Du *et al. EJNMMI Physics* (2025) 12:34
https://doi.org/10.1186/s40658-025-00750-7

EJNMMI Physics

ORIGINAL RESEARCH

Open Access

Impact of harmonization and oversampling methods on radiomics analysis of multi-center imbalanced datasets: application to PET-based prediction of lung cancer subtypes

Dongyang Du[1,2,3,4], Isaac Shiri[5], Fereshteh Yousefirizi[4], Mohammad R. Salmanpour[4], Jieqin Lv[2,3], Huiqin Wu[2,3], Wentao Zhu[6], Habib Zaidi[5], Lijun Lu[2,3,7*] and Arman Rahmim[4,8]

*Correspondence:
Lijun Lu
ljlubme@gmail.com
[1]College of Computer Science, Inner Mongolia University, Hohhot, Inner Mongolia 010021, China
[2]School of Biomedical Engineering, Guangdong Provincial Key Laboratory of Medical Image Processing, Southern Medical University, Guangzhou, Guangdong 510515, China
[3]Guangdong Province Engineering Laboratory for Medical Imaging and Diagnostic Technology, Southern Medical University, Guangzhou, Guangdong 510515, China
[4]Department of Integrative Oncology, BC Cancer Research Institute, Vancouver, BC V5Z 1L3, Canada
[5]Division of Nuclear Medicine and Molecular Imaging, Geneva University Hospital, Geneva 4 CH-1211, Switzerland
[6]Research Center for Healthcare Data Science, Zhejiang Lab, Hangzhou, Zhejiang 311121, China
[7]Pazhou Lab, Guangzhou, Guangdong 510330, China
[8]Departments of Radiology and Physics, University of British Columbia, Vancouver, BC V5Z 1M9, Canada

Abstract

Background Medical imaging data frequently encounter image-generation heterogeneity and class imbalance properties, challenging strong generalized predictive performances with data-driven machine-learning methods. The purpose of this study was to investigate the impact of harmonization and oversampling methods on multi-center imbalanced datasets, with specific application to PET-based radiomics modeling for histologic subtype prediction in non-small cell lung cancer (NSCLC).

Methods The retrospective study included 245 patients with adenocarcinoma (ADC) and 78 patients with squamous cell carcinoma (SCC) from 4 centers. Utilizing 1502 radiomics features per patient, we trained, validated, and tested 4 machine-learning classifiers, to investigate the effect of no harmonization (NoH) or 4 feature harmonization methods, paired with no oversampling (NoO) or 5 oversampling methods on subtype prediction. Model performance was evaluated using the average area under the ROC curve (AUROC) and G-mean via 5 times 5-fold cross-validations. Statistical comparisons of the combined models against baseline (NoH + NoO) were performed for each fold of cross-validation using the DeLong test.

Results The number of cross-combinations with both AUROC and G-mean outperforming baseline in validation and testing was 15, 4, 2, and 7 (out of 29) for random forest (RF), linear discriminant analysis (LDA), logistic regression (LR), and support vector machine (SVM), respectively. ComBat harmonization combined with oversampling (SMOTE) via RF yielded better performance than baseline (AUROC and G-mean of validation: 0.725 vs. 0.608 and 0.625 vs. 0.398; testing: 0.637 vs. 0.567 and 0.506 vs. 0.287), though statistical significances were not observed.

Conclusions Applying harmonization and oversampling methods in multi-center imbalanced datasets can improve NSCLC-subtype prediction, but the effect varies widely across classifiers. We have created open-source comparisons of harmonization and oversampling on different classifiers for comprehensive evaluations in different studies.

Keywords PET radiomics, Harmonization, Oversampling, Multi-center imbalanced datasets, NSCLC

Introduction

Lung cancer is the leading cause of cancer death worldwide, among which approximately 82% cases present with non-small cell lung cancers (NSCLC) [1, 2]. Adenocarcinoma (ADC) and squamous cell carcinoma (SCC) are the main histologic subtypes of NSCLC, accounting for ~60% and ~35%, respectively [3]. Accurate histologic classification and clinical staging are of paramount importance in treatment planning and prognostication for lung cancer patients. For instance, a pemetrexed-based regimen was found to significantly prolong overall survival and progression-free survival in ADC while having the opposite effect on SCC [4]. Hence, accurately distinguishing SCC from ADC is an essential step in the therapeutic decision-making process.

The field of radiomics, a non-invasive framework to extract high-dimensional quantitative features from medical images and to build predictive/prognostic models via machine-learning pipelines, is being extensively applied in positron emission tomography (PET) [5], including lung cancer [6]. Although promising results have been reported in the differential diagnosis of lung cancer [7], histological subtyping [8], treatment response assessment [9], and prognosis analysis [10], still, there remains significant concerns about (i) the poor reproducibility, (ii) inaccurate prediction, and (iii) limited generalizability of radiomics studies caused by variations of imaging parameter, imbalanced distributions of data, and the lack of multi-center validation [11–13].

Imaging data generated from different scanner models, acquisition protocols, and reconstruction settings introduced non-biological systematic-related differences in radiomics features, strongly affecting the reproducibility and predictive performance of radiomics [14, 15]. To eliminate the imaging effect and facilitate the clinical translation of radiomics, harmonization has been proposed as a useful pre-processing step to solve the data heterogeneity problem [16]. In particular, ComBat harmonization based on empirical Bayes framework is widely used to align the feature distributions between different centers/scanners and has produced satisfactory results in different cancer types [17, 18]. However, harmonization transformations have been commonly estimated and assessed within entire cohorts increasing the risk of information leakage between training and testing cohorts, and the ability to harmonize testing cohorts remains to be established. Overall, the generalization performance of ComBat for radiomics features harmonization needs to be further evaluated [19].

Imaging data frequently suffers from class-imbalanced problem where the uneven distribution of majority and minority classes can lead to the misclassification of data-driven machine-learning algorithms [20]. Because fewer minority examples are mistaken for noise and discarded, the classifier is biased toward the prediction of the majority class, which has a larger decision region. Since the accurate prediction of minority classes is crucial for clinical practice, re-sampling techniques that generate more minority examples, like Oversampling, have been shown to be effective in improving prediction accuracy [21]. However, only a limited number of studies have investigated the efficacy of re-sampling approaches on imbalanced PET radiomics [22–24]. In particular, Xie et al. [23] conducted a multi-center study and highlighted the improvement of re-sampling techniques combined with classifiers on PET radiomics-based prognostic performance. Meanwhile, in that highly heterogenous multicentric dataset, the impact of harmonization methods on prognostication performance was not investigated.

Variabilities in generated images in current clinical practice are unavoidable, particularly for multi-center modeling where the data are acquired using different machines and protocols, and furthermore, class imbalance characteristics are commonly present in imaging data. It remains to investigate how pre-processing techniques of harmonization combined with the oversampling impact the performance of radiomics models in a multi-center imbalanced context, which we pursue. In the present work, we study 323 patients combined from 3 datasets involving 4 centers with intrinsic imaging heterogeneity between the centers and manufacturers in terms of acquisition and reconstruction parameters, conducting a comparative study to evaluate the impact of harmonization and oversampling methods on the performance of PET radiomics, applying to histologic subtype prediction of NSCLC. The prediction performance for 4 harmonization methods and 5 oversampling methods against baseline (no harmonization and no oversampling) were compared via 4 machine learning classifiers in validation and testing cohorts.

Materials and methods

Patient selection and image acquisition

Three NSCLC datasets were enrolled in this study, following specific inclusion and exclusion criteria (Fig. 1). The inclusion criteria were: (1) patients with ADC or SCC proven by pathological diagnosis; (2) available pre-treatment PET images. The exclusion criteria were: (1) PET images with no uptake; (2) PET images with low quality (excluded by previous studies [18, 25]); (3) patients with more than one ADC and/or SCC simultaneously. The imaging dataflow diagram, which provides institution, manufacturer, and reconstruction flow information for each dataset, is shown in Fig. 2. More detailed information concerning acquisition and reconstruction protocols are described in supplementary Table S1.

Dataset 1 [26, 27] was retrieved from The Cancer Imaging Archive (TCIA), which contains 159 eligible patients from Palo Alto Veterans Affairs Healthcare System (hereafter, referred to as center 1) and Stanford University School of Medicine (hereafter, referred to as center 2). PET images were acquired on multi-vendor (i.e., GE, Philips, and Siemens) PET/CT scanners and reconstructed using various methods (e.g., ordered

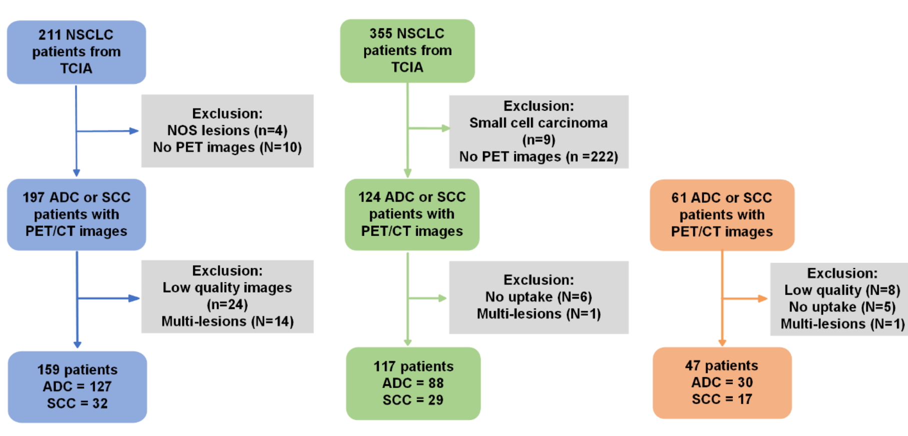

Fig. 1 Dataflow diagram of patient selection for the three datasets

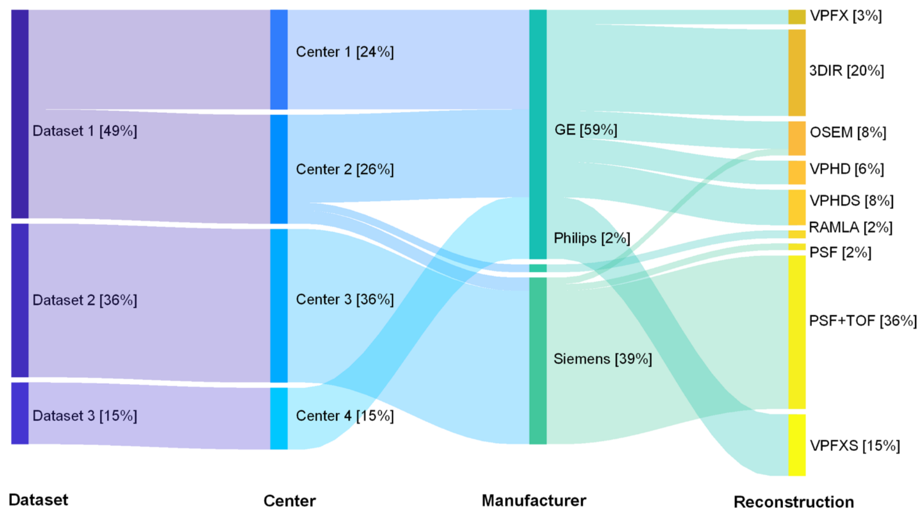

Fig. 2 Dataflow diagram of imaging characteristic for each dataset. The number in brackets shows the percentage of cases in each parameter or batch

subset expectation maximization (OSEM), row action maximum likelihood algorithm (RAMLA), and VUE-point HD (VPHD)).

Dataset 2 [28] also obtained from TCIA includes 117 patients from an independent institute (hereafter, referred to as center 3) according to the inclusion and exclusion criteria. PET scans were performed with a Siemens Biograph 64 mCT scanner, and the acquired time-of-flight (TOF)-PET data were reconstructed using the OSEM algorithm with point spread function (PSF) modeling. This study was performed in accordance with the TCIA Data Usage Policy; datasets 1 and 2 were de-identified, as such, did not require Institutional Review Board (IRB) approval, and patient informed consent was waived.

Dataset 3 was collected from Hainan Cancer Hospital, including 47 patients who met the enrollment criteria (hereafter, referred to as center 4). All PET images acquired on the GE Discovery 710 scanner were reconstructed using the VUE-point FX with PSF modeling (VPFXS) method. The IRB approved the usage of dataset 3, and patient informed consent was waived due to the retrospective design of the present study.

Image pre-processing and feature extraction

ITK-SNAP software (version 3.6.0; www.itksnap.org) was used to manually delineate region of interest (ROI) for each patient. The manual delineation was performed on PET images, while also referring to CT images to determine the metabolic location of the tumor. The segmented tumors were finally checked and verified slice by slice by senior radiologists with more than 10 years of experience in thoracic radiology [18, 25]. A comprehensive open-source Image Biomarkers Standardization Initiative (IBSI) compliant package called Pyradiomics (version 3.0.1) was applied for customizing the feature extraction [29]. A total of 1502 radiomics features, including 107 original, 744 wavelet, 186 Laplacian of Gaussian (LOG), 93 gradient, 93 square, 93 square root, 93 logarithm, and 93 exponential features, were extracted from each ROI (See supplementary Table S2). More details for image preprocessing and feature extraction are described in supplementary material. The codes for reproducible feature extraction were shared to https://github.com/dudongyangsmu/HarmonizationOversampling.

Data splitting

To assess the generalizability of harmonization techniques, 10% (32 samples) of the total samples were taken as the testing set using a stratified random sampling method based on subtype. The remaining 90% (291 samples) were further divided into training and validation sets using 5-fold cross-validation to assess the effect of harmonization and oversampling methods on model performance. No significant demographic differences between the training and validation or testing cohort. This process was repeated 5 times on the whole dataset to get more reliable results (Fig. 3).

Assessment of radiomics feature robustness

To build more robust models and allow wider applicability of the results, we evaluated the robustness of radiomics features to different segmentations and noise levels. The original segmentations of all patients in the training set were randomly perturbed by the combination of tumor growth and/or shrinkage, rotation, and translation to generate random segmentations; 40% of SUV_{max} was used in original segmentations to generate threshold-based segmentations. Two noisy images were created for each patient of the training set by adding Gaussian noise with random standard deviations (SD, ranges from 0.1 to 0.5 with the step of 0.05 SUV). Supplementary Figure S1 shows the generated segmentations and noise images. The 2-way random effect intraclass correlation coefficient (ICC) with a 95% confidence interval (CI) was calculated for each feature to assess the absolute agreement of feature extraction between different segmentations and between different noise images, respectively, and features with an ICC 95% CI lower bound < 0.5 were considered as non-robust and removed from the analysis [30].

Harmonization methods

To remove the feature variations caused by different manufacturers, scanners, and acquisition and reconstruction parameters, especially in the multi-center study (also called batch effect), we investigated 4 popular harmonization methods [31, 32]: ComBat, centering-scaling, Singular Value Decomposition (SVD)-based, and Independent

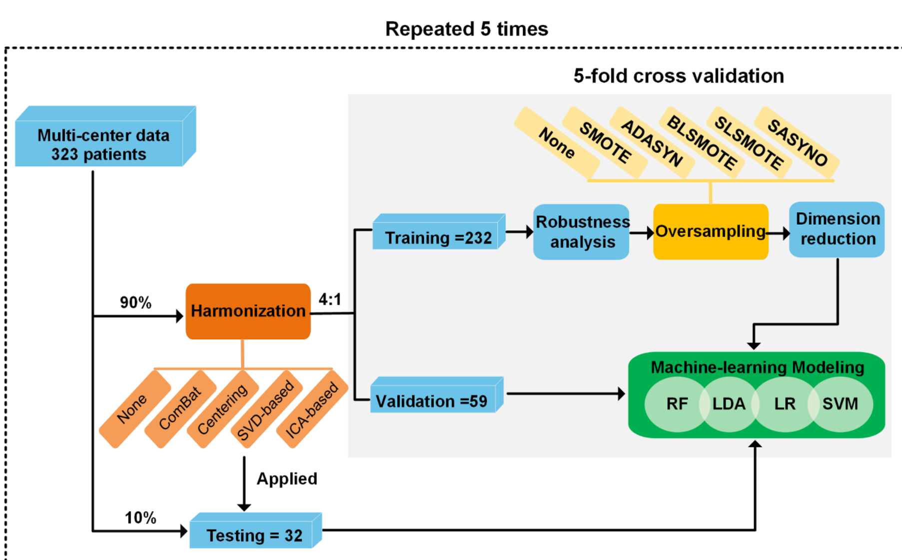

Fig. 3 Workflow of the radiomics study adopted in this protocol

Component Analysis (ICA)-based matrix factorization methods. The first two are location-scale methods that transform the features to produce similar mean and/or variance values in each batch. The last two are matrix factorization methods that filter out the components associated with batch to reconstruct clean data. To explore the feasibility of harmonizing previously unseen data (e.g., a new patient or cohort from a known batch), we applied the learned harmonization transformation from internal data (training and validation data) to testing data, assuming the batch effect in the testing data was identical with that of the internal data. Since imaging characteristics were highly heterogeneous among the three datasets (Fig. 2), different batch division strategies were determined based on center, manufacturer, reconstruction algorithm, and k-means clustering in this study. Details of harmonization methods are available in Supplementary Material.

Oversampling methods

The motivation behind oversampling is to generate new samples, modify imbalance distribution between minority and majority classes in the training set, increase intra-class diversity and enhance model's classification ability, especially for the minority class. We applied 5 well-known oversampling methods in the training set, namely synthetic minority oversampling technique (SMOTE) [33], adaptive synthetic (ADASYN) [34], borderline-SMOTE (BSMOTE) [35], safe-level-SMOTE (SSMOTE) [36], and self-adaptive synthetic oversampling (SASYNO) [37]. Details of oversampling methods are described in Supplementary Material.

Model construction

To avoid model overfitting introduced by the curse of dimensionality, the minimum redundancy-maximum relevance (MRMR) method was used to rank the features according to the relevance with the target class and the redundancy between the features. The top k-features ($k < 15$) were fed into classifiers, and k was determined based on the average area under the receiver operating characteristics curve (AUROC) of the 5-fold cross-validation within the training set. Four widely used classifiers were investigated: random forest (RF), linear discriminant analysis (LDA), logistic regression (LR), and support vector machine (SVM) with linear kernel. The predictive performances of 30 cross-combinations derived from no harmonization (NoH) or 4 harmonization methods coupled with no oversampling (NoO) or 5 oversampling methods were evaluated on each classifier. The AUROC and G-mean, which are suitable for imbalanced classification, were used as evaluation metrics [23]. Open-source solutions were created and shared via GitHub (https://github.com/dudongyangsmu/HarmonizationOversampling) for comprehensive comparisons of harmonization and oversampling techniques via different machine-learning classifiers.

Statistical analysis

The ROC curve comparisons of machine learning models on cross-validated datasets are known to be challenging [38] because of the inherent dependence between the various cross-validation folds and the complex relationships between the trained models. In this study, we used a framework proposed by Van De Wiel et al. [39]: the DeLong test [40] was used to perform pairwise comparisons of the ROC curves in each fold, and the median *p*-value over the different folds was reported as the final *p*-value. $P < 0.05$

indicated a statistically significant difference. Statistical analyses were conducted using the R package "Daim" (Version 1.1.0).

Results

Patients

A total of 323 patients (mean age, 65.16 ± 9.35 years) from 4 centers were enrolled in this study. Of these, 206 (63.8%) and 117 (36.2%) were male and female, respectively. Furthermore, 245 (75.9%) and 78 (24.1%) patients were with ADC and SCC, respectively. The imbalance ratio between ADC and SCC was 3.14:1. More details of the demographic and clinical characteristics of patients in each center are shown in Table 1.

Radiomics feature robustness

The percentage of selected robust features (categorized by different filters) in the robustness assessment is summarized in Fig. 4, with additional details summarized in Supplementary Table S3. A total of 842 (56.1%) radiomics features [range: 802–875 (53.4 – 58.3%)] were robust to both diverse segmentations and noise levels, which were applied to subsequent analysis. Segmentation had a moderate effect on wavelet features (76.3%); other classes of features, except for gradient (87.1%), square root (80.7%), and original features (80.4%), were strongly affected (square to LOG features, 51.6 – 70.4%). Noise mainly affected wavelet features, with only 58.1% of wavelet features were robust, while other classes of features were less affected, with robustness percentages ranging from 88.2 to 97.9%.

Impact of harmonization and oversampling on prediction performance

In this study, features were harmonized among 4 centers, 3 manufacturers, 9 reconstruction algorithms, and 2 clusters of k-means clustering, respectively. Many models presented the highest performance when using the center to define the batch for

Table 1 The demographic and clinical characteristics of patients

Characteristic	Dataset 1		Dataset 2	Dataset 3	All patients
	Center 1	Center 2	Center 3	Center 4	
Patient no.	76	83	117	47	323
Age (years)	68.76 ± 7.61	66.94 ± 10.83	61.96 ± 9.71	62.96 ± 9.25	65.16 ± 9.35
Sex					
Male	76	32	70	28	206
Female	0	51	47	19	117
T stage					
Tis/T1/T2/T3/T4	3/35/25/11/2	1/22/26/5/3	0/63/31/18/4	-	4/120/82/34/9
Not collected	0	26	1	47	74
N stage					
N0/N1/N2/N3	61/7/8/0	45/5/7/0	73/27/5/11	-	179/39/20/11
Not collected	0	26	1	47	74
M stage					
M0/M1	75/1	54/3	85/31	-	214/35
Not collected	0	26	1	47	74
Subtype					
ADC	53	74	88	30	245
SCC	23	9	29	17	78

Data are the number of patients, except for age depicted as mean ± standard deviation

ADC, adenocarcinoma; SCC, squamous cell carcinoma

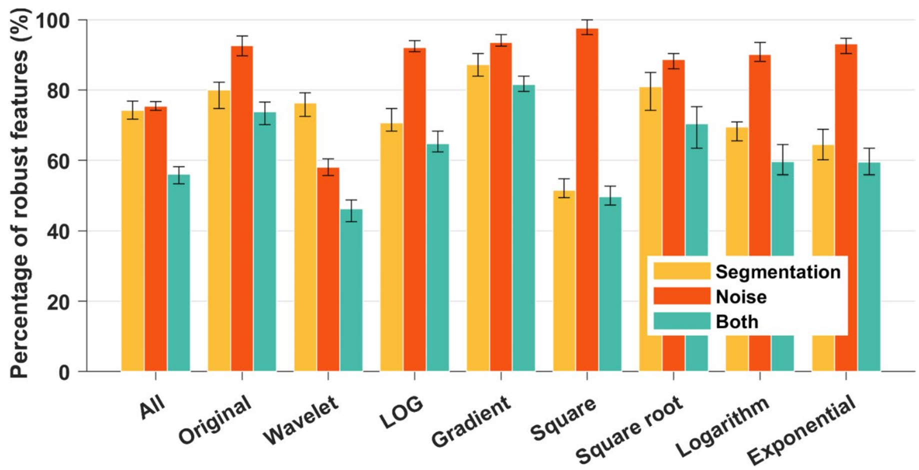

Fig. 4 The percentage of robust features categorized by different filters in the robustness assessment

harmonization, followed by reconstruction-based harmonization, whereas manufacturer-based harmonization and clustering-based harmonization showed the lowest AUROC values. Figure 5 depicts the AUROC and G-mean values of paired harmonization and oversampling methods for four machine-learning classifiers in the validation set where the center was used as a batch. The results for reconstruction, manufacturer, and cluster-based harmonization are provided in Supplementary Figures S3, S4, and S5. In the validation cohort, the RF classifier, ComBat, and Centering-scaling harmonization methods showed highest predictive performances in combination with the majority of oversampling methods, with AUROC values ranging from 0.712 to 0.737 and G-mean from 0.573 to 0.626 (Fig. 5a and b). Compared with the baseline (NoH + NoO; AUROC: 0.608; G-mean: 0.398), all the combinations (29/29) exhibited improved performance in both AUROC and G-mean (Fig. 5c). For LDA, LR, and SVM classifiers, only 13 (45%), 6 (21%), and 23 (79%) combinations of harmonization and oversampling methods showed higher AUROC values than baseline (AUROCs for LDA: 0.650–0.672 vs. 0.647; AUROCs for LR: 0.653–0.667 vs. 0.651; AUROCs for SVM, 0.625–0.662 vs. 0.623), respectively. However, G-mean values consistently increased from 0.128 to 0.216 to 0.532–0.626 when oversampling methods were applied (Fig. 5d and l).

Combinations with both AUROC and G-mean outperforming the baseline in the validation cohort were considered as effective. Figure 6 further shows the testing performance of these effective combinations. Fourteen of the 29 (52%) combinations on the RF classifier, 4 of the 13 (31%) combinations on the LDA classifier, 2 of the 6 (33%) combinations on the LR classifier, and 7 of the 23 (30%) combinations on the SVM classifier had generalization ability with both AUROC and G-mean better than baseline. The best result was achieved by ComBat + BSMOTE (AUROC: 0.654; G-mean: 0.494), followed by ComBat + ADASYN (AUROC: 0.651; G-mean: 0.500), ComBat + SSMOTE (AUROC: 0.649: G-mean: 0.453), and ComBat + SMOTE (AUROC: 0.637; G-mean: 0.506) via RF classifier. However, no statistically significant differences were observed with the baseline in the validation and testing cohorts.

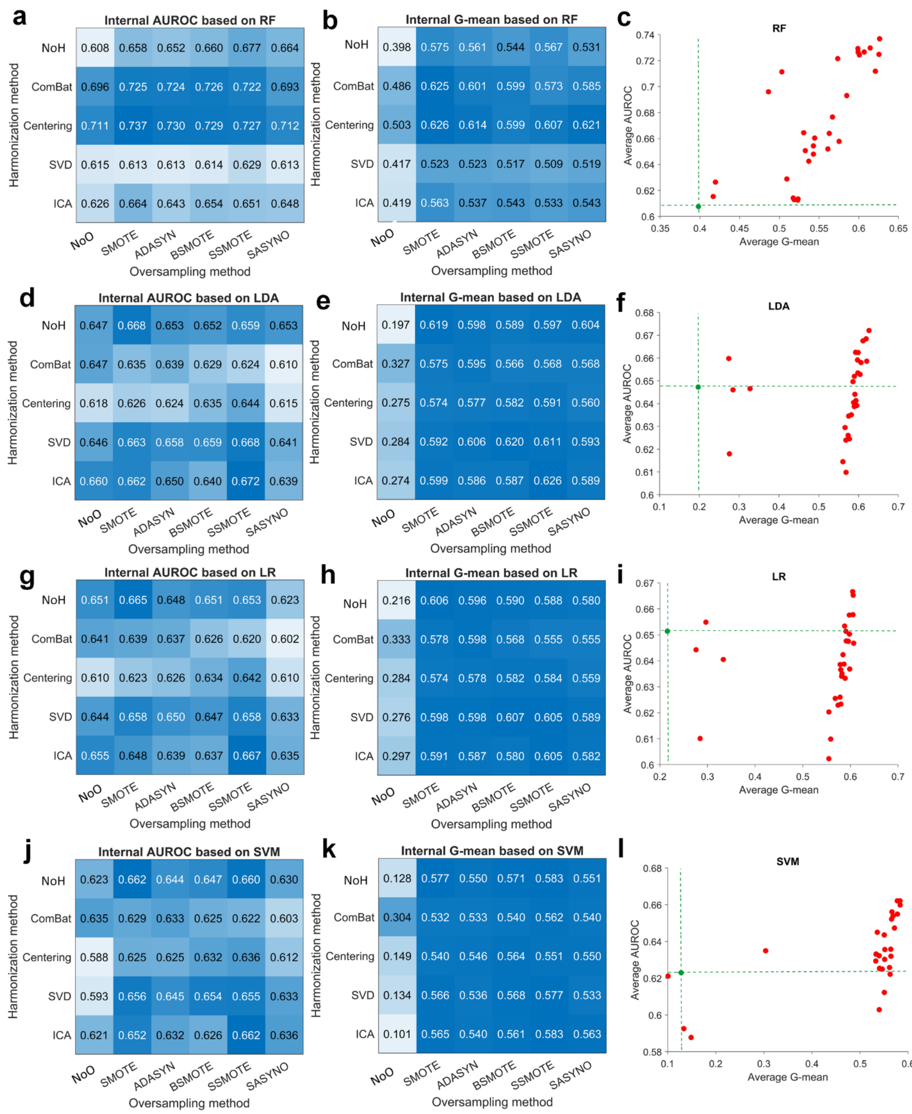

Fig. 5 AUROCs (left column), G-means (median column), and scatterplots of AUROC and G-mean (right column) for 30 combinations of harmonization and oversampling methods on (**a-c**) RF, (**d-f**) LDA, (**g-i**) LR, and (**j-l**) SVM classifiers in the validation cohort. Batch for harmonization was defined based on center. The green point shows the AUROC and G-mean of baseline (NoH + NoN) in the validation cohort. Combinations with both AUROC and G-mean greater than baseline in the validation cohort were considered effective and high-performing combinations, which correspond to the right upper region of green line

Discussion

Variabilities in generated images and class imbalance characteristics are frequently present in medical imaging data, making it challenging to achieve good generalized radiomics performance using data-driven machine learning methods [41]. In the present study, utilizing 1502 radiomics features per patient, we evaluated the effect of the image pre-processing techniques harmonization and oversampling on multi-center imbalanced PET radiomics for the identification of ADC and SCC. The results showed that performance prediction can be improved by applying harmonization and oversampling methods in multi-center imbalanced datasets though we did not find statistical significance in small cohorts of patients. To the best of our knowledge, this is the first

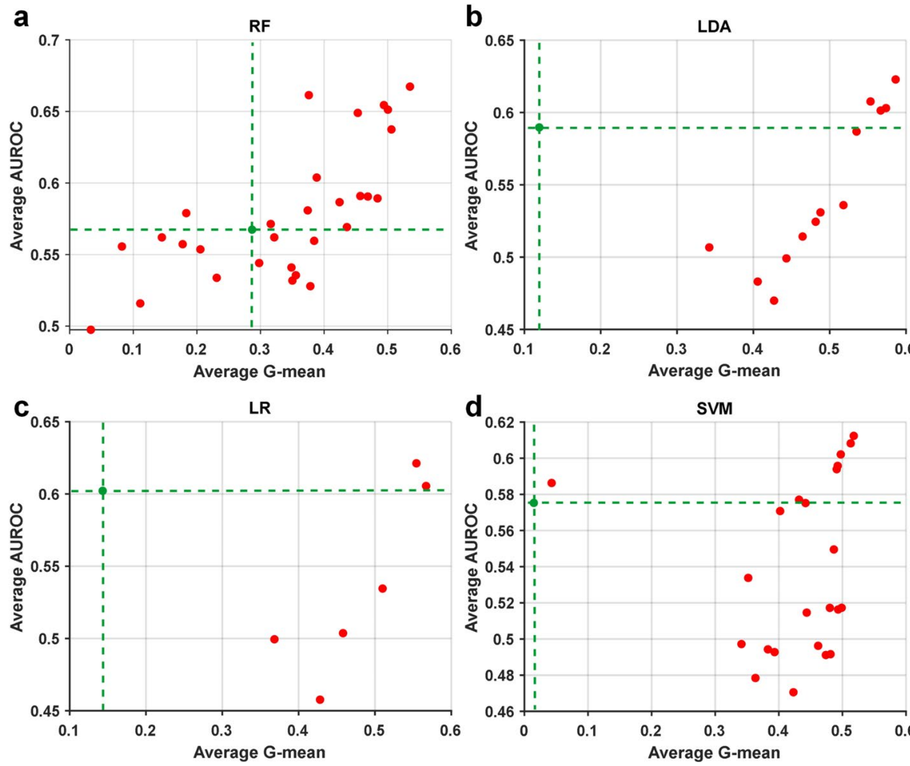

Fig. 6 Scatterplots of AUROC and G-mean for the effective high-performing combinations on (**a**) RF, (**b**) LDA, (**c**) LR, and (**d**) SVM classifiers in the testing cohort. The green point shows the AUROC and G-mean of baseline in the testing cohort. The right upper region of green line corresponds to the combinations with both AUROC and G-mean greater than baseline in both validation and testing cohorts

study comprehensively studying the impact of cross-combinations of harmonization and oversampling for multi-center imbalanced radiomics.

To determine whether harmonization can enhance the prediction/prognosis performance of radiomics models, previous studies [18, 42] estimated harmonization transformations and carried out feature correction in the entire dataset, increasing the risk of information leakage between training and testing cohorts. By contrast, the harmonization efforts in our current study were constructed using internal data and then applied to testing data, which not only assessed the impact of harmonization on radiomics but also emphasized the generalizability of harmonization. Batch determination is essential for feature harmonization. It is well established that variability of a number of acquisition and reconstruction factors can influence the PET radiomics feature values, However, these studies also highlighted that the sensitivity of radiomic features to different factors can vary greatly [43, 44]. In this study, a great variability was observed in terms of manufactures and reconstruction settings, while the injected activity, uptake time and scanning time are in accordance with the EANM standard, with slight differences among centers. However, image quality was synergistically affected by all these factors. It was thus challenging to determine the batch labels. We used k-means clustering to determine the batch, patients with similar feature distributions had same batch label. Furthermore, with the assumption that the center, manufacturer, reconstruction method relative to other factors has predominant impact on radiomics features, we also performed center-based, manufacturer-based and reconstruction-based harmonization to investigate the efficiency of harmonization. The findings indicated that the employed harmonization methods are batch-dependent because different batch divisions produced varying effects on the prediction performance of cross-combinations.

It is worth noting that the validation performance demonstrated the impact of feature harmonization on radiomics modelling, while the testing performance depicted the generalization ability of the feature harmonization. In this study, applying harmonization and oversampling showed a trend toward improving NSCLC-subtype prediction. However, the improvements were limited, and statistical significance was not demonstrated. Possible reasons for this are: (i) The sample size is relatively small, and complex sources of variability exist in the data. (ii) ComBat corrects for the center effect and may also remove crucial biological information, as the subtypes (ADC and SCC) do not present with the same frequencies between different centers (Table 1). (iii) ComBat can introduce a covariate to explain the different distributions [45], but this study aims to differentiate ADC and SCC. Hence, the histological subtype cannot be added to the ComBat [46]. (iv) ComBat, to be useful, assumes the feature distributions across centers must be similar except for the additive and multiplicative effects [45]. However, due to the existence of complex confounders and the computation of higher-order features, this linearity assumption does not hold for some features [47].

The effects of combinations of harmonization and oversampling on prediction performance differed in different classifiers, resulting in 15, 4, 2, and 7 (out of 29) combinations outperformed the baseline for RF, LDA, LR, and SVM classifiers, respectively. Additionally, Ferreira et al. [48] reported that applying ComBat did not improve the disease-free survival prediction of cervical cancer when PET radiomics features were combined with clinical features, although the model performance varied across scanners. Lv et al. [42] found that ComBat had no impact on imbalance-adjusted PET radiomics models used to predict survival of four-center head and neck cancer patients. However, many studies also demonstrated the positive effect of ComBat, for example, Shiri et al. [18] and Dissaux et al. [49] observed that ComBat improved the prediction performance of PET-radiomics analysis in multicentric settings. Therefore, there is no consensus in the radiomics analysis regarding ComBat being the best harmonization as it was impacted by various machine-learning classifiers, data intrinsic characteristics, and different clinical tasks. When retrospectively dealing with complex PET datasets with high SUV inconsistencies, post-processing harmonization with ComBat—which constructs a linear model based on a statistical framework to correct for these inconsistencies across scanners—may face limitations. Instead, the use of deep learning-based harmonization approaches may be more effective in voxel-level harmonization than ComBat [50, 51].

Harmonization improved data overall-consistency by eliminating variations of feature values attributed to differences in imaging devices, acquisition protocols, and reconstruction settings, while oversampling increased intra-class diversity by generating new minority class samples. In this study, we have shown that when oversampling methods were applied, G-mean values were consistently increased for all combined models on four classifiers, while AUROC values were slightly improved for part of cross-combinations on four classifiers (Figs. 5 and 6). Similarly, a previous study found that applying re-sampling methods statistically significantly increased G-means but not AUROCs [23]. Also, in a recent large study on radiomics, Demircioğlu et al. [52] demonstrated that applying resampling methods did not improve the average predictive performance on fifteen radiomics datasets, only slight improvements in AUC were observed on specific dataset. On the other hand, G-mean shows an accurate prediction for both classes, indicating that the models have a more balanced accuracy for the majority and minority classes. Therefore, in imbalanced radiomics, oversampling techniques need to be applied more frequently for balanced accuracy.

In the field of radiomics, the high-throughput features were usually extracted to comprehensively quantify the heterogeneity of tumor. In terms of the curse of high dimensionality, two steps of feature selection were performed in this study. Firstly, the features were ranked according to the relevance with the target class and the redundancy between the features. Secondly, considering the relationship between the number of modeling variables and the sample size [53, 54], only the top sqrt(n) features (n equals to the training sample size, herein sqrt(n) $\approx$ 15) were used to construct the optimal feature subsets based to the average performance of the cross-validation in the training set. It must be emphasized that the 30 combined models selected different sets of features with agreement ranging from 1 to 26% (Figure S6), hindering the interpretability of features. Despite that, it should be noted that the selected features were highly correlated between the different oversampling methods with the Pearson correlation coefficients > 0.8, when fixing the harmonization method (Figure S7). This provided an explanation as to why we observed many marginal performance differences between different oversampling methods in Figs. 5 and 6. Moreover, we found that the features involved into the ComBat-related models also exhibited a high correlation with those in Centering-related models, suggesting that the ComBat was likely to play a similar role to center-based z-score normalization.

Since the data imbalance was primarily driven by the data from center 2, the effect of the ComBat method on PET radiomics analysis was further validated in centers 1, 3 and 4. Notice that, the predominant trends of feature distributions present in the data were associated with the center (Figure S8a), which recapitulated the need to have feature harmonization in PET radiomics. The center effect was removed after Combat harmonization (Figure S8b). Additionally, the random forest classifier was trained with any two centers, and tested using the remaining one. We observed that applying ComBat yield higher performance than the baseline model with AUROC of 0.728 ± 0.037 vs. 0.674 ± 0.066, G-mean of 0.644 ± 0.054 vs. 0.461 ± 0.141, sensitivity of 0.620 ± 0.207 vs. 0.414 ± 0.395, specificity of 0.706 ± 0.168 vs. 0.703 ± 0.255. Based on these findings, we believe that radiomics feature harmonization in modern nuclear medicine imaging is necessary.

We also paid attention to other parameters, such as image noise [55], tumor segmentation [56], and feature extraction software [57], which have been demonstrated to affect the reproducibility and robustness of radiomics features. To this end, multiple sets of noise images and tumor segmentations were created for each patient, and all the feature calculations were performed using an open-source package, Pyradiomics, which is well-documented and easy-replicated. According to the guideline of ICC for robustness analysis, ICC 95% CI lower bounds less than 0.5 are indicative of poor reliability [30]. Higher thresholds highlight the effects of noise and segmentation on radiomics features but at the expense of removing the biologically important features. Therefore, in this trade-off, our approach using 0.5 as a threshold to select robust features seems to be a safe option.

The respiratory motion can lead to degradation of image quality and imprecision of tumor contouring, particularly in PET chest images. Therefore, the goodness of the segmentation and the reproducibility of the radiomics features was influenced by the respiratory motion [58]. With the advance in PET/CT imaging technology, the new commercial PET system is equipped with data-driven gating (DDG) technique for motion correction [59]. However, due to the nature of retrospective study, DDG was not employed for all PET images in this study. Although we assessed the robustness of radiomics features by image perturbation (noise addition and contour transformation), which may be used to mimic the reproducibility of radiomics features to test-retest imaging [60], the impact of respiratory motion on the reproducibility and clinical performance of radiomic models before and after harmonization need further evaluation. Future studies using respiratory-gated PET images and ungated images to explore the clinical value of harmonization are our actively pursued.

Our work still had some limitations. First, despite being gathered from several institutions, the sample size for our dataset was not very large, especially results were validated and tested using relatively small sizes of datasets, which may lead to the non-significance of results and limit the extrapolation of results to other cancer types using different scanners and imaging protocols. Future validation will require more extensive prospective studies, which are currently unavailable. Secondly, we mainly concentrated on feature-level harmonization and oversampling techniques. Further exploration can include image-level methods such as image harmonization based on deep learning [61] and image oversampling based on geometric transformation [62].

Conclusions

We investigated the impact of harmonization methods jointly with oversampling methods to enable improved prediction performance in multi-center imbalanced datasets. In particular, we focused on utilizing PET radiomics models based on machine-learning algorithms for subtype prediction in NSCLC. ComBat combined with SMOTE, BSMOTE, SSMOTE and ADASYN oversampling via the RF classifier performed the best, although improvements were not statistically significant. Of note, depending on the clinical tasks and data characteristics, the effect of harmonization and oversampling methods on prediction performance varies across different machine-learning classifiers. To this end, we have constructed and shared open-source codes so that future research efforts can comprehensively compare the combinations of harmonization and oversampling methods based on different classifiers and drive the development of new harmonization and/or oversampling techniques.

Abbreviations

PET	Positron emission tomography
NSCLC	Non-small cell lung cancer
ADC	Adenocarcinoma
SCC	Squamous cell carcinoma
TCIA	Cancer Imaging Archive
OSEM	Ordered subset expectation maximization
RAMLA	Row action maximum likelihood algorithm
VPHD	VUE-point HD
TOF	Time-of-flight
PSF	Point spread function
VPFXS	VUE-point FX with PSF modeling
IBSI	Image Biomarkers Standardization Initiative
ICC	Intraclass correlation coefficient
CI	Confidence interval
SVD	Singular Value Decomposition
ICA	Independent Component Analysis
NoO	No oversampling

NoH	No harmonization
SMOTE	Synthetic minority oversampling technique
ADASYN	Adaptive synthetic
BSMOTE	Borderline-SMOTE
SSMOTE	Safe-level-SMOTE
SASYNO	Self-adaptive synthetic oversampling
AUROC	Area under the ROC curve
RF	Random forest
LDA	Linear discriminant analysis
LR	Logistic regression
SVM	Support vector machine

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40658-025-00750-7.

Supplementary Material 1

Acknowledgements

The authors acknowledge helpful discussions with Drs. Babak Saboury and Carlos Uribe.

Author contributions

(I) Conception and design: D Du, I Shiri, L Lu, and A Rahmim; (II) Administrative support: L Lu, A Rahmim; (III) Provision of study materials or patients: I Shiri, W Zhu, H Zaidi, L Lu, and A Rahmim; (IV) Collection and assembly of data: I Shiri, J Lv, H Wu, D Du; (V) Data analysis and interpretation: D Du, I Shiri, F Yousefirizi, MR. Salmanpour, A Rahmim; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Funding

This study was supported by the National Natural Science Foundation of China (Nos. 62371221 and 12326616), the Science and Technology Program of Guangdong Province (No. 2022A0505050039), the Natural Sciences and Engineering Research Council of Canada (NSERC) (No. RGPIN-2019-06467), and the Inner Mongolia Autonomous Region Natural Science Foundation (No. 2024QN08063).

Data availability

Datasets 1 and 2 are available from The Cancer Imaging Archive (TCIA) NSCLC Radiogenomics and Lung-PET-CT-Dx datasets, respectively. Dataset 3 is available from the corresponding author on reasonable request.

Code availability

https://github.com/dudongyangsmu/HarmonizationOversampling.

Declarations

Ethics approval and consent to participate

The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study abided by the The Cancer Imaging Archive Data Usage Policy, public data were de-identified thus did not require Institutional Review Board approval, and patient informed consent was waived. The usage of local data was approved by the Institutional Review Board of Hainan Cancer Hospital and individual consent for this retrospective analysis was waived.

Consent for publication

Not applicable.

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Received: 18 July 2024 / Accepted: 24 March 2025
Published online: 07 April 2025

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71:209–49.
2. Miller KD, Nogueira L, Devasia T, Mariotto AB, Yabroff KR, Jemal A, Kramer J, Siegel RL. Cancer treatment and survivorship statistics, 2022. CA Cancer J Clin. 2022;72:409–36.
3. Ji Y, Qiu Q, Fu J, Cui K, Chen X, Xing L, Sun X. Stage-specific PET radiomic prediction model for the histological subtype classification of non-small-cell lung cancer. Cancer Manag Res. 2021;13:307–17.
4. Scagliotti G, Hanna N, Fossella F, Sugarman K, Blatter J, Peterson P, Simms L, Shepherd FA. The differential efficacy of pemetrexed according to NSCLC histology: a review of two phase III studies. Oncologist. 2009;14:253–63.
5. Orlhac F, Nioche C, Klyuzhin I, Rahmim A, Buvat I. Radiomics in PET imaging: a practical guide for newcomers. PET Clin. 2021;16:597–612.

6. Manafi-Farid R, Askari E, Shiri I, Pirich C, Asadi M, Khateri M, Zaidi H, Beheshti M. [18F]FDG-PET/CT radiomics and artificial intelligence in lung cancer: technical aspects and potential clinical applications. Semin Nucl Med. 2022;52:759–80.
7. Du D, Gu J, Chen X, Lv W, Feng Q, Rahmim A, Wu H, Lu L. Integration of PET/CT radiomics and semantic features for differentiation between active pulmonary tuberculosis and lung cancer. Mol Imaging Biol. 2021;23:287–98.
8. Han Y, Ma Y, Wu Z, Zhang F, Zheng D, Liu X, Tao L, Liang Z, Yang Z, Li X, Huang J, Guo X. Histologic subtype classification of non-small cell lung cancer using PET/CT images. Eur J Nucl Med Mol Imaging. 2021;48:350–60.
9. Shao D, Du D, Liu H, Lv J, Cheng Y, Zhang H, Lv W, Wang S, Lu L. Identification of stage IIIC/IV EGFR-mutated non-small cell lung cancer populations sensitive to targeted therapy based on a PET/CT radiomics risk model. Front Oncol. 2021;11:721318.
10. Zhao M, Kluge K, Papp L, Grahovac M, Yang S, Jiang C, Krajnc D, Spielvogel CP, Ecsedi B, Haug A, Wang S, Hacker M, Zhang W, Li X. Multi-lesion radiomics of PET/CT for non-invasive survival stratification and histologic tumor risk profiling in patients with lung adenocarcinoma. Eur Radiol 2022.
11. Ketabi A, Ghafarian P, Mosleh-Shirazi MA, Mahdavi SR, Rahmim A, Ay MR. Impact of image reconstruction methods on quantitative accuracy and variability of FDG-PET volumetric and textural measures in solid tumors. Eur Radiol. 2019;29:2146–56.
12. Naseri H, Skamene S, Tolba M, Faye MD, Ramia P, Khriguian J, Patrick H, Andrade Hernandez AX, David M, Kildea J. Radiomics-based machine learning models to distinguish between metastatic and healthy bone using lesion-center-based geometric regions of interest. Sci Rep. 2022;12:9866.
13. Piñeiro-Fiel M, Moscoso A, Pubul V, Ruibal Á, Silva-Rodríguez J, Aguiar P. A systematic review of PET textural analysis and radiomics in cancer. Diagnostics. 2021;11:380.
14. Hosseini SA, Shiri I, Hajianfar G, Bahadorzadeh B, Ghafarian P, Zaidi H, Ay MR. Synergistic impact of motion and acquisition/reconstruction parameters on ^{18}F-FDG PET radiomic features in non-small cell lung cancer: Phantom and clinical studies. Med Phys. 2022;49:3783–96.
15. Xu Y, Lu L, Sun SH, Lian EL, Yang W, Schwartz H, Yang LH, Zhao Z. Effect of CT image acquisition parameters on diagnostic performance of radiomics in predicting malignancy of pulmonary nodules of different sizes. Eur Radiol. 2022;32:1517–27.
16. Da-Ano R, Visvikis D, Hatt M. Harmonization strategies for multicenter radiomics investigations. Phys Med Biol. 2020;65:24TR02.
17. Orlhac F, Boughdad S, Philippe C, Stalla-Bourdillon H, Nioche C, Champion L, Soussan M, Frouin F, Frouin V, Buvat I. A postreconstruction harmonization method for multicenter radiomic studies in PET. J Nucl Med. 2018;59:1321–8.
18. Shiri I, Amini M, Nazari M, Hajianfar G, Haddadi Avval A, Abdollahi H, Oveisi M, Arabi H, Rahmim A, Zaidi H. Impact of feature harmonization on radiogenomics analysis: prediction of EGFR and KRAS mutations from non-small cell lung cancer PET/CT images. Comput Biol Med. 2022;142:105230.
19. Da-ano R, Lucia F, Masson I, Abgral R, Alfieri J, Rousseau C, Mervoyer A, Reinhold C, Pradier O, Schick U, Visvikis D, Hatt M. A transfer learning approach to facilitate comBat-based harmonization of multicentre radiomic features in new datasets. PLoS ONE. 2021;16:e0253653.
20. He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl DATA Eng. 2009;21:22.
21. Mohammed R, Rawashdeh J, Abdullah M. Machine learning with oversampling and undersampling techniques: overview study and experimental results. In: 2020 11th International Conference on Information and Communication Systems (ICICS). Irbid, Jordan: IEEE. 2020:243–248.
22. Lv J, Chen X, Liu X, Du D, Lv W, Lu L, Wu H. Imbalanced data correction based PET/CT radiomics model for predicting lymph node metastasis in clinical stage T1 lung adenocarcinoma. Front Oncol. 2022;12:788968.
23. Xie C, Du R, Ho JW, Pang HH, Chiu KW, Lee EY, Vardhanabhuti V. Effect of machine learning re-sampling techniques for imbalanced datasets in 18F-FDG PET-based radiomics model on prognostication performance in cohorts of head and neck cancer patients. Eur J Nucl Med Mol Imaging. 2020;47:2826–35.
24. Krajnc D, Spielvogel CP, Grahovac M, Ecsedi B, Rasul S, Poetsch N, Traub-Weidinger T, Haug AR, Ritter Z, Alizadeh H, Hacker M, Beyer T, Papp L. Automated data Preparation for in vivo tumor characterization with machine learning. Front Oncol. 2022;12:1017911.
25. Amini M, Hajianfar G, Hadadi Avval A, Nazari M, Deevband MR, Oveisi M, Shiri I, Zaidi H. Overall survival prognostic modelling of non-small cell lung cancer patients using positron emission tomography/computed tomography harmonized radiomics features: the quest for the optimal machine learning algorithm. Clin Oncol. 2022;34:114–27.
26. Bakr S, Gevaert O, Echegaray S, Ayers K, Zhou M, Shafiq M, Zheng H, Benson JA, Zhang W, Leung ANC, Kadoch M, Hoang CD, Shrager J, Quon A, Rubin DL, Plevritis SK, Napel S. A radiogenomic dataset of non-small cell lung cancer. Sci Data. 2018;5:180202.
27. Gevaert O, Xu J, Hoang CD, Leung AN, Xu Y, Quon A, Rubin DL, Napel S, Plevritis SK. Non–small cell lung cancer: identifying prognostic imaging biomarkers by leveraging public gene expression microarray data—methods and preliminary results. Radiology. 2012;264:387–96.
28. [Dataset] Li P, Wang S, Li T, Lu J, HuangFu Y, Wang D. A large-scale CT and PET/CT dataset for lung cancer diagnosis. Cancer Imaging Archive Cancer Imaging Arch 2020.
29. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, Beets-Tan RGH, Fillion-Robin J-C, Pieper S, Aerts HJWL. Computational radiomics system to Decode the radiographic phenotype. Cancer Res. 2017;77:e104–7.
30. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15:155–63.
31. Renard E, Absil PA. Comparison of batch effect removal methods in the presence of correlation between outcome and batch. PLoS ONE. 2018;13:e0202947.
32. Ligero M, Jordi-Ollero O, Bernatowicz K, Garcia-Ruiz A, Delgado-Muñoz E, Leiva D, Mast R, Suarez C, Sala-Llonch R, Calvo N, Escobar M, Navarro-Martin A, Villacampa G, Dienstmann R, Perez-Lopez R. Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis. Eur Radiol. 2021;31:1460–70.
33. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.
34. Haibo He Y, Bai, Garcia EA, Shutao Li ADASYN. Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). Hong Kong, China: IEEE. 2008:1322–1328.

35. Han H, Wang W-Y, Mao B-H, Borderline. -SMOTE: a new over-sampling method in imbalanced data sets learning. In: Huang D-S, Zhang X-P, Huang G-B, editors. Advances in intelligent computing. Lecture Notes in Computer Science. Volume 3644. Berlin, Heidelberg: Springer Berlin Heidelberg; 2005. pp. 878–87.
36. Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C, Safe-Level. -SMOTE: safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In: Theeramunkong T, Kijsirikul B, Cercone N, Ho T-B, editors. Advances in knowledge discovery and data mining. Berlin, Heidelberg: Springer Berlin Heidelberg; 2009. pp. 475–82.
37. Gu X, Angelov PP, Soares EA. A self-adaptive synthetic over-sampling technique for imbalanced classification. Int J Intell Syst. 2020;35:923–43.
38. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput. 1998;10:1895–923.
39. van de Wiel MA, Berkhof J, van Wieringen WN. Testing the prediction error difference between 2 predictors. Biostatistics. 2009;10:550–60.
40. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44:837– 45.
41. Castiglioni I, Rundo L, Codari M, Di Leo G, Salvatore C, Interlenghi M, Gallivanone F, Cozzi A, D'Amico NC, Sardanelli F. AI applications to medical images: from machine learning to deep learning. Phys Med. 2021;83:9–24.
42. Lv W, Feng H, Du D, Ma J, Lu L. Complementary value of intra- and peri-tumoral PET/CT radiomics for outcome prediction in head and neck cancer. IEEE Access. 2021;9:81818–27.
43. Pfaehler E, Beukinga RJ, de Jong JR, Slart RHJA, Slump CH, Dierckx RAJO, Boellaard R. Repeatability of ^{18}F-FDG PET radiomic features: a Phantom study to explore sensitivity to image reconstruction settings, noise, and delineation method. Med Phys. 2019;46:665–78.
44. Yan J, Chu-Shern JL, Loi HY, Khor LK, Sinha AK, Quek ST, Tham IWK, Townsend D. Impact of image reconstruction settings on texture features in ^{18}F-FDG PET. J Nucl Med Publ Soc Nucl Med. 2015;56:1667–73.
45. Orlhac F, Eertink JJ, Cottereau A-S, Zijlstra JM, Thieblemont C, Meignan M, Boellaard R, Buvat I. A guide to combat harmonization of imaging biomarkers in multicenter studies. J Nucl Med. 2022;63:172–9.
46. Mali SA, Ibrahim A, Woodruff HC, Andrearczyk V, Müller H, Primakov S, Salahuddin Z, Chatterjee A, Lambin P. Making radiomics more reproducible across scanner and imaging protocol variations: A review of harmonization methods. J Pers Med. 2021;11:842.
47. Liu Y, Liu Z, Du D, Luna JM, Rahmim A, Jha A. Assessing linearity of PET-derived radiomics features across scanners: implications for combat harmonization [abstract]. J Nucl Med. 2022;63(supplement 2):3174.
48. Ferreira M, Lovinfosse P, Hermesse J, Decuypere M, Rousseau C, Lucia F, Schick U, Reinhold C, Robin P, Hatt M, Visvikis D, Bernard C, Leijenaar RTH, Kridelka F, Lambin P, Meyer PE, Hustinx R. [18F]FDG PET radiomics to predict disease-free survival in cervical cancer: a multi-scanner/center study with external validation. Eur J Nucl Med Mol Imaging. 2021;48:3432–43.
49. Dissaux G, Visvikis D, Da-ano R, Pradier O, Chajon E, Barillot I, Duvergé L, Masson I, Abgral R, Santiago Ribeiro M-J, Devillers A, Pallardy A, Fleury V, Mahé M-A, De Crevoisier R, Hatt M, Schick U. Pretreatment ^{18}F-FDG PET/CT radiomics predict local recurrence in patients treated with stereotactic body radiotherapy for early-stage non–small cell lung cancer: a multicentric study. J Nucl Med. 2020;61:814–20.
50. Haberl D, Spielvogel CP, Jiang Z, Orlhac F, Iommi D, Carrió I, Buvat I, Haug AR, Papp L. Multicenter PET image harmonization using generative adversarial networks. Eur J Nucl Med Mol Imaging. 2024;51:2532–46.
51. Du D, Lv W, Lv J, Chen X, Wu H, Rahmim A, Lu L. Deep learning-based harmonization of CT reconstruction kernels towards improved clinical task performance. Eur Radiol. 2023;33:2426–38.
52. Demircioğlu A. The effect of data resampling methods in radiomics. Sci Rep. 2024;14:2858.
53. Foley D. Considerations of sample and feature size. IEEE Trans Inf Theory. 1972;18:618–26.
54. Chalkidou A, O'Doherty MJ, Marsden PK. False discovery rates in PET and CT studies with texture features: a systematic review. PLoS ONE. 2015;10:e0124165.
55. Oliver JA, Budzevich M, Hunt D, Moros EG, Latifi K, Dilling TJ, Feygelman V, Zhang G. Sensitivity of image features to noise in conventional and respiratory-gated PET/CT images of lung cancer: uncorrelated noise effects. Technol Cancer Res Treat. 2017;16:595–608.
56. Lu L, Lv W, Jiang J, Ma J, Feng Q, Rahmim A, Chen W. Robustness of radiomic features in [11 C]Choline and [18F]FDG PET/CT imaging of nasopharyngeal carcinoma: impact of segmentation and discretization. Mol Imaging Biol. 2016;18:935–45.
57. Fornacon-Wood I, Mistry H, Ackermann CJ, Blackhall F, McPartlin A, Faivre-Finn C, Price GJ, O'Connor JPB. Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform. Eur Radiol. 2020;30:6241–50.
58. Chen YH, Kan KY, Liu SH, Lin HH, Lue KH. Impact of respiratory motion on ^{18}F-FDG PET radiomics stability: clinical evaluation with a digital PET scanner. J Appl Clin Med Phys. 2023;24:e14200.
59. Aide N, Lasnon C, Desmonts C, Armstrong IS, Walker MD. McGowan DR. Advances in PET/CT technology: an update. Semin Nucl Med. 2022;52:286–301.
60. Zwanenburg A, Leger S, Agolli L, Pilz K, Troost EGC, Richter C, Löck C. Assessing robustness of radiomic features by image perturbation. Sci Rep. 2019;9:614.
61. Tixier F, Jaouen V, Hognon C, Gallinato O, Colin T, Visvikis D. Evaluation of conventional and deep learning based image harmonization methods in radiomics studies. Phys Med Biol. 2021;66:245009.
62. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data. 2019;6:60.

Publisher's note