dose target region, the PTV mainly captures dose deviations within the prescribed treatment volume. However, treatment errors can produce spatially distributed dose deviations that extend beyond the target volume and involve surrounding normal tissues or organs at risk (OARs). Therefore, feature extraction restricted to the PTV may fail to capture dose deviations occurring in surrounding normal tissues or OARs, such as the bladder, rectum, and small bowel, which are important for comprehensive treatment safety assessment. In addition, high-dose target regions may be susceptible to minor dose fluctuations and high-frequency noise [34], which could reduce the stability of extracted delta-dosiomics features.

Although $V_{10\%}$ is commonly used as the evaluation volume in traditional gamma analysis, several publications [29, 35] have suggested that $V_{50\%}$ is a more sensitive volume for PSQA. $V_{50\%}$, also named the hot area, encompasses most regions directly exposed to the treatment beams while excluding a substantial proportion of the peripheral low-dose regions. This may allow $V_{50\%}$ to serve as a balanced evaluation volume, being less strict than the high-dose regions, such as the PTV, while remaining more sensitive to delivery errors and less affected by background noise commonly observed in broader low-dose regions, such as $V_{10\%}$. This observation is consistent with the findings of Ishizaka et al. [36], who similarly demonstrated that high-dose regions may be advantageous for PSQA prediction. From a delta-dosiomics perspective, these findings implied that $V_{50\%}$ may represent a more clinically informative evaluation volume for PSQA.

Model performance was also influenced by the complexity of the classification task. The feasibility of multilevel error classification was demonstrated by Wolfs et al. [37], who used convolutional neural networks (CNNs) to analyze 2D EPID-based gamma maps. Consistent with their findings, our results showed a gradual decrease in model performance from Level 1 to Level 3, indicating that classification became more challenging when the task progressed from identifying main error types to distinguishing error directions and magnitudes. In some classification levels, the models achieved comparable or slightly higher accuracy. This improvement may be partly explained by the use of delta-dosiomics features extracted from 3D dose-difference maps, which preserve richer spatial and textural information for characterizing complex treatment error patterns.

Effective identification of treatment errors relies on capturing their underlying dose deviation patterns. As described by van Elmpt et al. [38], PSQA can detect three major categories of treatment errors: machine-related, plan-related, and patient-related errors. In our study, plan-related and patient-related errors generally showed relatively robust classification performance. These errors often induce more systematic and spatially coherent dose deviations, resulting in clearer and more distinguishable patterns in 3D dose-difference maps. In contrast, machine-related errors were more difficult to classify, particularly random MLC errors and gantry position errors [37, 39]. This observation is consistent with the findings of Kimura et al [39]. Random MLC errors are stochastic in nature and may generate diffuse, spatially unstructured noise in dose-difference maps. It is difficult for machine learning models to extract consistent spatial features. Gantry position errors may produce relatively subtle and symmetric dose deviations, which makes their direction and magnitude more difficult to distinguish. These findings suggest that treatment errors can be more accurately classified when they produce spatially consistent and distinguishable dose deviation patterns.

A limited number of studies have highlighted the value of 3D modeling for identifying radiotherapy treatment errors. van den Berg et al. [40] used ensemble CNNs on 3D dose verification data to detect radiotherapy treatment errors. However, their model used relative dose-difference maps as input, it may not fully reflect absolute dose deviations, which are more closely correlated with tumor control probability (TCP) and normal tissue complication probability (NTCP). Additionally, the black-box nature of deep learning models may limit interpretability. Conversely, our study employed 3D absolute dose-difference maps for delta-dosiomics feature extraction, thereby preserving spatial information on dose deviations. Although dosiomics has been increasingly used for clinical outcome prediction [41–43], its application in treatment error identification remains relatively limited. This approach may provide a more interpretable basis for identifying treatment errors and assessing the potential clinical impact of dose deviations. This consideration is supported by recent meta-analyses showing that deviations from radiotherapy protocols are associated with adverse clinical outcomes [44]. It is also consistent with the perspective proposed by Keall et al. [45], who emphasize that accurate dose accumulation is essential for correlating delivered treatments with clinical outcomes. Furthermore, by applying hierarchical classification to all evaluation volumes, this framework enabled a comparison of volume-specific sensitivity to treatment errors with varying types and magnitudes. Therefore, delta-dosiomics may serve as an interpretable complementary tool for conventional gamma analysis. In cases of suboptimal gamma passing rates, it may help further characterize the underlying error patterns and

support more targeted error investigation. Integrating the efficiency of gamma analysis with the interpretability of delta-dosiomics may improve the clinical utility of PSQA for detecting treatment error.

While this study supports the feasibility of delta-dosiomics for identifying radiotherapy treatment errors, several limitations should be acknowledged. First, the data were collected from a single institution with a relatively limited sample size, which may restrict the model's generalizability across patient cohorts, treatment systems, and equipment configurations. Future work should include multicenter and multidevice cohorts, as well as data from other anatomical sites, to enable external validation and further evaluate model robustness and clinical translatability. Second, although using the actually delivered dose distributions as the basis for error simulation improved the dataset's clinical relevance, it still relies on simulated error scenarios. These simulations may not fully capture the complexity and variability of real clinical errors. Future studies should aim to collect and annotate real treatment error cases to further improve model robustness and clinical applicability. Third, the current models relied exclusively on delta-dosiomics features extracted from 3D dose-difference maps. Integrating multimodal PSQA information, such as treatment planning parameters, machine log files, and radiomics, may further improve model performance and interpretability. Finally, as with all omics-based approaches, delta-dosiomics also faces standardization challenges. Future work should follow the recommendations of the Image Biomarker Standardization Initiative (IBSI) to evaluate feature standardization, reproducibility, stability, and redundancy. This will be essential for improving feature comparability and facilitating robust multi-center application.

5 Conclusion

In conclusion, this study demonstrated the feasibility of using delta-dosiomic features combined with machine learning for multilevel classification of radiotherapy treatment errors. The results highlighted the significant influence of evaluation volume, classification level, and machine learning algorithm on model performance. The analysis identified $V_{50\%}$ as the optimal evaluation volume for feature extraction. With further data optimization, multimodal integration, and algorithm refinement, this approach may support the development of intelligent, automated PSQA workflows and contribute to improved radiotherapy safety and quality.

Abbreviations

2D
two-dimensional
3D
three-dimensional
AI
artificial intelligence
ANN
artificial neural network
AUC
area under the receiver operating characteristic curve
AXB
Acuros XB
BODY
external body contour
CNNs
convolutional neural networks
GLCM
gray level co-occurrence matrix
GLDM
gray level dependence matrix
GLRLM
gray level run length matrix
GLSZM
gray level size zone matrix
GPRs

gamma passing rates
IBSI
Image Biomarker Standardization Initiative
IMRT
intensity-modulated radiation therapy
MLC
multileaf collimator
MU
monitor unit
NGTDM
neighboring gray tone difference matrix
NTCP
normal tissue complication probability
OARs
organs at risk
PSQA
patient-specific quality assurance
PTV
planning target volume
RF
random forest
RFE
recursive feature elimination
ROC
receiver operating characteristic
SSIMs
structural similarity index maps
SVM
support vector machine
TCP
tumor control probability
TPS
treatment planning system
$V_{10\%}$
10% prescription isodose volume
$V_{50\%}$
50% prescription isodose volume
VMAT
volumetric modulated arc therapy
XGBoost
extreme gradient boosting

Declarations

Ethics approval and consent to participate

This study was approved by the Research Ethics Committee of the First Affiliated Hospital of Chongqing Medical University (No. K2024-224-01). This study was conducted in accordance with the Declaration of Helsinki. The requirement for patient-informed consent was waived due to the study's retrospective nature and the collection of existing, de-identified medical data.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by
Chongqing medical scientific research project (Joint project of Chongqing Health Commission and Science and Technology Bureau) (No. 2024MSXM104).

Authors' contributions

J.S., X.Y., Y.S., and L.T. conceived and designed the study. J.S. and X.Y developed the methodology and performed model development. J.S., X.Y., W.L., H.Z., H.C., and W.C. participated in data collection, data curation, validation, and investigation. J.S. prepared the figures and tables and drafted the original manuscript. X.Y., Y.S., and L.T. contributed to data interpretation and critically revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors acknowledge the assistance of the Department of Radiation Oncology, the First Affiliated Hospital of Chongqing Medical University, for support in data collection and treatment data extraction.

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, Jemal A: **Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries**. *CA Cancer J Clin* 2024, **74**(3):229-263.
2. Mayadev JS, Ke G, Mahantshetty U, Pereira MD, Tarnawski R, Toita T: **Global challenges of radiotherapy for the treatment of locally advanced cervical cancer**. *Int J Gynecol Cancer* 2022, **32**(3):436-445.
3. Lizar JC, Yaly CC, Colello Bruno A, Viani GA, Pavoni JF: **Patient-specific IMRT QA verification using machine learning and gamma radiomics**. *Phys Med* 2021, **82**:100-108.
4. Osman AFI, Maalej NM: **Applications of machine and deep learning to patient-specific IMRT/VMAT quality assurance**. *Journal of Applied Clinical Medical Physics* 2021, **22**(9):20-36.
5. Miften M, Olch A, Mihailidis D, Moran J, Pawlicki T, Molineu A, Li H, Wijesooriya K, Shi J, Xia P *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-e83.
6. Kamperis E, Kodona C, Hatziioannou K, Giannouzakos V: **Complexity in Radiation Therapy: It's Complicated**. *Int J Radiat Oncol Biol Phys* 2020, **106**(1):182-184.
7. Low DA, Harms WB, Mutic S, Purdy JA: **A technique for the quantitative evaluation of dose distributions**. *Med Phys* 1998, **25**(5):656-661.
8. Defoor DL, Stathakis S, Roring JE, Kirby NA, Mavroidis P, Obeidat M, Papanikolaou N: **Investigation of error detection capabilities of phantom, EPID and MLC log file based IMRT QA methods**. *J Appl Clin Med Phys* 2017, **18**(4):172-179.
9. Stasi M, Bresciani S, Miranti A, Maggio A, Sapino V, Gabriele P: **Pretreatment patient-specific IMRT quality assurance: a correlation study between gamma index and patient clinical dose volume histogram**. *Med Phys* 2012, **39**(12):7626-7634.
10. Nelms BE, Zhen H, Tome WA: **Per-beam, planar IMRT QA passing rates do not predict clinically relevant patient dose errors**. *Med Phys* 2011, **38**(2):1037-1044.
11. Garcia-Romero A, Hernandez-Vitoria A, Millan-Cebrian E, Alba-Escorihuela V, Serrano-Zabaleta S, Ortega-Pardina P: **On the new metrics for IMRT QA verification**. *Medical Physics* 2016, **43**(11):6058-6071.

12. Song JH, Kim MJ, Park SH, Lee SR, Lee MY, Lee DS, Suh TS: **Gamma analysis dependence on specified low-dose thresholds for VMAT QA**. *Journal of Applied Clinical Medical Physics* 2015, **16**(6):263-272.
13. Heilemann G, Poppe B, Laub W: **On the sensitivity of common gamma-index evaluation methods to MLC misalignments in rapidarc quality assurance**. *Medical Physics* 2013, **40**(3):031702.
14. Park JM, Kim JI, Park SY, Oh DH, Kim ST: **Reliability of the gamma index analysis as a verification method of volumetric modulated arc therapy plans**. *Radiation Oncology* 2018, **13**(1):175.
15. Wootton LS, Nyflot MJ, Chaovalitwongse WA, Ford E: **Error Detection in Intensity-Modulated Radiation Therapy Quality Assurance Using Radiomic Analysis of Gamma Distributions**. *Int J Radiat Oncol Biol Phys* 2018, **102**(1):219-228.
16. Steers JM, Fraass BA: **The effect of dose gradients on gamma comparison insensitivity in patient specific QA comparisons**. *Med Phys* 2023, **50**(6):3671-3686.
17. Kimura Y, Kadoya N, Oku Y, Kajikawa T, Tomori S, Jingu K: **Error detection model developed using a multi-task convolutional neural network in patient-specific quality assurance for volumetric-modulated arc therapy**. *Medical Physics* 2021, **48**(9):4769-4783.
18. Kry SF, Molineu A, Kerns JR, Faught AM, Huang JY, Pulliam KB, Tonigan J, Alvarez P, Stingo F, Followill DS: **Institutional patient-specific IMRT QA does not predict unacceptable plan delivery**. *Int J Radiat Oncol Biol Phys* 2014, **90**(5):1195-1201.
19. Wootton LS, Nyflot MJ, Chaovalitwongse WA, Ford E: **Error detection in intensity-modulated radiation therapy quality assurance using radiomic analysis of gamma distributions**. *International Journal of Radiation Oncology*Biology*Physics* 2018, **102**(1):219-228.
20. Ma C, Wang R, Zhou S, Wang M, Yue H, Zhang Y, Wu H: **The structural similarity index for IMRT quality assurance: radiomics-based error classification**. *Med Phys* 2021, **48**(1):80-93.
21. Wolfs CJA: **What is the optimal input information for deep learning-based pre-treatment error identification in radiotherapy?** *Physics and Imaging in Radiation Oncology* 2022.
22. Zhang D, Wang B, Zhang G, Ma C, Deng X: **Comparison of 3D and 2D gamma passing rate criteria for detection sensitivity to IMRT delivery errors**. *Journal of Applied Clinical Medical Physics* 2018, **19**(4):230-238.
23. Visser R, Wauben DJL, De Groot M, Godart J, Langendijk JA, Van'T Veld AA, Korevaar EW: **Efficient and reliable 3D dose quality assurance for IMRT by combining independent dose calculations with measurements: Efficient and reliable 3D dose QA for IMRT**. *Medical Physics* 2013, **40**(2):021710.
24. Mayerhoefer ME, Materka A, Langs G, Haggstrom I, Szczypinski P, Gibbs P, Cook G: **Introduction to Radiomics**. *J Nucl Med* 2020, **61**(4):488-495.
25. Lam D, Zhang X, Li H, Deshan Y, Schott B, Zhao T, Zhang W, Mutic S, Sun B: **Predicting gamma passing rates for portal dosimetry-based IMRT QA using machine learning**. *Med Phys* 2019, **46**(10):4666-4675.
26. Zhu H, Zhu Q, Wang Z, Yang B, Zhang W, Qiu J: **Patient-specific quality assurance prediction models based on machine learning for novel dual-layered MLC linac**. *Med Phys* 2023, **50**(2):1205-1214.
27. Sivabhaskar S, Li R, Roy A, Fakhreddine M, Papanikolaou N, Kirby N: **Machine learning models to predict the delivered positions of Elekta multileaf collimator leaves for volumetric modulated arc therapy**.
28. Sakai M, Nakano H, Kawahara D, Tanabe S, Takizawa T, Narita A, Yamada T, Sakai H, Ueda M, Sasamoto R *et al*: **Detecting MLC modeling errors using radiomics-based machine learning in patient-specific QA with an EPID for intensity-modulated radiation therapy**. *Medical Physics* 2021, **48**(3):991-1002.
29. Steers JM, Fraass BA: **IMRT QA: Selecting gamma criteria based on error detection sensitivity**. *Medical Physics* 2016, **43**(4):1982-1994.
30. Steers JM, Fraass BA: **IMRT QA and gamma comparisons: The impact of detector geometry, spatial sampling, and delivery technique on gamma comparison sensitivity**. *Med Phys* 2021, **48**(9):5367-5381.
31. Timakova E, Zavgorodni SF: **Effect of modulation factor and low dose threshold level on gamma pass rates of single isocenter multi-target SRT treatment plans**. *Journal of Applied Clinical Medical Physics* 2024, **25**(9):e14459.
32. Lu W, Li Y, Huang W, Cui H, Zhang H, Yi X: **Optimizing the Region for Evaluation of Global Gamma Analysis for Nasopharyngeal Cancer (NPC) Pretreatment IMRT QA by COMPASS: A Retrospective Study**. *Front Oncol* 2022, **12**:859415.
33. Malhotra A, Carpentier EE, Duzenli C: **Recommendations for using analytical anisotropic algorithm and AcurosXB for epidermal dose calculations in breast radiotherapy from an in vivo Gafchromic film study**. *Journal of Applied Clinical Medical Physics* 2024,

25(8):e14416.

34. Huang Y, Pi Y, Cai R, Ma K, Wang H, Chen H, Gu H, Shao Y, Feng A, Duan Y *et al*: **Prediction of three-dimensional dose distribution for patient-specific quality assurance based on log files using WingsNet**. *Radiation Oncology* 2025, **20**(1):171.
35. Han C, Yu W, Zheng X, Zhou Y, Gong C, Xie C, Jin X: **Composite QA for intensity-modulated radiation therapy using individual volume–based 3D gamma indices**. *Journal of Radiation Research* 2018, **59**(5):669-676.
36. Ishizaka N, Kinoshita T, Sakai M, Tanabe S, Nakano H, Tanabe S, Nakamura S, Mayumi K, Akamatsu S, Nishikata T *et al*: **Prediction of patient-specific quality assurance for volumetric modulated arc therapy using radiomics-based machine learning with dose distribution**. *Journal of Applied Clinical Medical Physics* 2024, **25**(1):e14215.
37. Wolfs CJA, Canters RAM, Verhaegen F: **Identification of treatment error types for lung cancer patients using convolutional neural networks and EPID dosimetry**. *Radiotherapy and Oncology* 2020, **153**:243-249.
38. van Elmpt W, McDermott L, Nijsten S, Wendling M, Lambin P, Mijnheer B: **A literature review of electronic portal imaging for radiotherapy dosimetry**. *Radiotherapy and Oncology* 2008, **88**(3):289-309.
39. Kimura Y, Kadoya N, Oku Y, Jingu K: **Development of a deep learning-based error detection system without error dose maps in the patient-specific quality assurance of volumetric modulated arc therapy**. *Journal of Radiation Research* 2023, **64**(4):728-737.
40. van den Berg K, Wolfs CJA, Verhaegen F: **A 3D transfer learning approach for identifying multiple simultaneous errors during radiotherapy**. *Phys Med Biol* 2024, **69**(3).
41. Chu H, De Vette SPM, Neh H, Sijtsema NM, Steenbakkers RJHM, Moreno A, Langendijk JA, Van Ooijen PMA, Fuller CD, Van Dijk LV: **Three-dimensional deep learning normal tissue complication probability model to predict late xerostomia in patients with head and neck cancer**. *International Journal of Radiation Oncology*Biology*Physics* 2025, **121**(1):269-280.
42. De Vette SPM, Neh H, Van Der Hoek L, MacRae DC, Chu H, Gawryszuk A, Steenbakkers RJHM, Van Ooijen PMA, Fuller CD, Hutcheson KA *et al*: **Deep learning NTCP model for late dysphagia after radiotherapy for head and neck cancer patients based on 3D dose, CT and segmentations**. *Radiotherapy and Oncology* 2025, **213**:111169.
43. Yang Z, Noble DJ, Shelley L, Berger T, Jena R, McLaren DB, Burnet NG, Nailon WH: **Machine-learning with region-level radiomic and dosimetric features for predicting radiotherapy-induced rectal toxicities in prostate cancer patients**. *Radiotherapy and Oncology* 2023, **183**:109593.
44. Jomy J, Sharma R, Lu R, Chen D, Ataalla P, Kaushal S, Liu ZA, Ye XY, Fairchild A, Nichol A *et al*: **Clinical impact of radiotherapy quality assurance results in contemporary cancer trials: a systematic review and meta-analysis**. *Radiother Oncol* 2025, **207**:110875.
45. Keall PJ, El Naqa I, Fast MF, Hewson EA, Hindley N, Poulsen P, Sengupta C, Tyagi N, Waddington DEJ: **Real-Time Dose-Guided Radiation Therapy**. *Int J Radiat Oncol Biol Phys* 2025, **122**(4):787-801.

Figures

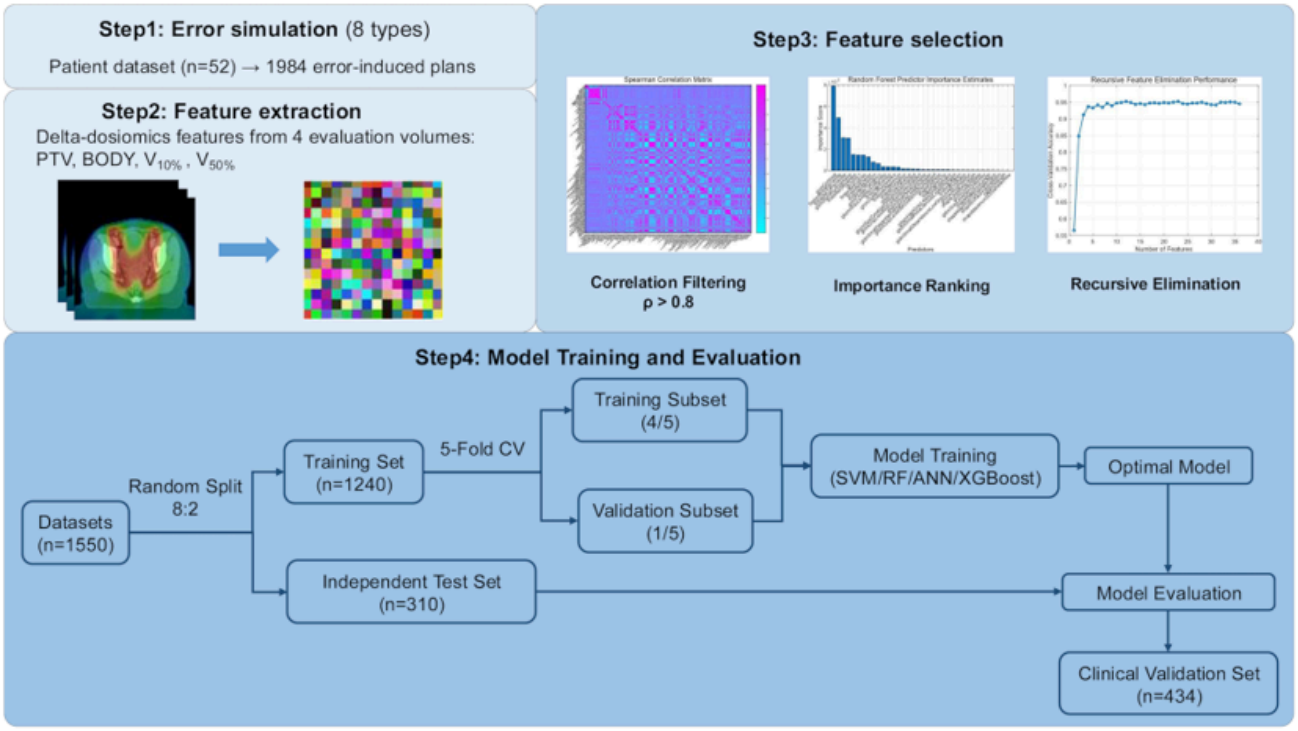


Figure 1

Workflow for radiotherapy treatment error identification in patient-specific quality assurance. BODY: external body contour; PTV: planning target volume; $V_{10\%}$: 10% prescription isodose volume; $V_{50\%}$: 50% prescription isodose volume; XGBoost: extreme gradient boosting; RF: random forest; SVM: support vector machine; ANN: artificial neural network.

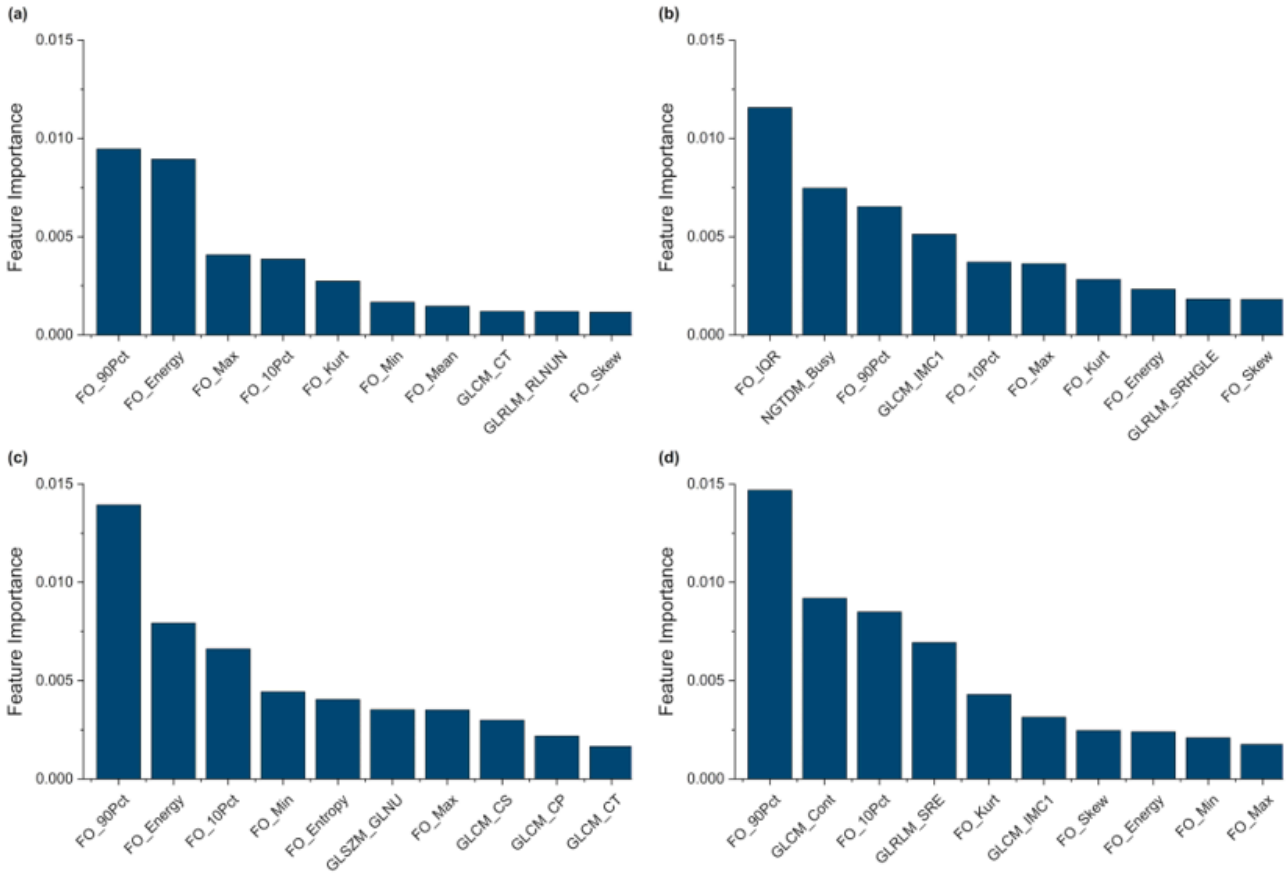


Figure 2

Feature importance ranking for four evaluation volumes. (a) BODY, (b) PTV, (c) $V_{10\%}$, and (d) $V_{50\%}$. FO: first-order; GLCM: gray level co-occurrence matrix; GLRLM: gray level run length matrix; GLSZM: gray level size zone matrix; NGTDM: neighboring gray tone difference matrix; Pct: percentile; Cont: contrast; SRE: short-run emphasis; Kurt: kurtosis; IMC1: informational measure of correlation 1; Skew: skewness; Min: minimum; Max: maximum; GLNU: gray-level non-uniformity; CS: cluster shade; CP: cluster prominence; CT: cluster tendency; RLNUN: run length non-uniformity normalized; IQR: interquartile range; Busy: busyness; SRHGLE: short-run high gray-level emphasis.

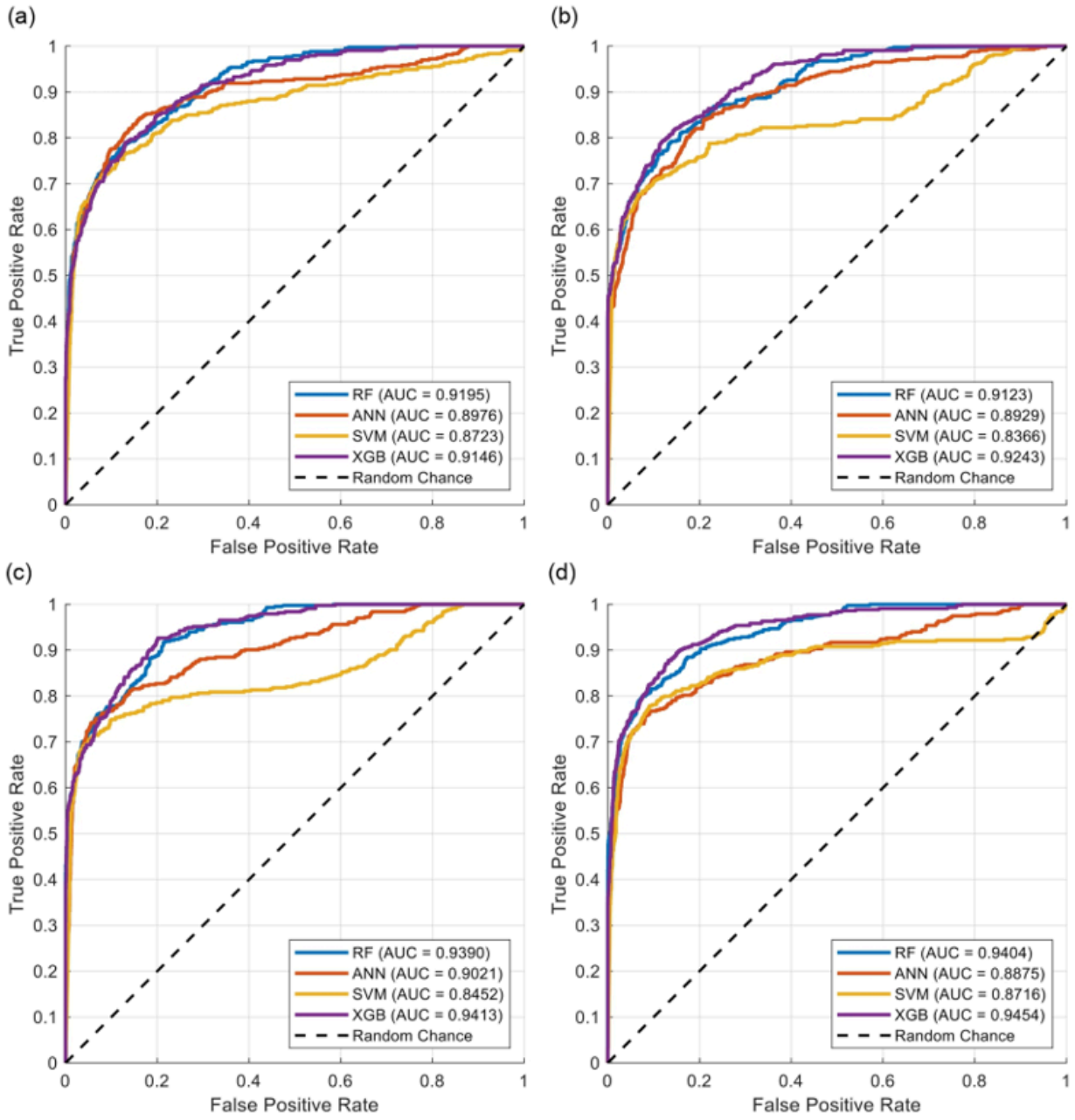


Figure 3

Receiver operating characteristic curves of the four machine learning algorithms for Level 1 error classification on the clinical validation set across different evaluation volumes. (a) BODY, (b) PTV, (c) $V_{10\%}$, and (d) $V_{50\%}$.

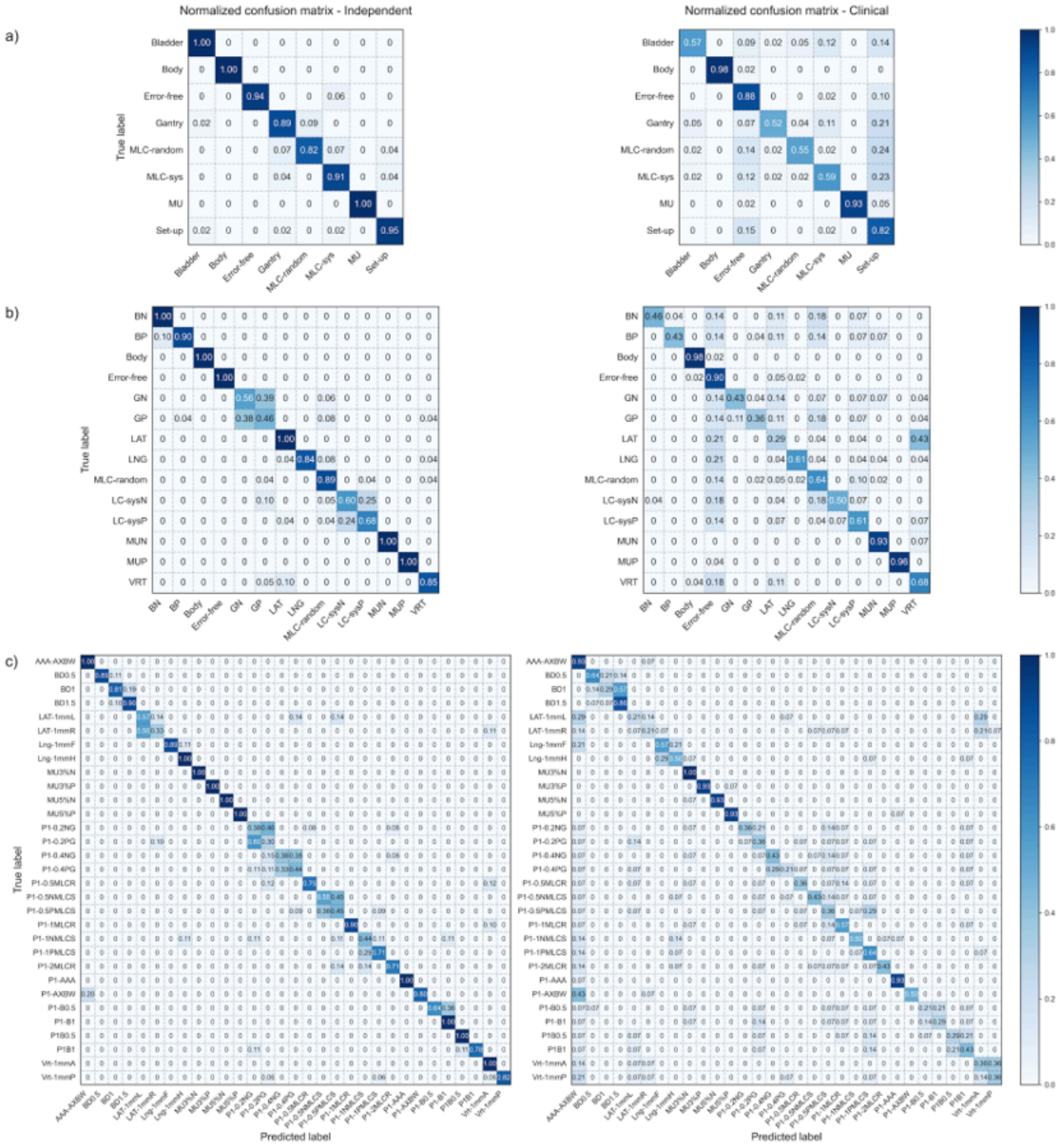


Figure 4

Normalized confusion matrices of the $V_{50\%}$-based XGBoost model on the independent test set (left column) and clinical validation set (right column). The rows correspond to three classification levels: (a) Level 1 (main error type), (b) Level 2 (error direction), and (c) Level 3 (error magnitude).