International Journal of Engineering Trends and Technology
ISSN: 2231-5381/ https://doi.org/10.14445/22315381/IJETT-V74I8P104

Volume 74 Issue 8, 50-67, August 2026

Original Article

Hybrid Resampling and Ensemble Learning for Predicting Radiation-Induced Pneumonitis in Head and Neck Cancer

Abhijit Nath[1], Sheikh Wakie Masood[1], Shahin Ara Begum[1*], Ravi Kannan[2]

[1]*Department of Computer Science, Assam University, Silchar, Assam, India.*
[2]*Cachar Cancer Hospital & Research Centre, Silchar, Assam, India.*

[1*]*Corresponding Author : shahin.ara.begum@aus.ac.in*

Received: 21 January 2026 Revised: 18 June 2026 Accepted: 24 June 2026 Published: 29 August 2026

Abstract - *Radiation-Induced Pneumonitis (RP) is a clinically important but difficult-to-predict toxicity in Head and Neck Cancer (HNC), particularly when only routinely collected hospital variables are available. This study developed a recall-oriented machine learning framework for RP prediction using clinical and treatment-related data from 2,482 HNC patients treated with radiotherapy between 2018 and 2024. Twenty-five baseline variables and four engineered interaction terms were analysed using a stratified 70/30 train–test split. Three phase-wise experiments were performed: Phase I on the original imbalanced training data, Phase II with SMOTETomek-based class balancing, and Phase III as an extension of Phase II with Gaussian-noise data augmentation. XGBoost, LightGBM, CatBoost, and SVM were evaluated as individual classifiers, while the proposed stacking classifier combined XGBoost, LightGBM, and CatBoost with Logistic Regression as the meta-learner. Hyperparameters were optimized by Bayesian search with stratified cross-validation, and thresholds were selected from training-only out-of-fold probabilities using F2-score as the primary criterion. In Phase I, the stacking classifier achieved the highest F2-score 0.57 and recall 0.86. In Phase II, the stacking classifier achieved the highest F2-score 0.58 and ROC-AUC 0.67. In Phase III, it again performed best overall, with a precision 0.24, recall 0.92, F2-score 0.59, and ROC-AUC 0.65. These findings support stacking-based prediction as a screening-oriented aid for identifying high-risk RP cases from routinely available hospital records.*

Keywords - *Head and Neck Cancer, Radiotherapy, Radiation-Induced Pneumonitis, Data augmentation, Ensemble learning.*

1. Introduction

Head and Neck Cancer (HNC) has been a significant issue of public-health concern, especially in regions with high tobacco and betel nut exposure, where many patients are at advanced stages and require intensive radiotherapy-based treatment [1, 2]. Even with the current radiotherapy methods, which have enhanced locoregional control, continuity of care and recovery is still affected by treatment-related toxicities. One of such complications is Radiation-Induced Pneumonitis (RP). Its early symptoms, such as cough, dyspnoea, and low-grade fever, are often nonspecific, which makes timely recognition difficult without structured surveillance [3, 4].

The majority of predictive investigations of RP have been produced in thoracic malignancies and often utilize dose-volume, imaging, or radiomics-based characteristics. However, these techniques have demonstrated promising performance in lung and oesophageal cohorts, but it is still unclear how such models can be directly applied to HNC treatment practice due to the differences in geometry of treatment fields, exposure to, incidental lung dose, comorbidity patterns, and supportive-care pathways [5-7]. Another issue is that in a common clinical database, the RP-positive cases are usually less common than the treatment courses with no complications. In this kind of class imbalance, models might seem fine in general terms but may fail to reveal the minority positive cases, which are most important in clinical aspects [8, 9]. Consequently, there is still a need to have more effective prediction strategies that have the potential to facilitate the prior detection of high-risk HNC patients based on the routinely available structured clinical and treatment data.

Recent studies have also pointed out the potential of machine learning as a tool to predict toxicity, especially when there is an integration of multiple sources of data. Thoracic oncology literature has proposed that multimodal approaches incorporating clinical, dosimetric, and imaging variables could be more discriminative and systematic reviews have shown that hybrid models can be more effective than single-source models in certain environments [10]. Nevertheless,

there are still some critical gaps in the HNC context. To begin with, there is a relative lack of specific predictive evidence on RP in HNC when compared to thoracic cancer patient groups. Second, the classification-imbalance treatment is not very consistent and is not addressed adequately. Third, the relative effect of sequential correction of imbalance and balanced data augmentation under a leakage-safe evaluation design has not been properly studied on routinely collected data. Moreover, the idea of ensemble methods is often discussed as a stable alternative, its practical contribution in this setting remains uncertain and may depend on the clinical objective and the operating threshold rather than on a single metric alone [9, 11].

The present study addresses these gaps through a three-Phase machine-learning framework of RP prediction in patients with HNC undergoing radiotherapy at a single facility. The analysis involves 25 regularly measured baseline variables along with four prespecified second-order interaction terms, which sum to 29 modelling features. The framework quantifies the performance at three phase-wise experiments, which were performed: Phase I on the original imbalanced training data, Phase II with SMOTETomek-based class balancing, and Phase III as an extension of Phase II with Gaussian-noise augmentation. XGBoost, LightGBM, CatBoost, and SVM were evaluated as individual classifiers, while the proposed stacking classifier combined XGBoost, LightGBM, and CatBoost with Logistic Regression as the meta-learner to determine whether ensemble learning would bring a more balanced profile of prediction. Bayesian search was used to optimize hyperparameters during cross-validation and performance was evaluated by considering recall-sensitive measures, F2-score, as the primary measure, ROC-AUC, precision, and recall as secondary measures [12-17].

The study was designed not only to compare individual classifiers, but also to examine the effect of Phase-wise imbalance handling and balanced data augmentation on minority-class detection under a leakage-safe workflow. In this respect, the analysis is a structured comparison of imbalanced learning at baseline, hybrid resampling, referring to SMOTETomek-based balancing in Phase II and its extension with Gaussian-noise-based balanced data augmentation in Phase III, and fixed balanced augmentation, where a held-out test set was used as a final evaluation. The design will enable a more objective evaluation of whether performance improvements are due to actually improved learning among the minority class as opposed to optimistic evaluation processes.

The rest of the paper is structured in the following way: Section 2 examines related research in the RP prediction and toxicity modelling. In Section 3, the study background is provided. Section 4 presents Dataset and features. Section 5 provides the methodology. The Phase-wise results and discussion are found in Section 6. Section 7 is a summary of the clinical and methodological implications, limitations and concluding findings.

2. Related Work

Radiation-Induced Pneumonitis (RP) has been a significant concern in radiotherapy studies since it offers a chance of identifying high-risk patients that may support treatment planning, early monitoring and institute timely intervention [5, 6]. The most published work in this field has been carried out in thoracic malignancies, particularly lung and oesophageal cancer, where lung-dose exposure is more direct and imaging and dosimetric data are usually available [5, 6]. Over time, the field has gradually moved beyond traditional statistical methods toward machine-learning, radiomics, and deep-learning models in an effort to improve predictive performance and capture more complex risk patterns [6, 10].

2.1. Prediction of RP in thoracic cancers

A number of studies have indicated that machine-learning models could be useful in the prediction of RP in thoracic cancer cohorts. Ye et al., constructed a machine-learning model based on equivalent uniform dose descriptors in lung cancer receiving volumetric modulated arc therapy and demonstrated that the selected dosimetric features might be useful to distinguish between the risk of RP [5]. Elsewhere, Wang et al., used the Support Vector Machine modelling with four-dimensional computed tomography-based lung function data and showed better performance than the traditional dose-volume histogram-based methods, suggesting that functional features could be used to enhance the prediction of RP [7].

More recent work has taken this line in terms of broader multimodal approaches. A hybrid deep-learning structure in oesophageal cancer was suggested by Sheng et al., that integrated image and clinical data, and demonstrated a high predictive power with regard to symptomatic RP [22]. Likewise, Tang et al., developed a dual-model transfer-learning model that combined computed tomography imaging data with clinical and dosimetric data that also favored the use of combined heterogeneous data sources instead of using one domain of features [18]. Suggestions indicate that the prediction of thoracic RP has been gradually shifting towards more combined modelling approaches.

2.2. Deep Learning, Radiomics, and Multimodal Modelling

Recent literature has also shown interest in increasing research in the area of deep learning and image-based prediction of RP. Choi et al., proposed a deep-learning model to predict RP following stereotactic body radiotherapy for pulmonary nodules and demonstrated that image-based approaches could be useful to give early risk information in thoracic practice [19]. In another study, Li et al., combined three-dimensional deep image features to dose-volume information in locally advanced non-small cell lung cancer

and found them useful in the two institutional cohorts, which has demonstrated the usefulness of multimodal modelling [20]. Lee et al., also demonstrated that clinical variables in combination with pretreatment chest computed tomography could be used to predict RP, which once again supports the potential of hybrid deep-learning designs [21].

Recent evidence synthesis also represents the greater trend in the literature. Chen et al., have reviewed the literature on machine-learning-based RP prediction and observed that the results varied with the cohort, features, and validation designs, although the general trend was promising [6].

In another systematic review and meta-analysis, Hegazy et al., suggested that radiomics and combined models had the potential to enhance discrimination, but that the extent of the advantage depends strongly on study design, data quality, and validation strategy [10]. Such reviews are significant in that they demonstrate that there are promising results, but these do not exist equally, and they should be viewed in the framework of model design and cohort characteristics.

2.3. Clinical Setting of RP and the Necessity of Prediction Tools

In addition to pure modelling studies, a number of authors have reported the clinical relevance of RP in modern oncology practice. Hanania et al., and Choi et al., have shown that RP remains a relevant toxicity following thoracic radiotherapy and that this risk depends on both treatment-related and patient-related variables, among them, the volumetric irradiation, dose exposure, smoking history, and concurrent therapy [3, 4].

More recently, real-world studies have indicated that the prevalence of pneumonitis can remain high, or even higher, in certain contemporary combined-treatment regimens, including chemoradiotherapy plus durvalumab, which again justifies the use of a reliable risk-prediction instrument [23]. Taken together, these studies show that RP is not only a modelling problem, but it is also still a clinical management problem.

2.4. Research Gaps

Despite the fact that the available literature has proven the potential of machine learning, radiomics, and deep-learning models in predicting RP, some research gaps are significant. First, the majority of studies have been conducted in populations with thoracic cancer, and their results cannot be extrapolated directly to the Head and Neck Cancer (HNC), in which the treatment geometry, incidental exposure to lungs, clinical pathways, and patient characteristics vary significantly. Second, a wide range of published models rely heavily on imaging and dosimetric characteristics, but more frequent structured clinical and treatment variables have been less investigated in this respect. Third, even where it is clearly described that class imbalance is being dealt with explicitly, there is no clear description of how the clinically important minority subgroup of RP-positive cases is dealt with. Lastly, there have not been many studies on Phase-wise performance variations on a leakage-safe workflow that isolates baseline imbalanced learning, class balancing, and balanced data augmentation in a more transparent manner.

It is on these research gaps that the current study is founded. Rather than using special imaging pipelines, the current study is based on routinely measured structured variables of an HNC radiotherapy cohort. It also evaluates a Phase-wise framework of evaluation of original imbalanced data training, SMOTETomek-based class balancing, and fixed balanced data augmentation.

In addition, the study gives greater emphasis to the F2-score so that minority-class detection remains central to model assessment. Thus, the work should add value not with its assertion of a single overall better classifier, but with a comparative assessment of RP prediction techniques, *viz*., ensemble learning in a methodologically clear and clinically-focused comparison of treatment settings in an under-researched HNC scenario.

3. Study Background

3.1. Patient Data

Cachar Cancer Hospital and Research Centre (CCHRC), Silchar, Assam, India, gave their approval to the research through its Institutional Ethics Committee. This retrospective study included 2,482 patients with a history of Head and Neck Cancer (HNC) who received radiotherapy in the institution and confirmed the diagnosis by the use of histopathology.

The qualifying criteria included a final histopathological diagnosis of HNC, receipt of radiotherapy at the centre, and full access to clinical and treatment data, including the data of Computed Tomography (CT) simulations and radiotherapy planning data.

The exclusion criteria included incomplete records, failure to follow the radiotherapy treatment regimen as prescribed, and shortened palliative treatment regimen (e.g. a 10-fraction regimen).

3.2. Patient Demographics

A total of 2,482 patients with HNC were included in the study, out of which 1,855 (75%) were males and 627 (25%) were females. The average age of the patients was 55 (range: 20-80 years).

3.3. Technique of Radiation Therapy

In order to obtain a conformal dose distribution, intensity-modulated techniques (e.g. IMRT/VMAT) were used in the linac due to the specific tumour site and stage based on the current clinical and radiological practices. Image guidance

and regular setup checks were also introduced into the therapeutic workflow.

3.3.1. Bhabhatron Cobalt-60

One feasible option for teletherapy is Bhabhatron, which uses Cobalt 60 (around 1.25MeV) for external beam teletherapy. Immobilization of patients and proper planning is required to provide a uniform target coverage and normal-tissue sparing [24].

3.3.2. Linear Accelerator or Linac Varian

High-energy photons/electrons aiming at conformal therapy were produced from the Linac. IMRT/VMAT plans using multileaf collimators were used to modulate the fluence and increase the conformity. The workflow consisted of immobilization, CT-based planning, and protocol-based image guidance.

3.3.3. Dose Prescription and Overall Treatment Time

The total dose of 60-70Gy administered in 1.82.0Gy fractions, 5 days per week in a course of 67 weeks, was adjusted to tumour site, stage, and clinical condition. The priorities of treatment were determined based on the target coverage and protection of organs in danger.

3.3.4. Radiation-Induced Toxicities

Skin reactions and mucositis were frequent. RP: Accidental dose to the lungs or coupled therapies occurred. Patients with RP experienced the following symptoms: cough, dyspnoea and low-grade fever, necessitating careful monitoring and dose-aware planning to mitigate risk [4].

3.4. Diagnosis of Radiation Pneumonitis

The diagnostic method of radiation pneumonitis is a combination of clinical analysis and diagnostic imaging. Primary diagnosis is based on chest radiographs, but is perfected through computed tomography, which offers a high-quality three-dimensional imaging of pulmonary alterations.

Early detection and rapid intervention are critical to improving outcomes for patients who develop radiation pneumonitis. By combining thorough clinical assessments with modern imaging technologies, medical practitioners can properly diagnose and treat this condition and therefore improve the care of their patients during and after radiotherapeutic treatment [4].

4. Dataset and Features

4.1. Data Source and Collection

A retrospective dataset from CCHRC was used in this study. Records spanning 2018–2024 were extracted from the hospital information system; older non-digitized entries were abstracted from departmental paper files. All data were de-identified before analysis. The end-to-end collection workflow is shown in Figure 1.

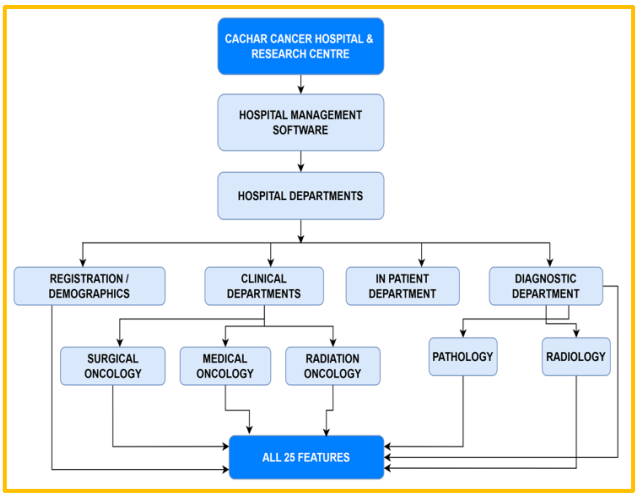


Fig. 1 Overview of the retrospective data collection

4.2. Variables

Variables were selected a priori in close discussion with the physician team based on clinical relevance and routine availability in practice, and were grouped as follows:

- Demographic: age (years), sex.
- Baseline clinical: performance status, primary site, stage (TNM I–IV), baseline weight at radiotherapy start, comorbidities (diabetes, tuberculosis, heart disease, asthma), behavioural history (smoking, alcohol), prior COVID-19 infection.
- Treatment-related: total prescribed dose (Gy), treatment machine (Bhabhatron Cobalt-60 or Linac), technique (conventional/IMRT/VMAT as applicable), admission status (inpatient/outpatient), duration of hospital stay (days), swallow therapy, Ryle's tube, tracheostomy, treatment intent (curative/palliative/haemostatic), and neoadjuvant chemotherapy.
- On-treatment toxicities: mucositis grade (routine clinical grading).

The outcome was RP, defined as a binary endpoint (Absent/Present) based on clinical assessment supported by imaging where indicated. Typical symptoms include cough, dyspnoea, and low-grade fever [4].

4.3. Feature Set for Modelling

Using routine data, 25 baseline features were identified. To capture clinically plausible non-additive relationships, four second-order interaction terms were pre-specified: age × site, age × duration of stay, baseline weight × age, and baseline weight × duration of stay.

The inclusion of these interaction effects was guided a priori by clinical plausibility and by the consistent availability of the required variables across the study sample to yield a comprehensive registry of 29 explanatory variables.

4.4. Data Quality and Handling

Observations without outcome annotations were excluded from further analyses. Missing explanatory variables were conservatively imputed by replacing missing values with the median of the observed numerical distribution or the modal category for categorical variables. Following ordinal encoding of categorical variables, continuous variables were rescaled to the [0,1] range using min–max normalisation.

Crucially, preprocessing operations such as imputation, encoding, and scaling were fitted within each cross-validation training fold and then applied to the corresponding validation fold and the held-out test set in order to avoid data leakage.

4.5. Descriptive Summary

Table 1. gives a representative picture of the feature matrix and descriptive statistics for the dataset, including details about columns like Feature name, Data type, Values, Count and Percentage.

The variables listed include Age bands, Sex, Performance status, Site, Stage, Swallow therapy, Hospital admission, Duration-of-stay bands, Ryle's tube use, Tracheostomy, Mucositis grade, History of smoking, History of TB/heart disease/asthma and the outcome of RP. Comprehensive distributions for all the variables are available on request in the supplementary appendix.

Table 1. Dataset features with statistics

Feature Name	Date Type	Values	Count	Percentage
Age	Numerical	7 - 92 Years	Mean: 59.3, Median: 59 Years	SD: 12.6 Years
Gender	Categorical	Male	1855	74.71
		Female	627	25.25
PS	Categorical	I	2302	92.71
		II	132	5.32
		III	48	1.93
Site	Categorical	Accessory Sinuses	2	0.08
		Eye and Adnexa	28	1.13
		Hypopharynx	561	22.61
		Larynx	257	10.35
		Lip	78	3.14
		Nasal Cavity and Middle Ear	9	0.36
		Nasopharynx	82	3.30
		Neck	15	0.60
		Oral Cavity	1015	40.90
		Oropharynx	305	12.29
		Thyroid	14	0.56
		Tonsil	116	4.67
Stage	Categorical	1	22	0.89
		2	250	10.07
		3	919	35.12
		4	1291	51.99
Swallow therapy	Categorical	Present	585	23.56
		Absent	1897	76.40
Hospital admission	Categorical	Present	1933	77.85
		Absent	549	22.11
Duration of stay	Categorical	0	1120	45.11
		01 - 30 (Days)	1210	48.73
		30 - 75 (Days)	152	6.12

Ryles tube	Categorical	Present	1399	56.34
		Absent	1083	43.62
Tracheostomy	Categorical	Present	343	13.81
		Absent	2139	86.15
Mucositis grade	Categorical	0	313	12.61
		I	94	3.79
		II	811	32.66
		III	1170	47.12
		IV	94	3.79
History of smoking	Categorical	Present	1201	48.37
		Absent	1282	51.63
History of TB	Categorical	Present	81	3.26
		Absent	2401	96.70
History of heart disease	Categorical	Present	45	1.81
		Absent	2437	98.15
History of asthma	Categorical	Present	53	2.13
		Absent	2431	97.91
Pneumonia	Categorical	Present	567	22.88
		Absent	1915	77.08

5. Methodology

The process of the methodology follows a well-defined pipeline, starting with the preprocessing of the data, followed by the feature engineering, and finally, the process of dealing with the issue of class imbalance through the study of balanced augmentation. This is followed by stratified splitting, model development, hyperparameter tuning, and a thorough evaluation, which will involve threshold setting. Subsequently, systematic checks for overfitting is performed, and finally, the feature importance is checked.

Any operation that has the potential to cause data leakage, such as imputation, encoding, scaling, resampling, model fitting, and threshold selection, is carefully confined to its specific Phase, thus maintaining data integrity in the validation process and in compliance with recognized best practices in predictive analytics.

Figure 2 shows the Phase-wise prediction pipeline for RP, which is leakage-safe, including data split, module preprocessing, handling imbalanced data, model construction, thresholding, and results.

5.1. Data Preprocessing

Records that did not have a documented outcome, labels were filtered out, especially when the dichotomous endpoint of pneumonia presence/absence was not documented. The other missing values in predictor variables were handled using median imputation for numerical variables and mode imputation for categorical ones. Subsequently, categorical variables were min–max normalized to the range [0, 1]. The avoidance of winsorization was also based on the fact that models based on trees, like boosted trees, tend to be relatively resistant to monotonic rescaling, whereas support vector machines enjoy the advantage of a homogeneous scale space after imputation. Each preprocessing object, such as imputers, encoders, and scalers, was fitted separately within each cross-validation training fold and applied to the matching validation and held-out test set. This preprocessing scheme was maintained the same in all the Phases of the experiment, whereas the imbalance-handling strategy was Phase-adaptive.

5.2. Feature Engineering

The modelling set consisted of 25 variables that are routinely collected, and they include demographic variables, baseline disease status, variables related to treatment, and on-treatment adverse events, as explained in Section 4. Four second-order interaction terms were specified a priori to capture potentially important non-additive relationships among clinically relevant variables. The interaction terms that were retained in the last modelling pipeline were: age x site, age x duration of stay, baseline weight x age, and baseline weight x duration of stay, which yielded a total of twenty-nine features. Feature elimination was not performed explicitly as collinearity screening. This choice has been taken due to the

relative robustness of the tree-based learners used in the study to redundant predictors, but the Support Vector Machine is regulated through cross-validation, with fold-wise preprocessing, scaling, and hyperparameter optimization. The feature engineering was thus added to the modelling pipeline as opposed to being a data-driven preselection process.

Figure 3 shows gain-based variable-importance diagnostics of XGBoost and LightGBM. These profiles of importance were only to be used in an interpretative manner, but not to determine the inclusion of features, the choice of threshold, or to rank models. Section 6.7 examines the extent to which these learned importance patterns align with clinical plausibility.

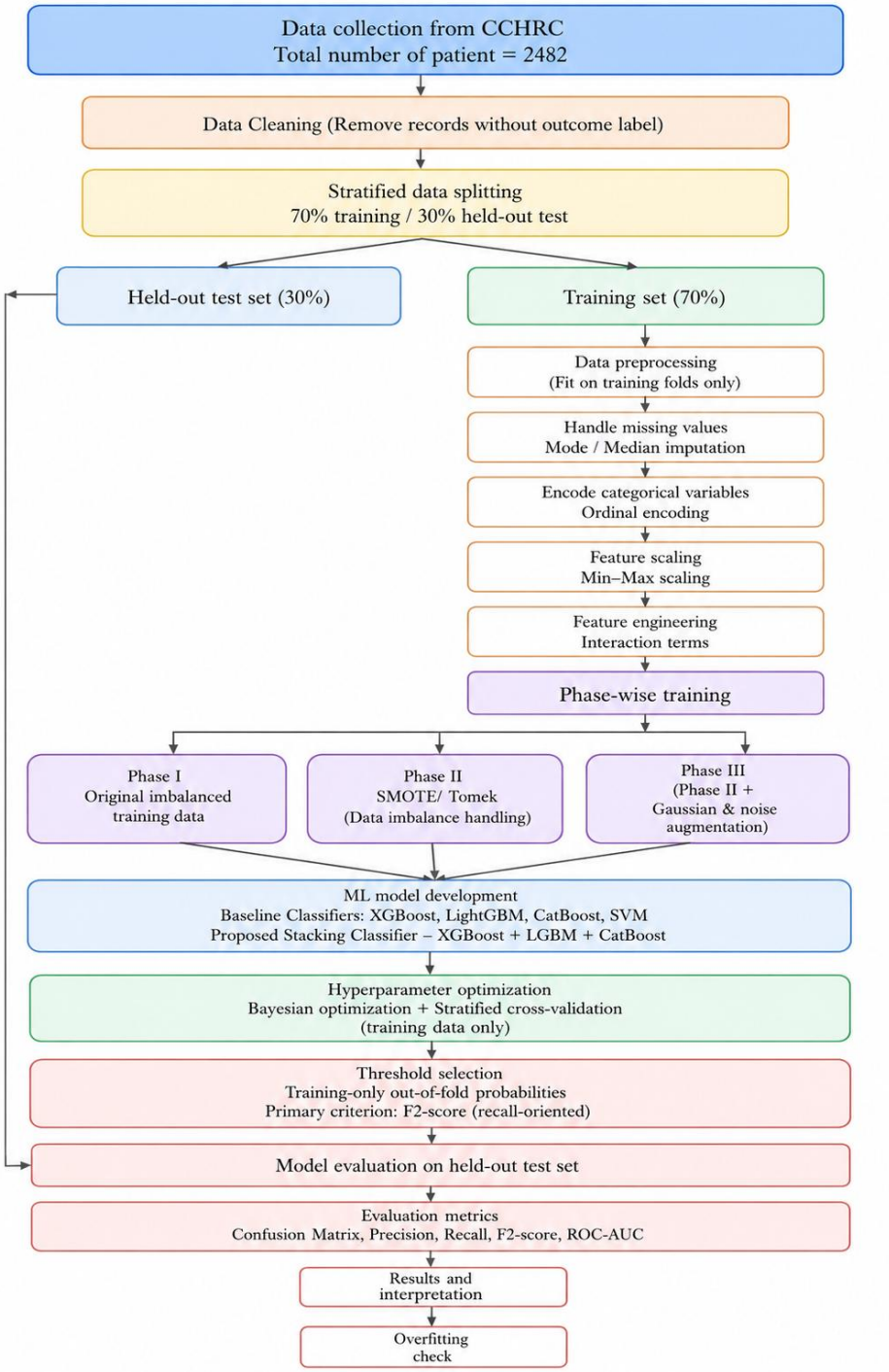


Fig. 2 Phase-wise prediction pipeline of RP

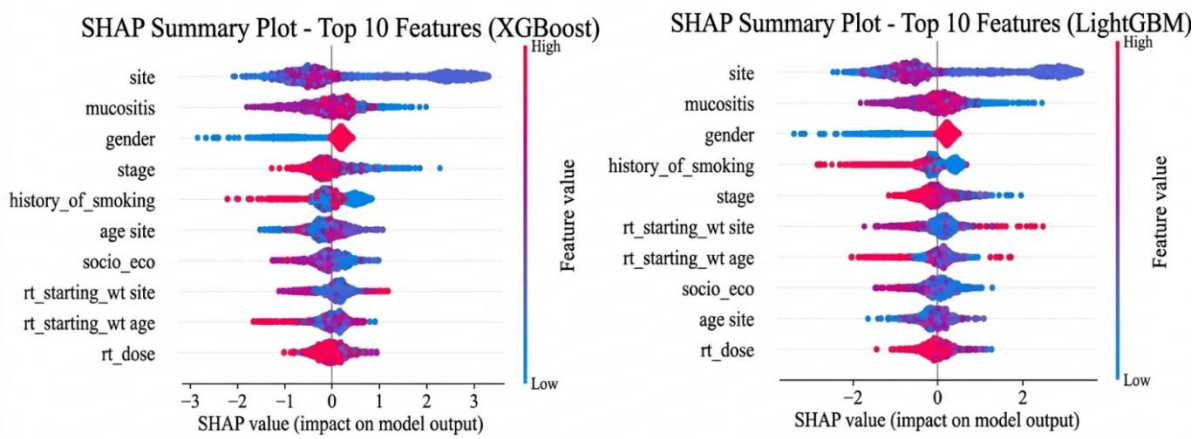


Fig. 3 Top 10 feature-importance profiles obtained from the baseline tree-based models for RP prediction

5.3. Phase-Wise Class Imbalance Handling and Augmentation

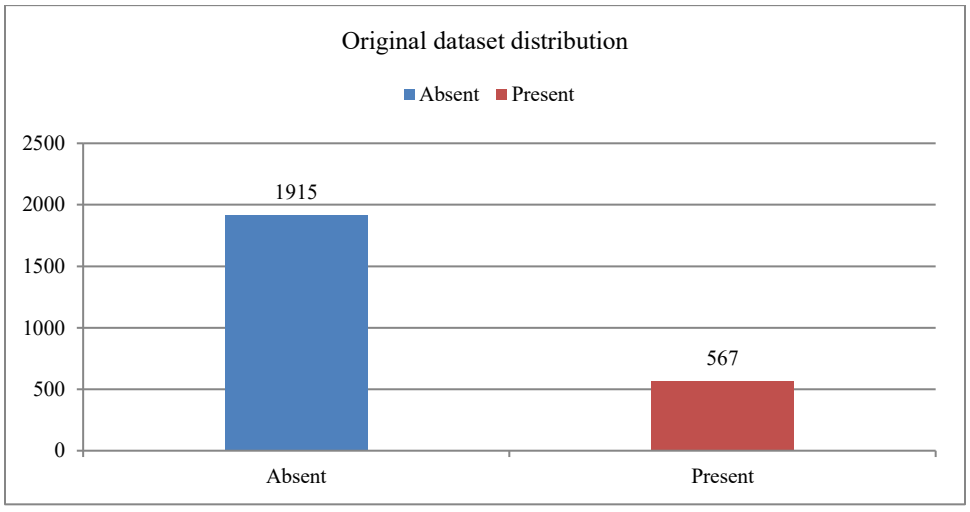
Since the number of pneumonia-positive cases was lower than the pneumonia-negative cases, direct model training on the original data might also bias the learning process towards the majority class and reduce the model's sensitivity to the minority class. To address this, phase-specific imbalance handling was applied only to the training data, and the held-out test set was untouched.

The SMOTETomek was the main balancing approach, where SMOTE created synthetic minority-class samples, and Tomek links eliminated borderline majority-minority pairs to enhance the separation of classes [22]. In Phase II, SMOTETomek alone was used as the imbalance-handling strategy. Phase III pipeline was kept and extended with Gaussian-noise-based data augmentations following SMOTETomek balancing to further augment training diversity and assess whether further improvements in recall-oriented performance were possible. Data augmentation was used in the last Phase III setup to achieve a fixed training size of 5,000 samples.

Figure 4(a)-4(d) shows the class distributions in the original data, stratified training split, post-SMOTETomek training set and final augmented training set. The impact of these phase-wise imbalance-handling and data augmentation strategies is analysed comparatively in Section 8.

5.4. Dataset Splitting, Leakage Control, and Cross-Validation

Stratified sampling was used to split the dataset into 70 percent training data and 30 percent held-out test data in such a way that it maintained the marginal distribution of the dependent variable between the split. Within the training set, stratified cross-validation was used for model selection and hyperparameter tuning. All preprocessing, resampling, and data augmentation processes were all restricted to the training section of every fold. The held-out test set was excluded from preprocessing, resampling, data augmentation, model selection, hyperparameter tuning, and threshold selection, and was used only once for final performance evaluation. The reason behind this design is to prevent the leakage and comparatively position the three experimental Phases detailed in Section 8.



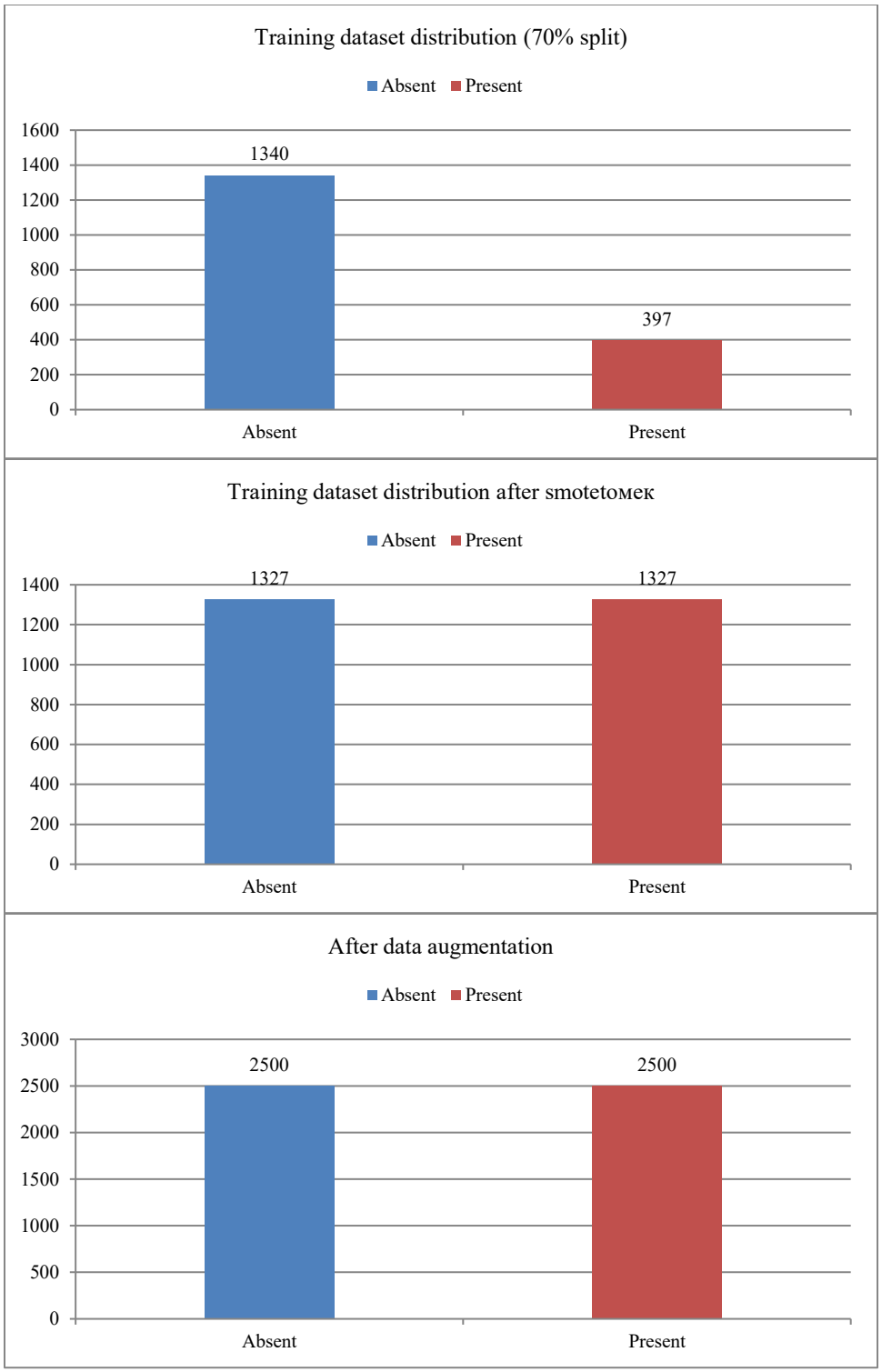


Fig. 4 (a) Class distribution in original dataset, (b) Stratified training split, (c) Post-SMOTETomek training set, and (d) After data augmentation.

5.5. Model Development and Stacking Classification Design

Four supervised classifiers were selected and evaluated as individual models: Extreme Gradient Boosting (XGBoost) [12], Light Gradient Boosting Machine (LightGBM) [13], CatBoost [15], and a Support Vector Machine (SVM) [14]. XGBoost and LightGBM build decision trees sequentially, with each tree correcting the residual errors of its predecessor. CatBoost follows the same principle but handles heterogeneous tabular data more reliably through symmetric tree growth and native categorical feature processing - a

characteristic that suits structured clinical datasets well. SVM was included as a non-tree-based comparator, providing a margin-based probabilistic classifier whose error profile differs structurally from the boosting models.

Hyperparameter optimisation was performed using Bayesian optimisation within cross-validation folds [12, 13]. For the tree-based classifiers, the search space covered learning rate, number of estimators, maximum tree depth, subsampling ratios, and regularisation terms [12, 13, 15]. For SVM, kernel type, regularisation parameter C, and kernel coefficient gamma were tuned within the same framework [14].

The model development pipeline followed a fixed five-step sequence within each training fold: (1) predictor preprocessing, (2) phase-specific class imbalance management and data augmentation, (3) hyperparameter optimisation of individual classifiers, (4) generation of out-of-fold probability estimates from the base learners, and (5) training of the meta-learner on these probability estimates. Steps 1 and 2 were applied exclusively within training folds at every stage. The held-out test set remained completely separate throughout, ensuring that no test data influenced model training or calibration — a principle central to the leakage-safe design of this study. A stacking ensemble was developed as the proposed primary model. XGBoost, LightGBM, and CatBoost were selected as base learners on the basis of their stable and mutually reinforcing predictive behaviour across the phase-wise evaluation. Their out-of-fold probability estimates were passed to a regularised Logistic Regression meta-learner [11], which determined the relative contribution of each base model to the final risk score. Logistic Regression was selected as the meta-learner for its well-calibrated probability outputs under regularisation and the interpretability of its coefficient vector. Prior ensemble learning research supports the rationale that combining classifiers with structurally different error profiles yields more consistent predictions than any single model in isolation [11].

SVM was retained as a standalone comparator given its higher recall tendency in certain evaluation stages. It was excluded from the final stacking ensemble, however, as its performance proved inconsistent across phases - reflected in its variable F2-score results - and it did not contribute a stable signal alongside the tree-based learners. All three tree-based classifiers completed training within acceptable time frames under cross-validation on the structured tabular dataset.

The coefficient vector of the Logistic Regression meta-learner is presented in Figure 5, and the stacking architecture is illustrated in Figure 6.

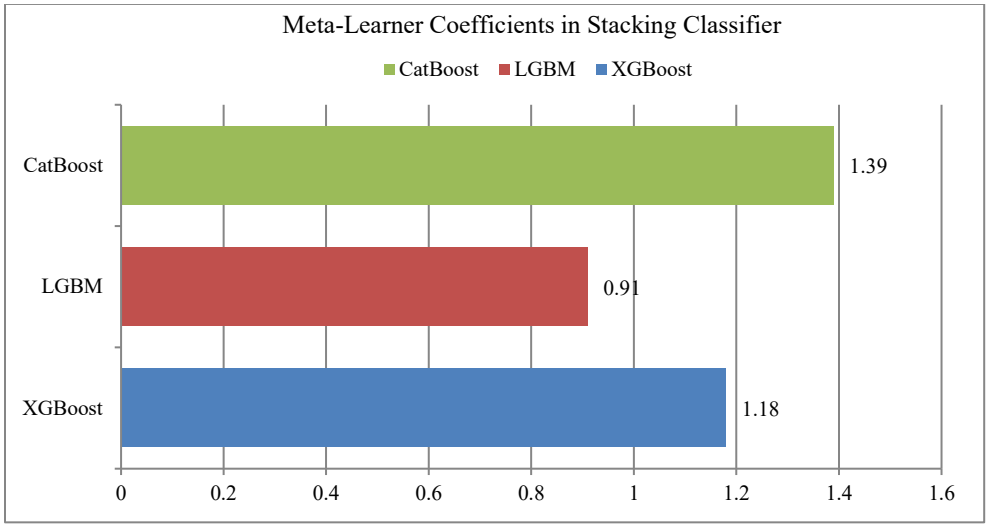


Fig. 5 Phase II logistic regression meta-learner coefficients

5.6. Hyperparameter Optimization

Bayesian optimisation was used for hyperparameter tuning, based on the cross-validated performance scores [15]. In the case of XGBoost and LightGBM, the search spaces were the number of trees, learning rate, maximum depth of trees or leaf complexities, row subsampling fraction, column subsampling fraction, and L1/L2 regularization parameters. The search space also encompassed tree depth, learning rate, boosting iterations, and regularization-related parameters for CatBoost. In the case of the Support Vector Machine, the regularization parameter C, type of kernel (linear and radial basis function) and the scaling parameter of the radial basis function kernel, γ, were part of the search space.

Hyperparameter optimization was performed on the training data through stratified cross-validation, with each model having a fixed search budget. The final model configurations are summarized in table 2. Within the stacking framework, the Logistic Regression meta-learner was regularized, and tuning considered both the strength of

regularization and ensemble-specific settings such as passthrough and class weighting wherever applicable. In Phase III, data augmentation-related parameters, particularly the Gaussian-noise level, were also optimized within the final training pipeline while preserving the fixed balanced data augmentation design.

Table 2. Selected key hyperparameter settings of the classifiers

Phase	Model	Key selected hyperparameters
Phase I	XGBoost	lr=0.1401, max_depth=5, n_estimators=204, min_child_weight=8
	LightGBM	lr=0.1451, max_depth=4, num_leaves=54, n_estimators=300
	CatBoost	lr=0.2000, depth=5, iterations=350, l2_leaf_reg=0.0862
	SVM	kernel=rbf, C=0.0283, gamma=0.3810
	Stacking Classifier	meta-C=0.0437, class_weight=balanced, passthrough=True
Phase II	XGBoost	lr=0.1213, max_depth=8, n_estimators=317, min_child_weight=3
	LightGBM	lr=0.2000, max_depth=5, num_leaves=15, n_estimators=350
	CatBoost	lr=0.2000, depth=7, iterations=260, l2_leaf_reg=1.885
	SVM	kernel=rbf, C=4.806, gamma=0.000487
	Stacking Classifier	meta-C=0.0117, class_weight=None, passthrough=True
Phase III	XGBoost	noise=0.005, lr=0.1415, max_depth=4, n_estimators=203, min_child_weight=7
	LightGBM	noise=0.005, lr=0.1415, max_depth=5, num_leaves=38, n_estimators=203
	CatBoost	noise=0.008, lr=0.0573, depth=6, iterations=220, l2_leaf_reg=7.327
	SVM	noise=0.005, kernel=rbf, C=0.0352, gamma=0.0163
	Stacking Classifier	noise=0.002, meta-C=2.502, class_weight=None, passthrough=True

5.7. Evaluation Metrics and Operating Threshold

The evaluation framework in this study was designed to prioritise the early identification of pneumonia-positive patients while maintaining overall discriminatory performance. The F2-score was selected as the primary evaluation metric because it places greater weight on recall than on precision [16]. In this clinical context, failing to identify a true positive case - a patient who will develop radiation-induced pneumonitis - was considered a more serious error than generating an additional false positive. The F2-score directly reflects this clinical priority by penalising missed detections more heavily, making it a clinically justified and appropriate metric for both model evaluation and operating threshold selection in this study.

The area under the Receiver Operating Characteristic Curve (ROC-AUC) was used as the secondary evaluation metric [17]. Unlike threshold-dependent measures, ROC-AUC assesses discrimination performance across all possible classification thresholds, summarising the model's overall capacity to assign higher predicted risk scores to pneumonia-positive cases than to pneumonia-negative cases, which is particularly important given the class imbalance present in this dataset. Precision and recall were additionally reported at the

selected operating threshold to characterise the practical screening trade-off at the point of clinical decision.

The operating threshold for each model was determined using training-only cross-validated probability outputs. The threshold that maximised the F2-score was identified from out-of-fold predicted probabilities generated exclusively within the training pipeline. Keeping threshold selection within the training process removes any risk of over-optimistic results and ensures a consistent decision rule across all models and phases. This selected threshold was then applied without modification to the held-out test set. The same thresholding procedure was applied uniformly across Phase I, Phase II, and Phase III to maintain methodological comparability throughout the phase-wise evaluation.

All threshold-dependent measures, precision, recall, F1-score, and F2-score were computed with respect to the pneumonia-positive class, referred to as the positive class (class 1) throughout this study. ROC-AUC was similarly computed based on the predicted probability assigned to the positive class. Given the class-imbalanced nature of the dataset, these metrics were interpreted relative to the phase-specific resampling and augmentation conditions applied during training, as described in Section 5.3.

5.8. Overfitting Analysis

Generalization performance was evaluated by comparing the cross-validated training performance with the held-out test-set performance, with a particular focus on the F2-score and ROC-AUC. In addition, cross-validation fold variability was also examined to determine the models whose learning behaviour was not stable. The purpose of this analysis was aimed not to maximize training performance, but to determine whether the performance gains were consistent when assessed on data that was not utilized to fit the model or to select the operating threshold.

When substantial discrepancies were observed between training-set and held-out test performance, the complexity of the model was viewed cautiously when optimizing hyperparameters by regulating the tree depth parameter, learning rate parameter, regularization strength parameter, and SVM regularization parameter. The data augmentation design in Phase III was also monitored carefully to prevent overly aggressive positive-leaning operating behavior. The main train/test performance summaries are illustrated in Figure 8 and are discussed along with the confusion-matrix analysis in Section 8.

5.9. Feature-Importance and Model Contribution Analysis

Feature importance derived from XGBoost and LightGBM was analyzed to identify the most influential predictors and to assess whether the most important variables were still clinically relevant in the experimental stages. In addition, a qualitative analysis of the fold-wise patterns of importance was also conducted to assess the consistency of such signals, as shown in Figure 3.

In the stacking ensemble, the base learners contribution was summarized by the coefficients of the Meta-learner as illustrated in Figure 5. These analyses were not applied in post hoc feature selection, adjusting a threshold, or ranking the model behavior, but only to interpret and discuss the model behavior. Section 6.7 explains how well the learned patterns of importance were in line with clinical expectations.

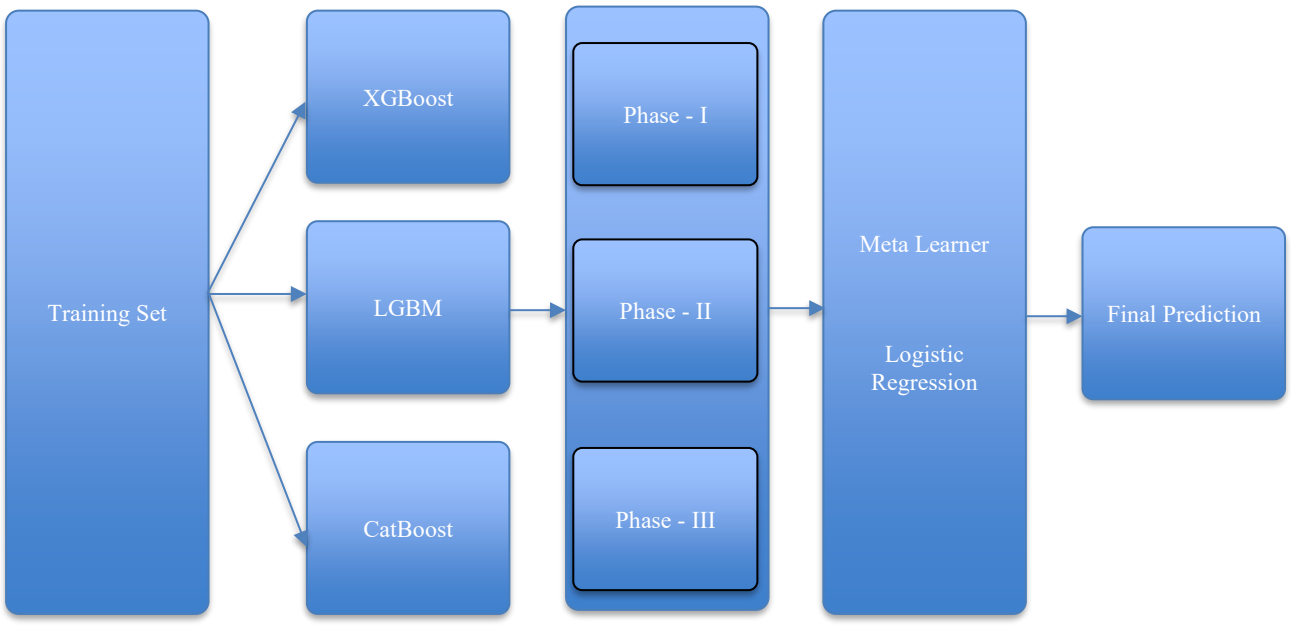


Fig. 6 Architecture of the stacking classifier integrating XGBoost, LightGBM and CatBoost through a logistic-regression meta-learner

6. Results and Discussion

This section presents the empirical findings obtained during the three phases of the proposed RP prediction framework. Phase I evaluated the classifiers on the original imbalanced dataset, Phase II assessed the impact of SMOTETomek-based class balancing, and Phase III examined the models over the balanced augmented dataset.

The results of the confusion matrix are summarized in Table 3, and phase-wise performance measures are shown in Table 4. Figure 7 shows the ROC curves, and Figure 8 compares the training-side and held-out test performance to review generalization. Across the three phases, the proposed stacking classifier outperformed, and the base classifier is seen to be a reliable recall-oriented model.

The improvements were not dramatic but quite modest, as is fitting since Phase I already represented a tuned, leakage-free, threshold-optimized baseline. Accordingly, the later phases are systematic enhancements that gradually improved recall-oriented behaviour, reduced the miss rates for pneumonia-positive cases, and strengthened the final model.

The operating threshold of the individual classifiers at which they were training in cross-validation was chosen and fixed on the held-out test set [16, 17].

6.1. Phase I: Results on the Original Imbalanced Dataset

Phase I, the classifier were evaluated over the original imbalanced dataset. The classifiers were tested with a recall-oriented thresholding strategy instead of a fixed default threshold, causing the baseline to be more clinically significant than a naive imbalanced-data test.

XGBoost was the highest performing standalone classifier with a precision of 0.25, a recall of 0.70, an F1-score of 0.37, an F2-score of 0.51 and an ROC-AUC of 0.61. LightGBM yielded a precision of 0.30, a recall of 0.60, an F1-score of 0.40, an F2-score of 0.50, and an ROC-AUC 0.64. CatBoost gave the best precision of the single models at 0.38, with a recall of 0.55, an F1-score of 0.45, an F2-score of 0.51 and an ROC-AUC of 0.65. SVM had a higher sensitivity and recall of 0.72 and an F2-score of 0.52, but poorer overall.

The stacking classifier performed best in this phase. It achieved the highest recall (0.86), the highest F2-score (0.57), and the highest ROC-AUC (0.66). The values in the corresponding confusion matrices in Table 4 reveal that the stacking classifier as well generated the least False Negatives (FN = 24) and the highest True Positives (TP = 146) in Phase I. Even though the false-positive cases was still high, the reduced number of false-negatives means that the suggested ensemble is more appropriate to screening-oriented applications than the individual models.

6.2. Phase II: Results after SMOTETomek-Based Imbalance Handling

Phase II examined whether imbalance handling through SMOTETomek could improve pneumonia-positive detection beyond the tuned imbalanced baseline. This phase maintained the leakage-safe evaluation system but altered the training distribution to minimize majority-class bias [22].

Among the standalone classifiers, XGBoost achieved the highest precision and recall, 0.35 and 0.55, respectively, and F1-score 0.43 and F2-score 0.50, as well as ROC-AUC 0.64. LightGBM was the strongest standalone tree-based model, with a precision of 0.35, a recall of 0.62, an F1-score of 0.44, an F2-score of 0.53 and a ROC-AUC of 0.65. CatBoost produced a precision 0.35, recall 0.53, F1-score 0.42, F2-score 0.48, and ROC-AUC 0.64. In this phase, SVM achieved the highest recall at 0.90, yet its ROC-AUC of 0.56, which implies that it is not separating its classes as well, despite having a relatively better sensitivity profile.

The stacking classifier is observed as the strongest overall model. It achieved precision 0.25, recall 0.88, F1-score 0.39, F2-score 0.58, and ROC-AUC 0.67. It obtained the maximum F2-score and ROC-AUC of the phase. Table 3 shows the confusion-matrix pattern, which further elucidates the outcome. Compared to the stacking classifier (FN = 21), SVM had a smaller number of false negatives (FN = 17), at the cost of a larger false-positive rate (FP = 500) and lower discrimination. However, the stacking classifier had a more desirable overall recall based balance, with high True-Positive recovery (TP = 149) and also the highest ROC-AUC. These findings suggest that SMOTETomek improved the minority-class and reinforced the suggested ensemble learning without altering the identity of the most successful model.

6.3. Phase III: Results on the Balanced Augmented Dataset

Phase III evaluated the balanced augmented dataset and was carried out as a follow-up of the tested Phase II framework. The learning pipeline over a balanced dataset was augmented to investigate the possibility of further enhancing recall-oriented performance [22].

The standalone classifier XGBoost had precision 0.24, recall 0.89, F1-score 0.37, F2-score 0.57 and ROC-AUC 0.60. LightGBM produced precision 0.24, recall 0.79, F1-score 0.37, F2-score 0.55, and ROC-AUC 0.62. CatBoost was still very sensitivity-based, with a precision of 0.23, a recall of 0.90, a F1-score of 0.37 and a F2-score of 0.58 and ROC-AUC of 0.60. SVM once again gave a high recall rate of 0.88, however, its ROC-AUC was low at 0.52.

As observed from the results, stacking classifier produced the strongest overall Phase III result. It achieved precision 0.24, recall 0.92, F1-score 0.38, F2-score 0.59, and ROC-AUC 0.65. These were the maximum value of recall, maximum of F2-score and maximum ROC-AUC in the phase. Table 3 also

validates the notion that the stacking classifier recorded the lowest number of False Negatives (FN = 14) and the highest number of True Positives (TP = 156) during Phase III. And this is among the most significant of the study findings, as it demonstrates that data augmentation enabled the proposed ensemble to achieve the most powerful final screening-oriented profile when Phase III was reconstructed with the Phase II framework.

The Phase III results, thus, affirm the ultimate design decision to adopt a tree-based stacking classifier as the model of choice to use in high-sensitivity RP risk flagging.

6.4. Result Interpretation and Practical Meaning

Based on the results of the three phases of experimentation, the following is observed: First, the stacking classifier is found to be the most robust model based on F2-score. Second, the advancements in Phase I to Phase III were not dramatic but gradual. In the stacking classifier, recall improved by 0.86 in Phase I to 0.88 in Phase II and 0.92 in Phase III, whereas F2-score improved by 0.57 to 0.58 and then to 0.59. This trend indicates that the primary value of balancing and data augmentation was not a significant overall upswing in each metric, but a gradual decrease in missed positive cases and a more recall-focused decision profile.

The confusion-matrix results in Table 3 also clarify the practical meaning of these gains. The stacking classifier in Phase I was more effective in reducing false negatives than the individual models. In Phase II, SVM showed slightly fewer false negatives, but the stacking classifier was the best performing model, combining a higher F2-score with a significantly improved ROC-AUC. Phase III, again saw the stacking classifier register the lowest false negatives and largest true positives.

The phase-wise findings also clarify the contribution of the individual models within the study. LightGBM and CatBoost were also good tree-based competitors, and SVM was performing consistently with high recall but poorly discriminative. This is why SVM was retained as an individual comparator but was not included in the final stacking ensemble. In contrast, XGBoost, LightGBM, and CatBoost provided a more stable and complementary base for ensemble learning.

Overall, the findings support the proposed stacking classifier as a robust model in the study under the stated clinical objective. The framework did not produce an unrealistic leap in performance, yet it was uniformly advantaged to recall and lowered false negatives at various stages. These features render it as the preferred model in screening-based prediction of radiation-induced pneumonitis.

Table 3. Confusion-matrix counts across the three experimental Phases

Phase	Model	TN	FP	FN	TP
Phase I	XGBoost	213	362	51	119
	LGBM	337	238	67	103
	CatBoost	424	151	76	94
	SVM	210	365	48	122
	Stacking Classifier	122	453	24	146
Phase II	XGBoost	401	174	76	94
	LGBM	378	197	65	105
	CatBoost	406	169	80	90
	SVM	75	500	17	153
	Stacking Classifier	125	450	21	149
Phase III	XGBoost	90	485	19	151
	LGBM	159	416	35	135
	CatBoost	73	502	16	154
	SVM	70	505	20	150
	Stacking Classifier	79	496	14	156

Table 4. Results for the baseline classifiers and the proposed stacking classifier

Model Name	Precision	Recall	F1-score	F2-score	ROC-AUC
Phase I : Results of ML classifiers over the Original Dataset (imbalanced dataset)					
XGBoost	0.25	0.70	0.37	0.51	0.61
LGBM	0.30	0.60	0.40	0.50	0.64
CatBoost	0.38	0.55	0.45	0.51	0.65
SVM	0.25	0.72	0.37	0.52	0.60
Stacking Classifier (XGBoost + LGBM + CatBoost)	0.24	0.86	0.38	0.57	0.66
Phase II: Results of ML classifiers over the Original Dataset with imbalance handling (SMOTETomek)					
XGBoost	0.35	0.55	0.43	0.50	0.64

LGBM	0.35	0.62	0.44	0.53	0.65
CatBoost	0.35	0.53	0.42	0.48	0.64
SVM	0.23	0.90	0.37	0.57	0.56
Stacking Classifier (XGBoost + LGBM + CatBoost)	0.25	0.88	0.39	0.58	0.67
Phase III: Results of ML classifiers over the balanced augmented dataset					
XGBoost	0.24	0.89	0.37	0.57	0.60
LGBM	0.24	0.79	0.37	0.55	0.62
CatBoost	0.23	0.90	0.37	0.58	0.60
SVM	0.23	0.88	0.36	0.56	0.52
Stacking Classifier (XGBoost + LGBM + CatBoost)	0.24	0.92	0.38	0.59	0.65

Figure 7. ROC-AUC curves of the evaluated classifiers across the three experimental phases

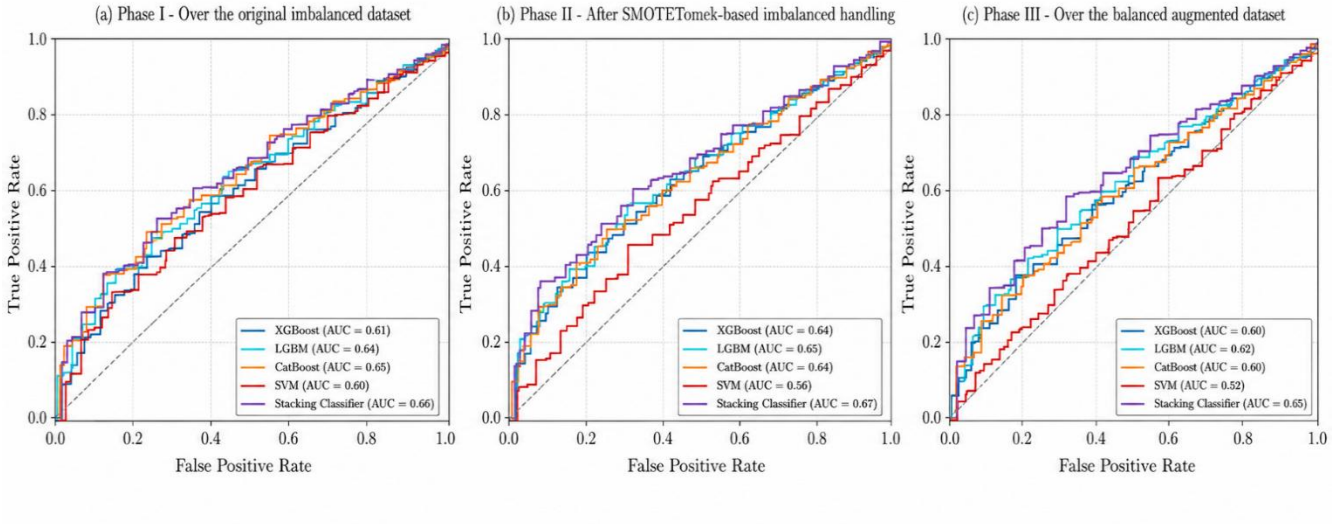


Fig. 7 ROC-AUC curves of the classifiers in the three experiment Phases, (a) Phase I, (b) Phase II, and (c) Phase III.

6.5. Error Patterns, Robustness, and Overfitting

The confusion-matrix analysis showed clear differences in error across the three phases. The initial imbalanced environment continued to impose a visible false-negative penalty on the standalone models in Phase I, CatBoost, LightGBM, and XGBoost missed 76, 67, and 51 cases of pneumonia, respectively, but SVM missed 48 cases. In comparison, the stacking classifier minimized the number of false-negatives (24) and found 146 true positives. That implies that even with the initial imbalanced distribution, the given ensemble was already able to offer the best screening-oriented error profile of the analyzed models.

In Phase II, the error pattern changed after SMOTETomek-based balancing. This decrease in false negatives underscores the importance of class balancing for recall-based detection, but this benefit came at the cost of increased false-positive rates in certain models. For example, SVM identified 153 pneumonia-positive cases, missed 17, and detected 500 false positives. The stacking classifier was highly recall-biased in this step, with 149 true positives and 21 false negatives, but also had a large number of false positives (450). Nonetheless, the combination of these counts with F2-score and ROC-AUC indicates that the stacking classifier outperformed the other considered Phase II models.

In Phase III, the fixed balanced data augmentation strategy altered the error trade-off further. XGBoost found 151 true positives and missed 19 cases, whereas CatBoost found 154 true positives with 16 false negatives. The stacking classifier produced the highest final sensitivity profile with 156 true positives and just 14 false negatives, and a false-positive value of 496. These results indicate that data augmentation further enhanced the recall-oriented detection, though the improvement occurred primarily due to increased sensitivity. Therefore, the Phase III design is useful for diagnosing cases.

The entire procedure of preprocessing, resampling, data augmentation, as well as threshold selection was limited to the training portion of the pipeline, with the held-out test set being utilized only in the final evaluation step as outlined in Section 5.8 [16, 17]. This interpretation of Phase II is also supported by the generalization summary in Figure 8. Specifically, the stacking classifier is observed to be the most consistent in cross-validated versus held-out test performance, with a cross-

validated F2-score at the chosen threshold of 0.58 and held-out test F2-score of 0.58, and the highest test ROC-AUC of 0.66. This pattern supports the view that the improvement observed in the preferred Phase II setting reflects genuine modelling gain.

6.6. Clinical Meaning and Alignment with Metrics

The use of the F2-score as the primary measure was driven by clinical reasons of minimizing false-negative predictions, and this contributes to minimizing the chance of missing patients with pneumonia-positive conditions and may require further observation or early intervention [16]. In this context, the clinical relevance of recall-oriented performance over and above that of accuracy is relevant since the detection of a high-risk patient might have more implications than a false alert.

The threshold-dependent measures used in this paper, such as precision, recall, F1-score and F2-score, were computed with respect to the pneumonia-positive class (class = 1), whereas ROC-AUC was determined based on the predicted probability of class 1. Simultaneously, secondary metrics are relevant in getting a better insight into how the models behaved. ROC-AUC was used to assess overall discrimination across thresholds, whereas precision and the confusion-matrix structure helped to interpret the practical effect of model performance [17]. The final findings indicate that various models proved to be better based on the clinical aim. The stacking classifier has the highest F2-score and ROC-AUC. These results suggest that the ultimate model selection must be based on the need of the target use case to focus more on sensitivity-oriented screening or a more balanced clinical decision-support role.

Since the methodology is based on the regularly gathered structured variables instead of the special-purpose imaging pipelines, the proposed framework may be more readily implemented into the current head-and-neck radiotherapy of the workflow as an early-warning support tool. Nevertheless, it is also seen that the present results suggest that the advantage of balancing and data augmentation is model-dependent instead of being consistently incremental, and should therefore be interpreted with appropriate clinical caution [22].

6.7. Overfitting, Generalization and ROC-AUC Analysis

Figure 8 summarizes the generalization behaviour of the Phase II models by comparing cross-validated training-side performance with held-out test performance. The findings suggest that the gains in the model development were high under independent assessment. The differences between cross-validated and held-out test performance between XGBoost, LightGBM, and CatBoost are relatively small, again in line with the consistent behaviour of regularized tree-based models in clinically noisy, but structured classification tasks. The operating threshold selected by SVM was more sensitive, but the stacking classifier had the most balanced overall profile in the desired Phase II setting with the least variation between cross-validated F2-score at the selected threshold and held-out test F2-score, as well as the highest test ROC-AUC.

The ROC-AUC comparison in Figure 7 further supports the phase-wise interpretation of model behaviour. In Phase I, the models exhibited small but distinct discrimination with the original imbalanced distribution, and the stacking classifier gave the highest ROC-AUC among the considered models. Phase II shows further improvement in ROC-AUC values with SMOTETomek-based balancing, suggesting the enhanced distribution between classes in the balanced dataset training environment. CatBoost was also competitive throughout the phases in terms of tree-based comparator, and the stacking classifier also presented the best overall balance in the Phase II environment. Phase III had a competitive discrimination, and the stacking classifier recorded the best ROC-AUC amongst the tested classifiers. Overall, these findings imply that the proposed modelling pipeline enhanced minority-class learning with no obvious signs of leakage-based overestimation.

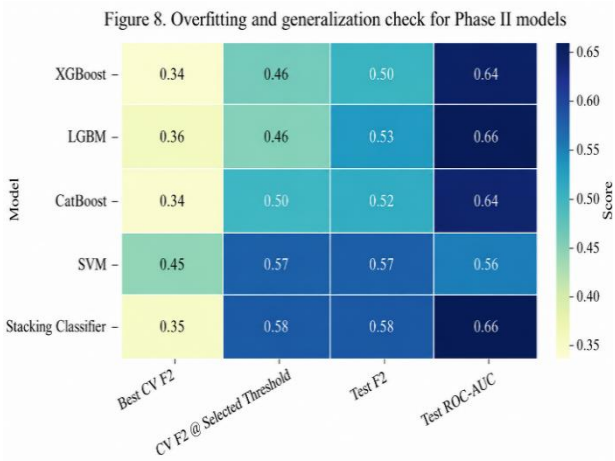


Fig. 8 Overfitting and generalization summary of the Phase II classifiers based on cross-validated, F2-score and held-out test performance

7. Conclusion

The present study developed and evaluated a leakage-safe, phase-based, predictive model of Radiation-Induced Pneumonitis (RP) in Head and Neck Cancer (HNC) patients, using routinely collected clinical and treatment variables. The framework was structured across three progressive stages: Phase I used the original imbalanced data to establish a preprocessed baseline using the original imbalanced data; Phase II introduced SMOTETomek-based class balancing is introduced; and Phase III applied balanced data augmentation on top of the Phase II pipeline. At each of three phases, four classifiers: XGBoost, LightGBM, CatBoost, and SVM were systematically compared. A tree-based stacking ensemble, which consists of XGBoost, LightGBM and CatBoost as

meta-learners, is proposed and evaluated as the primary classifier. The results show that the stacking classifier provided the strongest and most consistent recall-oriented performance across the study. However, the phase-to-phase gain is not substantial, and this is understandable, as the Phase I baseline is already optimized, threshold, and leakage safe. Phase II is the most balanced result of the three phases, as it included a robust F2-score performance, better discrimination, and generalization stability following balancing with SMOTETomek. While Phase III further enhanced sensitivity, it resulted in a high false-positive rate. This is why the results of Phase II are presented as most stable in the present study. From a clinical perspective, the main value of the framework lies in its ability to reduce missed pneumonia-positive cases and to support early risk flagging in a screening-oriented setting. The suggested approach is not designed to substitute clinical judgment; instead, it will be used to help in recognizing patients who might need more careful observation during or after radiotherapy. As the framework is based on regularly collected hospital data, it can be practically applied to real-world oncology workflows as well as to build scalable healthcare solutions. While this study provides valuable insights, its scope is defined by a retrospective, single-centre design, which necessitates broader external validation. The way forward in future work will be multicenter testing, prospective evaluation and addition of more predictor sets with more treatment and follow-up information. Despite such shortcomings, the current results indicate that a recall-based stacking model may be clinically significant and methodologically sound to help predict the risk of radiation-induced pneumonitis in head and neck cancer.

Acknowledgements

The authors thank the Cachar Cancer Hospital and Research Centre, Silchar, Assam, for providing the data, tools, and facilities essential for this research. The assistance and advice of the healthcare professionals involved in this study are particularly significant and greatly appreciated. The authors gratefully acknowledge the Department of Computer Science, Assam University, Silchar, for providing the facilities necessary to carry out this work.

Author Contributions

The study was designed and conceived by all the authors. The preparations of the material, data collection and analysis were conducted by Abhijit Nath, Sheikh Wakie Masood, Shahin Ara Begum and Ravi Kannan. Abhijit Nath wrote the first draft of the manuscript, and all the authors commented on the drafts of the manuscript. None of the authors rejected the final manuscript.

Funding

The authors do not have financial conflicts of interest to disclose.

Data Availability

The information gathered for this research is not open to the public because of patient privacy.

Ethics Approval

Approval was obtained from the IRB/Ethics Committee of Cachar Cancer Hospital and Research Centre, Silchar, Assam, India (18th IRB meeting, 11th May 2019, File No.: CCHRC/IRB/01/2019/92).

Consent to Participate

Informed consent to participate was sought, and anonymity and voluntary participation was ensured.

Competing Interests

The authors declare no competing interests.

References

[1] Suzanne Tanya Nethan, Shalini Gupta, and Saman Warnakulasuriya, *Risk Factors for Oral Squamous Cell Carcinoma in the Indian Population*, Microbes and Oral Squamous Cell Carcinoma, Springer, Singapore, pp. 9-40, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[2] Debojyoti Roy et al., "Profiling of Oral Squamous Cell Carcinoma of Lower Assam Region – A Hospital-based Retrospective Study," *Contemporary Clinical Dentistry*, vol. 15, no. 4, pp. 240-243, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[3] Feihu Chen et al., "Re-Evaluating the Risk Factors for Radiation Pneumonitis in the Era of Immunotherapy," *Journal of Translational Medicine*, vol. 21, no. 1, pp. 1-15, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[4] Marisol Arroyo-Hernández et al., "Radiation-Induced Lung Injury: Current Evidence," *BMC Pulmonary Medicine*, vol. 21, no. 1, pp. 1-12, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[5] Fengsong Ye et al., "Predicting Radiation Pneumonitis in Lung Cancer: A EUD-based Machine Learning Approach for VMAT Patients," *Frontiers in Oncology*, vol. 14, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[6] Zhi Chen et al., "Predicting Radiation Pneumonitis in Lung Cancer using Machine Learning and Multimodal Features: A Systematic Review and Meta-Analysis of Diagnostic Accuracy," *BMC Cancer*, vol. 24, no. 1, pp. 1-18, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[7] Yoshiyuki Katsuta et al., "Prediction of Radiation Pneumonitis with Machine Learning using 4D-CT-based Dose-Function Features," *Journal of Radiation Research*, vol. 63, no. 1, pp. 71-79, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[8] N.V. Chawla et al., "SMOTE: Synthetic Minority Over-Sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002. [CrossRef] [Google Scholar] [Publisher Link]

[9] Haibo He, and Edwardo A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263-1284, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[10] Heesoon Sheen et al., "Radiomics-based Hybrid Model for Predicting Radiation Pneumonitis: A Systematic Review and Meta-Analysis," *Physica Medica*, vol. 123, pp. 1-9, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[11] David H. Wolpert, "Stacked Generalization," *Neural Networks*, vol. 5, no. 2, pp. 241-259, 1992. [CrossRef] [Google Scholar] [Publisher Link]

[12] Tianqi Chen, and Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, New York, NY, United States, pp. 785-794, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[13] Guolin Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *Advances in Neural Information Processing Systems*, vol. 30, 2017. [Google Scholar] [Publisher Link]

[14] Corinna Cortes, and Vladimir Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273-297, 1995. [CrossRef] [Google Scholar] [Publisher Link]

[15] Peter I. Frazier, "A Tutorial on Bayesian Optimization," *arXiv preprint*, pp. 1-22, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[16] Alaa Tharwat, "Classification Assessment Methods," *Applied Computing and Informatics*, vol. 17, no. 1, pp. 168-192, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[17] Tom Fawcett, "An Introduction to ROC Analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861-874, 2006. [CrossRef] [Google Scholar] [Publisher Link]

[18] Jingli Tang et al., "Radiation Pneumonitis Prediction using Dual-Modal Data Fusion based on Med3D Transfer Network," *Journal of Digital Imaging*, vol. 38, no. 4, pp. 2085-2101, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[19] Kyu Hye Choi et al., "Using Deep Learning to Predict Radiation Pneumonitis in Patients Treated with Stereotactic Body Radiotherapy (SBRT) for Pulmonary Nodules: Preliminary Results," *Journal of the Korean Physical Society*, vol. 81, no. 5, pp. 460-470, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[20] Yan Kong et al., "Enhancing the Prediction of Symptomatic Radiation Pneumonitis for Locally Advanced Non-Small-Cell Lung Cancer by Combining 3D Deep Learning-Derived Imaging Features with Dose–Volume Metrics: A Two-Center Study," *Radiation Oncology and Oncology*, vol. 201, no. 3, pp. 274-282, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[21] Jang Hyung Lee et al., "Deep-Learning Model Prediction of Radiation Pneumonitis using Pretreatment Chest CT and Clinical Factors," *Technology in Cancer Research and Treatment*, vol. 23, pp. 1-11, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[22] Liming Sheng et al., "Radiation Pneumonia Predictive Model for Radiotherapy in Oesophageal Carcinoma Patients," *BMC Cancer*, vol. 23, no. 1, pp. 1-11, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[23] Rutger H. Stoffers et al., "Rising Incidence of Radiation Pneumonitis After Adjuvant Durvalumab in NSCLC Patients Treated with Concurrent Chemoradiotherapy," *Acta Oncologica*, vol. 64, pp. 267-275, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[24] M. Athiyaman et al., "A Physicist Description of Indigenous Telecobalt Machine Bhabhatron-II TAW," *Global Journal of Science Frontier Research*, vol. 15, no. 3, pp. 105-110, 2015. [Google Scholar]

[25] National Cancer Institute, Common Terminology Criteria for Adverse Events (CTCAE) Version 5.0, U.S. Department of Health and Human Services, NIH, NCI, 2017. [Online]. Available: https://dctd.cancer.gov/research/ctep-trials/for-sites/adverse-events/ctcae-v5-5x7.pdf