IMPROVING RADIOMIC MODEL RELIABILITY AND GENERALIZABILITY USING PERTURBATIONS IN HEAD AND NECK CARCINOMA

TENG XINZHI

PhD

The Hong Kong Polytechnic University

2023

The Hong Kong Polytechnic University

Department of Health Technology and Informatics

IMPROVING RADIOMIC MODEL RELIABILITY AND GENERALIZABILITY USING PERTURBATIONS IN HEAD AND NECK CARCINOMA

TENG XINZHI

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

March 2023

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____(Signed)

Teng Xinzhi (Name of Student)

Abstract

Background: Radiomic models for clinical applications need to be reliable. However, the model reliability is conventionally established in prospective settings, requiring proposal and special design of a separate study. As prospective studies are rare, the reliability of most proposed models is unknown. Facilitating the assessment of radiomic model reliability during development would help to identify the most promising models for prospective studies.

Purpose: This thesis aims to propose a framework to build reliable radiomic models using perturbation method. The aim was separated to three studies: 1) develop a perturbation-based assessment method to quantitatively evaluate the reliability of radiomic models, 2) evaluate perturbation-based method against test-retest method for developing reliable radiomic model, and 3) evaluate radiomic model reliability and generalizability after removing low-reliable radiomics features.

Methods and Materials: Four publicly available head-and-neck carcinoma (HNC) datasets and one breast cancer dataset, in total of 1,641 patients, were retrospectively recruited from The Cancer Image Archive (TCIA). The computed tomography (CT) images, their gross tumor volume (GTV) segmentations, distant metastasis (DM) and local-/regional- recurrence (LR) after definitive treatment were collected from HNC datasets. Multi-parametric diffusion-weighted images (DWI), test-retest DWI scans, pathological complete response (pCR) were collected from breast cancer dataset. For the development of reliability assessment method for radiomic model, one dataset with

DM outcome as clinical task was used to build the survival model. Sixty perturbed datasets were simulated by randomly translating, rotating, and adding noise to the original image and randomizing GTV segmentation. The perturbed features were subsequently extracted from the perturbed datasets. The radiomic survival model was developed for DM risk prediction, and its reliability was quantified with intra-class coefficient of correlation (ICC) to evaluate the model prediction consistency on perturbed features. In addition, the sensitivity analysis was performed to verify the variation between input feature reliability and output prediction reliability. Then, a new radiomic model to predict pCR with DWI-derived apparent diffusion coefficient (ADC) map was developed, and its reliability was quantified with ICC to quantify the model prediction consistency on perturbed image features and test-retest image features respectively. Following the establishment of perturbation-based model reliability assessment (ICC), the model reliability and generalizability after removing low-reliable features (ICC thresholds of 0, 0.75 and 0.95) was evaluated under a repeated stratified cross-validation with HNC datasets. The model reliability is evaluated with perturbation-based ICC and the model generalizability is evaluated by the average train-test area under the receiver operating characteristic curve (AUC) difference in cross-validation. The experiment was conducted on all four HNC datasets, two clinical outcomes and five classification algorithms.

Results: In development of model reliability assessment method, the reliability index ICC was used to quantify the model output consistency in features extracted from the perturbed images and segmentations. In a six-feature radiomic model, the concordance indexes (C-indexes) of the survival model were 0.742 and 0.769 for the training and

testing cohorts, respectively. For the perturbed training and testing datasets, the respective mean C-indexes were 0.686 and 0.678. This yielded ICC values of 0.565 (0.518–0.615) and 0.596 (0.527–0.670) for the perturbed training and testing datasets, respectively. When only highly reliable features were used for radiomic modeling, the model's ICC increased to 0.782 (0.759–0.815) and 0.825 (0.782–0.867) and its C-index decreased to 0.712 and 0.642 for the training and testing data, respectively. It shows our assessment method is sensitive to the reliability of the input. In the comparison experiment between perturbation-based and test-retest method, the perturbation method achieved radiomic model with comparable reliability (ICC: 0.90 vs. 0.91, P-value > 0.05) and classification performance (AUC: 0.76 vs. 0.77, P-value > 0.05) to test-retest method. For the model reliability and generalizability evaluation after removing low-reliable features, the average model reliability ICC showed significant improvements from 0.65 to 0.78 (ICC threshold 0 vs 0.75, P-value < 0.01) and 0.91 (ICC threshold 0 vs. 0.95, P-value < 0.01) under the increasing reliability thresholds. Additionally, model generalizability has increased substantially, as the mean train-test AUC difference was reduced from 0.21 to 0.18 (P-value < 0.01) and 0.12 (P-value < 0.01), and the testing AUCs were maintained at the same level (P-value > 0.05).

Conclusions: We proposed a perturbation-based framework to evaluate radiomic model reliability and to develop more reliable and generalizable radiomic model. The perturbation-based method is a practical alternative to test-retest scans in assessing radiomic model reliability. Our results also suggest the pre-screening of low-reliable radiomics features prior to modeling is a necessary step to improve final model reliability and generalizability to the unseen dataset.

Research Output

1. **Teng, X.**, Zhang, J., Zwanenburg, A., Sun, J., Huang, Y., Lam, S., Zhang, Y., Li, B., Zhou, T., Xiao, H., Liu, C., Li, W., Han, X., Ma, Z., Li, T., & Cai, J. (2022). Building reliable radiomic models using image perturbation. *Scientific Reports*, *12*(1), Article 1. https://doi.org/10.1038/s41598-022-14178-x. (**Published**) (IF: 4.996, 11 / 120 in Multidisciplinary)

2. **Teng, X.**, Zhang, J., Ma, Z., Zhang, Y., Lam, S., Li, W., Xiao, H., Li, T., Li, B., Zhou, T., Ren, G., Lee, F. K.-H., Au, K.-H., Lee, V. H.-F., Chang, A. T. Y., & Cai, J. (2022). Improving radiomic model reliability using robust features from perturbations for head-and-neck carcinoma. Frontiers in Oncology, 12, 974467. https://doi.org/10.3389/fonc.2022.974467. (**Published**) (IF: 5.738, 170 / 360 in Oncology)

3. **Teng, X.**, Zhang, J., Han, X., Sun, J., Lam, S, Ai, Q., Ma, Z., Lee, F. K.-H., Au, K.-H., Yip, C. W.-Y., Chow J., Lee, V. H., Cai, J. (2023). Explainable Machine Learning via Intra-Tumoral Radiomics Feature Mapping for Patient Stratification in Adjuvant Chemotherapy for Locoregionally Advanced Nasopharyngeal Carcinoma. (**Published**)

4. **Teng, X.**, Zhang J., Zhang, X., Fan, X., Lee, E. Y. P., Cai, J. (2023). Virtual Biopsy of HER2 and HR using ADC for Neoadjuvant Chemotherapy Response in Breast Cancer. (**Accepted**)

5. Lam, S.-K., Zhang, Y., Zhang, J., Li, B., Sun, J.-C., Liu, C. Y.-T., Chou, P.-H., **Teng, X.**, Ma, Z.-R., Ni, R.-Y., Zhou, T., Peng, T., Xiao, H.-N., Li, T., Ren, G., Cheung, A. L.-Y., Lee, F. K.-H., Yip, C. W.-Y., Au, K.-H., … Cai, J. (2022). Multi-Organ Omics-Based Prediction for Adaptive Radiation Therapy Eligibility in Nasopharyngeal Carcinoma Patients Undergoing Concurrent Chemoradiotherapy. *Frontiers in Oncology*, *11*. https://www.frontiersin.org/articles/10.3389/fonc.2021.792024

6. Li, B., Ren, G., Guo, W., Zhang, J., Lam, S.-K., Zheng, X., **Teng, X.**, Wang, Y., Yang, Y., Dan, Q., Meng, L., Ma, Z., Cheng, C., Tao, H., Lei, H., Cai, J., & Ge, H. (2022). Function-Wise Dual-Omics analysis for radiation pneumonitis prediction in lung cancer patients. *Frontiers in Pharmacology*, *13*. https://www.frontiersin.org/articles/10.3389/fphar.2022.971849

7. Li, B., Zheng, X., Zhang, J., Lam, S., Guo, W., Wang, Y., Cui, S., **Teng, X**., Zhang, Y., Ma, Z., Zhou, T., Lou, Z., Meng, L., Ge, H., & Cai, J. (2022). Lung Subregion Partitioning by Incremental Dose Intervals Improves Omics-Based Prediction for Acute Radiation Pneumonitis in Non-Small-Cell Lung Cancer Patients. *Cancers*, *14*(19), Article 19. https://doi.org/10.3390/cancers14194889

8. Li, W., Lam, S., Li, T., Cheung, A. L.-Y., Xiao, H., Liu, C., Zhang, J., **Teng, X.**, Zhi, S., Ren, G., Lee, F. K., Au, K., Lee, V. H., Chang, A. T. Y., & Cai, J. (2022). Multi-institutional Investigation of Model Generalizability for Virtual Contrast-Enhanced MRI Synthesis. In L. Wang, Q. Dou, P. T.

Fletcher, S. Speidel, & S. Li (Eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022* (pp. 765–773). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-16449-1_73

9. Liu, C., Li, M., Xiao, H., Li, T., Li, W., Zhang, J., **Teng, X.**, & Cai, J. (2022). Advances in MRI-guided precision radiotherapy. *Precision Radiation Oncology*, *6*(1), 75–84. https://doi.org/10.1002/pro6.1143

10. Sun, H., Ren, G., **Teng, X.**, Song, L., Li, K., Yang, J., Hu, X., Zhan, Y., Wan, S. B. N., Wong, M. F. E., Chan, K. K., Tsang, H. C. H., Xu, L., Wu, T. C., Kong, F.-M. (Spring), Wang, Y. X. J., Qin, J., Chan, W. C. L., Ying, M., & Cai, J. (2023). Artificial intelligence-assisted multistrategy image enhancement of chest X-rays for COVID-19 classification. *Quantitative Imaging in Medicine and Surgery*, *13*(1), 39416–39416. https://doi.org/10.21037/qims-22-610

11. Xiao, H., Ni, R., Zhi, S., Li, W., Liu, C., Ren, G., **Teng, X.**, Liu, W., Wang, W., Zhang, Y., Wu, H., Lee, H.-F. V., Cheung, L.-Y. A., Chang, H.-C. C., Li, T., & Cai, J. (2022). A dual-supervised deformation estimation model (DDEM) for constructing ultra-quality 4D-MRI based on a commercial low-quality 4D-MRI for liver cancer radiation therapy. *Medical Physics*, *49*(5), 3159–3170. https://doi.org/10.1002/mp.15542

12. Xiao, H., **Teng, X.**, Liu, C., Li, T., Ren, G., Yang, R., Shen, D., & Cai, J. (2021). A review of deep learning-based three-dimensional medical image

registration methods. *Quantitative Imaging in Medicine and Surgery*, *11*(12), 4895–4916. https://doi.org/10.21037/qims-21-175

13. Xu, X.-X., Teh, F. C., Lin, C.-J., Lee, J., Yang, F., Guo, Z.-Q., Guo, T.-S., Sun, L.-J., **Teng, X.-Z.**, & Liu, J.-J. (2018). Characterization of CIAE developed double-sided silicon strip detector for charged particles. *Nuclear Science and Techniques*, *29*(5), 1–6.

14. Zhang, J., **Teng, X.**, Lam, S., Sun, J., Cheung, A. L.-Y., Ng, S. C.-Y., Lee, F. K.-H., Au, K.-H., Yip, C. W.-Y., Lee, V. H.-F., Lin, Z., Liang, Y., Yang, R., Han, Y., Zhang, Y., Kong, F.-M. (Spring), & Cai, J. (2023). Quantitative Spatial Characterization of Lymph Node Tumor for N Stage Improvement of Nasopharyngeal Carcinoma Patients. *Cancers*, *15*(1), Article 1. https://doi.org/10.3390/cancers15010230

15. Zhang, L., Yin, F.-F., Li, **T., Teng,** X., Xiao, H., Harris, W., Ren, L., Kong, F.-M. S., Ge, H., Mao, R., & Cai, J. (2021). Multi-contrast four-dimensional magnetic resonance imaging (MC-4D-MRI): Development and initial evaluation in liver tumor patients. *Medical Physics*, *48*(12), 7984–7997. https://doi.org/10.1002/mp.15314

16. Zhang, Y., Lam, S., Yu, T., **Teng, X.**, Zhang, J., Lee, F. K., Au, K., Yip, C. W., Wang, S., & Cai, J. (2022). Integration of an imbalance framework with novel high-generalizable classifiers for radiomics-based distant metastases

prediction of advanced nasopharyngeal carcinoma. *Knowledge-Based Systems*, *235*, 107649. https://doi.org/10.1016/j.knosys.2021.107649

17. Zhang, Y., Yang, D., Lam, S., Li, B., **Teng, X.**, Zhang, J., Zhou, T., Ma, Z., Ying, T.-C. (Michael), & Cai, J. (2022). Radiomics-Based Detection of COVID-19 from Chest X-ray Using Interpretable Soft Label-Driven TSK Fuzzy Classifier. *Diagnostics*, *12*(11), Article 11. https://doi.org/10.3390/diagnostics12112613

Conference abstract

1. **Teng X.**, Zhang J., Ma Z., Han X., Lam S., Xiao H., Liu C., Huang Y., Lee F., Yip W., Cheung A., Lee H., Cai J. (2022). Patient-Specific Adjuvant Chemotherapy Decision Making Using Radiomics for Locally-advanced NPC: A multicenter Retrospective Study. The 64th Annual Meeting & Exhibition of the American Association of Physicists in Medicine (AAPM) (**Oral Presentation**)

Acknowledgement

I would like to take this opportunity to thank all the people who gave me help and support during my Ph.D. period.

Among all, I would like to first express my sincere thanks to my chief supervisor, Prof. Cai Jing, for his kind support, patient guidance, valuable advice, and precious opportunities during the Ph.D. period. For my project, Prof. Cai provided a lot of valuable suggestions and comments from the clinical and practical perspectives, which helped my project become more comprehensive and made sure my project was in the right direction. To make us well-prepared for future challenges, Prof. Cai tried his best to exercise us by all manner of means, such as help us improve the presentation skills and writing skills throughout the group meeting, as well as encourage us sharing our mind during the group meeting. In my daily research, Prof. Cai also provided me with the largest freedom to explore my project. I truly appreciate Prof. Cai for all the support.

Next, I would like to thank my group members, especially Zhang Jiang, Lam Saikit, Liu Chenyang, Xiao Haonan, and Ma Zongrui for their help in my project. When I encountered research problems, Zhang Jiang, Li Wen, Liu Chenyang, and Xiao Haonan were always friendly to share their ideas with me, their suggestions always helped me solve my problems. My thesis would be not possible without their kind help.

The biggest thanks must go to my family members, especially my parents. They always give me the most freedom, encourage me to do what I want to do, and always give me the best unconditionally. I hope I would not let them down and become

someone they can be proud of.

At last, I would like to thank the thesis committee and the external examiners for their valuable time to coordinate and assess my thesis. I hope my research could contribute to the Radiomic community.

Table of Contents

List of Figures

List of Tables

Abbreviations

CT	Computed Tomography
HU	Hounsfield unit
GLCM	gray-level co-occurrence matrix
PET	Positron Emission Tomography
MRI	Magnetic Resonance Imaging
VIM	International Vocabulary of Metrology
ROI	Region of Interest
GTV	Gross Tumor Volume
IBSI	Image Biomarker Standardization Initiative
3D/4D	Three-/four-dimensional
CAD	Computer Aided Detection
C-index	Concordance Index

ICC	Intra-class Coefficient of Correlation
TR	Repetition Time
TE	Echo Time
AUC	Area under the Receiver Operating Characteristic Curve
LR	Local-/Regional- Recurrence
DM	Distant Metastasis
HNC	Head-and-Neck Carcinoma
OPC	Oropharyngeal Carcinoma
SVC	Support Vector Classifier
KNN	K-nearest Neighbors Algorithm
MLP	Multilayer Perceptron Network
AP	Anterior-Posterior
SI	Superior-Inferior

LM	Lateral-Medial
TCIA	The Cancer Imaging Archive
DECT	Dual Energy Computed Tomography
IRB	Institutional Review Board
RQS	Radiomics Quality Score
pCR	Pathological Complete Response
HNSCC	Head and Neck Squamous Cell Carcinoma
Dice	Dice Similarity Coefficients
ADC	Apparent Diffusion Coefficients

Chapter 1. Literature Review

1.1. Introduction

Radiomic is a flourishing field utilizing machine learning to associate cancer imaging phenotype to cancer genotype or clinical outcome for precision medicine [1–3]. The number of publications in radiomics rose dramatically in recent years [4]. Radiomics strives to characterize differences in tumor phenotypes based on non-invasive medical images such as computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET) with hand-crafted features. Also, it has demonstrated that it can capture the heterogeneity of a tumor [5], associates it with tumor characteristics for diagnosis [6] and treatment prognostication [7], and improve the overall decision-making during the treatment [8]. Despite the great potential of radiomics, the reliability and generalizability of radiomic models are the major concerns taking radiomics from bench to bedside. This study, therefore, attempts to facilitate the reliability and generalizability of radiomic models with simulated perturbations.

1.2. Clinical Value of Radiomics

Radiomics is a new research regime for the field of radiation oncology, derived from a combination of 'radio-' meaning medical imaging and '-omic' meaning collective characterization and quantification of the object of interest [9]. Radiomics is an analytic tool containing feature extraction [10] and analysis with machine learning techniques. Feature extraction refers to a large number of features [11] are collectively extracted

from tomographic images, conversion of digital medical images into mineable high-dimensional data. With a large amount of data, machine learning techniques are implemented to perform inferences based on advanced computational algorithms [12]. The relationship between radiomic feature extraction and machine learning is dependent on each other. A large amount of the features requires an analytical tool to empower it from data to information. Machine learning, empowered by computer technology advances, is designed to infer information from available data and predict outcomes accurately.

In summary, radiomics is an analytical tool in the medical imaging field. The high-throughput exploitation of quantitative image features from routinely acquired medical imaging enables data to be extracted and applied with machine learning techniques within clinical-decision support systems.

Radiomics has become an uprising research regime involving medical imaging, such as radiology and radiation oncology [2]. Conventionally, the medical imaging data is evaluated visually or qualitatively, for example, the staging of the tumor, and it leaves a large amount of data unused [10]. With the leverage in image feature quantification and advances in machine learning, the value of a large amount of unused image data has opportunities to be mined.

A typical and simplified scenario for explaining the need for radiomics would be that two lung cancer patients end up with different survival despite having similar TNM staging, histology, and ages, shown in **Table 1**. TNM staging, histology, and ages are the common prognosis factors for oncologists to decide the treatment plan and predict

the patient's prognosis after treatment. However, the current characterization of the tumor and patients are not enough to provide sufficiently accurate information to predict patient outcome. It turns out that the unmined and vast amounts of medical image data may hide some useful information, which may provide different patient stratifications with the information hidden in the images. One of the earliest radiomic papers [13] showed the possibility of it, *Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach*. It shows that cancer represents strong phenotypic differences that can be visualized non-invasively by medical imaging. Phenotype means a set of observable characteristics of an organism ("Phenotype," 2021). The tumor phenotypes can be decoded with the extraction of radiomics features. **Figure 1** shows the CT images with tumor segmentation in the axial view, the shape of tumors, Hounsfield unit (HU) value intensity, and texture differences. Such differences have the potential to stratify patients into different subgroups and optimize the treatment for each subgroup for an optimal outcome. Machine learning techniques such as a decision tree [13] can infer useful information from quantified data. Consequently, these data could be able to be used to support the clinical decision. Stratification of patients into subtypes for better outcomes is precision medicine. The concept of precision medicine is gaining popularity, and radiomics is a potential way to achieve it by having a deep look at the image data.

Table 1. The example of insufficient biomarker for survival prediction of clinical characteristics of non-small cell lung cancer patients. Two non-small cell lung cancer patients with similar age, same TNM staging, histology, and gender, yet end up with different survival. The conventional clinical characteristics did not show consistent risk stratification between two patients.

ID	**Age**	**T**	**N**	**M**	**Overall Stage**	**Histology**	**Gender**	Survival (days)	Survival Status
A	71	4	3	0	IIIb	SCC	female	2119	alive
B	62	4	2	0	IIIb	SCC	female	261	dead

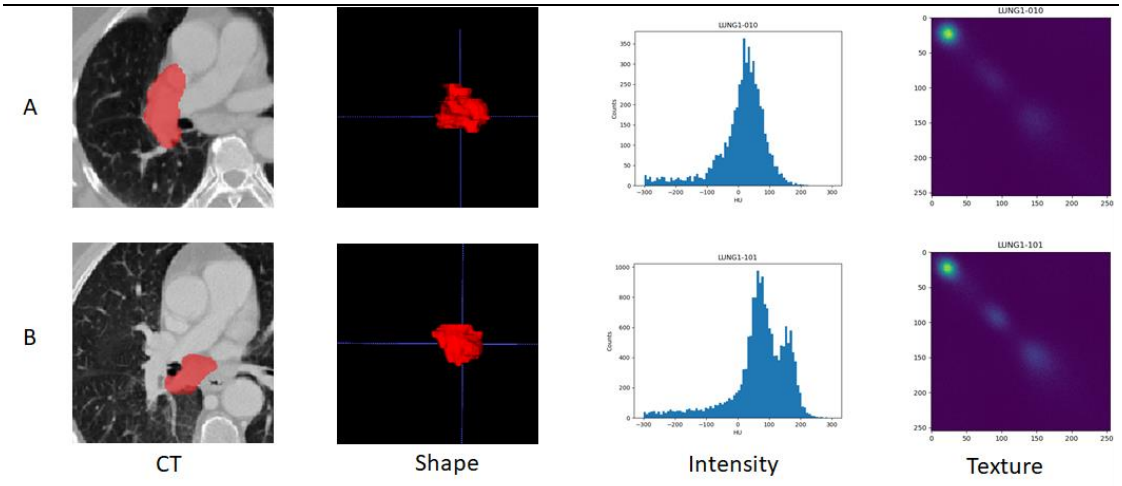

Figure 1. The demonstration of axial CT slices, tumor contour shape, histogram of CT HU value within the segmentation, and GLCM visualized with heatmap. Conceptually, the shape, histogram and GLCM texture maps are different in two patients. The comparison visualized the potential of quantitative image feature characterization for patient risk stratification.

1.3. Challenges in Radiomics Workflow

Although radiomics is gaining popularity and publications have been exploded in recent years, the challenge of radiomics is the reliability of radiomic models [14]. Radiomics starts with quantifying features from medical images within a region of interest (ROI) with handcrafted definitions [15,16]. The quantification is intrinsically sensitive to the variations on the image and ROI, and the workflow of radiomics inevitably introduces variations into medical images from image acquisition to processing of the image. Therefore, the reliability of radiomics has gained awareness since the very beginning of this field, for example, the first radiomic-based article by Hugo et al. intentionally utilized a set of test-retest image data, RIDER Lung CT [17], to evaluate the repeatability and reproducibility of the radiomics features. However, the nature of medical imaging, expensive and harm to patient, limited the acquisition of test-retest data, and the intensive repeated scans are not viable due to the radiation dose of CT and PET to the patient and the long scanning time of MRI. Furthermore, many factors are affecting the radiomic feature value precisions, such as imaging protocol and acquisition parameters. Therefore, the repeatability and reproducibility of radiomics features were also intensively studied in recent years. The factors affecting radiomics features' reproducibility such as imaging protocol and acquisition parameters are categorized as controllable factors [18], meaning that with a well-controlled scanner and image acquisition protocol, the factors can be minimized. However, in real life, a standard image acquisition protocol and reconstruction algorithm is difficult to achieve across institutions since such variations have no significant impact on the routine

function of medical images.

Radiomics workflow involves several independent steps, **Figure 2**, and each step has its challenges. In the following sections, the variations in each step of the radiomics workflow were review and discussed. These steps are (a) image acquisition and reconstruction, (b) segmentation of the ROI, (c) feature extraction, and (d) radiomic modeling.

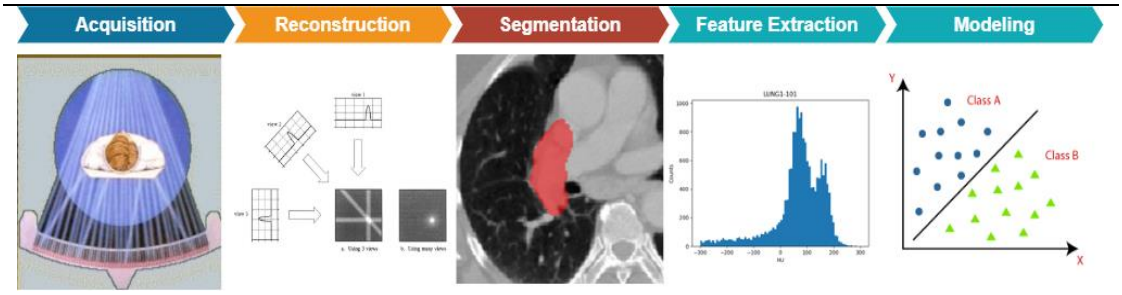

Figure 2. The complete workflow of Radiomics, including image data acquisition, 3D volume reconstruction, ROI segmentation, feature extraction, and modeling with machine learning algorithms. Each step consists of variation sources and may affect the reproducibility and repeatability of extracted features as well as the output consistency of radiomic models.

1.3.1. Challenges in Image Acquisition and Reconstruction

The widely used imaging modalities in current radiomic studies are CT, MR, and PET, allowing for image acquisition and reconstructions. The standardization of the image acquisition and reconstructions protocols is typically different from institution to institution. The lack of imaging protocol consistency does not affect the conventional diagnosis of medical imaging. In contrast, differences in acquisition and reconstruction protocols would affect radiomics since radiomics involves quantification of the images on the voxel level [19], which affects the image noise and textures. Such variability in imaging and reconstruction would affect inconsistent prediction results when validating developed radiomic models using other institutions' datasets. The reconstruction algorithms [20] would also affect the quantification of the images and, therefore, affect the predictive model's performance.

Institution-independent features is the starting point to build a model which is more likely to be generalized between institutions. Study [21] intensively scanned the phantom, shown in **Figure 3**, with 17 CT scans with varied manufacturers and thoracic imaging protocols. They reported the observation of variation in radiomics features for different acquisition parameters.The studies are often focused on single tumor site, and single modality, the generalizability of the reported feature reliability to the dataset of interest is still unknown. Paper [22] conducted a comprehensive study in the feature reliability against multi-center and multi-vendor using apparent diffusion coefficient (ADC) maps. A study [23] conducted experiments on three settings, reporting the radiomic feature reproducibility and repeatability on T2-weighted MRI of cervical

cancer patients. These publications thoroughly investigated the impact of acquisition parameters on radiomics features, and the reliability of each feature has been reported.

Figure 3. The Credence Cartridge Radiomics phantom with 10 cartridges for radiomic feature robust analysis against scanners, image acquisition protocols, and reconstruction algorithms. Each cartridge contains unique textures. The top four cartridges are 3D printed ABS plastic with 20% to 50% honeycomb fill, which provide regular and periodic textures. Then, the following cartridges are sycamore wood, cork, extra dense cork, solid acrylic, natural cork, and plaster resin. The lower six cartridges provide natural textures for the radiomic. The image is adapted from TCIA [24].

Literatures on the image acquisition and reconstruction was tabulated in **Table 2**. In total, 38 literatures performed experiments with scanners and analyzed the impact of inter-scanner, test-retest, image acquisition parameters, and reconstruction algorithms on the reproducibility of radiomics features. Most literatures identified that the texture features are more vulnerable to variations in the image acquisition and reconstruction than intensity (or first order) features.

The major limitation of the reviewed literatures is that their result is hardly be applied to the clinical-oriented studies. In clinical-oriented studies, the feature reliability needs to be quantified and plays a role in feature selection. A well-known study [3] used feature reliability index and feature relevant outcome index to rank the radiomics features

The first reason is that majority of the studies did not tabulate the feature reproducibility index for individual features. The second reason is that the scanner is required to perform such study, which is a major barrier to most research groups.

Table 2. The literature investigating the image acquisition and reconstruction on CT and MRI.

Author	Year	Sites	Modalities	Sources of variation	Feature categories
Cabini [25]	2022	Lung	CT	Image acquisition parameters	Shape, Intensity, GLCM, GLRLM, NGLDM, GLZLM
Carbonell [26]	2022	Liver	T1-w MR, T2-w MR, ADC	1.Test-retest repeatability 2. Inter-scanner 3. Inter-observer segmentation	Shape, Intensity, GLCM, GLSZM, GLDM, GLRLM, NGTDM
Chen [27]	2021	Hematoma	CT	1. Test-retest repeatability 2. Image acquisition parameters	Intensity, GLCM, GLRLM, NGTDM
Chen [28]	2022	Phantom	CT	1. Test-retest repeatability 2. Scanning modes 3. Inter-scanners	Intensity, GLCM, GLDM, GLRLM, GLSZM, NGTDM
Crombe [29]	2021	Abdomen	T2-w MR	T2-w acquisition methods	Intensity, GLCM, GLRLM, NGLDM, GLZLM
Denzler [30]	2021	NSCLC, MPM, SSc-ILD Lung	CT	Reconstruction kernels	Intensity, GLCM, NGTDM, GLRLM, GLSZM, NGLDM
Emaminejad [31]	2021	CA Lung Lung	CT	1. Dose level variation 2. Reconstruction kernel 3. Slice thickness variation	Intensity, GLCM, GLRLM, GLSZM, NGTDM, GLDM
Euler [32]	2021	Phantom	CT	1. Image acquisition parameters 2. Radiation dose 3. DECT approach	Shape, Intensity, GLCM, GLRLM, GLSZM, NGTDM, GLDM
Fiset [23]	2019	Cervix	T2-w MR	1. Test-retest repeatability 2. Acquisition protocols 3. Inter-observer segmentation	Shape, Intensity, GLCM, GLSZM GLDM, GLRLM, NGTDM
Gao [33]	2022	Pulmonary nodules Lung	CT	1. Radiation dose 2. Reconstruction kernels	Shape, Intensity, GLCM, GLSZM GLDM, GLRLM, NGTDM
Granzier [34]	2022	Breast (Healthy)	T1-w MR, T2-w MR, ADC	Test-retest repeatability	Intensity, GLCM, GLSZM GLDM, GLRLM, NGTDM

Ibrahim [35]	2021	HCC Liver	CT	Imaging phases	shape, Intensity, GLCM, GLDM, GLRLM, GLSZM, NGTDM
Ibrahim [36]	2021	Phantom	CT	1. Inter-scanners 2. Scanning parameters	Shape, Intensity, GLCM, GLDM, GLRLM, GLSZM, NGTDM
Lee [37]	2021	Phantom	T1-w MR and T2-w MR	1. MRI scanning protocols parameters 2. Scanner types	Intensity, GLCM, GLRLM, NGDTM
Lennartz [38]	2022	Phantom & human (Abdomen)	DECT	Inter-scanners	Intensity, GLCM, GLDM, GLRLM, GLZLM
Mahon [39]	2019	NSCLC Lung	4DCT, T1-w MR	Test-retest repeatability	Intensity, GLCM, GLRLM, GLSZM, NGTDM
McHugh [40]	2021	Colorectal Cancer Liver Metastases	T1-w MR, T2-w MR, qT1-wMR	1. MR sequences 2. Pre- and post-contrast 3. Image normalization	Shape, Intensity, GLCM, GLDM, GLRLM, GLSZM, NGTDM
Meyer [41]	2019	Metastatic liver lesions	CT	1. Radiation dose 2. Reconstruction settings	Shape, Intensity, GLCM, GLDM, GLRLM, GLSZM, NGTDM
Mitchell-Hay [42]	2022	Brain	T1-w MR	1. Inter-scanner 2. Test-retest repeatability for weeks	Intensity, GLCM, GLSZM, GLRLM, NGTDM
Nazeri [43]	2021	Brown adipose tissue	CT (and PET)	Test-retest repeatability within 14 days	Shape, Intensity, GLCM, GLSZM, GLRLM, GLDM
Pandey [44]	2020	Healthy Brain	T2-w MR	1. Age and gender 2. Test-retest repeatability 3. Inter-scanner	Intensity, GLCM, GLSZM, GLRLM, GLDM
Perrin [45]	2018	Liver malignancy	CE-CT	1. Contrast injection rate 2. pixel resolution 3. Scanner model	Intensity, GLCM
Prayer [46]	2020	Fibrosing interstitial lung disease Lung	CT	1. Test-retest repeatability 2. Inter-scanner	Intensity, GLCM, GLRLM, GLSZM, GLDM
Raisi-Estabragh [47]	2020	Healthy volunteer, Myocardial infarction	MR	Different Heart pathology Multi-centre & multi-vendor test-retest	Shape, Intensity, GLCM, GLRLM, GLSZM, NGTDM, GLDM

Refaee [48]	2022	Phantom	CT	Reconstruction kernel normalization	Intensity, GLCM, GLRLM, GLSZM, GLDM, NGTDM
Reiazi [49]	2021	Oropharyngeal Cancer Oropharynx	CT	Inter-scanners	Shape, Intensity, GLRLM, GLSZM, GLCM, GLDM, NGTDM
Rinaldi [50]	2022	NSCLC Lung	CT	1. Tube voltage, scanner model 2. Reconstruction algorithm	Shape, Intensity, GLCM, GLRLM, GLSZM, GLDM, NGTDM
Sanchez [51]	2021	Liver Tumor & Muscle Liver	CT	1. Voxel sizes 2. Reconstruction slice thickness	Shape, Intensity, GLCM, GLRLM, GLSZM, GLDM, NGTDM
Sun [52]	2022	jellies, fruit/vegetables, phantom	T2-w MR	1. Test and retest repeatability 2. Inter-observer segmentation 3. Resampling on slice thickness	Shape, Intensity, GLCM, GLRLM, GLSZM, NGTDM, GLDM
Xue [53]	2021	CA Prostate (PSA)	T2-w MR	Inter-scanners	Intensity, GLCM, GLDM, GLRLM, GLSZM, NGTDM
Xue [54]	2020	Phantom	T2-w MR	Image reconstruction settings	Intensity, GLCM, GLDM, GLRLM, GLSZM, NGTDM
Alis [55]	2020	Heart	MR	1. Inter-observer reproducibility of radiomics features 2. Cardiac cycle	Intensity, GLCM, GLRLM, GLDM, GLSZM
Bologna [56]	2019	Virtual Phantom	T1-w MR	Acquisition parameters, TR and TE	Intensity, GLCM, GLRLM
Hoebel [57]	2021	Glioblastoma Brain	T1-w MR	1. Test-retest repeatability 2. Normalization strategy 3. Image intensity quantization	Shape, Intensity, GLCM
Hu [58]	2022	Phantom + Human (Healthy)	T1-w, fluid-attenuated T1-w, T2-w MR	Integrated parallel acquisition technologies	GLDM, GLCM, GLSZM, NGTDM
Lee [59]	2022	Abdominal phantom with liver nodules	CT	1. Reconstruction protocols 2. Reconstruction kernels	Intensity, GLCM, GLRLM, GLSZM, GLDM
Muenzfeld [60]	2021	3D printed anthropomorphic	CT	Reconstruction kernels	Intensity, GLCM, GLRLM, GLDM, GLSZM

Whitney [61]	2021	CA Breast	T1-w DCE MR	Field strength of MRI	Shape, Intensity

1.3.2. Challenges in Segmentation

Segmentation of the ROI is the next critical step in radiomics since the radiomics features are primarily extracted from the region that interests us. Majority of radiomic studies in radiation oncology use the visible tumor volume or the gross tumor volume (GTV) [62]. The manual segmentations by the oncologist are considered the gold standard. However, besides being time-consuming and labor-intensive with manual contour, the segmentations are subject to inter-observer and intra-observer variations. **Figure 4** shows the variability of the GTV segmentation by oncologists on CT images of prostate cancer. The inconsistency between different oncologists on the medical images would lead to variability in extracted radiomics features.

Besides the manual segmentations, the automatic or semi-automatic method for tumor volume delineation is considered better than manual segmentation in terms of stability and cost-effectiveness. Publication by [63] shows stable feature extraction when using automatic segmentation than manual segmentation with non-small cell lung cancer patients. The limitations of auto-segmentation are also apparent in two ways. Firstly, the auto segmentation is primarily applied to a simple scenario such as lung cancer or prostate cancer, where the anatomic structure is simple and the contrast of the tumor to surrounding tissue is significant. For a more complex case such as head-and-neck carcinoma (HNC), the accuracy of the auto segmentation will not be as good as manual segmentation, thus the scope of auto segmentation is limited.

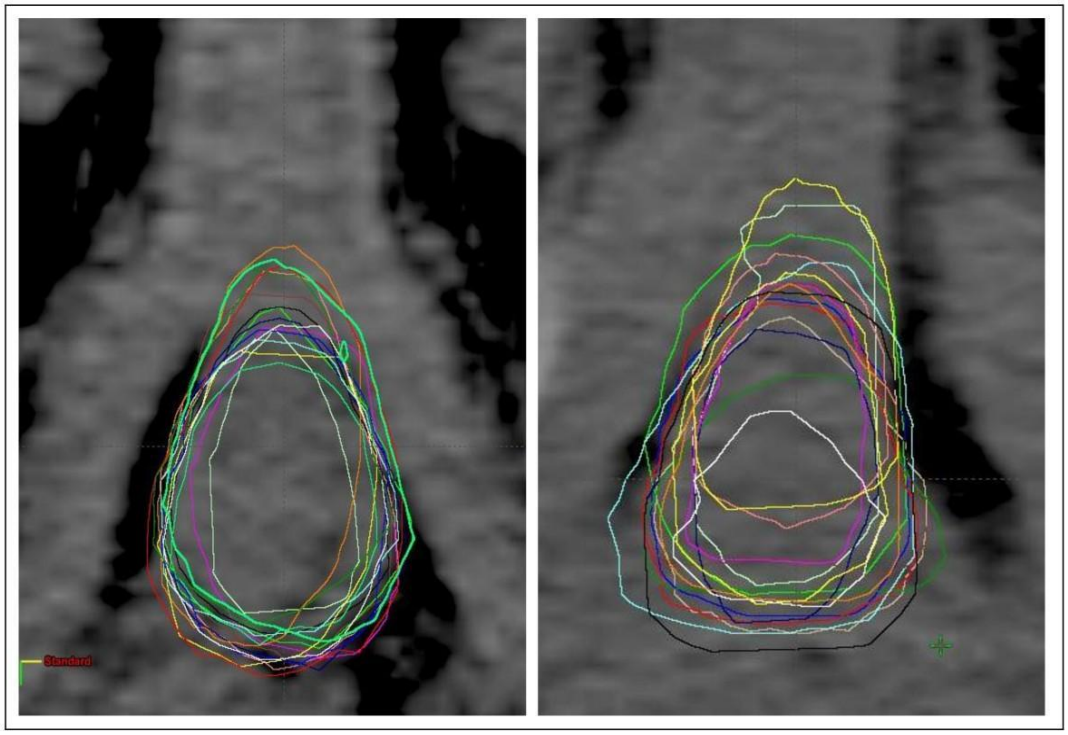

Figure 4. A demonstration of the variability of contour inconsistency by different oncologists or inter-observer variability on penile bulb. The images are the CT image central slice with manual segmentation on penile bulb by different oncologists on two patients [64].

In order to minimize the effects of variations of radiomic models from inter-observer segmentation, some studies assessed the feature variability with either with multiple segmentations by different oncologists (inter-observer variability) or same oncologist (intra-observer variability). Besides the labor-intensive feature reproducibility assessment method, Zwanenburg et al.[65] proposed a super-voxel approach to estimate the possible segmentation variations. The limitation of both methods is that it does not account for the variations in images. For manual segmentation, this method is labor-intensive and may cause difficulties when applying such a method in research.

Table 3 tabulated 25 studies focusing on the feature reproducibility study by segmentation variability. Majority of the study focuses on the impact of inter-observer variability on radiomics features, some studies also studied the impact of intra-observer variability, only few of them studied the impact of inter-observer variability with combination of test-retest images. The literatures of ROI variability on radiomics features shared similar limitation as we discussed in section 1.3.1 that most studies did not explicitly tabulate their quantitative results and clinical-oriented study cannot directly use their results. Furthermore, it is labor-intense for each radiomic study to carry out an analysis on the impact of inter-observer variability on radiomics features.

Table 3. Literatures investigating the impact of segmentation variability on radiomic feature reproducibility and repeatability.

Author	Year	Site	Modalities	Sources of variation	Feature category
Bianconi [66]	2021	Lung Lesions	CT	1. Inter-observer variability 2. Image quantization method	Intensity, GLCM, GLDM, GLRLM, GLSZM, NGTDM
Carbonell [26]	2022	Liver	T1-w MR, T2-w MR, ADC	1.Test-retest repeatability 2. Inter-scanner 3. Inter-observer segmentation	Shape, Intensity, GLCM, GLSZM, GLDM, GLRLM, NGTDM
Chen [67]	2021	CA Cervix	DWI	1. Inter-observer segmentation 2. Intra-observer segmentation	Shape, Intensity, GLCM, GLDM, GLRLM, GLSZM, NGTDM
Duan [68]	2022	HCC Liver	CT, T1-w MR, T2-w MR	Inter-observer segmentation	Intensity, GLCM, GLDM, GLRLM, GLSZM, NGTDM
Fiset [23]	2019	Cervix	T2-w MR	1. Test-retest repeatability 2. Acquisition protocols 3. Inter-observer segmentation	Shape, Intensity, GLCM, GLSZM GLDM, GLRLM, NGTDM
Gitto [69]	2021	Cartilaginous bone tumors	CT, T1-w MR, T2-w MR	Inter-observer segmentation	Shape, Intensity, GLCM, GLSZM GLDM, GLRLM
Gitto [70]	2022	Spine bone tumor	T2-w MR	Small geometrical transformations of the ROIs	Shape, Intensity, GLCM, GLSZM GLDM, GLRLM, NGTDM
Granzier [71]	2020	Bresat (Maligant)	DCE T1-w MR	Inter-observer variability in VOIs segmentation	Shape, Intensity, GLCM, GLSZM GLDM, GLRLM, NGTDM
Haarburger [72]	2020	Lesions and tumors of Lung, Kidney, Liver	CT	1. Inter-observer segmentation 2. Manual vs. automated segmentations	Shape, GLCM, GLSZM, GLRLM, NDGTM
Haniff [73]	2021	HCC Liver	T1-w MR	segmentation method (semi-automatic vs manual)	Shape, Intensity, GLCM, GLDM, GLRLM
Jensen [74]	2021	Phantom	CT, T1-w MR, T2-w MR	Size of ROI	Intesnity, GLCM, GLRLM, GLSZM, GLDM, NGTDM

Kelahan [75]	2022	Liver metastasis from CA colorectal	CT	Size of ROI	Intensity, GLCM, GLRLM, NGLDM, GLZSM
Kocak [76]	2019	Clear cell renal cell carcinoma Kidney	CT	1. Inter-observer segmentation 2. Intra-observer segmentation	Intensity, GLCM, GLDM, GLRLM, GLSZM, NGTDM
Le [77]	2021	Carotid H&N	CT	1. ROI Segmentation 2. ROI Perturbations 3. Pre-Processing 4. Image discretization	Intensity, GLCM, GLDM, GLRLM, GLSZM, NGTDM
Müller-Franzes [78]	2022	Tumour of Lung, Liver, Kidney, Brain	CT, FLAIR MR	1. Inter-observer segmentation 2. Intra-observer segmentation	Shape, Intensity, GLCM, GLSZM, GLRLM, GLDM
Schurink [79]	2022	CA Rectum	T2-w MR and ADC	1. Inter-observer variability in VOIs segmentation 2. Feature extraction software	Shape, Intensity, GLCM, GLRLM, GLSZM, NGTDM
Sun [52]	2022	jellies, fruit/vegetables, phantom	T2-w MR	1. Test and retest repeatability 2. Inter-observer segmentation 3. Resampling on slice thickness	Shape, Intensity, GLCM, GLRLM, GLSZM, NGTDM, GLDM
Tixier [80]	2019	Glioblastoma Brain	T1-w MR, FLARE MR	Inter-observer segmentation	GLCM, GLSZM
Tunali [81]	2019	CA Lung (Peritumoral regions of lesions)	CT	ROI segmentation	Shape, Intensity, GLCM, GLRLM, GLSZM, NGTDM
Urraro [82]	2021	CA Prostate (PSA)	T2-w MR, ADC	Inter-observer segmentation	Shape, Intensity, GLCM, NGLDM, GLRLM, GLZLM
Wang [83]	2020	Stomach (Gastric Cancer with liver metastasis)	CT	1. Intra-observer segmentation 2. Inter-observer segmentation	Shape, Intensity, GLDM, GLCM, GLRLM, GLSZM, NGTDM
Wong [84]	2021	Phantom	T1-w MR	1. Intra-observer segmentation 2. Inter-observer segmentation 3. Test-retest repeatability	Shape, Intensity, GLCM
Alis [55]	2021	Heart	MR	1. Inter-observer reproducibility of radiomics features 3. Cardiac cycle	Intensity, GLCM, GLRLM, GLDM, GLSZM
Gruzdev [85]	2020	Pancreatic neuroendocrine neoplasms	CT	1. Intra-observer segmentation 2. Inter-observer segmentation	Not Specified

Konik [86]	2021	Cystic renal masses Kidney	CT	Inter-observer segmentation	Intensity, GLCM, GLRLM, GLSZM, NGTDM

1.3.3. Challenges in Image Preprocessing

With the images and segmentations of ROI, the features can be extracted from the image within the ROI. There are primarily four common radiomic feature extraction platforms, PyRadiomics, LIFEx, CERR and IBEX. Publication [87] shows that the feature reliability is highly dependent on the choice of feature extraction platform as well as the parameters associated with feature calculation. Despite an international collaboration in the standardization of the feature calculation [88], it does not address the parameters associated with feature calculations, such as resampling, image interpolation algorithms, and bias correction in MR images. These extraction parameters would also affect the feature values.

The image biomarker standardization initiative (IBSI) [88], an independent international collaboration, has been proposed to standardize the definition and implementation of qualitative image features, and a calibration dataset has been provided for image feature consensus calculation. This initiative is a comprehensive project with collaborations of 25 research teams using different software, and finally, more than 97% of the features studied reached excellent reproducibility. It shows the reduction of feature value after standardization between different extraction software. Although the IBSI provided guidelines in calculating radiomics features, it does not standardize the features from filtered images, such as the Log-Gaussian filter and wavelet filter image. Both filtered images have been proved to be useful in radiomic studies.

Table 4 tabulated 13 literatures focusing on the impact of image preprocessing on

radiomics features. Most literatures study the impact of image discretization method, some literatures study the impact of normalization methods, and few studies focus on impact of resampling method on the images. The literatures of the image preprocessing algorithms also shared similar limitations as shown in Sections 1.3.1 and 1.3.2, however, the barrier of replicate similar variations is not as high as studying image acquisition variability and ROI variations.

Table 4. Literatures investigating the impact of image preprocessing on radiomic feature reproducibility and repeatability.

Author	Year	Site	Modalities	Sources of variation	Feature category
Duron [89]	2019	lachrymal gland, breast	T1-w MR, T2-w MR	Image discretization methods	GLCM, GLSZM, GLRLM, GLDM, NGTDM
Fornacon-Wood [87]	2020	CA H&N CA Lung	CT	Different feature extraction platforms	Shape, Intensity, GLCM, NGTDM
Gao [33]	2022	Pulmonary nodules Lung	CT	1. Radiation dose 2. Image preprocessing	Shape, Intensity, GLCM, GLSZM GLDM, GLRLM, NGTDM
Hoebel [57]	2021	Glioblastoma Brain	T1-w MR, T2-w MR	Pre-processing techniques	Shape, Intensity, GLCM
Ibrahim [90]	2021	Phantom	CT	1. In-plane spatial resolution, 2. Interpolation and resampling	Intensity, GLCM, GLRLM, GLSZM, GLDM, NGTDM
Li [91]	2020	Phantom	CT	Image preprocessing parameters [re-segmentation, width of bins]	Shape, Intensity, GLCM, GLRLM, GLSZM
McHugh [40]	2021	Colorectal Cancer Liver Metastases	T1-w MR, T2-w MR, qT1-w MR	1. MR sequence 2. contrast enhancement pre- and post-contrast 3. Normalization	Shape, Intensity, GLCM, GLRLM, GLSZM, GLDM, NGTDM
Moradmand [92]	2020	Glioblastoma Brain	T1-w MR, T2-w MR, FLARE	1. Intensity inhomogeneity correction [N4Bias Correction] 2. Noise filtering [SUSAN Denoise]	Shape, Intensity, GLCM, GLRLM, GLSZM, GLDM, NGTDM
Scalco [93]	2020	CA Prostate	T2-w MR	Image normalization techniques	Intensity, GLCM, GLRLM, GLSZM, GLDM, NGTDM
Schwier [94]	2019	CA Prostate	T2-w MR, ADC	1. Image normalization techniques 2. Image filters 3. Image discretization	Shape, Intensity, GLCM, GLRLM, GLSZM, NGTDM
Sun [52]	2022	jellies, fruit/vegetables, phantom	T2-w MR	1. Test and retest repeatability 2. Inter-observer segmentation 3. Resampling on slice thickness	Shape, Intensity, GLCM, GLRLM, GLSZM, NGTDM, GLDM

Hoebel [57]	2021	Glioblastoma Brain	T1-w MR, FLARE	1. Test-retest repeatability 2. Normalization strategy 3. Image intensity quantization	Shape, Intensity, GLCM
Simpsons [95]	2020	Phantom + Human	MR	Image discretization methods	GLCM, GLRLM, GLSZM, NGTDM

1.3.4. Challenges in Modeling

The radiomic modeling includes two parts, feature selection and model building. The feature selection aims to reduce dimension of high-dimensional radiomics features. The feature selection follows two principles: 1) the selected features need to be correlated with the outcome and 2) the selected features need to be less correlated with each other. The modeling building aims to use the advanced machine learning algorithms for identify the optimal model maximizing the correlation with outcome.

With the extraction of radiomics features, advanced machine learning techniques were used to infer useful information and help the clinical decision. However, different feature selection and modeling methods would affect the performance of the model in terms of its sensitivity. The study by Parmar et al. [96] showed various feature model performances on an unseen testing cohort with different feature selection and classifiers. However, limited publications focused on feature selection and modeling methods comparison in the field of radiomics. There are still no universal modeling methods, and the optimal modeling selection method may differ dataset to datasets [97]. It is worth noting that the variations in feature selection and model building are different from the variations in image acquisition and reconstruction section 1.3.1, segmentation section 1.3.2 and image preprocessing 1.3.3. The variations in feature selection and model building are the varied radiomic model performance with different feature selection and model building methods, while the previous sections discussed the feature value variations under different settings. The variations in feature selection and model building are unlikely to be the issue as the variations of selected features reflects that

different feature selection methods focused on different aspects of data.

1.3.5. Summary of Current Challenge

The fundamental question to ask is that can radiomic model have an output with a reasonable error range for routine clinical application. As discussed in previous sections, multiple sources of variation in each step of radiomic workflow create an essential weakness of radiomics. This weakness was recognized in the very early days of radiomics, and **Figure 5** visualizes the variations in each step. Despite an explosive increase in the radiomics literature, researchers frequently failed to adequately consider sources of variation and report the reliability individual radiomics features. The concerns on reliability and reproducibility slow the pace of innovation in radiomics and limit its translational potential.

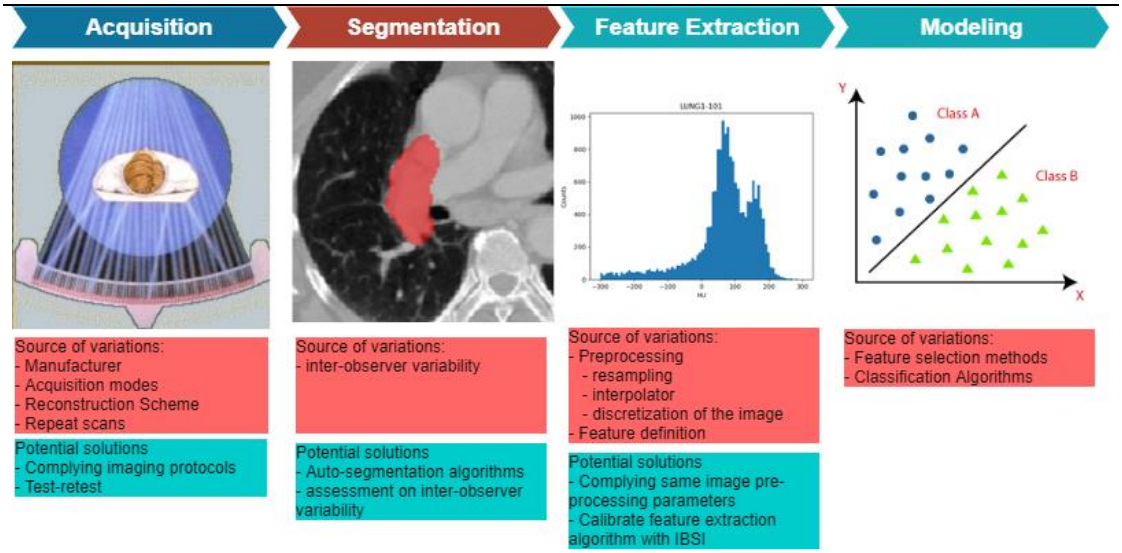

Figure 5. Summary of variations sources in each step of radiomics workflow.

Chapter 2. Research Objectives

2.1. Research Gap

A key goal of clinically-oriented radiomics studies is to build a radiomic score that can reliably describe tumour characteristics using medical images. The first step towards achieving this goal is to identify reliable features that can be used under different circumstances. Despite numerous publications on radiomic feature reproducibility, the lack of a feature reliability index for individual features has prevented researchers from directly referencing previous results and building a reliable radiomic model. Test-retest reliability measurements, achieved by scanning patients twice, are among the most straightforward methods for evaluating radiomic feature reliability. However, it is near impossible to perform test-retest scans for every radiomics study, as extra scans incur additional costs and can harm patients. Several practical methods have been proposed to evaluate the feature reliability of radiomics studies. For example, radiomic feature reliability can be evaluated using already available test-retest scans. RIDER lung [17], is an available test-retest dataset that can be accessed through The Cancer Image Archive (TCIA). The limitation of this method that feature reproducibility may not be generalisable across diseases and imaging modalities. Timmeren et al. has raised concerns about generalisability across datasets as well [98]. Feature reproducibility can also be evaluated using multiple segmentations of the same medical image, acquired by several experienced oncologists [64]; many clinically-oriented radiomics studies have already adopted this method. However, the method is labour-intensive and can only be used to evaluate the effect of ROI on feature

reliability. Therefore, a method to assess dataset-specific radiomic feature reliability is warranted.

However, a comprehensive evaluation of radiomic feature reliability is challenging due to the limited availability of imaging resources. A review by Zhao et al. [18] discussed the sources of radiomic variations and posited that these sources could be categorised into controllable and random factors. Controllable factors are sources that can be minimised by retaining the same parameters for each step of the radiomics workflow. For example, during image acquisition, the manufacturer, imaging protocol, reconstruction kernel, reconstruction slice thickness, reconstruction algorithm and reconstruction algorithm parameters can be controlled. During feature calculation, the resampling resolution, image interpolation algorithm, discretisation and compliance with internationally recognised standards for feature value calculation can be maintained. These variabilities can be minimised given sufficiently transparent reporting for each step. Recognition of the need to evaluate the scientific integrity and clinical utility of radiomics studies led to the development of a radiomics quality score (RQS) in 2017 [99]. The RQS comprises a checklist that researchers can use to score radiomics studies based on predefined guidelines. In contrast to controllable factors, random factors are sources of variations that cannot be minimised by controlling reporting details as they are intrinsic to the medical imaging process. For example, during image acquisition, patient position may differ across scans, leading to slight translations or rotations that affect image intensity values and, in turn, the feature values. During segmentation, uncertainties are unavoidable due to ambiguous tumour boundaries and limited knowledge of tumour micro-aggression. For instance, one study

showed that the GTV contours of prostate cancers could not be exactly reproduced by the same oncologist in a one-week interval [71]. The above evidence strengthens the categorisation of the sources of variability into random and controllable factors. The evaluation of random factors is more essential than that of controllable factors.

Recently, a simulation-based method to evaluate feature reliability was proposed by Zwanenburg et al. [65]. They suggested applying linear transformations, namely translation, rotation and noise addition, followed by randomised perturbations of the contours, to original images to obtain perturbed images. Then, the perturbed features were used to evaluate feature reliability against random factors. They further compared the simulation method with the test-retest method for the evaluation of radiomic model reproducibility.

The advantage of the above simulation-based perturbation method is feasibility. Most radiomics studies can harness such perturbations to evaluate feature reliability and, importantly, evaluate it in a dataset-specific manner. However, this method has a clear limitation in that it can only be used to evaluate the reliability of radiomics features. This contrasts the purposes of prospective studies that investigate the factors affecting radiomic feature reliability. The key issue here is that the reported feature reliability may not be generalisable to other datasets. Some studies have used phantoms to investigate feature reliability but have not provided precise reports of specific features; this has deterred the applicability of their results to radiomics research. Other studies have assessed the inter-observer variability of segmentations and yielded data-specific reliability assessments, but such results do not reflect all of the sources of variation in

radiomics studies. As most of these methods are not data-specific, the direct reliability assessment of radiomic models has not been achieved.

Many studies on radiomic feature reliability assessment have evaluated the effects of manufacturers, imaging protocols and reconstruction parameters. It is clear that these studies have provided valuable information on the generalisation of radiomic models across institutions. However, the most fundamental variations, i.e., random factors that cannot be minimised, have not been adequately investigated.

The simulation-based method of perturbation is a perfect solution to this problem as the perturbation was designed to simulate randomness in the image and can be applied to any dataset without extra costs or resources, such as imaging resources and labour. Furthermore, the method offers a new regime that can be explored and potentially utilised for many purposes. Therefore, in this thesis, a novel simulation method in radiomics has been used to improve radiomic model reliability. Model reliability is defined as the model output variability after slight variations on images and masks.

2.2. Research Aim

The aim of this project was to develop and investigate a perturbation-based method of assessing model reliability to improve model reliability and generalisability. Perturbations have been used to generate internal validation datasets that account for randomisations in the radiomic workflow. To develop and validate the previously described method, we retrospectively collected four publicly available HNC datasets.

Once fully developed, this method can be applied to any radiomics study to evaluate radiomic model reliability and further improve model reliability and generalisability. To the best of our knowledge, this is the first recorded attempt to improve radiomic model reliability using perturbations.

2.3. Research Objectives

2.3.1. Objective 1: Develop a novel perturbation-based framework for the evaluation of radiomic model reliability.

Here, we aimed to develop and evaluate a reliability assessment framework based on image perturbation for the evaluation of radiomic model reliability. Although the radiomics community has been aware of the significance of radiomic model reliability, the lack of materials has precluded the development of a method to directly measure model reliability. We used a publicly available dataset and developed two radiomic models to validate our method. One model was developed using the full feature set, and the other was developed with at minimum number of good, reliable features. Model variability in the perturbation datasets and reliability evaluation metrics was compared for validation purposes. To the best of our knowledge, this is the first study to provide a data-specific and practical method for the direct evaluation of model reliability. Despite the perturbation-based framework being more practical for implementation than the test–retest method, the latter has been the gold-standard method for assessing measurement reliability. Therefore, in the second objective described below, we aimed to check whether the perturbation-based method can substitute the test–retest method in evaluating radiomic model reliability.

2.3.2. Objective 2: Compare the perturbation-based method with the test-retest method for the evaluation of radiomic model reliability.

Model reliability assessment is solely based on simulated perturbations. Therefore, it is important to evaluate the perturbation-based method against the test-retest method of model reliability assessment. The breast multiparametric MRI for prediction of NAC response challenge was used to obtain test-retest data for ADC mapping; the pathological complete response (pCR) was set as the outcome. This dataset provided a platform to compare the effect of model reliability by weighing the test-retest dataset against the simulated perturbation dataset.

2.3.3. Objective 3: Explore the utility of perturbed image features in developing reliable and generalisable radiomic models.

Here, we aimed to optimise the utility of perturbed image features in improving the reliability and generalisability of radiomic models. The reliability evaluation framework, achieved in the first objective, can help to evaluate dataset-specific radiomic reliability. We hypothesised that removing low-reliability features quantified by image perturbation improves radiomic model reliability and generalisability, with model reliability quantifying the model output consistency in terms of perturbed features and model generalisability quantifying the difference between a testing and training area under the receiver operating characteristic curve (AUC). This hypothesis was tested on four publicly available HNC datasets, two classification tasks and five classifiers.

Chapter 3. Development of Perturbation-based Radiomic Model Reliability Assessment Framework

3.1. Introduction

Radiomics is a flourishing field in which machine learning is used to associate cancer imaging phenotypes with cancer genotypes or clinical outcomes for precision medicine [10,13,100]. Radiomics strives to characterize the differences in tumor phenotypes based on non-invasive medical images, such as CT, MRI, and PET. Furthermore, radiomics can be used to capture the heterogeneity of a tumor [5], associate heterogeneity with tumor characteristics for diagnosis [101] and treatment prognostication [102], and improve the overall decision-making during treatment [103].

Despite the potential of radiomics, the unknown reliability of reported radiomics features and signatures against the variability of image acquisition, reconstruction, and segmentation is one of the major challenges in translating radiomic models from bench to bedside [14,104]. Lafata et al. [105] reported the variability of a classification model for non-small-cell lung cancer histology with respect to free-breathing three-dimensional (3D)-CT and phases of four-dimensional (4D)-CT imaging. In addition to radiomic model applications, the deep-learning model variability caused by variations in analyzed images should be considered. Blazis et al. [20] reported the impact of CT reconstruction parameters on the performance of a lung nodule computer-aided diagnosis (CAD) system based on deep learning. They found that the performance of the CAD system increased when the iterative reconstruction levels or the image quality

were also increased. Both publications suggest that the impact of imaging variations on the reliability of radiomic models need to be better understood.

To our knowledge, no study has compared the reliability of radiomic models with that of features against imaging variations. Multiple scans of the same patients obtained within a short interval are necessary to conduct a model reliability study, where the predicted outcomes from different scan sets could reflect the model variability. As obtaining such datasets is resource intensive and increases the burden on the patient, they are only obtained for research purposes. To obtain multiple datasets, Zwanenburg et al. [65] proposed perturbing the images and contours to simulate the acquisition of multiple image sets. They validated this method by comparing the feature reliability with that in two test-retest datasets.

Following this idea, we propose a reliability assessment method of the radiomic model using perturbations. In addition to traditional radiomic modeling methods, we simulated multiple internal validation datasets by adding plausible perturbations to the original images and segmentations. The perturbed data were then used to validate the reliability of the radiomic model against randomization, and reliability was indicated by the intra-class coefficient of correlation (ICC), which was used to describe the consistency of model prediction outcomes within the same patient across all perturbations.

3.2. Materials and Methods

3.2.1. Overview

The overview of the workflow used to demonstrate our model reliability assessment method is illustrated in **Figure 6**. First, we collected pre-treatment CT images and clinical outcomes from a publicly available HNC dataset and randomly split the data into training (70%) and testing cohorts (30%), with similar outcome ratios between the two cohorts. Second, a radiomic survival model was built to assess distant metastasis (DM)-free survival. Third, internal validation datasets with perturbations were simulated [65,106]. The simulated perturbation datasets were used to extract perturbed radiomics features and validate the survival model's reliability against randomizations, as shown in **Figure 6** (b). Finally, the ICC was used to quantify the model's reliability, reflecting its prediction consistency when using the perturbed data.

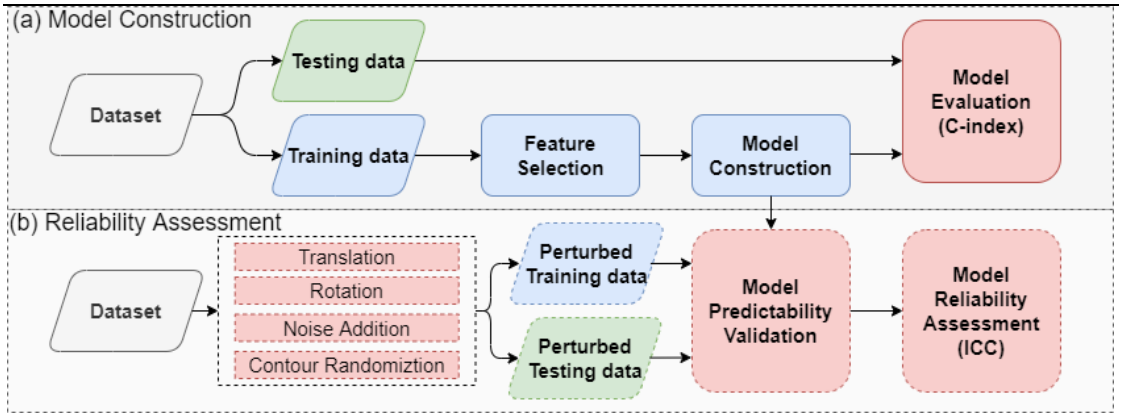

Figure 6. The general workflow of the study. The part (a) shows the model construction workflow with first randomly split the cohort into the training and testing data, in which the training part is for the model development and test part is for the model performance validation. The part (b) shows the reliability assessment workflow. The entire cohort is used to simulate the perturbed cohort by adding the randomizations to image through translation, rotation and noise addition, and to contour. Then, the perturbed data was used to validate the model for the reliability against randomization. Finally, the model reliability is quantitatively evaluated with ICC.

3.2.2. Materials

The dataset, Head-Neck-PET-CT [107], was collected in TCIA [24]. This dataset consists of 298 patients with head and neck squamous cell carcinoma (HNSCC) with a median follow-up of 43 months. The patients were treated at four different centers and received only radiation (n = 48, 16%) or chemo-radiation (n = 250, 84%) with curative intent. The informed consent has been waived due to retrospect nature of the study.

The ROI for feature extraction was the primary GTV, which was the primary treatment target of radiation therapy. The GTV is the most reliable region for predictive feature extraction [108] and has been used in several predictive radiomics studies of HNSCC [13,109,110].

Distant metastasis-free survival, defined as the interval from the first day of treatment to the date of the event, was the clinical endpoint in this study to demonstrate the reliability assessment of the radiomic model [111]. Previous studies of binary classification models of HNC [110,112] have achieved good prediction results but were limited because the time-to-event was neglected during model development.

3.2.3. Image Preprocessing and Radiomic Feature Extraction

The CT images and their GTV contours were preprocessed before their features were extracted to maintain the features' reproducibility and consistency [92,113]. First, the GTV contours were interpolated to a voxel-based segmentation mask. Second, an isotropic resampler (1 mm × 1 mm × 1 mm) was applied to the images and masks, with B-spline interpolation on the image and nearest-neighbor interpolation on the mask to

enhance the reproducibility of the radiomics features [114]. The preprocessing steps were implemented on Python v3.8 using the SimpleITK v1.2.4 [115] and OpenCV [116] packages.

The radiomics features were then extracted using the Pyradiomics v2.2.0 [11] package, which is Image Biomarker Standardization Initiative-compliant [87,88]. A total of 5,486 radiomics features were extracted from the GTV of each patient's CT scan. Twelve images were included in the feature extraction, including one unfiltered image, three Laplacian-of-Gaussian filtered images (with sigma values of 1 mm, 3 mm, and 6 mm), and eight Coiflet1 wavelet filtered images (LLL, HLL, LHL, LLH, LHH, HLH, HHL, HHH). In addition to the 14 shape features from GTV segmentation, 18 first-order and 73 second-order features were extracted from the ROI of each filtered image. A re-segmentation of the soft-tissue range (−150 to 180) [65] and discretization, with fixed bin counts of 4, 8, 16, 32, 64, and 128, were specified for the texture feature extraction.

3.2.4. Radiomic Modeling Summary

Patients were randomly assigned to the training and testing cohorts (70/30 split) with stratification by DM status [102,117]. The data in the training cohort were used for feature selection and subsequent model training, while the data in the testing cohort were used to evaluate the model's performance, shown in **Figure 6**.

3.2.5. Feature Selection

A filter-based feature selection method was adopted in our analysis [118]. This

process has two steps: feature–outcome relevance filtering and feature–feature redundancy filtering. Identifying the most relevant and less redundant features is a common practice in radiomics studies, regardless of the evaluation metric [96].

Relevance filtering. Relevance filtering aims to identify the radiomics features that are correlated with the outcomes [110]. First, the outcome relevance of each feature was repeatedly evaluated by log-rank test P-values under downsample bootstrapping (imbalanced-learn 0.8.0 [119]) without replacement over 100 iterations on the training dataset. Downsampling can be used to capture useful information in an imbalanced dataset [120]. Second, features with P-values less than 0.1 were selected in each iteration and ranked by their frequencies, with the top 10% of features with the highest frequencies selected.

Redundancy filtering. Redundancy filtering aims to remove features correlated with each other [121]. First, the feature pairs with Pearson correlation coefficients higher than 0.6 were identified. Then, the features with higher mean correlation coefficients than the rest of the features were removed. The removal of these redundant features should improve the predictive ability of the classifiers [122].

3.2.6. Model Building

To build the survival model, the optimal features for model building were identified using backward recursive feature elimination based on the penalized Cox proportional hazard model [123]. This approach maximizes the validation concordance index (C-index) curve by using repeated three-fold cross-validation in the training set. After identifying the optimal features, a penalized Cox proportional hazard survival model

was built for DM-free survival. The hyperparameter of the model was fine-tuned with five-fold cross-validation to maximize the C-index for the survival model. Thus, the model's performance with the training and testing cohorts was evaluated.

3.2.7. Reliability Assessment

This section describes the method to evaluate the model reliability using perturbations and the workflow shown in **Figure 6(b)**. First, the internal validation datasets were simulated with the perturbations by adding plausible randomizations to the original images and segmentations. Second, the survival model was evaluated using both the perturbed training and testing data. Third, the model reliability against simulated randomization was quantified using the reliability index ICC.

3.2.8. Validation Data Simulation

The internal validation data sets were simulated using the perturbation method [65,124]. For each perturbation, both the image and mask were translated and rotated simultaneously by a random amount. This simulation aimed to mimic variations in the patient's position during imaging. Then, a random Gaussian noise field was added to the image to mimic the noise level variations between different image acquisitions [125]. Next, the GTV mask was also perturbed by a randomly generated deformable vector field, which aimed to simulate uncertainties in inter-observer delineations on the same target [126]. In total, 60 sets of perturbed images and contours were simulated, with the corresponding radiomics features extracted as the internal validation sets to evaluate the model reliability under randomization.

3.2.9. Model Validation

The model performance was validated and reported on the original and perturbed datasets using the C-index as the evaluation metric. Two observations may warrant attention. First, the model performance consistency between the original and perturbed datasets might be a qualitative indicator of model performance reliability against the simulated randomizations. Second, the model performance variance with perturbed datasets may reflect the model's sensitivity to slight fluctuations. A quantitative assessment of model reliability could be performed by comparing the model performance on the original and perturbed data.

3.2.10. Model Reliability Quantification

In addition to the qualitative analysis of model reliability, a quantification metric, the ICC, was proposed to evaluate model reliability under randomization. The ICC is often used as a reliability index for inter-rater reliability analysis [127], and several radiomic studies have used this measure to quantify feature reproducibility [37,106,128].

The model reliability ICC reflects the extent to which the measurements can be replicated. We aimed to determine whether model predictions can be repeatedly measured/produced after adding plausible randomizations to the images and segmentations both for the same patient and across the entire dataset. As each perturbed dataset was simulated randomly and the model was expected to yield an identical outcome, the one-way random effects with absolute agreement, ICC(1, 1), were calculated to quantify the model's reliability, with patients as the subjects and perturbations as the raters [129]. ICC values range between 0 and 1, with values closer

to 1 representing more robust reliability. Typically, ICC values less than 0.5, between 0.5 and 0.75, between 0.75 and 0.9, and greater than 0.9 indicate poor, moderate, good, and excellent reliability, respectively [129].

3.2.11. Model Reliability Validation

To validate the calculation of model reliability, the same experiment was repeated with highly reliable features (ICC > 0.75). This validation aimed to verify the sensitivity of the ICC in response to changes in model input reliability. An increase in feature reliability was expected to increase the model ICC.

3.3. Results

First, the optimal features and associated characteristics for model building are reported. Second, the model's performance on the original and perturbed dataset are evaluated. Third, the reliability of the radiomic model is computed.

The first step was to identify the features relevant to the outcome and remove redundant features. After filtering, 17 of 5486 features were selected. Then, a backward recursive feature elimination based on a penalized Cox proportional hazard model was used to find the optimal feature set for model building. **Figure 7** shows the changes in training and validation C-indexes of a 10-times-repeated, three-fold cross-validation of the training dataset with respect to the number of features in the recursive feature elimination process. The feature set with the highest validation C-index was identified as the optimal feature set, and thus six features were identified as the optimal feature set and used for model building. The characteristics of these six selected features are

tabulated in **Table 5**.

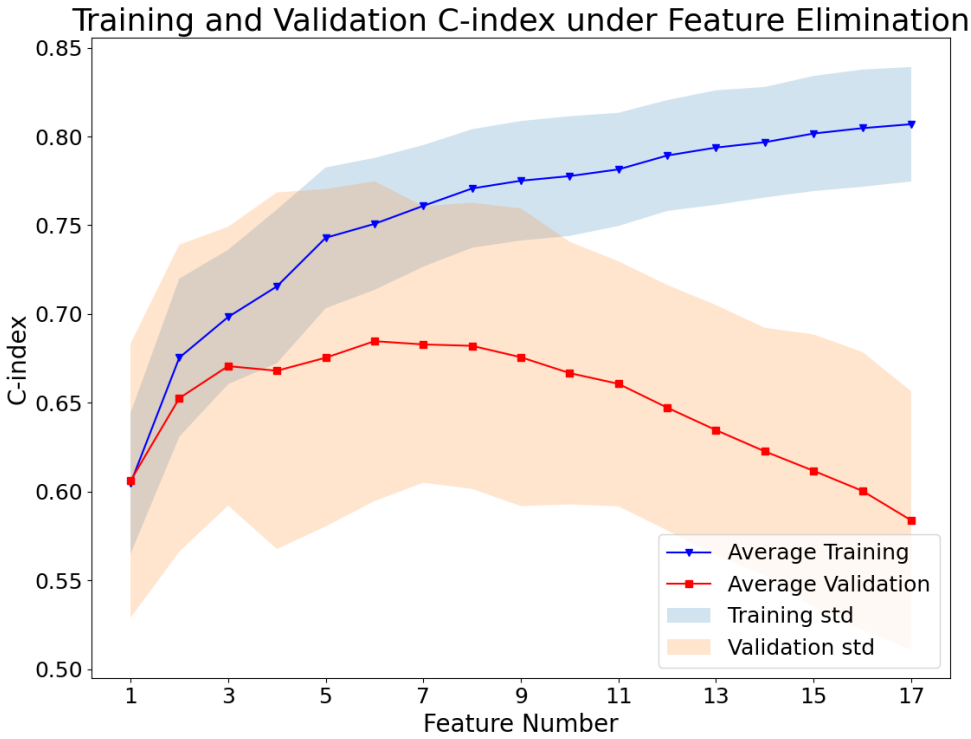

Figure 7. Changes in the training and validation C-indexes with respect to feature numbers in the stepwise backward feature elimination method under three-fold cross-validation, repeated 10 times. The points indicate the averaged C-index over cross-validation folds, and the shaded area indicates the range of one standard deviation (std). The curve indicates the feature number (N=6) yielding an optimal validation performance.

Table 5. The characteristics of selected features for model building. The univariate C-index, P-value, and ICC were tabulated. Feature names indicate the feature, the bin count (if applicable), and the image used to compute it.

features	C-index	P-value	ICC
log-sigma-6-0-mm-gldm_LargeDependenceLowGrayLevelEmphasis_64_binCount	0.619	0.045	0.747
wavelet-HHL_glrlm_LongRunLowGrayLevelEmphasis_128_binCount	0.587	0.169	0.454
original_glszm_LargeAreaLowGrayLevelEmphasis_128_binCount	0.614	0.066	0.610
wavelet-LLL_glrlm_RunEntropy_128_binCount	0.608	0.064	0.900
wavelet-LHL_glszm_LowGrayLevelZoneEmphasis_64_binCount	0.572	0.091	0.491
wavelet-HLL_glszm_SmallAreaHighGrayLevelEmphasis_128_binCount	0.604	0.085	0.542

After identifying the six optimal features, the radiomic survival model was constructed and validated. The C-indexes of the survival radiomic model in the training and testing cohorts were 0.742 and 0.769, respectively. The averaged model performance C-indexes (standard deviation) over the perturbed training and testing cohorts were 0.686 (0.038) and 0.678 (0.065), respectively.

The model performance on the original and perturbed cohorts is visualized in **Figure 8**, which shows that the original training and testing C-indexes probably overestimate the model's performance compared with the perturbed cohort evaluation.

Furthermore, the model performance variations on the perturbed cohorts are significant, with C-indexes ranging from 0.609 to 0.758 in training and from 0.514 to 0.794 in testing.

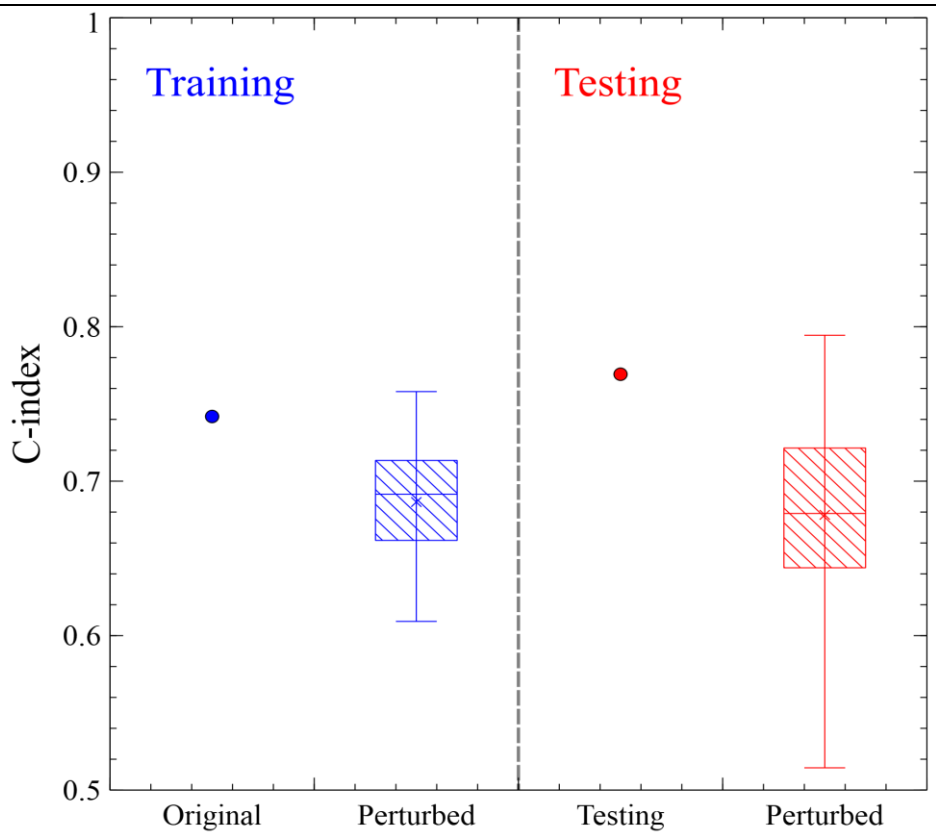

Figure 8. Visualization of model performance on the original and perturbed data. The training and testing C-index on the original data is within the performance of perturbed data, indicating that the original dataset could be a subset of the perturbed subset. Furthermore, Although the averaged C-index for the perturbed training and testing did not show a statistically significant difference (P-value = 0.418), the variations in the testing data (STD = 0.065, ICC = 0.565) is larger than the training data (STD = 0.038, ICC = 0.596).

After evaluating the model's performance, the quantified model performance using ICC was calculated with a 95% confidence interval. The model reliability ICC was 0.565 (0.518–0.615) on the training set and 0.596 (0.527–0.670) on the testing set. According to the convention [129], this model's reliability is moderate ($0.5 < ICC < 0.75$), and it is consistent with the significant variations in model performance with the perturbed datasets as shown in **Figure 8**.

An additional experiment was performed to validate the sensitivity of the reliability ICC, using the highly reliable features ($ICC > 0.75$) to repeat the radiomic modeling process. After prescreening the reliable features, 67% (3667 / 5486) of features were retained; these were reduced to four optimal features for model building after feature selection. The new model performance C-indexes for the original training and testing cohorts were 0.711 and 0.641, respectively, while the averaged perturbed training and testing C-indexes (standard deviation) were 0.640 (0.029) and 0.625 (0.042). The model reliability ICC values, with a 95% confidence interval, were 0.782 (0.749–0.815) and 0.825 (0.782–0.867) for the perturbed training and testing sets, respectively.

An additional experiment, starting with highly reliable features, led to a significant increase in the model reliability ICC values from moderate to good. This result demonstrated the sensitivity of our method to input reliability.

The subgroup analysis based on filtered images was also performed. The median value radiomic feature ICC (range) for the original image group, log-sigma image group, and wavelet image group is 0.87 (0.42-1.00), 0.91 (0.35-0.99), and 0.77 (0.14-0.99). **Table 6** showed the subgroup analysis results based on the filtered image groups. In

general, the trend of improving model reliability is maintained, which also indicates that our method can be used to quantify radiomic model reliability for quantitative analysis using filtered or non-filtered images.

Table 6. The model performance in discrimination and reliability. An improvement in model reliability is observed after removing non-robust radiomics features.

		Training C-index	Testing C-index	Model Reliability ICC
No Filtering	Original features	0.67	0.71	0.72
	Log-sigma features	0.72	0.54	0.59
	Wavelet features	0.80	0.56	0.51
Feature ICC > 0.75	Original features	0.65	0.74	0.85
	Log-sigma features	0.58	0.54	0.91
	Wavelet features	0.62	0.55	0.89

The feature maps of feature wavelet-LLL_glrlm_RunEntropy and wavelet-HHL_glrlm_LongRunLowGrayLevelEmphasis were calculated across perturbed images to interpret the results visually. As shown in **Figure 9**, the feature map of RunEntropy showed a homogeneous pattern across perturbed images than the feature map of LongRunLowGrayLevelEmphasis, which is consistent with the feature reliability ICC calculated using perturbation images.

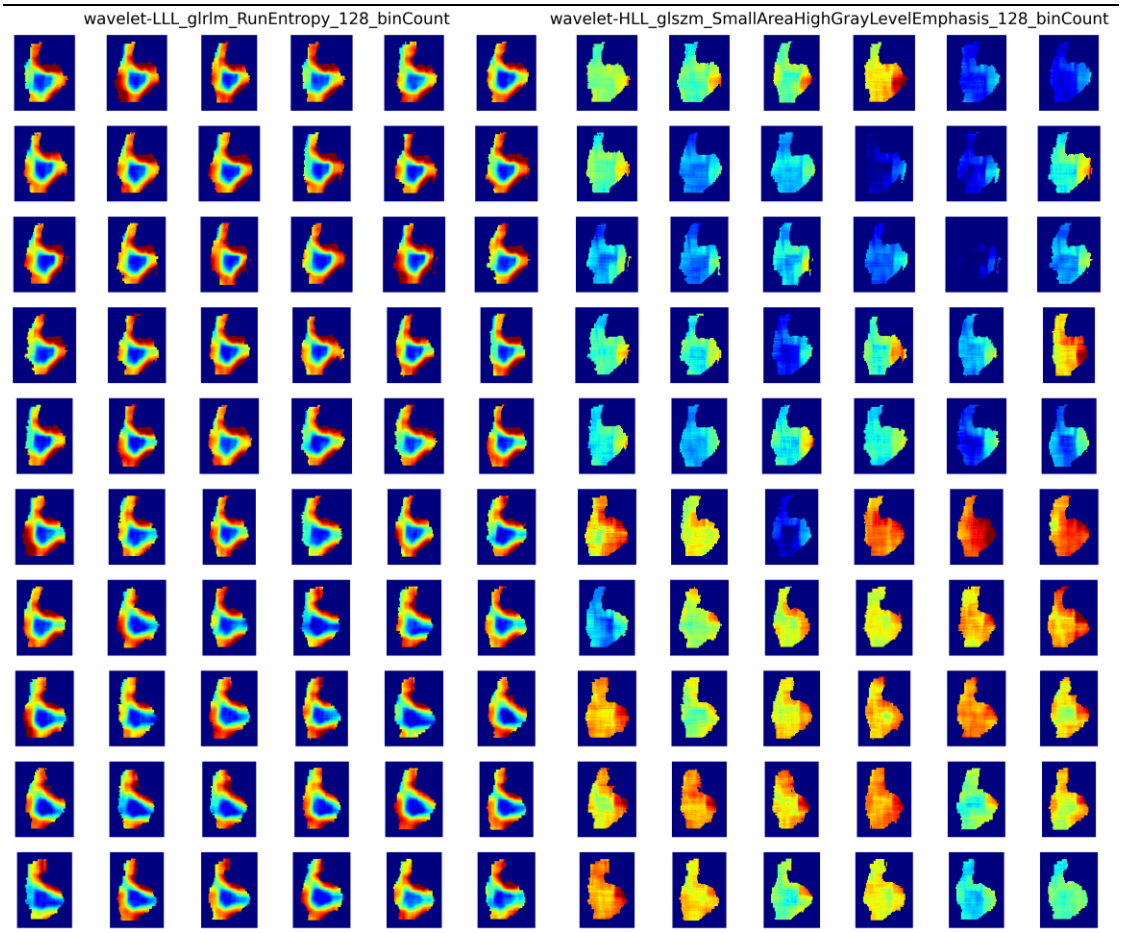

Figure 9. The feature map of wavelet-LLL_glrlm_RunEntropy (left) and wavelet-HLL_glszm_SmallAreaHighGrayLevelEmphasis_128_binCount (right) for same patient with identical axial slice. The window is fixed between 1 percentile and 99 percentile of the feature map to eliminate the effects of noise. The visualization of feature maps revealed the radiomic feature reliability against perturbations.

After evaluating the model's discriminatory power, the quantified model performance using ICC was calculated with a 95% confidence interval. The model reliability ICC was 0.565 (0.518–0.615) on the training set and 0.596 (0.527–0.670) on the testing set. According to convention [130], this model's reliability is moderate (0.5 < ICC < 0.75), and it is consistent with the significant variations in model performance with the perturbed datasets, as shown in **Figure 8**.

An additional experiment was performed to validate the sensitivity of the reliability ICC, using the highly reliable features (ICC > 0.75) to repeat the radiomic modeling process. After prescreening the reliable features, 67% (3667 / 5486) of features were retained; these were reduced to four optimal features for model building after feature selection. The new model performance C-indexes for the original training and testing cohorts were 0.711 and 0.641, respectively, while the averaged perturbed training and testing C-indexes (standard deviation) were 0.640 (0.029) and 0.625 (0.042). The model reliability ICC values, with a 95% confidence interval, were 0.782 (0.749–0.815) and 0.825 (0.782–0.867) for the perturbed training and testing sets, respectively. The univariable analysis result has been tabulated in **Table 6**.

3.4. Discussion

This study proposed a radiomic model reliability evaluation method using data perturbations. We demonstrated this method using a publicly available dataset and by building radiomic models to predict DM-free survival. To our knowledge, this is the first study to describe a method to assess the reliability of radiomics models based on image perturbation. Our method evaluates model reliability against randomization in a

radiomic workflow using the perturbation method. This study may provide a new perspective on model assessment for the radiomic community. Our results showed that model performance can be overestimated, despite the decent model predictability achieved using an independent testing set. Moreover, simulated perturbation data can serve as an internal validation method for a model reliability assessment.

This study is also the first to assess radiomic model reliability. Currently, there is no radiomic model reliability assessment method, despite consensus on the importance of building reliable radiomic models within the community [131]. This paradox may be due to several reasons. First, the reliability of a model covers a wide range of aspects, as radiomics is a multi-step process and uncertainties may be introduced in each step [104,132]. Therefore, it is challenging to characterize the stability of radiomic models. Second, limited medical resources, such as re-scanned images, prevent the internal validation of model reliability. If multiple scanned image sets obtained over a short time interval and inter-observer delineations of different scans were available, the model could be validated internally to account for random variations in parameters such as patient positioning and inter-observer delineation. Third, it is challenging to characterize a model's reliability against controllable factors, such as different scanners and acquisition parameters, because such medical resources are inaccessible. These factors have been shown to affect radiomic feature reproducibility and, potentially, model reliability. To tackle some of these challenges, our study used the perturbation method to simulate perturbed datasets, thereby accounting for randomized factors in the radiomic workflow. For example, rotation and translation mimic variations in the patient's positioning during the scans and resampling uncertainties, noise addition

mimics fluctuations in the voxel values caused by statistical uncertainties, and contour randomization mimics inter-observer uncertainty in region-of-interest delineation. These simulated datasets play a crucial role in assessing radiomic model reliability.

This study also evaluated the reliability of the model against randomness. The majority of reliability studies in radiomics publications have focused on the reproducibility and reliability of controllable factors, such as the scanner brand [133], image acquisition parameters [134], reconstruction kernels [135], and preprocessing parameters [136]. However, the effects of these controllable factors can be minimized with sufficiently transparent reporting [104]. In contrast, random and natural variations persist in every radiomic study and are difficult to address by harmonization or standardization. Therefore, understanding the impact of randomness on radiomics features and models is crucial for establishing clinical radiomic applications.

Our results revealed the vulnerability of our radiomic model to randomness. In our results, the model performance evaluation using perturbed data showed lower training and testing C-indexes for the survival model and considerable variability in its distribution under perturbations. The lower training C-index for the perturbed data reveals that evaluating models using their original data results in overfitting to noise in the original data and over-estimation of the model's learning. If a model is unable to achieve a similar performance using the same data with plausible randomization, it is unlikely that it could be translated to the clinic. Careful assessment of radiomic models' reliability is therefore essential.

A potential solution to this issue is to evaluate the reliability of features under

randomization and integrate this information into radiomic modeling. Despite plenty of discussion and studies of radiomic feature reliability and reliability under various circumstances, only two methods have been implemented in a few clinical studies. The first method uses a test-retest dataset and evaluates radiomic feature reliability using two consecutive scans in a short interval, followed by incorporating this reliability into the dataset. This method may reflect realistic feature reliability under test and retest settings. However, the acquisition of test-retest imaging is rarely conducted outside of a research context, and most medical imaging datasets therefore lack complimentary test-retest image data. Although some studies have adopted the test-retest RIDER Lung dataset [137] to assess feature reliability in an attempt to build reliable models, the generalizability of feature reliability from the RIDER Lung data to the dataset being studied has been criticized [98]. The second method assesses feature reliability using inter-observer variability on the contours. The ROI on the images is delineated multiple times by independent oncologists, and feature reliability is evaluated from the inter-observer consistency of feature values. This method is more practically accessible than test-retest images to assess feature reliability. However, this method also has limitations in terms of the insufficient identification of non-robust features and high medical personnel costs [65]. The shortcomings of these two methods for assessing feature reliability limit their effectiveness for removing non-robust radiomics features during radiomic modeling, potentially resulting in radiomic models that are vulnerable to randomization. Therefore, simulated randomization of a dataset via the perturbation method may enable estimation of the impact of randomness on radiomic modeling. Multiple perturbed datasets can be generated with perturbations, and their feature

values can be determined. Feature reliability can be quantified using the ICC for each feature by considering its variability within a single subject and across the dataset. Then, removing the less reliable features can improve the reliability of radiomic models against randomizations. In contrast to test-retest and inter-observer variability, simulation methods may be more versatile for evaluating feature reliability with no additional clinical resource costs and could enable data-specific feature reliability evaluations. Moreover, perturbations can provide additional validation data to evaluate model reliability and safeguard it against randomization.

In addition to these contributions, some aspects of our approach could be explored to enhance the impact of this study. First, image and contour perturbation via simulation is a new method in radiomics, so comparisons between this and established methods (e.g., test-retest and inter-observer variability) could be studied further to identify their respective advantages and disadvantages. Second, our validation results showed a decline in model predictability performance from the testing data when poorly and moderately reliable features were removed. A future study could investigate how to balance the model's predictive performance with its reliability.

3.5. Conclusions

This study proposed a radiomic model reliability assessment method using perturbations. This method identifies unreliable models by comparing the model's performance on the training dataset with the performance achieved on random perturbations of the training dataset. Using this approach could help the radiomics community to build more reliable models for future clinical applications.

Chapter 4. Comparing Effectiveness of Image Perturbation and Test-retest Imaging Towards Establishment of Reliable Radiomic Models

4.1. Introduction

Radiomics is one of the most up-to-date quantitative imaging techniques nowadays. Quantitative features, which are believed to represent tumor phenotypes that are imperceptible to human eyes, are extracted in a high-throughput manner from routine medical imaging, such as CT, MR, or PET. Morphological, histogram, as well as textural information could be included in different classes of radiomics features. They are then selected and built into different models to help noninvasive diagnosis [138–140], prognosis [141–143], and treatment response prediction [144–146]. Despite the promising potential of radiomics, the reliability of radiomic models is one of the major concerns when translating into routine clinical practice.

Radiomic feature reliability refers to the feature's ability to keep stable when the same subject is imaged several times under the same acquisition settings. It is believed to be the first and foremost criteria to ensure model reliability and has been studied extensively by previous research. Test-retest imaging is one of the most popular approaches by repeatedly scanning each patient within a short period of time, and feature reliability is assessed by comparing the feature values between the two different scans. For example, Granzier et al. identified repeatable radiomics features within breast tissues using a two-day interval test-retest data with fixed scanner and clinical

breast protocol [147]. However, test-retest imaging is not a standard clinical procedure and requires additional medical resources and potential extra dose to patients. Consequently, the existing test-retest study include only a limited number of patients, which further reduced the significance of their findings. In addition, the conclusions of feature reliability are hardly generalizable across image modalities and cancer sites [148], rendering the necessity of specific reliability analysis for different radiomic studies.

Several methods have been proposed to assess radiomic feature reliability through perturbations. Marco et al. first applied random translations of the ROIs to assess the radiomic feature reliability on ADC images [149]. They found an overall satisfactory reliability and a high site dependency. Zwanenburg et al. proposed to generate pseudo-retest images by random translation, rotation, noise addition and contour randomizations, and demonstrated the similar patterns of feature reliability to test-retest imaging [150]. Further studies have demonstrated the potential of perturbed images in quantifying radiomic model output reliability and improving the model generalizability and reliability by removing low-repeatable features [151]. Although perturbation methods have been proven to be capable of capturing most non-repeatable features in test-retest images, it is still unknown if image perturbation could replace test-retest imaging in building a reliable radiomic model.

This study aimed to compare radiomic model reliability after removing non-repeatable radiomics features assessed by image perturbation and test-retest imaging. A unique breast cancer dataset with available test-retest ADC images derived from

diffusion weighted MRI (DWI) scans and pCR outcome was retrospectively collected. Patients were randomly split into one training and testing set for model development and validation. We compared model reliability, including both generalizability and reliability, between models built from repeatable feature assessed by image perturbation and test-retest under a wide range of reliability thresholds. The overall study workflow is summarized by **Figure 10**. This study could provide the radiomic community direct evidence of the benefit of image perturbation on building reliable radiomic models. Most importantly, whether image perturbation is equivalent to test-retest imaging in building a reliable radiomic model could be directly validated.

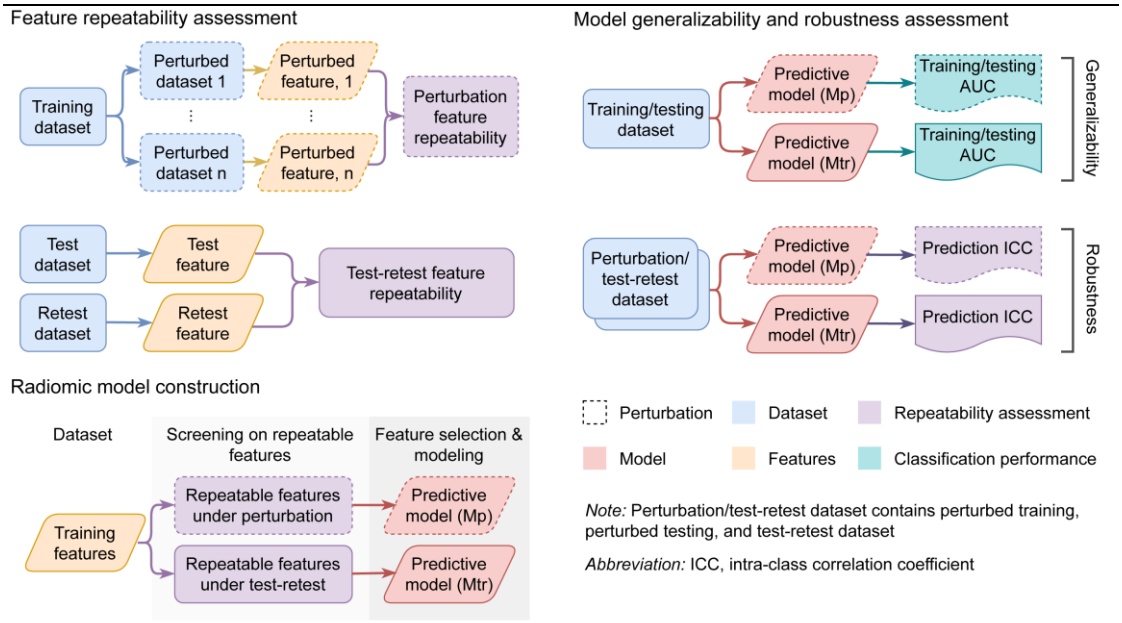

Figure 10. Study workflow. We conducted our study by radiomic feature reliability assessment by test-retest and perturbation, radiomic model development using high-repeatable features from the two assessments, and generalizability and reliability analysis of the two models.

4.2. Material and Methods

4.2.1. Patient Data

We retrospectively collected 191 patients from the publicly available BMMR2 challenge dataset [34,152]. It was derived from the ACRIN 6698 trail where female patients with invasive breast cancer were prospectively enrolled from ten institutions between 2012 and 2015 [153]. Institutional review board (IRB) approval is waived due to the solely use of public data. Pre-treatment DWI-derived ADC maps and manual tumor segmentations were downloaded from TCIA [24] for radiomics model development. pCR at the time of surgery [154] was used as the prediction endpoint. We adopted the same train-test split as the BMMR2 challenge with 60% (n=117) randomly chosen as the training set and the remaining 40% (n=74) set as the testing set. We also downloaded 71 test-retest pre-treatment ADC map pairs scanned within a "coffee-break" with 41 overlapped with the primary patient cohort. The tumor volume was manually drawn on dynamic contrast-enhanced MR subtraction images [153], and migrated to the ADC map. The number of 74 testing patients is selected to ensure a sufficient statistical power to test the association between image feature and pCR.

4.2.2. Radiomics Feature Extraction

A comprehensive set of Radiomics features was extracted from the original and filtered DWI images within the tumor volume. All the images were preprocessed by isotropic resampling (1x1x1mm) and 32-bin-number discretization before feature extraction. Both first-order (n=18) and texture features (n=70) were extracted from each preprocessed image, and shape features (n=14) were extracted from the tumor

segmentation. The definitions and extraction of radiomics features follow the standardization by the IBSI [155]. In total, 1316 radiomics features were extracted for each patient. Detailed settings of the image preprocessing and feature extraction parameters are listed in **Table 7**. All the image preprocessing and feature extraction procedures were performed by the Python package PyRadiomics [11].

4.2.3. Feature Reliability Assessment

Table 7. Image perturbation, preprocessing, and radiomic feature extraction parameters.

Parameters	Specifications
Pixel value offset	0
Resample pixel size (mm)	[1,1,1]
Image/mask interpolation algorithm	B-spline
Mask partial volume threshold	0.5
Interpolation grid alignment	Align grid origins
Translation distances (pixel)	[0.0, 0.2, 0.4, 0.6, 0.8]
Rotation angles (degree)	[-5,0,5]
Rotation axis	Mask bounding box center, axial direction
Contour randomization smoothing sigma (mm)	[10,10,10]
Contour randomization intensity (mm)	[1,1,1]
Perturbation times	40
Image discretization bin number	32
Image filters	Unfiltered, Laplacian-of-Gaussian (3D), Wavelet
Kernel size of Laplacian-of-Gaussian filter (mm)	[1,2,3,4,5]
Wavelet filter starting level	0
Wavelet filter total level	1
Wavelet filter type	Coilf1
Wavelet filter decompositions	[LLL, HLL, LHL, LLH, LHH, HLH, HHL, HHH]
Feature class	shape, firstorder, glcm, glrlm, glszm, gldm, ngtdm

Radiomics feature reliability was assessed from both perturbed images and test-retest images for model reliability comparisons, shown in **Figure 10**. We performed 40 image perturbations independently for each patient by random combinations of rotations, translations, and contour randomizations, same as the methodology adopted by Teng et al [151]. Detailed image perturbation parameters can be found in **Table 7**. The same set of radiomics features were extracted from each perturbed or test/retest image with the same preprocessing procedure. One-way, random, absolute, single rater ICC [156] was calculated for each radiomic feature under both image perturbation and test-retest due to the random choice of perturbation parameters and scanning condition for each patient. The ICC calculation was provided by the Python package Pingouin (version 0.5.2) [157].

4.2.4. Radiomics Model Construction

Two radiomics models were separately constructed from the repeatable features under image perturbation (M_p) and test-retest (M_{tr}), as shown in **Figure 10**. Volume dependent radiomics features were first removed to minimize its confounding effect on the comparison results, as tumor volume is more stable by definition. Radiomics features that had a Pearson correlation r > 0.6 to the tumor mass volume was removed from subsequent analysis. Repeatable features were determined from the pre-set ICC thresholds of 0, 0.5, 0.75, 0.9, and 0.95. They were further filtered by redundancy and outcome relevancy before model training. We adopted the minimum Redundancy - Maximum Relevance (mRMR) feature selection algorithm to rank the repeatable features based on the redundancy and outcome relevancy [158]. Finally, 5 top-ranked

features were selected for model development. The majority pCR group (non-event) was randomly down-sampled by 500 times and an ensemble of logistic regression models were trained. The final prediction probability of each patient was given by the average of the individual model predictions. This easy-ensemble approach could reduce the training bias from the heavily imbalanced outcome [159]. It was implemented by the publicly available python package imbalance-learn (version 0.9.1) [160].

4.2.5. Model Reliability Assessment

We assessed radiomics model reliability in both generalizability and reliability (**Figure 10**). Model generalizability was assessed by comparing training and testing classification performance evaluated by AUC. Model reliability was assessed by the model prediction reliability under the setting of both training perturbation, testing perturbation, and test-retest. Probability predictions of either model was generated on the perturbed training, perturbed testing, and test-retest images, and the one-way, random, absolute ICCs were calculated for the prediction reliability using the same rationale of feature reliability. Both generalizability and reliability were compared between M_p and M_{tr} with different ICC threshold settings, as shown in **Figure 10**.

4.2.6. Statistical Analyses

Each classification performance metric was evaluated under 1000-iteration patient bootstrapping to acquire 95% confidence interval (95CI). Two-way P-values for comparing the classification performance were calculated by permutation test with 1000 iterations using the function “permutation_test” provided by the open-source Python package SciPy (version 1.9.1) [161]. The comparison was performed between

each pair of models with and without feature reliability filtering as well as M_p and M_{tr}. A P-value < 0.05 was considered significant. The 95CI of the model prediction ICC was evaluated according to the formulas presented by McGraw et al [156].

4.3. Results

4.3.1. Feature Reliability and Predictability

A systematic larger feature reliability based on image perturbation was found compared to test-retest imaging. **Figure 11** visualizes the distribution of feature ICCs assessed by training perturbation versus test-retest. Among all the 1120 volume-independent radiomics features, only 143 showed lower ICC under training perturbation than test-retest, which can be visualized as scarce scattered points above the diagonal line in **Figure 12**. However, they demonstrated a strong correlation with Pearson correlation r = 0.79 (P-value<0.001).

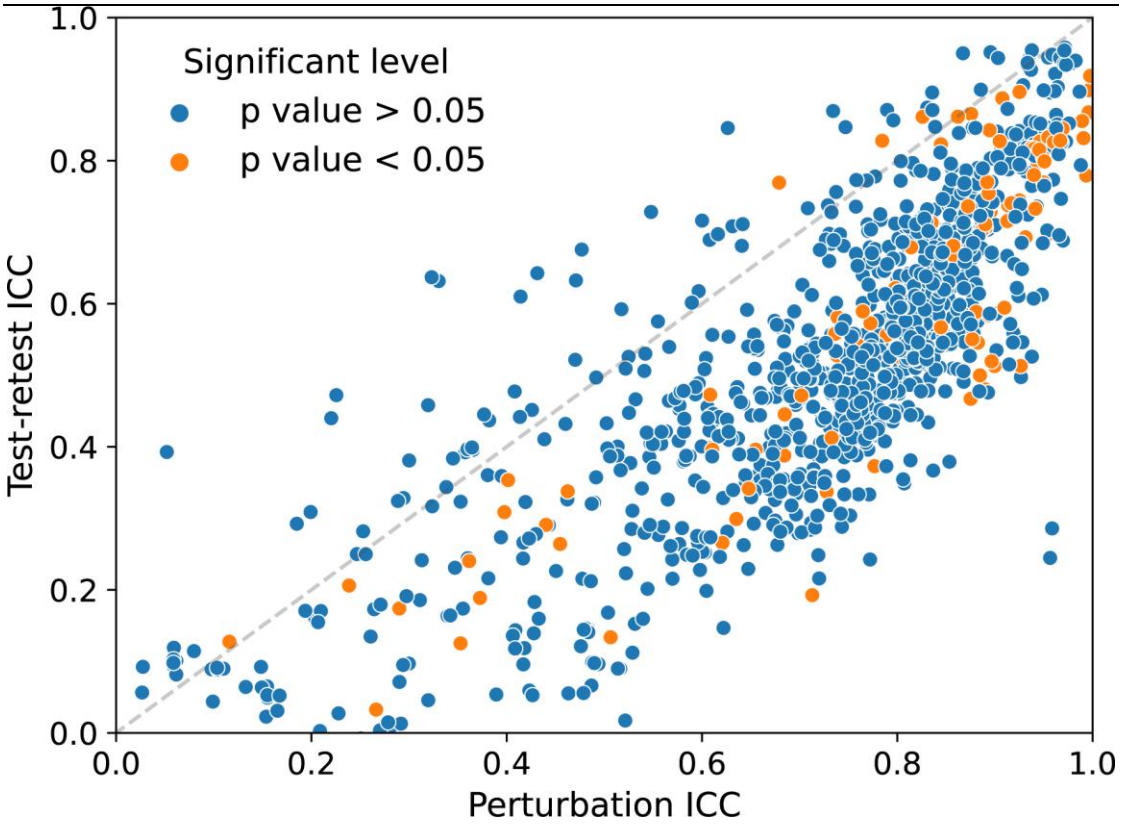

Figure 11. Scatter plots showing the reliability of volume independent features measured by ICC under test-retest imaging (y-axis) and image perturbation (x-axis). The perturbation method yielded higher ICC values than test-retest method in general. Furthermore, features that had significant univariate correlations with the outcome, pCR, where colored as orange while the rest as blue.

The feature reliability agreement between perturbation and test-retest showed a strong dependence on ICC thresholds, as shown in **Figure 12**. In general, the number of commonly repeatable and non-repeatable features between the two ICC measures increased with higher ICC thresholds. Specifically, the number of mutually agreed repeatable features decreased from 621 to 141, 18, and 2 with ICC threshold increased from 0.5 to 0.75, 0.9, and 0.95, as suggested by the shrinking blue bars in **Figure 12**. In contrast, the number of mutually disagreed repeatable features increased from 151 to 484, 989, and 1072 (green bars). For disagreements between perturbation and test-retest evaluation, very few (<0.7%) features are repeatable against test-retest variations while unrepeatable against perturbations (red bars), and a considerate number of features are repeatable against perturbation while unrepeatable under test-retest settings (orange bars).

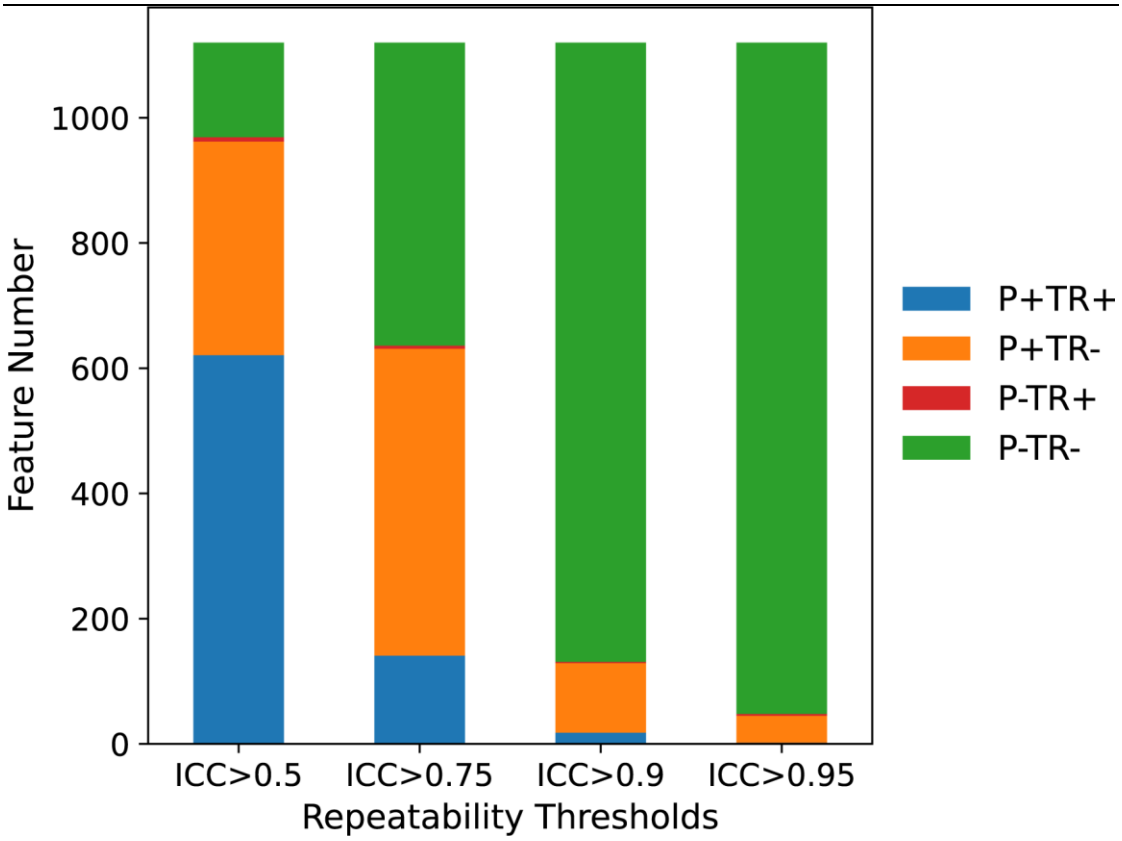

Figure 12. Stacked bar plot displaying the feature reliability agreement between perturbation and test-retest. P+/- indicates the repeatable/unrepeatable feature group by the perturbation method and TR+/- for repeatable/unrepeatable feature group in the test-retest method.

Only a small portion of all the volume-independent radiomics features demonstrated strong univariate correlation with the prediction outcome with an inclination towards high repeatable features (**Figure 12**). Quantitatively, 111 radiomics features reached statistical significance (P-value < 0.05) when correlating with pCR. With the ICC threshold of 0.5, 11% (n=71) of the high-repeatable features under test-retest had statistical significance and 10% (n=93). The percentage increased to 23% at ICC threshold of 0.75 but decreased to 5% (n=1) and 0% (n=0) at 0.9 and 0.95 for test-retest. However, a continuous increase to 11% (n=72), 25% (n=32), and 27% (n=12) for perturbation was discovered.

4.3.2. Model Generalizability and Reliability

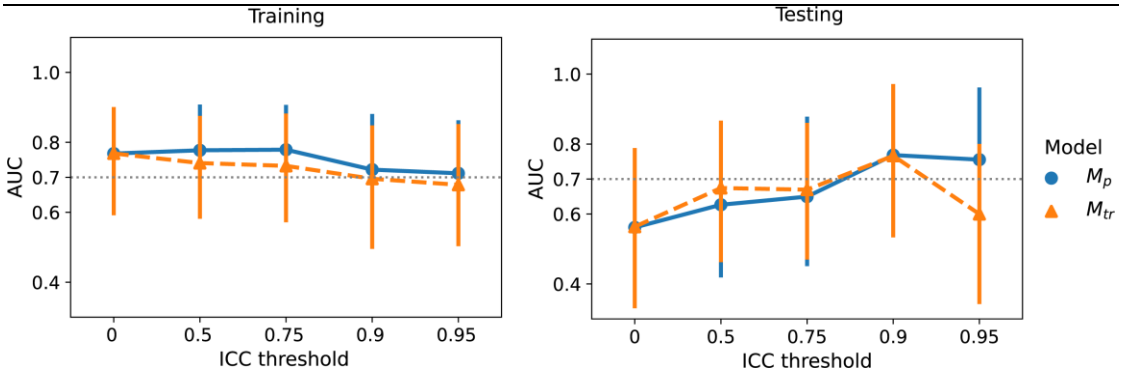

Figure 13. Comparison of generalizability between models based on repeatable features assessed by image perturbations (M_p, blue) and the test-retest imaging (M_{tr}, orange) under varying thresholds. ICC was used to quantify the feature reliability under perturbation for M_p and under tests-retest imaging for M_{tr}. Training and testing classification performance were quantified by AUC. The error bars indicate 95% confidence intervals acquired from 1000-iteration bootstrapping.

An overall trend of increasing generalizability and reliability was observed with increasing ICC thresholds. **Figure 13** presented the overall trend and comparisons of training and testing AUCs of M_p and M_{tr} under varying feature ICC thresholds. Testing AUC increased significantly from 0.56 (0.41-0.70) at baseline (ICC threshold = 0) to the maximum of 0.76 (0.64-0.88, p=0.021) at ICC threshold = 0.9 under perturbation and 0.77 (0.64-0.88, p=0.018) under test-retest. On the other hand, both M_p and M_{tr} demonstrated steady decreases of the training AUCs under increasing ICC thresholds without statistically significant differences to the baseline. Similarly, the baseline model had the lowest reliability with prediction ICC = 0.51 (0.45-0.58) on training perturbation, 0.57 (0.49-0.66) on testing perturbation, and 0.45 (0.25-0.62) on test-retest, as indicated by the lowest bars in **Figure 13**. Significant improvement can be already observed when increasing the feature ICC threshold to 0.5 for both M_p and M_{tr}. The prediction ICC of M_{tr} grew faster than M_p at higher ICC thresholds before reaching the maximum at ICC threshold = 0.9.

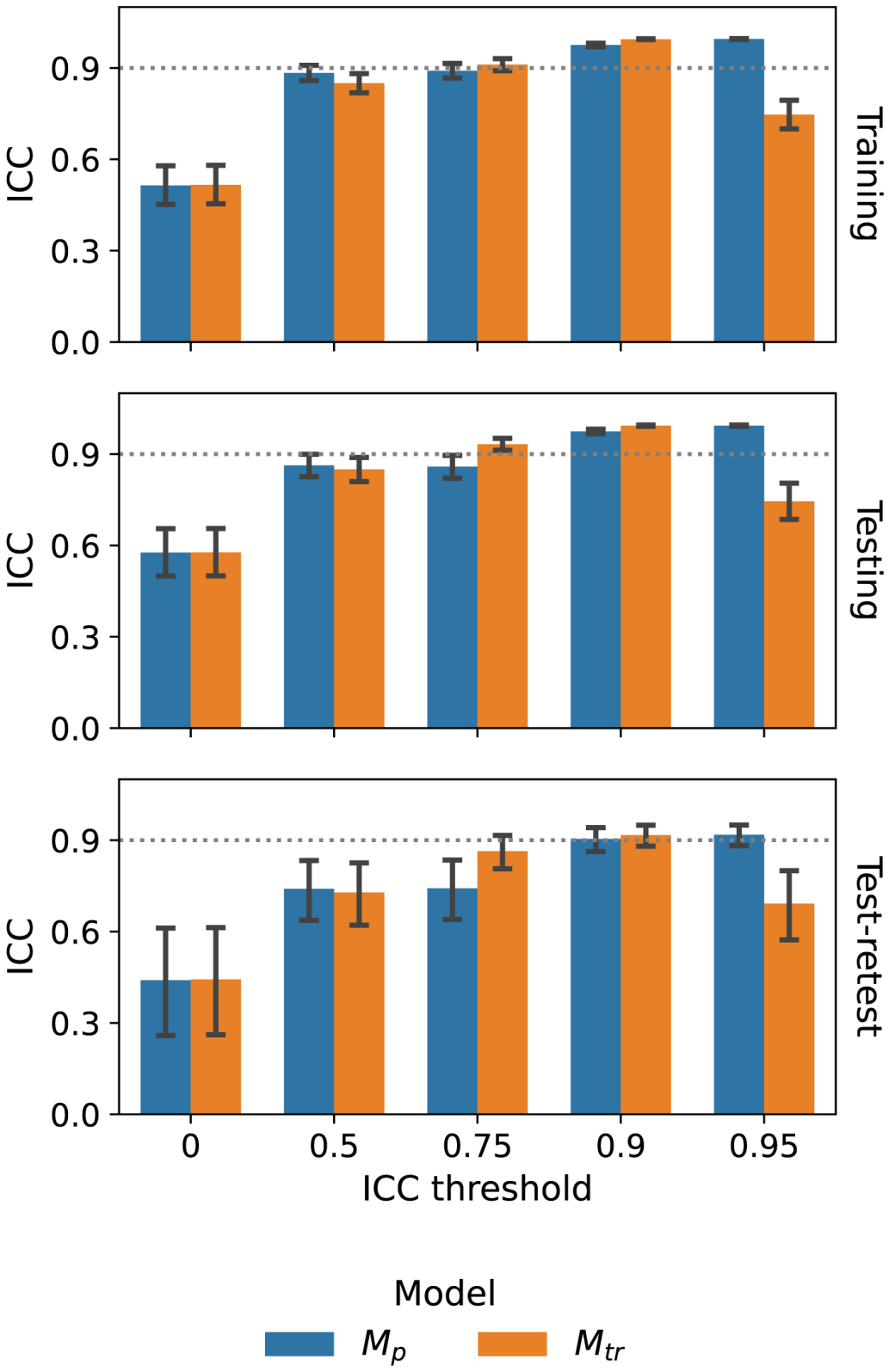
Training
Testing
Test-retest
ICC
0.0
0.3
0.6
0.9
0
0.5
0.75
0.9
0.95
ICC threshold
Model
M_p
M_{tr}

Figure 14. Bar plots for comparing reliability between models based on repeatable features assessed by image perturbations (M_p, blue) and the test-retest imaging (M_{tr}, orange) under varying thresholds. ICC was used to quantify the feature reliability under perturbation for M_p and under tests-retest imaging for M_{tr}. Model reliability was evaluated by probability prediction ICC under perturbation or tests-retest. The error bars indicate 95% confidence intervals acquired during ICC calculation.

M_{tr} demonstrated slightly higher generalizability and significantly higher reliability than M_p on multiple feature ICC filtering thresholds. We observed a smaller training AUCs and higher testing AUCs of M_{tr} at ICC thresholds of 0.5 and 0.75 (**Figure 14**). The AUC differences between M_p and M_{tr} were kept small with the absolute values below 0.05 (p>0.05). Under the ICC threshold of 0.75, M_{tr} had a significantly higher prediction ICC on both testing perturbation (M_{tr}=0.93 (0.91-0.95), M_p=0.86 (0.82-0.90)) and test-retest (M_{tr}=0.87 (0.80-0.92), M_p=0.75 (0.63-0.84)), while no obvious difference found on training perturbation (M_{tr}=0.91 (0.89-0.93), M_p=0.90 (0.86-0.92)), as demonstrated by **Figure 14**. Both the ICC threshold of 0.5 and 0.9 demonstrated minimum model reliability deviations between M_p and M_{tr} with absolute prediction ICC difference < 0.03.

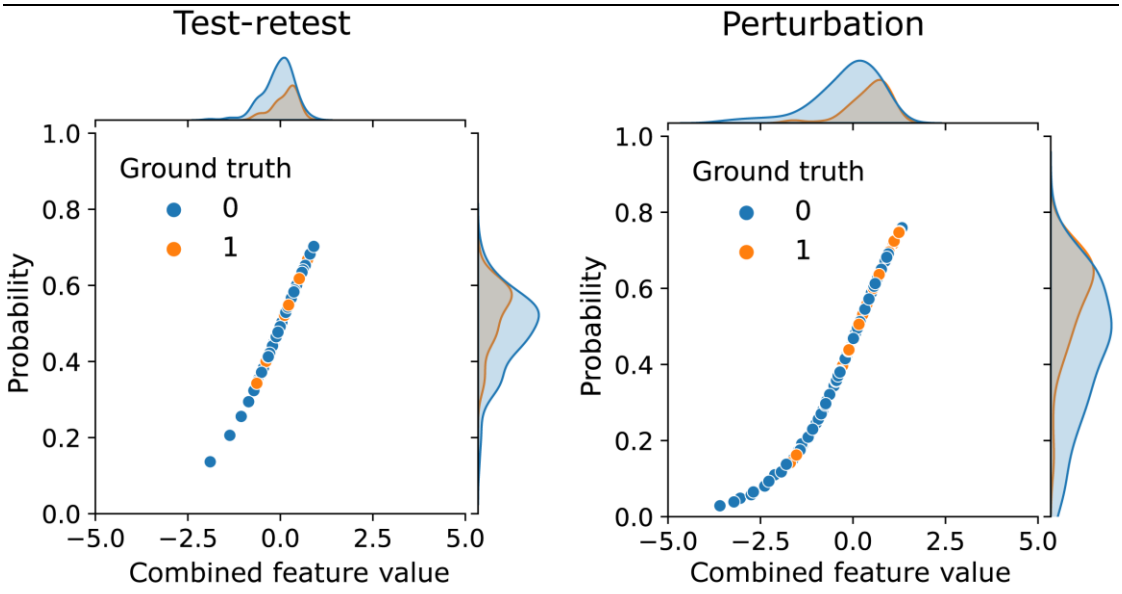

Figure 15. Distributions of the linearly combined feature values and predicted probabilities of the logistic regression models developed from test-retest repeatable features and perturbation repeatable features using the feature ICC threshold of 0.95. The predicted probabilities follow the sigmoid mapping of the logistic regression. Samples with ground-truth of non-event are colored by blue and event by orange. Predictions of the test-retest model were aggregated in the high-slop region whereas a wider spread is found for the perturbation model.

Both M_{tr} generalizability and reliability dropped significantly when increasing the ICC threshold from 0.9 to 0.95. The training AUCs of both M_p and M_{tr} remained table, while a much larger decrease of testing AUC to 0.59 (0.45-0.73) was found for M_{tr} at ICC threshold = 0.95. On the contrary, M_p had a slightly reduced testing AUC to 0.75 (0.62-0.86). Similar to model generalizability, the prediction ICC of M_{tr} fell significantly from above 0.9 to below 0.75 on training perturbation, testing perturbation, and test-retest. On the other hand, the prediction ICC of M_p increased continuously and maximized at ICC threshold = 0.95. **Figure 15** presents the distributions of the predicted probabilities of M_{tr} and M_p combining both training and testing samples at ICC threshold of 0.95. As expected, they both followed the sigmoid mapping as logistic regression from the linearly combined features values. The predictions of M_{tr} are more aggregated in the high-slop region in comparison with M_p with more spread to the lower tail.

4.4. Discussion

This is the first study that directly compared the reliability of radiomic models based on repeatable radiomics features selected by image perturbation and test-retest imaging using ADC maps derived from a publicly available breast cancer DWI dataset. We observed systematically lower radiomic feature reliability assessed by test-retest than perturbations with better binary agreement at higher ICC thresholds. In general, model generalizability and reliability increased continuously with higher ICC thresholds. Similar optimal generalizability and reliability were achieved by the classification model based on perturbation (M_p) and test-retest (M_{tr}) at the ICC threshold of 0.9

simultaneously. Notably, increasing the ICC threshold to 0.95 resulted in significant drops of testing AUC and prediction ICCs for M_{tr}. Our results provide the direct evidence that our perturbation method could replace test-retest method in building a reliable radiomic model with optimal generalizability and reliability.

The lower radiomic feature reliability under test-retest could be largely attributed by the larger variations of tumor segmentations. We further evaluated the segmentation similarities by the Dice similarity coefficients (Dice) and Hausdorff distances (HD) with rigid registrations between test and retest images. The tumor segmentations were less similar between test and retest images (Dice = 0.51(±0.16), HD = 12.47mm(±10.95mm)) than training perturbations (Dice = 0.71(±0.11), HD = 2.72mm(±0.90mm)). Previous research by Saha et al. has also suggested less stable radiomics features from breast MRI within the tumor volume due to a large inter-reader variability (Dice = 0.60) [162]. They also emphasized the necessity of standardization in breast tumor segmentation through precise instructions or auto-contouring, where Dice can be increased to 0.77.

We discovered a positive impact of higher feature reliability on model reliability, as suggested by the increasing testing AUCs and prediction ICCs under higher ICC thresholds. Our results are consistent with the findings by Teng et al. that image perturbation could enhance radiomic model reliability on multiple HNC datasets [151]. A higher model output reliability is generally guaranteed with increased input reliability when using a linear logistic regression model, as long as the predictability is ensured. For model generalizability, a higher ICC threshold could result in less final selected

features with decreased variabilities, thus enhancing the probability of true discovery.

However, the extremely high feature ICC threshold of 0.95 resulted in a much lower model generalizability and reliability under test-retest. During feature selection, only 5 features remained as repeatable for M_{tr}, and none of them showed significant univariate correlation with pCR in training. Consequently, the final selected features had a minimum probability of being truly predictive, and the constructed model was largely overfitted on training with significantly reduced testing AUC. Meanwhile, the predicted probabilities were confounded within the high-slop region (**Figure 15**). Although the selected features and their linear combinations are guaranteed to be highly repeatable (ICC >= 0.95), they could result in larger variations of the prediction values due to the sigmoid transformation. Although generalizability and reliability describe model reliability from two different perspectives, a model built from features with low sensitivity to the prediction target is more likely to have low performances on both due to the previous discussed reasons. Such findings underline the importance of careful selection on repeatable feature criteria when optimizing the final predictive model. A balance between sensitivity and reliability needs to be achieved depending on the level of data standardization during application.

Despite the different reliabilities of M_p and M_{tr} under multiple feature reliability criteria, they both achieved similar optimal generalizability and high reliability at the ICC threshold of 0.9. Such observation provides the direct evidence that perturbation could replace test-retest imaging while achieving the similar optimal model performance. It is advised to incorporate radiomic feature reliability analysis using

image perturbation when test retest is less achievable due to limited medical resources. Nevertheless, the optimal ICC threshold discovered by this study may not be generalizable to other radiomics applications where different image modalities and cancer site were studied and different radiomics features were extracted.

Our study has several limitations that need to be addressed by future investigations. First, only one public dataset was used to conduct this experiment. Further investigations on the generalizability of our findings need to be conducted on other image modalities, cancer sites, and radiomic feature categories. Second, previous studies have also suggested the impact of scanning settings and image preprocessing parameters on radiomic feature reliability [21,37,163]. Therefore, a comprehensive test-retest dataset including different scanners, image acquisition protocols and preprocessing settings is needed to further evaluate the role of perturbation in building a reliable radiomic model. Third, we evaluated model reliability in terms of internal generalizability and model reliability without considering external validation performance. Patient data from multiple institutions could be recruited to further enhance our understandings of the impact of feature reliability on cross-institutional reliability.

4.5. Conclusion

We systematically compared the radiomic model reliability, including both generalizability and reliability, between using repeatable radiomics features assessed by image perturbation and test-retest imaging. The same optimal reliability can be achieved by image perturbation as test-retest imaging. Higher feature reliability

resulted in higher model reliability in general but may have an opposite effect at extremely high reliability threshold. We recommend the radiomic community to include feature reliability analysis using image perturbation in any radiomic study when test-retest is not feasible, but care should be taken when deciding the optimal feature reliability criteria.

Chapter 5. Improving Reliable Radiomics Model Reliability using Image Perturbation for Head and Neck Carcinoma

5.1. Introduction

Radiomics is an emerging artificial intelligence technology that utilizes high-throughput features extracted from imaging features for divulging cancer biological and genetic characteristics [10,13,100,164]. It has demonstrated promises and offered insights with its defined radiomic signatures into cancer diagnosis [165], prognostication [166], treatment response [167] as well as toxicity prediction [168]. Despite a wide range of potential applications in the clinic, a primary concern of radiomics modeling is the reliability of radiomic models. Its clinical applicability has largely been hindered by lacking the assessment on the reliability of features and models [169]. In particular, highly robust radiomics features and models shall remain stable under various imaging conditions and contouring preferences [170,171] if the patient conditions remain unchanged. Features with poor reliability against unexpected changes in imaging and delineation could lead to uncertainties in the downstream radiomics models, leading to non-robust prediction even for the same set of patients [172]. Therefore, using robust radiomics features in the predictive model should be the first and foremost criterion towards clinical application.

Recently, there has been an increasing interest in the investigation of feature reliability. Jin et al. [173] conducted a phantom study to evaluate CT radiomics features' reliability on acquisition parameters and feature extraction parameters. They further

verified the results using patient data. Lee et al. [37] conducted a similar study on MRI radiomics features, and they focused on the feature variability on different MRI scanners and imaging protocol parameters. Sofia et al. [174] assessed the feature reliability of day-by-day test-retest MRI of 14 patients with rectal cancer and reported a wide range of poor repeatable secondary/texture features with the fixed imaging machine, imaging protocol, and machine operator. Lu et al. [175] evaluated the feature reliability of test-retest MRI using a public prostate cancer dataset. They also reported poor repeatable features in different sequences of MRI. Recently, Zwanenburg et al. [65] proposed an image perturbation method to quantify feature reliability and compared their results under test-retest imaging.

Despite extensive research on feature reliability, few studies analyzed the reliability of radiomic models, which includes model reliability and generalizability. One example is the research performed by Parmar et al. [96] on model reliability evaluation against patient subsampling. They observed varied model reliability under different feature selection methods and classifiers. However, there has been no study attempting to correlate feature reliability with the downstream radiomic models by explicitly demonstrating the impact of feature reliability on the model reliability. We hypothesized that radiomic feature reliability could provide additional knowledge in feature reproducibility that improves model reliability which includes model reliability and generalizability. The main challenge of performing a large-scale patient data analysis to test this hypothesis is the high demand for medical resources if the traditional test-retest imaging approach is adopted. Subsequently, we designed our study using the image perturbation method proposed by Zwanenburg et al. [150] to

evaluate both radiomic feature reliability and model reliability. We constructed multiple groups of radiomic models under three robust feature selection criteria using four publicly available HNC datasets, two clinic endpoints, and five classifiers. Both model reliability and generalizability were analyzed and compared their results under the three robust feature selection criteria.

5.2. Materials and Methods

5.2.1. Overview

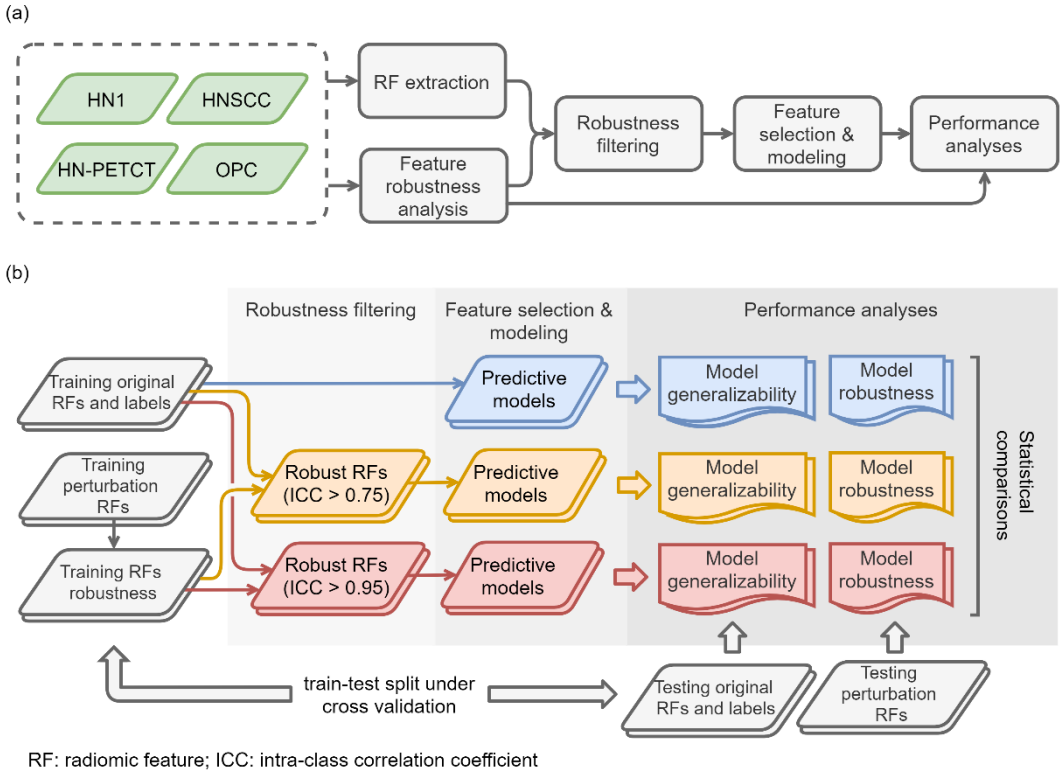

Figure 16. The overall study workflow (a) and model construction and performance analyses workflow (b).

The overall study workflow is summarized in **Figure 16(a)**. Four publicly available datasets of HNC named 1) Head-Neck-Radiomics-HN1 (HN1) [13,110], 2) Head-Neck-PET-CT (HN-PETCT) [107,110], 3) HNSCC [108], 4) Oropharyngeal Carcinoma (OPC)-Radiomics [176,177] were analyzed, and each dataset was used to test our hypothesis independently. Two prediction outcomes, including DM and local-/regional- recurrence (LR), were modeled using five representative commonly used classifiers. The five classifiers include Ridge [178], Support Vector Classifier (SVC) [106], classifiers implementing the k-nearest neighbors algorithm (KNN) [179], Decision Tree [180], and Multilayer Perceptron Network (MLP) [181]. Each dataset was randomly split into multiple training and testing cohorts for repeated stratified cross-validation, and the training cohorts underwent reliability analysis, reliability filtering, and feature selection and modeling. During each cross-validation iteration, the reliability of each radiomic feature was analyzed by image perturbations on the training samples and quantified by ICC. Features with high reliability scores were filtered out and further selected based on outcome relevance and redundancy before model training. To validate the improvements of both model generalizability and reliability using radiomics features with increasing reliability, three groups of radiomic models were constructed without feature reliability filtering, with filtering threshold of 0.75, and with filtering threshold of 0.95, as shown in **Figure 16(b)**. The reliability and generalizability of the three groups of radiomic models were compared statistically. The comparisons were performed independently for the 4 datasets, 2 outcomes, and 5 classifiers, resulting in 40 experiments in total. The improvements of the final selected radiomic feature reliability were also validated through statistical comparisons.

5.2.2. Patient Population

A total of 1,419 HNC patients were recruited from the four publicly available datasets from TCIA [24]. Pre-treatment CT images and their corresponding structure-sets were downloaded in DICOM format from the TCIA website. DM and LR records were also collected as predictive targets for radiomic modeling. They are two critical oncological endpoints in cancer treatment prognosis [182,183] and the common predictive outcomes in many radiomics studies [166,184,185]. The patient consent form has been waived due to the retrospective nature of the study.

In order to ensure data consistency, a set of inclusion criteria were applied. Only patients with available 1) pre-treatment CT images, 2) clinical outcomes record of both DM and LR, and 3) primary GTV contours were included in the study. The identifier of the selected image and the GTVs are also shared in GitHub for replication purposes. Each dataset was split into 60 training and testing set using repeated stratified cross-validation. The folder numbers were chosen in a way that at least two patients in the minority group and 100 patients in total are left for testing to ensure the reliability of the testing performance. The final selected patient numbers, patient distributions for the two prediction outcomes, and train-test split cross-validation methods for the five datasets are listed in **Table 8**.

Table 8. Summary of patient numbers, patient distributions of the two binary prediction outcomes, and the train-test cross-validation methods of the screened patient cohort of the four public datasets.

Dataset name	Total patient No.	Distant metastasis		Local-/regional-recurrence		Cross-validation method
		Event	Non-event	Event	Non-event	
HN1	137	8	129	34	103	Stratified 2-fold, 30 repetitions
HN-PETCT	298	40	258	43	255	Stratified 3-fold, 20 repetitions
HNSCC	460	39	421	65	395	Stratified 4-fold, 15 repetitions
OPC	524	74	450	73	451	Stratified 4-fold, 15 reptations

5.2.3. Image Preprocessing and Radiomic Feature Extraction

Radiomics features were extracted from the pre-treatment CTs within GTVs. The images and GTV contours were preprocessed before extracting features to maintain feature reproducibility and consistency [92,113]. First, CT images were isotopically resampled into 1mm x 1mm x 1mm using B-spline interpolation. The GTV contours were converted into voxel-based segmentation masks according to the resampled CT image grids. Additionally, a re-segmentation mask of the HU range of [-150, 180] was generated for each image to limit the texture feature extraction within soft tissue. All the mentioned preprocessing steps were implemented on Python (3.8) using SimpleITK (1.2.4) [115] and OpenCV [116] packages.

The rest of image preprocessing and radiomic feature extraction were performed using Pyradiomics (2.2.0) [11] package. In addition to the original image, features were extracted from 11 filtered images, including three Laplacian-of-Gaussian (LoG) filtered images (with a sigma value of 1, 3, and 6 mm), and eight coilf1 wavelet filtered images (LLL, HLL, LHL, LLH, LHH, HLH, HHL, HHH). The image intensities of both the original and filtered images were discretized into multiple fixed bin counts of 50, 100, 150, 200, 250, 300, and 350 for texture feature extraction to reduce the feature susceptibility to image noise. A total of 5,486 radiomics features were extracted for each patient. The radiomic feature extraction parameter file for Pyradiomics can be found in the GitHub link.

5.2.4. Feature Reliability Analysis and Filtering

Table 9. The parameters of perturbation modes. AP: anterior-posterior, SI: superior-inferior, LM: lateral-medial.

Perturbation modes	Perturbation range	Reference axis	Perturbation number	Total number
Translation distance (mm)	0 to 3 with a 0.2 step size	AP, SI, LM	4,096	4,423,680
Rotation angles (degree)	-20 to 20 with a 5-step size	SI	9	
Noise addition level	0, 1, 2, 3	-	4	
Contour Randomization	30	-	30	

The reliability of radiomics features were analyzed via the image perturbations in four modes proposed by Zwanenburg et al. [65] with slight modifications. For each perturbation, both the image and mask were translated and rotated simultaneously by a random amount. They aim to simulate the patient position variation during imaging. A random Gaussian noise field was added to the image to mimic the noise level variations between different imaging acquisitions. The GTV mask was also deformed by a randomly generated deformable vector field. It aims to mimic the inter-observer variability during GTV delineation. Multiple parameters of the different perturbation modes were chosen. The translation distances, rotation angles, noise addition levels, and contour randomization parameters were listed in **Table 9**. To explore the perturbations within the specified range as much as possible, 60 perturbations among the entire 4,423,680 combinations of perturbation modes were randomly chosen independently for each patient. The complete set of radiomics features were extracted for the chosen perturbations, and the feature reliability was calculated for each training set using the one-way, random ICC [129,186] with patients as subjects and perturbations as raters. The ICC scores were used to filter out the robust features based on a pre-defined threshold before feature selection and modeling.

5.2.5. Feature Selection and Modeling

A two-step feature selection approach was adopted to obtain the top features that are less redundant and more relevant to the outcome for modeling. First, the outcome relevance of each feature was evaluated by one-way ANOVA P-value repeatedly under down-sampled bootstrapping (imbalanced-learn 0.8.0 [119]) without replacement with

100 iterations on the training set. Features with P-values less than 0.1 were picked out in each iteration and ranked by their frequencies, and the top 10% features with the highest frequencies were chosen. Second, the feature with a higher mean correlation with the rest of the features in each highly correlated feature pair was removed. Pearson correlation coefficient was used to evaluate inter-feature correlation, and the threshold of 0.6 was chosen to identify the feature pairs with high correlations. A maximum of 10 features was further filtered based on the outcome relevance frequency ranking acquired in the previous step. The predictive models were trained from the final selected features using five different classification methods with automatic hyper-parameter tunning. All the model trainings were implemented with the scikit-learn (0.24.0) [187] package.

5.2.6. Performance Analyses

The reliability of the predictive models was evaluated in both generalizability and reliability. Model generalizability evaluates model predictability consistency between the training cohort and unseen cohort. It is quantified as the difference between training and testing predictability which is scored by the AUC. The model reliability metric was designed to evaluate prediction reliability of one patient under varying imaging and contouring conditions. In this study, it is defined as the reliability of the predicted event probability on the perturbed testing samples and calculated by the one-way random ICC. These two performance scores were calculated for all the models generated from the 60 cross-validation iterations and statistically compared between each of the two feature reliability filtering thresholds (ICC > 0.75, ICC > 0.95) and the performance of models

constructed without reliability filtering using pairwise t-test. The comparisons were performed for each dataset, prediction outcome, and modeling classifier independently. Additionally, the reliability of the final selected features with and without reliability filtering was statistically compared by pairwise t-test for each dataset and prediction outcome.

5.3. Results

5.3.1. Feature Reliability and Model Reliability

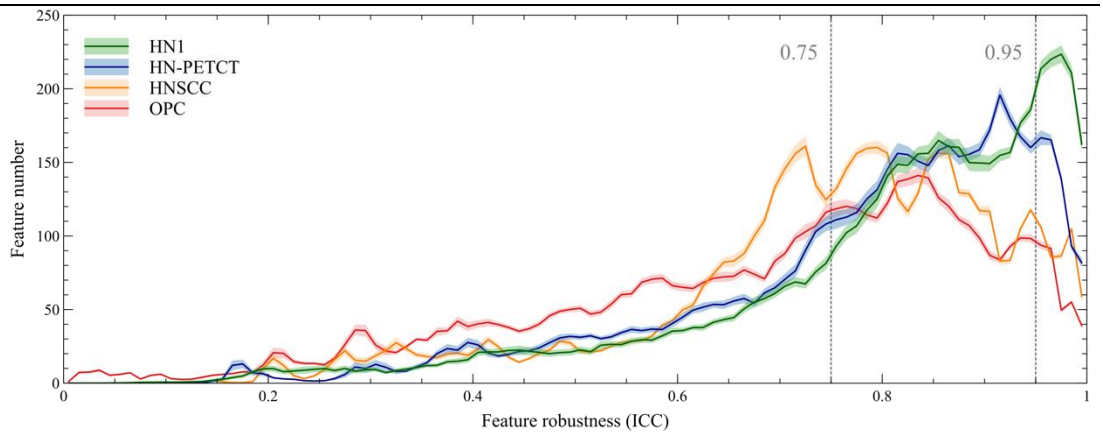

Figure 17. Histograms of the reliability of all the extracted radiomics features for the four analyzed datasets averaged under cross-validations. Feature reliability is quantified as ICC. The shaded areas indicate the 95% confidence interval of the average histogram curves. In general, there are more high-robust features than ones with low reliability. Different datasets show distinctive patterns of feature reliability distributions. HN1 and HN-PETCT have more features with high reliability, whereas HNSCC and OPC have the histograms skewed towards the lower end.

The radiomic feature reliability was quantified by the ICC under image perturbations. The distributions of all the extracted radiomics features show a strong skewness towards higher reliability, as shown by the histograms of feature ICCs for the four datasets in **Figure 17**. Different datasets show distinctive patterns of feature reliability distributions. HN1 (median = 0.84) and HN-PETCT (median = 0.82) has more features with high reliability whereas HNSCC (median = 0.77) and OPC (median = 0.74) have the histograms skewed towards the lower end. On average, 3320/5486 radiomics features remained after being filtered by the threshold of 0.75 and 605/5486 remained for the threshold of 0.95. The final selected radiomics features after the subsequent feature selection procedures showed a significant increase (P-value < 10^{-11}) in mean ICC with increasing feature reliability filtering thresholds. On average, the ICC of the final selected features improved by 0.18 under the filtering threshold of 0.75, and the improvement increased to 0.30 under the threshold of 0.95, as shown by the first column of the heatmaps in **Figure 19(a)**.

The radiomic model reliability showed significant improvements with feature reliability filtering before feature selection and modeling. Model reliability was evaluated by the testing prediction ICC under the same set of image perturbations performed during feature reliability analysis. The prediction ICC of radiomic models constructed without feature reliability filtering is 0.65 averaged overall all the datasets, outcomes, and classifiers. It is increased to 0.78 and 0.91 after feature reliability filtering with ICC > 0.75 and ICC > 0.95 respectively. The detailed results in model reliability improvements and their statistical tests for the four datasets (row) and five classifiers (column) are visualized in the last five columns of the heatmaps in **Figure**

19 separated by outcome and reliability filtering thresholds. Heterogeneous model reliability improvements can be observed in different datasets, classifiers, and prediction outcomes. Higher (ICC > 0.75: 0.045~0.24, ICC > 0.95: 0.11~0.47) and more statistically significant (ICC > 0.75: P-value=9.8 × 10-35~1.1 × 10-2, ICC > 0.95: P-value=8.9 × 10-48~1.2 × 10-8) prediction ICC increases were found with the higher feature reliability filtering threshold in general.

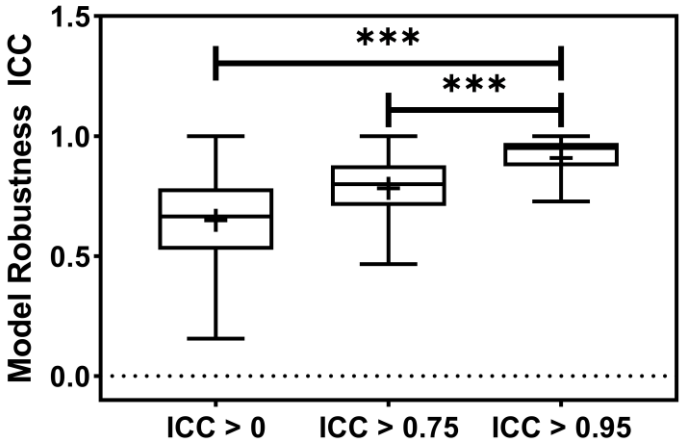

Figure 18. The boxplot shows the model reliability ICC distribution for three feature reliability filtering groups, ICC > 0, ICC > 0.75, and ICC > 0.95. The feature reliability filtering of ICC > 0.95 yields the most robust model. *** indicates the P-value is smaller than 0.0001.

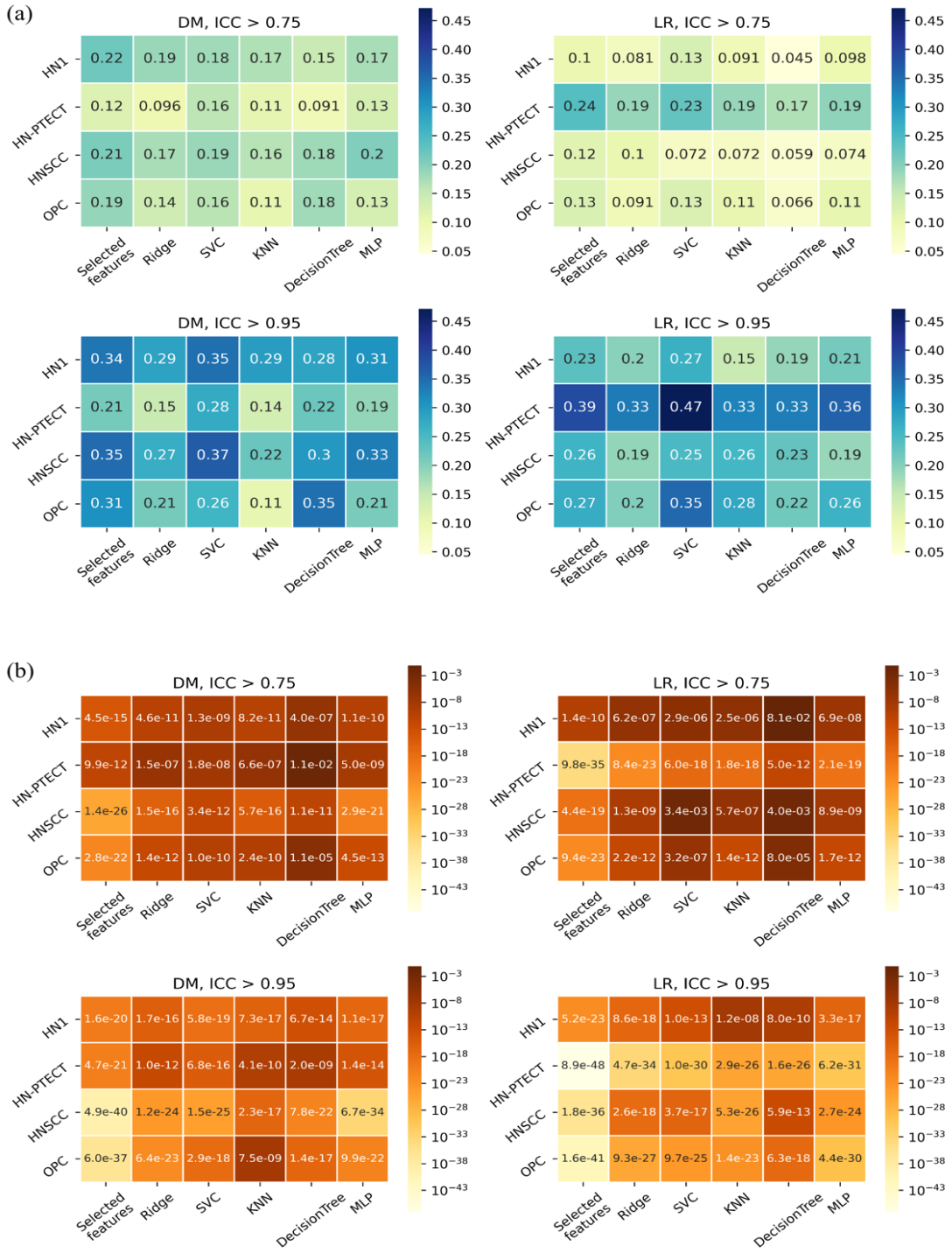

Figure 19. Average ICC improvement (a) and t-test P-values (b) of the final selected features and testing predictions after robust feature pre-selection shown in heatmaps.

Each heatmap contains the results of one prediction outcome and one feature reliability filtering threshold. The first column of each heatmap represents the improvements of the final selected radiomics features, and the remaining five columns are the improvements of the testing prediction reliability using different classifiers. Results of the four datasets are recorded in rows. All the experiments showed positive improvements in ICC. A higher and more statistically significant increase in average ICC improvements can be observed with a higher filtering threshold.

5.3.2. Model Generalizability

Model generalizability is quantified as the difference between the training and testing AUCs, and a lower score indicates better generalizability. The model generalizability score averaged over all the datasets, outcomes, and classifiers are 0.21, 0.18, and 0.12 without reliability filtering, with the filtering threshold of 0.75, and the threshold of 0.95 respectively. In general, model generalizability showed statistically significant improvements after feature reliability filtering on most experiments, as shown by the majority of negative mean generalizability differences and small t-test P-values in **Figure 19**.

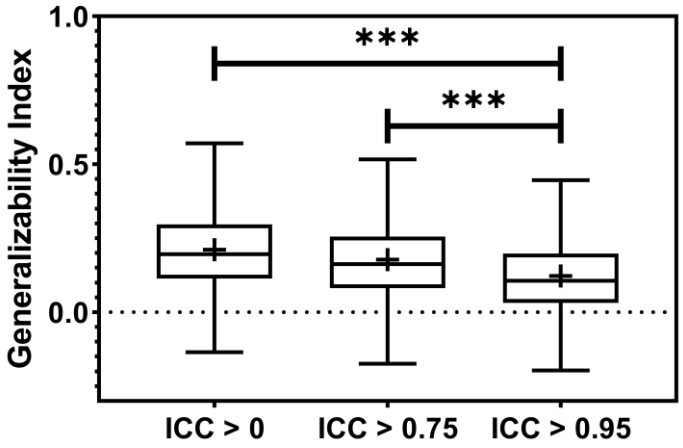

Figure 20. The boxplot showed the train-test performance differences. The most restricted feature reliability filtering provides the most generalizable models. *** indicates the P-value is smaller than 0.0001.

However, the prediction of LR on HN-PETCT had positive mean generalizability differences (ICC > 0.75: -0.026~0.013, ICC > 0.95: -0.025~0.016) for most of the classifiers under both filtering thresholds. Despite the heterogeneous results among datasets, outcomes, and classifiers, larger improvements with higher statistical significance in mode generalizability were observed with the higher feature reliability filtering threshold (ICC > 0.75: -0.06~-0.02, P-value = 7.2×10^{-7}~2.1×10^{-1}; ICC > 0.95: -0.19~-0.054, P-value=4.8×10^{-15}~6.5×10^{-1}) apart from LR models for HN-PETCT. **Figure 20** shows the comparisons of average training and testing AUCs along with its 95% confident interval across the cross-validation models with increasing feature reliability filtering thresholds. Each subfigure contains the results of all the five classifiers shown in different colors and separated by datasets and clinic outcomes. Decreasing training AUCs were observed with increasing filtering thresholds. Specifically, the training AUCs averaged over all the datasets and prediction outcomes without feature reliability filtering, with reliability filtering on ICC > 0.75, and with filtering on ICC > 0.95 are 0.78, 0.76, and 0.69 respectively. Significant drops of training AUCs (pairwise t-test P-values < 0.05) were observed in 33/40 experiments from no feature reliability filtering to the threshold of 0.75 and 40/40 experiments to the threshold of 0.95. Meanwhile, the average testing AUCs are 0.57, 0.58, 0.57 with 18/40 experiments showing statistically significant difference (pairwise t-test P-values < 0.05) for ICC > 0.75 and 24/40 for ICC > 0.95. Different classifiers showed heterogeneous trends of testing AUCs under increasing thresholds. Notably, the testing AUCs of LR radiomic models on HN-PETCT showed significant decrease for feature reliability filtering with ICC > 0.75 (mean decrease: 0.026, 5/5 classifiers with P-value

< 0.05) and ICC > 0.95 (mean decrease: 0.102, 4/5 classifiers with P-value < 0.05), shown in **Figure 21**.

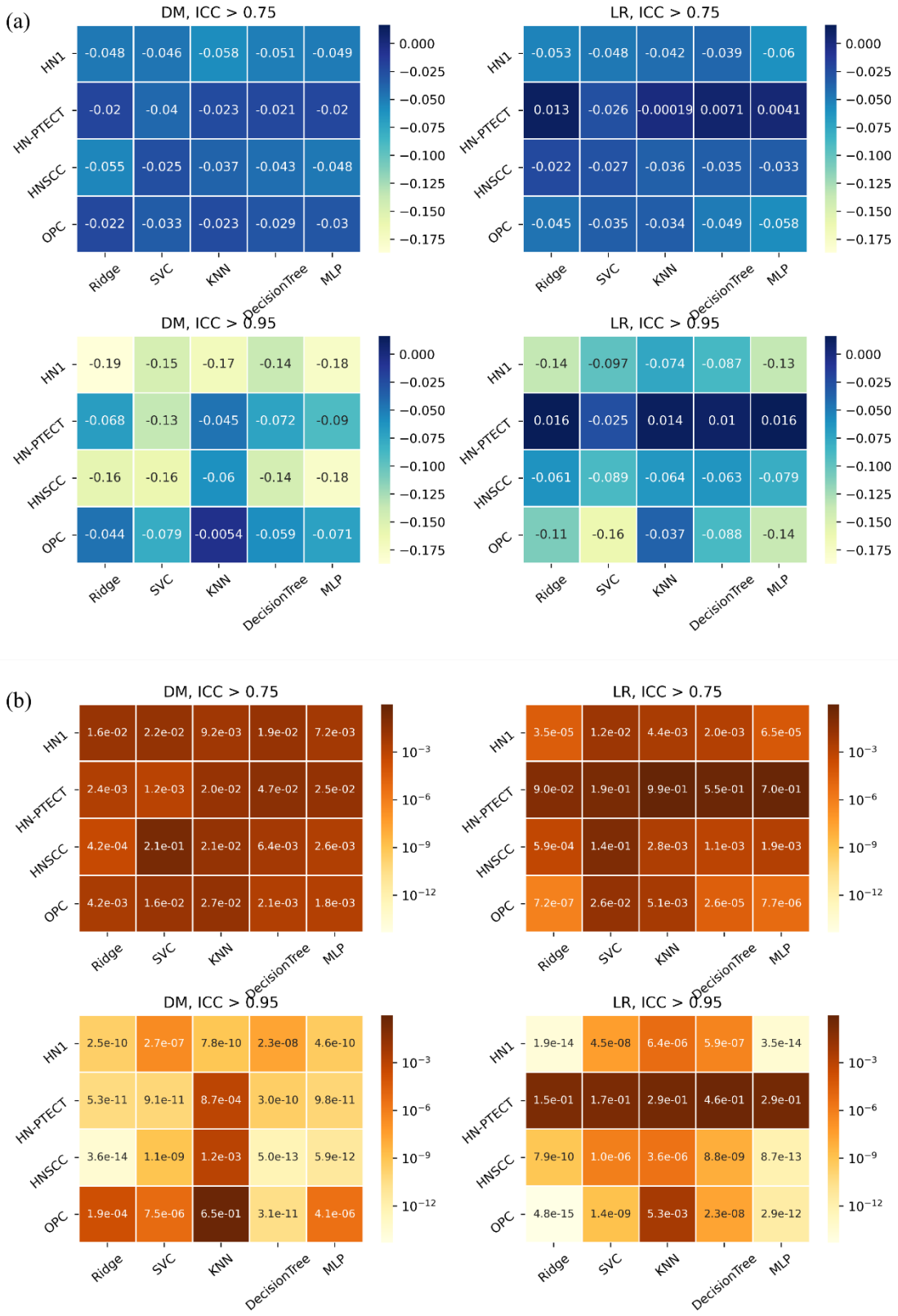
(a)
DM, ICC > 0.75
HN1 -0.048 -0.046 -0.058 -0.051 -0.049
HN-PTECT -0.02 -0.04 -0.023 -0.021 -0.02
HNSCC -0.055 -0.025 -0.037 -0.043 -0.048
OPC -0.022 -0.033 -0.023 -0.029 -0.03
Ridge SVC KNN DecisionTree MLP
0.000 −0.025 −0.050 −0.075 −0.100 −0.125 −0.150 −0.175
LR, ICC > 0.75
HN1 -0.053 -0.048 -0.042 -0.039 -0.06
HN-PTECT 0.013 -0.026 -0.00019 0.0071 0.0041
HNSCC -0.022 -0.027 -0.036 -0.035 -0.033
OPC -0.045 -0.035 -0.034 -0.049 -0.058
Ridge SVC KNN DecisionTree MLP
0.000 −0.025 −0.050 −0.075 −0.100 −0.125 −0.150 −0.175
DM, ICC > 0.95
HN1 -0.19 -0.15 -0.17 -0.14 -0.18
HN-PTECT -0.068 -0.13 -0.045 -0.072 -0.09
HNSCC -0.16 -0.16 -0.06 -0.14 -0.18
OPC -0.044 -0.079 -0.0054 -0.059 -0.071
Ridge SVC KNN DecisionTree MLP
0.000 −0.025 −0.050 −0.075 −0.100 −0.125 −0.150 −0.175
LR, ICC > 0.95
HN1 -0.14 -0.097 -0.074 -0.087 -0.13
HN-PTECT 0.016 -0.025 0.014 0.01 0.016
HNSCC -0.061 -0.089 -0.064 -0.063 -0.079
OPC -0.11 -0.16 -0.037 -0.088 -0.14
Ridge SVC KNN DecisionTree MLP
0.000 −0.025 −0.050 −0.075 −0.100 −0.125 −0.150 −0.175
(b)
DM, ICC > 0.75
HN1 1.6e-02 2.2e-02 9.2e-03 1.9e-02 7.2e-03
HN-PTECT 2.4e-03 1.2e-03 2.0e-02 4.7e-02 2.5e-02
HNSCC 4.2e-04 2.1e-01 2.1e-02 6.4e-03 2.6e-03
OPC 4.2e-03 1.6e-02 2.7e-02 2.1e-03 1.8e-03
Ridge SVC KNN DecisionTree MLP
10^{-3} 10^{-6} 10^{-9} 10^{-12}
LR, ICC > 0.75
HN1 3.5e-05 1.2e-02 4.4e-03 2.0e-03 6.5e-05
HN-PTECT 9.0e-02 1.9e-01 9.9e-01 5.5e-01 7.0e-01
HNSCC 5.9e-04 1.4e-01 2.8e-03 1.1e-03 1.9e-03
OPC 7.2e-07 2.6e-02 5.1e-03 2.6e-05 7.7e-06
Ridge SVC KNN DecisionTree MLP
10^{-3} 10^{-6} 10^{-9} 10^{-12}
DM, ICC > 0.95
HN1 2.5e-10 2.7e-07 7.8e-10 2.3e-08 4.6e-10
HN-PTECT 5.3e-11 9.1e-11 8.7e-04 3.0e-10 9.8e-11
HNSCC 3.6e-14 1.1e-09 1.2e-03 5.0e-13 5.9e-12
OPC 1.9e-04 7.5e-06 6.5e-01 3.1e-11 4.1e-06
Ridge SVC KNN DecisionTree MLP
10^{-3} 10^{-6} 10^{-9} 10^{-12}
LR, ICC > 0.95
HN1 1.9e-14 4.5e-08 6.4e-06 5.9e-07 3.5e-14
HN-PTECT 1.5e-01 1.7e-01 2.9e-01 4.6e-01 2.9e-01
HNSCC 7.9e-10 1.0e-06 3.6e-06 8.8e-09 8.7e-13
OPC 4.8e-15 1.4e-09 5.3e-03 2.3e-08 2.9e-12
Ridge SVC KNN DecisionTree MLP
10^{-3} 10^{-6} 10^{-9} 10^{-12}

Figure 21. Heatmaps on mean model generalizability improvements (a) and statistical test results (b) after feature reliability filtering. Model generalizability is defined as the difference between training and testing AUCs, $AUC_{testing}$ - $AUC_{training}$. A score closer to zero shows better generalizability. In general, model generalizability improved after feature reliability filtering, as shown by the negative values on the heatmaps (a) for both filtering thresholds. Greater improvements were observed with the higher filtering threshold (ICC > 0.95). Moreover, more significant differences are shown by the smaller P-value. However, the predictions of LR on the dataset HN-PETCT showed worse generalizability after feature reliability filtering and the opposite trend of generalizability change and statistical test results with increasing filtering thresholds.

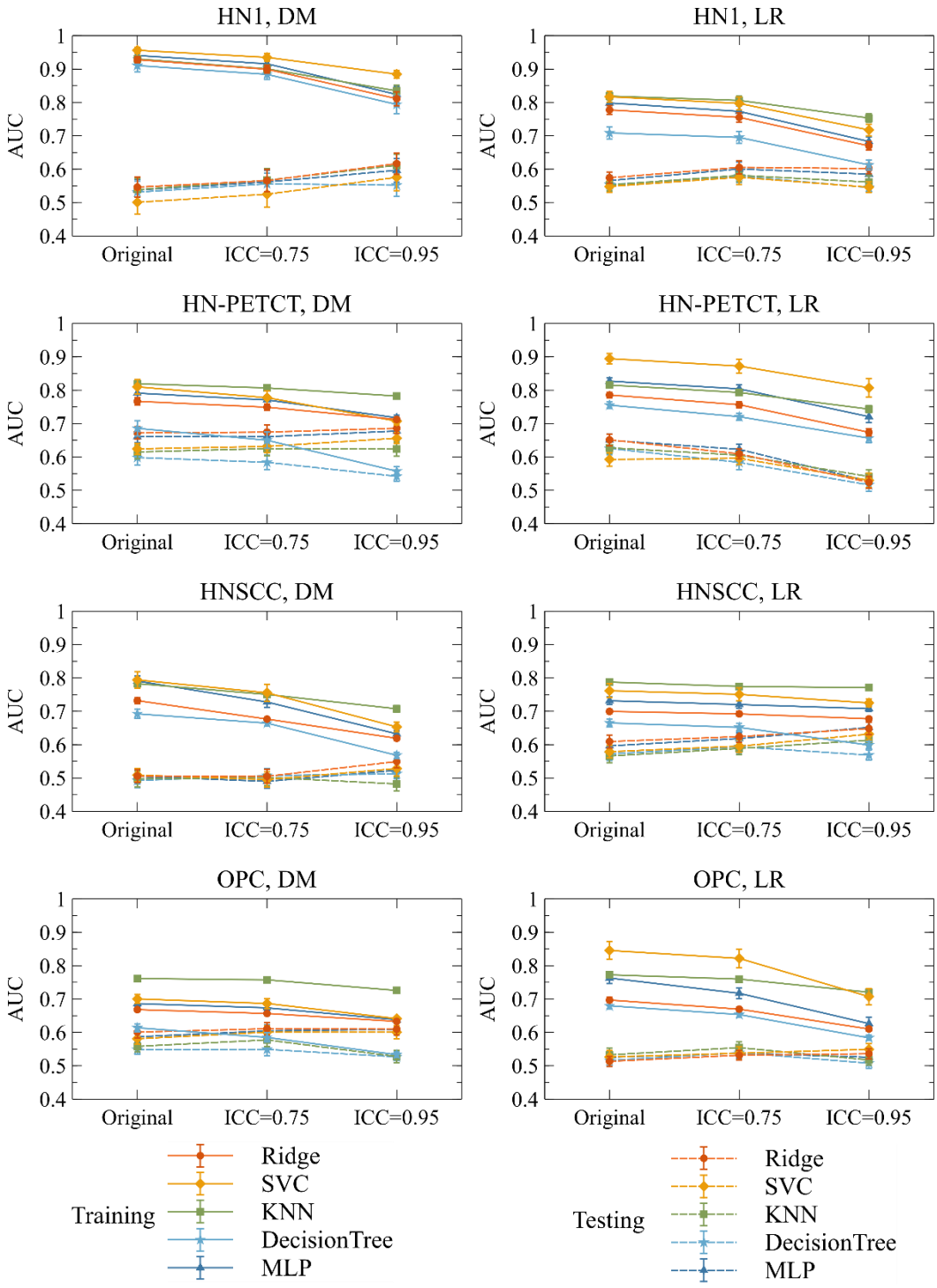

Figure 22. The mean and its 95% confidence interval of the training and testing AUCs of the final constructed models. Each color represents one classifier for modeling. The

solid lines represent the training performances, and the dashed lines represent the testing performances. The 95% confidence intervals are drawn by the error bars. Each subfigure contains the evolution of training/testing AUCs with increasing feature reliability filtering thresholds for one dataset and prediction outcome. A decreasing trend of training AUCs were observed with increasing thresholds for all the datasets, prediction outcomes, and classifiers. The testing AUCs remain stable except for local-regional recurrence prediction on HN-PETCT dataset.

5.3.3. Bias Evaluation

The model reliability improved significantly with the improved feature reliability via the mRMR feature selection, as shown in Table 10, which is consistent with the model reliability improvement with filer-based feature selection. The training AUC also showed a consistent drop with the increase in the threshold of feature reliability, shown in **Table 11**. In contrast, the testing AUC showed an increase or maintaining the same level, resulting in the improved model generalizability.

The bias analysis against the feature selection method showed consistent results between the filter-based and mRMR feature selection methods in improving model reliability and generalizability with robust radiomics features. Therefore, it is unlikely that different feature selection algorithms would affect the conclusion.

Table 10. The model reliability (ICC) for different feature reliability pre-screening thresholds.

Outcomes		ICC > 0	ICC > 0.75	ICC > 0.95
DM	HN1	0.73 (0.66 - 0.79)	0.88 (0.84 - 0.91)	0.95 (0.94 - 0.96)
	HN-PETCT	0.76 (0.71 - 0.80)	0.92 (0.90 - 0.94)	0.92 (0.97 - 0.98)
	HNSCC	0.69 (0.64 - 0.75)	0.78 (0.93 - 0.82)	0.94 (0.93 - 0.96)
	OPC	0.74 (0.70 - 0.79)	0.91 (0.90 - 0.93)	0.99 (0.99 - 0.99)
LR	HN1	0.70 (0.64 - 0.77)	0.86 (0.82 - 0.90)	0.96 (0.95 - 0.98)
	HN-PETCT	0.63 (0.57 - 0.70)	0.81 (0.77 - 0.85)	0.94 (0.92 - 0.95)
	HNSCC	0.73 (0.68 - 0.78)	0.89 (0.86 - 0.91)	0.98 (0.97 - 0.98)
	OPC	0.70 (0.66 - 0.75)	0.84 (0.81 - 0.87)	0.97 (0.97 - 0.98)

Table 11. The training and testing AUC between different feature reliability pre-screening thresholds.

Outcomes		ICC > 0		ICC > 0.75		ICC > 0.95	
		Training AUC	Testing AUC	Training AUC	Testing AUC	Training AUC	Testing AUC
DM	HN1	0.96	0.52	0.92	0.53	0.82	0.60
	HN-PETCT	0.84	0.69	0.82	0.70	0.74	0.70
	HNSCC	0.76	0.53	0.68	0.50	0.63	0.53
	OPC	0.72	0.60	0.68	0.62	0.64	0.62
LR	HN1	0.86	0.57	0.82	0.60	0.70	0.60
	HN-PETCT	0.83	0.62	0.79	0.63	0.70	0.54
	HNSCC	0.74	0.62	0.72	0.64	0.68	0.65
	OPC	0.72	0.52	0.69	0.54	0.61	0.54

5.3.4. Results Summary

Overall, the results demonstrated that both the reliability and generalizability of the final constructed radiomic model are increased when applying feature reliability filtering before feature selection and modeling.

5.4. Discussion

This study demonstrated the impact of feature reliability filtering on the generalizability and reliability of the final constructed HNC radiomic models by comparing the performances under different filtering thresholds. To reduce the bias on patient cohort, prediction outcome, and classifier, four publicly HNC datasets, two prediction outcomes, and five classifiers were used to conduct the experiment. The results showed that pre-screening on the feature reliability before radiomic modeling could increase the reliability of the final selected features and the constructed model against image perturbations. Model generalizability also increased as the consistency of the model predictability between the training and testing datasets has improved, shown in **Figure 22**. Our results confirmed the hypothesis that image perturbations could provide additional knowledge in radiomic feature stability and improve the radiomic models' reliability and generalizability.

Previous literature has discussed the positive impact of robust feature pre-selection on radiomic model generalizability and reliability. For instance, Haarburger et al. [72] envisioned that robust-only features are more likely to lead to a more reliable radiomic model. Vuong et al. [188] obtained a radiomic model with multi-institutional datasets,

which performed equally well as a model on a standardized dataset by including pre-screening on the robust features. Our results confirmed their envision and findings with quantifiable measurements on the improvements in model reliability and generalizability, providing concrete evidence of increased model stability after feature reliability filtering.

The improved model reliability can be explained by the reduced variability of the final selected features after pre-screening on feature reliability, as indicated by the statistically smaller mean feature ICCs. Model output variability is thus reduced as the final selected features are the direct model input. On the other hand, some non-robust features remained after feature selection without feature reliability filtering beforehand. They are more likely to be related to the outcome in the training cohort by chance and less likely to be predictive on the testing cohort or the entire population. Thus, the final constructed models tend to have high AUCs in training but low in testing. The high type I error caused by low feature reliability reduces the power of feature selection in identifying the truly predictive features and lowers the generalizability of the final constructed models. However, a statistically significant reduction (mean: 0.007, P-value < 0.001) in LR prediction generalizability and testing AUCs (mean: 0.1, P-value < 0.001) with pre-selection of robust features on the HN-PETCT dataset is discovered, as shown in **Figure 22**. We found out that one non-robust feature - *wavelet-LHH_glszm_ZoneEntropy* - demonstrated a significant correlation with LR in the entire HN-PETCT cohort with P-value < 0.001. Meanwhile, it is vulnerable against the image perturbations with an ICC of 0.36 (95% CI: [0.32, 0.42]) and thus removed from modeling, resulting in a reduction in overall model predictability and generalizability.

This raises the concern in the limited reliability of testing predictability in representing the model generalizability on the unseen population. To further explain the reduced testing performance, we have calculated the distribution of testing AUCs on the perturbed data and compared with the results on the original data for dataset HN-PETCT and SVC, as visualized in **Figure 23**. Compared with DM predictions, the testing AUCs for LR demonstrated higher variabilities, and the original testing AUCs deviated more to the averaged AUCs under perturbations. Although the original testing AUCs increased statistically (ICC < 0.75: mean increase = 0.02, P-value < 0.01; ICC < 0.95: mean increase = 0.019, P-value < 0.01) after feature reliability filtering for LR, the average testing AUCs showed the opposite trend. The high variability of testing AUCs on LR increases the risk of under-representative testing performance evaluation on the original data, which can be alleviated by feature reliability filtering. Our new findings also support Zwanenburge et al.'s recommendation of using the averaged feature values under image perturbations for modeling.

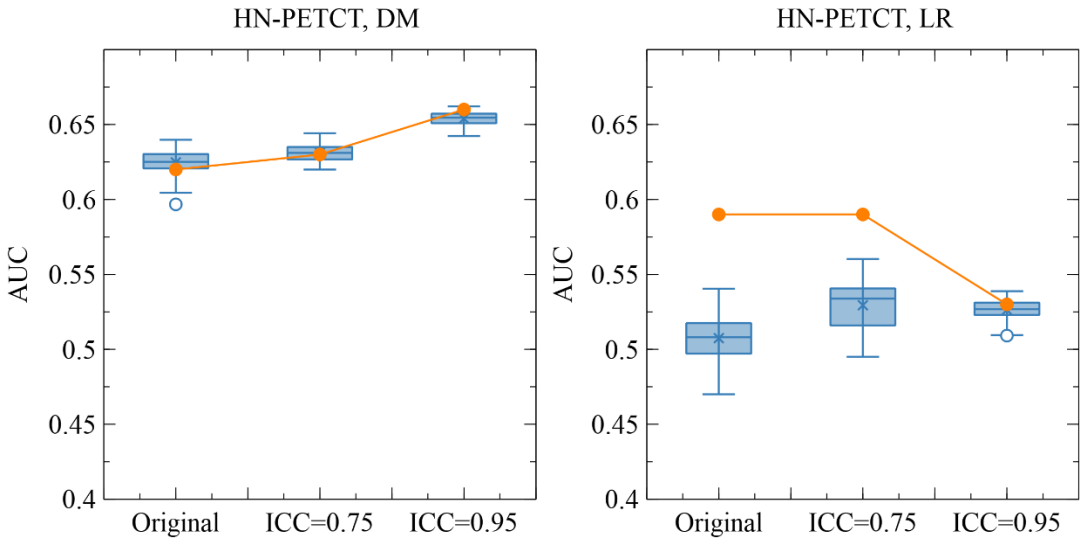

Figure 23. The comparison of the original and perturbed testing AUCs of HN-PETCT-298 averaged over train-test splits for the prediction of DM (a) and LR (b) using SVC. The testing AUCs showed high consistencies between the original images and perturbed images for the prediction of DM while large deviations were observed for the prediction of LR.

Notably, we applied a comprehensive evaluation framework to assess model reliability and generalizability under repeated cross-validations. Instead of only splitting the entire cohort into a single training-testing pair and generating a single model for evaluation, multiple independent train-test splits can give statistical and unbiased evaluations of the impact of radiomic feature reliability on model reliability and generalizability. The main drawback of this method is the high heterogeneity in training and testing performance among iterations [189], which may reduce the statistical significance of our results. We used image perturbations to assess both radiomic feature reliability and model reliability. Although the scope of the image perturbations applied in this study might be limited, and the resulting feature reliability and model reliability is not guaranteed to be as sensitive as test-retest imaging and manual re-contouring, they are rather conservative simulations that impose no additional cost in medical resources and can be easily applied to any dataset. Comprehensive validations of the proposed perturbation method in the future are warranted to increase the credibility of this work. There are other limitations of this study. First, we only considered four datasets of HNC datasets from TCIA and our results may only be generalizable to HNC data. To further generalize the findings to other sites, it is encouraged to test our method on more cancer sites. Second, bias could arise from the single feature selection method, as different criteria and techniques in feature selection have different power in identifying truly predictive radiomics features. It is also suggested to validate our methods with different feature selection methods. The clinical task in this study also poses challenges to this study. As the definition of DM and LR would be affected by the interval of follow-up and diagnostic methods.

5.5. Conclusion

In this study, we aimed to assess the impact of feature reliability on the reliability and generalizability of radiomic models. We found that using robust features improved model reliability and strengthened model generalizability. This work highlighted the significance of radiomic feature reliability when it comes to developing reliable prediction models.

Chapter 6. Discussion

6.1. Advances in Radiomics

The major contribution of this work to the radiomics field is the development of a perturbation-based framework for the evaluation of radiomic model reliability. This has added a brand-new dimension for radiomic model evaluation. Traditional radiomic model evaluation only determines the accuracy of model prediction with respect to the ground truth, whereas the reliability evaluation method developed in this study can also determine the consistency of model prediction when randomness is introduced in the image. Notably, the perturbation-based framework evaluates model reliability against random factors. Most of the reliability studies in the field of radiomics have focused on the reproducibility and reliability of controllable factors such as the scanner brand [133], image acquisition parameters [134], reconstruction kernels [135], and pre-processing parameters [136]. The effects of these factors on radiomic models can be minimised by carefully managing imaging protocols and harmonising image pre-processing. However, the effect of randomness cannot be minimised using current radiomics workflows. Therefore, the effect of randomness should be a priority when evaluating the reliability of radiomic models. Furthermore, the results reported in Chapter 3 revealed vulnerabilities of radiomic models under the influence of random factors. In Chapter 3, model classification performance showed a low training and testing C-index with a perturbed dataset. It appeared that developing a model using only original features would lead to overfitting on randomness if the number of samples is limited. Without perturbed features, the model was unlikely to identify such an overfitting issue

even when an external validation cohort was used. The perturbation-based method can thus serve as a safeguard when evaluating the reliability of radiomic models against randomness. Chapter 3 provides a foundation for further investigations of radiomic model reliability and generalisability.

The standard method of evaluating the reliability of image-based biomarkers involves comparing feature values between test and retest scans. Test-retest scans reflect the true variations introduced by multiple scans, whereas image perturbation methods are based on simulation. Therefore, in Chapter 4, we checked whether perturbation-based methods could replace test-retest scans for evaluating radiomic model reliability. Despite observing systematically lower radiomic feature reliability using the test–retest method than using the perturbation-based method, model generalisability and reliability showed consistent pattern, where an increase of model generalizability and reliability as the ICC thresholds increased. Similar optimal generalisability and reliability were achieved by the classification model based on perturbation (M_p) and test-retest (M_{tr}) scans at an ICC threshold of 0.9. Notably, increasing the ICC threshold to 0.95 significantly reduced the testing AUC and predicted ICCs for M_{tr}. Our results directly prove that the perturbation-based method can replace the test–retest method for building a reliable radiomic model with optimal generalisability and reliability. Furthermore, a positive effect of high feature reliability on model reliability was observed upon increasing both the AUC and ICC of the test set. However, an extremely high feature ICC threshold of 0.95 drastically lowered model generalisability and reliability for the test-retest method. During feature selection, only five features remained repeatable for M_{tr}, none of which showed significant

univariate correlation with pCR during training. Therefore, the selection of reliable radiomics features needs to be investigated to obtain reliable and generalisable models.

Subsequent studies should aim to improve model reliability and generalisability using only reliable features. The results reported in Chapter 4 show that perturbed features can not only be used to evaluate radiomic model reliability but can also help to improve model generalisability by removing low-reliability features. Therefore, in Chapter 5, we aimed to thoroughly evaluate the effect of radiomic feature reliability on radiomic model reliability and generalisability. Our results showed that removing low-robustness features during radiomic model development substantially improved model generalisability to unknown dataset. We used four publicly available datasets to verify the above. Thus, these results show that removing low-reliability radiomics features during radiomic model development improve model generalisability.

Five publicly available datasets from an open respiratory TCIA were included this to improve the transparency and replicability of our conclusions. These datasets minimise barriers to access and to testing the reproducibility of experiments. Furthermore, we repeated our experiments on four datasets, two clinically relevant tasks and five classic classification algorithms. All of the experimental results were consistent, strengthening our results and confirming the positive effect of perturbations in radiomic modelling.

The conclusions of this thesis are not limited to the radiomics or machine learning fields but have strong implications for deep learning and medical image analysis as well. Perturbations in radiomics are analogous to augmented data in deep learning. Instead

of using augmented or perturbed data to train a deep learning model and improve generalisability [190], we used augmented or perturbed data to quantify feature reliability and then enhance generalisability by removing low-reliability features. Similar results have been achieved with different approaches. Importantly, this implies that both fields can cross-reference each other. For example, augmented images can be used to evaluate the reproducibility of deep learning models. In addition, robust deep learning models can be expected to reproduce similar results. Such methods can enhance the explainability of end-to-end models as the inputs and outputs are both images. For instance, if a model can exactly reproduce performance after the translation, rotation or cropping of input images, then the model is deemed translation- or rotation-invariant and can partially explain the outputs of deep learning models.

6.2. Limitations

The main limitation of this study is that the method cannot be implemented using an automated pipeline or software. Although Zwanenburger et al. [155] have thoroughly described the mathematical details of the perturbation-based method, its implementation requires some training in programming and testing. However, the purpose of this study was to not only demonstrate the potential of perturbation in radiomics studies but also highlight its generalisability to other studies. Although the codes used in this thesis and other studies have been shared on GitHub, it may be difficult for others to implement this technique. As in other radiomics studies, the core objective of our work was to extract radiomics features within an ROI; most researchers achieve this using integrated packages in Python, such as PyRadiomics, or in MATLAB,

such as Radiomics Toolbox. Only when the toolbox or integrated packages are made more accessible and user-friendly can the scope of radiomics research expand. One solution to this problem is straightforward packaging by the software, which could allow more uses with different levels of experience to use the technique.

Another limitation of our method is that perturbation is not a comprehensive simulation of all of the variations possible in a radiomic workflow. Perturbations can only simulate patient position variations, contour variations and resampling errors and not variations such as in scanners, reconstruction protocols and inter-observer segmentations. Therefore, the perturbation-based method can only be used to evaluate a subset of variations. However, as there are currently no other alternatives to the test–retest method for such evaluations, the perturbation-based method is the next best option.

Chapter 7. Conclusion

In this study, we have successfully developed a radiomic model reliability evaluation method using perturbation and validated the improvement of radiomic models' reliability and generalizability with removing the low reliable features from radiomic modeling. The method proposed is dataset-specific which does not rely on other medical or human resources to achieve, it takes the randomization factors into the radiomic workflow and improve the radiomic model's reliability against randomizations. To the best of our knowledge, it is the first time the radiomic model reliability is studies in a simulation setting, and it is expected to be one of the solutions facilitating radiomic model reliability problem.

Chapter 8. References

[1] Y. Jiang, H. Wang, J. Wu, C. Chen, Q. Yuan, W. Huang, T. Li, S. Xi, Y. Hu, Z. Zhou, Y. Xu, G. Li, R. Li, Noninvasive imaging evaluation of tumor immune microenvironment to predict outcomes in gastric cancer, Annals of Oncology. 31 (2020) 760–768. https://doi.org/10.1016/j.annonc.2020.03.295.

[2] P. Lambin, R.T.H. Leijenaar, T.M. Deist, J. Peerlings, E.E.C. de Jong, J. van Timmeren, S. Sanduleanu, R.T.H.M. Larue, A.J.G. Even, A. Jochems, Y. van Wijk, H. Woodruff, J. van Soest, T. Lustberg, E. Roelofs, W. van Elmpt, A. Dekker, F.M. Mottaghy, J.E. Wildberger, S. Walsh, Radiomics: the bridge between medical imaging and personalized medicine, Nat Rev Clin Oncol. 14 (2017) 749–762. https://doi.org/10.1038/nrclinonc.2017.141.

[4] M. Sollini, L. Antunovic, A. Chiti, M. Kirienko, Towards clinical application of image mining: a systematic review on artificial intelligence and radiomics, Eur J Nucl Med Mol Imaging. 46 (2019) 2656–2672. https://doi.org/10.1007/s00259-019-04372-x.

[5] M. Fan, P. Zhang, Y. Wang, W. Peng, S. Wang, X. Gao, M. Xu, L. Li, Radiomic analysis of imaging heterogeneity in tumours and the surrounding parenchyma based on unsupervised decomposition of DCE-MRI for predicting molecular subtypes of breast cancer, Eur Radiol. 29 (2019) 4456–4467. https://doi.org/10.1007/s00330-018-5891-3.

[6] A. Saha, M.R. Harowicz, L.J. Grimm, C.E. Kim, S.V. Ghate, R. Walsh, M.A. Mazurowski, A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features, Br J Cancer. 119 (2018) 508–

516. https://doi.org/10.1038/s41416-018-0185-8.

[7] H. Shen, J. Yin, R. Niu, Y. Lian, Y. Huang, C. Tu, D. Liu, X. Wang, X. Lan, X. Yuan, J. Zhang, MRI-based radiomics to compare the survival benefit of induction chemotherapy plus concurrent chemoradiotherapy versus concurrent chemoradiotherapy plus adjuvant chemotherapy in locoregionally advanced nasopharyngeal carcinoma: A multicenter study, Radiotherapy and Oncology. 171 (2022) 107–113. https://doi.org/10.1016/j.radonc.2022.04.017.

[8] P. Yongfeng, J. Chuner, W. Lei, Y. Fengqin, Y. Zhimin, F. Zhenfu, J. Haitao, J. Yangming, W. Fangzheng, The Usefulness of Pretreatment MR-Based Radiomics on Early Response of Neoadjuvant Chemotherapy in Patients With Locally Advanced Nasopharyngeal Carcinoma, Oncol Res. 28 (2021) 605–613. https://doi.org/10.3727/096504020X16022401878096.

[9] M. Vailati-Riboni, V. Palombo, J.J. Loor, What Are Omics Sciences?, in: B.N. Ametaj (Ed.), Periparturient Diseases of Dairy Cows: A Systems Biology Approach, Springer International Publishing, Cham, 2017: pp. 1–7. https://doi.org/10.1007/978-3-319-43033-1_1.

[10] R.J. Gillies, P.E. Kinahan, H. Hricak, Radiomics: Images Are More than Pictures, They Are Data, Radiology. 278 (2016) 563–577. https://doi.org/10.1148/radiol.2015151169.

[11] J.J.M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R.G.H. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, H.J.W.L. Aerts, Computational Radiomics System to Decode the Radiographic Phenotype, Cancer Research. 77 (2017) e104–e107. https://doi.org/10.1158/0008-

5472.CAN-17-0339.

[12] J.H. Thrall, X. Li, Q. Li, C. Cruz, S. Do, K. Dreyer, J. Brink, Artificial Intelligence and Machine Learning in Radiology: Opportunities, Challenges, Pitfalls, and Criteria for Success, Journal of the American College of Radiology. 15 (2018) 504–508. https://doi.org/10.1016/j.jacr.2017.12.026.

[13] H.J.W.L. Aerts, E.R. Velazquez, R.T.H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M.M. Rietbergen, C.R. Leemans, A. Dekker, J. Quackenbush, R.J. Gillies, P. Lambin, Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach, Nat Commun. 5 (2014) 4006. https://doi.org/10.1038/ncomms5006.

[14] A. Ibrahim, S. Primakov, M. Beuque, H.C. Woodruff, I. Halilaj, G. Wu, T. Refaee, R. Granzier, Y. Widaatalla, R. Hustinx, F.M. Mottaghy, P. Lambin, Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework, Methods. 188 (2021) 20–29. https://doi.org/10.1016/j.ymeth.2020.05.022.

[15] R.M. Haralick, K. Shanmugam, I. Dinstein, Textural Features for Image Classification, IEEE Transactions on Systems, Man, and Cybernetics. SMC-3 (1973) 610–621. https://doi.org/10.1109/TSMC.1973.4309314.

[16] C. Sun, W.G. Wee, Neighboring gray level dependence matrix for texture classification, Computer Vision, Graphics, and Image Processing. 23 (1983) 341–352. https://doi.org/10.1016/0734-189X(83)90032-4.

[17] B. Zhao, L.H. Schwartz, M.G. Kris, Data From RIDER_Lung CT, (2015).

https://doi.org/10.7937/K9/TCIA.2015.U1X8A5NR.

[18] B. Zhao, Understanding Sources of Variation to Improve the Reproducibility of Radiomics, Frontiers in Oncology. 11 (2021). https://www.frontiersin.org/article/10.3389/fonc.2021.633176 (accessed May 3, 2022).

[19] B.A. Varghese, D. Hwang, S.Y. Cen, J. Levy, D. Liu, C. Lau, M. Rivas, B. Desai, D.J. Goodenough, V.A. Duddalwar, Reliability of CT-based texture features: Phantom study, Journal of Applied Clinical Medical Physics. 20 (2019) 155–163. https://doi.org/10.1002/acm2.12666.

[20] S.P. Blazis, D.B.M. Dickerscheid, P.V.M. Linsen, C.O.M. Jarnalo, Effect of CT reconstruction settings on the performance of a deep learning based lung nodule CAD system, European Journal of Radiology. 136 (2021). https://doi.org/10.1016/j.ejrad.2021.109526.

[21] D. Mackin, X. Fave, L. Zhang, D. Fried, J. Yang, B. Taylor, E. Rodriguez-Rivera, C. Dodge, A.K. Jones, L. Court, Measuring Computed Tomography Scanner Variability of Radiomics Features, Investigative Radiology. 50 (2015) 757. https://doi.org/10.1097/RLI.0000000000000180.

[22] J. Peerlings, H.C. Woodruff, J.M. Winfield, A. Ibrahim, B.E. Van Beers, A. Heerschap, A. Jackson, J.E. Wildberger, F.M. Mottaghy, N.M. DeSouza, P. Lambin, Stability of radiomics features in apparent diffusion coefficient maps from a multi-centre test-retest trial, Sci Rep. 9 (2019) 4800. https://doi.org/10.1038/s41598-019-41344-5.

[23] S. Fiset, M.L. Welch, J. Weiss, M. Pintilie, J.L. Conway, M. Milosevic, A. Fyles,

A. Traverso, D. Jaffray, U. Metser, J. Xie, K. Han, Repeatability and reproducibility of MRI-based radiomic features in cervical cancer, Radiotherapy and Oncology. 135 (2019) 107–114. https://doi.org/10.1016/j.radonc.2019.03.001.

[24] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, F. Prior, The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository, J Digit Imaging. 26 (2013) 1045–1057. https://doi.org/10.1007/s10278-013-9622-7.

[25] R.F. Cabini, F. Brero, A. Lancia, C. Stelitano, O. Oneta, E. Ballante, E. Puppo, M. Mariani, E. Alì, V. Bartolomeo, M. Montesano, E. Merizzoli, D. Aluia, F. Agustoni, G.M. Stella, R. Sun, L. Bianchini, E. Deutsch, S. Figini, C. Bortolotto, L. Preda, A. Lascialfari, A.R. Filippi, Preliminary report on harmonization of features extraction process using the ComBat tool in the multi-center "Blue Sky Radiomics" study on stage III unresectable NSCLC, Insights into Imaging. 13 (2022) 38. https://doi.org/10.1186/s13244-022-01171-1.

[26] G. Carbonell, P. Kennedy, O. Bane, A. Kirmani, M. El Homsi, D. Stocker, D. Said, P. Mukherjee, O. Gevaert, S. Lewis, S. Hectors, B. Taouli, Precision of MRI radiomics features in the liver and hepatocellular carcinoma, Eur Radiol. 32 (2022) 2030–2040. https://doi.org/10.1007/s00330-021-08282-1.

[27] K. Chen, L. Deng, Q. Li, L. Luo, Are computed-tomography-based hematoma radiomics features reproducible and predictive of intracerebral hemorrhage expansion? an in vitro experiment and clinical study, BJR. 94 (2021) 20200724. https://doi.org/10.1259/bjr.20200724.

[28] Y. Chen, J. Zhong, L. Wang, X. Shi, W. Lu, J. Li, J. Feng, Y. Xia, R. Chang, J. Fan, L. Chen, Y. Zhu, F. Yan, W. Yao, H. Zhang, Robustness of CT radiomics features: consistency within and between single-energy CT and dual-energy CT, Eur Radiol. 32 (2022) 5480–5490. https://doi.org/10.1007/s00330-022-08628-3.

[29] A. Crombé, X. Buy, F. Han, S. Toupin, M. Kind, Assessment of Repeatability, Reproducibility, and Performances of T2 Mapping-Based Radiomics Features: A Comparative Study, Journal of Magnetic Resonance Imaging. 54 (2021) 537–548. https://doi.org/10.1002/jmri.27558.

[30] S. Denzler, D. Vuong, M. Bogowicz, M. Pavic, T. Frauenfelder, S. Thierstein, E.I. Eboulet, B. Maurer, J. Schniering, H.S. Gabryś, I. Schmitt-Opitz, M. Pless, R. Foerster, M. Guckenberger, S. Tanadini-Lang, Impact of CT convolution kernel on robustness of radiomic features for different lung diseases and tissue types, BJR. 94 (2021) 20200947. https://doi.org/10.1259/bjr.20200947.

[31] N. Emaminejad, M.W. Wahi-Anwar, G.H.J. Kim, W. Hsu, M. Brown, M. McNitt-Gray, Reproducibility of lung nodule radiomic features: Multivariable and univariable investigations that account for interactions between CT acquisition and reconstruction parameters, Medical Physics. 48 (2021) 2906–2919. https://doi.org/10.1002/mp.14830.

[32] A. Euler, F.C. Laqua, D. Cester, N. Lohaus, T. Sartoretti, D. Pinto dos Santos, H. Alkadhi, B. Baessler, Virtual Monoenergetic Images of Dual-Energy CT—Impact on Repeatability, Reproducibility, and Classification in Radiomics, Cancers. 13 (2021) 4710. https://doi.org/10.3390/cancers13184710.

[33] Y. Gao, M. Hua, J. Lv, Y. Ma, Y. Liu, M. Ren, Y. Tian, X. Li, H. Zhang,

Reproducibility of radiomic features of pulmonary nodules between low-dose CT and conventional-dose CT, Quant Imaging Med Surg. 12 (2022) 2368–2377. https://doi.org/10.21037/qims-21-609.

[34] R. w. y. Granzier, A. Ibrahim, S. Primakov, S. a. Keek, I. Halilaj, A. Zwanenburg, S. m. e. Engelen, M. b. i. Lobbes, P. Lambin, H. c. Woodruff, M. l. Smidt, Test–Retest Data for the Assessment of Breast MRI Radiomic Feature Repeatability, Journal of Magnetic Resonance Imaging. 56 (2022) 592–604. https://doi.org/10.1002/jmri.28027.

[35] A. Ibrahim, Y. Widaatalla, T. Refaee, S. Primakov, R.L. Miclea, O. Öcal, M.P. Fabritius, M. Ingrisch, J. Ricke, R. Hustinx, F.M. Mottaghy, H.C. Woodruff, M. Seidensticker, P. Lambin, Reproducibility of CT-Based Hepatocellular Carcinoma Radiomic Features across Different Contrast Imaging Phases: A Proof of Concept on SORAMIC Trial Data, Cancers. 13 (2021) 4638. https://doi.org/10.3390/cancers13184638.

[36] A. Ibrahim, T. Refaee, R.T.H. Leijenaar, S. Primakov, R. Hustinx, F.M. Mottaghy, H.C. Woodruff, A.D.A. Maidment, P. Lambin, The application of a workflow integrating the variable reproducibility and harmonizability of radiomic features on a phantom dataset, PLOS ONE. 16 (2021) e0251147. https://doi.org/10.1371/journal.pone.0251147.

[37] J. Lee, A. Steinmann, Y. Ding, H. Lee, C. Owens, J. Wang, J. Yang, D. Followill, R. Ger, D. MacKin, L.E. Court, Radiomics feature robustness as measured using an MRI phantom, Sci Rep. 11 (2021) 3973. https://doi.org/10.1038/s41598-021-83593-3.

[38] S. Lennartz, A. O'Shea, A. Parakh, T. Persigehl, B. Baessler, A. Kambadakone, Robustness of dual-energy CT-derived radiomic features across three different scanner types, Eur Radiol. 32 (2022) 1959–1970. https://doi.org/10.1007/s00330-021-08249-2.

[39] R.N. Mahon, G D Hugo, E. Weiss, Repeatability of texture features derived from magnetic resonance and computed tomography imaging and use in predictive models for non-small cell lung cancer outcome, Phys. Med. Biol. 64 (2019) 145007. https://doi.org/10.1088/1361-6560/ab18d3.

[40] D.J. McHugh, N. Porta, R.A. Little, S. Cheung, Y. Watson, G.J.M. Parker, G.C. Jayson, J.P.B. O'Connor, Image Contrast, Image Pre-Processing, and T1 Mapping Affect MRI Radiomic Feature Repeatability in Patients with Colorectal Cancer Liver Metastases, Cancers. 13 (2021) 240. https://doi.org/10.3390/cancers13020240.

[41] M. Meyer, J. Ronald, F. Vernuccio, R.C. Nelson, J.C. Ramirez-Giraldo, J. Solomon, B.N. Patel, E. Samei, D. Marin, Reproducibility of CT Radiomic Features within the Same Patient: Influence of Radiation Dose and CT Reconstruction Settings, Radiology. 293 (2019) 583–591. https://doi.org/10.1148/radiol.2019190928.

[42] R.N. Mitchell-Hay, T.S. Ahearn, A.D. Murray, G.D. Waiter, Investigation of the Inter- and Intrascanner Reproducibility and Repeatability of Radiomics Features in T1-Weighted Brain MRI, Journal of Magnetic Resonance Imaging. 56 (2022) 1559–1568. https://doi.org/10.1002/jmri.28191.

[43] A. Nazeri, J.P. Crandall, T.J. Fraum, R.L. Wahl, Repeatability of Radiomic

Features of Brown Adipose Tissue, Journal of Nuclear Medicine. 62 (2021) 700–706. https://doi.org/10.2967/jnumed.120.248674.

[44] U. Pandey, J. Saini, M. Kumar, R. Gupta, M. Ingalhalikar, Normative Baseline for Radiomics in Brain MRI: Evaluating the Robustness, Regional Variations, and Reproducibility on FLAIR Images, Journal of Magnetic Resonance Imaging. 53 (2021) 394–407. https://doi.org/10.1002/jmri.27349.

[45] T. Perrin, A. Midya, R. Yamashita, J. Chakraborty, T. Saidon, W.R. Jarnagin, M. Gonen, A.L. Simpson, R.K.G. Do, Short-term reproducibility of radiomic features in liver parenchyma and liver malignancies on contrast-enhanced CT imaging, Abdom Radiol. 43 (2018) 3271–3278. https://doi.org/10.1007/s00261-018-1600-6.

[46] F. Prayer, J. Hofmanninger, M. Weber, D. Kifjak, A. Willenpart, J. Pan, S. Röhrich, G. Langs, H. Prosch, Variability of computed tomography radiomics features of fibrosing interstitial lung disease: A test-retest study, Methods. 188 (2021) 98–104. https://doi.org/10.1016/j.ymeth.2020.08.007.

[47] Z. Raisi-Estabragh, P. Gkontra, A. Jaggi, J. Cooper, J. Augusto, A.N. Bhuva, R.H. Davies, C.H. Manisty, J.C. Moon, P.B. Munroe, N.C. Harvey, K. Lekadir, S.E. Petersen, Repeatability of Cardiac Magnetic Resonance Radiomics: A Multi-Centre Multi-Vendor Test-Retest Study, Frontiers in Cardiovascular Medicine. 7 (2020). https://www.frontiersin.org/articles/10.3389/fcvm.2020.586236 (accessed January 5, 2023).

[48] T. Refaee, Z. Salahuddin, Y. Widaatalla, S. Primakov, H.C. Woodruff, R.

Hustinx, F.M. Mottaghy, A. Ibrahim, P. Lambin, CT Reconstruction Kernels and the Effect of Pre- and Post-Processing on the Reproducibility of Handcrafted Radiomic Features, Journal of Personalized Medicine. 12 (2022) 553. https://doi.org/10.3390/jpm12040553.

[49] R. Reiazi, C. Arrowsmith, M. Welch, F. Abbas-Aghababazadeh, C. Eeles, T. Tadic, A.J. Hope, S.V. Bratman, B. Haibe-Kains, Prediction of Human Papillomavirus (HPV) Association of Oropharyngeal Cancer (OPC) Using Radiomics: The Impact of the Variation of CT Scanner, Cancers. 13 (2021) 2269. https://doi.org/10.3390/cancers13092269.

[50] L. Rinaldi, S.P. De Angelis, S. Raimondi, S. Rizzo, C. Fanciullo, C. Rampinelli, M. Mariani, A. Lascialfari, M. Cremonesi, R. Orecchia, D. Origgi, F. Botta, Reproducibility of radiomic features in CT images of NSCLC patients: an integrative analysis on the impact of acquisition and reconstruction parameters, Eur Radiol Exp. 6 (2022) 2. https://doi.org/10.1186/s41747-021-00258-6.

[51] L. Escudero Sanchez, L. Rundo, A.B. Gill, M. Hoare, E. Mendes Serrao, E. Sala, Robustness of radiomic features in CT images with different slice thickness, comparing liver tumour and muscle, Sci Rep. 11 (2021) 8262. https://doi.org/10.1038/s41598-021-87598-w.

[52] M. Sun, A. Baiyasi, X. Liu, X. Shi, X. Li, J. Zhu, Y. Yin, J. Hu, Z. Li, B. Li, Robustness and reproducibility of radiomics in T2 weighted images from magnetic resonance image guided linear accelerator in a phantom study, Physica Medica. 96 (2022) 130–139. https://doi.org/10.1016/j.ejmp.2022.03.002.

[53] C. Xue, J. Yuan, D.M. Poon, Y. Zhou, B. Yang, S.K. Yu, Y.K. Cheung,

Reliability of MRI radiomics features in MR-guided radiotherapy for prostate cancer: Repeatability, reproducibility, and within-subject agreement, Medical Physics. 48 (2021) 6976–6986. https://doi.org/10.1002/mp.15232.

[54] C. Xue, Y. Zhou, G.G. Lo, O.L. Wong, S.K. Yu, K.Y. Cheung, J. Yuan, Reliability of radiomics features due to image reconstruction using a standardized T2-weighted pulse sequence for MR-guided radiotherapy: An anthropomorphic phantom study, Magnetic Resonance in Medicine. 85 (2021) 3434–3446. https://doi.org/10.1002/mrm.28650.

[55] D. Alis, M. Yergin, O. Asmakutlu, C. Topel, E. Karaarslan, The influence of cardiac motion on radiomics features: radiomics features of non-enhanced CMR cine images greatly vary through the cardiac cycle, Eur Radiol. 31 (2021) 2706–2715. https://doi.org/10.1007/s00330-020-07370-y.

[56] M. Bologna, V.D.A. Corino, L.T. Mainardi, Assessment of the effect of intensity standardization on the reliability of T1-weighted MRI radiomic features: experiment on a virtual phantom, in: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2019: pp. 413–416. https://doi.org/10.1109/EMBC.2019.8857825.

[57] K.V. Hoebel, J.B. Patel, A.L. Beers, K. Chang, P. Singh, J.M. Brown, M.C. Pinho, T.T. Batchelor, E.R. Gerstner, B.R. Rosen, J. Kalpathy-Cramer, Radiomics Repeatability Pitfalls in a Scan-Rescan MRI Study of Glioblastoma, Radiology: Artificial Intelligence. 3 (2021) e190199. https://doi.org/10.1148/ryai.2020190199.

[58] K. Hu, W. Deng, N. Li, Q. Cai, Z. Yuan, L. Li, Y. Liu, Impact of Parallel

Acquisition Technology on the Robustness of Magnetic Resonance Imaging Radiomic Features, Journal of Computer Assisted Tomography. 46 (2022) 906. https://doi.org/10.1097/RCT.0000000000001344.

[59] S.B. Lee, Y.J. Cho, Y. Hong, D. Jeong, J. Lee, S.-H. Kim, S. Lee, Y.H. Choi, Deep Learning-Based Image Conversion Improves the Reproducibility of Computed Tomography Radiomics Features : A Phantom Study, Investigative Radiology. 57 (2022) 308–317. https://doi.org/10.1097/RLI.0000000000000839.

[60] H. Muenzfeld, C. Nowak, S. Riedlberger, A. Hartenstein, B. Hamm, P. Jahnke, T. Penzkofer, Intra-scanner repeatability of quantitative imaging features in a 3D printed semi-anthropomorphic CT phantom, European Journal of Radiology. 141 (2021) 109818. https://doi.org/10.1016/j.ejrad.2021.109818.

[61] H.M. Whitney, K. Drukker, A. Edwards, J. Papaioannou, M. Medved, G. Karczmar, M.L. Giger, Robustness of radiomic features of benign breast lesions and hormone receptor positive/HER2-negative cancers across DCE-MR magnet strengths, Magnetic Resonance Imaging. 82 (2021) 111–121. https://doi.org/10.1016/j.mri.2021.06.021.

[62] J.E. van Timmeren, D. Cester, S. Tanadini-Lang, H. Alkadhi, B. Baessler, Radiomics in medical imaging—"how-to" guide and critical reflection, Insights into Imaging. 11 (2020) 91. https://doi.org/10.1186/s13244-020-00887-2.

[63] C. Parmar, E.R. Velazquez, R. Leijenaar, M. Jermoumi, S. Carvalho, R.H. Mak, S. Mitra, B.U. Shankar, R. Kikinis, B. Haibe-Kains, P. Lambin, H.J.W.L. Aerts, Robust Radiomics Feature Quantification Using Semiautomatic Volumetric Segmentation, PLOS ONE. 9 (2014) e102107.

https://doi.org/10.1371/journal.pone.0102107.

[64] L. Perna, C. Cozzarini, E. Maggiulli, G. Fellin, T. Rancati, R. Valdagni, V. Vavassori, S. Villa, C. Fiorino, Inter-observer variability in contouring the penile bulb on CT images for prostate cancer treatment planning, Radiation Oncology. 6 (2011) 123. https://doi.org/10.1186/1748-717X-6-123.

[65] A. Zwanenburg, S. Leger, L. Agolli, K. Pilz, E.G.C. Troost, C. Richter, S. Löck, Assessing robustness of radiomic features by image perturbation, Sci Rep. 9 (2019) 614. https://doi.org/10.1038/s41598-018-36938-4.

[66] F. Bianconi, M.L. Fravolini, I. Palumbo, G. Pascoletti, S. Nuvoli, M. Rondini, A. Spanu, B. Palumbo, Impact of Lesion Delineation and Intensity Quantisation on the Stability of Texture Features from Lung Nodules on CT: A Reproducible Study, Diagnostics. 11 (2021) 1224. https://doi.org/10.3390/diagnostics11071224.

[67] H. Chen, Y. He, C. Zhao, L. Zheng, N. Pan, J. Qiu, Z. Zhang, X. Niu, Z. Yuan, Reproducibility of radiomics features derived from intravoxel incoherent motion diffusion-weighted MRI of cervical cancer, Acta Radiol. 62 (2021) 679–686. https://doi.org/10.1177/0284185120934471.

[68] J. Duan, Q. Qiu, J. Zhu, D. Shang, X. Dou, T. Sun, Y. Yin, X. Meng, Reproducibility for Hepatocellular Carcinoma CT Radiomic Features: Influence of Delineation Variability Based on 3D-CT, 4D-CT and Multiple-Parameter MR Images, Front Oncol. 12 (2022) 881931. https://doi.org/10.3389/fonc.2022.881931.

[69] S. Gitto, R. Cuocolo, I. Emili, L. Tofanelli, V. Chianca, D. Albano, C. Messina,

M. Imbriaco, L.M. Sconfienza, Effects of Interobserver Variability on 2D and 3D CT- and MRI-Based Texture Feature Reproducibility of Cartilaginous Bone Tumors, J Digit Imaging. 34 (2021) 820–832. https://doi.org/10.1007/s10278-021-00498-3.

[70] S. Gitto, M. Bologna, V.D.A. Corino, I. Emili, D. Albano, C. Messina, E. Armiraglio, A. Parafioriti, A. Luzzati, L. Mainardi, L.M. Sconfienza, Diffusion-weighted MRI radiomics of spine bone tumors: feature stability and machine learning-based classification performance, Radiol Med. 127 (2022) 518–525. https://doi.org/10.1007/s11547-022-01468-7.

[71] R.W.Y. Granzier, N.M.H. Verbakel, A. Ibrahim, J.E. van Timmeren, T.J.A. van Nijnatten, R.T.H. Leijenaar, M.B.I. Lobbes, M.L. Smidt, H.C. Woodruff, MRI-based radiomics in breast cancer: feature robustness with respect to inter-observer segmentation variability, Sci Rep. 10 (2020) 14163. https://doi.org/10.1038/s41598-020-70940-z.

[72] C. Haarburger, G. Müller-Franzes, L. Weninger, C. Kuhl, D. Truhn, D. Merhof, Radiomics feature reproducibility under inter-rater variability in segmentations of CT images, Sci Rep. 10 (2020) 12688. https://doi.org/10.1038/s41598-020-69534-6.

[73] N.S.M. Haniff, M.K. Abdul Karim, N.H. Osman, M.I. Saripan, I.N. Che Isa, M.J. Ibahim, Stability and Reproducibility of Radiomic Features Based Various Segmentation Technique on MR Images of Hepatocellular Carcinoma (HCC), Diagnostics. 11 (2021) 1573. https://doi.org/10.3390/diagnostics11091573.

[74] L.J. Jensen, D. Kim, T. Elgeti, I.G. Steffen, B. Hamm, S.N. Nagel, Stability of

Radiomic Features across Different Region of Interest Sizes—A CT and MR Phantom Study, Tomography. 7 (2021) 238–252. https://doi.org/10.3390/tomography7020022.

[75] L.C. Kelahan, D. Kim, M. Soliman, R.J. Avery, H. Savas, R. Agrawal, M. Magnetta, B.P. Liu, Y.S. Velichko, Role of hepatic metastatic lesion size on inter-reader reproducibility of CT-based radiomics features, Eur Radiol. 32 (2022) 4025–4033. https://doi.org/10.1007/s00330-021-08526-0.

[76] B. Kocak, E.S. Durmaz, O.K. Kaya, E. Ates, O. Kilickesmez, Reliability of Single-Slice–Based 2D CT Texture Analysis of Renal Masses: Influence of Intra- and Interobserver Manual Segmentation Variability on Radiomic Feature Reproducibility, American Journal of Roentgenology. 213 (2019) 377–383. https://doi.org/10.2214/AJR.19.21212.

[77] E.P.V. Le, L. Rundo, J.M. Tarkin, N.R. Evans, M.M. Chowdhury, P.A. Coughlin, H. Pavey, C. Wall, F. Zaccagna, F.A. Gallagher, Y. Huang, R. Sriranjan, A. Le, J.R. Weir-McCall, M. Roberts, F.J. Gilbert, E.A. Warburton, C.-B. Schönlieb, E. Sala, J.H.F. Rudd, Assessing robustness of carotid artery CT angiography radiomics in the identification of culprit lesions in cerebrovascular events, Sci Rep. 11 (2021) 3499. https://doi.org/10.1038/s41598-021-82760-w.

[78] G. Müller-Franzes, S. Nebelung, J. Schock, C. Haarburger, F. Khader, F. Pedersoli, M. Schulze-Hagen, C. Kuhl, D. Truhn, Reliability as a Precondition for Trust—Segmentation Reliability Analysis of Radiomic Features Improves Survival Prediction, Diagnostics. 12 (2022) 247. https://doi.org/10.3390/diagnostics12020247.

[79] N.W. Schurink, S.R. van Kranen, S. Roberti, J.J.M. van Griethuysen, N. Bogveradze, F. Castagnoli, N. el Khababi, F.C.H. Bakers, S.H. de Bie, G.P.T. Bosma, V.C. Cappendijk, R.W.F. Geenen, P.A. Neijenhuis, G.M. Peterson, C.J. Veeken, R.F.A. Vliegen, R.G.H. Beets-Tan, D.M.J. Lambregts, Sources of variation in multicenter rectal MRI data and their effect on radiomics feature reproducibility, Eur Radiol. 32 (2022) 1506–1516. https://doi.org/10.1007/s00330-021-08251-8.

[80] F. Tixier, H. Um, R.J. Young, H. Veeraraghavan, Reliability of tumor segmentation in glioblastoma: Impact on the robustness of MRI-radiomic features, Medical Physics. 46 (2019) 3582–3591. https://doi.org/10.1002/mp.13624.

[81] I. Tunali, L.O. Hall, S. Napel, D. Cherezov, A. Guvenis, R.J. Gillies, M.B. Schabath, Stability and reproducibility of computed tomography radiomic features extracted from peritumoral regions of lung cancer lesions, Medical Physics. 46 (2019) 5075–5085. https://doi.org/10.1002/mp.13808.

[82] F. Urraro, V. Nardone, A. Reginelli, C. Varelli, A. Angrisani, V. Patanè, L. D'Ambrosio, P. Roccatagliata, G.M. Russo, L. Gallo, M. De Chiara, L. Altucci, S. Cappabianca, MRI Radiomics in Prostate Cancer: A Reliability Study, Front Oncol. 11 (2021) 805137. https://doi.org/10.3389/fonc.2021.805137.

[83] L. Wang, J. Tan, Y. Ge, X. Tao, Z. Cui, Z. Fei, J. Lu, H. Zhang, Z. Pan, Assessment of liver metastases radiomic feature reproducibility with deep-learning-based semi-automatic segmentation software, Acta Radiol. 62 (2021) 291–301. https://doi.org/10.1177/0284185120922822.

[84] O.L. Wong, Ji. Yuan, Y. Zhou, S.K. Yu, K.Y. Cheung, Longitudinal acquisition repeatability of MRI radiomics features: An ACR MRI phantom study on two MRI scanners using a 3D T1W TSE sequence, Medical Physics. 48 (2021) 1239–1249. https://doi.org/10.1002/mp.14686.

[85] I.S. Gruzdev, K.A. Zamyatina, V.S. Tikhonova, E.V. Kondratyev, A.V. Glotov, G.G. Karmazanovsky, A.Sh. Revishvili, Reproducibility of CT texture features of pancreatic neuroendocrine neoplasms, European Journal of Radiology. 133 (2020) 109371. https://doi.org/10.1016/j.ejrad.2020.109371.

[86] A. Könik, N. Miskin, Y. Guo, A.B. Shinagare, L. Qin, Robustness and performance of radiomic features in diagnosing cystic renal masses, Abdom Radiol. 46 (2021) 5260–5267. https://doi.org/10.1007/s00261-021-03241-2.

[87] I. Fornacon-Wood, H. Mistry, C.J. Ackermann, F. Blackhall, A. McPartlin, C. Faivre-Finn, G.J. Price, J.P.B. O'Connor, Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform, Eur Radiol. 30 (2020) 6241–6250. https://doi.org/10.1007/s00330-020-06957-9.

[88] A. Zwanenburg, M. Vallières, M.A. Abdalah, H.J.W.L. Aerts, V. Andrearczyk, A. Apte, S. Ashrafinia, S. Bakas, R.J. Beukinga, R. Boellaard, M. Bogowicz, L. Boldrini, I. Buvat, G.J.R. Cook, C. Davatzikos, A. Depeursinge, M.-C. Desseroit, N. Dinapoli, C.V. Dinh, S. Echegaray, I. El Naqa, A.Y. Fedorov, R. Gatta, R.J. Gillies, V. Goh, M. Götz, M. Guckenberger, S.M. Ha, M. Hatt, F. Isensee, P. Lambin, S. Leger, R.T.H. Leijenaar, J. Lenkowicz, F. Lippert, A. Losnegård, K.H. Maier-Hein, O. Morin, H. Müller, S. Napel, C. Nioche, F. Orlhac, S. Pati, E.A.G. Pfaehler, A. Rahmim, A.U.K. Rao, J. Scherer, M.M.

Siddique, N.M. Sijtsema, J. Socarras Fernandez, E. Spezi, R.J.H.M. Steenbakkers, S. Tanadini-Lang, D. Thorwarth, E.G.C. Troost, T. Upadhaya, V. Valentini, L.V. van Dijk, J. van Griethuysen, F.H.P. van Velden, P. Whybra, C. Richter, S. Löck, The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping, Radiology. 295 (2020) 328–338. https://doi.org/10.1148/radiol.2020191145.

[89] L. Duron, D. Balvay, S.V. Perre, A. Bouchouicha, J. Savatovsky, J.-C. Sadik, I. Thomassin-Naggara, L. Fournier, A. Lecler, Gray-level discretization impacts reproducible MRI radiomics texture features, PLOS ONE. 14 (2019) e0213459. https://doi.org/10.1371/journal.pone.0213459.

[90] A. Ibrahim, T. Refaee, S. Primakov, B. Barufaldi, R.J. Acciavatti, R.W.Y. Granzier, R. Hustinx, F.M. Mottaghy, H.C. Woodruff, J.E. Wildberger, P. Lambin, A.D.A. Maidment, The Effects of In-Plane Spatial Resolution on CT-Based Radiomic Features' Stability with and without ComBat Harmonization, Cancers. 13 (2021) 1848. https://doi.org/10.3390/cancers13081848.

[91] Y. Li, G. Tan, M. Vangel, J. Hall, W. Cai, Influence of feature calculating parameters on the reproducibility of CT radiomic features: a thoracic phantom study, Quant Imaging Med Surg. 10 (2020) 1775–1785. https://doi.org/10.21037/qims-19-921.

[92] H. Moradmand, S.M.R. Aghamiri, R. Ghaderi, Impact of image preprocessing methods on reproducibility of radiomic features in multimodal magnetic resonance imaging in glioblastoma, Journal of Applied Clinical Medical Physics.

21 (2020) 179–190. https://doi.org/10.1002/acm2.12795.

[93] E. Scalco, A. Belfatto, A. Mastropietro, T. Rancati, B. Avuzzi, A. Messina, R. Valdagni, G. Rizzo, T2w-MRI signal normalization affects radiomics features reproducibility, Medical Physics. 47 (2020) 1680–1691. https://doi.org/10.1002/mp.14038.

[94] M. Schwier, J. van Griethuysen, M.G. Vangel, S. Pieper, S. Peled, C. Tempany, H.J.W.L. Aerts, R. Kikinis, F.M. Fennessy, A. Fedorov, Repeatability of Multiparametric Prostate MRI Radiomics Features, Sci Rep. 9 (2019) 9441. https://doi.org/10.1038/s41598-019-45766-z.

[95] G. Simpson, J.C. Ford, R. Llorente, L. Portelance, F. Yang, E.A. Mellon, N. Dogan, Impact of quantization algorithm and number of gray level intensities on variability and repeatability of low field strength magnetic resonance image-based radiomics texture features, Physica Medica. 80 (2020) 209–220. https://doi.org/10.1016/j.ejmp.2020.10.029.

[96] C. Parmar, P. Grossmann, J. Bussink, P. Lambin, H.J.W.L. Aerts, Machine Learning methods for Quantitative Radiomic Biomarkers, Sci Rep. 5 (2015) 13087. https://doi.org/10.1038/srep13087.

[97] D.A.P. Delzell, S. Magnuson, T. Peter, M. Smith, B.J. Smith, Machine Learning and Feature Selection Methods for Disease Classification With Application to Lung Cancer Screening Image Data, Frontiers in Oncology. 9 (2019). https://www.frontiersin.org/articles/10.3389/fonc.2019.01393 (accessed January 4, 2023).

[98] J.E. van Timmeren, R.T.H. Leijenaar, W. van Elmpt, J. Wang, Z. Zhang, A.

Dekker, P. Lambin, Test–Retest Data for Radiomics Feature Stability Analysis: Generalizable or Study-Specific?, Tomography. 2 (2016) 361–365. https://doi.org/10.18383/j.tom.2016.00208.

[99] W. Rogers, S. Thulasi Seetha, T.A.G. Refaee, R.I.Y. Lieverse, R.W.Y. Granzier, A. Ibrahim, S.A. Keek, S. Sanduleanu, S.P. Primakov, M.P.L. Beuque, D. Marcus, A.M.A. van der Wiel, F. Zerka, C.J.G. Oberije, J.E. van Timmeren, H.C. Woodruff, P. Lambin, Radiomics: from qualitative to quantitative imaging, BJR. 93 (2020) 20190948. https://doi.org/10.1259/bjr.20190948.

[100] P. Lambin, R.T.H. Leijenaar, T.M. Deist, J. Peerlings, E.E.C. de Jong, J. van Timmeren, S. Sanduleanu, R.T.H.M. Larue, A.J.G. Even, A. Jochems, Y. van Wijk, H. Woodruff, J. van Soest, T. Lustberg, E. Roelofs, W. van Elmpt, A. Dekker, F.M. Mottaghy, J.E. Wildberger, S. Walsh, Radiomics: the bridge between medical imaging and personalized medicine, Nat Rev Clin Oncol. 14 (2017) 749–762. https://doi.org/10.1038/nrclinonc.2017.141.

[101] Y. Bian, H. Jiang, C. Ma, L. Wang, J. Zheng, G. Jin, J. Lu, CT-Based Radiomics Score for Distinguishing Between Grade 1 and Grade 2 Nonfunctioning Pancreatic Neuroendocrine Tumors, American Journal of Roentgenology. 215 (2020) 852–863. https://doi.org/10.2214/AJR.19.22123.

[102] T.P. Coroller, P. Grossmann, Y. Hou, E.R. Velazquez, R.T.H. Leijenaar, G. Hermann, P. Lambin, B. Haibe-Kains, R.H. Mak, H.J.W.L. Aerts, CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma, Radiotherapy and Oncology. 114 (2015) 345–350. https://doi.org/10.1016/j.radonc.2015.02.015.

[103] A. Guerrisi, E. Loi, S. Ungania, M. Russillo, V. Bruzzaniti, F. Elia, F. Desiderio, R. Marconi, F.M. Solivetti, L. Strigari, Novel cancer therapies for advanced cutaneous melanoma: The added value of radiomics in the decision making process–A systematic review, Cancer Medicine. 9 (2020) 1603–1612. https://doi.org/10.1002/cam4.2709.

[104] B. Zhao, Understanding Sources of Variation to Improve the Reproducibility of Radiomics, Frontiers in Oncology. 11 (2021). https://www.frontiersin.org/articles/10.3389/fonc.2021.633176 (accessed January 4, 2023).

[105] K. Lafata, J. Cai, C. Wang, J. Hong, C.R. Kelsey, F.-F. Yin, Spatial-temporal variability of radiomic features and its effect on the classification of lung cancer histology, Phys. Med. Biol. 63 (2018) 225003. https://doi.org/10.1088/1361-6560/aae56a.

[106] Y. Suter, U. Knecht, M. Alão, W. Valenzuela, E. Hewer, P. Schucht, R. Wiest, M. Reyes, Radiomics for glioblastoma survival analysis in pre-operative MRI: exploring feature robustness, class boundaries, and machine learning techniques, Cancer Imaging. 20 (2020) 55. https://doi.org/10.1186/s40644-020-00329-8.

[107] M. Vallières, E. Kay-Rivest, L. Perrin, X. Liem, C. Furstoss, N. Khaouam, P. Nguyen-Tan, C.-S. Wang, K. Sultanem, Data from Head-Neck-PET-CT, (2017). https://doi.org/10.7937/K9/TCIA.2017.8OJE5Q00.

[108] T. Fh, C. Cyw, C. Eyw, Radiomics AI prediction for head and neck squamous cell carcinoma (HNSCC) prognosis and recurrence with target volume approach, BJR|Open. 3 (2021) 20200073. https://doi.org/10.1259/bjro.20200073.

[109] M. Bogowicz, S. Tanadini-Lang, P. Veit-Haibach, M. Pruschy, S. Bender, A. Sharma, M. Hüllner, G. Studer, S. Stieb, H. Hemmatazad, S. Glatz, M. Guckenberger, O. Riesterer, Perfusion CT radiomics as potential prognostic biomarker in head and neck squamous cell carcinoma, Acta Oncologica. 58 (2019) 1514–1518. https://doi.org/10.1080/0284186X.2019.1629013.

[110] M. Vallières, E. Kay-Rivest, L.J. Perrin, X. Liem, C. Furstoss, H.J.W.L. Aerts, N. Khaouam, P.F. Nguyen-Tan, C.-S. Wang, K. Sultanem, J. Seuntjens, I. El Naqa, Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer, Sci Rep. 7 (2017) 10117. https://doi.org/10.1038/s41598-017-10371-5.

[111] E. Lombardo, C. Kurz, S. Marschner, M. Avanzo, V. Gagliardi, G. Fanetti, G. Franchin, J. Stancanello, S. Corradini, M. Niyazi, C. Belka, K. Parodi, M. Riboldi, G. Landry, Distant metastasis time to event analysis with CNNs in independent head and neck cancer cohorts, Sci Rep. 11 (2021) 6418. https://doi.org/10.1038/s41598-021-85671-y.

[112] A. Diamant, A. Chatterjee, M. Vallières, G. Shenouda, J. Seuntjens, Deep learning in head & neck cancer outcome prediction, Sci Rep. 9 (2019) 2764. https://doi.org/10.1038/s41598-019-39206-1.

[113] X. Fave, L. Zhang, J. Yang, D. Mackin, P. Balter, D. Gomez, D. Followill, A.K. Jones, F. Stingo, L.E. Court, Impact of image preprocessing on the volume dependence and prognostic potential of radiomics features in non-small cell lung cancer, Transl. Cancer Res. 5 (2016) 349–363. https://doi.org/10.21037/tcr.2016.07.11.

[114] M. Shafiq-ul-Hassan, G.G. Zhang, K. Latifi, G. Ullah, D.C. Hunt, Y. Balagurunathan, M.A. Abdalah, M.B. Schabath, D.G. Goldgof, D. Mackin, L.E. Court, R.J. Gillies, E.G. Moros, Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels, Medical Physics. 44 (2017) 1050–1062. https://doi.org/10.1002/mp.12123.

[115] Z. Yaniv, B.C. Lowekamp, H.J. Johnson, R. Beare, SimpleITK Image-Analysis Notebooks: a Collaborative Environment for Education and Reproducible Research, J Digit Imaging. 31 (2018) 290–303. https://doi.org/10.1007/s10278-017-0037-8.

[116] G. Bradski, The OpenCV Library, Dr. Dobb's. (n.d.). http://www.drdobbs.com/open-source/the-opencv-library/184404319 (accessed January 4, 2023).

[117] J. Cai, J. Zheng, J. Shen, Z. Yuan, M. Xie, M. Gao, H. Tan, Z. Liang, X. Rong, Y. Li, H. Li, J. Jiang, H. Zhao, A.A. Argyriou, M.L.K. Chua, Y. Tang, A Radiomics Model for Predicting the Response to Bevacizumab in Brain Necrosis after Radiotherapy, Clinical Cancer Research. 26 (2020) 5438–5447. https://doi.org/10.1158/1078-0432.CCR-20-1264.

[118] L. Yu, H. Liu, Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution, (n.d.).

[119] G. Lemaıtre, F. Nogueira, Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning, (n.d.).

[120] A.-S. Dirand, F. Frouin, I. Buvat, A downsampling strategy to assess the predictive value of radiomic features, Sci Rep. 9 (2019) 17869.

https://doi.org/10.1038/s41598-019-54190-2.

[121] Q. Qiu, J. Duan, Z. Duan, X. Meng, C. Ma, J. Zhu, J. Lu, T. Liu, Y. Yin, Reproducibility and non-redundancy of radiomic features extracted from arterial phase CT scans in hepatocellular carcinoma patients: impact of tumor segmentation variability, Quantitative Imaging in Medicine and Surgery. 9 (2019) 45364–45464. https://doi.org/10.21037/qims.2019.03.02.

[122] A. Appice, M. Ceci, S. Rawles, P. Flach, Redundant feature elimination for multi-class problems, in: Proceedings of the Twenty-First International Conference on Machine Learning, Association for Computing Machinery, New York, NY, USA, 2004: p. 5. https://doi.org/10.1145/1015330.1015397.

[123] X. Zhang, X. Xu, Q. Tian, B. Li, Y. Wu, Z. Yang, Z. Liang, Y. Liu, G. Cui, H. Lu, Radiomics assessment of bladder cancer grade using texture features from diffusion-weighted imaging, Journal of Magnetic Resonance Imaging. 46 (2017) 1281–1288. https://doi.org/10.1002/jmri.25669.

[124] M. Mottola, S. Ursprung, L. Rundo, L.E. Sanchez, T. Klatte, I. Mendichovszky, G.D. Stewart, E. Sala, A. Bevilacqua, Reproducibility of CT-based radiomic features against image resampling and perturbations for tumour and healthy kidney in renal cancer patients, Sci Rep. 11 (2021) 11542. https://doi.org/10.1038/s41598-021-90985-y.

[125] S. Rizzo, F. Botta, S. Raimondi, D. Origgi, C. Fanciullo, A.G. Morganti, M. Bellomi, Radiomics: the facts and the challenges of image analysis, European Radiology Experimental. 2 (2018) 36. https://doi.org/10.1186/s41747-018-0068-z.

[126] M. Pavic, M. Bogowicz, X. Würms, S. Glatz, T. Finazzi, O. Riesterer, J. Roesch, L. Rudofsky, M. Friess, P. Veit-Haibach, M. Huellner, I. Opitz, W. Weder, T. Frauenfelder, M. Guckenberger, S. Tanadini-Lang, Influence of inter-observer delineation variability on radiomics stability in different tumor sites, Acta Oncologica. 57 (2018) 1070–1074. https://doi.org/10.1080/0284186X.2018.1445283.

[127] L. Daly, G.J. Bourke, Interpretation and Uses of Medical Statistics, John Wiley & Sons, 2008.

[128] S.-H. Park, H. Lim, B.K. Bae, M.H. Hahm, G.O. Chong, S.Y. Jeong, J.-C. Kim, Robustness of magnetic resonance radiomic features to pixel size resampling and interpolation in patients with cervical cancer, Cancer Imaging. 21 (2021) 19. https://doi.org/10.1186/s40644-021-00388-5.

[129] T.K. Koo, M.Y. Li, A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research, Journal of Chiropractic Medicine. 15 (2016) 155–163. https://doi.org/10.1016/j.jcm.2016.02.012.

[130] Y. Suter, U. Knecht, M. Alão, W. Valenzuela, E. Hewer, P. Schucht, R. Wiest, M. Reyes, Radiomics for glioblastoma survival analysis in pre-operative MRI: exploring feature robustness, class boundaries, and machine learning techniques, Cancer Imaging. 20 (2020) 55. https://doi.org/10.1186/s40644-020-00329-8.

[131] A.H. Ree, K.R. Redalen, Personalized radiotherapy: concepts, biomarkers and trial design, BJR. 88 (2015) 20150009. https://doi.org/10.1259/bjr.20150009.

[132] R. Reiazi, E. Abbas, P. Famiyeh, A. Rezaie, J.Y.Y. Kwan, T. Patel, S.V. Bratman, T. Tadic, F.-F. Liu, B. Haibe-Kains, The impact of the variation of imaging

parameters on the robustness of Computed Tomography radiomic features: A review, Computers in Biology and Medicine. 133 (2021) 104400. https://doi.org/10.1016/j.compbiomed.2021.104400.

[133] F. Orlhac, A. Lecler, J. Savatovski, J. Goya-Outi, C. Nioche, F. Charbonneau, N. Ayache, F. Frouin, L. Duron, I. Buvat, How can we combat multicenter variability in MR radiomics? Validation of a correction procedure, Eur Radiol. 31 (2021) 2272–2280. https://doi.org/10.1007/s00330-020-07284-9.

[134] J.J. Foy, H.A. Al-Hallaq, V. Grekoski, T. Tran, K. Guruvadoo, S.G.A. III, W.F. Sensakovic, Harmonization of radiomic feature variability resulting from differences in CT image acquisition and reconstruction: assessment in a cadaveric liver, Phys. Med. Biol. 65 (2020) 205008. https://doi.org/10.1088/1361-6560/abb172.

[135] M. Ligero, O. Jordi-Ollero, K. Bernatowicz, A. Garcia-Ruiz, E. Delgado-Muñoz, D. Leiva, R. Mast, C. Suarez, R. Sala-Llonch, N. Calvo, M. Escobar, A. Navarro-Martin, G. Villacampa, R. Dienstmann, R. Perez-Lopez, Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis, Eur Radiol. 31 (2021) 1460–1470. https://doi.org/10.1007/s00330-020-07174-0.

[136] Y. Li, S. Ammari, C. Balleyguier, N. Lassau, E. Chouzenoux, Impact of Preprocessing and Harmonization Methods on the Removal of Scanner Effects in Brain MRI Radiomic Features, Cancers. 13 (2021) 3000. https://doi.org/10.3390/cancers13123000.

[137] B. Zhao, L.P. James, C.S. Moskowitz, P. Guo, M.S. Ginsberg, R.A. Lefkowitz,

Y. Qin, G.J. Riely, M.G. Kris, L.H. Schwartz, Evaluating Variability in Tumor Measurements from Same-day Repeat CT Scans of Patients with Non–Small Cell Lung Cancer, Radiology. 252 (2009) 263–272. https://doi.org/10.1148/radiol.2522081593.

[138] S. Li, J. Liu, Y. Xiong, P. Pang, P. Lei, H. Zou, M. Zhang, B. Fan, P. Luo, A radiomics approach for automated diagnosis of ovarian neoplasm malignancy in computed tomography, Sci Rep. 11 (2021) 8730. https://doi.org/10.1038/s41598-021-87775-x.

[139] M. Bang, J. Eom, C. An, S. Kim, Y.W. Park, S.S. Ahn, J. Kim, S.-K. Lee, S.-H. Lee, An interpretable multiparametric radiomics model for the diagnosis of schizophrenia using magnetic resonance imaging of the corpus callosum, Transl Psychiatry. 11 (2021) 462. https://doi.org/10.1038/s41398-021-01586-2.

[140] H. Liu, H. Ren, Z. Wu, H. Xu, S. Zhang, J. Li, L. Hou, R. Chi, H. Zheng, Y. Chen, S. Duan, H. Li, Z. Xie, D. Wang, CT radiomics facilitates more accurate diagnosis of COVID-19 pneumonia: compared with CO-RADS, J Transl Med. 19 (2021) 29. https://doi.org/10.1186/s12967-020-02692-3.

[141] J. Shin, J.S. Lim, Y.-M. Huh, J.-H. Kim, W.J. Hyung, J.-J. Chung, K. Han, S. Kim, A radiomics-based model for predicting prognosis of locally advanced gastric cancer in the preoperative setting, Sci Rep. 11 (2021) 1879. https://doi.org/10.1038/s41598-021-81408-z.

[142] E. Pak, K.S. Choi, S.H. Choi, C.-K. Park, T.M. Kim, S.-H. Park, J.H. Lee, S.-T. Lee, I. Hwang, R.-E. Yoo, K.M. Kang, T.J. Yun, J.-H. Kim, C.-H. Sohn, Prediction of Prognosis in Glioblastoma Using Radiomics Features of Dynamic

Contrast-Enhanced MRI, Korean J Radiol. 22 (2021) 1514. https://doi.org/10.3348/kjr.2020.1433.

[143] X. Xu, J. Zhang, K. Yang, Q. Wang, X. Chen, B. Xu, Prognostic prediction of hypertensive intracerebral hemorrhage using CT radiomics and machine learning, Brain Behav. 11 (2021) e02085. https://doi.org/10.1002/brb3.2085.

[144] A. Delli Pizzi, A.M. Chiarelli, P. Chiacchiaretta, M. d'Annibale, P. Croce, C. Rosa, D. Mastrodicasa, S. Trebeschi, D.M.J. Lambregts, D. Caposiena, F.L. Serafini, R. Basilico, G. Cocco, P. Di Sebastiano, S. Cinalli, A. Ferretti, R.G. Wise, D. Genovesi, R.G.H. Beets-Tan, M. Caulo, MRI-based clinical-radiomics model predicts tumor response before treatment in locally advanced rectal cancer, Sci Rep. 11 (2021) 5379. https://doi.org/10.1038/s41598-021-84816-3.

[145] J. Gu, T. Tong, C. He, M. Xu, X. Yang, J. Tian, T. Jiang, K. Wang, Deep learning radiomics of ultrasonography can predict response to neoadjuvant chemotherapy in breast cancer at an early stage of treatment: a prospective study, Eur Radiol. 32 (2022) 2099–2109. https://doi.org/10.1007/s00330-021-08293-y.

[146] R.R. Colen, C. Rolfo, M. Ak, M. Ayoub, S. Ahmed, N. Elshafeey, P. Mamindla, P.O. Zinn, C. Ng, R. Vikram, S. Bakas, C.B. Peterson, J. Rodon Ahnert, V. Subbiah, D.D. Karp, B. Stephen, J. Hajjar, A. Naing, Radiomics analysis for predicting pembrolizumab response in patients with advanced rare cancers, J Immunother Cancer. 9 (2021) e001752. https://doi.org/10.1136/jitc-2020-001752.

[147] R.W.Y. Granzier, A. Ibrahim, S. Primakov, S.A. Keek, I. Halilaj, A. Zwanenburg, S.M.E. Engelen, M.B.I. Lobbes, P. Lambin, H.C. Woodruff, M.L. Smidt, Test–

Retest Data for the Assessment of Breast MRI Radiomic Feature Repeatability, Magnetic Resonance Imaging. 56 (2022) 592–604. https://doi.org/10.1002/jmri.28027.

[148] J.E. van Timmeren, R.T.H. Leijenaar, W. van Elmpt, J. Wang, Z. Zhang, A. Dekker, P. Lambin, Test–Retest Data for Radiomics Feature Stability Analysis: Generalizable or Study-Specific?, Tomography. 2 (2016) 361–365. https://doi.org/10.18383/j.tom.2016.00208.

[149] M. Bologna, V.D.A. Corino, E. Montin, A. Messina, G. Calareso, F.G. Greco, S. Sdao, L.T. Mainardi, Assessment of Stability and Discrimination Capacity of Radiomic Features on Apparent Diffusion Coefficient Images, J Digit Imaging. 31 (2018) 879–894. https://doi.org/10.1007/s10278-018-0092-9.

[150] A. Zwanenburg, S. Leger, L. Agolli, K. Pilz, E.G.C. Troost, C. Richter, S. Löck, Assessing robustness of radiomic features by image perturbation, Scientific Reports. 9 (2019) 1–10. https://doi.org/10.1038/s41598-018-36938-4.

[151] X. Teng, J. Zhang, A. Zwanenburg, J. Sun, Y. Huang, S. Lam, Y. Zhang, B. Li, T. Zhou, H. Xiao, C. Liu, W. Li, X. Han, Z. Ma, T. Li, J. Cai, Building reliable radiomic models using image perturbation, Sci Rep. 12 (2022) 10035. https://doi.org/10.1038/s41598-022-14178-x.

[152] D.C. Newitt, S.C. Partridge, Z. Zhang, J. Gibbs, T. Chenevert, M. Rosen, P. Bolan, H. Marques, J. Romanoff, L. Cimino, B.N. Joe, H. Umphrey, H. Ojeda-Fournier, B. Dogan, K.Y. Oh, H. Abe, J. Drukteinis, L.J. Esserman, N.M. Hylton, ACRIN 6698/I-SPY2 Breast DWI, (2021). https://doi.org/10.7937/TCIA.KK02-6D95.

[153] S.C. Partridge, Z. Zhang, D.C. Newitt, J.E. Gibbs, T.L. Chenevert, M.A. Rosen, P.J. Bolan, H.S. Marques, J. Romanoff, L. Cimino, B.N. Joe, H.R. Umphrey, H. Ojeda-Fournier, B. Dogan, K. Oh, H. Abe, J.S. Drukteinis, L.J. Esserman, N.M. Hylton, Diffusion-weighted MRI Findings Predict Pathologic Response in Neoadjuvant Treatment of Breast Cancer: The ACRIN 6698 Multicenter Trial, Radiology. 289 (2018) 618–627. https://doi.org/10.1148/radiol.2018180273.

[154] D.M. Wolf, C. Yau, J. Wulfkuhle, L. Brown-Swigart, R.I. Gallagher, P.R.E. Lee, Z. Zhu, M.J. Magbanua, R. Sayaman, N. O'Grady, A. Basu, A. Delson, J.P. Coppé, R. Lu, J. Braun, S.M. Asare, L. Sit, J.B. Matthews, J. Perlmutter, N. Hylton, M.C. Liu, P. Pohlmann, W.F. Symmans, H.S. Rugo, C. Isaacs, A.M. DeMichele, D. Yee, D.A. Berry, L. Pusztai, E.F. Petricoin, G.L. Hirst, L.J. Esserman, L.J. van 't Veer, Redefining breast cancer subtypes to guide treatment prioritization and maximize response: Predictive biomarkers across 10 cancer therapies, Cancer Cell. 40 (2022) 609-623.e6. https://doi.org/10.1016/j.ccell.2022.05.005.

[155] A. Zwanenburg, M. Vallières, M.A. Abdalah, H.J.W.L. Aerts, V. Andrearczyk, A. Apte, S. Ashrafinia, S. Bakas, R.J. Beukinga, R. Boellaard, M. Bogowicz, L. Boldrini, I. Buvat, G.J.R. Cook, C. Davatzikos, A. Depeursinge, M.-C. Desseroit, N. Dinapoli, C.V. Dinh, S. Echegaray, I. El Naqa, A.Y. Fedorov, R. Gatta, R.J. Gillies, V. Goh, M. Götz, M. Guckenberger, S.M. Ha, M. Hatt, F. Isensee, P. Lambin, S. Leger, R.T.H. Leijenaar, J. Lenkowicz, F. Lippert, A. Losnegård, K.H. Maier-Hein, O. Morin, H. Müller, S. Napel, C. Nioche, F.

Orlhac, S. Pati, E.A.G. Pfaehler, A. Rahmim, A.U.K. Rao, J. Scherer, M.M. Siddique, N.M. Sijtsema, J. Socarras Fernandez, E. Spezi, R.J.H.M. Steenbakkers, S. Tanadini-Lang, D. Thorwarth, E.G.C. Troost, T. Upadhaya, V. Valentini, L.V. van Dijk, J. van Griethuysen, F.H.P. van Velden, P. Whybra, C. Richter, S. Löck, The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping, Radiology. 295 (2020) 328–338. https://doi.org/10.1148/radiol.2020191145.

[156] K.O. McGraw, S.P. Wong, Forming inferences about some intraclass correlation coefficients, Psychological Methods. 1 (1996) 30–46. https://doi.org/10.1037/1082-989X.1.1.30.

[157] R. Vallat, Pingouin: statistics in Python, Journal of Open Source Software. 3 (2018) 1026. https://doi.org/10.21105/joss.01026.

[158] C. Ding, H. Peng, Minimum Redundancy Feature Selection From Microarray Gene Expression Data, 2003. https://doi.org/10.1109/CSB.2003.1227396.

[159] X.-Y. Liu, J. Wu, Z.-H. Zhou, Exploratory Undersampling for Class-Imbalance Learning, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 39 (2009) 539–550. https://doi.org/10.1109/TSMCB.2008.2007853.

[160] G. Lemaître, F. Nogueira, C.K. Aridas, Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning, J. Mach. Learn. Res. 18 (2017) 559–563.

[161] P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D.

Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S.J. van der Walt, M. Brett, J. Wilson, K.J. Millman, N. Mayorov, A.R.J. Nelson, E. Jones, R. Kern, E. Larson, C.J. Carey, İ. Polat, Y. Feng, E.W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E.A. Quintero, C.R. Harris, A.M. Archibald, A.H. Ribeiro, F. Pedregosa, P. van Mulbregt, SciPy 1.0: fundamental algorithms for scientific computing in Python, Nat Methods. 17 (2020) 261–272. https://doi.org/10.1038/s41592-019-0686-2.

[162] A. Saha, M.R. Harowicz, M.A. Mazurowski, Breast cancer MRI radiomics: An overview of algorithmic features and impact of inter-reader variability in annotating tumors, Med Phys. 45 (2018) 3076–3085. https://doi.org/10.1002/mp.12925.

[163] R.T.H. Leijenaar, G. Nalbantov, S. Carvalho, W.J.C. van Elmpt, E.G.C. Troost, R. Boellaard, H.J.W.L. Aerts, R.J. Gillies, P. Lambin, The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis, Sci Rep. 5 (2015) 11075. https://doi.org/10.1038/srep11075.

[164] R. Thawani, M. McLane, N. Beig, S. Ghose, P. Prasanna, V. Velcheti, A. Madabhushi, Radiomics and radiogenomics in lung cancer: A review for the clinician, Lung Cancer. 115 (2018) 34–41. https://doi.org/10.1016/j.lungcan.2017.10.015.

[165] J.R. Ferreira Junior, M. Koenigkam-Santos, F.E.G. Cipriano, A.T. Fabro, P.M. de Azevedo-Marques, Radiomics-based features for pattern recognition of lung cancer histopathology and metastases, Computer Methods and Programs in

Biomedicine. 159 (2018) 23–30. https://doi.org/10.1016/j.cmpb.2018.02.015.

[166] A. Mouraviev, J. Detsky, A. Sahgal, M. Ruschin, Y.K. Lee, I. Karam, C. Heyn, G.J. Stanisz, A.L. Martel, Use of radiomics for the prediction of local control of brain metastases after stereotactic radiosurgery, Neuro-Oncology. 22 (2020) 797–805. https://doi.org/10.1093/neuonc/noaa007.

[167] L. Shi, Y. He, Z. Yuan, S. Benedict, R. Valicenti, J. Qiu, Y. Rong, Radiomics for Response and Outcome Assessment for Non-Small Cell Lung Cancer, Technol Cancer Res Treat. 17 (2018) 1533033818782788. https://doi.org/10.1177/1533033818782788.

[168] I. Desideri, M. Loi, G. Francolini, C. Becherini, L. Livi, P. Bonomo, Application of Radiomics for the Prediction of Radiation-Induced Toxicity in the IMRT Era: Current State-of-the-Art, Frontiers in Oncology. 10 (2020). https://www.frontiersin.org/articles/10.3389/fonc.2020.01708 (accessed January 4, 2023).

[169] Z. Liu, Z. Li, J. Qu, R. Zhang, X. Zhou, L. Li, K. Sun, Z. Tang, H. Jiang, H. Li, Q. Xiong, Y. Ding, X. Zhao, K. Wang, Z. Liu, J. Tian, Radiomics of Multiparametric MRI for Pretreatment Prediction of Pathologic Complete Response to Neoadjuvant Chemotherapy in Breast Cancer: A Multicenter Study, Clinical Cancer Research. 25 (2019) 3538–3547. https://doi.org/10.1158/1078-0432.CCR-18-3190.

[170] B. Baeßler, K. Weiss, D. Pinto dos Santos, Robustness and Reproducibility of Radiomics in Magnetic Resonance Imaging: A Phantom Study, Investigative Radiology. 54 (2019) 221. https://doi.org/10.1097/RLI.0000000000000530.

[171] M. Vallières, A. Zwanenburg, B. Badic, C.C.L. Rest, D. Visvikis, M. Hatt, Responsible Radiomics Research for Faster Clinical Translation, Journal of Nuclear Medicine. 59 (2018) 189–193. https://doi.org/10.2967/jnumed.117.200501.

[172] M. Avanzo, L. Wei, J. Stancanello, M. Vallières, A. Rao, O. Morin, S.A. Mattonen, I. El Naqa, Machine and deep learning methods for radiomics, Medical Physics. 47 (2020) e185–e202. https://doi.org/10.1002/mp.13678.

[173] H. Jin, J.H. Kim, Evaluation of Feature Robustness Against Technical Parameters in CT Radiomics: Verification of Phantom Study with Patient Dataset, J Sign Process Syst. 92 (2020) 277–287. https://doi.org/10.1007/s11265-019-01496-z.

[174] S. Gourtsoyianni, G. Doumou, D. Prezzi, B. Taylor, J.J. Stirling, N.J. Taylor, M. Siddique, G.J.R. Cook, R. Glynne-Jones, V. Goh, Primary Rectal Cancer: Repeatability of Global and Local-Regional MR Imaging Texture Features, Radiology. 284 (2017) 552–561. https://doi.org/10.1148/radiol.2017161375.

[175] H. Lu, N.A. Parra, J. Qi, K. Gage, Q. Li, S. Fan, S. Feuerlein, J. Pow-Sang, R. Gillies, J.W. Choi, Y. Balagurunathan, Repeatability of Quantitative Imaging Features in Prostate Magnetic Resonance Imaging, Frontiers in Oncology. 10 (2020). https://www.frontiersin.org/articles/10.3389/fonc.2020.00551 (accessed January 4, 2023).

[176] J.Y.Y. Kwan, J. Su, S.H. Huang, L.S. Ghoraie, W. Xu, B. Chan, K.W. Yip, M. Giuliani, A. Bayley, J. Kim, A.J. Hope, J. Ringash, J. Cho, A. McNiven, A.

Hansen, D. Goldstein, J.R. de Almeida, H.J. Aerts, J.N. Waldron, B. Haibe-Kains, B. O'Sullivan, S.V. Bratman, F.-F. Liu, Radiomic Biomarkers to Refine Risk Models for Distant Metastasis in HPV-related Oropharyngeal Carcinoma, International Journal of Radiation Oncology, Biology, Physics. 102 (2018) 1107–1116. https://doi.org/10.1016/j.ijrobp.2018.01.057.

[177] J.Y.Y. Kwan, J. Su, S.H. Huang, L.S. Ghoraie, W. Xu, B. Chan, K.W. Yip, M. Giuliani, A. Bayley, J. Kim, A.J. Hope, J. Ringash, J. Cho, A. McNiven, A. Hansen, D. Goldstein, J.R. De Almeida, H.J. Aerts, J.N. Waldron, B. Haibe-Kains, B. O'Sullivan, S.V. Bratman, F.-F. Liu, Data from Radiomic Biomarkers to Refine Risk Models for Distant Metastasis in Oropharyngeal Carcinoma, (2019). https://doi.org/10.7937/TCIA.2019.8DHO2GLS.

[178] L. Fournier, L. Costaridou, L. Bidaut, N. Michoux, F.E. Lecouvet, L.-F. de Geus-Oei, R. Boellaard, D.E. Oprea-Lager, N.A. Obuchowski, A. Caroli, W.G. Kunz, E.H. Oei, J.P.B. O'Connor, M.E. Mayerhoefer, M. Franca, A. Alberich-Bayarri, C.M. Deroose, C. Loewe, R. Manniesing, C. Caramella, E. Lopci, N. Lassau, A. Persson, R. Achten, K. Rosendahl, O. Clement, E. Kotter, X. Golay, M. Smits, M. Dewey, D.C. Sullivan, A. van der Lugt, N.M. deSouza, European Society of Radiology, Incorporating radiomics into clinical trials: expert consensus endorsed by the European Society of Radiology on considerations for data-driven compared to biologically driven quantitative biomarkers, Eur Radiol. 31 (2021) 6001–6012. https://doi.org/10.1007/s00330-020-07598-8.

[179] Y. Zhang, A. Oikonomou, A. Wong, M.A. Haider, F. Khalvati, Radiomics-based Prognosis Analysis for Non-Small Cell Lung Cancer, Sci Rep. 7 (2017) 46349.

https://doi.org/10.1038/srep46349.

[180] R. Forghani, P. Savadjiev, A. Chatterjee, N. Muthukrishnan, C. Reinhold, B. Forghani, Radiomics and Artificial Intelligence for Biomarker and Prediction Model Development in Oncology, Computational and Structural Biotechnology Journal. 17 (2019) 995–1008. https://doi.org/10.1016/j.csbj.2019.07.001.

[181] J. Yun, J.E. Park, H. Lee, S. Ham, N. Kim, H.S. Kim, Radiomic features and multilayer perceptron network classifier: a robust MRI classification strategy for distinguishing glioblastoma from primary central nervous system lymphoma, Sci Rep. 9 (2019) 5746. https://doi.org/10.1038/s41598-019-42276-w.

[182] S. Irani, Distant metastasis from oral cancer: A review and molecular biologic aspects, Journal of International Society of Preventive and Community Dentistry. 6 (2016) 265. https://doi.org/10.4103/2231-0762.186805.

[183] E.A. Mittendorf, T.A. Buchholz, S.L. Tucker, F. Meric-Bernstam, H.M. Kuerer, A.M. Gonzalez-Angulo, I. Bedrosian, G.V. Babiera, K. Hoffman, M. Yi, M.I. Ross, G.N. Hortobagyi, K.K. Hunt, Impact of Chemotherapy Sequencing on Local-Regional Failure Risk in Breast Cancer Patients Undergoing Breast-Conserving Therapy, Annals of Surgery. 257 (2013) 173. https://doi.org/10.1097/SLA.0b013e3182805c4a.

[184] L.-L. Zhang, M.-Y. Huang, Y. Li, J.-H. Liang, T.-S. Gao, B. Deng, J.-J. Yao, L. Lin, F.-P. Chen, X.-D. Huang, J. Kou, C.-F. Li, C.-M. Xie, Y. Lu, Y. Sun, Pretreatment MRI radiomics analysis allows for reliable prediction of local recurrence in non-metastatic T4 nasopharyngeal carcinoma, EBioMedicine. 42 (2019) 270–280. https://doi.org/10.1016/j.ebiom.2019.03.050.

[185] Z. Zhou, K. Wang, M. Folkert, H. Liu, S. Jiang, D. Sher, J. Wang, Multifaceted radiomics for distant metastasis prediction in head & neck cancer, Phys. Med. Biol. 65 (2020) 155009. https://doi.org/10.1088/1361-6560/ab8956.

[186] K.O. McGraw, S.P. Wong, Forming Inferences About Some Intraclass Correlation Coefficients, (n.d.) 18.

[187] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. Vanderplas, A. Joly, B. Holt, G. Varoquaux, API design for machine learning software: experiences from the scikit-learn project, (2013). https://doi.org/10.48550/arXiv.1309.0238.

[188] D. Vuong, M. Bogowicz, S. Denzler, C. Oliveira, R. Foerster, F. Amstutz, H.S. Gabryś, J. Unkelbach, S. Hillinger, S. Thierstein, A. Xyrafas, S. Peters, M. Pless, M. Guckenberger, S. Tanadini-Lang, Comparison of robust to standardized CT radiomics models to predict overall survival for non-small cell lung cancer patients, Medical Physics. 47 (2020) 4045–4053. https://doi.org/10.1002/mp.14224.

[189] B. Efron, R. Tibshirani, Improvements on Cross-Validation: The .632+ Bootstrap Method, Journal of the American Statistical Association. 92 (1997) 548–560. https://doi.org/10.2307/2965703.

[190] C. Shorten, T.M. Khoshgoftaar, A survey on Image Data Augmentation for Deep Learning, Journal of Big Data. 6 (2019) 60. https://doi.org/10.1186/s40537-019-0197-0.