ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

# Characterization of radiation-induced esophagitis using explainable machine learning algorithms for lung cancer patients treated in the NRG/RTOG 0617clinical trial

**Taman Upadhaya*, Indrin J. Chetty, Joseph Lui, and Katelyn M. Atkin**

Department of Radiation Oncology, Cedars-Sinai Medical Center, Los Angeles, California, United States of America

## Abstract

Limiting acute esophagitis remains a clinical challenge in the treatment of locally advanced non-small cell lung cancer using chemoradiotherapy. In this study, machine learning (ML) algorithms were used to predict acute esophagitis in patients ($n = 451$; training set 70% and independent test set 30%) treated in the Radiation Therapy Oncology Group 0617 clinical trial. Multiple ML models were trained to predict grade ≥ 2 esophagitis using clinical, radiomics, and dosimetric/dosiomics features. Model performance was evaluated using the area under the receiver operating characteristic curve (AUC). The best-performing model was explained using the Shapley additive explanation framework and decision curve analysis. The highest-accuracy model was based on radiomics–dosiomics features using the least absolute shrinkage and selection operator model (AUC = 0.70; 95% confidence interval (CI): 0.60–0.78). A novel model that accounts for esophageal geometry relative to tumor location achieved an AUC of 0.67 (95% CI: 0.58–0.76) using dose–volume histogram relationships. The ML models revealed important relationships that can guide radiation dose planning to minimize individual risk and facilitate personalized radiation therapy.

***Keywords:*** Machine learning; Esophagitis; Radiomics; Dosiomics; Locally advanced non-small cell lung cancer; RTOG 0617 trial

***Corresponding author:**
Taman Upadhaya
(Taman.Upadhaya@csmc.edu)

**Citation:** Upadhaya T, Chetty IJ, Lui J, Atkin KM. Characterization of radiation-induced esophagitis using explainable machine learning algorithms for lung cancer patients treated in the NRG/RTOG 0617clinical trial. *Artif Intell Health*. doi: 10.36922/AIH026220054

**Received:** May 27, 2026

**Revised:** July 1, 2026

**Accepted:** July 7, 2026

**Published online:** October 8, 2026

**Copyright:** © 2026 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

**Publisher's Note:** AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## 1. Introduction

In patients with unresectable locally advanced non-small cell lung cancer (LA-NSCLC), definitive chemoradiotherapy (CRT) improves outcomes but is associated with clinically significant side effects.[1] Acute esophagitis can be of major concern, resulting in hospitalization, feeding-tube placement, treatment interruptions, and poorer clinical outcomes.[2] Therefore, robust methods to predict acute esophagitis are warranted and may help transform radiotherapy dose exposure into a modifiable risk factor during CRT treatment planning.

The Quantitative Analyses of Normal Tissue Effects in the Clinic[3] report provides guidance on aggregate, population-based dose constraints, such as dose–volume constraints, to limit esophagitis (grade ≥ 2) to <30%. However, these broad estimates are population-based and do not consider patient-specific factors. Multiple studies have

shown that a wide range of dose–volume metrics (absolute volume, absolute area, or percentage of a reference volume, or reference area) correlate with severe acute esophagitis after radiotherapy, but no single threshold has been established.[4-6] While some studies have identified significant associations across doses ranging from 10–80 Gy in univariate analyses, they were not consistent multivariate predictors.[7-8] A meta-analysis conducted on 1,082 individual patient data identified esophagus V60 Gy as the most reliable predictor, with V60 Gy < 0.07% best identifying low esophagitis risk and V60 ≥ 17% denoting high risk.[9] In another study, researchers identified contralateral (CE) esophagus-sparing as a method to reduce esophagitis among patients treated with chemoradiation therapy. Using CE-sparing, albeit with a limited sample size ($n$ = 27), the researchers reported no grade ≥ 3 esophagitis despite some tumors being located within 1 cm of the esophagus.[10] While traditional dose–volume metrics are informative for predicting esophageal toxicity, they do not account for patient-specific factors and imaging biomarkers, which may improve model predictive accuracy.

In recent years, radiomic imaging biomarkers have been shown to serve as signatures of clinical endpoints such as overall survival, disease progression, and tumor response.[11,12] Radiomics refers to the high-throughput extraction of high-dimensional quantitative features from medical images available to the oncology imaging community.[13-15] Radiomics and dosiomics characterize voxel images and three-dimensional (3D) dose intensities, respectively, facilitating detection of subtle tissue differences and a better understanding of treatment response in tumor and healthy tissue. Studies have shown that inclusion of radiomics and/or dosiomics biomarkers improves esophagitis predictive performance relative to traditional metrics.[16-19] Although these results are encouraging, studies are often limited by single-center datasets that may not be representative of broader population cohorts, lack of curated data, small sample sizes, and limited validation.[16-18]

In this study, we performed multivariate analysis to predict the risk of esophagitis for LA-NSCLC patients treated with concurrent CRT in a cooperative group (NRG Oncology [NRG]/Radiation Therapy Oncology Group [RTOG] 0617) clinical trial.[20-21] The multivariate features included radiomics, dosiomics (3D dose distributions), and patient-specific parameters (e.g., demographics, performance status, comorbidities, smoking status). Additionally, we incorporated a novel geometry-based analysis in which the esophagus was divided into four quadrants, and tumor location and dose–volume parameters were analyzed with respect to the specific quadrant. We assessed the individual predictive value of each feature set and systematically compared predictive performance across six machine learning (ML) models.

## 2. Data and methods

### 2.1. Clinical cohort description

This study utilized 451 patient datasets from the National Clinical Trial (NCT) NCT00533949-D1, -D2, and -D3, available through the National Cancer Institute's National Clinical Trials Network (NCTN)/National Cancer Institute Community Oncology Research Program data archive.[20-21] This clinical trial (NCT00533949, NRG Oncology/RTOG 0617) compared standard- versus high-dose CRT with or without cetuximab for unresectable stage III NSCLC. Data included pre-treatment non-contrast computed tomography (CT) images (slice thickness 2.5–3.2 mm), radiotherapy (RT) plans, and clinical variables such as tumor histology, age, gender, smoking history, and esophagitis status (Table 1).[21-22] Outcomes data included tumor control and normal tissue toxicities, such as esophagitis rates, scored according to the Common Terminology Criteria for Adverse Events, version 3.[23] All analyses and interpretations presented in this study are solely those of the authors and do not necessarily represent the views of the clinical trial investigators, the NCTN, or the National Cancer Institute.

### 2.2. Esophageal volume-of-interest

The esophagus was divided and analyzed in relation to the tumor location. Subdivided regions included the four quadrants (**Figure S1**). Radiomics and dosiomics feature extraction was limited to the entire esophageal region, whereas the novel dose–volume histogram (DVH) analysis encompassed the entire esophagus and its four quadrants. Additional details on the esophageal regional volume-of-interest (VOI) extraction process are provided in **Section A1**.

### 2.3. Dose–volume histograms

In addition, DVHs for the planning target volume (PTV), clinical target volume, and esophageal normal tissue were incorporated. Detailed dosimetric metrics were extracted for Vx (volume receiving ≥ x Gy) and Dx (dose delivered to x% of the organ volume) in 5-Gy increments, spanning V5 Gy to V100 Gy and D5% to D100%.[24-25]

### 2.4. Radiomics and dosiomics features

Contours of the esophageal tissue (esophagus–gross tumor volume) or of esophageal VOIs defined on the planning CT images and 3D dose maps were analyzed. Handcrafted radiomics features were extracted from VOIs on the CT

images, whereas dosiomics features were extracted from 3D dose maps according to the guidelines defined by the Image Biomarker Standardization Initiative (IBSI).[26] These included 29 morphological features, 50 intensity-based features (18 statistical features, 23 intensity histogram, 7 intensity-volume histogram, 2 local intensity features), and 94 textural features (25 grey level co-occurrence matrix, 16 grey level run length matrix, 16 grey level size zone matrix, 21 neighboring grey level dependence matrix, 16 grey level distance zone matrix), extracted in 3D according to the guidelines defined by the IBSI.[27] Intensity histogram and textural features were computed using 16 different parameter combinations to optimize features.[26] This process resulted in different "variants" for each feature, which were analyzed and optimized as (i) isotropic resampling to voxel sizes of 1 and 2 $mm^3$ (i.e., images were first interpolated to these dimensions using linear interpolation), (ii) fixed bin number and fixed bin size gray-level discretization, and (iii) number of gray levels: 16, 32, 64, and 128. A total of 9,296 features were extracted from the esophagus minus clinical tumor volume. VOI's in both RT dose units and CT Hounsfield units (intensities) were acquired. Image filtering (mean, Laplacian of Gaussian, Laws, Gabor, and wavelets [high-pass and low-pass]) was applied in accordance with the IBSI convolutional filter definitions.[28] Additional details on the filter parameter step can be found in **Table S1**.

### 2.5. Univariate analysis

Univariate analysis was performed on the training data to investigate associations between individual features and the outcome (grade ≥ 2 esophagitis) to reduce dimensionality. First, the Chi-square test was applied for categorical features, whereas the Wilcoxon rank-sum test (or Welch's *t*-test as appropriate) was applied for numerical features to identify statistically significant features. Second, area under the receiver operating characteristic curve (AUC) was calculated to identify and rank the initial features. For all AUCs, 95% confidence intervals (CIs) were estimated using 1,000 bootstrap iterations.[29] Finally, correlation analysis (correlation coefficient [CC] ≤ 0.8) was applied to identify independent predictors. Among correlated features, those with the highest AUC rank value were selected. Additionally, univariate logistic regression models were fitted for each feature, and odds ratios were computed.

### 2.6. Multivariate analysis

Multivariate analysis was conducted to assess the interactions among features and to quantify their relevance and redundancy. This was a multi-step workflow comprising feature ranking, feature subset selection, and classification (Figure 1C).

#### *2.6.1. Workflow*

We began by retaining features that were statistically significant ($p$-value < 0.05) and uncorrelated (CC ≤ 0.80) in the univariate analysis, providing unsupervised dimensionality reduction, which improves predictive performance and interpretability.[13,14] We then performed model-based (supervised) feature selection within a nested cross-validation loop, ranking and selecting features sequentially to reduce dimensionality and model complexity. We evaluated six feature-selection algorithms (as detailed in **Section 2.6.2**). Pre-processing was applied to remove zero-variance features and to center the data (mean of zero and a standard deviation of one). Class imbalance was handled using the synthetic minority over-sampling technique, and parameter tuning was performed using nested 10-fold cross-validation (CV) via stratified random sampling.[30] Parameter estimation for each ML model was performed via a grid search with nested CV over hundreds of iterations on the training/validation samples. Models were fit independently within each algorithm's CV loop, yielding six models per feature type and 42 models in total. All model-building steps (univariate feature selection, parameter tuning, feature subset selection) were strictly applied only to the training dataset to prevent data leakage and overly optimistic performance estimates. The final model for testing was chosen based on the highest area under the curve in CV. The detailed workflow is shown in Figure 1C.

#### *2.6.2. Machine learning algorithms*

The six ML models for classification included, i.e., generalized linear model (GLM) with least absolute shrinkage and selection operator (LASSO) penalty (LASSO-GLM)[31], random forest (RF)[32], extreme gradient boosting (XGBoost)[33], multilayer perceptron neural network algorithm (MLP-ANN)[34], support vector machine (SVM) with linear and SVM with radial basis function (RBF) kernel.[35] Finally, we developed an ensemble model using weighted soft voting across all of the above classifiers.[36] Details about the model types are presented in **Section A2**.

#### *2.6.3. Feature selection and ranking*

For feature subset selection, we utilized importance and permutation measures, as appropriate for each algorithm, within the CV folds to avoid leakage.[29] Specifically, we ranked features by absolute weighted magnitude vector for the linear SVM, the total decrease in node impurities (Gini index[32]) averaged over all trees for RF, and the coefficients ranked by magnitude for the LASSO-GLM. For models

without direct, stable feature importances, such as XGBoost, SVM with an RBF kernel, and the MLP, we applied permutation importance, estimating each feature's contribution from the drop in validation performance (e.g., ΔAUC) after random shuffling with scores averaged over multiple iterations. In the final step, we used forward sequential selection based on the previous ranking to identify the smallest feature set achieving the highest AUC.

#### *2.6.4. Performance evaluation*

To assess the findings of this study, we reported accuracy, precision (positive predictive value), recall/sensitivity, specificity, false positive rate (1−specificity), F1 score, Matthews correlation coefficient (MCC), and the AUC.[37] All metrics were computed on the CV and holdout independent test datasets. The pairwise differences in AUC between classifiers were assessed using the DeLong test.[38] Finally, to understand the overall ML algorithms, Shapley additive explanation (SHAP) summary plots were generated for the esophageal toxicity models.[39]

#### *2.6.5. Software*

Image processing and feature extraction were performed with our collaboratively developed in-house toolkit MEDimage (documentation: https://medimage.readthedocs.io/en/latest/), an open-source framework for streamlined radiomics workflows. ML analyses were conducted in accordance with standard safeguards to prevent any data leakage. Visualization, statistical analysis, model development, and evaluation were conducted with an in-house ML codebase built on open-source Python packages: pandas, matplotlib, imbalanced-learn, and XGBoost. All models were implemented in Python 3.9 and the scikit-learn library (v1.2.2).[40,41]

### 2.7. Decision curve analysis

We performed decision curve analysis (DCA) to assess the clinical utility of our models by estimating net benefit across a range of clinically plausible threshold probabilities in both the validation and independent test cohorts.[42] DCA compares each model against the default "treat-all" and "treat-none" decisions, enabling direct evaluation of whether using the model would improve decision-making in practice. This approach is widely used to compare prediction models, biomarkers, and diagnostic tests, and assessing net benefit is explicitly recommended by the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis reporting guidelines for prediction models.[43] We plotted decision curves (net benefit versus threshold) for the best model and summarized the range of thresholds to facilitate comparison.

## 3. Results

### 3.1. Patient characteristics

In total, we included 451 patients. Among them, 315 were assigned to the training set and 136 to the independent test set, which was unseen. The rate of ≥grade 2 esophagitis was 39% in the training set (124/315) and 40% in the independent test set (54/136). The median age was 64 years in both cohorts. The gender distribution was approximately 60% male and 40% female in the training set, and 59% male and 41% female in the independent test set. Chemotherapy was administered to 48% of the training set and 45% of the independent test set. In the training set, 57% received 60 Gy and 43% received 70 Gy, and in the independent test set, 55% received 60 Gy and 45% received 70 Gy. All treatments were delivered using 3D conformal radiation therapy or intensity-modulated radiation therapy techniques. Additional patient characteristics are shown in Table 1**,** and univariate analysis results are provided in **Table S2**.

### 3.2. Univariate analysis

Results of the univariate analysis are provided in **Tables S2–S5**. Several features from various feature categories (DVH, radiomics, dosiomics) were associated with ≥ grade 2 esophagitis. We used correlation screening to control for redundancy and identify independent predictors.

Among the traditional clinical features, age and PTV maximum dose emerged as significant predictors of grade ≥ 2 esophagitis (**Table S2**).

For conventional whole-esophagus DVH metrics, 30 DVH features were significant, and six (V35, V50, V60, V70, D10cc, and D10) were retained as independent predictors of grade ≥ 2 esophagitis. **Table S3** lists all significant features.

Extending the analysis to our novel geometry-based DVH metrics for significant predictors of esophagitis, the 5-mm esophageal ring-based approach identified 34 statistically significant DVH features, with two (D15 and V65) remaining independent predictors of esophagitis (**Table S3**). The esophageal quadrant-based approach identified 155 statistically significant DVH features, of which 19 emerged as independent predictors of esophagitis. Features included: Quadrant I: mean and maximum dose, V65, D5cc, and D10; Quadrant II: mean and maximum dose, V65, V80, D15, and D5cc; Quadrant III: mean and maximum dose, D15, D5cc, and V65; and Quadrant IV: mean and maximum dose and D5. Detailed results are provided in **Table S3**.

Among dosiomics features extracted from the whole esophagus, 553 were significantly associated with ≥ grade

**Table 1. Tumor and treatment characteristics of the study participants**

| Characteristic | Description | |
|---|---|---|
| | Training no. (%) | Testing no. (%) |
| Number of patients | 315 (70%) | 136 (30%) |
| Age at diagnosis, median (range) | 64 (37–82) | 64 (40–83) |
| Esophagitis ≥ grade 2 | 124 (39%) | 54 (40%) |
| Gender | | |
| Male | 188 (60%) | 80 (59%) |
| Female | 127 (40%) | 56 (41%) |
| AJCC stage (6th edition, 2002) | | |
| IIIA, or N2/undetectable primary | 210 (67%) | 89 (65%) |
| IIIB, or N3/undetectable primary | 105 (33%) | 47 (35%) |
| Zubrod performance status | | |
| Normal activity | 179 (57%) | 85 (63%) |
| Symptoms, but nearly fully ambulatory | 136 (43%) | 51 (37%) |
| Smoking history | | |
| Non-smoker | 22 (7%) | 6 (4%) |
| Former light smoker | 26 (8%) | 13 (10%) |
| Former heavy smoker | 112 (36%) | 44 (32%) |
| Current smoker | 141 (45%) | 65 (48%) |
| Unknown | 14 (4%) | 8 (6%) |
| Chemotherapy (cetuximab) | | |
| Yes | 150 (48%) | 61 (45%) |
| No | 165 (52%) | 75 (55%) |
| Radiotherapy | | |
| 60 Gy | 180 (57%) | 75 (55%) |
| 70 Gy | 135 (43%) | 61 (45%) |
| RT technique | | |
| 3D-CRT | 165 (52%) | 68 (50%) |
| IMRT | 150 (48%) | 68 (50%) |
| Histology | | |
| Squamous cell carcinoma | 139 (44%) | 57 (42%) |
| Adenocarcinoma | 126 (40%) | 52 (38%) |
| Large cell undifferentiated | 9 (3%) | 3 (2%) |
| Non-small cell lung cancer, NOS | 41 (13%) | 24 (18%) |
| Race | | |
| American Indian/Alaskan native | 2 (0.6%) | 0 (0%) |
| Asian | 7 (2.2%) | 3 (2%) |
| Black or African American | 28 (9%) | 15 (11%) |
| Native Hawaiian/other Pacific Islander | 1 (0.3%) | 0 (0%) |
| White | 273 (86.6%) | 118 (87%) |
| Unknown | 4 (1.3%) | 0 (0%) |

Abbreviations: 3D-CRT: Three-dimensional conformal radiation therapy; AJCC: American Joint Committee on Cancer; IMRT: Intensity-modulated radiation therapy; NOS: Not otherwise specified; RT: Radiotherapy.

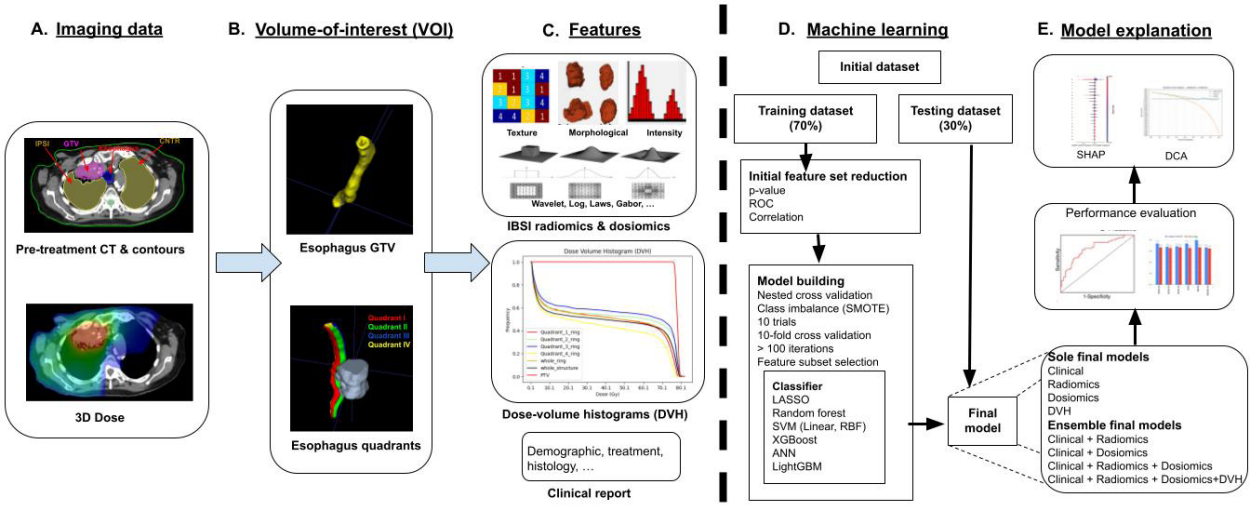



**Figure 1**. Workflow steps include: (A) collection of patient image (CT and dose) and clinical information dataset, (B) segmentation and definition of the esophagus regions (whole organ, 5-mm ring and four cross-sectional quadrants), (C) application of the IBSI convolution filters, and extraction of clinical, DVH, radiomics, dosiomics, and clinical report (D) feature selection and machine learning model for the development of the predictive models, and (E) model interpretation using DCA and SHAP.
Abbreviations: 3D: Three-dimensional; ANN: Artificial neural network; CT: Computed tomography; DCA: Decision curve analysis; DVH: Dose–volume histogram; GBM: Gradient boosting machine; GTV: Gross tumor volume; IBSI: Image Biomarker Standardization Initiative; LASSO: Least absolute shrinkage and selection operator; RBF: Radial basis function; ROC: Receiver operating characteristic; SHAP: Shapley additive explanation; SMOTE: Synthetic minority over-sampling technique; SVM: Support vector machine; XGBoost: Extreme gradient boosting.

2 esophagitis, and 39 of these emerged as independent predictors after controlling for redundancy (**Table S5**). There were seven intensity-based features and 32 texture-based features. Similarly, among radiomics features from the whole esophagus, 730 were significant, and 43 were independent predictors of esophagitis (**Table S4**), including nine intensity-based and 34 texture-based features. Detailed results are provided in **Tables S4** and **S5**.

### 3.3. Model performance and interpretation

We evaluated predictive performance using the AUC. Figure 2 summarizes AUCs across feature categories (radiomics, dosiomics, DVH), and other performance metrics are presented in **Figures S3** and **S4**. For each feature category, we paired six classifiers with their corresponding feature-selection schemes, using the top-ranked feature subset, and we evaluated six representative AUC values per category-classifier combination. In total, we implemented and evaluated 42 different ML models (**Figure S2**). These representative AUCs for all classifier pairs and SHAP[39] values are reported in **Figures S4**–**S5**.

The highest-accuracy model used radiomics and dosiomics features with the GLM/LASSO algorithm, yielding AUCs of 0.70/0.70 for the validation/test datasets, respectively. Figure 2B shows the corresponding SHAP analysis. The top-ranked features consisted of two dosiomics features and two radiomics features. Comprehensive performance metrics (including precision, recall, F1 score, MCC, and calibration) are provided in **Figure S3**.

The second-highest-performing model used geometric DVH features and the GLM/LASSO algorithm with 13 features, and achieved AUCs of 0.69/0.67 for the validation and test datasets, respectively. In comparison, the radiomics-only model achieved lower AUCs of 0.63 (CV) and 0.66 (independent testing), while the dosiomics-only model achieved AUCs of 0.68 and 0.67, respectively. The corresponding SHAP analysis is shown in **Figure S4**, with the top-ranked features for all the models and comprehensive metrics provided in **Figures S4** and **S5**.

We also performed analysis using a soft-voting ensemble model, which combines the probabilistic outputs of multiple (individually weaker) classifiers into a single model. This ensemble achieved AUCs of 0.71/0.69 for the validation and test datasets, respectively. Improvement in these scores (including an increase in the F1 score from 0.54 to 0.58) suggests that complementary information from different individual classifiers can enhance predictive performance. Pairwise DeLong test $p$-values for CV and independent test are provided in **Tables S6** and **S7**.

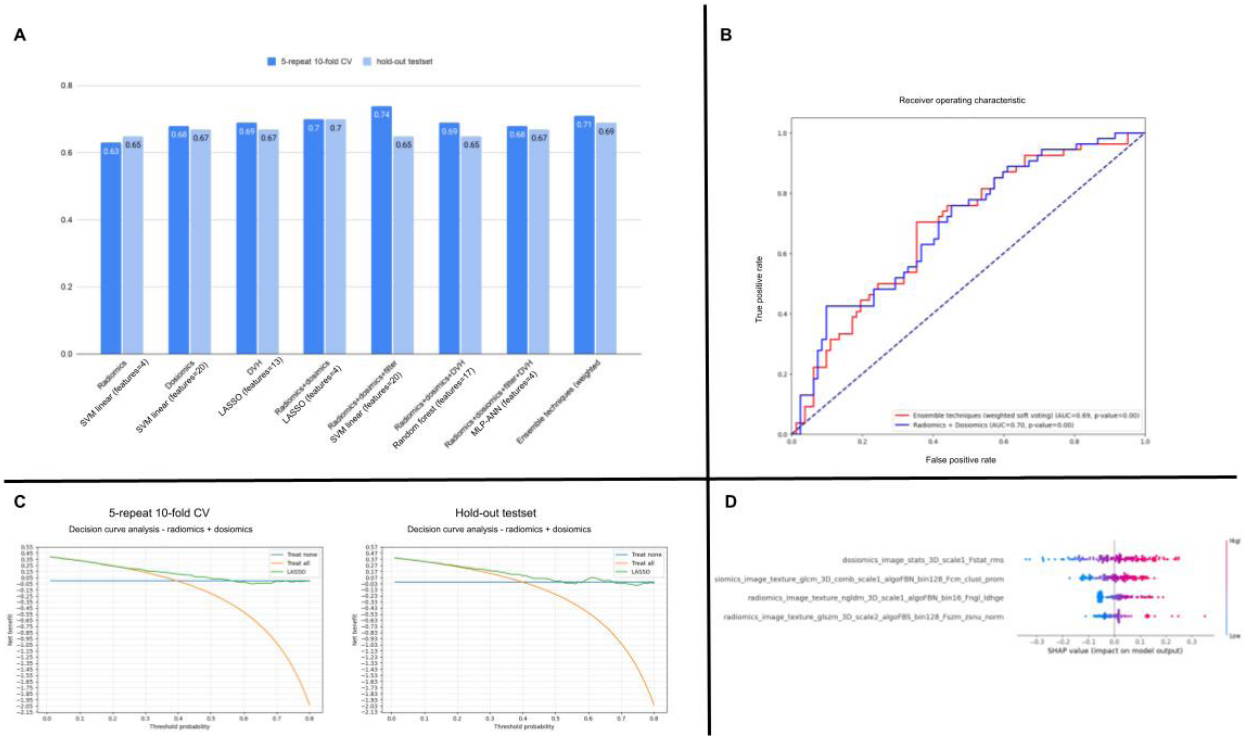



**Figure 2.** (A) Performance of models based on features from CT (radiomics) and 3D dose (dosiomics) using esophagus contours to predict RE ≥ grade 2. (B) The ROC curve for the highest-performing model on an unseen independent test set. Experiments were performed using 10-fold nested CV and are reported for the unseen, independent test dataset. Other performance metrics (precision, recall/sensitivity, specificity, MCC, and F1 score) are provided in the supplementary materials (**Figures S3–S5)**. (C) Decision curve analysis shows that our model for predicting grade ≥ 2 radiation esophagitis is in the clinically relevant threshold range. (D) SHAP plot of overall feature importance to the final machine learning model, ordered by mean absolute value of feature SHAP values.

Abbreviations: 3D: Three-dimensional; AUC: Area under the receiver operating characteristic curve; CT: Computed tomography; CV: Cross-validation; DVH: Dose–volume histogram; LASSO: Least absolute shrinkage and selection operator; MCC: Matthews correlation coefficient; MLP-ANN: Multilayer perceptron neural network algorithm; ROC: Receiver operating characteristic; RE: Radiation esophagitis; SHAP: Shapley additive explanation; SVM: Support vector machine.

### *3.3.1. Decision curve analysis*

The decision curves in Figure 2C show that radiomics- and dosiomics-combined models provide higher net benefit than the "treat-all" and "treat-none" strategies across a broad range of threshold probabilities. When tested using the validation datasets, the model threshold range was 0.040–0.560 for predicting ≥ grade 2 esophagitis; it was 0.09–0.67 for the test dataset. Within these ranges, using the model improves decision-making compared with default strategies. This DCA demonstrated that the proposed models provided a positive net clinical benefit over a broad range of threshold probabilities. In this context, the threshold probability represents the predicted risk of radiation-induced esophagitis at which a clinician would consider modifying the treatment plan or implementing preventive interventions. For example, clinicians with a lower risk tolerance may choose to intervene at lower threshold probabilities (e.g., 4%–10%), whereas others may reserve intervention for patients with substantially higher predicted risks. The observed net benefit across threshold ranges of 0.04–0.56 and 0.09–0.67 suggests the models may support clinical decision-making under a wide range of physician preferences and risk thresholds. This broad range suggests that the models could facilitate individualized treatment planning by identifying patients who may benefit from esophageal dose reduction or proactive toxicity mitigation while avoiding unnecessary treatment modifications for patients at low predicted risk.

## 4. Discussion

Recent advances in radiomics—the high-throughput extraction of quantitative imaging features—have enabled characterization of tissues relevant to malignancy or treatment response. In this study, we analyzed numerous

radiomics and dosiomics features to quantify associations between esophageal tissue phenotype and outcomes (esophagitis ≥ grade 2). The cohort consisted of 451 patients with LA-NSCLC. We also evaluated DVH metrics defined relative to esophageal anatomy and tumor location.

Prior work on radiomics and dosiomics as signatures of radiation esophagitis in LA-NSCLC is still limited. Zheng *et al.*[44] reported top performance in predicting ≥ grade 2 esophagitis in a 161-patient cohort (51 positives) from a radiomics-only model (AUC = 0.74) and a hybrid clinical + radiomics + dosiomics model (AUC = 0.75) on the test set. In postoperative RT (PORT-C), Ma *et al.*[16] showed that convolutional neural network-derived features outperformed dosimetric and handcrafted features (validation AUCs ≈ 0.785 versus 0.724), noting limitations including modest sample size, omission of key clinical covariates (e.g., age, chemotherapy), and a low incidence of esophagitis (11.6%). Luna *et al.*[17] studied grade ≥ 3 esophagitis using clinical and dosimetric factors in 202 NSCLC patients and achieved a maximum AUC of 0.56. Wang *et al.*[45] combined clinical, dosimetric, and pre-treatment cytokines for grade ≥ 3 esophagitis and achieved an AUC of 0.75. Nair *et al.*[18] recently conducted a multimodal positron emission tomography (PET)–CT dose-based study ($n$ = 179; 27 positives) using CT/PET radiomics and dosiomic DVH features with an RF classifier. Herr *et al.*[19] (Multi-Institutional Radiomics and Quality Control; $n$ = 1,760) analyzed clinical variables using only LASSO/GLM to predict grade ≥ 2 and grade ≥ 3 radiation esophagitis. Ladbury *et al.*[11] analyzed DVH and clinical data to predict grade ≥ 3 esophagitis, achieving an AUC of 0.706 on the NRG/RTOG clinical trial datasets but did not include radiomics/dosiomics features. Overall, radiomics and dosimetric predictors have shown promising results, but most prior studies rely on retrospective datasets with relatively small sample sizes. In contrast, our work uses prospectively collected data, which may reduce biases commonly seen in retrospective cohorts. To address key gaps in the literature, we analyzed a larger cohort with more esophagitis events ($n$ = 178), introduced novel geometry-aware DVH features, and compared multiple algorithms and feature families. Importantly, the planning CT datasets were obtained from a randomized clinical trial and were rigorously curated. Finally, because the data were collected across multiple institutions, the cohort is inherently more heterogeneous than single-center datasets, which can help reduce overly optimistic performance estimates and better reflect real-world variability.

Our study has some important limitations. Because the CT images were anonymized, we did not have institution or scanner details, which prevented feature harmonization (e.g., via the ComBat method). However, we generated probability density plots of CT intensity values across all patients and observed limited variation, partially mitigating this concern. Other than basic clinical variables and imaging features (radiomics and dosiomics) available in the RTOG 0617 dataset, potentially relevant factors such as genetic characteristics and tumor-specific biomarkers, which might play a key role in determining the risk of radiation esophagitis, were unavailable for analysis.

While other studies have reported the value of radiomics in predicting radiation esophagitis, to our knowledge, only two other studies have included geometric and dosimetric factors centered on tumor location relative to the esophagus. Monti *et al.*[24] divided the esophagus longitudinally into four subsites: cervical (cricoid to the sternal notch), upper thoracic (the sternal notch to the inferior azygos vein), middle thoracic (inferior azygos vein to lower pulmonary veins), and lower thoracic (lower pulmonary veins to gastroesophageal junction). In a prospective clinical trial (202 patients) comparing intensity-modulated RT with passive-scattering proton therapy, they found that the regions most strongly associated with the dose effect for esophagitis were the upper and middle thoracic esophagus. Kamran *et al.*[10] contoured the contralateral esophagus as an avoidance structure to create a steep dose falloff gradient across the esophagus. This technique was evaluated in a Phase I clinical trial of 27 patients. This approach was associated with a reduced risk of esophagitis in patients uniformly treated with chemoradiation therapy, with no esophagitis ≥ grade 3 events observed even when tumors were within 1 cm of the esophagus. Here, we further subdivided the esophagus into cross-sectional quadrants to identify potential locations more susceptible to radiation damage. The geometry-based analysis identified two esophageal ring-based metrics and 19 quadrant-based DVH metrics that were independent predictors of esophagitis ≥ grade 2 (**Table S3**), in addition to four conventional esophageal dose metrics. Moreover, the ML models trained on geometric DVH features achieved slightly better performance than dosiomics- or radiomics-only models, suggesting they provide strong complementary information. Our geometric algorithm derives the esophageal subdivision directly from the esophagus contour, making it simple to compute DVH metrics and requiring no manual labeling. Each quadrant was labeled counterclockwise on every slice using a consistent convention (e.g., starting from posterior-right to anterior-right), and this labeling was done from the superior to the inferior extent of the esophagus. This information can be used to devise treatment plans that take such regions into account during dose optimization, potentially reducing dose and reducing

the risk of developing esophagitis. These features can also be incorporated into future clinical workflows and trials for pre-treatment risk stratification (e.g., identifying high- versus low-risk patients) to support individualized treatment planning.

## 5. Conclusion

We explored a multi-regional and multi-omics approach to extract CT-based radiomics and 3D dosiomics features to predict radiation esophagitis in patients with LA-NSCLC. We developed a geometric model of the esophagus based on cross-sectional quadrants, enabling us to examine whether specific esophageal regions may be more susceptible to radiation-induced damage. These results suggest that radiomics and dosiomics on planning CT scans and 3D dose distributions provide complementary information to enhance model performance, implying that such underlying features serve as signatures for predicting esophagitis. The ensemble soft voting classifier improved model performance (based on F1 score). Although statistical methods were employed to minimize bias, other limitations prevent definitive conclusions, underscoring the need for further research in this space.

## Acknowledgments

The authors would like to thank Dr. Jeffrey Bradley, the entire NRG Oncology/RTOG 0617 team, and NCTN for making this data publicly available to the scientific community. The authors are indebted to the Cedars-Sinai Medical Center Biomedical Sciences Research Information and Scientific Computing Core for use of their institutional computational cluster for the quantitative imaging calculations reported in this work.

## Funding

This work was supported in part by a research grant from Varian Medical Systems/Siemens Healthineers for research unrelated to this topic and by Cedars-Sinai Cancer Developmental Funds.

## Conflict of interest

The authors declare no conflicts of interest.

## Author contributions

*Conceptualization:* Taman Upadhaya, Indrin J. Chetty, Katelyn M. Atkin
*Investigation:* Taman Upadhaya, Indrin J. Chetty, Katelyn M. Atkin
*Methodology:* Taman Upadhaya
*Formal analysis:* Taman Upadhaya
*Writing–original draft:* Taman Upadhaya, Indrin J. Chetty, Katelyn M. Atkin
*Writing–review & editing:* All authors

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Availability of data

NRG Oncology/RTOG 0617 team and NCTN make this data publicly available to the scientific community (https://www.cancerimagingarchive.net/collection/nsclc-cetuximab/).

## References

1. Curran WJ, Paulus R, Langer CJ, *et al.* Sequential vs Concurrent Chemoradiation for Stage III Non–Small Cell Lung Cancer: Randomized Phase III Trial RTOG 9410. *J Natl Cancer Inst.* 2011;103(19):1452-1460.

   doi: 10.1093/jnci/djr325

2. Rubin P, Constine LS, Marks LB, editors. ALERT—Adverse Late Effects of Cancer Treatment. Vol 1: General Concepts and Specific Precepts. Springer; 2014.

   doi: 10.1007/978-3-540-72314-1

3. Bentzen SM, Constine LS, Deasy JO, *et al.* Quantitative analyses of normal tissue effects in the clinic (QUANTEC): an introduction to the scientific issues. *Int J Radiat Oncol Biol Phys*. 2010;76(3):S3-S9.

   doi: 10.1016/j.ijrobp.2009.09.040

4. Chapet O, Kong FM, Lee JS, Hayman JA, Ten Haken RK. Normal tissue complication probability modeling for acute esophagitis in patients treated with conformal radiation therapy for non-small cell lung cancer. *Radiother Oncol*. 2005;77(2):176-181.

   doi: 10.1016/j.radonc.2005.10.001

5. Bradley J, Deasy JO, Bentzen SM, El Naqa I. Dosimetric correlates for acute esophagitis in patients treated with radiotherapy for lung carcinoma. *Int J Radiat Oncol Biol Phys*. 2004;58(4):1106-1113.

   doi: 10.1016/j.ijrobp.2003.09.080

6. El Naqa I, Bradley J, Blanco AI, *et al.* Multivariable modeling of radiotherapy outcomes, including dose-volume and clinical factors. *Int J Radiat Oncol Biol Phys*. 2006;64(4):1275-1286.

   doi: 10.1016/j.ijrobp.2005.11.022

7. Wei X, Liu HH, Tucker SL, *et al.* Risk factors for acute esophagitis in non-small-cell lung cancer patients treated with concurrent chemotherapy and three-dimensional

conformal radiotherapy. *Int J Radiat Oncol Biol Phys*. 2006;66(1):100-107.

doi: 10.1016/j.ijrobp.2006.04.022

8. Rodríguez N, Algara M, Foro P, *et al.* Predictors of acute esophagitis in lung cancer patients treated with concurrent three-dimensional conformal radiotherapy and chemotherapy. *Int J Radiat Oncol Biol Phys*. 2009;73(3):810-817.

doi: 10.1016/j.ijrobp.2008.04.064

9. Palma DA, Senan S, Oberije C, *et al.* Predicting esophagitis after chemoradiation therapy for non-small cell lung cancer: an individual patient data meta-analysis. *Int J Radiat Oncol Biol Phys*. 2013;87(4):690-696.

doi: 10.1016/j.ijrobp.2013.07.029

10. Kamran SC, Yeap BY, Ulysse CA, *et al.* Assessment of a contralateral esophagus-sparing technique in locally advanced lung cancer treated with high-dose chemoradiation: a phase 1 nonrandomized clinical trial. *JAMA Oncol*. 2021;7(6):910-914.

doi: 10.1001/jamaoncol.2021.0281

11. Ladbury C, Li R, Danesharasteh A, *et al.* Explainable artificial intelligence to identify dosimetric predictors of toxicity in patients with locally advanced non-small cell lung cancer: a secondary analysis of RTOG 0617. *Int J Radiat Oncol Biol Phys*. 2023;117(5):1287-1296.

doi: 10.1016/j.ijrobp.2023.06.019

12. Lambin P, Leijenaar RTH, Deist TM, *et al.* Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol*. 2017;14(12):749-762.

doi: 10.1038/nrclinonc.2017.141

13. Leger S, Zwanenburg A, Pilz K, *et al.* A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling. *Sci Rep*. 2017;7(1):13206.

doi: 10.1038/s41598-017-13448-3

14. Giger ML. Computer-aided detection and diagnosis/radiomics/machine learning/deep learning in medical imaging. *Med Phys*. 2023;50(S1):50-53.

doi: 10.1002/mp.16025

15. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology*. 2016;278(2):563-577.

doi: 10.1148/radiol.2015151169

16. Ma Z, Liang B, Wei R, *et al.* Enhanced prediction of postoperative radiotherapy-induced esophagitis in non-small cell lung cancer: dosiomic model development in a real-world cohort and validation in the PORT-C randomized controlled trial. *Thorac Cancer*. 2023;14(28):2839-2845.

doi: 10.1111/1759-7714.15068

17. Luna JM, Chao HH, Shinohara RT, *et al.* Machine learning highlights the deficiency of conventional dosimetric constraints for prevention of high-grade radiation esophagitis in non-small cell lung cancer treated with chemoradiation. *Clin Transl Radiat Oncol*. 2020;22:69-75.

doi: 10.1016/j.ctro.2020.03.007

18. Nair SS, Salazar RM, Xu T, *et al.* Multiomic random forest toxicity modeling of radiation esophagitis. *Phys Imaging Radiat Oncol*. 2025;36:100838.

doi: 10.1016/j.phro.2025.100838

19. Herr DJ, Yin H, Bergsma D, *et al.* Factors associated with acute esophagitis during radiation therapy for lung cancer. *Radiother Oncol*. 2024;197:110349.

doi: 10.1016/j.radonc.2024.110349

20. Bradley JD, Hu C, Komaki RR, *et al.* Long-term results of NRG Oncology RTOG 0617: standard- versus high-dose chemoradiotherapy with or without cetuximab for unresectable stage III non–small-cell lung cancer. *J Clin Oncol*. 2020;38(7):706-714.

doi: 10.1200/JCO.19.01162

21. Bradley JD, Paulus R, Komaki R, *et al.* Standard-dose versus high-dose conformal radiotherapy with concurrent and consolidation carboplatin plus paclitaxel with or without cetuximab for patients with stage IIIA or IIIB non-small-cell lung cancer (RTOG 0617): a randomised, two-by-two factorial phase 3 study. *Lancet Oncol*. 2015;16(2):187-199.

doi: 10.1016/S1470-2045(14)71207-0

22. Clark K, Vendt B, Smith K, *et al.* The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging*. 2013;26(6):1045-1057.

doi: 10.1007/s10278-013-9622-7

23. Trotti A, Colevas AD, Setser A, *et al.* CTCAE v3.0: development of a comprehensive grading system for the adverse effects of cancer treatment. *Semin Radiat Oncol*. 2003;13(3):176-181.

doi: 10.1016/S1053-4296(03)00031-6

24. Monti S, Xu T, Mohan R, Liao Z, Palma G, Cella L. Radiation-induced esophagitis in non-small-cell lung cancer patients: voxel-based analysis and normal tissue complication probability modeling. *Cancers (Basel)*. 2022;14(7):1833.

doi: 10.3390/cancers14071833

25. Drzymala RE, Mohan R, Brewster L, *et al.* Dose-volume histograms. *Int J Radiat Oncol Biol Phys*. 1991;21(1):71-78.

doi: 10.1016/0360-3016(91)90168-4

26. Zwanenburg A, Vallières M, Abdalah MA, *et al.* The Image Biomarker Standardization Initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*. 2020;295(2):328-338.

doi: 10.1148/radiol.2020191145

27. Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative. *arXiv*. Preprint posted online 2016.

doi: 10.48550/arXiv.1612.07003

28. Whybra P, Zwanenburg A, Andrearczyk V, *et al.* The Image Biomarker Standardization Initiative: standardized convolutional filters for reproducible radiomics and enhanced clinical insights. *Radiology*. 2024;310(2):e231319.

doi: 10.1148/radiol.231319

29. Dixon WJ, Massey FJ Jr. Introduction to Statistical Analysis. 2nd ed. McGraw-Hill; 1957.

30. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321-357.

doi: 10.1613/jair.953

31. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Methodol*. 1996;58(1):267-288.

doi: 10.1111/j.2517-6161.1996.tb02080.x

32. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32.

doi: 10.1023/A:1010933404324

33. Wade C, Glynn K. Hands-On Gradient Boosting with XGBoost and scikit-learn: Perform Accessible Machine Learning and Extreme Gradient Boosting With Python. Packt Publishing; 2020.

34. Aitkin M, Foxall R. Statistical modelling of artificial neural networks using the multi-layer perceptron. *Stat Comput*. 2003;13(3):227-239.

doi: 10.1023/A:1024218716736

35. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach Learn*. 2002;46(1-3):389-422.

doi: 10.1023/A:1012487302797

36. Li Y, Luo Y. Performance-weighted-voting model: an ensemble machine learning method for cancer type classification using whole-exome sequencing mutation. *Quant Biol*. 2020;8(4):347-358.

doi: 10.1007/s40484-020-0226-1

37. Rainio O, Teuho J, Klén R. Evaluation metrics and statistical tests for machine learning. *Sci Rep*. 2024;14(1):6086.

doi: 10.1038/s41598-024-56706-x

38. Demler OV, Pencina MJ, D'Agostino RB Sr. Misuse of DeLong test to compare AUCs for nested models. *Stat Med*. 2012;31(23):2577-2587.

doi: 10.1002/sim.5328

39. Lundberg S, Lee SI. A unified approach to interpreting model predictions. *arXiv*. Preprint posted online 2017.

doi: 10.48550/arXiv.1705.07874

40. He H, Garcia EA. Learning from imbalanced data. *IEEE Trans Knowl Data Eng*. 2009;21(9):1263-1284.

doi: 10.1109/TKDE.2008.239

41. Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: Machine learning in Python. *arXiv*. Preprint posted online 2012.

doi: 10.48550/arXiv.1201.0490

42. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26(6):565-574.

doi: 10.1177/0272989X06295361

43. Moons KGM, Altman DG, Reitsma JB, *et al.* Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. 2015;162(1):W1-W73.

doi: 10.7326/M14-0698

44. Zheng X, Guo W, Wang Y, *et al.* Multi-omics to predict acute radiation esophagitis in patients with lung cancer treated with intensity-modulated radiation therapy. *Eur J Med Res*. 2023;28(1):126.

doi: 10.1186/s40001-023-01041-6

45. Wang K, Zhao J, Duan J, *et al.* Radiomic and dosimetric parameter-based nomogram predicts radiation esophagitis in patients with non-small cell lung cancer undergoing combined immunotherapy and radiotherapy. *Front Oncol*. 2024;14:1490348.

doi: 10.3389/fonc.2024.1490348

# Appendix

## Supplementary methods

### A1. Esophagus volume-of-interest

The VOIs represent the contours of non-cancerous esophageal tissue, defined by subtracting the PTV from the defined esophagus contours. There are various studies describing the radiobiology of the esophagus variably, suggesting it may behave as a serial organ, a near-parallel organ, or even as a fully parallel organ[30-32]. Previously, it has been shown in preclinical research that upper portions of the esophagus may represent a risk factor for the occurrence of acute radiation esophagitis compared to other segments[33-35]. In this study, we analyzed the esophagus using three region definitions: the whole esophagus, a 5-mm esophageal ring, and subdivision into four quadrants. Radiomics and dosiomics feature extraction was limited to the entire esophageal region, whereas the novel DVH analysis incorporated the whole esophagus, ring, and the quadrants. From the whole-esophagus contour, we constructed a 5-mm ring by offsetting the boundary by 2.5 mm inward and 2.5 mm outward (i.e., morphological erosion and dilation) on each axial-plane slice; then the slice-wise rings were stacked to form the 3D ring volume. For the quadrant partition, we computed the slice-wise centroid of the esophagus and drew two orthogonal lines (horizontal and vertical) through the centroid to divide the esophagus into four regions in the axial plane. Then, each quadrant was labeled counterclockwise on every slice using a consistent convention (e.g., starting from posterior-right to anterior-right). **Figure S1** illustrates an example of the novel esophageal geometric region derived in this work. The entire esophagus VOI, derived from the CT images, was used to generate different versions of filtered images. These filtered images were then used as inputs to the radiomics feature-extraction pipelines.

### A2. Machine learning algorithms

The six ML models for classification included, i.e., LASSO–GLM[30], RF[31], XGBoost[32], MLP-ANN[33], SVM with linear kernel, and SVM with RBF kernel[34].

- Generalized linear model with LASSO penalty is a generalized linear model with a LASSO penalty function and is considered a supervised learning algorithm for classification. It estimates the probability of a binary outcome using a sigmoid function and adds an L1 (LASSO) penalty to the coefficients[30]. The penalty shrinks many coefficients toward zero, thereby controlling overfitting, reducing the probability of correlated predictors, and producing a sparse and interpretable model. We used the "Lasso" method from the scikit-learn library.
- Random forest is an ensemble supervised learning method for classification and regression[31]. It builds decision trees on bootstrapped samples of the training data and, at each split, considers a random subset of features. Predictions are aggregated by majority vote. Increasing the number of trees typically reduces variance and stabilizes performance, thereby lowering the risk of overfitting relative to a single tree. We used the "RandomForestClassifier" method from the scikit-learn library.
- Extreme gradient boosting is an ensemble, supervised learning method based on a gradient boosting framework over decision trees built sequentially to fit the residuals (negative gradients) of a chosen loss function. It includes L1/L2 regularization, learning rate (shrinkage), and row/column subsampling to improve generalization and speed[32]. Predictions are the additive sum of all trees. We used the "XGBClassifier" method from the scikit-learn library.
- A multilayer perceptron neural network is a supervised learning method that uses a feed-forward, fully connected neural network for classification[33]. It consists of an input layer, one or more hidden layers, and an output layer, with each neuron connected to all neurons in the preceding layer. The neurons in an MLP-ANN typically use nonlinear activation functions, enabling the network to learn complex patterns, including nonlinear relationships, which can be beneficial for classification tasks. The parameters of the MLP-ANN are learned by backpropagation to minimize cross-entropy, with the output layer producing class probabilities via a sigmoid (binary) mapping. We used the "MLPClassifier" method from the scikit-learn library.
- Support vector machine is a supervised learning method that determines a decision boundary (hyperplane) separating classes while maximizing the margin to the nearest training points (support vectors)[34]. We used a soft-margin SVM with a regularization parameter to trade off margin size against classification errors. We implemented an SVM with a linear and RBF kernel to model class boundaries. We used the "SVC" method from the sklearn library.
- Ensemble techniques by weighted soft voting of algorithms[35] constitute a variable technique enabling assignment of different levels of importance to each classifier. This weighted voting is preferred over relying on a single classifier because it leverages the "wisdom of the crowd" to build a more accurate, robust, and generalizable model. It also

mitigates the weaknesses of any single model and capitalizes on their collective strengths by combining the predictions of multiple, diverse classifiers and assigning weights based on their individual performance. The formula for the final ensemble probability for each class j, denoted as $P_j$, is calculated as:

$$P_j = \frac{\sum_{(i=1)}^{n} W_i \times P_{ij}}{\sum_{(i=1)}^{n} W_i} \tag{A1}$$

where $n$ is the number of individual classifiers, $W_j$ is the weight assigned to classifier $i$ based on the classifier performance, i.e., AUCs. $P_{ij}$ is the probability that classifier $i$ predicted for class j, $\sum_{i=1}^{n} W_i \times P_{ij}$ is the weighted sum of probabilities for class j across all classifiers, and $\sum_{i=1}^{n} W_i$ is the sum of all weights.