Quantitative imaging analysis

QUANTITATIVE IMAGING ANALYSIS:
CHALLENGES AND POTENTIALS

Turkey Refaee

QUANTITATIVE IMAGING ANALYSIS:
CHALLENGES AND POTENTIALS

Turkey Abdullah Refaee

Cover: Stefanie van den Herik
Layout: Dennis Hendriks || ProefschriftMaken.nl
Printed by: ProefschriftMaken.nl

ISBN: 978-94-6423-883-9

QUANTITATIVE IMAGING ANALYSIS:
CHALLENGES AND POTENTIALS

Dissertation

to obtain the degree of Doctor
at Maastricht University
by the authority of the Rector Magnificus Prof.dr. Pamela Habibović
in accordance with the decision of the Board of Deans,
to be defended in public on
Tuesday 12 July 2022 at 16:00 hours

by
Turkey Abdullah Refaee

Promotor:
Prof.dr. Philippe Lambin

Co-promoters:
Dr. Julien Guiot (University of Liège)
Dr. Henry C. Woodruff

Thesis Assessment Committee:
Prof.dr. Manon van Engeland (Chair)
Prof.dr. Wesseling Geertjan (Respiratory Medicine, MUMC)
Prof.dr. Henning Muller (University of Applied Sciences Western Switzerland)
Dr. Karen Zegers (Maastro Clinic)

Contents

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$minimum = \min(\mathbf{X})$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2$$

$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_s} \mathbf{P}(i,j)\right)^2}{N_p}$$

$$correlation = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) ij - \mu_x \mu_y}{\sigma_x(i) \sigma_y(j)}$$

$$MAD = \frac{1}{N_p} \sum_{i=1}^{N_p} |\mathbf{X}(i) - \bar{X}|$$

$$minimum = \min(\mathbf{X})$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_s} \mathbf{P}\right.}{N_z}$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2$$

Part I

$$variance = \frac{1}{N_p}\sum_{i=1}^{N_p}(\mathbf{X}(i)-\bar{X})^2$$

$$total\ energy = V_{voxel}\sum_{i=1}^{N_p}(\mathbf{X}(i)+c)^2$$

$$SDE = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}\frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$minimum = \min(\mathbf{X})$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2$$

$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_s} \mathbf{P}(i,j)\right)^2}{N_p}$$

$$correlation = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) ij - \mu_x \mu_y}{\sigma_x(i) \sigma_y(j)}$$

$$MAD = \frac{1}{N_p} \sum_{i=1}^{N_p} |\mathbf{X}(i) - \bar{X}|$$

$$minimum = \min(\mathbf{X})$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

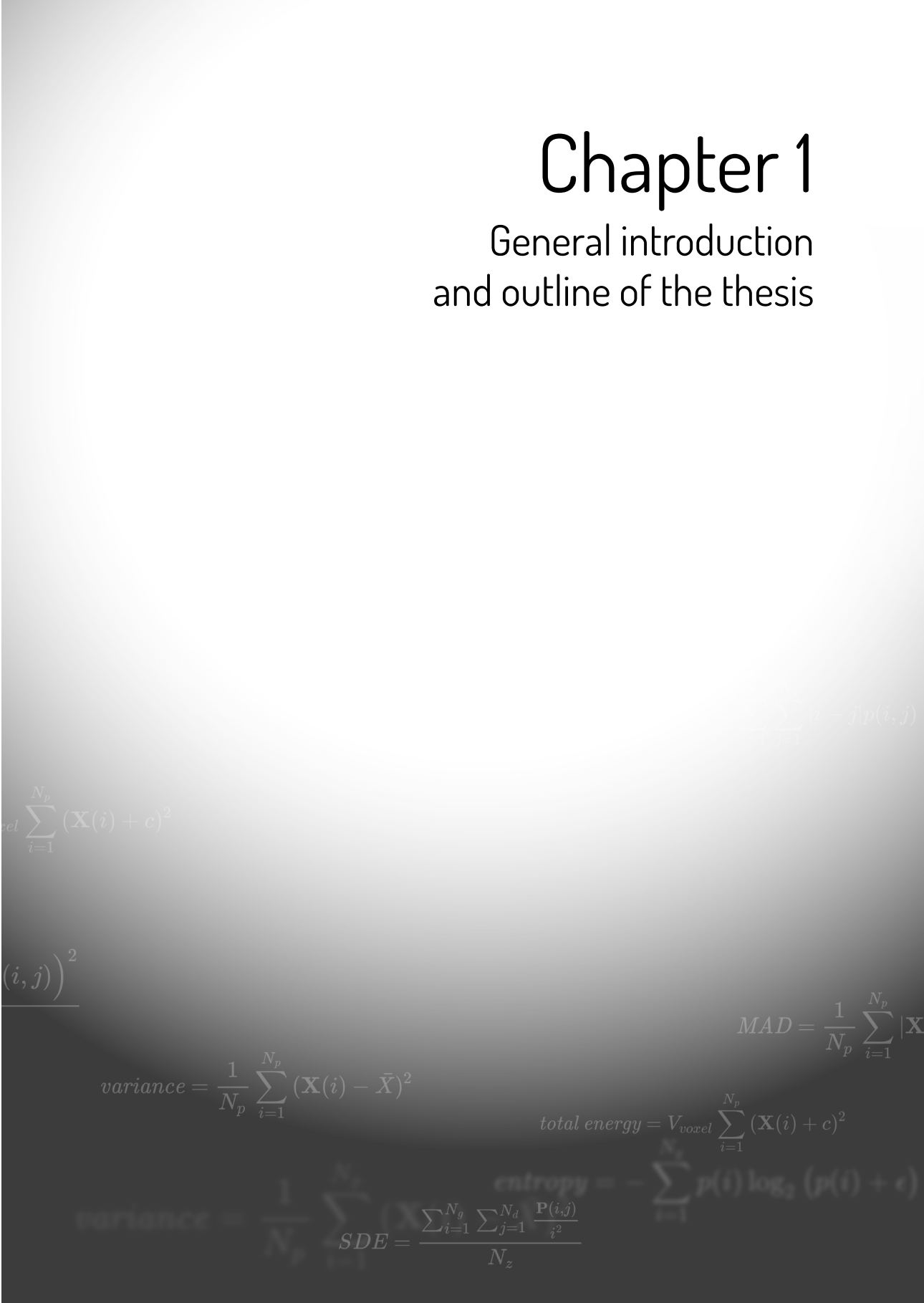
$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_s} \mathbf{P}\right.}{N_z}$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2$$

Chapter 1

General introduction and outline of the thesis



Technological advancements, especially in recent times, have resulted in a plethora of innovations in many different scientific fields. This is especially the case with regard to the new diagnostic procedures and imaging modalities being made available to the field of oncology [1]. However, the genetic and micro-environmental heterogeneity found in tumors and between patients adds a stark layer of complexity [2,3]. Currently, due to the sheer abundance and complexity of oncology-related datasets, new strategies for facilitating clinical decision-making are becoming increasingly necessary [4].

Precision Medicine

Precision medicine refers to preventative and therapeutic approaches that focus on accounting for an individual patient's characteristics, as well as their specific ailments [5]. Data mining is a typical method of precision medicine. It involves detecting patterns in and across big datasets of diverse populations, using powerful computational techniques such as machine learning. Across the variety of patient populations, patterns may be established that allow for the categorization of patient groups and the identification of the best therapy for each patient, hence improving therapeutic outcomes [6,7]. However, in order to cover as many of the variables within a population as feasible, vast patient datasets are required. Radiological images obtained during routine examinations are an essential source of large-scale data that may be employed; nevertheless, imaging in a clinical setting is mostly used qualitatively to make diagnoses, but not kept for later analysis. The method of handcrafted radiomics provides a quantitative approach for measuring tumor heterogeneity by extracting a very large number of image characteristics from imaging data, using various mathematical techniques [8].

Handcrafted radiomics

The term "radiomics" refers to a set of mathematical formulas (handcrafted characteristics) that are extracted from regions of interest (ROI) in medical imaging [8]. A significant number of quantitative parameters can be rapidly retrieved with the use of high-throughput computing, providing for a more thorough description of lung diseases. Radiomic characteristics, in principle, can extract information not apparent to the naked eye and is capable of offering better predictions than other approaches. Hand-crafted radiomic features contain first-order statistics shape, texture, fractal dimension, and filter-based features [9]. To perform a radiomics study, a set of processes (workflow) has to be applied, and this includes image acquisition, segmentation, feature extraction, feature selection, and modeling (**Figure 1).**

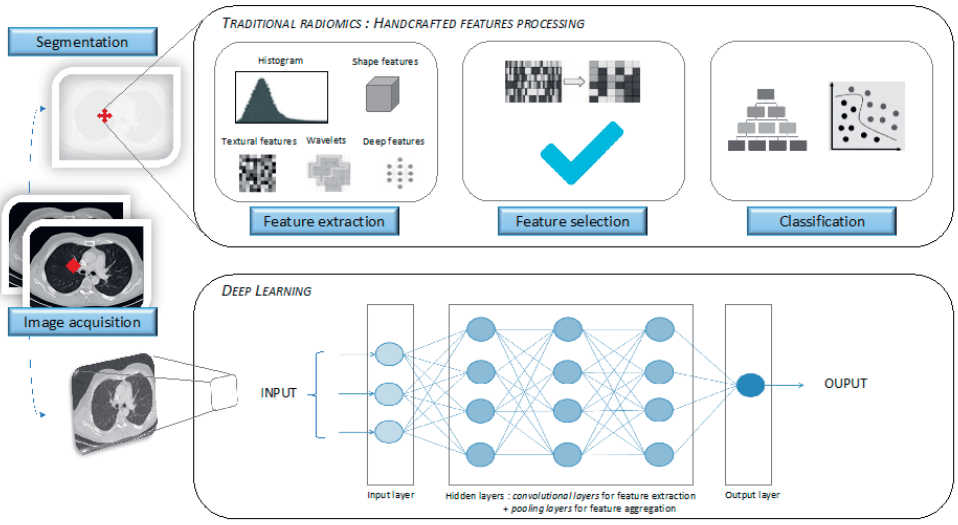


Figure 1. A standard radiomic analysis workflow for handcrafted features (top) and Deep Learning method (bottom) (Courtesy of Frix, A.N, 2021).

The handcrafted radiomics workflow starts with obtaining medical images which is the most important stage in any radiomics study. Images in radiomics studies are frequently gathered retrospectively, which implies that the images were collected in a non-controlled setting, using a range of different acquisition settings. As a result, image heterogeneities are frequently found inside and across datasets.

Segmentation and features extraction is the next step once the data has been collected and arranged. The regions of interest (ROIs) in the images are segmented for study. The ROI determines the area from which radiomic characteristics will be retrieved [10]. The ROIs considered in this thesis are a complete lung and sectorized lungs. The segmented ROI is used to calculate a collection of handcrafted image characteristics. Intensity, shape/volume, and texture features, as well as higher order features like radial-gradient and filtered features, are examples of these characteristics [11]. Intensity features are calculated from the histogram within the ROI – including the mean, median, standard deviation, and skewness. Texture features explore the relationship between one voxel and its neighbor inside the ROI, such as the quantification of the number of consecutive intensity values that occur in a certain direction. Shape features use the ROI to describe certain features, such as the sphericity and maximum diameter. Filtered features are computed after applying image filtering techniques (e.g., wavelet or Gaussian (LoG)).

Once the segmentation and features extraction are done, the next critical stage in the radiomics workflow is to reduce the number of features. Many of the retrieved features have no correlation with the outcome or have a substantial correlation with other radiomic

or clinical variables [10]. These features add no new information and should thus be avoided. Following the selection of the best features, radiomics models can be generated using a variety of machine learning techniques. Several modeling algorithms can be implemented, such as decision trees and logistic regressions. The performance of the models can be assessed using various metrics – including the area under the receiver operating characteristic curve (AUC) [12] and calibration plot [13] – which illustrate the connection between the true class of samples and the model prediction probabilities.

The approach of machine learning (ML) is a field of artificial intelligence in which an algorithm learns from a dataset via inference [14]. Its primary goal is to create a model capable of classifying, predicting, and estimating a scenario using the available data. Consequently, the technology may help clinicians make better decisions, since ML systems are able to consider a greater number of variables than human beings. Clinical observations, biology, genetics, and radiomics data may also be used to help improve decision-making. Deep learning is part of the machine learning field and uses the principles of simplified neuron interactions [15] and has already been shown to be extremely useful for solving image-processing tasks (**Figure 1)**.

Deep learning, or deep radiomics, is an alternative to (handcrafted) radiomics [16]. Deep learning methods usually involve feeding images directly into convolutional neural networks (CNN). Neural networks (NN) are models that have an input layer, many hidden layers, and an output layer. Each layer is made up of nodes that link to all the nodes in the preceding and subsequent layers [17]. Each node has a weight, and if its output exceeds a given threshold, it activates and transfers information to the next layer, finally going to the output layer, which provides a specific prediction. Deep learning has been successfully implemented in several different studies, with the use of medical imaging data.

Challenges in radiomics

Numerous publications highlighted the potential of radiomics in facilitating precision medicine. However, multiple obstacles hinder the generalizability of radiomics signatures which, therefore, influences the clinical translation. The most obvious limitation is the lack of reproducibility of radiomic biomarkers. Several studies have investigated the stability of radiomic features with test-retest or phantom experiments, and have reported that a considerable percentage of features are not reproducible, i.e. using the same acquisition and reconstruction parameters on the same vendor for acquiring the scan [17].

The first part of this thesis is devoted to the challenges facing radiomic features. Chapter 3 investigated the reproducibility of radiomic features across different scanners and scanning parameters. Chapter 4 evaluated the reproducibility of handcrafted radiomics across the arterial and portal venous phases of contrast-enhanced computed tomography images that depict hepatocellular carcinomas, as well as the potential of ComBat harmonization to

correct for these differences. In Chapter 5, we look at the reproducibility of HRFs derived from phantom CT scans taken with various reconstruction kernels on various imaging vendors, as well as the possibility of Reconstruction Kernel Normalization (RKN) and ComBat harmonization techniques to address the variations. Finally, chapter 6 evaluated the effects of differences in in-plane spatial resolution (IPR) on handcrafted radiomics, using a phantom dataset acquired on two scanner models.

The diseases with radiomics

The second part of this thesis is focused on the potential application of both handcrafted radiomics and deep learning in different lung disorders such as interstitial lung diseases (ILD) and chronic obstructive pulmonary diseases (COPD).

The term "interstitial lung disease" or ILD refers to a set of diffuse parenchymal lung disorders that are linked with high morbidity and mortality. Patients with fibrotic ILD often experience a decline in lung function with progressive symptoms, poor therapeutic response, and a lower quality of life. Idiopathic pulmonary fibrosis (IPF) is the most common, progressive, and severe subtype of ILD [18]. Although the disease was once thought to be rare, it now occurs with the same frequency as stomach, brain, and testicular tumors [19,20]. The prevalence of IPF has grown over time, with estimates ranging from 28 to 18 instances per 100,000 persons each year in Europe and North America [20,21]. IPF is more frequent in men and rare in those under the age of 50. (median age at diagnosis is about 65 years) [18,22,23]. Despite the fact that the disease's progression is diverse and somewhat unpredictable, the median survival period following diagnosis is 2–4 years [24]. IPF is usually associated with usual interstitial pneumonia (UIP) patterns on histology [25]. Although UIP is a defining feature of IPF, it is not unique to IPF and can be present in other ILDs, including connective tissue disease-associated ILD (CTD-ILD), hypersensitivity pneumonitis (HP), and sarcoidosis [26]. Accurate identification of IPF and UIP is important for prompt initiation of antifibrotic treatment and, when applicable, enrollment in clinical trials. According to the most recent ATS-ERS recommendations [25], radiologists only recognized a UIP pattern on thin-section CT with a sensitivity of 34% in a recent research that included a cohort of patients with pathologically verified UIP patterns [27]. Furthermore, radiographic evaluation of fibrotic lung disorders remains difficult and frequently varies amongst specialists [28–31]. Consequently, an automated technique that aids radiologists (particularly less experienced ones) in avoiding needless biopsies in the context of a multidisciplinary discussion might be extremely beneficial.

Chronic obstructive pulmonary disease (COPD) is one of the most common lung disorders, affecting an estimated 328 million people worldwide, and it is anticipated to become the leading cause of mortality in the world over the next two decades [21]. COPD is characterized by a restriction in airflow, which may be assessed via spirometry tests. It is not completely reversible

and is frequently induced by noxious particle or gas exposure (e.g., cigarette smoking), which causes an inflammatory reaction in the lungs [22,23]. COPD is a multicomponent disease, comprising a combination of bronchiolitis, emphysema, and extrapulmonary effects.

Role of computed tomography (CT)

In the majority of clinically suspected cases, high-resolution computed tomography (HRCT) can significantly reduce the differential diagnosis of interstitial lung disease (ILD) (**Figure 2**). In addition, HRCT can sometimes yield a precise diagnosis without requiring a surgical biopsy. HRCT may also be used to count the number of lung abnormalities and provide composite scores that can be used to assess disease severity and prognosis [24–26]. HRCT is a valuable tool for evaluating patients with suspected idiopathic pulmonary fibrosis (IPF) and is increasingly being used as a surrogate measure for monitoring therapeutic response in various drug trials [27–29].

A CT scan with adequate technical quality is necessary for the effective interpretation of ILDs findings [30]. The following parameters should be utilized to acquire a volumetric image: a) thin collimation; b) thin-slice thickness reconstructions (≤ 1.5 mm) with the use of a high-resolution filter; c) shortest rotation time and highest pitch, to reduce the motion artifacts and the acquisition time; and d) use of optimization tools to reduce radiation dose [19,31].

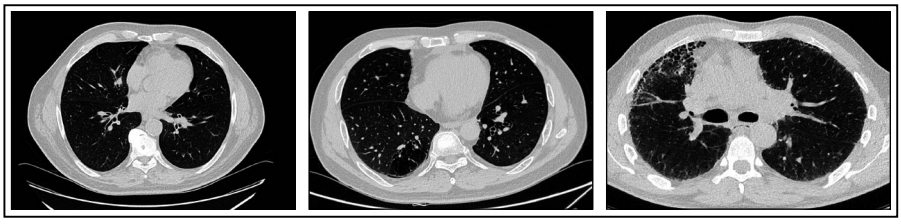


Figure 2. Figure shows CT of A) normal lungs; B) COPD lungs; C) ILD lungs.

Objectives and outline of the thesis

The overall aims of this thesis are two-fold; 1) to evaluate the effects of different scanners and scanning parameters on the reproducibility of radiomic features; 2) to investigate the use of radiomics in the classification between different ILDs. To this end, this thesis is divided into two parts. The first focuses on the challenges that the field of radiomics faces. The objective is to evaluate the reproducibility of radiomic features extracted from the same scanner, or from different scanners with different CT acquisition parameters. The second part concerns the application of handcrafted radiomics and deep learning in the classification of different types of lung disorders. The objective is to explore how the power of AI can be harnessed for the classification between different ILDs, potentially overcoming some of the current difficulties in the decision-making surrounding lung diseases. The outline of this thesis is shown in **Figure 3** and elaborated on below.

Part I: General introduction and outline of the thesis

Chapter 1 provides a general introduction to different lung diseases, the role of CT images, handcrafted radiomics, machine learning, and challenges regarding radiomics studies.

Chapter 2 provides a general overview and update on the recent rapidly expanding work in the field of handcrafted radiomics and deep learning, describing some of their limitations and providing examples of emerging clinical applications.

Part II: Challenges in handcrafted radiomics

Chapter 3 provides a study that attempts to test the repeatability of handcrafted radiomics using phantom scans. For this, a total of 13 scans were included and examined. These were collected with the use of various imaging vendors and reconstruction settings. The utilization of the ComBat harmonization approach was also explored.

Chapter 4 looks at the reproducibility of handcrafted radiomics derived from CT-based hepatocellular carcinoma in two imaging phases: arterial and portal venous. ComBat harmonization methods were also explored in order to evaluate their efficacy in reducing the impact of phase differences.

Chapter 5 investigates the reproducibility of HRFs extracted from phantom CT scans acquired with different reconstruction kernels on different imaging vendors. We also investigate the potential of ComBat harmonization, Reconstruction Kernel Normalization (RKN) and the combination of both methods to reduce the variations in HRF values attributed to differences solely in the reconstruction kernels of the original scans.

Chapter 6 assesses how differences in in-plane resolution can affect the reproducibility of handcrafted radiomics, when all other parameters are kept at a constant level. This study included two sets of phantom scans which were collected in the same manner except for the in-plane resolution. In addition, we explored the impact of various resampling methods and the application of ComBat harmonization on the reproducibility of handcrafted radiomics.

Part III: Application of handcrafted radiomics and deep learning on lung disease

Chapter 7 presents a review of the emerging role of radiomics in COPD and lung cancer. The review outlines the main applications of radiomics in lung cancer and briefly reviews the workflow from image acquisition to the evaluation of model performance. Furthermore, the current assessments of COPD and the potential application of radiomics in COPD were also discussed.

Chapter 8 investigates the use of handcrafted radiomics to classify between IPF with UIP presentation in HRCT or confirmed by lung biopsy and non-IPF ILD with the absence of UIP patterns (confirmed by lung biopsy). Furthermore, we examine the difference in trachea volume and use it as a predictor for IPF.

Chapter 9 compares the use of handcrafted radiomics and deep learning to diagnose diverse lung disorders, such as IPF, and non-IPF ILDs patients. Interpretability approaches were also utilized to explain the performance of handcrafted radiomics and deep learning. Furthermore, the suggested handcrafted radiomics and deep learning outcomes were compared to the performance of medical imaging experts.

Part V: General discussion and future perspective of the thesis

The thesis is concluded by **chapter 10**, in which the work in this thesis is discussed and the directions for future research are provided.

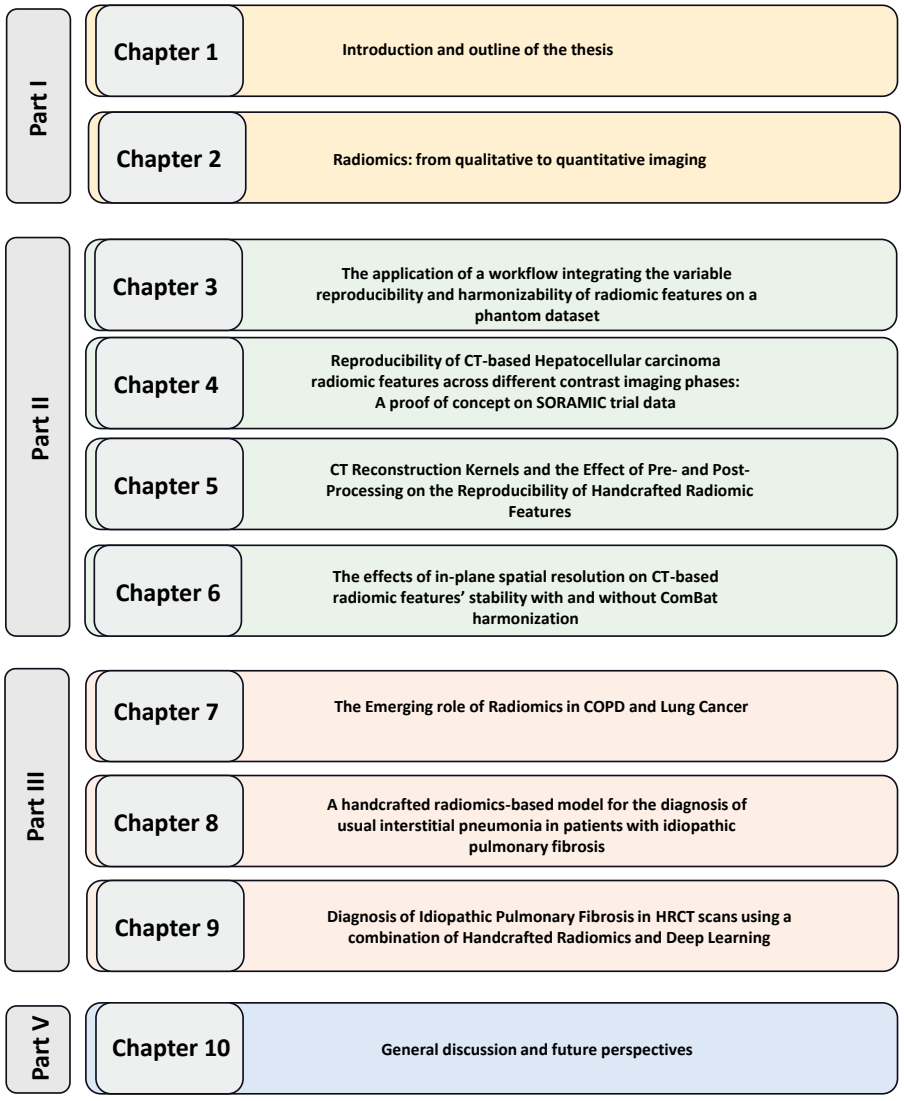


Figure 3. Outline of the thesis

References

1. Burstein, H.J.; Krilov, L.; Aragon-Ching, J.B.; Baxter, N.N.; Chiorean, E.G.; Chow, W.A.; De Groot, J.F.; Devine, S.M.; DuBois, S.G.; El-Deiry, W.S.; et al. Clinical Cancer Advances 2017: Annual Report on Progress Against Cancer From the American Society of Clinical Oncology. *J. Clin. Oncol.* **2017**, *35*, 1341–1367.
2. Gerlinger, M.; Rowan, A.J.; Horswell, S.; Math, M.; Larkin, J.; Endesfelder, D.; Gronroos, E.; Martinez, P.; Matthews, N.; Stewart, A.; et al. Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *N. Engl. J. Med.* **2012**, *366*, 883–892.
3. Curtis, C.; Shah, S.P.; Chin, S.-F.; Turashvili, G.; Rueda, O.M.; Dunning, M.J.; Speed, D.; Lynch, A.G.; Samarajiwa, S.; Yuan, Y.; et al. The Genomic and Transcriptomic Architecture of 2,000 Breast Tumours Reveals Novel Subgroups. *Nature* **2012**, *486*, 346–352.
4. Abernethy, A.P.; Etheredge, L.M.; Ganz, P.A.; Wallace, P.; German, R.R.; Neti, C.; Bach, P.B.; Murphy, S.B. Rapid-Learning System for Cancer Care. *J. Clin. Oncol.* **2010**, *28*, 4268–4274.
5. Garraway, L.A.; Verweij, J.; Ballman, K.V. Precision Oncology: An Overview. *J. Clin. Oncol.* **2013**, *31*, 1803–1805.
6. Collins, F.S.; Varmus, H. A New Initiative on Precision Medicine. *N. Engl. J. Med.* **2015**, *372*, 793–795.
7. Aerts, H.J.W.L.; van Baardwijk, A.A.W.; Petit, S.F.; Offermann, C.; van Loon, J.; Houben, R.; Dingemans, A.-M.C.; Wanders, R.; Boersma, L.; Borger, J.; et al. Identification of Residual Metabolic-Active Areas within Individual NSCLC Tumours Using a Pre-Radiotherapy (18) Fluorodeoxyglucose-PET-CT Scan. *Radiother. Oncol.* **2009**, *91*, 386–392.
8. Lambin, P.; Rios-Velazquez, E.; Leijenaar, R.; Carvalho, S.; van Stiphout, R.G.P.M.; Granton, P.; Zegers, C.M.L.; Gillies, R.; Boellard, R.; Dekker, A.; et al. Radiomics: Extracting More Information from Medical Images Using Advanced Feature Analysis. *Eur. J. Cancer* **2012**, *48*, 441–446.
9. van Griethuysen, J.J.M.; Fedorov, A.; Parmar, C.; Hosny, A.; Aucoin, N.; Narayan, V.; Beets-Tan, R.G.H.; Fillion-Robin, J.-C.; Pieper, S.; Aerts, H.J.W.L. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res.* **2017**, *77*, e104–e107.
10. Ibrahim, A.; Vallières, M.; Woodruff, H.; Primakov, S.; Beheshti, M.; Keek, S.; Refaee, T.; Sanduleanu, S.; Walsh, S.; Morin, O.; et al. Radiomics Analysis for Clinical Decision Support in Nuclear Medicine. *Semin. Nucl. Med.* **2019**, *49*, 438–449.
11. Aerts, H.J.W.L.; Velazquez, E.R.; Leijenaar, R.T.H.; Parmar, C.; Grossmann, P.; Carvalho, S.; Bussink, J.; Monshouwer, R.; Haibe-Kains, B.; Rietveld, D.; et al. Decoding Tumour Phenotype by Noninvasive Imaging Using a Quantitative Radiomics Approach. *Nat. Commun.* **2014**, *5*, 4006.
12. Huang, J.; Ling, C.X. Using AUC and Accuracy in Evaluating Learning Algorithms. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 299–310.
13. Steyerberg, E.W.; Vickers, A.J.; Cook, N.R.; Gerds, T.; Gonen, M.; Obuchowski, N.; Pencina, M.J.; Kattan, M.W. Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. *Epidemiology* **2010**, *21*, 128–138.
14. Rowe, M. An Introduction to Machine Learning for Clinicians. *Acad. Med.* **2019**, *94*, 1433–1436.

15. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444.
16. Schmidhuber, J. Deep Learning in Neural Networks: An Overview. *Neural Netw.* **2015**, *61*, 85–117.
17. Ibrahim, A.; Primakov, S.; Beuque, M.; Woodruff, H.C.; Halilaj, I.; Wu, G.; Refaee, T.; Granzier, R.; Widaatalla, Y.; Hustinx, R.; et al. Radiomics for Precision Medicine: Current Challenges, Future Prospects, and the Proposal of a New Framework. *Methods* **2021**, *188*, 20–29.
18. American Thoracic Society. Idiopathic Pulmonary Fibrosis: Diagnosis and Treatment. International Consensus Statement. American Thoracic Society (ATS), and the European Respiratory Society (ERS). *Am. J. Respir. Crit. Care Med.* **2000**, *161*, 646–664.
19. Parkin, D.M. *Cancer Incidence in Five Continents*; International Agency for Research on Cancer, 2002; ISBN 9789283221555.
20. Hutchinson, J.; Fogarty, A.; Hubbard, R.; McKeever, T. Global Incidence and Mortality of Idiopathic Pulmonary Fibrosis: A Systematic Review. *Eur. Respir. J.* **2015**, *46*, 795–806.
21. Hopkins, R.B.; Burke, N.; Fell, C.; Dion, G.; Kolb, M. Epidemiology and Survival of Idiopathic Pulmonary Fibrosis from National Data in Canada. *Eur. Respir. J.* **2016**, *48*, 187–195.
22. Raghu, G.; Collard, H.R.; Egan, J.J.; Martinez, F.J.; Behr, J.; Brown, K.K.; Colby, T.V.; Cordier, J.-F.; Flaherty, K.R.; Lasky, J.A.; et al. An Official ATS/ERS/JRS/ALAT Statement: Idiopathic Pulmonary Fibrosis: Evidence-Based Guidelines for Diagnosis and Management. *Am. J. Respir. Crit. Care Med.* **2011**, *183*, 788–824.
23. Raghu, G.; Weycker, D.; Edelsberg, J.; Bradford, W.Z.; Oster, G. Incidence and Prevalence of Idiopathic Pulmonary Fibrosis. *Am. J. Respir. Crit. Care Med.* **2006**, *174*, 810–816.
24. Ley, B.; Collard, H.R.; King, T.E., Jr Clinical Course and Prediction of Survival in Idiopathic Pulmonary Fibrosis. *Am. J. Respir. Crit. Care Med.* **2011**, *183*, 431–440.
25. Raghu, G.; Remy-Jardin, M.; Myers, J.L.; Richeldi, L.; Ryerson, C.J.; Lederer, D.J.; Behr, J.; Cottin, V.; Danoff, S.K.; Morell, F.; et al. Diagnosis of Idiopathic Pulmonary Fibrosis. An Official ATS/ERS/JRS/ALAT Clinical Practice Guideline. *Am. J. Respir. Crit. Care Med.* **2018**, *198*, e44–e68.
26. Chen, L.; Halai, V.; Leandru, A.; Wallis, A. Interstitial Lung Disease: Update on the Role of Computed Tomography in the Diagnosis of Idiopathic Pulmonary Fibrosis. *J. Comput. Assist. Tomogr.* **2019**, *43*, 898–905.
27. Richeldi, L.; Scholand, M.B.; Lynch, D.A.; Colby, T.V.; Myers, J.L.; Groshong, S.D.; Chung, J.H.; Benzaquen, S.; Nathan, S.D.; Davis, J.R.; et al. Utility of a Molecular Classifier as a Complement to High-Resolution Computed Tomography to Identify Usual Interstitial Pneumonia. *Am. J. Respir. Crit. Care Med.* **2021**, *203*, 211–220.
28. Gruden, J.F. CT in Idiopathic Pulmonary Fibrosis: Diagnosis and Beyond. *AJR Am. J. Roentgenol.* **2016**, *206*, 495–507.
29. Tominaga, J.; Sakai, F.; Johkoh, T.; Noma, S.; Akira, M.; Fujimoto, K.; Colby, T.V.; Ogura, T.; Inoue, Y.; Taniguchi, H.; et al. Diagnostic Certainty of Idiopathic Pulmonary Fibrosis/usual Interstitial Pneumonia: The Effect of the Integrated Clinico-Radiological Assessment. *Eur. J. Radiol.* **2015**, *84*, 2640–2645.

30. Walsh, S.L.F.; Calandriello, L.; Sverzellati, N.; Wells, A.U.; Hansell, D.M.; UIP Observer Consort Interobserver Agreement for the ATS/ERS/JRS/ALAT Criteria for a UIP Pattern on CT. *Thorax* **2016**, *71*, 45–51.
31. Walsh, S.L.F.; Wells, A.U.; Desai, S.R.; Poletti, V.; Piciucchi, S.; Dubini, A.; Nunes, H.; Valeyre, D.; Brillet, P.Y.; Kambouchner, M.; et al. Multicentre Evaluation of Multidisciplinary Team Meeting Agreement on Diagnosis in Diffuse Parenchymal Lung Disease: A Case-Cohort Study. *Lancet Respir Med* **2016**, *4*, 557–565.
32. Quaderi, S.A.; Hurst, J.R. The Unmet Global Burden of COPD. *Glob Health Epidemiol Genom* **2018**, *3*, e4.
33. Rabe, K.F.; Hurd, S.; Anzueto, A.; Barnes, P.J.; Buist, S.A.; Calverley, P.; Fukuchi, Y.; Jenkins, C.; Rodriguez-Roisin, R.; van Weel, C.; et al. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease: GOLD Executive Summary. *Am. J. Respir. Crit. Care Med.* **2007**, *176*, 532–555.
34. Celli, B.R.; MacNee, W.; ATS/ERS Task Force Standards for the Diagnosis and Treatment of Patients with COPD: A Summary of the ATS/ERS Position Paper. *Eur. Respir. J.* **2004**, *23*, 932–946.
35. Wells, A.U.; Desai, S.R.; Rubens, M.B.; Goh, N.S.L.; Cramer, D.; Nicholson, A.G.; Colby, T.V.; du Bois, R.M.; Hansell, D.M. Idiopathic Pulmonary Fibrosis: A Composite Physiologic Index Derived from Disease Extent Observed by Computed Tomography. *Am. J. Respir. Crit. Care Med.* **2003**, *167*, 962–969.
36. Best, A.C.; Meng, J.; Lynch, A.M.; Bozic, C.M.; Miller, D.; Grunwald, G.K.; Lynch, D.A. Idiopathic Pulmonary Fibrosis: Physiologic Tests, Quantitative CT Indexes, and CT Visual Scores as Predictors of Mortality. *Radiology* **2008**, *246*, 935–940.
37. Sumikawa, H.; Johkoh, T.; Colby, T.V.; Ichikado, K.; Suga, M.; Taniguchi, H.; Kondoh, Y.; Ogura, T.; Arakawa, H.; Fujimoto, K.; et al. Computed Tomography Findings in Pathological Usual Interstitial Pneumonia: Relationship to Survival. *Am. J. Respir. Crit. Care Med.* **2008**, *177*, 433–439.
38. Lynch, D.A.; Godwin, J.D.; Safrin, S.; Starko, K.M.; Hormel, P.; Brown, K.K.; Raghu, G.; King, T.E., Jr; Bradford, W.Z.; Schwartz, D.A.; et al. High-Resolution Computed Tomography in Idiopathic Pulmonary Fibrosis: Diagnosis and Prognosis. *Am. J. Respir. Crit. Care Med.* **2005**, *172*, 488–493.
39. Demedts, M.; Behr, J.; Buhl, R.; Costabel, U.; Dekhuijzen, R.; Jansen, H.M.; MacNee, W.; Thomeer, M.; Wallaert, B.; Laurent, F.; et al. High-Dose Acetylcysteine in Idiopathic Pulmonary Fibrosis. *N. Engl. J. Med.* **2005**, *353*, 2229–2242.
40. Bankier, A.A.; O'Donnell, C.R.; Boiselle, P.M. Quality Initiatives. Respiratory Instructions for CT Examinations of the Lungs: A Hands-on Guide. *Radiographics* **2008**, *28*, 919–931.
41. Lynch, D.A.; Sverzellati, N.; Travis, W.D.; Brown, K.K.; Colby, T.V.; Galvin, J.R.; Goldin, J.G.; Hansell, D.M.; Inoue, Y.; Johkoh, T.; et al. Diagnostic Criteria for Idiopathic Pulmonary Fibrosis: A Fleischner Society White Paper. *Lancet Respir Med* **2018**, *6*, 138–153.

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$minimum = \min(\mathbf{X})$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2$$

$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_s} \mathbf{P}(i,j)\right)^2}{N_p}$$

$$correlation = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) ij - \mu_x \mu_y}{\sigma_x(i) \sigma_y(j)}$$

$$MAD = \frac{1}{N_p} \sum_{i=1}^{N_p} |\mathbf{X}(i) - \bar{X}|$$

$$minimum = \min(\mathbf{X})$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_s} \mathbf{P}\right.}{N_z}$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

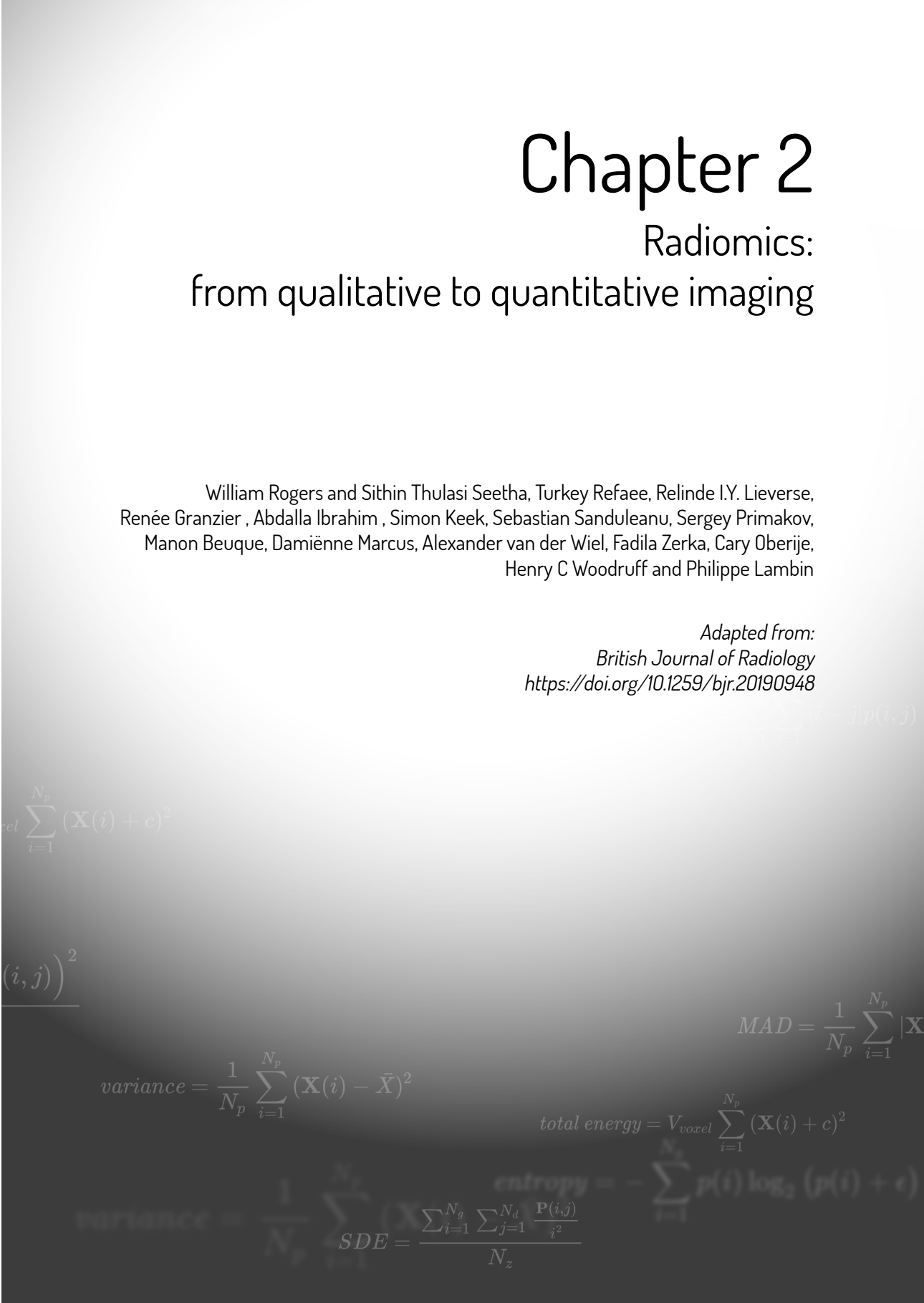
$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2$$

Chapter 2

Radiomics: from qualitative to quantitative imaging

William Rogers and Sithin Thulasi Seetha, Turkey Refaee, Relinde I.Y. Lieverse, Renée Granzier , Abdalla Ibrahim , Simon Keek, Sebastian Sanduleanu, Sergey Primakov, Manon Beuque, Damiënne Marcus, Alexander van der Wiel, Fadila Zerka, Cary Oberije, Henry C Woodruff and Philippe Lambin

Adapted from:
British Journal of Radiology
https://doi.org/10.1259/bjr.20190948



Abstract

Historically, medical imaging has been a qualitative or semi-quantitative modality. It is difficult to quantify what can be seen in an image, and to turn it into valuable predictive outcomes. As a result of advances in both computational hardware and machine learning algorithms, computers are making great strides in obtaining quantitative information from imaging and correlating it with outcomes. This opens a new "omics" field, radiomics, adding new input avenues for precision medicine, beyond genomics. Radiomics, in its two forms "handcrafted and deep", is an emerging field that translates medical images into quantitative data to yield biological information and enable radiologic phenotypic profiling for diagnosis, theragnosis, decision support, and monitoring. Within this review, we describe the steps of handcrafted radiomics, a multistage process in which features based on shape, pixel intensities, and texture are extracted from radiographs. The application of deep learning, the second arm of radiomics, and its place in the radiomics workflow is discussed, along with its advantages and disadvantages. To better illustrate the technologies being used, we provide real-world clinical applications of radiomics in oncology and other diseases, showcasing research on the applications of radiomics, as well as covering its limitations and its future direction towards precision medicine.

Keywords: Radiomics; Oncology; Machine Learning; Deep Learning; Medical Imaging.

Introduction

Medical imaging technologies in healthcare have expanded remarkably from the discovery of X-Rays 124 years ago to the use of Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and Positron Emission Tomography (PET), among others in modern-day clinical practice [1] (see Figure 1). These tools have become an integral part in detection and diagnosis for many diseases due to several factors, including: the minimally invasive nature of imaging, rapid technological developments, lower costs compared to alternatives, the high information density of images, and the hardware can be used for multiple diseases and sites [2,3].

Medical imaging in its infancy generated analogue images, which underwent subjective interpretation based on visual inspection and verbal communication. By the end of the 20th century, information technology has brought radiology to the digital world [4], although the interpretation of radiographs remained mostly qualitative. Humans excel at recognising patterns through visual inspection, however, they are often lacking when performing complex quantitative assessments [5,6]. In the early 1960s, researchers started to focus on computerized quantitative analysis of medical data for aiding clinical diagnosis [7–9], what later came to be known as Computer Aided Decision (CAD) systems. However, these systems were using a classical approach using statistical analysis and probability theories, and the volume of available data was low, so the results were often too inaccurate for clinical use. Later in the 1980s, further advances in theoretical computer science and digital imaging lead to the development of advanced machine learning and pattern recognition algorithms, which when integrated with CAD systems were able to generate clinically reliable results [10] [11].

In recent decades, simple quantitative image analysis has been adopted by clinicians (e.g. RECIST [12]), and has been primarily focused on assisting qualitative observations [13]. For instance, CAD systems can be found in health care worldwide, aiding radiologists and clinicians in making diagnostic and theragnostic decisions [14]. One of the most typical applications of CAD systems is in recognizing abnormalities during cancer screening [15]. Notable contributions are in the area of lung and breast cancer research. For example, there are many CAD studies which focus on detecting and diagnosing lung nodules [16,17] (as benign or malignant) on CT and chest radiographs. Similarly, many such studies have been conducted in breast mammography images for highlighting microcalcifications [18], architectural distortions, and the prediction of mass type [19,20].

It is conceivable that the lack of quantitative information leads to increased follow-ups or invasive biopsies that would be deemed unnecessary given the unused information in medical images [21]. Even though there have been various developments in quantitative

image analysis, traditionally radiologists are trained to understand the behaviour of the underlying disease through visual inspection of radiographic images [21]. This partially explains why most of the developments in imaging technology are in optimising the visual representation of the generated images, with vendors competing to generate the highest quality images. With the exception of CT, with its semi-parametric calibrated Hounsfield Units, and some particular MRI sequences, individual voxel values do not correlate with the underlying biology without further calibration and modelling. Furthermore, qualitative analysis is not so dependent on reproducible voxel values, while machines on the other hand only process numerical values and rely on the standardisation of image acquisition and reconstruction to yield reproducible results. The lack of standardisation of medical images has been a major hurdle in the development of quantitative image analysis (QIA) in medical imaging [22–25]. However, in recent years, quantitative imaging is becoming more popular with the advent of, e.g., quantitative FDG-PET [26,27] or quantitative MRI [28,29] for treatment response assessment.

The ubiquitous computer, vast amounts of data, and advanced algorithms have opened a new era in medical imaging. The high information density of images allows for many quantitative metrics since intricate pixel and voxel relationships can be captured by complex operations. Radiomics involves the process of extraction of quantifiable features from vast amounts of data that might correlate with the underlying biology or clinical outcomes using advanced machine learning analysis techniques [30,31]. Radiomics has two main arms, based on how imaging information is transformed into mineable data: handcrafted radiomics and deep learning. Handcrafted features are formulas mostly based on intensity histograms, shape attributes, and texture, that can be used to fingerprint phenotypical characteristics of the radiograph [32] while in deep learning a complex network “creates” its own features. Various statistical and machine learning models have been widely researched, and are envisioned to be complementary to best medical practice by aiding in making informed clinical decisions in both oncological and non-oncological diseases [33–36].

Since the 1990s predictions were being made that genomics, spearheaded by the Human Genome Project, would completely transform therapeutic medicine, heralding precision medicine [37]. Precision medicine, also termed personalized medicine, originally referred to the view that incorporating genomic information in the clinical workflow will lead to marked improvements in the prediction, diagnosis, and treatment of diseases. Recently, the scope of precision medicine has expanded to incorporate inputs beyond the genome [38]. Radiomics and other “-omic” developments, such as metabolomics and proteomics, are contributing to this a paradigm shift in medicine, where the focus has changed from standard clinical protocols based on trial populations to a personalised treatment tailored not only to the disease and site but also the patient, further enabling precision medicine.

In this review, we provide a broad overview and update on the fast-growing field of quantitative imaging research, focussing on the two arms "handcrafted radiomics and deep learning" describing some of its caveats and giving examples of the budding clinical implementation, the stepping stones towards precision medicine.

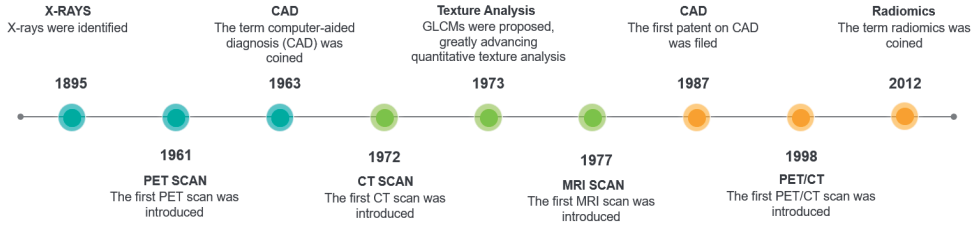


Figure 1. Timeline highlighting key developments in medical imaging.

Radiomics: from feature extraction to correlation with outcomes

Performing feature extraction of textures in medical imaging is nothing new and in fact serious research had begun in the early 1980s at Kurt Rossmann Laboratories for Radiologic Image Research in the Department of Radiology at the University of Chicago to develop CAD systems for the detection of lung nodules as well as detection of clustered microcalcifications in mammograms [39,40]. The first CAD patent was filed all the way back in 1987 using a method of pixel thresholding and contiguous pixel area thresholding [40].

The radiomic workflow begins with the medical image, which can be represented in two, three, or four dimensions [32,41]. Images contain quantitative data in the form of signals that are captured at different scales and variation across medical machines [42,43]. Normalisation techniques are used to evenly distribute pixel intensities across a dataset and within a standardized range [42,43]. Next, a region of interest (ROI) is defined so that only information related to the lesion can be extracted, and the useful information that can be extracted are called features. There are competing methods to extract features both in 2D and 3D. One such method is the manual segmentation of the lesion or the creation of a bounding box, as seen in Figure 2 [45,46]. This can also be performed using automated segmentation algorithms. Methods for automated segmentation include deep learning architectures such as U-Net, or semi-automatic methods like click-and-grow algorithms [45,46].

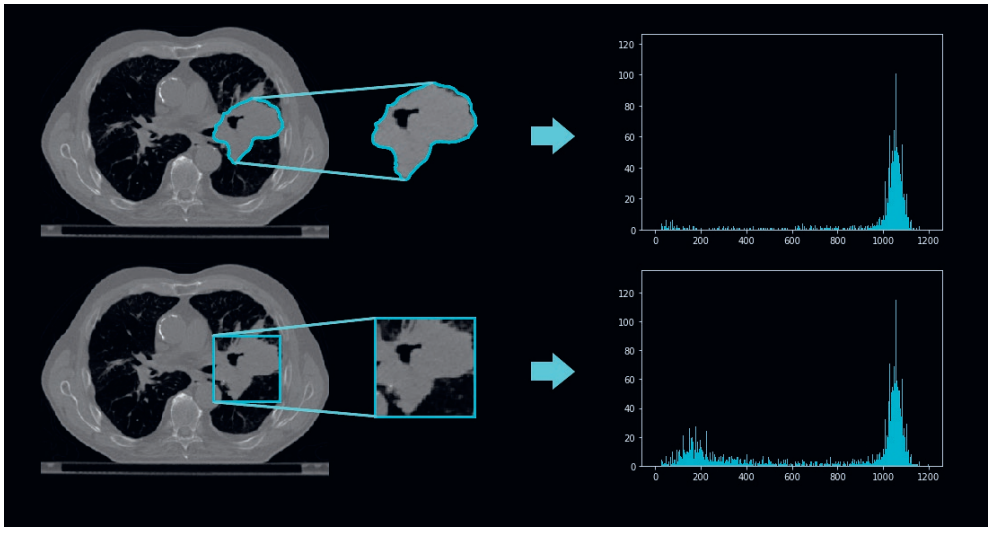


Figure 2. The difference between using A) a contoured binary mask, and B) using a bounding box.

Once the ROI is defined, the choice of features to be extracted depend on the information being sought. Shape features such as volume relate only to the definition of the ROI, and if this is manually created, suffer from inter-and intra-observer variability [47]. First-order features give insight into the distribution of pixel intensities, e.g. histograms of pixel intensities are quantified by a large number of statistical methods, including variance, skewness, and kurtosis. These features, however, are unable to quantify how pixels are positioned in relation to each other. Second and higher-order features may capture this relationship, with second-order features obtained based on the average relationship between two pixels/voxels, and higher-order features for more than two pixels/voxels. An example of a second-order feature extraction method is the grey level co-occurring matrix (GLCM). GLCMs are co-occurring pixels in each defined direction (see Figure 3) and are counted and recorded (see Figure 4) into a matrix. Statistical analysis such as contrast, correlation, and homogeneity, as well as tailored formulae can then be applied on the GLCM to extract independent features [48]. Features extracted in this manner are considered “hand crafted” features as they are features that are pre-defined by specially designed formulae.

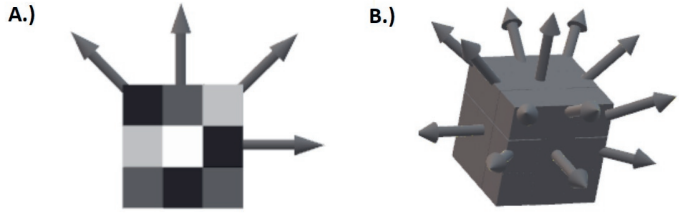


Figure 3. Possible angles for the calculation of co-occurrence matrices in two and three dimensions. A.) shows the four possible directions in two dimensions while B.) shows the thirteen possible directions in three dimensions.

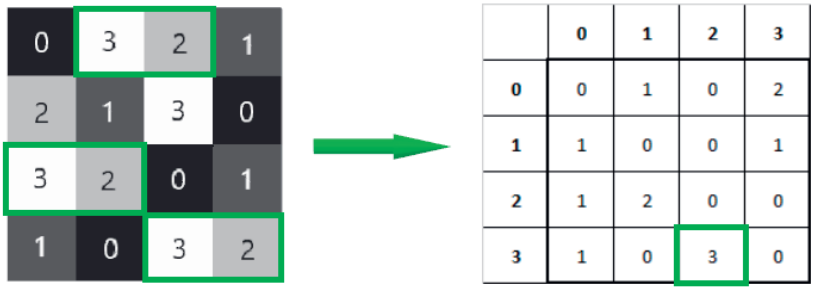


Figure 4. Calculating a GLCM for horizontal co-occurring pixel intensities. In total three co-occurring pixel intensities of 3 and 2 that are next to each other on a horizontal plane can be totalled and tracked in the corresponding matrix.

After features have been extracted from all the images in a database, a subset of features needs to be selected that go into the final model. To make a model generalisable, it is important to avoid finding spurious correlations in the data that do not generalise to other similar datasets, an occurrence termed overfitting [49–51]. If a model has learned to recognize noise, outliers, or other kinds of variance, it is unlikely to perform well when presented new data. The larger the number of predictors, the larger the chance to find spurious correlations, a major problem in the realm of machine learning [52]. To detect overfitting, ideally, a model's performance is validated in external datasets with similar population and outcome distributions, but from different centres -- if the model performs significantly better on the training set than on the validation set, overfitting is likely [53,54]. In the absence of an external validation dataset, data can be split into different subsets, and the model trained in one group and validated on the other(s) in a process called cross-validation (see Figure 4)[55]. During this process, the model hyper-parameters (settings within the model itself, e.g. degree of polynomial fitting) can be further tuned to increase performance in the training and validation sets [56].

Figure 5. An example of fivefold cross-validation which can be used to evaluate machine learning models. Cross-validation gives the ability to test the result across the entirety of a dataset, giving a better estimation of a model's overall performance.

A method to overcome overfitting is to reduce the number of predictors, in this case, imaging features. Feature selection is the process of reducing the number of predictors while retaining the core important information that correlates with outcomes or the underlying biology [32]. Many feature reduction methods exist, but none are known to work well on all kinds of datasets, and they can be combined in many ways [32]. This remains an active field of research [57]. Similar features can also be grouped to achieve dimensionality reduction, and methods such as principal component analysis and independent component analysis are employed to this end [58].

Once features are selected, the task is to correlate these features - individually or in groups - to diagnostic and prognostic outcomes or to the underlying biology. There are numerous methods to find and test such models, from simple linear regression and curve-fitting to advanced machine learning methods such as decision trees, support vector machines (SVM), random forests, boosted trees, or neural networks [59]. Ensembling is the combination of models that get trained on random samples of data from the training set called bags and then combined as a whole using a voting system. This is the basis for algorithms such as Random Forests, AdaBoost, and Gradient Boosting [60]. An intuitive explanation is that even though the individual models can show a large amount of variance due to being trained on small subsets of the data, their averaging or voting smooths out the variance while improving the ability to better generalise [60].

Once a generalisable model has been trained and externally validated, it might be desirable to expand the interoperability of the model to all hardware, acquisition, and reconstruction parameters found in general clinical practice. Instead of relying on the standardisation of images, the features themselves can be harmonized to a common frame-of-reference using combined batch methods such as ComBat [44,60,61], originally developed for similar problems encountered in gene sequencing assays [62].

Deep learning for fully automated workflows

Artificial neural networks (ANN) are a class of machine learning architecture that are loosely based on how biological brains work [63]. With the exception of unsupervised learning (such as autoencoders), deep learning architectures usually rely on information regarding the outcome in order to craft their features, and unlike in handcrafted radiomics, feature extraction and correlation are intertwined [64]. Also, unlike radiomics, there is generally no need for image segmentation, as the whole image can be presented to a deep learning model, both during training and in clinical routine.

An ANN is able to use a collection of neurons and weights, one for each of the inputs preceding the neuron [65]. These weights get continuously updated, or corrected, in steps called epochs

that work together to create a very complex function able to make predictions. The weights are inputs for each neuron and are multiplied and averaged, resulting in a transfer function, which is converted to an output via a function called an activation function [66]. These activation functions are often a sigmoidal function such as a hyperbolic tangent or sigmoid, or a function called a rectified linear unit (ReLU) that can be represented as the maximum of the product of the coefficient and zero or one. A representation of a single neuron, including the activation function, can be seen in Figure 6 [67]. Multiple neurons can then be stacked to create a single layer referred to as a "hidden layer" and hidden layers (were inputs and outputs all connect) can be stacked to create larger networks, see Figure 7 [65]. The term deep learning is used to describe a neural network that has many layers, which is considered deep. For a binary classifier or regression, the final layer should contain only a single neuron and use a sigmoid activation function to make a prediction with a binary outcome (zero or one). If the problem is categorical, the network's final layer should contain the same number of neurons as there are categories to be classified and the final activation will be a "softmax" function, which is the average of the exponentials of the inputs [68], yielding the probabilities of each category. Deep learning for image vision employs convolutional neural networks (CNN) which are a type of ANN that have an automated feature extractor designed specifically for images [69]. CNNs employ a filtering technique, which convolves the image with a kernel (sliding window), creating a new pixel/voxel value (and hence new image) by sliding a matrix of numbers over the image, see Figure 8. It is possible to make a variety of different filters using these types of convolutions, such as blurring, sharpening, edge detection, and gradient detection [69,70], and CNNs are able to learn filters that are best suited to extracting features needed for making predictions.

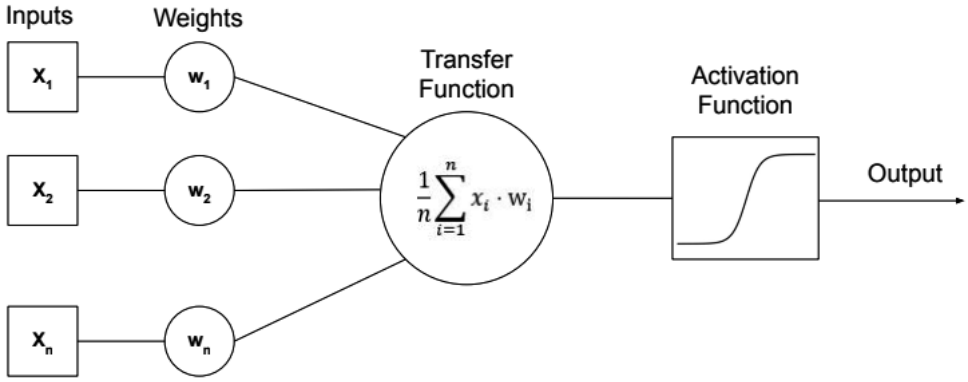


Figure 6. The architecture of a single neuron with a transfer function and a sigmoid activation function visualised.

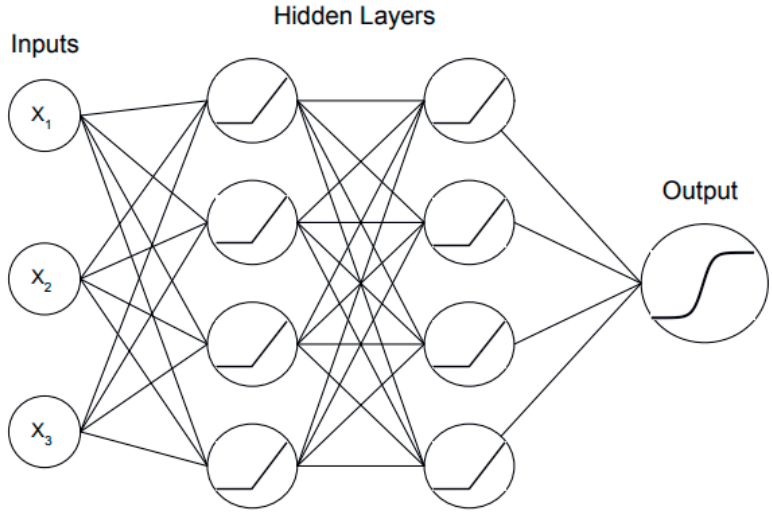


Figure 7. A three-layer neural network that is a binary classifier with three inputs. Nodes with X_n refer to inputs while other nodes refer to activation functions. The connecting lines between the nodes represent weights.

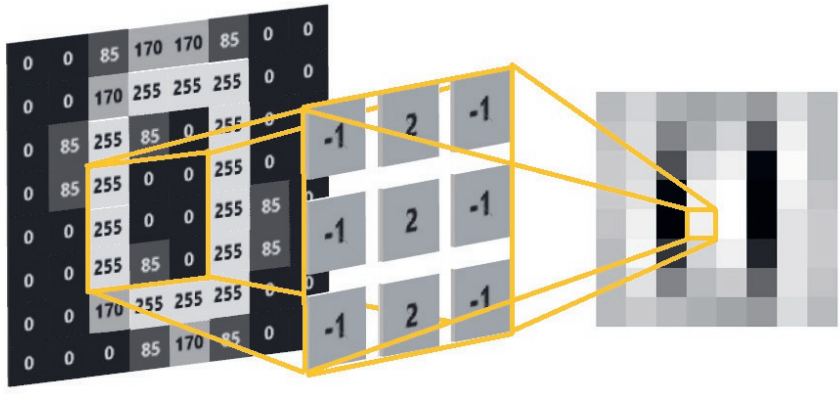


Figure 8. A filter that is able to filter out vertical lines. The yellow lines represent the kernel or sliding window, while the image on the right is the result of performing convolutions across the entirety of the original image.

ANNs do have some drawbacks compared to using hand crafted features alongside other machine learning techniques. The main drawback is the intrinsic need for much larger datasets to train the models, since feature creation is contingent on the training data, as opposed to handcrafted radiomics. Another drawback to using ANNs is interpretability. ANNs build ultra-complex functions that can be extremely difficult for practitioners to make sense of. Although CNNs have performed very well in image recognition, they have been less successful learning texture features, since texture information inherently has a higher dimensionality compared to other types of datasets, making them more difficult for neural networks to master [69,71]. According to Basu et al (2018), a redesign of neural network

architectures is required to extract features in a similar manner as GLCM and other features based on spatial correlation.

Currently, the main application of deep learning in the radiomics workflow still lies in the automated detection and localization of organs and lesions, removing the major burden in dataset curation. While there is no algorithm that can solve every problem, deep learning still has its place and is able to work as additional methods for delineation and feature extraction that compliments handcrafted radiomics. There is active research in combining both deep learning features and radiomics features that shows improved results [72–74].

Potential Clinical Applications

Radiomics in Oncology

Radiomics has been widely studied for application in diagnosis and treatment prognosis/selection in oncology, primarily due to the existence of large imaging datasets used for staging, often containing delineations of tumours and organs at risk necessary for radiation treatment planning. These datasets can be used to train diagnostic and prognostic models for a variety of cancer types and sites. Using clinical reports, pathology/histology, and genetic information along with radiomics analysis can give a global outlook on the biology of the disease [48]. In this section, an overview of notable studies published in this area will be discussed.

Lung:

Lung cancer is by far the leading cause of cancer-related deaths among both men and women worldwide [75]. Recent studies have shown that radiomics can determine the risk of lung cancer from screening scans [76–78]. Radiomic features found to have a strong association to decode tumour heterogeneity for risk stratification [79,80], concluding that patients with heterogeneous tumours tend to have a worse prognosis. In addition to that, Yoon et al. were able to show the association of radiomic analysis with gene expression [81]. Radiomic features were also found to correlate with TNM staging for lung and head-and-neck cancer [31,82]. Later studies further validated the strong predictive power of radiomics for distant metastasis [83–85].

Radiomics may also play a role in lung cancer treatment planning by evaluating tumour response to a specific treatment. Several studies focused on analysing the tumour response to radiation therapy [86,87]. For instance, Mattonen et al. developed a radiomics signature for treatment response to stereotactic ablative radiation therapy that was able to predict lung cancer recurrence post-therapy [86], while Fave et al. used multiple time point information referred to as delta-radiomic analysis to evaluate the change of radiomic features as a predictor for tumour response to radiation therapy [87]. The results suggest

that delta radiomic features are in fact a good indicator of treatment response. Another interesting study by Mattonen et al. found that radiomic analysis can identify features associated with local recurrence of lung cancer after radiation therapy [88], while physicians usually have great difficulty to distinguish local recurrence from radiation-induced sequelae.

Besides the traditional handcrafted feature extraction approach followed in the radiomics pipeline, deep learning radiomics is also gaining popularity among researchers. A deep learning-based approach followed by Shen et al. yielded more accurate malignancy prediction of nodules compared to previous methods [89]. Pham et al. used a two-step deep learning approach for evaluating lymph node metastases with accurate cancer detection [90]. Instead of using data from a single time point, deep recurrent convolutional network architectures can be used to analyse data from multiple time points to monitor treatment response [91].

Brain:
Brain tumours are usually graded based on clinical or pathological analysis to define their malignancy. Radiomics may be able to non-invasively perform grade assessment, as reported by Coroller et al. in meningioma patients, suggesting a strong correlation between certain imaging features and histopathologic grade [92]. Zhang et al. were able to classify between low-grade gliomas and high-grade gliomas with high accuracy [93]. Chen et al. investigated the prediction of brain metastases (BM) in T1 lung adenocarcinoma patients and found that the predictive performance for the radiomics model was significantly better compared to clinical models and could potentially be used for BM screening [94]. Fetit et al. performed radiomic analysis for the classification of brain tumours in childhood suggesting that radiomics can aid in the classification of tumour subtype [95]. However, the scalability of the techniques used in these studies needs to be assessed further by extensions to multicentric cohorts using different acquisition protocols and vendors.

Radiation therapy can lead to necrosis, which is difficult to distinguish from tumour recurrence on imaging. Larroza et al. were able to develop a high classification accuracy model to distinguish between brain metastasis and radiation necrosis using radiomic analysis [96]. Some radiomic studies successfully investigated the treatment response in recurrent glioblastoma patients with a radiomics approach [97–99]. An iterative study by radiomic researchers found strong evidence of radiomic features in predicting survival and treatment response of patients with glioblastoma using pre-treatment imaging data [100–102].

Deep learning has also made some other interesting contributions in this area. Chang et al. used residual deep convolutional network for predicting the genotype in grade II-IV glioma with high accuracy [103]. Deep learning can also be used complementary to traditional hand crafted radiomics studies. For example, studies [72,73] focused on using deep networks for segmentation, followed by radiomics analysis for survival prediction.

Breast:

Among women, breast cancer is the second leading cause of death for cancer worldwide [75]. However, earlier diagnosis can lead to a better prognosis. Radiomics in the field of breast cancer has been applied to several imaging modalities including (PET)-MRI, (contrast-enhanced) mammography, ultrasound, and digital breast tomosynthesis (DBT) focusing on tumour classification, molecular subtypes, tumour response prediction to neoadjuvant systemic therapy (NST), lymph node metastasis, overall survival, and recurrence risks. For example, a large number of radiomics studies have been used for the prediction of malignant breast cancers [104–107]. Besides the prediction of tumour malignancy, several radiomics studies examined the prediction of breast cancer molecular subtypes with the aim of leaving out liquid biopsies in the future [108–111]. Lymph node metastasis identification is an important prognostic factor and often determines treatment. In all clinically node negative patients, a sentinel lymph node procedure is the basis of the axillary treatment [112]. Dong et al. was able to provide an alternative to this invasive approach by successfully applying radiomics for the prediction of lymph node metastasis in the sentinel lymph node using imaging data [113].

In addition to the prediction of breast tumour malignancy, tumour molecular subtypes and sentinel lymph node metastasis identification, radiomics studies have also made some significant contributions to treatment planning. Chan et al. investigated the power of radiomics to discriminate between patients with low and high treatment failure risk on pre-treatment imaging data [114]. There are multiple studies that predict tumour response to neoadjuvant systemic therapy using radiomic analysis. For instance, Braman et al. found a combination of intratumoural and peritumoural radiomics features as a robust and strong indicator for pathologic complete tumour response using pre-treatment imaging data [115]. Two other studies [116,117] found similar evidence on serial imaging data containing follow-up scans. The use of multiparametric MRI for the prediction of tumour response to NST showed promising results [118,119].

Deep learning approaches have also been adopted in breast cancer research. The study of Huynh et al. investigated tumour classification capacity of deep features extracted from convolutional networks trained on a different dataset to analytically extracted features [120]. The results suggested a higher performance of deep features. Similarly, another study [121], used deep learning for risk assessment and found higher performance compared to conventional texture analysis.

Other sites and diseases

While cancers of the lung, brain, and breast have received wide attention from the radiomics research community, any site is open to QIA research. Diagnostic and prognostic radiomics research is ongoing for cancers of the head and neck [122], ovaries [38], prostate

[123], kidney [124], liver [125], colon and rectum [126], and many other sites. The main requirements for a radiomics study are the presence of a radiologic phenotype which allows for the clustering of patients based on differences within that phenotype or some correlation to the underlying biology, and the availability of imaging and clinical data. While not nearly as prevalent [127], this has meant that non-oncological diseases which require medical imaging as part of the standard of care have also been the subject of radiomics analysis, such as in the fields of neurology [35], ophthalmology [128], and dentistry [129].

Limitations of radiomics and future directions towards precision medicine

While radiomics facilitates new possibilities in the field of personalised medicine, some challenges remain. One of the primary obstacles is the lack of big and standardised clinical data. Although large amounts of medical imaging data are stored, these data are dispersed across different centres and acquired using different protocols. Access for research purposes is highly restricted by law and ethics. An exhaustive data curation and harmonization process is still necessary to make it usable for research. Radiomics will potentially enable imaging-based clinical decision support systems, however, the current black box approach, particularly in deep learning, makes it less acceptable for clinical application. In certain cases, hand crafted radiomic features have already been correlated with biological processes [130–132], but it is essential to work further in the direction of interpretable AI to make it more accessible for clinical implementation [33].

In recent years, various countries have already adopted many measures to control variability in clinical trial protocols, data acquisition, and analysis [133,134]. For example, across Europe consistent protocol guidance was adopted with the help of European Association of Nuclear Medicine [135]. The Quantitative Imaging Biomarker Alliance initiative also aims to achieve the same task in a much broader level [136,137]. On the other hand, algorithmically, developments in deep learning allow for automated quality check, clustering of data, and automated detection and contouring of organs and lesions, vastly improving data curation times. Generative adversarial networks open up the possibility of generating synthetic data [138] or domain adaptive algorithms [139,140] might be able to deal with the shortage of standardized data. Techniques like distributed learning provide the ability to train machine learning models using distributed data without the data ever leaving their original locations. Distributed learning has already been applied across several medical institutions to build predictive and segmentation models [141–144]. Furthermore, this approach can be coupled with other technologies such as blockchain to trace back data provenance and monitor the use of the final models [145]. Various techniques to visualize deep features have already been put forward by researchers to generate an intuitive understanding. A completely new research area of Artificial Intelligence called explainable AI aims to track the decisions made by the intelligent algorithms so that it can be better understood by humans. Companies

like Google, IBM, Microsoft and Facebook are at the forefront in this research. This will not only helps to build trust of AI systems among medical professionals but also unlocks new possibilities in understanding a disease [146,147].

The implementation of precision medicine itself has its own limitations and has drawn criticism due to the lack of a "transformation in therapeutic medicine" in the last two decades [148]. So far life expectancies or other public health measures have not shown any dramatic improvements, regardless of the vast amounts of precision medicine research being conducted. Contentious points remain such as excessive costs (e.g. gene therapy), although new developments such as radiomics promise to reduce costs in the long run. Furthermore the diagnostic and prognostic power of complex "omics-driven" models is still to be determined in specific populations, and evidence needs to be produced that such methods improve health outcomes [149]. Precision medicine is likely to mature and translate to clinical workflows over the next decade and will change the way health services are delivered and evaluated. Healthcare systems will need to adjust their methods and processes to accommodate for these changes.

Conclusion

Radiomics, whether handcrafted or deep, is an emerging field that translates medical images into quantitative data to give biological information and enable phenotypic profiling for diagnosis, theragnosis, decision support, and monitoring. Radiomics, in essence, allows personalised care by identifying features or signatures correlated with a disease or a treatment response with high precision and in a non-invasive way. Recent developments in genomics and deep learning have pushed radiomics researchers to focus more on extracting deep features and explore new possibilities in artificial intelligence modelling. In the future, radiomics will be a valued addition to precision medicine workflows by facilitating earlier and more accurate diagnosis, providing prognostic information, aiding in treatment choice, monitoring disease and treatment non-invasively, and enabling routine dynamic treatment based on individual responses. But the road to this vision is long, and many technical, regulatory, and ethical problems still need to be solved.

References

1. Scatliff JH, Morris PJ. From Roentgen to magnetic resonance imaging: the history of medical imaging. N C Med J. 2014 Mar;75(2):111–3.
2. Giakos GC, Pastorino M, Russo F, Chowdhury S, Shah N, Davros W. Noninvasive imaging for the new century [Internet]. Vol. 2, IEEE Instrumentation & Measurement Magazine. 1999. p. 32–5, 49. Available from: http://dx.doi.org/10.1109/5289.765967
3. Prince J, Links J. Medical Imaging: Signals and Systems (Prince, J.L. and Links, J.M.; 2006) [Book Review] [Internet]. Vol. 25, IEEE Signal Processing Magazine. 2008. p. 152–3. Available from: http://dx.doi.org/10.1109/msp.2008.4408454
4. Kesner A, Laforest R, Otazo R, Jennifer K, Pan T. Medical imaging data in the digital innovation age. Med Phys. 2018 Apr;45(4):e40–52.
5. Miller GA. The magical number seven plus or minus two: some limits on our capacity for processing information. Psychol Rev. 1956 Mar;63(2):81–97.
6. Wang Y-XJ, Ng CK. The impact of quantitative imaging in medicine and surgery: Charting our course for the future. Quant Imaging Med Surg. 2011 Dec;1(1):1–3.
7. Lodwick GS, Haun CL, Smith WE, Keller RF, Robertson ED. Computer Diagnosis of Primary Bone Tumors [Internet]. Vol. 80, Radiology. 1963. p. 273–5. Available from: http://dx.doi.org/10.1148/80.2.273
8. Meyers PH, Nice CM. Automated Computer Analysis of Radiographic Images [Internet]. Vol. 8, Archives of Environmental Health: An International Journal. 1964. p. 774–5. Available from: http://dx.doi.org/10.1080/00039896.1964.10663755
9. Winsberg F, Elkin M, Macy J, Bordaz V, Weymouth W. Detection of Radiographic Abnormalities in Mammograms by Means of Optical Scanning and Computer Analysis [Internet]. Vol. 89, Radiology. 1967. p. 211–5. Available from: http://dx.doi.org/10.1148/89.2.211
10. Summers RM. Road maps for advancement of radiologic computer-aided detection in the 21st century. Radiology. 2003 Oct;229(1):11–3.
11. Doi K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Comput Med Imaging Graph. 2007 Jun;31(4-5):198–211.
12. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer. 2009 Jan;45(2):228–47.
13. Zheng B. Identifying and testing new quantitative image, an analysis based clinical markers to predict breast cancer risk and prognosis [Internet]. Vol. 05, OMICS Journal of Radiology. 2016. Available from: http://dx.doi.org/10.4172/2167-7964.c1.009
14. Kobayashi T, Xu XW, MacMahon H, Metz CE, Doi K. Effect of a computer-aided diagnosis scheme on radiologists' performance in detection of lung nodules on radiographs. Radiology. 1996 Jun;199(3):843–8.
15. Halalli B, Makandar A. Computer Aided Diagnosis - Medical Image Analysis Techniques [Internet]. Breast Imaging. 2018. Available from: http://dx.doi.org/10.5772/intechopen.69792

16. The Robust Computer Aided Diagnostic System for Lung Nodule Diagnosis [Internet]. Vol. 8, International Journal of Recent Technology and Engineering. 2019. p. 5670–5. Available from: http://dx.doi.org/10.35940/ijrte.d8169.118419
17. Ziyad SR, Radha V, Vayyapuri T. Overview of Computer Aided Detection and Computer Aided Diagnosis Systems for Lung Nodule Detection in Computed Tomography [Internet]. Vol. 16, Current Medical Imaging Formerly Current Medical Imaging Reviews. 2020. p. 16–26. Available from: http://dx.doi.org/10.2174/1573405615666190206153321
18. Rizzi M, D'Aloia M, Castagnolo B. Computer aided detection of microcalcifications in digital mammograms adopting a wavelet decomposition [Internet]. Vol. 16, Integrated Computer-Aided Engineering. 2009. p. 91–103. Available from: http://dx.doi.org/10.3233/ica-2009-0306
19. Gibbs P, Turnbull LW. Textural analysis of contrast-enhanced MR images of the breast. Magn Reson Med. 2003 Jul;50(1):92–8.
20. Murakami R, Kumita S, Tani H, Yoshida T, Sugizaki K, Kuwako T, et al. Detection of breast cancer with a computer-aided detection applied to full-field digital mammography. J Digit Imaging. 2013 Aug;26(4):768–73.
21. Liu Z, Wang S, Dong D, Wei J, Fang C, Zhou X, et al. The Applications of Radiomics in Precision Diagnosis and Treatment of Oncology: Opportunities and Challenges. Theranostics. 2019 Feb 12;9(5):1303–22.
22. Hunter LA, Krafft S, Stingo F, Choi H, Martel MK, Kry SF, et al. High quality machine-robust image features: Identification in nonsmall cell lung cancer computed tomography images. Med Phys. 2013;40(12):121916.
23. Lu L, Ehmke RC, Schwartz LH, Zhao B. Assessing Agreement between Radiomic Features Computed for Multiple CT Imaging Settings. PLoS One. 2016 Dec 29;11(12):e0166550.
24. He L, Huang Y, Ma Z, Liang C, Liang C, Liu Z. Effects of contrast-enhancement, reconstruction slice thickness and convolution kernel on the diagnostic performance of radiomics signature in solitary pulmonary nodule [Internet]. Vol. 6, Scientific Reports. 2016. Available from: http://dx.doi.org/10.1038/srep34921
25. van Timmeren JE, Leijenaar RTH, van Elmpt W, Wang J, Zhang Z, Dekker A, et al. Test-Retest Data for Radiomics Feature Stability Analysis: Generalizable or Study-Specific? Tomography. 2016 Dec;2(4):361–5.
26. El Naqa I, Grigsby P, Apte A, Kidd E, Donnelly E, Khullar D, et al. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. Pattern Recognit. 2009 Jun 1;42(6):1162–71.
27. Ulaner GA. Measuring Treatment Response on FDG PET/CT [Internet]. Fundamentals of Oncologic PET/CT. 2019. p. 225–9. Available from: http://dx.doi.org/10.1016/b978-0-323-56869-2.00022-3
28. Xu Q-G, Xian J-F. Role of quantitative magnetic resonance imaging parameters in the evaluation of treatment response in malignant tumors. Chin Med J . 2015 Apr 20;128(8):1128–33.
29. Degnan AJ, Chung CY, Shah AJ. Quantitative diffusion-weighted magnetic resonance imaging assessment of chemotherapy treatment response of pediatric osteosarcoma and Ewing sarcoma

malignant bone tumors [Internet]. Vol. 47, Clinical Imaging. 2018. p. 9–13. Available from: http://dx.doi.org/10.1016/j.clinimag.2017.08.003
30. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RGPM, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012 Mar;48(4):441–6.
31. Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014 Jun 3;5:4006.
32. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine [Internet]. Vol. 14, Nature Reviews Clinical Oncology. 2017. p. 749–62. Available from: http://dx.doi.org/10.1038/nrclinonc.2017.141
33. Sanduleanu S, Woodruff HC, de Jong EEC. Tracking tumor biology with radiomics: A systematic review utilizing a radiomics quality score. Radiother Oncol [Internet]. 2018; Available from: https://www.sciencedirect.com/science/article/pii/S0167814018301798
34. Deist TM, Dankers FJWM, Valdes G, Wijsman R, Hsu I-C, Oberije C, et al. Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers. Med Phys. 2018 Jul;45(7):3449–59.
35. Ibrahim A, Vallières M, Woodruff H, Primakov S, Beheshti M, Keek S, et al. Radiomics Analysis for Clinical Decision Support in Nuclear Medicine [Internet]. Vol. 49, Seminars in Nuclear Medicine. 2019. p. 438–49. Available from: http://dx.doi.org/10.1053/j.semnuclmed.2019.06.005
36. Refaee T, Wu G, Ibrahim A, Halilaj I, Leijenaar RTH, Rogers W, et al. The Emerging Role of Radiomics in COPD and Lung Cancer. Respiration [Internet]. 2020 Jan 28; Available from: http://dx.doi.org/10.1159/000505429
37. Collins FS. Medical and Societal Consequences of the Human Genome Project [Internet]. Vol. 341, New England Journal of Medicine. 1999. p. 28–37. Available from: http://dx.doi.org/10.1056/nejm199907013410106
38. Nougaret S, Tardieu M, Vargas HA, Reinhold C, Vande Perre S, Bonanno N, et al. Ovarian cancer: An update on imaging in the era of radiomics. Diagn Interv Imaging. 2019 Oct;100(10):647–55.
39. Giger ML, Doi K, MacMahon H, Dwyer SJ III, Schneider RH. Computerized Detection Of Lung Nodules In Digital Chest Radiographs [Internet]. Medical Imaging. 1987. Available from: http://dx.doi.org/10.1117/12.967022
40. Giger ML, Doi K, MacMahon H. Image feature analysis and computer-aided diagnosis in digital radiography. 3. Automated detection of nodules in peripheral lung fields. Med Phys. 1988 Mar;15(2):158–66.
41. Larue RTHM, Van De Voorde L, van Timmeren JE, Leijenaar RTH, Berbée M, Sosef MN, et al. 4DCT imaging to assess radiomics feature stability: An investigation for thoracic cancers. Radiother Oncol. 2017 Oct;125(1):147-53.
42. Bagher-Ebadian H, Siddiqui F, Liu C, Movsas B, Chetty IJ. On the impact of smoothing and noise on robustness of CT and CBCT radiomics features for patients with head and neck cancers. Med Phys. 2017 May;44(5):1755–70.

43. Haga A, Takahashi W, Aoki S, Nawa K, Yamashita H, Abe O, et al. Standardization of imaging features for radiomics analysis. J Med Invest. 2019;66(1.2):35–7.
44. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007 Jan;8(1):118–27.
45. Kalpathy-Cramer. Radiomics of Lung Nodules: A Multi-Institutional Study of Robustness and Agreement of Quantitative Imaging Features. TOMOGRAPH. 2016 Dec;2(4).
46. Parmar C, Rios Velazquez E, Leijenaar R, Jermoumi M, Carvalho S, Mak RH, et al. Robust Radiomics feature quantification using semiautomatic volumetric segmentation. PLoS One. 2014 Jul 15;9(7):e102107.
47. Leijenaar RTH, Carvalho S, Velazquez ER, van Elmpt WJC, Parmar C, Hoekstra OS, et al. Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability. Acta Oncol. 2013 Oct;52(7):1391–7.
48. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology. 2016 Feb;278(2):563–77.
49. Rizzo S, Botta F, Raimondi S, Origgi D, Fanciullo C, Morganti AG, et al. Radiomics: the facts and the challenges of image analysis. Eur Radiol Exp. 2018 Nov 14;2(1):36.
50. Vial A, Stirling D, Field M, Ros M, Ritz C, Carolan M, et al. The role of deep learning and radiomic feature extraction in cancer-specific predictive modelling: a review [Internet]. Vol. 7, Translational Cancer Research. 2018. p. 803–16. Available from: http://dx.doi.org/10.21037/tcr.2018.05.02
51. Fan J, Shao Q-M, Zhou W-X. ARE DISCOVERIES SPURIOUS? DISTRIBUTIONS OF MAXIMUM SPURIOUS CORRELATIONS AND THEIR APPLICATIONS. Ann Stat. 2018 Jun;46(3):989–1017.
52. Lever J, Krzywinski M, Altman N. Model selection and overfitting [Internet]. Vol. 13, Nature Methods. 2016. p. 703–4. Available from: http://dx.doi.org/10.1038/nmeth.3968
53. Parmar C, Barry JD, Hosny A, Quackenbush J, Aerts HJWL. Data Analysis Strategies in Medical Imaging. Clin Cancer Res. 2018 Aug 1;24(15):3492–9.
54. Chatterjee A, Vallières M, Dohan A, Levesque IR, Ueno Y, Bist V, et al. An Empirical Approach for Avoiding False Discoveries When Applying High-Dimensional Radiomics to Small Datasets. IEEE Transactions on Radiation and Plasma Medical Sciences. 2019 Mar;3(2):201–9.
55. Wong T-T. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation [Internet]. Vol. 48, Pattern Recognition. 2015. p. 2839–46. Available from: http://dx.doi.org/10.1016/j.patcog.2015.03.009
56. Duarte E, Wainer J. Empirical comparison of cross-validation and internal metrics for tuning SVM hyperparameters [Internet]. Vol. 88, Pattern Recognition Letters. 2017. p. 6–11. Available from: http://dx.doi.org/10.1016/j.patrec.2017.01.007
57. Solorio-Fernández S, Ariel Carrasco-Ochoa J, Martínez-Trinidad JF. A review of unsupervised feature selection methods [Internet]. Artificial Intelligence Review. 2019. Available from: http://dx.doi.org/10.1007/s10462-019-09682-y
58. Zhang D, Zou L, Zhou X, He F. Integrating Feature Selection and Feature Extraction Methods With Deep Learning to Predict Clinical Outcome of Breast Cancer [Internet]. Vol. 6, IEEE Access. 2018. p. 28936–44. Available from: http://dx.doi.org/10.1109/access.2018.2837654

59. Choudhary R, Gianey HK. Comprehensive Review On Supervised Machine Learning Algorithms [Internet]. 2017 International Conference on Machine Learning and Data Science (MLDS). 2017. Available from: http://dx.doi.org/10.1109/mlds.2017.11
60. Ren Y, Zhang L, Suganthan PN. Ensemble Classification and Regression-Recent Developments, Applications and Future Directions [Review Article] [Internet]. Vol. 11, IEEE Computational Intelligence Magazine. 2016. p. 41–53. Available from: http://dx.doi.org/10.1109/mci.2015.2471235
61. Lovinfosse P, Visvikis D, Hustinx R, Hatt M. FDG PET radiomics: a review of the methodological aspects [Internet]. Vol. 6, Clinical and Translational Imaging. 2018. p. 379–91. Available from: http://dx.doi.org/10.1007/s40336-018-0292-9
62. Lucia F, Visvikis D, Vallières M, Desseroit M-C, Miranda O, Robin P, et al. External validation of a combined PET and MRI radiomics model for prediction of recurrence in cervical cancer patients treated with chemoradiotherapy. Eur J Nucl Med Mol Imaging. 2019 Apr;46(4):864–77.
63. Kriegeskorte N. Deep neural networks: a new framework for modelling biological vision and brain information processing [Internet]. Available from: http://dx.doi.org/10.1101/029876
64. Bengio Y, Delalleau O, Le Roux N. The Curse of Highly Variable Functions for Local Kernel Machines. Adv Neural Inf Process Syst. 2005;18:107–14.
65. Hinton GE. Learning multiple layers of representation. Trends Cogn Sci. 2007 Oct;11(10):428–34.
66. LeCun Y. Deep learning & convolutional networks [Internet]. 2015 IEEE Hot Chips 27 Symposium (HCS). 2015. Available from: http://dx.doi.org/10.1109/hotchips.2015.7477328
67. LeCun Y, Bengio Y, Hinton G. Deep learning [Internet]. Vol. 521, Nature. 2015. p. 436–44. Available from: http://dx.doi.org/10.1038/nature14539
68. Aggarwal CC. Neural Networks and Deep Learning: A Textbook. Springer; 2018. 497 p.
69. Guo Y, Liu Y, Oerlemans A, Lao S, Wu S, Lew MS. Deep learning for visual understanding: A review. Neurocomputing. 2016 Apr 26;187:27–48.
70. Tyagi V. Introduction to Digital Image Processing [Internet]. Understanding Digital Image Processing. 2018. p. 1–12. Available from: http://dx.doi.org/10.1201/9781315123905-1
71. Basu S, Mukhopadhyay S, Karki M, DiBiano R, Ganguly S, Nemani R, et al. Deep neural networks for texture classification—A theoretical analysis [Internet]. Vol. 97, Neural Networks. 2018. p. 173–82. Available from: http://dx.doi.org/10.1016/j.neunet.2017.10.001
72. Yogananda CGB, Nalawade SS, Murugesan GK, Wagner B, Pinho MC, Fei B, et al. Fully Automated Brain Tumor Segmentation and Survival Prediction of Gliomas using Deep Learning and MRI [Internet]. Available from: http://dx.doi.org/10.1101/760157
73. Sun L, Zhang S, Chen H, Luo L. Brain Tumor Segmentation and Survival Prediction Using Multimodal MRI Scans With Deep Learning. Front Neurosci. 2019 Aug 16;13:810.
74. A. Jochems, R. T. H. Leijenaar, M. Bogowicz, F. J. P. Hoebers, F. Wesseling, S. H. Huang, B. Chan, J. N. Waldron, B. O'Sullivan, D. Rietveld, C. R. Leemans, O. Riesterer, S. Tanadini-Lang, M. Guckenberger, K. Ikenberg, P. Lambin. Combining deep learning and radiomics to predict HPV status in oropharyngeal squamous cell carcinoma. Radiotherapy & Oncology. 2018 Apr;127:S504–5.

75. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018 Nov;68(6):394–424.
76. Hawkins S, Wang H, Liu Y, Garcia A, Stringfield O, Krewer H, et al. Predicting Malignant Nodules from Screening CT Scans. J Thorac Oncol. 2016 Dec;11(12):2120–8.
77. Kumar D, Chung AG, Shaifee MJ, Khalvati F, Haider MA, Wong A. Discovery Radiomics for Pathologically-Proven Computed Tomography Lung Cancer Prediction [Internet]. Lecture Notes in Computer Science. 2017. p. 54–62. Available from: http://dx.doi.org/10.1007/978-3-319-59876-5_7
78. Liu Y, Balagurunathan Y, Atwater T, Antic S, Li Q, Walker RC, et al. Radiological Image Traits Predictive of Cancer Status in Pulmonary Nodules. Clin Cancer Res. 2017 Mar 15;23(6):1442–9.
79. Ganeshan B, Panayiotou E, Burnand K, Dizdarevic S, Miles K. Tumour heterogeneity in non-small cell lung carcinoma assessed by CT texture analysis: a potential marker of survival. Eur Radiol. 2012 Apr;22(4):796–802.
80. Yoon HJ, Sohn I, Cho JH, Lee HY, Kim J-H, Choi Y-L, et al. Decoding Tumor Phenotypes for ALK, ROS1, and RET Fusions in Lung Adenocarcinoma Using a Radiomics Approach. Medicine . 2015 Oct;94(41):e1753.
81. Yoon HJ, Sohn I, Cho JH, Lee HY, Kim J-H, Choi Y-L, et al. Decoding Tumor Phenotypes for ALK, ROS1, and RET Fusions in Lung Adenocarcinoma Using a Radiomics Approach. Medicine . 2015 Oct;94(41):e1753.
82. Parmar C, Leijenaar RTH, Grossmann P, Rios Velazquez E, Bussink J, Rietveld D, et al. Radiomic feature clusters and prognostic signatures specific for Lung and Head & Neck cancer. Sci Rep. 2015 Jun 5;5:11044.
83. Coroller TP, Grossmann P, Hou Y, Rios Velazquez E, Leijenaar RTH, Hermann G, et al. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. Radiother Oncol. 2015 Mar;114(3):345–50.
84. Wu J, Aguilera T, Shultz D, Gudur M, Rubin DL, Loo BW Jr, et al. Early-Stage Non-Small Cell Lung Cancer: Quantitative Imaging Characteristics of (18)F Fluorodeoxyglucose PET/CT Allow Prediction of Distant Metastasis. Radiology. 2016 Oct;281(1):270–8.
85. Zhou H, Dong D, Chen B, Fang M, Cheng Y, Gan Y, et al. Diagnosis of Distant Metastasis of Lung Cancer: Based on Clinical and Radiomic Features. Transl Oncol. 2018 Feb;11(1):31–6.
86. Mattonen SA, Tetar S, Palma DA, Senan S, Ward AD. Automated Texture Analysis for Prediction of Recurrence After Stereotactic Ablative Radiation Therapy for Lung Cancer [Internet]. Vol. 93, International Journal of Radiation Oncology*Biology*Physics. 2015. p. S5–6. Available from: http://dx.doi.org/10.1016/j.ijrobp.2015.07.019
87. Fave X, Zhang L, Yang J, Mackin D, Balter P, Gomez D, et al. Delta-radiomics features for the prediction of patient outcomes in non-small cell lung cancer. Sci Rep. 2017 Apr 3;7(1):588.
88. Mattonen SA, Palma DA, Johnson C, Louie AV, Landis M, Rodrigues G, et al. Detection of Local Cancer Recurrence After Stereotactic Ablative Radiation Therapy for Lung Cancer: Physician Performance Versus Radiomic Assessment. Int J Radiat Oncol Biol Phys. 2016 Apr 1;94(5):1121–8.

89. Shen W, Zhou M, Yang F, Yu D, Dong D, Yang C, et al. Multi-crop Convolutional Neural Networks for lung nodule malignancy suspiciousness classification [Internet]. Vol. 61, Pattern Recognition. 2017. p. 663–73. Available from: http://dx.doi.org/10.1016/j.patcog.2016.05.029
90. Pham HHN, Futakuchi M, Bychkov A, Furukawa T, Kuroda K, Fukuoka J. Detection of lung cancer lymph node metastases from whole-slide histopathological images using a two-step deep learning approach. Am J Pathol [Internet]. 2019 Sep 18; Available from: http://dx.doi.org/10.1016/j.ajpath.2019.08.014
91. Xu Y, Hosny A, Zeleznik R, Parmar C, Coroller T, Franco I, et al. Deep Learning Predicts Lung Cancer Treatment Response from Serial Medical Imaging. Clin Cancer Res. 2019 Jun 1;25(11):3266–75.
92. Coroller TP, Bi WL, Huynh E, Abedalthagafi M, Aizer AA, Greenwald NF, et al. Radiographic prediction of meningioma grade by semantic and radiomic features. PLoS One. 2017 Nov 16;12(11):e0187908.
93. Zhang X, Yan L-F, Hu Y-C, Li G, Yang Y, Han Y, et al. Optimizing a machine learning based glioma grading system using multi-parametric MRI histogram and texture features. Oncotarget. 2017 Jul 18;8(29):47816–30.
94. Chen A, Lu L, Pu X, Yu T, Yang H, Schwartz LH, et al. CT-Based Radiomics Model for Predicting Brain Metastasis in Category T1 Lung Adenocarcinoma. AJR Am J Roentgenol. 2019 Apr 1;1–6.
95. Fetit AE, Novak J, Peet AC, Arvanitits TN. Three-dimensional textural features of conventional MRI improve diagnostic classification of childhood brain tumours. NMR Biomed. 2015 Sep;28(9):1174–84.
96. Larroza A, Moratal D, Paredes-Sánchez A, Soria-Olivas E, Chust ML, Arribas LA, et al. Support vector machine classification of brain metastasis and radiation necrosis based on texture analysis in MRI. J Magn Reson Imaging. 2015 Nov;42(5):1362–8.
97. Kickingereder P, Götz M, Muschelli J, Wick A, Neuberger U, Shinohara RT, et al. Large-scale Radiomic Profiling of Recurrent Glioblastoma Identifies an Imaging Predictor for Stratifying Anti-Angiogenic Treatment Response. Clin Cancer Res. 2016 Dec 1;22(23):5765–71.
98. Chang K, Zhang B, Guo X, Zong M, Rahman R, Sanchez D, et al. Multimodal imaging patterns predict survival in recurrent glioblastoma patients treated with bevacizumab. Neuro Oncol. 2016 Dec;18(12):1680–7.
99. Grossmann P, Narayan V, Chang K, Rahman R, Abrey L, Reardon DA, et al. Quantitative imaging biomarkers for risk stratification of patients with recurrent glioblastoma treated with bevacizumab [Internet]. Vol. 19, Neuro-Oncology. 2017. p. 1688–97. Available from: http://dx.doi.org/10.1093/neuonc/nox092
100. Pérez-Beteta J, Molina D, Martínez-González A, Arregui E, Asenjo B, Iglesias L, et al. P09.43 Novel geometrical imaging biomarkers predict survival and allow for patient selection for surgery in glioblastoma patients [Internet]. Vol. 19, Neuro-Oncology. 2017. p. iii80–iii80. Available from: http://dx.doi.org/10.1093/neuonc/nox036.299

101. Pérez-Beteta J, Molina-García D, Ortiz-Alhambra JA, Fernández-Romero A, Luque B, Arregui E, et al. Tumor Surface Regularity at MR Imaging Predicts Survival and Response to Surgery in Patients with Glioblastoma. Radiology. 2018 Jul;288(1):218–25.
102. Pérez-Beteta J, Molina-García D, Martínez-González A, Henares-Molina A, Amo-Salas M, Luque B, et al. Correction to: Morphological MRI-based features provide pretreatment survival prediction in glioblastoma [Internet]. Vol. 29, European Radiology. 2019. p. 2729–2729. Available from: http://dx.doi.org/10.1007/s00330-018-5870-8
103. Chang K, Bai HX, Zhou H, Su C, Bi WL, Agbodza E, et al. Residual Convolutional Neural Network for the Determination of Status in Low- and High-Grade Gliomas from MR Imaging. Clin Cancer Res. 2018 Mar 1;24(5):1073–81.
104. Bickelhaupt S, Paech D, Kickingereder P, Steudle F, Lederer W, Daniel H, et al. Prediction of malignancy by a radiomic signature from contrast agent-free diffusion MRI in suspicious breast lesions found on screening mammography [Internet]. Vol. 46, Journal of Magnetic Resonance Imaging. 2017. p. 604–16. Available from: http://dx.doi.org/10.1002/jmri.25606
105. Parekh VS, Jacobs MA. Integrated radiomic framework for breast cancer and tumor biology using advanced machine learning and multiparametric MRI [Internet]. Vol. 3, npj Breast Cancer. 2017. Available from: http://dx.doi.org/10.1038/s41523-017-0045-3
106. Bickelhaupt S, Jaeger PF, Laun FB, Lederer W, Daniel H, Kuder TA, et al. Radiomics Based on Adapted Diffusion Kurtosis Imaging Helps to Clarify Most Mammographic Findings Suspicious for Cancer. Radiology. 2018 Jun;287(3):761–70.
107. Whitney HM, Taylor NS, Drukker K, Edwards AV, Papaioannou J, Schacht D, et al. Additive Benefit of Radiomics Over Size Alone in the Distinction Between Benign Lesions and Luminal A Cancers on a Large Clinical Breast MRI Dataset. Acad Radiol. 2019 Feb;26(2):202–9.
108. Guo W, Li H, Zhu Y, Lan L, Yang S, Drukker K, et al. Prediction of clinical phenotypes in invasive breast carcinomas from the integration of radiomics and genomics data. J Med Imaging (Bellingham). 2015 Oct;2(4):041007.
109. Li H, Zhu Y, Burnside ES, Huang E, Drukker K, Hoadley KA, et al. Quantitative MRI radiomics in the prediction of molecular classifications of breast cancer subtypes in the TCGA/TCIA data set. NPJ Breast Cancer [Internet]. 2016 May 11;2. Available from: http://dx.doi.org/10.1038/npjbcancer.2016.12
110. Fan M, Li H, Wang S, Zheng B, Zhang J, Li L. Radiomic analysis reveals DCE-MRI features for prediction of molecular subtypes of breast cancer [Internet]. Vol. 12, PLOS ONE. 2017. p. e0171683. Available from: http://dx.doi.org/10.1371/journal.pone.0171683
111. Ma W, Zhao Y, Ji Y, Guo X, Jian X, Liu P, et al. Breast Cancer Molecular Subtype Prediction by Mammographic Radiomic Features. Acad Radiol. 2019 Feb;26(2):196–201.
112. Dong Y, Feng Q, Yang W, Lu Z, Deng C, Zhang L, et al. Preoperative prediction of sentinel lymph node metastasis in breast cancer based on radiomics of T2-weighted fat-suppression and diffusion-weighted MRI. Eur Radiol. 2018 Feb;28(2):582–91.

113. Chang K, Bai HX, Zhou H, Su C, Bi WL, Agbodza E, et al. Residual Convolutional Neural Network for the Determination of Status in Low- and High-Grade Gliomas from MR Imaging. Clin Cancer Res. 2018 Mar 1;24(5):1073–81.
114. Chan HM, van der Velden BHM, Loo CE, Gilhuijs KGA. Eigentumors for prediction of treatment failure in patients with early-stage breast cancer using dynamic contrast-enhanced MRI: a feasibility study [Internet]. Vol. 62, Physics in Medicine & Biology. 2017. p. 6467–85. Available from: http://dx.doi.org/10.1088/1361-6560/aa7dc5
115. Braman NM, Etesami M, Prasanna P, Dubchuk C, Gilmore H, Tiwari P, et al. Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast DCE-MRI. Breast Cancer Res. 2017 May 18;19(1):57.
116. Henderson S, Purdie C, Michie C, Evans A, Lerski R, Johnston M, et al. Interim heterogeneity changes measured using entropy texture features on T2-weighted MRI at 3.0 T are associated with pathological response to neoadjuvant chemotherapy in primary breast cancer [Internet]. Vol. 27, European Radiology. 2017. p. 4602–11. Available from: http://dx.doi.org/10.1007/s00330-017-4850-8
117. Parikh J, Selmi M, Charles-Edwards G, Glendenning J, Ganeshan B, Verma H, et al. Changes in primary breast cancer heterogeneity may augment midtreatment MR imaging assessment of response to neoadjuvant chemotherapy. Radiology. 2014 Jul;272(1):100–12.
118. Liu Z, Li Z, Qu J, Zhang R, Zhou X, Li L, et al. Radiomics of Multiparametric MRI for Pretreatment Prediction of Pathologic Complete Response to Neoadjuvant Chemotherapy in Breast Cancer: A Multicenter Study [Internet]. Vol. 25, Clinical Cancer Research. 2019. p. 3538–47. Available from: http://dx.doi.org/10.1158/1078-0432.ccr-18-3190
119. Xiong Q, Zhou X, Liu Z, Lei C, Yang C, Yang M, et al. Multiparametric MRI-based radiomics analysis for prediction of breast cancers insensitive to neoadjuvant chemotherapy [Internet]. Clinical and Translational Oncology. 2019. Available from: http://dx.doi.org/10.1007/s12094-019-02109-8
120. Huynh BQ, Li H, Giger ML. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. J Med Imaging 2016; 3: 034501. doi: https://doi.org/10.1117/1.JMI.3. 3.034501
121. Li H, Giger ML, Huynh BQ, Antropova NO. Deep learning in breast cancer risk assessment: evaluation of convolutional neural networks on a clinical dataset of full-field digital mammograms. J Med Imaging (Bellingham). 2017 Oct;4(4):041304.
122. Wong AJ, Kanwar A, Mohamed AS, Fuller CD. Radiomics in head and neck cancer: from exploration to application. Transl Cancer Res. 2016 Aug;5(4):371–82.
123. Toivonen J, Montoya Perez I, Movahedi P, Merisaari H, Pesola M, Taimen P, et al. Radiomics and machine learning of multisequence multiparametric prostate MRI: Towards improved non-invasive prostate cancer characterization. PLoS One. 2019 Jul 8;14(7):e0217702.
124. Kocak B, Ates E, Durmaz ES, Ulusan MB, Kilickesmez O. Influence of segmentation margin on machine learning–based high-dimensional quantitative CT texture analysis: a reproducibility

study on renal clear cell carcinomas [Internet]. Vol. 29, European Radiology. 2019. p. 4765–75. Available from: http://dx.doi.org/10.1007/s00330-019-6003-8

125. Saini A, Breen I, Pershad Y, Naidu S, Knuttinen MG, Alzubaidi S, et al. Radiogenomics and Radiomics in Liver Cancers. Diagnostics (Basel) [Internet]. 2018 Dec 27;9(1). Available from: http://dx.doi.org/10.3390/diagnostics9010004
126. Badic B, Hatt M, Durand S, Jossic-Corcos CL, Simon B, Visvikis D, et al. Radiogenomics-based cancer prognosis in colorectal cancer. Sci Rep. 2019 Jul 5;9(1):9743.
127. Sollini M, Antunovic L, Chiti A, Kirienko M. Towards clinical application of image mining: a systematic review on artificial intelligence and radiomics. Eur J Nucl Med Mol Imaging [Internet]. 2019 Jun 18; Available from: http://dx.doi.org/10.1007/s00259-019-04372-x
128. Tian Y, Liu Z, Tang Z, Li M, Lou X, Dong E, et al. Radiomics Analysis of DTI Data to Assess Vision Outcome After Intravenous Methylprednisolone Therapy in Neuromyelitis Optic Neuritis. J Magn Reson Imaging. 2019 May;49(5):1365–73.
129. Bianchi J, Gonçalves JR, Ruellas AC de O, Vimort J-B, Yatabe M, Paniagua B, et al. Software comparison to analyze bone radiomics from high resolution CBCT scans of mandibular condyles. Dentomaxillofac Radiol. 2019 Sep;48(6):20190049.
130. Panth KM, Leijenaar RTH, Carvalho S, Lieuwes NG, Yaromina A, Dubois L, et al. Is there a causal relationship between genetic changes and radiomics-based image features? An in vivo preclinical experiment with doxycycline inducible GADD34 tumor cells [Internet]. Vol. 116, Radiotherapy and Oncology. 2015. p. 462–6. Available from: http://dx.doi.org/10.1016/j.radonc.2015.06.013
131. Leijenaar RT, Bogowicz M, Jochems A, Hoebers FJ, Wesseling FW, Huang SH, et al. Development and validation of a radiomic signature to predict HPV (p16) status from standard CT imaging: a multicenter study. Br J Radiol. 2018 Jun;91(1086):20170498.
132. Ou D, Blanchard P, Rosellini S, Levy A, Nguyen F, Leijenaar RTH, et al. Predictive and prognostic value of CT based radiomics signature in locally advanced head and neck cancers patients treated with concurrent chemoradiotherapy or bioradiotherapy and its added value to Human Papillomavirus status. Oral Oncol. 2017 Aug;71:150–5.
133. Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: Evolving Considerations for PET Response Criteria in Solid Tumors [Internet]. Vol. 50, Journal of Nuclear Medicine. 2009. p. 122S – 150S. Available from: http://dx.doi.org/10.2967/jnumed.108.057307
134. Fukukita H, Senda M, Terauchi T, Suzuki K, Daisaki H, Matsumoto K, et al. Japanese guideline for the oncology FDG-PET/CT data acquisition protocol: synopsis of Version 1.0 [Internet]. Vol. 24, Annals of Nuclear Medicine. 2010. p. 325–34. Available from: http://dx.doi.org/10.1007/s12149-010-0377-7
135. Boellaard R, O'Doherty MJ, Weber WA, Mottaghy FM, Lonsdale MN, Stroobants SG, et al. FDG PET and PET/CT: EANM procedure guidelines for tumour PET imaging: version 1.0 [Internet]. Vol. 37, European Journal of Nuclear Medicine and Molecular Imaging. 2010. p. 181–200. Available from: http://dx.doi.org/10.1007/s00259-009-1297-4

136. Kinahan PE, Perlman ES, Sunderland JJ, Subramaniam R, Wollenweber SD, Turkington TG, et al. The QIBA Profile for FDG PET/CT as an Imaging Biomarker Measuring Response to Cancer Therapy. Radiology. 2020 Jan 7;191882.
137. Mankoff DA. Quantitative imaging as cancer biomarker [Internet]. Medical Imaging 2015: Physics of Medical Imaging. 2015. Available from: http://dx.doi.org/10.1117/12.2085907
138. Frid-Adar M, Klang E, Amitai M, Goldberger J, Greenspan H. Synthetic data augmentation using GAN for improved liver lesion classification [Internet]. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). 2018. Available from: http://dx.doi.org/10.1109/isbi.2018.8363576
139. Mahmood F, Chen R, Durr NJ. Unsupervised Reverse Domain Adaptation for Synthetic Medical Images via Adversarial Training. IEEE Trans Med Imaging. 2018 Dec;37(12):2572–81.
140. Wang Q, Milletari F, Nguyen HV, Albarqouni S, Jorge Cardoso M, Rieke N, et al. Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data: First MICCAI Workshop, DART 2019, and First International Workshop, MIL3ID 2019, Shenzhen, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13 and 17, 2019, Proceedings. Springer Nature; 2019. p. 254.
141. Jochems A, Deist TM, El Naqa I, Kessler M, Mayo C, Reeves J, et al. Developing and Validating a Survival Prediction Model for NSCLC Patients Through Distributed Learning Across 3 Countries. Int J Radiat Oncol Biol Phys. 2017 Oct 1;99(2):344–52.
142. Deist TM, Jochems A, van Soest J, Nalbantov G, Oberije C, Walsh S, et al. Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. Clin Transl Radiat Oncol. 2017 Jun;4:24–31.
143. Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis IC, Shi W. Federated learning of predictive models from federated Electronic Health Records. Int J Med Inform. 2018 Apr;112:59–67.
144. Sheller MJ, Anthony Reina G, Edwards B, Martin J, Bakas S. Multi-institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation [Internet]. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. 2019. p. 92–104. Available from: http://dx.doi.org/10.1007/978-3-030-11723-8_9
145. Lugan S, Desbordes P, Tormo LXR, Legay A, Macq B. Secure Architectures Implementing Trusted Coalitions for Blockchained Distributed Learning (TCLearn) [Internet]. 2019 [cited 2019 Oct 17]. Available from: http://arxiv.org/abs/1906.07690
146. Holzinger A. Explainable AI (ex-AI) [Internet]. Vol. 41, Informatik-Spektrum. 2018. p. 138–43. Available from: http://dx.doi.org/10.1007/s00287-018-1102-5
147. Khedkar S, Subramanian V, Shinde G, Gandhi P. Explainable AI in Healthcare [Internet]. SSRN Electronic Journal. Available from: http://dx.doi.org/10.2139/ssrn.3367686
148. Joyner MJ, Paneth N. Promises, promises, and precision medicine [Internet]. Vol. 129, Journal of Clinical Investigation. 2019. p. 946–8. Available from: http://dx.doi.org/10.1172/jci126119
149. Saracci R. Epidemiology in wonderland: Big Data and precision medicine. Eur J Epidemiol. 2018 Mar;33(3):245–57.

Acknowledgments

Authors acknowledge financial support from ERC advanced grant (ERC-ADG-2015, n° 694812 - Hypoximmuno), ERC-2018-PoC (n° 81320 - CL-IO). This research is also supported by the Dutch Technology Foundation STW (grant n° P14-19 Radiomics STRaTegy), which is the applied science division of NWO, and the Technology Programme of the Ministry of Economic Affairs. Authors also acknowledge financial support from SME Phase 2 (RAIL - n°673780), EUROSTARS (DART, DECIDE), the European Program H2020-2015-17 (BD2Decide - PHC30-689715, ImmunoSABR - n° 733008, PREDICT - ITN - n° 766276), TRANSCAN Joint Transnational Call 2016 (JTC2016 "CLEARLY"- n° UM 2017-8295), Interreg V-A Euregio Meuse-Rhine ("Euradiomics"). This work was supported by the Dutch Cancer Society (KWF Kankerbestrijding), Project number 12085/2018-2

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$minimum = \min(\mathbf{X})$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2$$

$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_s} \mathbf{P}(i,j)\right)^2}{N_p}$$

$$correlation = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) ij - \mu_x \mu_y}{\sigma_x(i) \sigma_y(j)}$$

$$MAD = \frac{1}{N_p} \sum_{i=1}^{N_p} |\mathbf{X}(i) - \bar{X}|$$

$$minimum = \min(\mathbf{X})$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_s} \mathbf{P}\right.}{N_z}$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2$$

Part II

$$variance = \frac{1}{N_p}\sum_{i=1}^{N_p}(\mathbf{X}(i)-\bar{X})^2$$

$$total\ energy = V_{voxel}\sum_{i=1}^{N_p}(\mathbf{X}(i)+c)^2$$

$$SDE = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}\frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$minimum = \min(\mathbf{X})$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2$$

$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_s} \mathbf{P}(i,j)\right)^2}{N_p}$$

$$correlation = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) ij - \mu_x \mu_y}{\sigma_x(i) \sigma_y(j)}$$

$$MAD = \frac{1}{N_p} \sum_{i=1}^{N_p} |\mathbf{X}(i) - \bar{X}|$$

$$minimum = \min(\mathbf{X})$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_s} \mathbf{P}\right.}{N_z}$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2$$

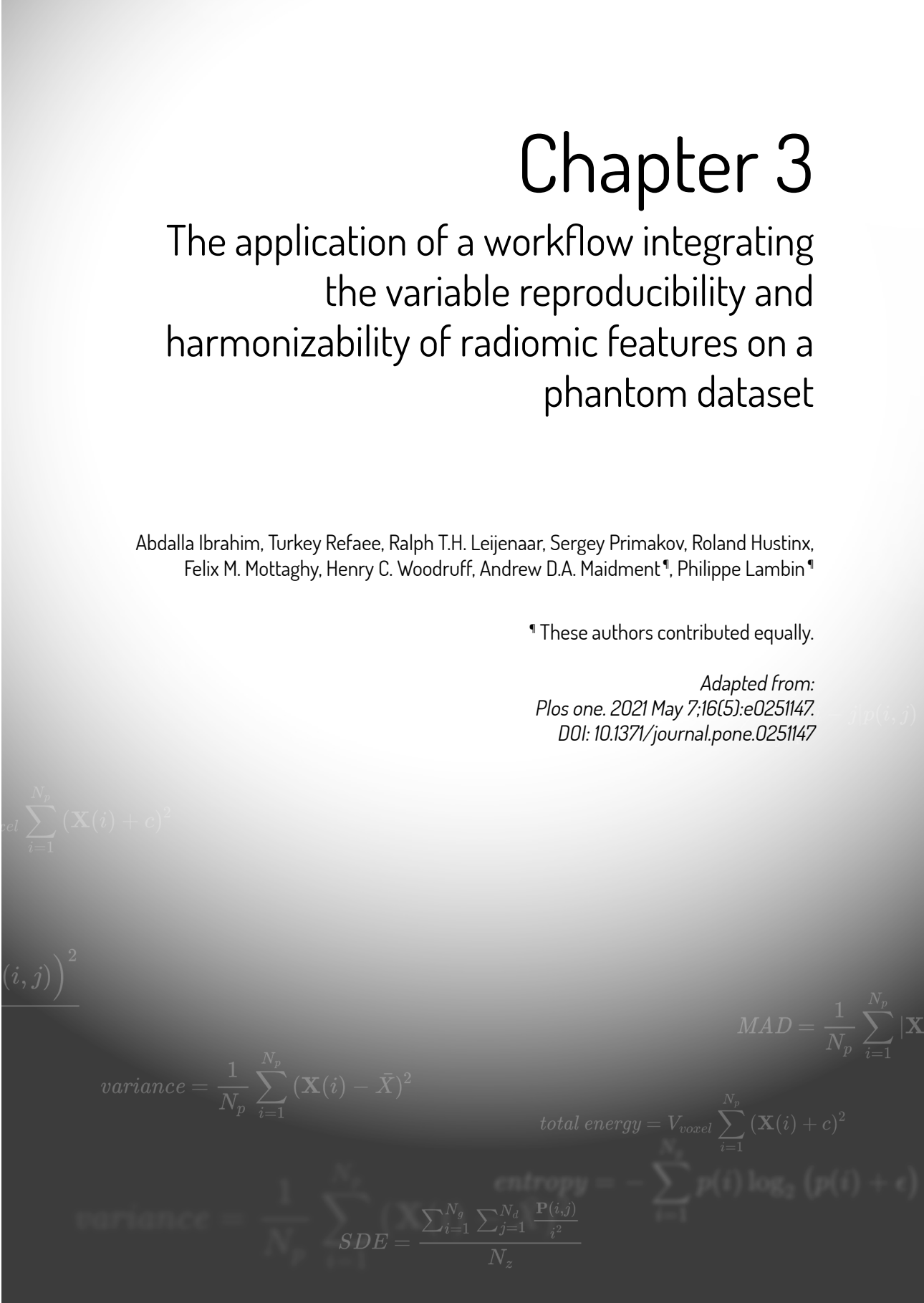
Chapter 3

The application of a workflow integrating the variable reproducibility and harmonizability of radiomic features on a phantom dataset

Abdalla Ibrahim, Turkey Refaee, Ralph T.H. Leijenaar, Sergey Primakov, Roland Hustinx, Felix M. Mottaghy, Henry C. Woodruff, Andrew D.A. Maidment¶, Philippe Lambin¶

¶ These authors contributed equally.

Adapted from:
Plos one. 2021 May 7;16(5):e0251147.
DOI: 10.1371/journal.pone.0251147



Abstract

Radiomics – the high throughput extraction of quantitative features from medical images and their correlation with clinical and biological endpoints- is the subject of active and extensive research. Although the field shows promise, the generalizability of radiomic signatures is affected significantly by differences in scan acquisition and reconstruction settings. Previous studies reported on the sensitivity of radiomic features (RFs) to test-retest variability, inter-observer segmentation variability, and intra-scanner variability. A framework involving robust radiomics analysis and the application of a post-reconstruction feature harmonization method using ComBat was recently proposed to address these challenges. In this study, we investigated the reproducibility of RFs across different scanners and scanning parameters using this framework. We analysed thirteen scans of a ten-layer phantom that were acquired differently. Each layer was subdivided into sixteen regions of interest (ROIs), and the scans were compared in a pairwise manner, resulting in seventy-eight different scenarios. Ninety-one RFs were extracted from each ROI. As hypothesized, we demonstrate that the reproducibility of a given RF is not a constant but is dependent on the heterogeneity found in the data under analysis. The number (%) of reproducible RFs varied across the pairwise scenarios investigated, having a wide range between 8 (8.8%) and 78 (85.7%) RFs. Furthermore, in contrast to what has been previously reported, and as hypothesized in the robust radiomics analysis framework, our results demonstrate that ComBat cannot be applied to all RFs but rather on a percentage of those – the "ComBatable" RFs – which differed depending on the data being harmonized. . The number (%) of reproducible RFs following ComBat harmonization varied across the pairwise scenarios investigated, ranging from 14 (15.4%) to 80 (87.9%) RFs, and was found to depend on the heterogeneity in the data. We conclude that the standardization of image acquisition protocols remains the cornerstone for improving the reproducibility of RFs, and the generalizability of the signatures developed. Our proposed approach helps identify the reproducible RFs across different datasets.

Keywords: Radiomics, Harmonization, Feature stability, Feature reproducibility

Introduction

With the advancement and involvement of artificial intelligence in performing high-level tasks, its application has been extensively researched in the field of medical imaging analysis [1]. Radiomics – the high throughput extraction of quantitative features from medical imaging to find correlations with biological or clinical outcomes [2-4] – is currently one of the most commonly used quantitative imaging analysis methods in medical imaging.

A major area of research in the field of radiomics is the selection of robust and informative image features to be used as input for machine learning models [5]. Evidence suggests that radiomic features (RFs) are sensitive to differences in several factors, including make and type of imaging scanner, reconstruction settings, and protocols used to acquire the images [6, 7]. Studies on the reproducibility of RFs across test-retest [8, 9]; or across scans of a phantom made on the same scanner using different exposure levels, while fixing other parameters [10]; or across scans of a phantom using different acquisition and reconstruction parameters [11] highlighted the high sensitivity of RFs to variations within datasets.

The above-mentioned studies focused on the reproducibility of RFs in limited settings, such as test-retest, inter-observer variability, and intra-scanner variability. As these studies reported significant differences in groups of RFs, it is only intuitive that adding more variation to image acquisition and reconstruction will further dampen the reproducibility of RFs. These findings indicate that ignoring data heterogeneity will influence the performance and generalizability of the models developed, especially in studies where training and validation sets are independent. Therefore, a global initiative – the Image Biomarkers Standardization Initiative (IBSI) – has been initiated in an effort to standardize the extraction of image biomarkers (RFs) from medical images [12]. The IBSI aims to standardize both the computation of RFs and the image processing steps required before RF extraction. However, little attention has been paid in the bulk of literature to date to the heterogeneity in image acquisition and reconstruction when performing radiomics analysis. As the goal of radiomics research is to employ quantitative imaging features as clinical biomarker, the issue of accurate measurement and reproducibility must be addressed [13]. Biomarkers are defined as “the objective indications of medical state observed from outside the patient – which can be measured reproducibly”. Therefore, reproducible measurement is a corner stone in choosing a biomarker. In essence, RFs that cannot be reproduced cannot be compared or selected as biomarkers.

Combining Batches (ComBat) harmonization is a method that was introduced for removing the effects of machinery and protocols used to extract gene expression data, in order to make gene expression data acquired at different centres comparable [14]. ComBat is a method that performs location and scale adjustments of the values presented to remove

the discrepancies in RF values introduced by technical differences in the images. These sources of variation are further referred to as batch effects. ComBat was subsequently adopted in radiomics analysis, and some studies reported that ComBat outperforms other harmonization methods (e.g, histogram-matching, voxel size normalization, and singular value decomposition) in radiomics analyses [15, 16]. Several radiomics studies have reported on the successful application of ComBat in removing the differences in RFs introduced by different vendors and acquisition protocols [17-21]. These studies investigated the differences in radiomic RF distributions across different batches following the application of ComBat harmonization. In contrast to gene expression arrays, RFs have different definitions, and the batch effect might vary for each RF. Using phantom data allows one to study the variations in a given RF extracted from scans acquired with different scanners/ reconstruction settings and to attribute these variations to the changes in acquisition and reconstruction, which in theory ComBat harmonization is designed to mitigate. However, we are not aware of any study that has performed a systematic evaluation of the performance of ComBat harmonization across variations between imaging parameters, which is the one of the objectives of this study.

Ibrahim et al. (2020) have proposed a new radiomics workflow (Fig 1) that tries to address the challenges current radiomics analyses face. The framework was proposed based on mathematical considerations of the complexity of medical imaging, and RFs' mathematical definitions. Our framework is based on the hypothesis that the reproducibility of a given RF is a not constant, but depends on the variations of image acquisition and reconstruction in the data under study. Furthermore, for ComBat to be applicable in radiomics, radiomic RF values for a given region of interest obtained after ComBat must be (nearly) identical, regardless of differences in acquisition and reconstruction.

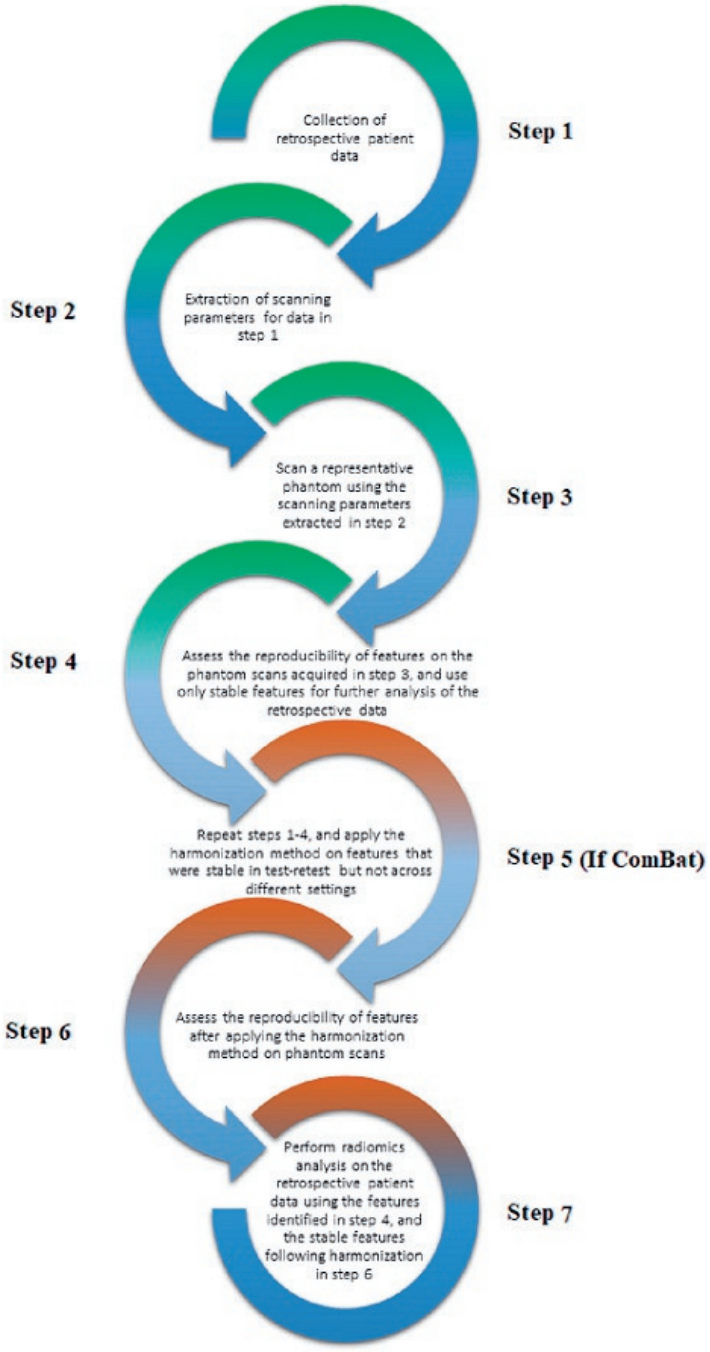


Figure 1. The proposed framework (reprinted with permission from [22]).

Our general objective is to set-up the requirements for selecting biomarkers from RFs, to ease their incorporation into clinical decision support systems. We hypothesize that variations in image acquisition and reconstruction will variably affect RFs reproducibility. Furthermore, the performance of ComBat on a given RF is dependent on those variations, i.e, a given RF can be successfully harmonized with ComBat with specific variations in the imaging parameters but not others. We investigate these hypotheses on CT scans using a ten-layer radiomics phantom, which was scanned with different acquisition and reconstruction parameters on various scanner models.

Methods

Phantom Data

The publicly available Credence Cartridge Radiomics (CCR) phantom data, found in The Cancer Imaging Archive (TCIA.org) [23, 24], was used. The CCR phantom is composed of 10 different layers that correspond to different texture patterns spanning a range of −900 to +700 Hounsfield units (HU). Each layer of the phantom was further subdivided into 16 distinct regions of interest (ROI) with cubic volume of 8 cm^3, resulting in a total of 2080 ROIs available for further analysis. The phantom was originally scanned using 17 different imaging protocols from four medical institutes using equipment from different vendors and a variety of acquisition and reconstruction parameters. Four of the scans lacked ROI definitions, thus to maintain consistency, these were not included. The remaining 13 scans are as follows: seven different scans acquired on GE scanners, five different scans acquired on Philips scanners, and one scan acquired on a Siemens scanner (Tables 1 and 2).

Table 2. CT reconstruction parameters*

Scan	Convolution Kernel	Filter Type	Slice thickness (mm)	Pixel spacing (mm)
CCR1-001	STANDARD	BODY FILTER	2.5	0.49
CCR1-002	STANDARD	BODY FILTER	2.5	0.70
CCR1-003	STANDARD	BODY FILTER	2.5	0.78
CCR1-004	STANDARD	BODY FILTER	2.5	0.98
CCR1-005	STANDARD	BODY FILTER	2.5	0.98
CCR1-006	STANDARD	BODY FILTER	2.5	0.98
CCR1-007	STANDARD	BODY FILTER	2.5	0.74
CCR1-008	B	B	3	0.98
CCR1-009	C	C	3	0.98
CCR1-010	B	B	3	1.04
CCR1-011	B	B	3	1.04
CCR1-012	B	B	3	0.98
CCR1-013	B31s	0	3	0.54

* Values are directly extracted from the publicly available imaging tags.

Table 1. CT acquisition parameters*

Scan	Vendor	Model	Scan Options	Effective mAs**	kVp
CCR1-001	GE	Discovery CT750 HD	HELICAL	81	120
CCR1-002	GE	Discovery CT750 HD	AXIAL	300	120
CCR1-003	GE	Discovery CT750 HD	HELICAL	122	120
CCR1-004	GE	Discovery ST	HELICAL	143	120
CCR1-005	GE	LightSpeed RT	HELICAL	1102	120
CCR1-006	GE	LightSpeed RT16	HELICAL	367	120
CCR1-007	GE	LightSpeed VCT	HELICAL	82	120
CCR1-008	Philips	Brilliance Big Bore	HELICAL	320	120
CCR1-009	Philips	Brilliance Big Bore	HELICAL	369	120
CCR1-010	Philips	Brilliance Big Bore	HELICAL	320	120
CCR1-011	Philips	Brilliance Big Bore	HELICAL	369	120
CCR1-012	Philips	Brilliance 64	HELICAL	372	120
CCR1-013	SIEMENS	Sensation Open	AXIAL	26-70	120

* Values are directly extracted from the publicly available imaging tags.

Radiomic features extraction

For each ROI, quantitative imaging features were calculated using the open source Pyradiomics (V 2.0.2). The software contains IBSI-compliant RFs, with deviations highlighted in the feature definitions. For the extraction step, no changes to the original slice thickness or pixel spacing of the scans were applied. To reduce noise and computational requirements, images were pre-processed by binning voxel greyscale values into bins with a fixed width of 25 HUs prior to extracting RFs. The extracted features included HU intensity features, shape features, and texture features describing the spatial distribution of voxel intensities using 5 texture matrices (i.e., grey-level co-occurrence (GLCM), grey-level run-length (GLRLM), grey-level size-zone (GLSZM), grey-level dependence (GLDM), and neighbourhood grey-tone difference matrix (NGTDM)). Detailed description of the features can be found online at https://pyradiomics.readthedocs.io/en/latest/features.html.

ComBat Harmonization

ComBat employs empirical Bayes methods to estimate the differences in feature values attributed to a batch effect. Empirical Bayes methods are able to estimate the prior distribution from a given dataset via statistical inference. In the context of radiomics, ComBat assumes that feature values can be approximated by the equation:

$$Y_{ij} = \alpha + \beta X_{ij} + \gamma_i + \delta_i \varepsilon_{ij} \quad (1)$$

where α is the average value for feature Y_{ij} for ROI j on scanner i; X is a design matrix of the covariates of interest; β is the vector of regression coefficients corresponding to each covariate; γ_i is the additive effect of scanner i on features, which is presupposed to follow a normal distribution; δ_i is the multiplicative scanner effect, which is presupposed to follow an

inverse gamma-distribution; and ε_{ij} is an error term, presupposed to be normally distributed with zero mean [17]. ComBat performs feature transformation based on the empirical Bayes prior estimates for γ and δ for each batch:

$$Y_{ij}^{ComBat} = \frac{(Y_{ij} - \hat{\alpha} - \hat{\beta} X_{ij} - \gamma_i^*)}{\delta_i^*} + \hat{\alpha} + \hat{\beta} X_{ij} \quad (2)$$

where $\hat{\alpha}$ and $\hat{\beta}$ are estimators of parameters α and β, respectively.$\gamma_i^*$and $\delta_i^*$are the empirical Bayes estimates of γ_i and δ_i, respectively [17].

Statistical analysis

To assess the agreement of a given RF for the same ROI scanned using different settings and scanners, the concordance correlation coefficient (CCC) was calculated using epiR (version 0.9-99) [25] on R [26] (version 3.5.1), using R studio (version 1.1.456) [27]. The CCC is used to evaluate the agreement between paired readings [28], and provides the measure of concordance as a value between 1 and -1, where 0 represents no concordance, 1 represents a perfect direct positive concordance, and -1 indicates a perfect inverse concordance. It further takes into account the rank and value of the RFs.

The analysis of the reproducibility before and after ComBat harmonization was performed in a pairwise manner, resulting in 78 different investigated scenarios. To assess differences in RF stability for differing data, the reproducibility of radiomics RFs across scans within a wide spectrum of scenarios was calculated. Data ranging from differences in a single acquisition or reconstruction parameter, to scans acquired using entirely different settings (See S1 table) were included. To identify reproducible radiomics, the CCC was calculated for all RFs for all ROIs across the 78 investigated scenarios. A cut-off of CCC>0.9, as found in the literature, suggests that a value < 0.9 indicates poor concordance [29].To identify the RFs that could be harmonized using ComBat, the pair-wise CCC was calculated following ComBat in each of the investigated 78 scenarios. We applied ComBat using R package “SVA” (version 3.30.1) [30]. As the RFs are calculated for the same ROI but for different scans, the agreement in RF value is expected to be high following ComBat harmonization. Thus, RFs that had a CCC<0.9 were considered to be not harmonizable with ComBat. The code used in this work is publicly available on https://github.com/AbdallaIbrahim/The-reproducibility-and-ComBatability-of-Radiomic-features.

Results

Table 3. The number (percentage) of concordant RFs before ComBat harmonization between pair wise combinations of scans with different acquisition and reconstruction.

	CCR1-001	CCR1-002	CCR1-003	CCR1-004	CCR1-005	CCR1-006	CCR1-007	CCR1-008	CCR1-009	CCR1-010	CCR1-011	CCR1-012
CCR1-002	38 (41.76%)											
CCR1-003	46 (50.55%)	59 (64.84%)										
CCR1-004	18 (19.78%)	34 (37.36%)	25 (27.47%)									
CCR1-005	13 (14.29%)	23 (25.27%)	17 (18.68%)	66 (72.53%)								
CCR1-006	16 (17.58%)	24 (26.37%)	18 (19.78%)	71 (78.02%)	69 (75.82%)							
CCR1-007	49 (53.85%)	65 (71.43%)	67 (73.63%)	21 (23.08%)	14 (15.38%)	14 (15.38%)						
CCR1-008	8 (8.79%)	12 (13.19%)	14 (15.38%)	41 (45.05%)	34 (37.36%)	47 (51.65%)	10 (10.99%)					
CCR1-009	9 (9.89%)	19 (20.88%)	13 (14.29%)	67 (73.63%)	65 (71.43%)	74 (81.32%)	11 (12.09%)	48 (52.75%)				
CCR1-010	8 (8.79%)	10 (10.99%)	13 (14.29%)	32 (35.16%)	21 (23.08%)	27 (29.67%)	11 (12.09%)	59 (64.84%)	34 (37.36%)			
CCR1-011	8 (8.79%)	11 (12.09%)	12 (13.19%)	45 (49.45%)	34 (37.36%)	42 (46.15%)	11 (12.09%)	57 (62.64%)	52 (57.14%)	78 (85.71%)		
CCR1-012	8 (8.79%)	13 (14.29%)	12 (13.19%)	21 (23.08%)	16 (17.58%)	22 (24.18%)	10 (10.99%)	61 (67.03%)	36 (39.56%)	71 (78.02%)	69 (75.82%)	
CCR1-013	51 (56.04%)	44 (48.35%)	47 (51.65%)	41 (45.05%)	34 (37.36%)	32 (35.16%)	48 (52.75%)	12 (13.19%)	23 (25.27%)	10 (10.99%)	9 (9.89%)	10 (10.99%)

Reproducible Radiomic features

For each ROI, a total of 91 RFs were extracted. The number (percentage) of reproducible RFs in each pair-wise comparison ranged from 9 (8.8%) to 78 (85.7%) RFs, depending on the variations in acquisition and reconstruction of the scans (table 3). The highest concordance in feature values (85.7%) was observed between the two Philips scans (CCR1-010 and CCR1-011) that were acquired using the same scanner model, and the same acquisition and reconstruction parameters except for the effective mAs, which differed by just 15% (tables 1 and 2).

The more profound the variations in scan acquisition parameters, the smaller the concordance of the extracted RFs (tables 1-3, S1).

As stated, in the best scenario (CCR1-010 and CCR1-011), 78 (85.7%) RFs were found to be reproducible, while 13 (14.3%) RFs were found not to be reproducible. Some RFs (n=8) were found to be concordant across all pairs. These RFs were histogram-based RFs that take into account the value of a single pixel/voxel, without looking at the relationship between neighbouring pixels/voxels. These RFs are (i) original first order 10Percentile; (ii) original

first order 90Percentile; (iii) original first order Maximum; (iv) original first order Mean (v) original first order Median; (vi) original first order Minimum; (vii) original first order Root Mean Squared; and (viii) original first order Total Energy. Nevertheless, the remainder (majority) of the RFs (including 10 histogram-based RFs) were not found to be reproducible across all pairs.

Looking at tables (1-3, S1), we can consider subgroups of scans. Scans CCR1-001-007 were all acquired using the same imaging vendor (GE), but different scanner models and scanning parameters. The highest number of concordant RFs in this group was found between CCR1-004 and CCR1-006 (71 RFs), which were acquired on two different scanner models, but were scanned with identical scanning parameters except for the mAs. The lowest number of concordant RFs in this group was found between scans CCR1-001 and CCR1-005 (13 RFs), which were acquired on two different scanner models, with the same scanning parameters except for the pixel spacing and mAs. Scans CCR1-007 to CCR1-012 were all acquired using one of two Philips imaging vendors. The highest number of concordant RFs is documented above. The lowest number of concordant RFs was found between CCR1-009 and CCR-010 (34 RFs), which differed in terms of the mAs, convolution kernel, filter type and pixel spacing. Looking at the group of scans that were reconstructed to the same pixel spacing (CCR1-004 to CCR1-006, CCR1-008, CCR1-009, and CCR-012), the highest number of concordant RFs was observed between CCR1-006 and CCR1-009 (74 RFs), which were acquired using two different imaging vendors, but using similar acquisition and reconstruction parameters except for the slice thickness, and kernel. The lowest number of concordant RFs was found between CCR1-005 and CCR1-012 (16 RFs), which were acquired using different imaging vendors, and different acquisition and reconstruction parameters except for the kVp. Finally, comparing scans acquired with different vendors resulted in a lower number of concordant RFs compared to scans acquired with the scanners from the same imaging vendor, except for the scenario when the majority of acquisition and reconstruction parameters were mostly identical (CCR1-006 vs CCR1-009).

ComBat harmonization

Table 4. The number (percentage) of concordant RFs after ComBat harmonization between pair wise combinations of scans with different acquisition and reconstruction.

	CCR1-001	CCR1-002	CCR1-003	CCR1-004	CCR1-005	CCR1-006	CCR1-007	CCR1-008	CCR1-009	CCR1-010	CCR1-011	CCR1-012
CCR1-002	63 (69.23%)											
CCR1-003	69 (75.82%)	75 (82.42%)										
CCR1-004	48 (52.75%)	72 (79.12%)	57 (62.64%)									
CCR1-005	43 (47.25%)	60 (65.93%)	54 (59.34%)	72 (79.12%)								
CCR1-006	50 (54.95%)	63 (69.23%)	59 (64.84%)	76 (83.52%)	72 (79.12%)							
CCR1-007	70 (76.92%)	69 (75.82%)	74 (81.32%)	56 (61.54%)	49 (53.85%)	57 (62.64%)						
CCR1-008	27 (29.67%)	36 (39.56%)	36 (39.56%)	61 (67.03%)	54 (59.34%)	56 (61.54%)	28 (30.77%)					
CCR1-009	40 (43.96%)	57 (62.64%)	53 (58.24%)	76 (83.52%)	74 (81.32%)	81 (89.01%)	52 (57.14%)	57 (62.64%)				
CCR1-010	18 (19.78%)	22 (24.18%)	19 (20.88%)	54 (59.34%)	48 (52.75%)	48 (52.75%)	17 (18.68%)	68 (74.73%)	53 (58.24%)			
CCR1-011	14 (15.38%)	23 (25.27%)	25 (27.47%)	67 (73.63%)	59 (64.84%)	59 (64.84%)	16 (17.58%)	65 (71.43%)	67 (73.63%)	80 (87.91%)		
CCR1-012	16 (17.58%)	29 (31.87%)	28 (30.77%)	56 (61.54%)	48 (52.75%)	49 (53.85%)	16 (17.58%)	70 (76.92%)	53 (58.24%)	72 (79.12%)	74 (81.32%)	
CCR1-013	65 (71.43%)	75 (82.42%)	69 (75.82%)	65 (71.43%)	55 (60.44%)	59 (64.84%)	67 (73.63%)	35 (38.46%)	58 (63.74%)	35 (38.46%)	36 (39.56%)	34 (37.36%)

As previously shown in the literature, we used each scan as a different batch in the ComBat equation. ComBat was applied pairwise (78 different pairs) and the concordance between RFs was measured for each pair (table 4). The percentage of RFs that became concordant following ComBat application ranged from 1.4% (71 concordant RFs increased to 72) to 344% (9 concordant RFs increased to 40).

The highest number of concordant RFs following ComBat application was 80 (87.9%) RFs. In this scenario, a single acquisition parameter differed between the two scans (Philips, CCR1-010 and CCR1-011). ComBat application improved the concordance of only two RFs (80 RFs after ComBat compared to 78 RFs before), and failed to improve the concordance of the remaining 11 RFs. On the other hand, in cases where the differences in acquisition and reconstruction parameters differed more (e.g., CCR1-001 (GE) vs CCR1-007 (Philips)), the application of ComBat improved the concordance of 31 RFs, resulting in a total of 40 concordant RFs (~44% of the total number of RFs), more than 3 times the number of concordant RFs before harmonization. Furthermore, the successful application of ComBat on RFs depended on the variations in the batches defined. Only two RFs were found to be concordant in all pairwise scenarios following ComBat harmonization: (i) original first order

Energy; and (ii) original gldm Small Dependence High Gray Level Emphasis; in addition to the 8 RFs mentioned above.

Discussion

In this work, for our first objective to investigate RFs reproducibility, we show that the majority of RFs are affected to different amounts depending upon the variations in acquisition and reconstruction parameters. We also show that the reproducibility of a given RF is not constant, but rather it is dependent on the variations in the data under study, as seen in table 3. We identified a number of RFs that were robust to the variations in scan acquisition in the dataset we analysed. These RFs could be used without any post–processing harmonization. While the same dataset has been analysed for similar purposes previously [11, 21], we analysed the data differently, and report different results than those studies. Our results show a substantial intra-scanner variability, and even greater inter-scanner variability, which is in line with other previous findings [10, 31, 32]. Only eight RFs (~9%) of the extracted RFs showed insensitivity to the differences in acquisition shown in tables 1 and 2, and could be directly used to build radiomic signatures. The rest of the RFs (91%) could not be used without addressing the acquisition differences. Our sub-groups analysis showed that changes in pixel spacing and convolution kernel have more profound effects on the reproducibility of RFs, compared to variations limited solely to the effective mAs, scanner model or imaging vendor used. While the percentages reported are representative of the reproducibility of RFs in the data analysed, it highlights the sensitive nature of RFs, and helps set guidelines to preselect meaningful and reproducible RFs. We deduce that the use of RFs extracted from scans acquired with different hardware and parameters, without addressing the issue of reproducibility and harmonization, can lead to spurious results as the vast majority of RFs are sensitive to even minor variations in image acquisition and reconstruction. Therefore, models developed using RFs with large unexplained variances will most likely not be generalizable.

As our second aim, we investigated the applicability of ComBat harmonization to removing differences in RF values attributed to batch effects. Studies [11, 21] have reported on the reproducibility of RFs on the same or a similar dataset to the one we analysed. However, our findings and conclusions vary significantly from theirs. In contrast to previous studies, we are the first to report that the reproducibility of RFs is dependent on the variations in the data under analysis. Previous studies referred to RFs as generally reproducible or non-reproducible. Our analysis shows that a given RF can be reproducible in some scenarios and not in the others, depending on the variations in acquisition and reconstruction parameters. Moreover, ComBat was mathematically defined to remove one (technical) batch effect at a time while considering all the biologic covariates at the same time. However, as our results show (tables 3 and 4), the variations in acquisition and reconstruction parameters within

one scanner, at least in some instances, have a stronger impact on the reproducibility of RFs than the variations between two scanners. As such, grouping the scans by the scanner type is not generally the way to define "batches" in the ComBat equation [14]. In contrast to what is reported in the literature, our analysis shows ComBat did not perform uniformly on most of the RFs when there were variations in the batches being harmonized. In contrast to those studies, we employed the concordance correlation coefficient (CCC) to assess the reproducibility of RFs, since the aim of harmonization is to improve the reproducibility of data. We did not use the increment of model performance as a measure for the success of harmonization for several reasons. First, the aim of harmonization is to improve the reproducibility of RFs, and ultimately the generalizability of the developed signatures, and not their model performance [33]. Second, by assuming that an increment in the model performance following harmonization is an indication that the harmonization is successful carries with it the assumption that radiomic models decode the information under analysis; this is against the essence of the study, which is to investigate whether radiomics has that potential or not. However, by using the CCC, we ensure that the results generated are based on reproducible RFs, and are therefore generalizable, regardless of the change in model performance. Furthermore, the aim of ComBat harmonization is only to remove the variance in RF values attributed to the batch effects, while maintaining the biologic information. As such, using ComBat to correct batch effects directly on patient data without providing the correct biological covariates that actually do have an effect on RF values will lead to loss of biological signals. This is because ComBat tries to harmonize the distribution of the RF across different batches, and without providing the correct biological covariates that have effects on RF values, ComBat assumes that the variations in RF value are only attributed to the defined batch, and thus would not perform uniformly as shown in table 3. In clinical settings, this is by default spurious, as the differences in RF values are attributed to both the machine and the biology/physiology. As the aim of radiomics studies is to investigate the biological correlations of RFs, we are unable to actually provide a list of biologic covariates that influence the values. In addition, each time an observation is added to the data being harmonized, ComBat has to be re-performed, and models have to be refitted, as the estimated batch effects will change each time. Therefore, the harmonization of patient RFs should follow the process of estimating fixed batch effects on phantom data, then applying the location/scale shift estimated from the phantom data on patient data, as previously described by Ibrahim et al [22].

The pairwise approach we used shows how the variations in scan acquisition and reconstruction parameters affect the reproducibility of RFs. Therefore, aside from probably a few RFs, the reproducibility of the majority of the RFs cannot be guessed in untested scenarios. The workflow (figure 1) addresses this problem by introducing the assessment of RF reproducibility on representative phantom data. This workflow differs from existing radiomics workflows by the addition of an intermediary RF pre-selection step between RF

extraction and RF selection by one of two approaches: (i) only extracting the reproducible RFs for analysis; (ii) extracting and harmonizing the 'ComBatable' RFs before RF selection and model building. The application of ComBat and the definition of what constitutes a 'batch' should be performed based on the data being analysed, as could be deduced from tables 3 and 4. For example, RFs extracted from scans acquired with different scanner models, but similar settings were found to be more concordant than RFs extracted with the same scanner model but with profound differences in acquisition and reconstruction parameters. Our proposed radiomics analysis workflow would ensure that the RFs being analysed are not affected by scan acquisition differences, and henceforth, signatures built would be more robust and generalizable. The first part of the model (steps 1-4), where only reproducible RFs are extracted and further analysed, might significantly limit the number of RFs used for further modelling. However, using the whole framework may significantly increase the number of RFs that can be used, depending on the data under study.

While the data used for this analysis are not representative of diagnostic clinical protocols and do not provide all technical details needed for proper analysis, our aim was to show that changes in scan acquisition and reconstruction parameters differently affect the majority of RFs. The variations in the reproducibility of RFs – as well as ComBat applicability – due to the heterogeneity in acquisition and reconstruction highlight the necessity of the standardization of image acquisition and reconstruction across centres. RFs have already been reported to be sensitive to test-retest [8, 34], which is the acquisition of two separate scans using the same parameters, as well as to the variations in the parameters within the same scanner [10]. Adding the variable sensitivity of RFs to different acquisition and reconstruction parameters significantly lowers the number of RFs that could be used for the analysis of heterogeneous data. As there is currently a pressing desire to analyse big data, a sound methodology is needed to address the heterogeneity introduced by machinery in retrospective data. Nevertheless, we strongly recommend the start of imaging protocol standardization across centres to facilitate future quantitative imaging analysis.

Recently, there has been an attempt to modify ComBat methodology in radiomics analysis [35]. The authors added a modification to ComBat (B-ComBat), which adds Bootstrapping and Monte Carlo to the original ComBat. The other functionality of ComBat the authors investigated was to use one of the batches as a reference (M-ComBat). The authors compared the performance of the four versions of ComBat by comparing the performance of radiomic models developed after the use of each method. The authors reported that all the methods are equally effective [35]. Therefore, we anticipate that the modified ComBat functions will have the same limitations of the original ComBat we discussed above.

Another method to harmonize RFs that is currently gaining momentum is deep learning based harmonization. A recent study developed deep learning algorithms, which were

reported to improve the reproducibility of RFs across variations in scanner type, acquisition protocols and reconstruction algorithms [36]. A more recent study [37] applied a similar approach to reduce the sensitivity of RFs to scanner types. The authors reported a significant improvement in the performance of radiomic models following harmonization. These studies highlight the potential efficacy of deep learning based harmonization methods.

One limitation of our study is in considering each scan as a separate batch effect (due to lack of data) while differences between pair batches are not similar (different numbers of varying parameters), which may have affected the performance of ComBat. Acquisition and reconstruction settings include a set of different parameters, which can singularly or collectively result in differences in RFs values. Another limitation is the lack of scans generated by other commonly used scanners and protocols in the clinics; and the lack of scans with the same settings acquired using different scanners, as the data currently available is limited to the changes introduced in the imaging parameters on the available scanners. While we did not investigate the added value of this approach on a clinical dataset, our focus in this study was in designing a framework to assess the reproducibility and 'ComBatability' of RFs. However, it is fair to assume that if RFs are not reproducible on phantom data, they would be equally, or possibly even more, unstable on patient datasets. For example, clinical data will be acquired at a variety of mAs values across a population of patients. Lastly, while Combat has been reported to outperform other harmonization methods in terms of apparent model performance, the systemic evaluation of the effects of these methods on the reproducibility of RFs, and the comparison with the effects of ComBat harmonization will be the aim of future studies, in addition to addressing the above mentioned limitations.

Conclusion

In conclusion, we demonstrate that the reproducibility of RFs is not a constant, but changes with variations in the data acquisition and reconstruction parameters. Moreover, ComBat cannot be successfully applied on all RFs, and its successful application on a given RF is dependent on the heterogeneity of the dataset. We conclude that ComBat harmonization should not be blindly performed on patient data, but following the estimation of adjustment parameters on a phantom dataset. We anticipate that radiomics studies will benefit from our proposed harmonization workflow, as it allows comparison of a greater number of RFs, and enhances the generalizability of radiomic models. Yet, standardization of imaging protocols remains the cornerstone for improving the generalizability of prospective quantitative image studies. We recommend the standardization of scan acquisition across centres, especially in prospective clinical trials that include medical imaging; and/or the development of a specific imaging protocols for scans acquired to be used for quantitative imaging analysis.

References

1. Walsh S, de Jong EE, van Timmeren JE, Ibrahim A, Compter I, Peerlings J, et al. Decision support systems in oncology. JCO clinical cancer informatics. 2019;3:1-9.
2. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature communications. 2014;5:4006.
3. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, Van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. European journal of cancer. 2012;48(4):441-6.
4. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology. 2015;278(2):563-77.
5. Reiazi R, Abbas E, Faima P, Kwan JY, Rezaie A, Bratman SV, et al. The Impact of the Variation of Imaging Factors on the Robustness of Computed Tomography Radiomic Features: A Review. medRxiv. 2020.
6. van Timmeren JE, Carvalho S, Leijenaar RT, Troost EG, van Elmpt W, de Ruysscher D, et al. Challenges and caveats of a multi-center retrospective radiomics study: an example of early treatment response assessment for NSCLC patients using FDG-PET/CT radiomics. PloS one. 2019;14(6):e0217536.
7. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and reproducibility of radiomic features: a systematic review. International Journal of Radiation Oncology* Biology* Physics. 2018;102(4):1143-58.
8. van Timmeren JE, Leijenaar RT, van Elmpt W, Wang J, Zhang Z, Dekker A, et al. Test–retest data for radiomics feature stability analysis: Generalizable or study-specific? Tomography. 2016;2(4):361.
9. Prayer F, Hofmanninger J, Weber M, Kifjak D, Willenpart A, Pan J, et al. Variability of computed tomography radiomics features of fibrosing interstitial lung disease: A test-retest study. Methods. 2020.
10. Zhovannik I, Bussink J, Traverso A, Shi Z, Kalendralis P, Wee L, et al. Learning from scanners: Bias reduction and feature correction in radiomics. Clinical and translational radiation oncology. 2019;19:33-8.
11. Mackin D, Fave X, Zhang L, Fried D, Yang J, Taylor B, et al. Measuring CT scanner variability of radiomics features. Investigative radiology. 2015;50(11):757.
12. Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative. arXiv preprint arXiv:161207003. 2016.
13. Lambin P, Leijenaar RT, Deist TM, Peerlings J, De Jong EE, Van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nature reviews Clinical oncology. 2017;14(12):749.
14. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118-27.

15. Ligero M, Jordi-Ollero O, Bernatowicz K, Garcia-Ruiz A, Delgado-Muñoz E, Leiva D, et al. Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis. European radiology. 2021;31(3):1460-70.
16. Foy JJ, Al-Hallaq HA, Grekoski V, Tran T, Guruvadoo K, Armato Iii SG, et al. Harmonization of radiomic feature variability resulting from differences in CT image acquisition and reconstruction: assessment in a cadaveric liver. Physics in Medicine & Biology. 2020;65(20):205008.
17. Orlhac F, Boughdad S, Philippe C, Stalla-Bourdillon H, Nioche C, Champion L, et al. A postreconstruction harmonization method for multicenter radiomic studies in PET. Journal of Nuclear Medicine. 2018;59(8):1321-8.
18. Fortin J-P, Parker D, Tunç B, Watanabe T, Elliott MA, Ruparel K, et al. Harmonization of multi-site diffusion tensor imaging data. Neuroimage. 2017;161:149-70.
19. Fortin J-P, Cullen N, Sheline YI, Taylor WD, Aselcioglu I, Cook PA, et al. Harmonization of cortical thickness measurements across scanners and sites. NeuroImage. 2018;167:104-20.
20. Orlhac F, Humbert O, Boughdad S, Lasserre M, Soussan M, Nioche C, et al. Validation of a harmonization method to correct for SUV and radiomic features variability in multi-center studies. Journal of Nuclear Medicine. 2018;59(supplement 1):288-.
21. Orlhac F, Frouin F, Nioche C, Ayache N, Buvat I. Validation of a method to compensate multicenter effects affecting CT Radiomics. Radiology. 2019:182023.
22. Ibrahim A, Primakov S, Beuque M, Woodruff H, Halilaj I, Wu G, et al. Radiomics for precision medicine: current challenges, future prospects, and the proposal of a new framework. Methods. 2020.
23. Mackin DF, Xenia; Zhang, Lifei; Fried, David; Yang, Jinzhong; Taylor, Brian; Rodriguez-Rivera, Edgardo; Dodge, Cristina; Jones, Aaron Kyle; and Court, Laurence. Data From Credence Cartridge Radiomics Phantom CT Scans. The Cancer Imaging Archive. 2017. doi: http://doi.org/10.7937/K9/TCIA.2017.zuzrml5b.
24. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. Journal of digital imaging. 2013;26(6):1045-57.
25. Stevenson M, Nunes T, Sanchez J, Thornton R, Reiczigel J, Robison-Cox J, et al. epiR: An R package for the analysis of epidemiological data. R package version 09-43. 2013.
26. Team RC. R: A language and environment for statistical computing. 2013.
27. Team R. RStudio: Integrated Development for R. Boston: RStudio Inc.; 2015. 2016.
28. Lawrence I, Lin K. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989:255-68.
29. McBride G. A proposal for strength-of-agreement criteria for Lin's concordance correlation coefficient. NIWA Client Report: HAM2005-062. 2005.
30. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28(6):882-3.

31. Kim H, Park CM, Lee M, Park SJ, Song YS, Lee JH, et al. Impact of reconstruction algorithms on CT radiomic features of pulmonary tumors: analysis of intra-and inter-reader variability and inter-reconstruction algorithm variability. PLoS One. 2016;11(10):e0164924.
32. Lu L, Ehmke RC, Schwartz LH, Zhao B. Assessing agreement between radiomic features computed for multiple CT imaging settings. PloS one. 2016;11(12):e0166550.
33. Vetter TR, Schober P. Agreement analysis: what he said, she said versus you said. Anesthesia & Analgesia. 2018;126(6):2123-8.
34. Leijenaar RT, Carvalho S, Velazquez ER, Van Elmpt WJ, Parmar C, Hoekstra OS, et al. Stability of FDG-PET radiomics features: an integrated analysis of test-retest and inter-observer variability. Acta oncologica. 2013;52(7):1391-7.
35. Da-ano R, Masson I, Lucia F, Doré M, Robin P, Alfieri J, et al. Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies. Scientific Reports. 2020;10(1):10248. doi: 10.1038/s41598-020-66110-w.
36. Andrearczyk V, Depeursinge A, Müller H. Neural network training for cross-protocol radiomic feature standardization in computed tomography. Journal of Medical Imaging. 2019;6(2):024008.
37. Marcadent S, Hofmeister J, Preti MG, Martin SP, Van De Ville D, Montet X. Generative Adversarial Networks Improve the Reproducibility and Discriminative Power of Radiomic Features. Radiology: Artificial Intelligence. 2020;2(3):e190035.

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$minimum = \min(\mathbf{X})$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2$$

$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_s} \mathbf{P}(i,j)\right)^2}{N_p}$$

$$correlation = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) ij - \mu_x \mu_y}{\sigma_x(i) \sigma_y(j)}$$

$$MAD = \frac{1}{N_p} \sum_{i=1}^{N_p} |\mathbf{X}(i) - \bar{X}|$$

$$minimum = \min(\mathbf{X})$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_s} \mathbf{P}\right.}{N_z}$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2$$

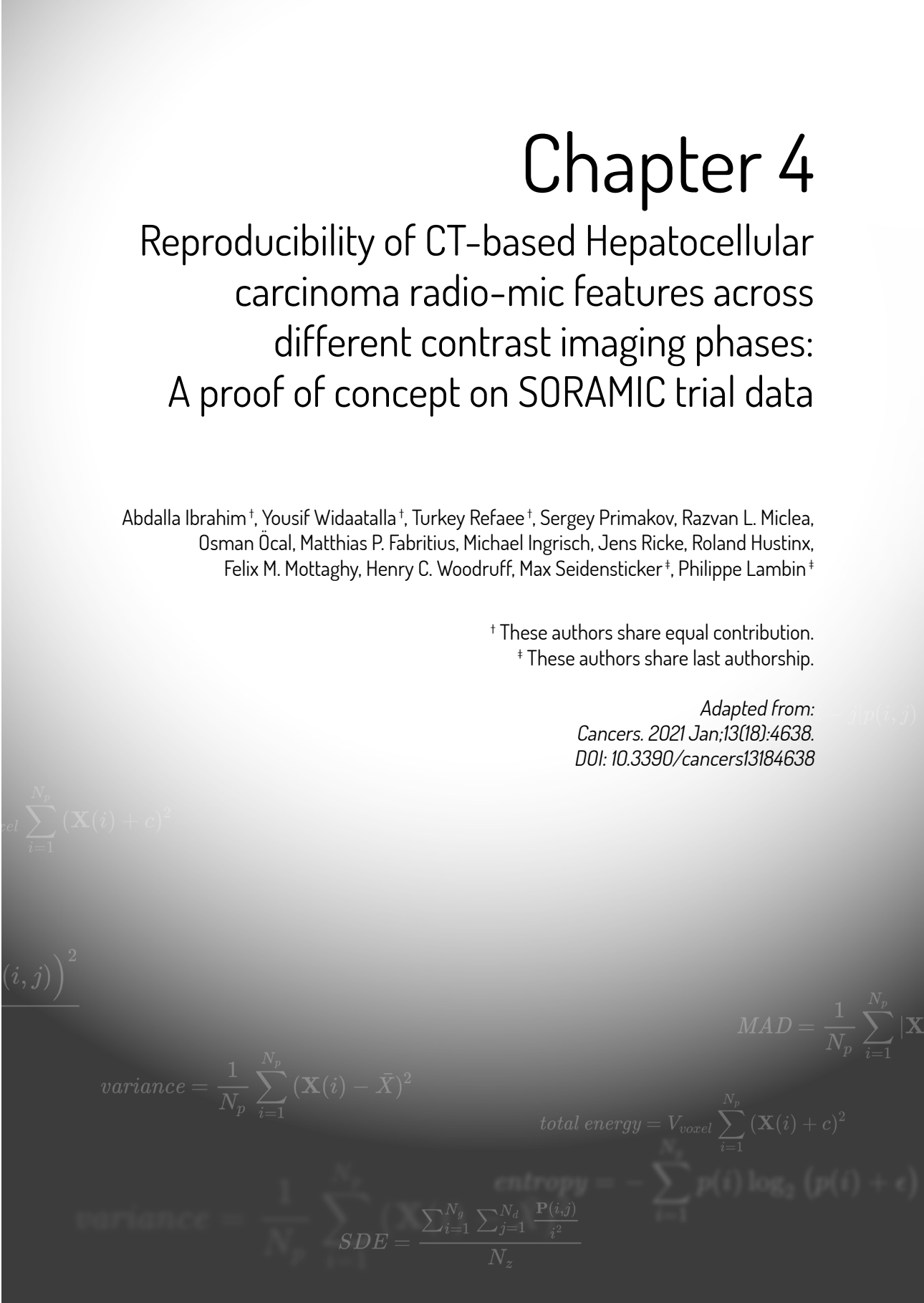
Chapter 4

Reproducibility of CT-based Hepatocellular carcinoma radio-mic features across different contrast imaging phases: A proof of concept on SORAMIC trial data

Abdalla Ibrahim [†], Yousif Widaatalla [†], Turkey Refaee [†], Sergey Primakov, Razvan L. Miclea, Osman Öcal, Matthias P. Fabritius, Michael Ingrisch, Jens Ricke, Roland Hustinx, Felix M. Mottaghy, Henry C. Woodruff, Max Seidensticker [‡], Philippe Lambin [‡]

[†] These authors share equal contribution.
[‡] These authors share last authorship.

Adapted from:
Cancers. 2021 Jan;13(18):4638.
DOI: 10.3390/cancers13184638



Abstract

Handcrafted radiomic features (HRFs) are quantitative imaging features extracted from regions of interest on medical images, which can be correlated with clinical outcomes and biologic characteristics. While HRFs have been used to train predictive and prognostic models, their reproducibility has been reported to be affected by variations in scan acquisition and reconstruction parameters, even within the same imaging vendor. In this work, we evaluated the reproducibility of HRFs across the arterial and portal venous phases of contrast enhanced computed tomography images depicting hepatocellular carcinomas, as well as the potential of ComBat harmonization to correct for this difference. ComBat harmonization is a method based on Bayesian estimates that was developed for gene expression arrays, and has been investigated as a potential method for harmonizing HRFs. Our results show that the majority of HRFs are not reproducible between the arterial and portal venous imaging phases, yet a number of HRFs could be used interchangeably between those phases. Furthermore, ComBat harmonization increased the number of reproducible HRFs across both phases by 1%. Our results guide the pooling of arterial and venous phases from different patients in an effort to increase cohort size, as well as joint analysis of the phases.

Keywords: Hepatocellular carcinoma; CT radiomics; domain translation; reproducibility.

Introduction

The recent decades witnessed vast advances in computational power, artificial intelligence, and medical imaging techniques [1], which provided a unique opportunity for transforming the abundant amounts of medical imaging into mineable quantitative data. The concept acquired much scientific attention recently, and a branch of medical imaging analysis -known as handcrafted radiomics- emerged as a result [2]. Handcrafted Radiomic features (HRFs) are quantitative features extracted with high throughput from medical imaging, with its varying modalities. The hypothesis is that medical images carry more data than can be seen by trained human eyes, and that these data can be decoded using the HRFs, i.e correlations between HRFs and underlying biology could potentially exist [3]. Since the introduction of the field, many studies reported on the potential of radiomic signatures to predict clinical endpoints, the majority of which were performed on computed tomography (CT) [4–7], magnetic resonance (MR) [8–10], and positron emission tomography (PET) scans [11,12].

Hepatocellular carcinoma (HCC) is the most common primary liver cancer, the fifth most common malignancy worldwide, and a leading cause of cancer-related mortality [13]. Different diagnostic approaches and treatment modalities are used clinically depending on the characteristics of the patient and the progression of the disease [14,15]. Contrast-enhanced computed tomography (CE-CT) scans are considered one of the main diagnostic tools for HCC. CE-CT can be acquired at different times following the injection of the contrast agent to acquire arterial, venous or late phase scans. Each phase shows specific characteristics for HCC lesions. However, there is still a clinical need for reliable non-invasive tools that could aid diagnosing and devising individualized treatment plans for HCC patients. Several studies investigated and reported on the potential of HRFs to aid clinical decision making in HCC patients [16–19].

While numerous studies have reported on the potential of HRFs in aiding clinical decision making on HCC and other diseases, several hurdles hindering the clinical translation of radiomic signatures to clinical decision support systems have been identified. These hurdles include the reproducibility of HRFs in test-retest studies, their sensitivity to variations in acquisition and reconstruction parameters of the scans, inter-observer variability, and the need for big data [20–26]. However, the need for big data in radiomics analysis necessitates the exploration of methods for combining and comparing retrospective medical imaging databases.

A number of studies tried to address the issue of reproducibility of HRFs using ComBat harmonization [27–30]. ComBat harmonization is a method that was developed to remove the batch effects in gene expression arrays [31]. The studies that investigated the application of ComBat in radiomics analyses reported on the improvement in performance metrics of

developed radiomic signatures after the application of ComBat compared to before, and recommended the use of the method. Other studies that investigated the reproducibility of HRFs on phantom datasets acquired with different settings [32], or with a single parameter difference [33], and reported that the performance of ComBat is dependent on the data under study and recommended a framework to assess the reproducibility of HRFs. Yet to date, no study reported on the agreement in HRFs across different phases or the potential of ComBat to remove the effects of different imaging phases from HRFs, which could allow the proper combination of phases in a single analysis, or the interchangeability of HRFs across phases to allow the use of different imaging scans per patient. Furthermore, no study performed a reproducibility analysis for HRFs following ComBat harmonization on patients' scans acquired with a single parameter difference.

We hypothesize that the time of acquisition after the injection of the contrast agents adds another level of complexity to be accounted for in the radiomics analysis, as HRFs might be affected by the appearance of contrast, due to the variations in the distribution of the contrast within the lesions. As a proof of concept, we investigate the sensitivity of HRFs extracted from CE-CT scans depicting HCC acquired during the arterial and portal venous phases, when all other acquisition and reconstruction parameters were fixed. Furthermore, we investigate the potential of the ComBat harmonization for domain translation of the HRFs extracted from these scans. Ultimately, we aim to (i) guide the identification of HRFs that can be used interchangeably between arterial and venous phase scans, which could increase the number of scans that can be included in a CE-CT based radiomics study; and (ii) identify the features that can be used in studies analyzing both phases simultaneously to maximize the information extracted from ROIs.

Materials and Methods

Patients and Imaging data

The imaging data were originally collected for the European multicenter clinical trial (SORAMIC) [34]. Imaging data for 424 patients diagnosed with HCC (using cyto-histological criteria, radiologic criteria, or a combination of both) were obtained for the SORAMIC trial, of which 338 scans were available for analysis in this study. Scans that contained artifacts were considered of poor quality (n=48). From the available 338 patients with both arterial and portal venous scans available, patients with scans that had any difference in the acquisition or reconstruction parameters, or lacked segmentations reviewed by an expert, were excluded. A total of 61 patients with 104 distinct lesions were finally included in this study (Figure 1). Scans included were acquired from different hospitals, using different vendors and protocols. In total, 9 scanner models from 4 different imaging vendors, and a range of scanning parameters, were included, as shown in Table 1. The imaging analysis was approved by the University of Magdeburg institutional review board (IRB00006099, EudraCT

no 2009-012576-27), and informed consent was obtained from all included patients. All methods were carried out in accordance with the relevant guidelines and regulations [35].

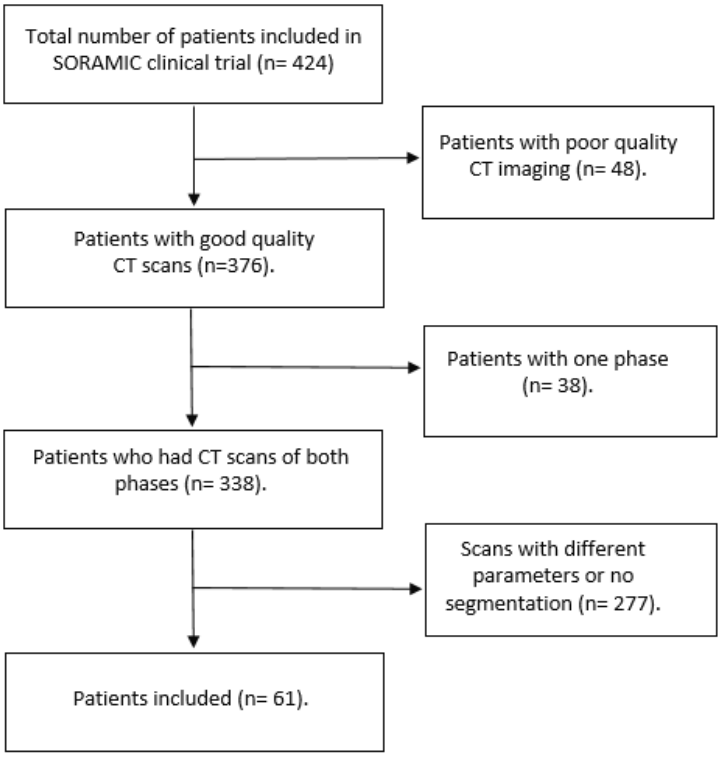


Figure 1. A flowchart showing the patients selection process.

Table 1. Acquisition and reconstruction parameters for the imaging dataset

Manufacturer	Scanner model	X-Ray Tube Current (kV)	Exposure (mAs)	Convolution kernels	Slice thickness (mm)	Pixel spacing (mm^2)
TOSHIBA	Aquilion	50 - 360	2-300	FC13	1-5	0.39x0.39 - 0.98x0.98
	Aquilion PRIME					
Philips	Brilliance 64			B		
GE	Discovery CT750 HD			STANDARD		
	Optima CT660					
SIEMENS	Sensation 16			B31f		
	SOMATOM Definition AS					
	SOMATOM Definition Flash			I30f , I40f		
	SOMATOM Force			Br40d		

Segmentation and HRFs extraction

The scans of a single patient were co-registered. The region of interest (ROI) was segmented on each scan while viewing both phases simultaneously and saved to both scans (Fig 2). The segmentations were performed using MIM software (MIM Software Inc., Cleveland, OH) by a medical doctor (Y.W) with 2 years of experience in image segmentation, and revised by a radiologist (R.M.) with 15 years of experience in medical radiology.

HRFs were extracted from these ROIs using the software RadiomiX Discovery Toolbox (version, October 2019; https://www.radiomics.bio), which calculates HRFs compliant with the Imaging Biomarkers Standardization Initiative (IBSI) [36], in addition to others. Image intensities were binned with a binwidth of 25 Hounsfield Units (HUs) in order to reduce noise levels and to reduce texture matrix sizes, and therewith computation power, with no resampling or further preprocessing of the images. The description of the extracted HRFs was published previously [24].

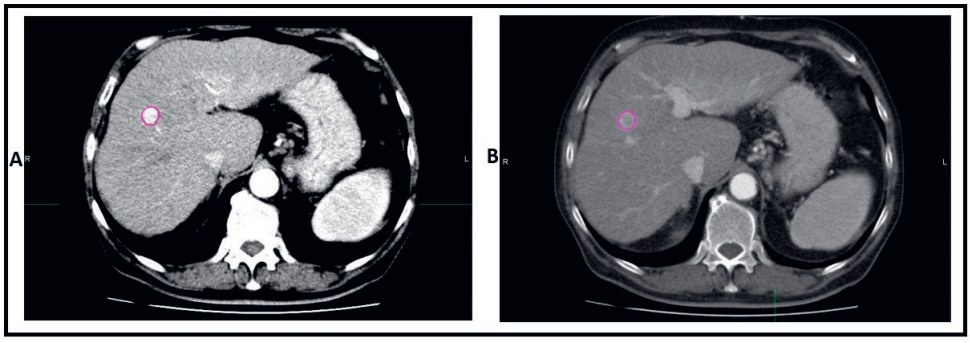


Figure 2. An example of ROI segmented in (A) the arterial phase and (B) portal venous phase.

ComBat Harmonization

ComBat method employs empirical Bayes to estimate the effects of assigned batches on the data being harmonized. For HRFs, ComBat assumes that a feature value can be approximated by the equation:

$$Y_{ij} = \alpha + \beta X_{ij} + \gamma_i + \delta_i \varepsilon_{ij} \quad (1)$$

where α is the average value for HRF Yij for ROI j on scanner i; X is a design matrix of the biologic covariates that are known to affect the value of HRFs; β is the vector of regression coefficients corresponding to each biologic covariate; γi is the additive effect of scanner i on HRFs, δi is the multiplicative scanner effect, and εij is an error term, presupposed to be normally distributed with zero mean. Based on the values estimated, ComBat performs feature transformation as given by the formula:

$$Y_{ij}^{ComBat} = \frac{\left(Y_{ij} - \hat{\alpha} - \hat{\beta}X_{ij} - \gamma_i^*\right)}{\delta_i^*} + \hat{\alpha} + \hat{\beta}X_{ij} \quad (2)$$

where $\hat{\alpha}$ and $\hat{\beta}$ are estimators of the parameters α and β, respectively; and γi* and δi* are the empirical Bayes estimates for the parameters γi and δi, respectively.

Statistical Analysis

All statistical analyses were performed using R language [37] on RStudio (V 3.6.3) [38]. To determine the reproducibility of HRFs, the concordance correlation coefficient (CCC) between the HRFs values across the two phases was calculated [39], using epiR package [40]. The CCC measures how concordant are the values of a given HRF and the rank of each data point relative to the rest in each batch. HRFs with CCC>0.9 were considered reproducible and could be interchangeably used between the arterial and venous phase CT scans.

To assess the performance of ComBat, shape features and HRFs with (near) zero variance (HRFs that have the same value in 95% or more of the observations) were removed. The phase of the scan was assigned as the batch for ComBat harmonization. The CCC was calculated after ComBat application and the cutoff of CCC>0.9 was applied to select the concordant HRFs. The correlation of concordant features with volume was assessed using Pearson correlation. Features that had a correlation coefficient > 0.85 were considered highly correlated. The analysis code used in this study can be found on: (https://github.com/AbdallaIbrahim/The-reproducibility-and-ComBatability-of-Radiomic-features).

Results

Patient characteristics

The patients included (n=61) had a median age of 66 years, mainly male (n=50, 81.9%), with cirrhotic livers (n=56, 91.8%), and a minority (n=11, 18.1%) had portal vein invasion. For more patient characteristics see Table 2.

Table 2. Patient characteristics.

Characteristic	N=61
Gender, male (%)	50 (81.9%)
Age, median (range)	66 (48-81)
Cirrhosis, yes (%)	56 (91.8%)
Child-Pugh grade	56 (91.8%)
A	5 (8.2%)
B	
Diameter of largest lesion, in mm, median (range)	37 (10-220)
Portal vein invasion, yes (%)	11 (18.1%)
Extrahepatic disease yes (%)	7 (11.4%)
BCLC staging	22 (36.1%)
A	22 (36.1%)
B	17 (27.8%)
C	
ECOG performance	58 (95.1%)
0	3 (4.9%)
1	

* Barcelona Clinic Liver cancer (BCLC) staging
** European Cooperative Oncology Group (ECOG) performance

Extracted HRFs

A total of 167 original HRFs were extracted from each of the available 104 ROIs. These HRFs are divided into 11 feature families: Fractal (n=3), Gray Level Co-occurence Matrix (GLCM; n= 26), Gray Level Distance Zone Matrix (GLDZM; n=16), Gray Level Run Length Matrix (GLRLM; n=15), Gray Level Size Zone Matrix (GLSZM, n=16), Intensity Histogram (IH; n=25), Local Intensity (LocInt, n=2), Neighbouring Gray Level Dependence Matrix (NGLDM; n=17), Neighbouring Gray Tone Difference Matrix (NGTDM, n=5), Shape (n=23), and Statistics (Stats, n=19).

The effects of differences in imaging phase on the reproducibility of HRFs

Out of the 167 extracted HRFs, 42 (25%) were reproducible (had a CCC>0.9) across both phases (Figure 3a, shape features were not included to ease the comparison between figures). These HRFs were divided into shape (n=22), NGTDM (n=1), NGLDM (n=4), IH (n=2), GLSZM (n=4), GLRLM (n=2) and GLDZM (n=7). The remaining HRFs had a CCC ranging from -0.07 and 0.85, with a median of 0.39.

Of the concordant 22 shape features, 8 features were highly correlated with volume (R>0.85), in addition to 1 feature from the NGLDM group (NGLDM_DN) and 2 features from the GLRLM group (GLRLM_RLN and GLRLM_GLN). The remaining features (31, 73.8%) had a correlation coefficient <0.85.

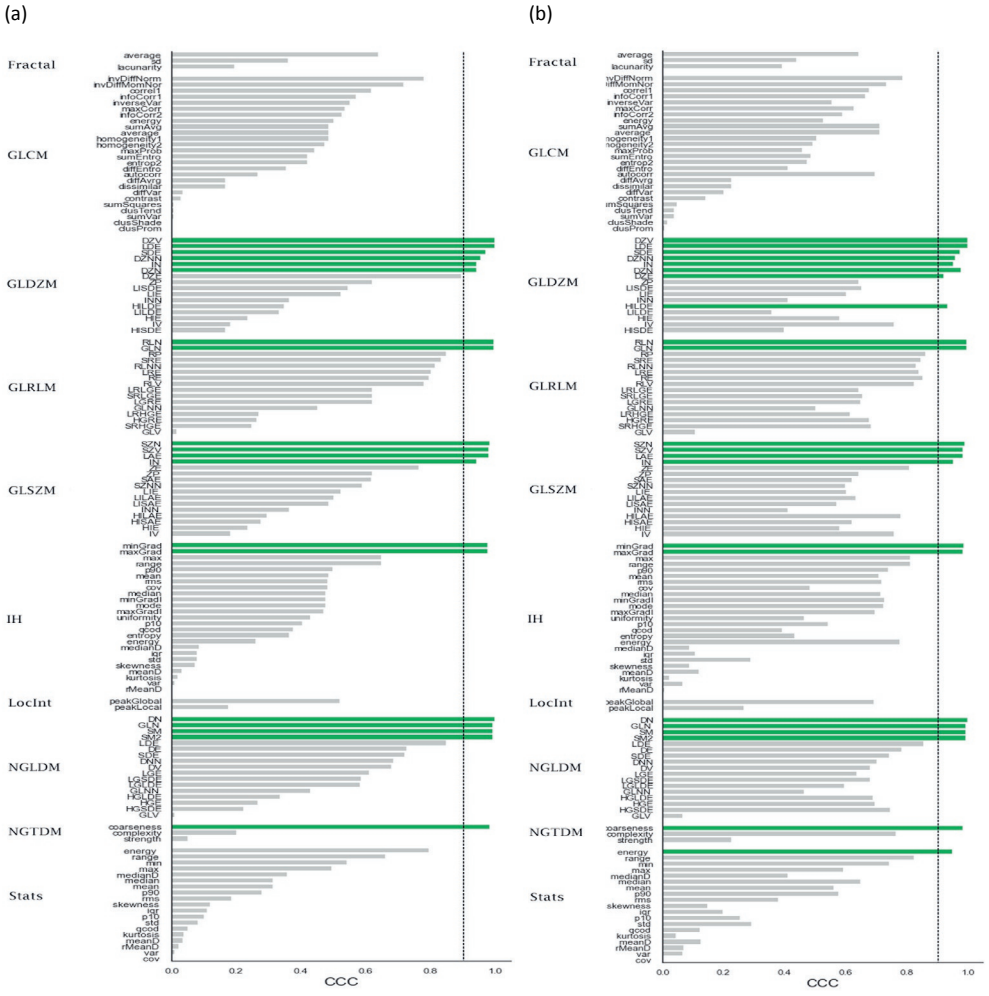


Figure 3. (a) The CCC values for the different HRFs before ComBat harmonization; **(b)** The CCC values for the different HRFs after ComBat harmonization

The effects of ComBat on the reproducibility of HRFs

The application of ComBat harmonization to remove the batch effects attributed to the difference in time between contrast injection and scan acquisition resulted in a total of 44 (26.1%) reproducible HRFs, i.e 2 extra HRFs became concordant following the application of ComBat: Stats_energy and GLDZM_HILDE (Fig 3b). The remaining 20 HRFs had a CCC>0.9 before and after ComBat harmonization, in addition to the shape features (n=22). The CCC of stats_energy increased from 0.8 to 0.95 following ComBat harmonization, and that of GLDZM_HILDE increased from 0.34 to 0.93.

The impact of ComBat on the CCC values had a wide range; 6 HRFs had an increment in CCC between 0.5 and 0.6; 42 HRFs had an increment in CCC between 0.1 and 0.49; 87 HRFs had an increment between 0 and 0.09; and 33 HRFs had a decrement in CCC between -0.001 and -0.06. Following ComBat harmonization, the number of highly correlated features with volume increased by one feature (Stats_energy). The concordant features before domain translation maintained their correlation with the volume.

Discussion

In this study, we investigated the reproducibility of HCC CT-based HRFs across the arterial and portal venous imaging phases when all other scanning parameters were fixed, and whether ComBat harmonization improves the reproducibility of HRFs in such a scenario. Uniquely, this is the first manuscript to investigate the potential of ComBat to remove batch effects attributed to the differences in imaging phase, and on patient data with a single parameter difference between the compared/harmonized scans. Our results show that the majority of HRFs were significantly affected by the difference in imaging phases, and only a quarter of the total extracted number of HRFs were reproducible across both phases. Moreover, ComBat harmonization did not successfully harmonize the majority of HRFs, even though the differences between the batches compared were limited to the variations in imaging phase.

HRFs are calculated using mathematical formulas applied on the array of values representing the medical image [41]. Changes in the value of units in this array are expected to have an impact on the value calculated by the same formula. Therefore, changes in the scanning parameters are expected to affect the reproducibility of different HRFs variably. Aside from HRFs that are not reproducible in test-retest studies, the sensitivity of the remaining HRFs to the imaging phase can be justified by the increased radio-opaqueness and the resulting perfusion patterns of contrast within the ROI, and thus, changes in the image array values based on which the HRFs are calculated. As expected, statistics and intensity histogram features, which are simple HRFs based on a single voxel value (e.g. minimum or maximum intensity value) or the description of their distribution (e.g. mean or median intensity value), were found to be the most significantly affected families. On the other hand, also according to expectations, HRFs that do not depend on the intensity values, but the shape of the segmentation (shape features), were found to be reproducible across both phases, with the exception of the shape feature centroid distance, which is based on the distribution of intensity values around the geometric center of the ROI. The copying of segmentations and the inclusion of scans that were acquired identically in both phases allowed isolating the effects of differences imaging phases on HRFs. However, in scenarios where acquisition and/or reconstruction parameters, or the segmentation of the ROI changes, the reproducibility

of HRFs is expected to be further impacted. This is also in line with what reported in a study that investigated the reproducibility of liver parenchyma and tumors HRFs extracted from two contrast enhanced scans (one phase) taken within a 14 days interval [42]. Therefore, the reproducibility analysis based on the data under study should be an integral part of each radiomics study.

Our study sheds the light on the methodology of combining HRFs from different modalities, either for the purpose of combining different phases/modalities per patient, or the combination of different phases for different patients. For merging different modalities per patient, we show that a number of HRFs is reproducible across the phases. Therefore, models that try to combine different imaging phases per patient are recommended to define which reproducible (test-retest) HRFs vary across the available phases, and preselect those for further analysis. Another implication of our findings is allowing the combination of different imaging phases per patient (e.g due to the lack of data), when only the reproducible HRFs across phases are extracted and compared between the different patients, regardless of the available imaging phase for each patient. This approach can significantly increase the number of data points in retrospective radiomics studies.

The correlation of radiomic features with the volume of the ROI has been considered one of the major points to be assessed in radiomics analysis, since some of the features were reported previously to be surrogates of volume [43]. In our analysis, we observed that the majority of the features identified as concordant (or domain-translatable with ComBat) between the arterial and venous CT scans was considerable, most of which were shape features. However, the majority of features were not found to be highly correlated with volume, which means that these features can decode additional information about the ROIs being investigated.

The number of features that had a CCC value higher than 0.9 was slightly higher after the application of ComBat on the HRFs extracted from the arterial and portal venous phases. ComBat successfully harmonized two additional HRFs compared to the number of concordant HRFs before domain translation. The majority of HRFs were not concordant across the phases even after the application of ComBat harmonization. The differences in ComBat performance per HRF (and feature families) are also expected, as in contrast to gene expression arrays, HRFs have different levels of complexity and are not expected to be uniformly affected by the batch defined for domain translation. The variant performance of ComBat on HRFs could be explained by the differences in the complexity of HRFs, compared to gene expression arrays [21]. The findings are in line with the reproducibility studies that assessed the performance of ComBat on phantom scans, which reported that ComBat harmonization does not successfully harmonize all HRFs, and that its performance is dependent on the variations between the batches [32,33]. As a consequence, we recommend

that the application of ComBat harmonization on HRFs follows a reproducibility analysis with reference values to assess its performance, as it is expected to vary with the variations in the dataset batches being harmonized [21]. Other deep learning based harmonization methods that have been recently investigated [44–47] might be more suitable for domain translation of images acquired in different phases. However, this is yet to be investigated.

While this study provides a proof of concept for the combination/replacement of different imaging phases, we speculate that the set of reproducible HRFs identified in this study is limited to HCC lesions extracted from scans acquired similarly to our dataset. Furthermore, the changes in reconstruction parameters (and sometimes acquisition parameters) between the two imaging phases in clinical routine significantly lowered the number of available scans to perform this analysis. Lastly, the reproducibility of the identified HRFs has to be investigated across different acquisition and reconstruction parameters. However, due to the lack of data, this was not performed. Nevertheless, this study serves as a guide for selecting and/or harmonizing the reproducible HRFs in future radiomic studies that utilize contrast enhanced imaging.

Conclusions

The majority of HRFs are significantly affected by changes in the imaging phase of the scan. Studies that investigate the potential of combining HRFs from different imaging phases or modalities must investigate the reproducibility and interoperability of the HRFs across the investigated phases for the lesions of interest. Furthermore, a number of HRFs can be interchangeably used between the arterial and portal venous phases, and these can be used to increase data points in retrospective imaging studies. ComBat harmonization increased the number of comparable CT based HRFs across the arterial and portal venous imaging phases for HCC lesions by 1% in our dataset.

References

1. Walsh, S.; de Jong, E.E.C.; van Timmeren, J.E.; Ibrahim, A.; Compter, I.; Peerlings, J.; Sanduleanu, S.; Refaee, T.; Keek, S.; Larue, R.T.H.M.; et al. Decision Support Systems in Oncology. JCO Clin Cancer Inform 2019, 3, 1–9, doi:10.1200/CCI.18.00001.
2. Lambin, P.; Rios-Velazquez, E.; Leijenaar, R.; Carvalho, S.; van Stiphout, R.G.P.M.; Granton, P.; Zegers, C.M.L.; Gillies, R.; Boellard, R.; Dekker, A.; et al. Radiomics: Extracting More Information from Medical Images Using Advanced Feature Analysis. Eur. J. Cancer 2012, 48, 441–446, doi:10.1016/j.ejca.2011.11.036.
3. Gillies, R.J.; Kinahan, P.E.; Hricak, H. Radiomics: Images Are More than Pictures, They Are Data. Radiology 2016, 278, 563–577, doi:10.1148/radiol.2015151169.
4. Refaee, T.; Wu, G.; Ibrahim, A.; Halilaj, I.; Leijenaar, R.T.H.; Rogers, W.; Gietema, H.A.; Hendriks, L.E.L.; Lambin, P.; Woodruff, H.C. The Emerging Role of Radiomics in COPD and Lung Cancer. Respiration 2020, 99, 99–107, doi:10.1159/000505429.
5. Aerts, H.J.W.L. The Potential of Radiomic-Based Phenotyping in Precision Medicine: A Review. JAMA Oncol 2016, 2, 1636–1642, doi:10.1001/jamaoncol.2016.2631.
6. van Timmeren, J.E.; Leijenaar, R.T.H.; van Elmpt, W.; Reymen, B.; Oberije, C.; Monshouwer, R.; Bussink, J.; Brink, C.; Hansen, O.; Lambin, P. Survival Prediction of Non-Small Cell Lung Cancer Patients Using Radiomics Analyses of Cone-Beam CT Images. Radiother. Oncol. 2017, 123, 363–369, doi:10.1016/j.radonc.2017.04.016.
7. Panth, K.M.; Leijenaar, R.T.H.; Carvalho, S.; Lieuwes, N.G.; Yaromina, A.; Dubois, L.; Lambin, P. Is There a Causal Relationship between Genetic Changes and Radiomics-Based Image Features? An in Vivo Preclinical Experiment with Doxycycline Inducible GADD34 Tumor Cells. Radiother. Oncol. 2015, 116, 462–466, doi:10.1016/j.radonc.2015.06.013.
8. Jethanandani, A.; Lin, T.A.; Volpe, S.; Elhalawani, H.; Mohamed, A.S.R.; Yang, P.; Fuller, C.D. Exploring Applications of Radiomics in Magnetic Resonance Imaging of Head and Neck Cancer: A Systematic Review. Front. Oncol. 2018, 8, 131, doi:10.3389/fonc.2018.00131.
9. Ursprung, S.; Beer, L.; Bruining, A.; Woitek, R.; Stewart, G.D.; Gallagher, F.A.; Sala, E. Radiomics of Computed Tomography and Magnetic Resonance Imaging in Renal Cell Carcinoma—a Systematic Review and Meta-Analysis. Eur. Radiol. 2020, 30, 3558–3566, doi:10.1007/s00330-020-06666-3.
10. Samiei, S.; Granzier, R.W.Y.; Ibrahim, A.; Primakov, S.; Lobbes, M.B.I.; Beets-Tan, R.G.H.; van Nijnatten, T.J.A.; Engelen, S.M.E.; Woodruff, H.C.; Smidt, M.L. Dedicated Axillary MRI-Based Radiomics Analysis for the Prediction of Axillary Lymph Node Metastasis in Breast Cancer. Cancers 2021, 13, doi:10.3390/cancers13040757.
11. Ibrahim, A.; Vallières, M.; Woodruff, H.; Primakov, S.; Beheshti, M.; Keek, S.; Refaee, T.; Sanduleanu, S.; Walsh, S.; Morin, O.; et al. Radiomics Analysis for Clinical Decision Support in Nuclear Medicine. Semin. Nucl. Med. 2019, 49, 438–449, doi:10.1053/j.semnuclmed.2019.06.005.

12. Lovinfosse, P.; Visvikis, D.; Hustinx, R.; Hatt, M. FDG PET Radiomics: A Review of the Methodological Aspects. Clinical and Translational Imaging 2018, 6, 379–391, doi:10.1007/s40336-018-0292-9.
13. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J. Clin. 2021, doi:10.3322/caac.21660.
14. Aubé, C.; Oberti, F.; Lonjon, J.; Pageaux, G.; Seror, O.; N'Kontchou, G.; Rode, A.; Radenne, S.; Cassinotto, C.; Vergniol, J.; et al. EASL and AASLD Recommendations for the Diagnosis of HCC to the Test of Daily Practice. Liver Int. 2017, 37, 1515–1525, doi:10.1111/liv.13429.
15. Finn, R.S.; Qin, S.; Ikeda, M.; Galle, P.R.; Ducreux, M.; Kim, T.-Y.; Kudo, M.; Breder, V.; Merle, P.; Kaseb, A.O.; et al. Atezolizumab plus Bevacizumab in Unresectable Hepatocellular Carcinoma. N. Engl. J. Med. 2020, 382, 1894–1905, doi:10.1056/NEJMoa1915745.
16. Mokrane, F.-Z.; Lu, L.; Vavasseur, A.; Otal, P.; Peron, J.-M.; Luk, L.; Yang, H.; Ammari, S.; Saenger, Y.; Rousseau, H.; et al. Radiomics Machine-Learning Signature for Diagnosis of Hepatocellular Carcinoma in Cirrhotic Patients with Indeterminate Liver Nodules. Eur. Radiol. 2020, 30, 558–570, doi:10.1007/s00330-019-06347-w.
17. Wu, J.; Liu, A.; Cui, J.; Chen, A.; Song, Q.; Xie, L. Radiomics-Based Classification of Hepatocellular Carcinoma and Hepatic Haemangioma on Precontrast Magnetic Resonance Images. BMC Med. Imaging 2019, 19, 23, doi:10.1186/s12880-019-0321-9.
18. Zhou, Y.; He, L.; Huang, Y.; Chen, S.; Wu, P.; Ye, W.; Liu, Z.; Liang, C. CT-Based Radiomics Signature: A Potential Biomarker for Preoperative Prediction of Early Recurrence in Hepatocellular Carcinoma. Abdom Radiol (NY) 2017, 42, 1695–1704, doi:10.1007/s00261-017-1072-0.
19. Wakabayashi, T.; Ouhmich, F.; Gonzalez-Cabrera, C.; Felli, E.; Saviano, A.; Agnus, V.; Savadjiev, P.; Baumert, T.F.; Pessaux, P.; Marescaux, J.; et al. Radiomics in Hepatocellular Carcinoma: A Quantitative Review. Hepatol. Int. 2019, 13, 546–559, doi:10.1007/s12072-019-09973-0.
20. Yip, S.S.F.; Aerts, H.J.W.L. Applications and Limitations of Radiomics. Phys. Med. Biol. 2016, 61, R150–66, doi:10.1088/0031-9155/61/13/R150.
21. Ibrahim, A.; Primakov, S.; Beuque, M.; Woodruff, H.C.; Halilaj, I.; Wu, G.; Refaee, T.; Granzier, R.; Widaatalla, Y.; Hustinx, R.; et al. Radiomics for Precision Medicine: Current Challenges,future Prospects, and the Proposal of a New Framework. Methods 2020, doi:10.1016/j.ymeth.2020.05.022.
22. Larue, R.T.H.M.; van Timmeren, J.E.; de Jong, E.E.C.; Feliciani, G.; Leijenaar, R.T.H.; Schreurs, W.M.J.; Sosef, M.N.; Raat, F.H.P.J.; van der Zande, F.H.R.; Das, M.; et al. Influence of Gray Level Discretization on Radiomic Feature Stability for Different CT Scanners, Tube Currents and Slice Thicknesses: A Comprehensive Phantom Study. Acta Oncol. 2017, 56, 1544–1553, doi:10.1080/0284186X.2017.1351624.
23. van Timmeren, J.E.; Leijenaar, R.T.H.; van Elmpt, W.; Wang, J.; Zhang, Z.; Dekker, A.; Lambin, P. Test-Retest Data for Radiomics Feature Stability Analysis: Generalizable or Study-Specific? Tomography 2016, 2, 361–365, doi:10.18383/j.tom.2016.00208.

24. Peerlings, J.; Woodruff, H.C.; Winfield, J.M.; Ibrahim, A.; Van Beers, B.E.; Heerschap, A.; Jackson, A.; Wildberger, J.E.; Mottaghy, F.M.; DeSouza, N.M.; et al. Stability of Radiomics Features in Apparent Diffusion Coefficient Maps from a Multi-Centre Test-Retest Trial. Sci. Rep. 2019, 9, 4800, doi:10.1038/s41598-019-41344-5.
25. Granzier, R.W.Y.; Verbakel, N.M.H.; Ibrahim, A.; van Timmeren, J.E.; van Nijnatten, T.J.A.; Leijenaar, R.T.H.; Lobbes, M.B.I.; Smidt, M.L.; Woodruff, H.C. MRI-Based Radiomics in Breast Cancer: Feature Robustness with Respect to Inter-Observer Segmentation Variability. Sci. Rep. 2020, 10, 14163, doi:10.1038/s41598-020-70940-z.
26. Leijenaar, R.T.H.; Carvalho, S.; Velazquez, E.R.; van Elmpt, W.J.C.; Parmar, C.; Hoekstra, O.S.; Hoekstra, C.J.; Boellaard, R.; Dekker, A.L.A.J.; Gillies, R.J.; et al. Stability of FDG-PET Radiomics Features: An Integrated Analysis of Test-Retest and Inter-Observer Variability. Acta Oncol. 2013, 52, 1391–1397, doi:10.3109/0284186X.2013.812798.
27. Orlhac, F.; Frouin, F.; Nioche, C.; Ayache, N.; Buvat, I. Validation of a Method to Compensate Multicenter Effects Affecting CT Radiomic Features. 2018.
28. Orlhac, F.; Boughdad, S.; Philippe, C.; Stalla-Bourdillon, H.; Nioche, C.; Champion, L.; Soussan, M.; Frouin, F.; Frouin, V.; Buvat, I. A Postreconstruction Harmonization Method for Multicenter Radiomic Studies in PET. J. Nucl. Med. 2018, 59, 1321–1328, doi:10.2967/jnumed.117.199935.
29. Da-Ano, R.; Masson, I.; Lucia, F.; Doré, M.; Robin, P.; Alfieri, J.; Rousseau, C.; Mervoyer, A.; Reinhold, C.; Castelli, J.; et al. Performance Comparison of Modified ComBat for Harmonization of Radiomic Features for Multicenter Studies. Sci. Rep. 2020, 10, 10248, doi:10.1038/s41598-020-66110-w.
30. Traverso, A.; Wee, L.; Dekker, A.; Gillies, R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. Int. J. Radiat. Oncol. Biol. Phys. 2018, 102, 1143–1158, doi:10.1016/j.ijrobp.2018.05.053.
31. Johnson, W.E.; Li, C.; Rabinovic, A. Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods. Biostatistics 2007, 8, 118–127, doi:10.1093/biostatistics/kxj037.
32. Ibrahim, A.; Refaee, T.; Leijenaar, R.T.H.; Primakov, S.; Hustinx, R.; Mottaghy, F.M.; Woodruff, H.C.; Maidment, A.D.A.; Lambin, P. The Application of a Workflow Integrating the Variable Reproducibility and Harmonizability of Radiomic Features on a Phantom Dataset. PLoS One 2021, 16, e0251147, doi:10.1371/journal.pone.0251147.
33. Ibrahim, A.; Refaee, T.; Primakov, S.; Barufaldi, B.; Acciavatti, R.J.; Granzier, R.W.Y.; Hustinx, R.; Mottaghy, F.M.; Woodruff, H.C.; Wildberger, J.E.; et al. The Effects of in-Plane Spatial Resolution on CT-Based Radiomic Features' Stability with and without ComBat Harmonization. Cancers 2021, 13, 1848, doi:10.3390/cancers13081848.
34. Ricke, J.; Bulla, K.; Kolligs, F.; Peck-Radosavljevic, M.; Reimer, P.; Sangro, B.; Schott, E.; Schütte, K.; Verslype, C.; Walecki, J.; et al. Safety and Toxicity of Radioembolization plus Sorafenib in Advanced Hepatocellular Carcinoma: Analysis of the European Multicentre Trial SORAMIC. Liver Int. 2015, 35, 620–626.

35. World Medical Association World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. JAMA 2013, 310, 2191–2194, doi:10.1001/jama.2013.281053.
36. Zwanenburg, A.; Vallières, M.; Abdalah, M.A.; Aerts, H.J.W.L.; Andrearczyk, V.; Apte, A.; Ashrafinia, S.; Bakas, S.; Beukinga, R.J.; Boellaard, R.; et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-Based Phenotyping. Radiology 2020, 191145, doi:10.1148/radiol.2020191145.
37. Team, R.C. R Language Definition. Vienna, Austria: R foundation for statistical computing 2000.
38. Gandrud, C. Reproducible Research with R and R Studio; CRC Press, 2013; ISBN 9781466572843.
39. Lin, L.I. A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics 1989, 45, 255–268.
40. Stevenson, M.; Stevenson, M.M.; BiasedUrn, I. Package "epiR." 2020.
41. Aerts, H.J.W.L.; Velazquez, E.R.; Leijenaar, R.T.H.; Parmar, C.; Grossmann, P.; Carvalho, S.; Bussink, J.; Monshouwer, R.; Haibe-Kains, B.; Rietveld, D.; et al. Decoding Tumour Phenotype by Noninvasive Imaging Using a Quantitative Radiomics Approach. Nat. Commun. 2014, 5, 4006, doi:10.1038/ncomms5006.
42. Perrin, T.; Midya, A.; Yamashita, R.; Chakraborty, J.; Saidon, T.; Jarnagin, W.R.; Gonen, M.; Simpson, A.L.; Do, R.K.G. Short-Term Reproducibility of Radiomic Features in Liver Parenchyma and Liver Malignancies on Contrast-Enhanced CT Imaging. Abdom Radiol (NY) 2018, 43, 3271–3278, doi:10.1007/s00261-018-1600-6.
43. Welch, M.L.; McIntosh, C.; Haibe-Kains, B.; Milosevic, M.F.; Wee, L.; Dekker, A.; Huang, S.H.; Purdie, T.G.; O'Sullivan, B.; Aerts, H.J.W.L.; et al. Vulnerabilities of Radiomic Signature Development: The Need for Safeguards. Radiother. Oncol. 2019, 130, 2–9, doi:10.1016/j.radonc.2018.10.027.
44. Andrearczyk, V.; Depeursinge, A.; Müller, H. Neural Network Training for Cross-Protocol Radiomic Feature Standardization in Computed Tomography. J Med Imaging (Bellingham) 2019, 6, 024008, doi:10.1117/1.JMI.6.2.024008.
45. Bashyam, V.M.; Doshi, J.; Erus, G.; Srinivasan, D.; Abdulkadir, A.; Habes, M.; Fan, Y.; Masters, C.L.; Maruff, P.; Zhuo, C.; et al. Medical Image Harmonization Using Deep Learning Based Canonical Mapping: Toward Robust and Generalizable Learning in Imaging. arXiv [eess.IV] 2020.
46. Modanwal, G.; Vellal, A.; Mazurowski, M.A. Normalization of Breast MRIs Using Cycle-Consistent Generative Adversarial Networks. arXiv [eess.IV] 2019.
47. Dewey, B.E.; Zhao, C.; Reinhold, J.C.; Carass, A.; Fitzgerald, K.C.; Sotirchos, E.S.; Saidha, S.; Oh, J.; Pham, D.L.; Calabresi, P.A.; et al. DeepHarmony: A Deep Learning Approach to Contrast Harmonization across Scanner Changes. Magn. Reson. Imaging 2019, 64, 160–170, doi:10.1016/j.mri.2019.05.041.

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$minimum = \min(\mathbf{X})$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2$$

$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_s} \mathbf{P}(i,j)\right)^2}{N_p}$$

$$correlation = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) ij - \mu_x \mu_y}{\sigma_x(i) \sigma_y(j)}$$

$$MAD = \frac{1}{N_p} \sum_{i=1}^{N_p} |\mathbf{X}(i) - \bar{X}|$$

$$minimum = \min(\mathbf{X})$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_s} \mathbf{P}\right.}{N_z}$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2$$

Chapter 5

CT Reconstruction Kernels and the Effect of Pre- and Post-Processing on the Reproducibility of Handcrafted Radiomic Features

T. Refaee, Z. Salahuddin [†], Y. Widaatalla [†], S. Primakov, H. C. Woodruff, R. Hustinx, F.M. Mottaghy, A. Ibrahim [‡] and P. Lambin [‡]

[†] Authors contributed equally.
[‡] Senior authors contributed equally.

Adapted from:
J. Pers. Med. 2022, 12,
https://doi.org/10.3390/jpm12040553

Abstract

Handcrafted radiomics features (HRFs) are quantitative features extracted from medical images to decode biological information to improve clinical decision making. Despite the potential of the field, limitations have been identified. The most important identified limitation, currently, is the sensitivity of HRF to variations in image acquisition and reconstruction parameters. In this study, we investigated the use of Reconstruction Kernel Normalization (RKN) and ComBat harmonization to improve the reproducibility of HRFs across scans acquired with different reconstruction kernels. A set of phantom scans (n = 28) acquired on five different scanner models was analyzed. HRFs were extracted from the original scans, and scans were harmonized using the RKN method. ComBat harmonization was applied on both sets of HRFs. The reproducibility of HRFs was assessed using the concordance correlation coefficient. The difference in the number of reproducible HRFs in each scenario was assessed using McNemar's test. The majority of HRFs were found to be sensitive to variations in the reconstruction kernels, and only six HRFs were found to be robust with respect to variations in reconstruction kernels. The use of RKN resulted in a significant increment in the number of reproducible HRFs in 19 out of the 67 investigated scenarios (28.4%), while the ComBat technique resulted in a significant increment in 36 (53.7%) scenarios. The combination of methods resulted in a significant increment in 53 (79.1%) scenarios compared to the HRFs extracted from original images. Since the benefit of applying the harmonization methods depended on the data being harmonized, reproducibility analysis is recommended before performing radiomics analysis. For future radiomics studies incorporating images acquired with similar image acquisition and reconstruction parameters, except for the reconstruction kernels, we recommend the systematic use of the pre- and post-processing approaches (respectively, RKN and ComBat).

Keywords: radiomics reproducibility; reconstruction kernel; ComBat harmonization; image harmonization

Introduction

Recent decades have witnessed an exponentially increasing number of studies investigating the potential of quantitative imaging features to extract additional information from medical images not detectable by human eyes [1,2]. Handcrafted radiomics refers to the high-throughput extraction of quantitative imaging features from medical images to decode biologic information [3,4] and, today, more than 5000 studies can be returned on the PubMed database using "radiomics" as a search word. The handcrafted radiomics approach "involves manual segmentation of the region of interest (eg, the tumor) on medical imaging and extraction of thousands of human-defined and curated quantitative features from the region of interest" [5].

The hypothesis in radiomics studies is that handcrafted radiomic features (HRFs) can be used singularly or collectively as clinical biomarkers [3]. Many studies have investigated and reported on the potential of HRFs to predict clinical endpoints, such as overall survival [6–8], tissue histology [9–13] and response to therapy [14,15]. These studies highlighted the potential of such approaches to be applied in clinical settings, since they could present non-invasive, reliable, readily available and cost-effective alternatives to current invasive clinical procedures, such as tissue biopsies. Moreover, with proper application, radiomics could provide reproducible predictions, which are quantitative and less dependent on the subjective interpretation of medical examinations [16,17].

With the development of handcrafted radiomics as a research field, the limitations the field faces have been increasingly investigated during recent years [4,18]. The most important identified limitation currently is the sensitivity of HRFs to variations in image acquisition and reconstruction parameters [19–24]. For an HRF to be used as a clinical biomarker (solely or in combination with other HRFs), it has to be reproducible across different imaging parameters for generalization purposes [24]. However, many studies have reported on the sensitivity of HRFs to variations in time (test–retest) [25–29] and to variations in imaging acquisition and reconstruction parameters [30–37]. Studies have also reported that the degree of variation in a single acquisition or reconstruction parameter affects the reproducibility of HRFs variably [31,34]. A number of studies have reported the significant effects of variations in reconstruction kernels on the reproducibility of HRFs [20,38].

Different methods have been investigated to address the issue of reproducibility of HRFs across scans acquired differently. ComBat harmonization [39] is one of the post-processing methods that have recently been extensively investigated in radiomics analyses [40–42]. ComBat harmonization is a method that was developed for removing batch effects—attributed to the use of different machinery—from gene expression arrays. A number of studies have reported on the applicability of ComBat harmonization in different scenarios, such as scans

acquired with varying degrees of differences in CT image acquisition and reconstruction parameters, scans acquired with a single variation in an image reconstruction parameter (in-plane resolution) and scans of different contrast-enhancement phases [31,35,43,44]. These studies reported that the performance of ComBat in radiomics analyses is dependent on the variations in the data being harmonized. A number of studies have also investigated the potential of ComBat in different scenarios [45–48]. However, the potential of ComBat to remove batch effects attributed solely to the variations in the reconstruction kernel has yet to be thoroughly investigated. Other investigated methods include pre-processing of the images to minimize effects due to differences in slice thickness, reconstruction with convolutional kernels, etc. Normalization of chest CT data minimized the variability that resulted from different reconstruction kernels [49]. The authors developed a method that targeted reducing the variations in the quantification of emphysema by normalizing the reconstruction kernel (Reconstruction Kernel Normalization—RKN). The CT scans obtained from different scanners that were reconstructed with varying kernels showed reduced variability in emphysema quantification after the proposed iterative normalization. However, the effect of this normalization method on the reproducibility of HRFs has not been investigated.

In this study, we hypothesize that the use of RKN and ComBat could improve the reproducibility of HRFs across scans acquired with different reconstruction kernels depending on the variations in the data being analyzed and/or harmonized. We further hypothesize that the combination of both methods (RKN and ComBat) would give superior results in terms of "number of reproducible HRFs" compared to no or only one harmonization method. Given that variations in the convolution kernel impact the reproducibility of HRFs the most, we investigate the reproducibility of HRFs extracted from phantom CT scans acquired with different reconstruction kernels on different imaging vendors. We also investigate the potential of ComBat harmonization, RKN and the combination of both methods to reduce the variations in HRF values attributed to differences solely in the reconstruction kernels of the original scans.

Materials and Methods

Imaging Data

The phantom data used in the study were obtained from the public Credence Cartridge Radiomics (CCR) phantom dataset [50] from the Cancer Imaging Archive site (TCIA.org) [51]. A total of 251 scans were acquired using different scanners, acquisition and reconstruction parameters. For this study, we included scans that were acquired using the same imaging acquisition and reconstruction parameters, except for the convolution kernel. After applying the inclusion criteria, 28 scans from five different scanner models were used in this study (Table 1).

Table 1. Acquisition and reconstruction parameters for the imaging dataset.

Manufacturer	Scanner Model	Number of Scans	X-Ray Tube Current (kV)	Convolution Kernels	Slice Thickness (mm)	Pixel Spacing (mm^2)
GE	Discovery STE	5	120	Standard, Detail, Edge, Soft, Lung	1.25	0.49 × 0.49
Philips	Brilliance 64	4	120	A, B, C, L	1.50	0.49 × 0.49
Siemens	Sensation 40	6	120	B10f, B20f, B31f, B50f, B60f, B70f	1.50	0.49 × 0.49
	Sensation 64	7	120	B10f, B20f, B30f, B31f, B50f, B60f, B70f	1.50	0.49 × 0.49
	SOMATOM Definition AS	6	120	I26f, I30f, I40f, I44f, I50f, I70f	1.50	0.49 × 0.49

Volume of Interest and HRFs Extraction

Each layer of the phantom was segmented as a single volume of interest (VOI), with the dimensions 8 × 8 × 2 cm^3. A total of 10 VOIs were segmented per scan, resulting in a total of 280 VOIs. HRFs were extracted using the open source PyRadiomics software version 2.2.0 [52]. HRFs were extracted at two different stages: directly from the original scans; and after image pre-processing. Image intensities were binned in all of the three scenarios with a binwidth of 25 Hounsfield units (HUs) to reduce noise levels and texture matrix sizes and the amount of computational power needed. No other image pre-processing was applied in any of the scenarios. Extracted HRFs were HU intensity features and texture features of five matrices: gray-level co-occurrence (GLCM); gray-level run-length (GLRLM); gray-level size zone (GLSZM); gray-level dependence (GLDM); and neighborhood gray-tone difference (NGTDM) matrices. A more detailed description of PyRadiomics HRFs can be found online at: https://pyradiomics.readthedocs.io/en/latest/features.html (accessed on 13 October 2021).

Reconstruction Kernel Normalization

The CT scan I_o is decomposed into a series of frequency components F^i. Image I_o is convoluted with the Gaussian filter at σ_i scale (σ_i = 0, 1, 2, 4, 8, 16) to get a filtered image $L_{\sigma i}$. The frequency component for i = 0, 1, 2, 3, 4 is given by $F^{i+1} = L_{\sigma+1} - L_{\sigma i+1}$ and for i = 5 it is given by $F^{i+1} = L_{\sigma i}$. The normalized image I_N is obtained by $I_N = F^6 + \sum_{i=1}^{5} \lambda_i \cdot F^5$. λ_i is given by $\frac{r_i}{e_i}$, where r_i and e_i are the standard deviations of the intensity values in the band F^i of the reference image and image I_o, respectively. This process is repeated until λ_i is within the range [0.95, 1.05]. This method was proposed for reducing the effects of varying reconstruction kernels for emphysema quantification in chest CT scans [49]. We investigated the effect of applying this normalization method on feature reproducibility.

Image Pre-Processing and HRF Post-Processing

Four scenarios were analyzed in this study (Figure 1): (i) HRFs extracted from original images; (ii) HRFs extracted from pre-processed scans with the method described in 2.3; (iii)

HRFs extracted from original images and harmonized with ComBat; and (iv) the combination of both methods. In scenario (ii), image pre-processing was performed using the method previously described in [49]. Each set of images (n = 5) was normalized to a reference scan from the set. HRFs were extracted following image pre-processing. In scenario (iii), ComBat harmonization was applied on HRFs extracted from the original scans without pre-processing. ComBat harmonization in radiomics has been previously described [43]. In scenario (iv), HRFs were extracted from images normalized with the RKN method and harmonized using ComBat harmonization.

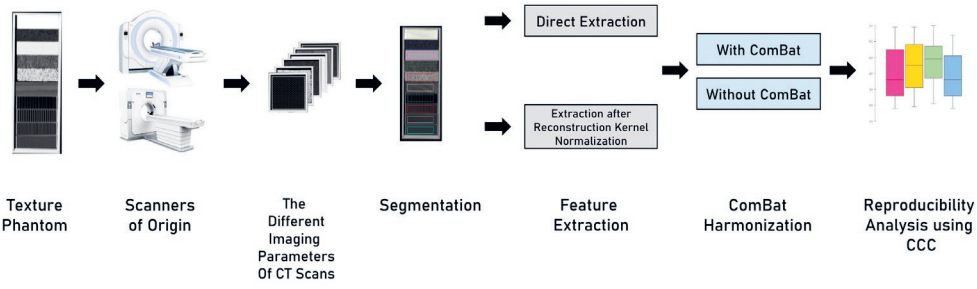


Figure 1. The study workflow.

Statistical Analysis

All statistical analyses were performed using R [53] on RStudio (V 3.6.3) [54]. For each scanner model, scans were compared in a pair-wise manner. The concordance correlation coefficient (CCC) was used to assess the reproducibility of HRFs across different pairs [55] (epiR package V. 2.0.26) [56]. The CCC assesses the agreement in the value and rank for each HRF across the pairwise scenarios. HRFs with CCC > 0.9 were considered reproducible in a given scenario. The CCC was calculated in each of the investigated scenarios described in Section 2.4.

To assess the statistical significance of the differences in the number of reproducible HRFs in each scenario, the McNemar test was used [57]**.** The McNemar test is used to assess whether marginal frequencies are equal before and after an intervention. In this study, we calculated McNemar's p-values using the HRFs extracted from the original images and after RKN, ComBat, and the combination of both. We also calculated the p-values among the methods, as well as the p-values for each method compared to the combination of methods. For each pair, the difference in the number of reproducible HRFs was labeled "significant" or "not significant" depending on the p-value.

Results

The Effect of Differences in Convolution Kernels on the Reproducibility of HRFs

The Pyradiomics toolbox provides a set of 91 original HRFs from each VOI. These HRFs are divided into First Order Statistics (*n* = 18), GLCM (*n* = 22), GLRLM (*n* = 16); GLSZM (*n* = 16), NGTDM (*n* = 5) and GLDM (*n* = 14). The number of reproducible HRFs varied across kernels and scanner models. Six HRFs were found to be robust to changes in convolution kernels across all scanner models: "Firstorder_10Percentile", "Firstorder_Energy", "Firstorder_Mean", "Firstorder_Median", "Firstorder_RootMeanSquared" and "Firstorder_TotalEnergy". On the Discovery STE scanner model (GE Medical Systems), the number of reproducible HRFs varied between 6 (6.59%) and 78 (85.71%). The greatest number of reproducible HRFs was observed across scans acquired with Detailed and Standard kernels (Figure 2).

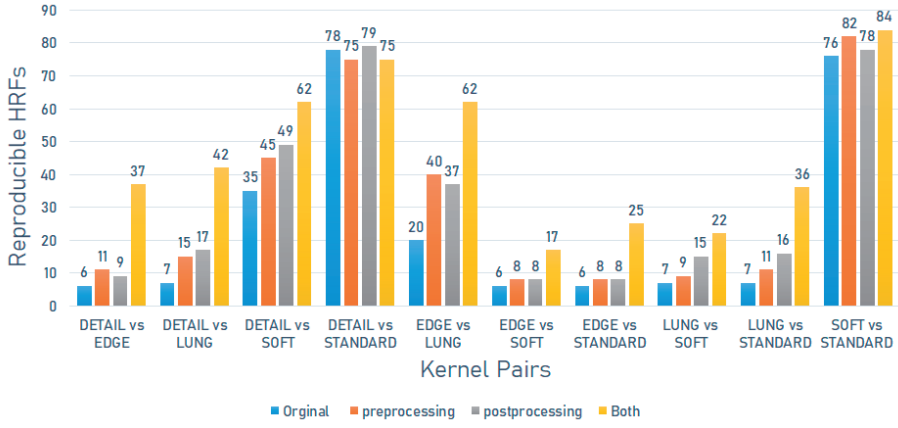


Figure 2. The number of reproducible HRFs across different kernels on the Discovery STE scanner model.

On the Sensation 40 scanner model (Siemens), the number of reproducible HRFs varied between 6 (6.59%) and 91 (100%). The greatest number of reproducible HRFs was observed across scans acquired with B60f and B70f kernels (Figure 3).

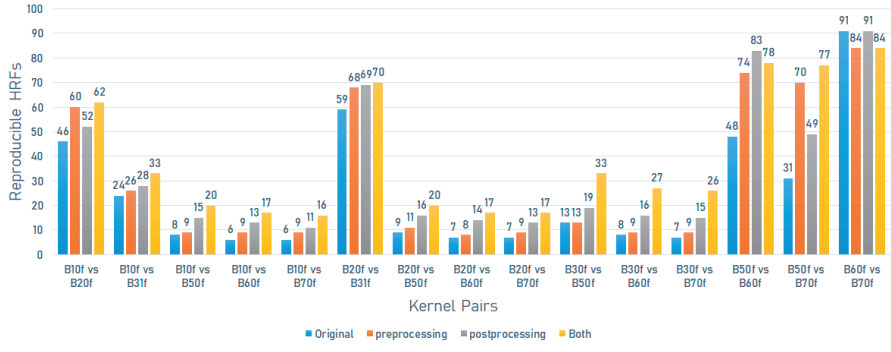


Figure 3. The number of reproducible HRFs across different kernels on the Sensation 40 scanner model.

On the SOMATOM definition scanner model (Siemens), the number of reproducible HRFs varied between 6 (6.59%) and 65 (71.4%). The greatest number of reproducible HRFs was observed across scans acquired with I44f and I50f kernels (Figure 4).

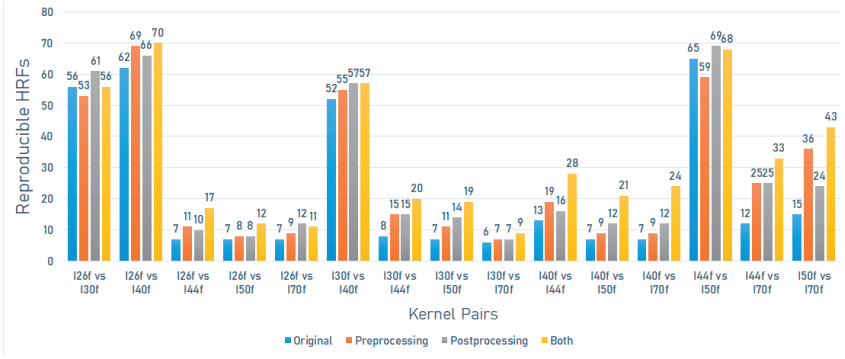


Figure 4. The number of reproducible HRFs across different kernels on the SOMATOM Definition scanner model.

On the Sensation 64 scanner model (Siemens), the number of reproducible HRFs varied between 6 (6.59%) and 91 (100%). The greatest number of reproducible HRFs was observed across scans acquired with B60f and B70f kernels (Figure 5).

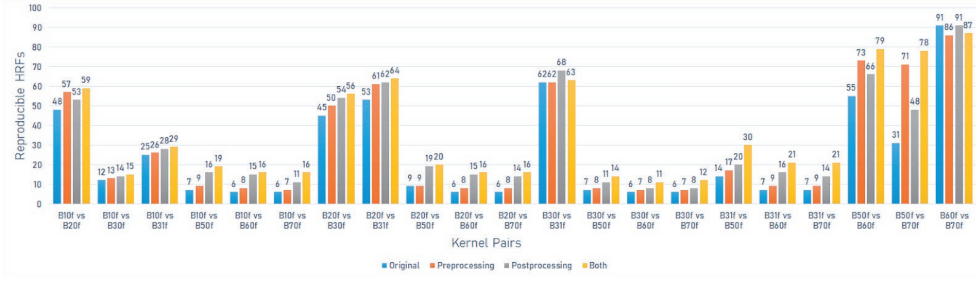


Figure 5. The number of reproducible HRFs across different kernels on the Sensation 64 scanner model.

On the Brilliance 64 scanner model (Philips), the number of reproducible HRFs varied between 14 (15.4%) and 48 (52.7%). The greatest number of reproducible HRFs was observed across scans acquired with A and B kernels (Figure 6).

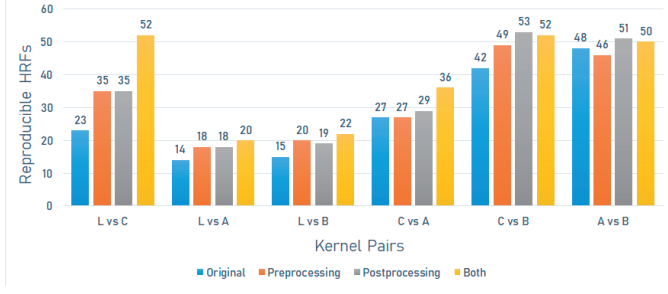


Figure 6. The number of reproducible HRFs across different kernels on the Brilliance 64 scanner model.

The Effects of Pre- and Post-Processing

Reconstruction Kernel Normalization (RKN)

The number of HRFs that became reproducible following the application of the described method varied with the variations in kernels being harmonized and the scanner model used. In most of the investigated scenarios (58 out of 67; 86.6%), the use of this method has resulted in an increment in the number of reproducible HRFs. However, only 19 scenarios (28.4%) showed statistically significant increments. In a number of scenarios (6 out of the analyzed 67 scenarios (9%)), there was a net loss in the number of reproducible HRFs compared to the original, 2 (3%) of which were statistically significant (Figures 2–6). In three (4.5%) scenarios, there was no difference between the number of reproducible HRFs extracted from the original and the normalized images.

On the Discovery STE scanner model (GE Medical Systems), the number of reproducible HRFs extracted from the scans after image pre-processing varied between 8 (8.8%) and 82 (90.1%). The greatest increment in the number of reproducible HRFs compared to the original images was observed across scans acquired with Edge and Lung kernels (Figure 2).

On the Sensation 40 scanner model (Siemens), the number of reproducible HRFs extracted from the scans after image pre-processing varied between 8 (8.8%) and 84 (92.3%). In this scenario, the highest number of reproducible HRFs decreased compared to those extracted from the original images for the scans acquired with B60f and B70f. The greatest increment in the number of reproducible HRFs compared to the original images was observed across scans acquired with B50f and B70f kernels (Figure 3).

On the SOMATOM definition scanner model (Siemens), the number of reproducible HRFs extracted from the scans after image pre-processing varied between 7 (7.7%) and 69 (75.8%). The greatest increment in the number of reproducible HRFs compared to the original images was observed across scans acquired with I50f and I70f kernels (Figure 4).

On the Sensation 64 scanner model (Siemens), the number of reproducible HRFs extracted from the scans after image pre-processing varied between 7 (7.7%) and 86 (94.5%). In this scenario, the highest number of reproducible HRFs decreased compared to those extracted from the original images (B60f vs. B70f) (Figure 5).

On the Brilliance 64 scanner model (Philips), the number of reproducible HRFs extracted from the scans after image pre-processing varied between 18 (19.8%) and 49 (53.8%). The greatest increment in the number of reproducible HRFs compared to the original images was observed across scans acquired with L and C kernels (Figure 6).

ComBat Harmonization

In 65 out of the 67 investigated scenarios (97%), there was a net increase in the number of reproducible HRFs compared to the original, with 36 (53.7%) scenarios witnessing significant statistical increments. In two scenarios, the same number of reproducible HRFs was found before and after ComBat harmonization. In 46 (68.7%) scenarios, ComBat harmonization outperformed the RKN method, 17 (25.4%) of which were statistically significant. In 13 (19.4%) scenarios, the RKN method outperformed ComBat harmonization, 5 (7.5%) of which were statistically significant increments.

On the Discovery STE scanner model (GE Medical Systems), the number of reproducible HRFs extracted from the scans after ComBat harmonization varied between 9 (9.9%) and 79 (86.8%). The greatest increment in the number of reproducible HRFs compared to the original images was observed across scans acquired with Edge and Lung kernels (Figure 2).

On the Sensation 40 scanner model (Siemens), the number of reproducible HRFs extracted from the scans after ComBat harmonization varied between 11 (12.1%) and 69 (75.8%). The greatest increment in the number of reproducible HRFs compared to the original images was observed across scans acquired with B50f and B60f kernels (Figure 3).

On the SOMATOM definition scanner model (Siemens), the number of reproducible HRFs extracted from the scans after ComBat harmonization pre-processing varied between 7 (7.7%) and 69 (75.8%). The greatest increment in the number of reproducible HRFs compared to the original images was observed across scans acquired with I44f and I70f kernels (Figure 4).

On the Sensation 64 scanner model (Siemens), the number of reproducible HRFs extracted from the scans after ComBat harmonization varied between 8 (8.8%) and 91 (100%). The greatest increment in the number of reproducible HRFs compared to the original images was observed across scans acquired with B50f and B70f kernels (Figure 5).

On the Brilliance 64 scanner model (Philips), the number of reproducible HRFs extracted from the scans after ComBat harmonization varied between 18 (19.8%) and 53 (58.8%). The greatest increment in the number of reproducible HRFs compared to the original images was observed across scans acquired with L and C kernels (Figure 6).

The Combination of Pre- and Post-Processing

In 63 (95.5%) out of the 67 investigated scenarios, there was a net increase in the number of reproducible HRFs compared to the original, 53 (79.1%) of which were statistically significant. Three (4.5%) showed a lower number of reproducible HRFs, with one (1.5%) scenario showing significantly fewer ($p < 0.05$). The same number of reproducible HRFs

was observed in one (1.5%) scenario. In 66 (98.5%) scenarios, the combination of methods outperformed the RKN method, with 42 (62.7%) being significantly higher. The same number of reproducible HRFs was observed in one (1.5%) scenario. With regards to ComBat harmonization, the combination of methods resulted in a higher number of reproducible HRFs in 56 (83.6%) scenarios, 27 (40.3%) of which were statistically significant. A higher number of reproducible HRFs was obtained using only ComBat harmonization in 10 (14.9%) scenarios, only one (1.5%) of which was statistically significant. The same number of reproducible HRFs was observed in one (1.5%) scenario.

On the Discovery STE scanner model (GE Medical Systems), the number of reproducible HRFs extracted from the normalized scans after ComBat harmonization varied between 17 (18.7%) and 84 (92.3%). The greatest increment in the number of reproducible HRFs compared to the original images was observed across scans acquired with Edge and Lung kernels (Figure 2).

On the Sensation 40 scanner model (Siemens), the number of reproducible HRFs extracted from the normalized scans after ComBat harmonization varied between 16 (17.6%) and 84 (92.3%). The greatest increment in the number of reproducible HRFs compared to the original images was observed across scans acquired with B50f and B70f kernels (Figure 3).

On the SOMATOM definition scanner model (Siemens), the number of reproducible HRFs extracted from the normalized scans after ComBat harmonization pre-processing varied between 9 (9.9%) and 70 (77%). The greatest increment in the number of reproducible HRFs compared to the original images was observed across scans acquired with I50f and I70f kernels (Figure 4).

On the Sensation 64 scanner model (Siemens), the number of reproducible HRFs extracted from the normalized scans after ComBat harmonization varied between 11 (12.1%) and 87 (95.7%). The greatest increment in the number of reproducible HRFs compared to the original images was observed across scans acquired with B50f and B70f kernels (Figure 5).

On the Brilliance 64 scanner model (Philips), the number of reproducible HRFs extracted from the normalized scans after ComBat harmonization varied between 20 (22%) and 52 (57.2%). The greatest increment in the number of reproducible HRFs compared to the original images was observed across scans acquired with L and C kernels (Figure 6).

Discussion

In this study, we analyzed the effects of difference in convolution kernels on five different scanner models, when all other CT acquisition and reconstruction parameters were fixed on a phantom dataset. We further investigated the ability of an image pre-processing (iterative normalization by frequency decomposition) method, and an HRF post-processing harmonization (using ComBat harmonization) method. Our results showed significant differences in the number of reproducible HRFs across the investigated scenarios. Scans reconstructed with similar convolution kernels showed a higher number of reproducible HRFs compared to scans reconstructed with significantly different convolution kernels. Similarly, the performance of both harmonization methods investigated varied with the differences in convolution kernels of the scans being harmonized.

Siemens scanner models (Sensation 40 and 64) have shown the reproducibility of all HRFs across the scans acquired with the higher end of convolution kernels (B60 and B70). Convolution kernels at the opposite end of the spectrum (for example, B10 and B70 on Siemens scanners) have shown the lowest number of reproducible HRFs. As such, our results are in line with previous studies that reported that the reproducibility of HRFs can be significantly affected by variations in convolution kernels [38,58-60].

The use of the RKN method on our dataset has resulted in a range of effects on the number of reproducible features, from negative to neutral to positive, depending on the scans being compared. We have observed a significant increase in the number of reproducible HRFs in most scenarios and a decrease in the number of reproducible HRFs in some other scenarios. This could be justified by the possibility that the analyzed data in this study included a wider range of convolution kernels than those used to develop the method.

The application of ComBat harmonization resulted in a higher number of reproducible HRFs compared to those before harmonization in almost all of the investigated scenarios, which is in line with previous reports [43,44,61]. Moreover, on average, ComBat harmonization outperformed the image pre-processing method. The performance of ComBat further depended on the differences in the convolution kernels of the scans being harmonized. In general, the number of reproducible HRFs after ComBat harmonization followed a similar pattern to that of the number of reproducible HRFs before post-processing. These findings are in line with previous studies that investigated the applicability of ComBat harmonization in radiomics analyses [31,34]. The results add to the evidence on the need for reproducibility analyses in radiomics studies, including scans acquired differently, as well as the need for radiomics-specific harmonization methods.

The combination of RKN and ComBat harmonization methods resulted in a higher number of reproducible HRFs across the majority of the investigated scenarios. This indicates that each method could be addressing the reproducibility of HRFs in different manners, with their having been shown to be complementary to each other in many of the investigated scenarios. Nevertheless, the combination resulted in a lower number of reproducible HRFs in an appreciated percentage of scenarios compared to ComBat harmonization only. This suggests the need for reproducibility analysis before applying harmonization methods in radiomics analyses.

We identified six HRFs that were robust with respect to variations in convolution kernels across all the investigated scenarios. These HRFs were first-order statistics, and their robustness could be justified by the standardization of HUs across scanners. However, the majority of texture HRFs were sensitive to the majority of variations in convolution kernels. Clear to the eye, the standardization of image acquisition and reconstruction parameters would be the cornerstone for the translation of radiomic signatures to clinical practice. The findings of this study, and previous experiments, have shown that the reproducibility of HRFs significantly depends on imaging acquisition and reconstruction parameters. Therefore, reproducibility analysis is needed for a proper understanding of their performance or generalizability [19]. Another potential solution would be the development of radiomic signatures specific to a set of imaging acquisition and reconstruction parameters. However, this solution limits the generalizability of radiomic signatures.

While we tried to analyze all the kernels used in clinical practice, we were limited by the available data. However, the results have shown a similar pattern across different scanner models. Future studies that include a wider spectrum of convolution kernels are recommended. Furthermore, we limited our analyses to the original HRFs as they are commonly standardized across radiomics platforms. Detailed full HRF reproducibility analysis could be beneficial for specific tasks. Furthermore, the analysis was performed on a phantom dataset that was designed to mimic human tissues. However, it only gives an idea about the reproducibility of HRFs in the given scenarios, and similar analysis is needed for patient datasets to gain a full understanding. The potential of other harmonization methods, for example, dynamic range limitation [62], could also be explored in future studies. Additionally, the sensitivity of HRFs to variations in segmentations could not be assessed in this study, due to the use of automated segmentations.

Conclusions

The reproducibility of the majority of HRFs depended on the variations in reconstruction kernels in the data being analyzed. Six HRFs were found to be reproducible across all investigated scenarios. Radiomics analysis of scans acquired with different reconstruction kernels is not recommended in the absence of reproducibility analysis. We recommend the systematic use of RKN and ComBat harmonization in future radiomics studies, including images acquired similarly except for the reconstruction kernel. Nevertheless, their application should follow a reproducibility analysis to identify the set of reproducible HRFs after harmonization. HRF-specific harmonization methods remain necessities in the field of radiomics.

References

1. Walsh, S.; de Jong, E.E.C.; van Timmeren, J.E.; Ibrahim, A.; Compter, I.; Peerlings, J.; Sanduleanu, S.; Refaee, T.; Keek, S.; Larue, R.T.H.M.; et al. Decision support systems in oncology. *JCO Clin. Cancer Inform.* **2019**, *3*, 1–9, https://doi.org/10.1200/CCI.18.00001.
2. Gillies, R.J.; Kinahan, P.E.; Hricak, H. Radiomics: Images are more than pictures, they are data. *Radiology* **2016**, *278*, 563–577, https://doi.org/10.1148/radiol.2015151169.
3. Lambin, P.; Rios-Velazquez, E.; Leijenaar, R.; Carvalho, S.; van Stiphout, R.G.P.M.; Granton, P.; Zegers, C.M.L.; Gillies, R.; Boellard, R.; Dekker, A.; et al. Radiomics: Extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* **2012**, *48*, 441–446, https://doi.org/10.1016/j.ejca.2011.11.036.
4. Yip, S.S.F.; Aerts, H.J.W.L. Applications and limitations of radiomics. *Phys. Med. Biol.* **2016**, *61*, R150–R166, https://doi.org/10.1088/0031-9155/61/13/R150.
5. Hosny, A.; Aerts, H.J.; Mak, R.H. Handcrafted versus deep learning radiomics for prediction of cancer therapy response. *Lancet Digit. Health* **2019**, *1*, e106–e107.
6. Aerts, H.J.W.L.; Velazquez, E.R.; Leijenaar, R.T.H.; Parmar, C.; Grossmann, P.; Carvalho, S.; Bussink, J.; Monshouwer, R.; Haibe-Kains, B.; Rietveld, D.; et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **2014**, *5*, 4006, https://doi.org/10.1038/ncomms5006.
7. Bae, S.; Choi, Y.S.; Ahn, S.S.; Chang, J.H.; Kang, S.-G.; Kim, E.H.; Kim, S.H.; Lee, S.-K. Radiomic MRI phenotyping of glioblastoma: Improving survival prediction. *Radiology* **2018**, *289*, 797–806, https://doi.org/10.1148/radiol.2018180200.
8. Oikonomou, A.; Khalvati, F.; Tyrrell, P.N.; Haider, M.A.; Tarique, U.; Jimenez-Juan, L.; Tjong, M.C.; Poon, I.; Eilaghi, A.; Ehrlich, L.; et al. Radiomics analysis at PET/CT contributes to prognosis of recurrence and survival in lung cancer treated with stereotactic body radiotherapy. *Sci. Rep.* **2018**, *8*, 4003, https://doi.org/10.1038/s41598-018-22357-y.
9. Wu, W.; Parmar, C.; Grossmann, P.; Quackenbush, J.; Lambin, P.; Bussink, J.; Mak, R.; Aerts, H.J.W.L. Exploratory study to identify radiomics classifiers for lung cancer histology. *Front. Oncol.* **2016**, *6*, 71, https://doi.org/10.3389/fonc.2016.00071.
10. Blüthgen, C.; Patella, M.; Euler, A.; Baessler, B.; Martini, K.; von Spiczak, J.; Schneiter, D.; Opitz, I.; Frauenfelder, T. Computed tomography radiomics for the prediction of thymic epithelial tumor histology, TNM stage and myasthenia gravis. *PLoS ONE* **2021**, *16*, e0261401, https://doi.org/10.1371/journal.pone.0261401.
11. Linning, E.; Lu, L.; Li, L.; Yang, H.; Schwartz, L.H.; Zhao, B. Radiomics for Classification of Lung Cancer Histological Subtypes Based on Nonenhanced Computed Tomography. *Acad. Radiol.* **2019**, *26*, 1245–1252, https://doi.org/10.1016/j.acra.2018.10.013.
12. Stefan, P.-A.; Puscas, M.E.; Csuak, C.; Lebovici, A.; Petresc, B.; Lupean, R.; Mihu, C.M. The utility of texture-based classification of different types of ascites on magnetic resonance. *J. BUON* **2020**, *25*, 1237–1244.

13. Csutak, C.; Ștefan, P.-A.; Lupean, R.-A.; Lenghel, L.M.; Mihu, C.M.; Lebovici, A. Computed tomography in the diagnosis of intraperitoneal effusions: The role of texture analysis. *Bosn. J. Basic Med. Sci.* **2021**, *21*, 488–494, https://doi.org/10.17305/bjbms.2020.5048.
14. Horvat, N.; Veeraraghavan, H.; Khan, M.; Blazic, I.; Zheng, J.; Capanu, M.; Sala, E.; Garcia-Aguilar, J.; Gollub, M.J.; Petkovska, I. MR imaging of rectal cancer: Radiomics analysis to assess treatment response after neoadjuvant therapy. *Radiology* **2018**, *287*, 833–843, https://doi.org/10.1148/radiol.2018172300.
15. Tharmalingam, H.; Tsang, Y.M.; Alonzi, R.; Beasley, W.; Taylor, N.J.; McWilliam, A.; Padhani, A.; Choudhury, A.; Hoskin, P.J. Changes in magnetic resonance imaging radiomic features in response to androgen deprivation therapy in patients with intermediate- and high-risk prostate cancer. *Clin. Oncol.* **2022**, https://doi.org/10.1016/j.clon.2021.12.020.
16. Ștefan, P.-A.; Lupean, R.-A.; Mihu, C.M.; Lebovici, A.; Oancea, M.D.; Hîțu, L.; Duma, D.; Csutak, C. Ultrasonography in the diagnosis of adnexal lesions: The role of texture analysis. *Diagnostics* **2021**, *11*, 812, https://doi.org/10.3390/diagnostics11050812.
17. Ștefan, R.-A.; Ștefan, P.-A.; Mihu, C.M.; Csutak, C.; Melincovici, C.S.; Crivii, C.B.; Maluțan, A.M.; Hîțu, L.; Lebovici, A. Ultrasonography in the differentiation of endometriomas from hemorrhagic ovarian cysts: The role of texture analysis. *J. Pers. Med.* **2021**, *11*, 611, https://doi.org/10.3390/jpm11070611.
18. Lohmann, P.; Bousabarah, K.; Hoevels, M.; Treuer, H. Radiomics in radiation oncology—Basics, methods, and limitations. *Strahlenther. Onkol.* **2020**, *196*, 848–855, https://doi.org/10.1007/s00066-020-01663-3.
19. Ibrahim, A.; Primakov, S.; Beuque, M.; Woodruff, H.C.; Halilaj, I.; Wu, G.; Refaee, T.; Granzier, R.; Widaatalla, Y.; Hustinx, R.; et al. Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework. *Methods* **2020**, *188*, 20–29, https://doi.org/10.1016/j.ymeth.2020.05.022.
20. Mali, S.A.; Ibrahim, A.; Woodruff, H.C.; Andrearczyk, V.; Müller, H.; Primakov, S.; Salahuddin, Z.; Chatterjee, A.; Lambin, P. Making radiomics more reproducible across scanner and imaging protocol variations: A review of harmonization methods. *J. Pers. Med.* **2021**, *11*, 842, https://doi.org/10.3390/jpm11090842.
21. Midya, A.; Chakraborty, J.; Gönen, M.; Do, R.K.G.; Simpson, A.L. Influence of CT acquisition and reconstruction parameters on radiomic feature reproducibility. *J. Med. Imaging* **2018**, *5*, 011020, https://doi.org/10.1117/1.JMI.5.1.011020.
22. Reiazi, R.; Abbas, E.; Famiyeh, P.; Rezaie, A.; Kwan, J.Y.Y.; Patel, T.; Bratman, S.V.; Tadic, T.; Liu, F.-F.; Haibe-Kains, B. The impact of the variation of imaging parameters on the robustness of computed tomography radiomic features: A review. *Comput. Biol. Med.* **2021**, *133*, 104400, https://doi.org/10.1016/j.compbiomed.2021.104400.
23. Espinasse, M.; Pitre-Champagnat, S.; Charmettant, B.; Bidault, F.; Volk, A.; Balleyguier, C.; Lassau, N.; Caramella, C. CT texture analysis challenges: Influence of acquisition and reconstruction parameters: A comprehensive review. *Diagnostics* **2020**, *10*, 258, https://doi.org/10.3390/diagnostics10050258.

24. Zhao, B. Understanding sources of variation to improve the reproducibility of radiomics. *Front. Oncol.* **2021**, *11*, 633176, https://doi.org/10.3389/fonc.2021.633176.
25. Granzier, R.W.Y.; Ibrahim, A.; Primakov, S.; Keek, S.A.; Halilaj, I.; Zwanenburg, A.; Engelen, S.M.E.; Lobbes, M.B.I.; Lambin, P.; Woodruff, H.C.; et al. Test-retest data for the assessment of breast MRI radiomic feature repeatability. *J. Magn. Reson. Imaging* **2021**, https://doi.org/10.1002/jmri.28027.
26. Shiri, I.; Abdollahi, H.; Shaysteh, S.; Mahdavi, S.R. Test-retest reproducibility and robustness analysis of recurrent glioblastoma MRI radiomics texture features. *Iran. J. Radiol.* **2017**, *5*.
27. Peerlings, J.; Woodruff, H.C.; Winfield, J.M.; Ibrahim, A.; van Beers, B.E.; Heerschap, A.; Jackson, A.; Wildberger, J.E.; Mottaghy, F.M.; DeSouza, N.M.; et al. Stability of radiomics features in apparent diffusion coefficient maps from a multi-centre test-retest trial. *Sci. Rep.* **2019**, *9*, 4800, https://doi.org/10.1038/s41598-019-41344-5.
28. Pfaehler, E.; Beukinga, R.J.; de Jong, J.R.; Slart, R.H.J.A.; Slump, C.H.; Dierckx, R.A.J.O.; Boellaard, R. Repeatability of 18 F-FDG PET radiomic features: A phantom study to explore sensitivity to image reconstruction settings, noise, and delineation method. *Med. Phys.* **2019**, *46*, 665–678, https://doi.org/10.1002/mp.13322.
29. Prayer, F.; Hofmanninger, J.; Weber, M.; Kifjak, D.; Willenpart, A.; Pan, J.; Röhrich, S.; Langs, G.; Prosch, H. Variability of computed tomography radiomics features of fibrosing interstitial lung disease: A test-retest study. *Methods* **2021**, *188*, 98–104, https://doi.org/10.1016/j.ymeth.2020.08.007.
30. Zhao, B.; Tan, Y.; Tsai, W.-Y.; Qi, J.; Xie, C.; Lu, L.; Schwartz, L.H. Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Sci. Rep.* **2016**, *6*, 23428, https://doi.org/10.1038/srep23428.
31. Ibrahim, A.; Refaee, T.; Primakov, S.; Barufaldi, B.; Acciavatti, R.J.; Granzier, R.W.Y.; Hustinx, R.; Mottaghy, F.M.; Woodruff, H.C.; Wildberger, J.E.; et al. The effects of in-plane spatial resolution on CT-based radiomic features' stability with and without ComBat harmonization. *Cancers* **2021**, *13*, 1848, https://doi.org/10.3390/cancers13081848.
32. Zhovannik, I.; Bussink, J.; Traverso, A.; Shi, Z.; Kalendralis, P.; Wee, L.; Dekker, A.; Fijten, R.; Monshouwer, R. Learning from scanners: Bias reduction and feature correction in radiomics. *Clin. Transl. Radiat. Oncol.* **2019**, *19*, 33–38, https://doi.org/10.1016/j.ctro.2019.07.003.
33. Ibrahim, A.; Primakov, S.; Barufaldi, B.; Acciavatti, R.J.; Granzier, R.W.Y.; Hustinx, R.; Mottaghy, F.M.; Woodruff, H.C.; Wildberger, J.E.; Lambin, P.; et al. Reply to Orlhac, F.; Buvat, I. Comment on "Ibrahim et Al. The Effects of In-Plane Spatial Resolution on CT-Based Radiomic Features' Stability with and without ComBat Harmonization. Cancers 2021, 13, 1848." Cancers 2021, 13, 3080.
34. Ibrahim, A.; Refaee, T.; Leijenaar, R.T.H.; Primakov, S.; Hustinx, R.; Mottaghy, F.M.; Woodruff, H.C.; Maidment, A.D.A.; Lambin, P. The application of a workflow integrating the variable reproducibility and harmonizability of radiomic features on a phantom dataset. *PLoS ONE* **2021**, *16*, e0251147, https://doi.org/10.1371/journal.pone.0251147.

35. Ibrahim, A.; Widaatalla, Y.; Refaee, T.; Primakov, S.; Miclea, R.L.; Öcal, O.; Fabritius, M.P.; Ingrisch, M.; Ricke, J.; Hustinx, R.; et al. Reproducibility of CT-based hepatocellular carcinoma radiomic features across different contrast imaging phases: A proof of concept on SORAMIC trial data. *Cancers* **2021**, *13*, 4638, https://doi.org/10.3390/cancers13184638.
36. Park, J.E.; Park, S.Y.; Kim, H.J.; Kim, H.S. Reproducibility and generalizability in radiomics modeling: Possible strategies in radiologic and statistical perspectives. *Korean J. Radiol.* **2019**, *20*, 1124–1137, https://doi.org/10.3348/kjr.2018.0070.
37. Meyer, M.; Ronald, J.; Vernuccio, F.; Nelson, R.C.; Ramirez-Giraldo, J.C.; Solomon, J.; Patel, B.N.; Samei, E.; Marin, D. Reproducibility of CT radiomic features within the same patient: Influence of radiation dose and CT reconstruction settings. *Radiology* **2019**, *293*, 583–591, https://doi.org/10.1148/radiol.2019190928.
38. Lu, L.; Ehmke, R.C.; Schwartz, L.H.; Zhao, B. Assessing agreement between radiomic features computed for multiple CT imaging settings. *PLoS ONE* **2016**, *11*, e0166550, https://doi.org/10.1371/journal.pone.0166550.
39. Johnson, W.E.; Li, C.; Rabinovic, A. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics* **2007**, *8*, 118–127, https://doi.org/10.1093/biostatistics/kxj037.
40. Ligero, M.; Jordi-Ollero, O.; Bernatowicz, K.; Garcia-Ruiz, A.; Delgado-Muñoz, E.; Leiva, D.; Mast, R.; Suarez, C.; Sala-Llonch, R.; Calvo, N.; et al. Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis. *Eur. Radiol.* **2021**, *31*, 1460–1470, https://doi.org/10.1007/s00330-020-07174-0.
41. Foy, J.J.; Al-Hallaq, H.A.; Grekoski, V.; Tran, T.; Guruvadoo, K.; Armato, S.G., III; Sensakovic, W.F. Harmonization of radiomic feature variability resulting from differences in CT image acquisition and reconstruction: Assessment in a cadaveric liver. *Phys. Med. Biol.* **2020**, *65*, 205008, https://doi.org/10.1088/1361-6560/abb172.
42. Arendt, C.T.; Leithner, D.; Mayerhoefer, M.E.; Gibbs, P.; Czerny, C.; Arnoldner, C.; Burck, I.; Leinung, M.; Tanyildizi, Y.; Lenga, L.; et al. Radiomics of high-resolution computed tomography for the differentiation between cholesteatoma and middle ear inflammation: Effects of post-reconstruction methods in a dual-center study. *Eur. Radiol.* **2021**, *31*, 4071–4078, https://doi.org/10.1007/s00330-020-07564-4.
43. Fortin, J.-P.; Parker, D.; Tunç, B.; Watanabe, T.; Elliott, M.A.; Ruparel, K.; Roalf, D.R.; Satterthwaite, T.D.; Gur, R.C.; Gur, R.E.; et al. Harmonization of multi-site diffusion tensor imaging data. *Neuroimage* **2017**, *161*, 149–170, https://doi.org/10.1016/j.neuroimage.2017.08.047.
44. Fortin, J.-P.; Cullen, N.; Sheline, Y.I.; Taylor, W.D.; Aselcioglu, I.; Cook, P.A.; Adams, P.; Cooper, C.; Fava, M.; McGrath, P.J.; et al. Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage* **2018**, *167*, 104–120, https://doi.org/10.1016/j.neuroimage.2017.11.024.
45. Crombé, A.; Kind, M.; Fadli, D.; le Loarer, F.; Italiano, A.; Buy, X.; Saut, O. Intensity harmonization techniques influence radiomics features and radiomics-based predictions in sarcoma patients. *Sci. Rep*. **2020**, *10*, 15496, https://doi.org/10.1038/s41598-020-72535-0.

46. Lucia, F.; Visvikis, D.; Vallières, M.; Desseroit, M.-C.; Miranda, O.; Robin, P.; Bonaffini, P.A.; Alfieri, J.; Masson, I.; Mervoyer, A.; et al. External validation of a combined PET and MRI radiomics model for prediction of recurrence in cervical cancer patients treated with chemoradiotherapy. *Eur. J. Nucl. Med. Mol. Imaging* **2019**, *46*, 864–877, https://doi.org/10.1007/s00259-018-4231-9.
47. Shiri, I.; Amini, M.; Nazari, M.; Hajianfar, G.; Haddadi Avval, A.; Abdollahi, H.; Oveisi, M.; Arabi, H.; Rahmim, A.; Zaidi, H. Impact of feature harmonization on radiogenomics analysis: Prediction of EGFR and KRAS mutations from non-small cell lung cancer PET/CT images. *Comput. Biol. Med.* **2022**, *142*, 105230, https://doi.org/10.1016/j.compbiomed.2022.105230.
48. Masson, I.; Da-Ano, R.; Lucia, F.; Doré, M.; Castelli, J.; Goislard de Monsabert, C.; Ramée, J.-F.; Sellami, S.; Visvikis, D.; Hatt, M.; et al. Statistical harmonization can improve the development of a multicenter CT-based radiomic model predictive of nonresponse to induction chemotherapy in laryngeal cancers. *Med. Phys.* **2021**, *48*, 4099–4109, https://doi.org/10.1002/mp.14948.
49. Gallardo-Estrella, L.; Lynch, D.A.; Prokop, M.; Stinson, D.; Zach, J.; Judy, P.F.; van Ginneken, B.; van Rikxoort, E.M. Normalizing computed tomography data reconstructed with different filter kernels: Effect on emphysema quantification. *Eur. Radiol.* **2016**, *26*, 478–486, https://doi.org/10.1007/s00330-015-3824-y.
50. Mackin, D.; Fave, X.; Zhang, L.; Fried, D.; Yang, J.; Taylor, B.; Rodriguez-Rivera, E.; Dodge, C.; Jones, A.K.; Court, L. Credence cartridge radiomics phantom CT scans—The cancer imaging archive (TCIA) public access—Cancer imaging archive wiki. *Cancer Imaging Arch.* **2017**. http://doi.org/10.7937/K9/TCIA.2017.zuzrml5b
51. Clark, K.; Vendt, B.; Smith, K.; Freymann, J.; Kirby, J.; Koppel, P.; Moore, S.; Phillips, S.; Maffitt, D.; Pringle, M.; et al. The cancer imaging archive (TCIA): Maintaining and operating a public information repository. *J. Digit. Imaging* **2013**, *26*, 1045–1057, https://doi.org/10.1007/s10278-013-9622-7.
52. Van Griethuysen, J.J.M.; Fedorov, A.; Parmar, C.; Hosny, A.; Aucoin, N.; Narayan, V.; Beets-Tan, R.G.H.; Fillion-Robin, J.-C.; Pieper, S.; Aerts, H.J.W.L. Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* **2017**, *77*, e104–e107, https://doi.org/10.1158/0008-5472.CAN-17-0339.
53. Team, R.C. *R Language Definition*; R Foundation for Statistical Computing: Vienna, Austria, 2000.
54. Gandrud, C. *Reproducible Research with R and R Studio*; CRC Press: Boca Raton, FL, USA, 2013; ISBN 9781466572843.
55. Lin, L.I. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **1989**, *45*, 255–268.
56. Stevenson, M.; Stevenson, M.M.; BiasedUrn, I. Package "epiR". Available online: https://vps.fmvz.usp.br/CRAN/web/packages/epiR/epiR.pdf (accessed on 15 January 2022).
57. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **1947**, *12*, 153–157, https://doi.org/10.1007/BF02295996.
58. Denzler, S.; Vuong, D.; Bogowicz, M.; Pavic, M.; Frauenfelder, T.; Thierstein, S.; Eboulet, E.I.; Maurer, B.; Schniering, J.; Gabryś, H.S.; et al. Impact of CT convolution kernel on robustness of

radiomic features for different lung diseases and tissue types. *Br. J. Radiol.* **2021**, *94*, 20200947, https://doi.org/10.1259/bjr.20200947.

59. He, L.; Huang, Y.; Ma, Z.; Liang, C.; Liang, C.; Liu, Z. Effects of contrast-enhancement, reconstruction slice thickness and convolution kernel on the diagnostic performance of radiomics signature in solitary pulmonary nodule. *Sci. Rep.* **2016**, *6*, 34921, https://doi.org/10.1038/srep34921.
60. Ibrahim, A.; Barufaldi, B.; Refaee, T.; Silva Filho, T.M.; Acciavatti, R.J.; Salahuddin, Z.; Hustinx, R.; Mottaghy, F.M.; Maidment, A.D.A.; Lambin, P. MaasPenn radiomics reproducibility score: A novel quantitative measure for evaluating the reproducibility of CT-based handcrafted radiomic features. *Cancers* **2022**, *14*, 1599.
61. Li, Y.; Ammari, S.; Balleyguier, C.; Lassau, N.; Chouzenoux, E. Impact of preprocessing and harmonization methods on the removal of scanner effects in brain MRI radiomic features. *Cancers* **2021**, *13*, 3000, https://doi.org/10.3390/cancers13123000.
62. Lupean, R.-A.; Ștefan, P.-A.; Csutak, C.; Lebovici, A.; Măluțan, A.M.; Buiga, R.; Melincovici, C.S.; Mihu, C.M. Differentiation of endometriomas from ovarian hemorrhagic cysts at magnetic resonance: The role of texture analysis. *Medicina* **2020**, *56*, 487, https://doi.org/10.3390/medicina56100487.

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$minimum = \min(\mathbf{X})$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2$$

$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_s} \mathbf{P}(i,j)\right)^2}{N_p}$$

$$correlation = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) ij - \mu_x \mu_y}{\sigma_x(i) \sigma_y(j)}$$

$$MAD = \frac{1}{N_p} \sum_{i=1}^{N_p} |\mathbf{X}(i) - \bar{X}|$$

$$minimum = \min(\mathbf{X})$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_s} \mathbf{P}\right.}{N_z}$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2$$

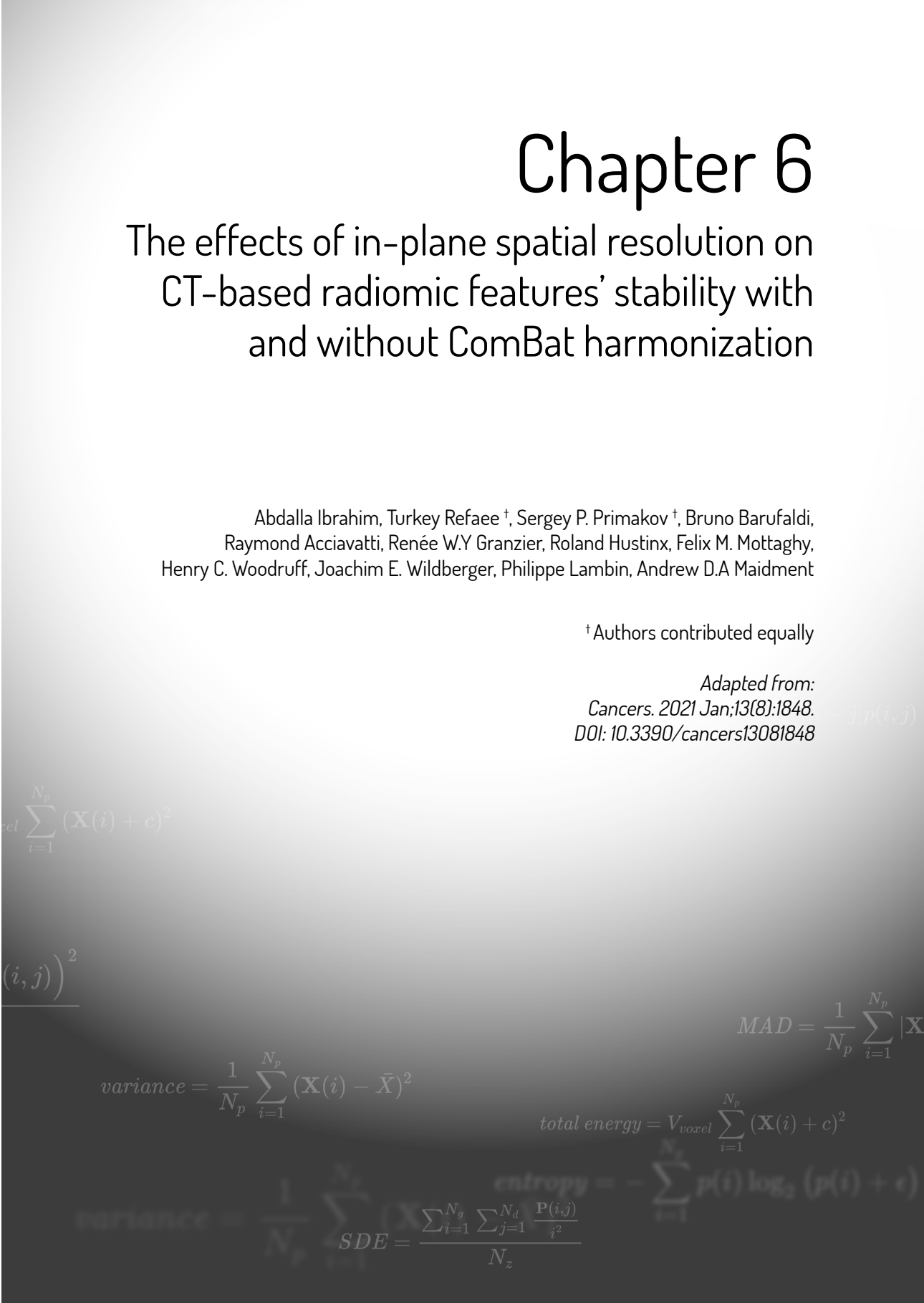
Chapter 6

The effects of in-plane spatial resolution on CT-based radiomic features' stability with and without ComBat harmonization

Abdalla Ibrahim, Turkey Refaee †, Sergey P. Primakov †, Bruno Barufaldi, Raymond Acciavatti, Renée W.Y Granzier, Roland Hustinx, Felix M. Mottaghy, Henry C. Woodruff, Joachim E. Wildberger, Philippe Lambin, Andrew D.A Maidment

† Authors contributed equally

Adapted from:
Cancers. 2021 Jan;13(8):1848.
DOI: 10.3390/cancers13081848



Abstract

While handcrafted radiomic features (HRFs) have shown promise in the field of personalized medicine, many hurdles hinders its incorporation into clinical practice, including but not limited to their sensitivity to differences in acquisition and reconstruction parameters. In this study, we evaluated the effects of differences in in-plane spatial resolution (IPR) on HRFs, using a phantom dataset (n=14) acquired on two scanner models. Further, we assessed the effects of interpolation methods (IMs), the choice of a new unified in-plane resolution (NUIR), and ComBat harmonization on the reproducibility of HRFs. The reproducibility of HRFs was significantly affected by variations in IPR, with pairwise concordant HRFs, as measured by the concordance correlation coefficient (CCC), ranging from 42% to 95%. The number of concordant HRFs (CCC > 0.9) after resampling varied depending on (i) the scanner model, (ii) the IM, and (iii) the NUIR. The number of concordant HRFs after ComBat harmonization depended on the variations between the batches harmonized. The majority of IMs resulted in a higher number of concordant HRFs compared to ComBat harmonization, and the combination of IMs and ComBat harmonization did not yield a significant benefit. Our developed framework can be used to assess reproducibility and harmonizability of RFs.

Keywords: Image Processing, Harmonization, Reproducibility, Radiomics biomarkers

Introduction

In recent years, quantitative medical imaging research using handcrafted radiomic features (HRFs) has been growing exponentially [1,2]. Radiomics refers to the high throughput extraction of quantitative imaging features that are expected to correlate with clinical and biological characteristics of patients [3,4]. For decades, it has been hypothesized that image texture analysis could potentially extract more information from an ROI than that solely perceived by the human eye [5,6]. Yet, the term radiomics has only been introduced recently [7,8]. HRFs are generally grouped into shape, intensity, and textural features. To date, many studies have reported on the potential of radiomics to predict various clinical endpoints [9,10]. However, major challenges, including the reproducibility of the HRFs across different acquisition and reconstruction parameters, have hindered the incorporation of radiomics in clinical decision support systems [11,12].

The essence of radiomics is that certain HRFs help decode biologic information [8], allowing these features to be treated as biomarkers. The mainstay of a biomarker is the ability to quantify it in a reproducible manner [13]. HRFs are mathematical equations applied to numeric arrays of intensity values which form the medical image. Therefore, it is intuitive that changes in the values in the array (due to differences in scan acquisition and reconstruction parameters), by the transitive property, lead to (potentially significant) quantitative changes in the HRFs. It is well established that changes in scan acquisition and reconstruction parameters affect the values in the array representing the medical image [14]. Therefore, it is a common clinical practice to scan a phantom to calibrate the CT scanner on a routine basis. Hence, similar practices are needed before radiomics studies are conducted, when the scans under analysis were acquired using heterogeneous acquisition and reconstruction parameters [15]. Many studies have already reported on the sensitivity of HRFs to different factors including: (i) temporal variability, or test-retest [16,17], in which two scans of a patient (or a phantom) are taken after a time interval using the exact scanning parameters; (ii) scanning parameters variability [11,18,19], in which an object (usually a phantom) is scanned multiple times using different scanning parameters. Variations in the majority of scanner/scanning parameter combinations were reported to impact the reproducibility of HRFs significantly [18-20].

One scan reconstruction parameter expected to have an effect on the reproducibility of HRFs is the in-plane spatial resolution (IPR), which is dictated in part by the pixel dimensions, while the through-plane spatial resolution is determined by the slice thickness and slice spacing. Resampling all the scans in a data set to a new unified in-plane spatial resolution (NUIR) before feature extraction has been employed as a method to reduce the variation in radiomic feature values [21,22]. The NUIR is usually decided based on the most frequent IPR in the dataset and different interpolation methods (IMs) can be used for this purpose.

Interpolation is a model-based method to recover continuous data from discrete data within a known range of data spacings (i.e., pixel size in images) [23]. The degree to which data recovery is possible is highly sensitive to the interpolation method and the underlying data structure. In the case of medical imaging analysis, interpolation is employed either to convert the spatial sampling rate (measured in pixel or voxel count per unit of length per dimension) to another, or to distort the image in the case of image registration [24]. Since the vast majority of HRFs are derived from pixel/voxel values and their distributions, interpolation to a common pixel spacing could potentially reduce variance introduced to these HRFs arising from differences in IPR.

As a rule, one must distinguish between interpolation methods that increase or reduce the image resolution. Interpolation from smaller pixels to larger pixels (i.e. reducing spatial resolution) usually involves some form of averaging, with the possible exception of modern deep learning-based methods.

Generally, while data acquired with small pixels will contain more noise, the process of averaging to large pixels will ameliorate the noise properties. As such, the process is less sensitive to the interpolation method/model. Interpolation from larger pixels to smaller pixels (i.e. increasing spatial resolution) on the other hand is fraught with challenges as the interpolated data can be highly sensitive to the interpolation model due to the need to create de novo pixel values. Larger pixels average the signal over a larger area than smaller ones, leading to the loss of variations in the original scene that occur over spatial frequencies smaller than the Nyquist limit and cannot be recovered exactly.

Certain methods, such as nearest neighbour interpolation (also called pixel replication), while fast, are less accurate than other methods such as sinc interpolation or deep-learning methods (which are trained with representative data). However, all such interpolation methods are sensitive to biases arising from the image [25]. The application of these methods to medical imaging has been evaluated qualitatively [26]. Yet, the effects of these methods on the reproducibility of HRFs is not well understood. Unlike humans, whose exposure to a vast assortment of scanners, patients, and acquisition conditions (including IPR) leads to a tolerance for such changes, IPR is likely to have more profound effects on HRFs.

A harmonization method that has become increasingly common in the field of radiomics is ComBat. ComBat was originally developed for the harmonization of gene expression arrays [27]. Several studies have investigated the potential of ComBat in radiomics analysis and recommended its use [28,29]. We hypothesize that ComBat, the chosen IM, and the selected NUIR will affect the reproducibility of HRFs differently. In this study, the reproducibility of HRFs was assessed across different IPRs, while keeping all other parameters fixed, using a public dataset of CT scans of a phantom. A thorough investigation of the applicability

of 10 different IMs was performed in an effort to identify suitable IMs for the purpose of increasing the number of reproducible HRFs in a heterogeneous dataset. In particular, we investigated whether data with discordant pixel sizes need to be interpolated to a common pixel size to perform radiomics analysis, and how the choice of IM and NUIR, as well as ComBat harmonization, affect the reproducibility of HRFs. Furthermore, we developed a generalizable workflow that assesses the impact of different harmonization techniques (Figure 1) on the reproducibility of RFs. Ultimately, the goal of our work is to guide robust radiomics analysis to ease its incorporation in clinical decision-making.

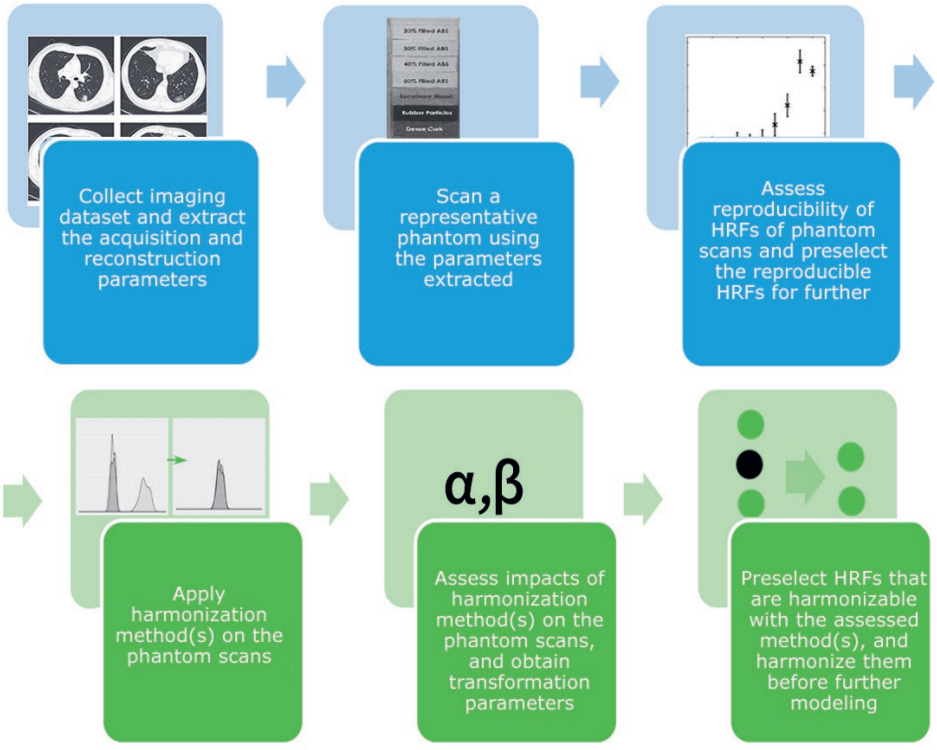


Figure 1. Proposed reproducible radiomic analysis workflow.

Materials and Methods

Phantom data

The publicly available Credence Cartridge Radiomics (CCR) phantom data [30] found in The Cancer Imaging Archive (TCIA.org) [31] was used. The CCR phantom is composed of 10 different layers that correspond to different texture patterns spanning a range of almost −900 to +700 HU (Figure S1). The publicly available dataset includes 251 scans of the phantom acquired using six scanner models manufactured by three different manufacturers. The

scans were acquired using various acquisition and reconstruction parameters to assess the reproducibility of HRFs. For the purpose of this study, 14 scans acquired using 2 different scanner models (Discovery STE & LightSpeed Pro 32) of the same manufacturer (GE), which were all acquired at a single slice thickness (1.25 mm), tube voltage (120 kV), tube current (250 mA), and convolution kernel (standard), but varying IPR (Table 1) were used. The reasoning behind this selection is multifold: (i) the effects of the variations are expected to be dependent on the heterogeneity in acquisition; (ii) the number and complexity of the different combinations available are too huge to be described, analyzed and presented in a single experiment; (iii) the data under analysis were acquired using the same scanner models, and the same acquisition and reconstruction parameters except for the in-plane resolution, which allows the assessment of the effect of variations in this single parameter.

Table 1. Scanning parameters of the phantom data.

Scanner		Pixel spacing (mm^2)
Discovery STE	**LightSpeed Pro 32**	
CCR-2-001	CCR-2-022	0.39*0.39
CCR-2-002	CCR-2-023	0.49*0.49
CCR-2-003	CCR-2-024	0.59*0.59
CCR-2-004	CCR-2-025	0.68*0.68
CCR-2-005	CCR-2-026	0.78*0.78
CCR-2-006	CCR-2-027	0.88*0.88
CCR-2-007	CCR-2-028	0.98*0.98

Interpolation and image resampling

The effects of the IMs included in the popular open-source radiomics toolbox PyRadiomics [33] were assessed in this study. The methods are based on the python library SimpleITK [33], and include (i) nearest neighbour (NN), (ii) linear, (iii) basis spline (B-spline), (iv) Gaussian, (v) Gaussian using labelling (mask) information (LabelGaussian), and windowed sinc interpolations using the following window types: (vi) Hamming (HammingWindowedSinc or HWS), (vii) Cosine (CosineWindowedSinc or CWS), (viii) Welch (WelchWindowedSinc or WWS), (ix) Lanczos window (LanczosWindowedSinc or LWS), and (x) Blackman (BlackmanWindowedSinc or BWS).

The simplest of these IMs, and the ones with the lowest computational costs, are (i) the NN interpolation, which functions by assigning any new voxel the same value as its closest neighbor in the original image; and (ii) linear interpolation, in which the values of new pixels are interpolated linearly between the two original values [26]. B-spline interpolation is more complex than NN or linear; the calculations span four pixels [34]. While the method performs well in terms of radiologic evaluation in which the aim is to convince human observers, it is known to unnecessarily over-smooth the image [26]. The windowed sinc functions are

complex convolution based interpolations that are based on multiplying the sinc function by a limited spatial support window to reduce unwanted effects on the resampled image [35], followed by filtering of the frequencies to avoid the injection of spurious frequency components. Windowed sinc functions are generally considered superior to other interpolation methods as little superfluous noise is injected into the interpolated images.

HRFs extraction

Each scan contained 10 independent regions of interest (ROIs) (one for each layer of the phantom) that occupy the same physical area of the phantom on each scan. For each ROI, HRFs were calculated using the open source software Pyradiomics V 2.1.2. HRFs were extracted multiple times to perform different experiments. First, to assess the effect of differences in in-plane resolution and ComBat harmonization on HRFs, no changes to the original in-plane resolution were made. Second, to assess the effect of different IMs and NUIRs and the combination of interpolation and ComBat, HRFs were extracted from the scans using all IMs and all available NUIRs in the dataset (Table 1).

For each set of scans (7 scans, with 10 ROIs per scan) from each scanner model (n=2), HRFs were extracted 71 times. The HRFs were extracted one time from the original scans, and 70 times with unique combinations of IM and NUIR. In each run, a total of 91 original RFs were extracted. In Pyradiomics, shape features are calculated on the original input image, and are not affected by the in-application resampling. Therefore, those HRFs were excluded.

To reduce noise and computational requirements, images were pre-processed by binning voxel grayscale values into bins with a fixed width of 25 HUs for extracting HRFs from unfiltered images. No other image pre-processing steps were performed. The extracted HRFs included HU intensity features, and texture features describing the spatial distribution of voxel intensities using 5 texture matrices (grey-level co-occurrence (GLCM), grey-level run-length (GLRLM), grey-level size-zone (GLSZM), grey-level dependence (GLDM), and neighborhood grey-tone difference (NGTDM) matrices). A more detailed description of the Pyradiomics HRFs can be found online (https://pyradiomics.readthedocs.io/en/latest/features.html).

ComBat harmonization

ComBat is an empirical Bayes based method used to estimate the effects of different batches on HRFs; in this scenario, variations in scan acquisition and reconstruction parameters were considered [27]. ComBat method assumes that a feature value can be approximated by the equation.

$$Y_{ij} = \alpha + \beta X_{ij} + \gamma_i + \delta_i \varepsilon_{ij} \quad (1)$$

where α is the average value for feature Y_{ij} for ROI j on scanner i; X is a design matrix of the biologic covariates known to affect the HRFs; β is the vector of regression coefficients corresponding to each biologic covariate; γ_i is the additive effect of scanner i on HRFs, δ_i is the multiplicative scanner effect, and εij is an error term, presupposed to be normally distributed with zero mean. Based on the values estimated, ComBat performs feature transformation in the form of:

$$Y_{ij}^{ComBat} = \frac{(Y_{ij} - \hat{\alpha} - \hat{\beta} X_{ij} - \gamma_i^*)}{\delta_i^*} + \hat{\alpha} + \hat{\beta} X_{ij} \quad (2)$$

where $\hat{\alpha}$ and $\hat{\beta}$ are estimators of parameters α and β, respectively. γ_i^* and δ_i^* are the empirical Bayes estimates of γ_i and δ_i, respectively [28].

Statistical analysis

To assess the agreement of a given HRF for the same ROI scanned using different settings and scanners, the concordance correlation coefficient (CCC) was calculated using the epiR package (Version 0.9-99) [36] and R language (Version 3.5.1) [37] with R studio (Version 1.1.456) [38]. The CCC is used to evaluate the agreement between paired readings [38], and provides the measure of concordance as a value between 1 and -1, where 0 represents no concordance and 1 or -1 represent a perfect direct positive or inverse concordance, respectively. The CCC metric further has the advantages of (i) robustness in small sample sizes, and (ii) taking the rank and value of the feature into consideration [39]. The cut-off of (CCC>0.9) was used to select reproducible HRFs, as the literature suggests that values < 0.9 indicate poor concordance [40].

Four different approaches for assessing concordances of HRFs were used (Figure 2): (i) HRFs extracted from the original scans; (ii) HRFs extracted from the original scans and harmonized using ComBat; (iii) HRFs extracted from resampled scans; and (iv) HRFs extracted from resampled scans harmonized using ComBat. For (i), the CCC was calculated for all HRFs of all ROIs across 7 different scans from each scanner. In each run, the CCC was calculated between a different pair of scans. For (ii), HRFs with nearly zero variance (i.e HRFs which have the same value in 95% or more of the data points) had to be removed before applying ComBat. Parametric prior estimations were used, and no reference batch was assigned for ComBat application. The CCC was calculated after harmonizing the remaining HRFs using ComBat. In each run, ComBat was applied on two batches (scans). For (iii), the CCC was calculated for the HRFs following feature extraction with each of the IMs. The effects of the NUIR were assessed by calculating the CCC for the HRFs after resampling all the scans to one of the available in-plane resolutions. For (iv), ComBat was applied after the same process in (iii), and the CCC was then calculated. To gauge an overall image of the reproducibility of HRFs across all pairs as well as the impact of IMs, NUIRs, and ComBat, the number (percentage) of HRFs that were reproducible by taking the intersection of HRFs that were reproducible in

each pairwise comparison of a certain scenario were compared (21 pairs in each scenario as shown in tables 2-5).

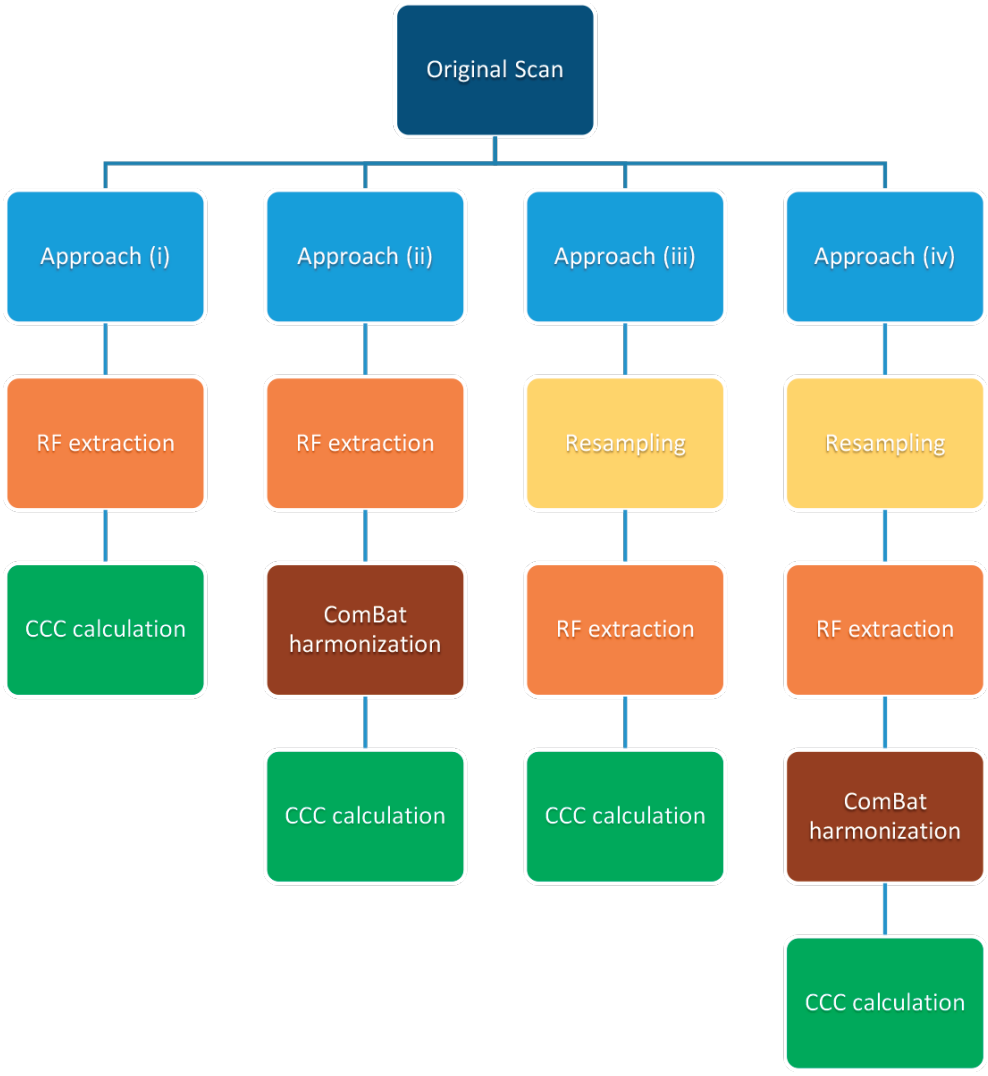


Figure 2. Reproducibility analysis approaches.

Further, we assessed the correlation between the HRFs that were concordant across all pairwise comparisons on each scanner model, using Spearman correlation [42]. HRFs were considered highly correlated if the Spearman's correlation coefficient had a value > 0.90.

Results

Approach (i): Effects of IPR on the reproducibility of HRFs

The number of HRFs insensitive to the variations in IPR depended on the scanner model (Tables 2 and S1). In pairwise comparisons, the number of concordant HRFs was lower when the difference in IPR between the scan pairs was greater. The lowest concordance was observed between the scan with the highest resolution and the scan with the lowest resolution.

Out of the 91 extracted HRFs, between 39 (42.9%) and 86 (94.5%) HRFs were concordant, varying pairwise and scanner wise. Some HRFs were robust to variations in IPR in one scanner model, and not in the other.

Table 2. Number of pair-wise concordant HRFs with a CCC > 0.9 before resampling, Discovery STE model.

Scan	CCR-2-001	CCR-2-002	CCR-2-003	CCR-2-004	CCR-2-005	CCR-2-006
CCR-2-002	75 (82.4%)					
CCR-2-003	57 (62.6%)	78 (85.7%)				
CCR-2-004	53 (58.2%)	64 (70.3%)	83 (91.2%)			
CCR-2-005	50 (54.9%)	61 (67.0%)	72 (79.1%)	86 (94.5%)		
CCR-2-006	51 (56.0%)	58 (63.7%)	68 (74.7%)	76 (83.5%)	85 (93.4%)	
CCR-2-007	39 (42.9%)	42 (46.2%)	44 (48.4%)	52 (57.1%)	60 (64.9%)	83 (91.2%)

On the Discovery STE model (GE), the number of concordant HRFs ranged between 39 (42.9%) and 86 (94.5%), with a median of 70 (39.6%) HRFs (Table 2). 36 (39.6%) HRFs were reproducible regardless of the IPR selected when all other scanning parameters were fixed (List S1). Of these 36 HRFs, nine remained after removing highly correlated HRFs (List S3), and none was highly correlated with volume. Overall, the Lightspeed Pro 32 model showed lower concordance than the Discovery STE model. The number of pairwise concordant HRFs on the Lightspeed Pro 32 model ranged between 39 (42.8%) and 82 (90.1%), with a median of 60 (65.9%) (Table S1). 27 (29.7%) HRFs were reproducible across all pairs (List S2). Of these 27 HRFs, nine remained after removing highly correlated HRFs (List S4), and none was highly correlated with volume. 26 (28.6%) HRFs were reproducible on both scanner models regardless of the IPR.

Approach (ii): ComBat harmonization of HRFs extracted from original scans

ComBat harmonization increased the number of concordant HRFs compared to before harmonization. On the Discovery model, the increment in the number (percentage) of HRFs ranged between 0 (0%) and 13 (14.3%), with a median of 6 (6.6%) of the total depending on the batches being harmonized (Table 3). 46 (50.5%) HRFs were found to be reproducible across all pairwise comparisons following ComBat harmonization, 35 of which were found to be highly correlated. The number of concordant HRFs decreased with the increment in IPR variation. Hence, the increment in the number of concordant HRFs was larger when the batches being harmonized had a larger difference in IPR.

Table 3. Number of pair-wise concordant HRFs with a CCC > 0.9 after ComBat harmonization, Discovery STE model.

Scan	CCR-2-001	CCR-2-002	CCR-2-003	CCR-2-004	CCR-2-005	CCR-2-006
CCR-2-002	79 (86.8%)					
CCR-2-003	65 (71.4%)	79 (86.8%)				
CCR-2-004	59 (64.8%)	70 (76.9%)	83 (91.2%)			
CCR-2-005	58 (63.7%)	66 (72.5%)	75 (82.4%)	87 (95.6%)		
CCR-2-006	57 (62.6%)	65 (71.4%)	70 (76.9%)	84 (92.3%)	86 (94.5%)	
CCR-2-007	48 (52.7%)	55 (60.4%)	57 (62.6%)	60 (65.9%)	73 (80.2%)	84 (92.3%)

The performance of ComBat had a similar pattern on both the Discovery STE and the Lightspeed Pro 32 models. The increment in the number (percentage) of concordant HRFs extracted from the scans acquired with the Lightspeed Pro 32 model following ComBat harmonization ranged between 1 (1.1%) and 14 (15.4%) HRFs with a median increment of 7 (7.7%) HRFs compared to before harmonization, depending on the batches being harmonized (Table S2). 41 (45.1%) HRFs were reproducible across all pairs following ComBat harmonization, 29 of which were found to be highly correlated.

Approach (iii): The effects of different IMs and NUIR on HRFs

Different interpolation methods showed different effects on the reproducibility of HRFs. These effects further depended on the selected NUIR and the scanner model (Figures 3 and S2). For the majority of combinations of scanner models, IMs and NUIRs, some HRFs were only concordant when extracted from the original scans, some HRFs became concordant only after resampling, while some lost their concordance following resampling (tables S5 and S6). CSW resampling to the highest and lowest resolutions are used below as detailed examples on both scanner models.

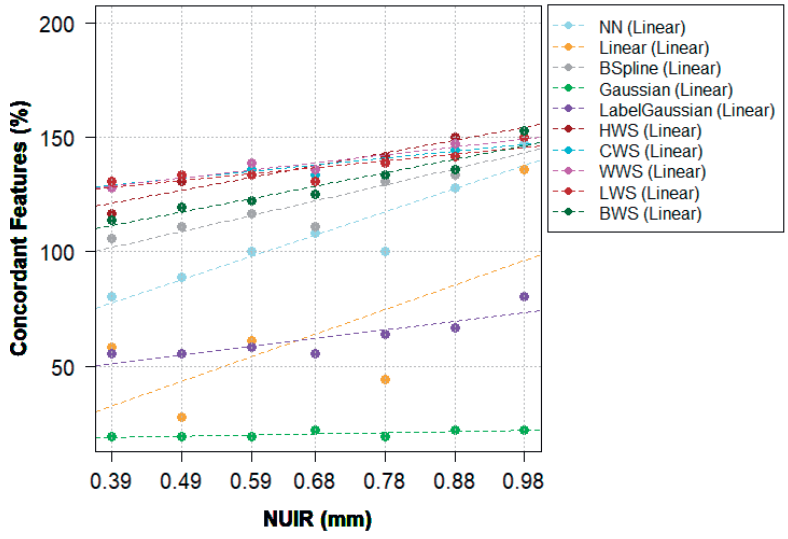


Figure 3. The percentage of concordant HRFs following resampling compared to no resampling with linear trendlines, Discovery STE model.

On the Discovery STE model, the use of windowed sinc IMs resulted in an overall increment in the number of reproducible HRFs, regardless of the NUIR selected. The range of HRFs that had an improved concordance across all pairs when using windowing sinc was between 14 (15.4%) and 20 (22%) HRFs, depending on the NUIR. When scans were resampled to the highest resolution using CWS, the increment in the number of concordant HRFs ranged between -2 (-2.2%) and 36 (39.6%), with a median of 12 (13.2%) HRFs. Moreover, 47 (51.6%) HRFs were concordant across all pairs. When scans were resampled to the lowest resolution using CWS, the increment in the number of concordant HRFs ranged between 4 (4.4%) and 35 (38.5%), with a median of 16 (17.6%) HRFs. 54 (59.3%) HRFs were concordant across all pairs. Table 4 shows the pairwise number (percentage) of reproducible HRFs following resampling to the median IPR value with CWS IM on the Discovery model, for comparison with table 5.

Table 4. Number of pair-wise concordant HRFs with a CCC > 0.9 after resampling* using CWS, Discovery model.

Scan	CCR-2-001	CCR-2-002	CCR-2-003	CCR-2-004	CCR-2-005	CCR-2-006
CCR-2-002	89 (97.8%)					
CCR-2-003	86 (94.5%)	88 (96.7%)				
CCR-2-004	86 (94.5%)	85 (93.4%)	88 (96.7%)			
CCR-2-005	86 (94.5%)	88 (96.7%)	91 (100%)	89 (97.8%)		
CCR-2-006	78 (85.7%)	77 (84.6%)	83 (91.2%)	79 (86.8%)	88 (96.7%)	
CCR-2-007	53 (58.2%)	53 (58.2%)	55 (60.4%)	54 (59.3%)	60 (65.9%)	85 (93.4%)

* All scans were resampled to the median pixel spacing value (0.49*0.49 mm2).

HWS performed the best when the images were resampled to a NUIR equal to or lower than the median (0.49*0.49 mm^2), while CWS, WWS and LWS methods performed better on NUIR values higher than the median. BSpline IM resulted in a minor to significant increment in the number of reproducible HRFs, with higher number of concordant features when higher NUIRs where chosen. Gaussian and Label-Gaussian IMs consistently resulted in lower numbers of concordant HRFs. The number of HRFs losing concordance across all pairs when using a Gaussian IM ranged between -29 (-31.9%) and -30 (-33%) HRFs, while the range for LabelGaussian was between -11 (-12.1%) and -19 (-20.9%) HRFs, depending on the NUIR. The rest of IMs (NN and Linear) resulted in an overall decrease in the number of concordant HRFs when a NUIR below the median resolution was selected, and a minor-significant improvement with NUIRs higher than the median resolution (Table S5).

On the Lightspeed Pro 32 model, windowed sinc IMs (except for BWS) showed a consistent increment in the number of reproducible HRFs, and varying depending on the NUIR. When scans were resampled to the highest resolution using CWS, the increment in the number of concordant HRFs ranged between -9 (-9.9%) and 36 (39.6%), with a median of 8 (8.8%) HRFs. 30 (33%) HRFs were concordant across all pairs. When scans were resampled to the lowest resolution using CWS, the increment in the number of concordant HRFs ranged between -3 (-3.3%) and 31 (34.1%), with a median of 16 (17.6%) HRFs. 38 (41.8 %) HRFs were concordant

across all pairs. Table S3 shows the pairwise number (percentage) of concordant HRFs following resampling to the median IPR value with CWS IM on the LightSpeed Pro 32 model, for comparison with table S4. The application of other IMs (BWS, NN, Linear, Gaussian and Label-Gaussian) with a NUIR other than the two lowest resolutions available resulted in an overall decrease in the number of concordant HRFs. However, when the lowest resolution was selected as NUIR, BSpline IM outperformed all other methods when the number of concordant HRFs across all pairs was considered (Table S6).

Approach (iv): The combination of IMs and ComBat harmonization

Approach (iii) resulted in a higher number of concordant HRFs in the majority of pairwise scenarios compared to approach (ii) for the majority of IMs that performed solely well (for example, table 3 vs table 4). The application of ComBat harmonization on HRFs extracted from resampled scans varied per scanner model, IMs, NUIRs, and batches. However, when the number of concordant HRFs across all pairs is considered, ComBat increased the number of concordant HRFs in almost all of the investigated scenarios (Figures 4 and S3; tables S7 and S8).

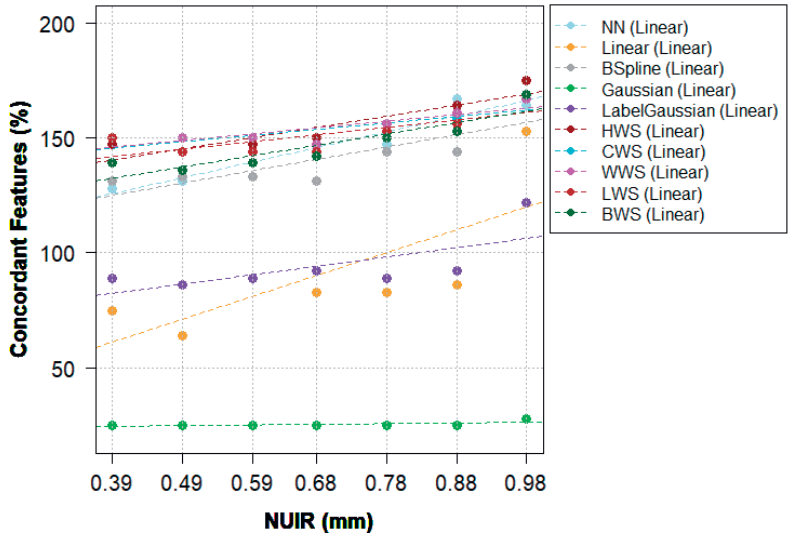


Figure 4. The percentage of concordant HRFs following resampling and ComBat harmonization compared to no resampling with linear trendlines, Discovery STE model.

On the Discovery model, the increment in the number (percentage) of concordant HRFs extracted from scans resampled to the highest resolution after ComBat harmonization ranged between 0 (0%) and 10 (11%), with a median increment of 0 (0%) of the total number of HRFs compared to before harmonization. 54 (59.3%) HRFs were concordant across all pairs. When ComBat was applied on HRFs extracted from scans resampled to the lowest resolution, the increment in the number (percentage) of HRFs ranged between -1 (-1.1%)

and 10 (11%) HRFs, with a median of 0 (0%), depending on the batches being harmonized. 61 (67%) were found to be stable across all pairs. Table 5 shows the Number of pair-wise concordant HRFs following the application of ComBat on scans acquired on the Discovery STE model, and resampled to the median IPR value using CWS IM.

Table 5. Number of pair-wise concordant HRFs with a CCC > 0.9 after ComBat following resampling* using CWS, Discovery STE model.

Scan	CCR-2-022	CCR-2-023	CCR-2-024	CCR-2-025	CCR-2-026	CCR-2-027
CCR-2-023	89 (97.8%)					
CCR-2-024	86 (94.5%)	88 (96.7%)				
CCR-2-025	86 (94.5%)	85 (93.4%)	88 (96.7%)			
CCR-2-026	86 (94.5%)	88 (96.7%)	91 (100%)	89 (97.8%)		
CCR-2-027	79 (86.8%)	78 (85.7%)	84 (92.3%)	84 (92.3%)	89 (97.8%)	
CCR-2-028	57 (62.6%)	61 (67.0%)	60 (65.9%)	59 (64.8%)	72 (79.1%)	85 (93.4%)

* All scans were resampled to the median pixel spacing value (0.49*0.49 mm2).

On the LightSpeed Pro 32 model, the increment in the number (percentage) of concordant HRFs after ComBat harmonization on HRFs extracted from scans resampled to the highest resolution (lowest concordance) ranged between -1 (-1.1%) and 13 (14.3%) HRFs, with a median of 3 (3.3%) of the total number of HRFs compared to before harmonization. 42 (46.2%) HRFs were concordant across all pairs. When ComBat was applied on HRFs extracted from scans resampled to the lowest resolution (highest concordance), the increment in the number (percentage) of HRFs ranged between 0 (0%) and 10 (11%) HRFs, with a median increment of 1 (1.1%) feature. 51 (56%) HRFs were concordant across all pairs. Table S4 shows the pairwise CCC following the application of ComBat on scans acquired with the LightSpeed Pro 32 model, and resampled to the median IPR value using CWS IM.

Discussion

In this study, the effects of variations in scans' IPR on the reproducibility of HRFs, the proper methodology of identifying HRFs that are reproducible across different IPRs, and how to properly adjust for these differences before performing radiomics analysis using image interpolation and/or ComBat harmonization were thoroughly investigated. Uniquely, this study evaluates the effects of all the different IMs and the choice of NUIRs on the reproducibility of HRFs. Previous studies usually investigated a single IM with a single NUIR [21,22].

While two batches of scans acquired with the same imaging parameters on two scanner models of the same vendor were used for analysis, the effects of IPR, ComBat, IMs, and NUIR on the reproducibility of HRFs varied on each of the scanner models. The CCC was calculated pairwise to assess the reproducibility of HRFs when different sets of data were used as batches. Calculating the pairwise CCC between HRF values extracted before resampling the images revealed that the reproducibility of HRFs in our data depended on several factors

including, but not limited to, the definition of the HRF, the degree of variation in IPR, and the scanner (hardware) make/model. Addressing the effects of these factors is crucial for performing robust radiomics analysis.

Without performing image preprocessing, the number of reproducible HRFs varied according to the batches being assessed. The aim of this study was to show that different investigated scenarios showed different numbers of reproducible HRFs. Therefore, although 36 HRFs for the Discovery STE scanner (27 HRFs for LightSpeed Pro 32 scanner) were always included in the set of concordant HRFs, it is difficult to conclude that these HRFs are insensitive to spatial resolution on all other scanner models based on our experiments. Yet, our framework guides the methodology of identifying reproducible HRFs according to the data under analysis. As we have shown, the number and type of HRFs is at least sensitive to the scanner model by the same manufacturer. Moreover, we anticipate based on their definition, that certain HRFs (such as histogram-based features) are less sensitive, while others (eg. texture features) are more sensitive to variations in scanning parameters and/or imaging vendors. Generally, scans with more similar original IPRs, and those of integer multiples of IPR showed higher numbers of concordant HRFs before and after resampling. This can be explained by the mechanisms by which a scan is acquired. When all other scanning parameters are fixed, the variations in IPR will result in variations in the number of pixels in 2D, while the other dimensions are preserved. Therefore, when all other parameters are fixed, the closer the IPR values are, the closer the values of the extracted HRFs.

For the IMs, the number of HRFs that had better/worse concordance after resampling was dependent on the NUIR chosen and scanner model. The window sinc interpolation family performed consistently better on both scanners and NUIRs investigated. In the field of radiology, both NN and linear are known to result in imprecisions [26,35]. A study into the reproducibility of HRFs investigated the performance of B-spline, linear and NN using a single image slice thickness, and concluded that NN is not a favorable method for the reproducibility of HRFs [42]. Our results support these previous reports by showing that NN and linear IMs are not the best candidates for improving the reproducibility of HRFs among scans acquired with different IPRs, and their use led to lower numbers of concordant HRFs in many of the investigated scenarios.

With regard to the selection of NUIR, a common trend of an inverse relationship between the NUIR and the number of concordant HRFs following resampling was observed. This trend was observed in both scanner models investigated. However, the percentage difference between the concordant HRFs is not significant at the lower end of the NUIR spectrum (Figures 3, 4, S2 and S3; tables S5 and S6). As the best NUIR is expected to be task dependent (for e.g classification of a lesion, predicting response to therapy or overall survival, etc), outcome-based analysis is needed to determine the best NUIR. Yet, as a general rule, the

smaller the NUIR, the better the concordance. In addition, while the number of non-highly correlated HRFs was found to be low on both scanner models (9 and 11 HRFs before and after ComBat harmonization, respectively), the exclusion of highly correlated HRFs should be performed based on the effects of the removal of these HRFs on the model performance.

A previous study investigated the effects on HRFs of voxel size resampling using linear interpolation. The authors resampled the scans of a phantom to a single voxel size, which was larger than the largest voxel size in the original scans, and reported that around 20% of the HRFs (N=213) became concordant after resampling [22]. Another study also investigated the effects of voxel size on HRFs of lung cancer patients [21]. The authors resampled all the scans to a single common voxel size using linear interpolation, and reported that resampling does not eliminate all the variations in feature values even when the only variation in scan acquisition and reconstruction parameters was the voxel size, but is favorable to no resampling. Another group investigated the effects of variation in several acquisition and reconstruction parameters on a 13-layer phantom using a different approach, and reported that resampling the scans to isotropic voxels increased the percentage of concordant HRFs from 59.5% to 89.3% [43]. In this study, we found a similar conclusion: the number of previously non-concordant HRFs that became concordant following resampling to the lowest resolution ranged between 1.1% and 22% depending on the IM, and not all HRFs benefit from image resampling.

In contrast to previous studies, we investigated more IMs and harmonization techniques, and propose a guideline on how to carefully approach HRFs reproducibility studies. Furthermore, we found that linear interpolation is not a good candidate for the purpose of improving the reproducibility of HRFs, when compared to other available IMs; and that the performance of an IM is dependent on the original IPR values and the chosen NUIR, as well as the imaging vendor.

When pairwise comparisons were considered, the performance of ComBat harmonization was found to be inferior to that of well-performing IMs, regardless of the NUIR. Moreover, the combination of ComBat and the well-performing IMs did not yield significantly better results compared to solely using the IM. Furthermore, the performance of ComBat varied depending on the batches used. Nevertheless, when the number of concordant HRFs across all pairs was considered, ComBat harmonization was of added value in almost all scenarios. Therefore, ComBat application on HRFs should follow a reproducibility study (phantom or tissue studies) to assess the impact of ComBat on the reproducibility of HRFs in those settings, and use only the harmonizable HRFs for further radiomics analyses [15], as described in the workflow (Figure 1). The application of ComBat without assessing HRFs' reproducibility as described may result in the inclusion of a high percentage of unreproducible HRFs, or even the loss of some of the HRFs that were originally reproducible, rendering the analysis of

these HRFs meaningless. This finding regarding ComBat harmonization is not in line with previous reports, which reported that ComBat successfully removes the batch effects for all HRFs [28,44]. This could be attributed to the differences in the radiomics software and/or the evaluation metrics used. In contrast to previous studies, and as the aim of harmonization is to improve reproducibility but necessarily the performance of generated radiomic models, we opted for the CCC. The CCC provides an accurate description of the reproducibility of HRFs, which is not reflected in neither the distribution of HRFs nor the performance of radiomics models [45]. If radiomic models are to be used clinically, it is expected to be applied to one patient per time. Therefore, the importance has been given in this study to the individual feature values, and not their distributions. HRFs with different values and order rank can share similar distributions, in which case the feature cannot be considered reproducible. In addition, different modeling techniques may yield significantly different results on the same dataset. Hence, the difference in the performance of a radiomic signature before and after harmonization does not necessarily inform about the performance of the harmonization method. Our proposed framework addresses this issue, and guides the selection of reproducible and harmonizable HRFs before developing a radiomic signature, which helps the translation and generalization of results, and ultimately the inclusion of radiomic signatures in clinical practice.

Of note, not all HRFs benefit from resampling all scans to a NUIR, or using ComBat harmonization. Some HRFs lost their concordance following resampling, depending on the IM employed and the chosen NUIR. The combination of IMs and NUIRs affected the HRFs differently on different scanner models. Some HRFs were not found to be concordant on one of the scanner models before or after resampling to any of the available NUIRs using any of the IMs, but were found to be concordant on the other scanner model. Other HRFs were found to be concordant across different scanner models and IPRs. These findings indicate the need for performing reproducibility studies depending on the data under study, and the fact that at this level, we are unable to provide a list of HRFs that can be used regardless of the acquisition and reconstruction parameters and scanner models used. However, it lays down the bases for identifying reproducible HRFs before performing data analysis. In real life scenarios, the variations between the imaging parameters in retrospective cohorts (especially multicentric) are usually not only limited to the IPR. Aside from the scanner/scanning parameters combination variations, some of the effects will be attributed to patient populations. Furthermore, while phantom studies reflect on the reproducibility of HRFs extracted from anthropomorphic phantoms, HRFs extracted from human tissue are expected to have a wider range of variations, due to the inclusion of biologic factors. This knowledge, combined with our findings, necessitate the critical investigation of the reproducibility of HRFs across the different scanning parameters/scanners before performing any statistical analysis, and future investigations into the effects of differences in acquisition and reconstruction parameters on the reproducibility of HRFs extracted from human tissues,

if feasible. Directly performing radiomics analysis on data acquired heterogeneously leads to spurious results, and lacks meaningful interpretation. Henceforth, we reiterate the need for using our proposed robust radiomics analysis framework for addressing differences in IPR. Furthermore, the workflow can be generalized to evaluate other harmonization methods.

Conclusions

The reproducibility of a given HRF, and its harmonizabilty with ComBat are not constants, but depended on the degree of variation in a single reconstruction parameter (the in-plane resolution) of the scans being analyzed. This implies that additional changes in the acquisition and reconstruction parameters could further reduce the number of reproducible and harmonizable HRFs. When scans acquired with different IPR values are to be analyzed, resampling the scans to a unified resolution can significantly improve the reproducibility of HRFs. Interpolation methods (CWS, HWS, BWS, WWS and B-spline) were found to be superior to ComBat harmonization alone in addressing the variations in HRFs attributed to differences in IPR, and the combination of an IM with ComBat following NUIR could increase the number of reproducible HRFs in some scenarios. The application of our proposed framework aids the selection of data- and outcome-specific interpolation and harmonization methods, and is expected to improve the translation and generalizability of radiomics analyses.

References

1. Walsh, S.; de Jong, E.E.C.; van Timmeren, J.E.; Ibrahim, A.; Compter, I.; Peerlings, J.; Sanduleanu, S.; Refaee, T.; Keek, S.; Larue, R.T.H.M.; et al. Decision Support Systems in Oncology. JCO Clin Cancer Inform 2019, 3, 1–9, doi:10.1200/CCI.18.00001.
2. Lambin, P.; Leijenaar, R.T.H.; Deist, T.M.; Peerlings, J.; de Jong, E.E.C.; van Timmeren, J.; Sanduleanu, S.; Larue, R.T.H.M.; Even, A.J.G.; Jochems, A.; et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat. Rev. Clin. Oncol. 2017, 14, 749–762, doi:10.1038/nrclinonc.2017.141.
3. Ibrahim, A.; Vallières, M.; Woodruff, H.; Primakov, S.; Beheshti, M.; Keek, S.; Refaee, T.; Sanduleanu, S.; Walsh, S.; Morin, O.; et al. Radiomics Analysis for Clinical Decision Support in Nuclear Medicine. Semin. Nucl. Med. 2019, 49, 438–449, doi:10.1053/j.semnuclmed.2019.06.005.
4. Kumar, V.; Gu, Y.; Basu, S.; Berglund, A.; Eschrich, S.A.; Schabath, M.B.; Forster, K.; Aerts, H.J.W.L.; Dekker, A.; Fenstermacher, D.; et al. Radiomics: the process and the challenges. Magn. Reson. Imaging 2012, 30, 1234–1248, doi:10.1016/j.mri.2012.06.010.
5. Miller, A.S.; Blott, B.H.; Hames, T.K. Review of neural network applications in medical imaging and signal processing. Med. Biol. Eng. Comput. 1992, 30, 449–464, doi:10.1007/BF02457822.
6. Kjaer, L.; Ring, P.; Thomsen, C.; Henriksen, O. Texture analysis in quantitative MR imaging. Tissue characterisation of normal brain and intracranial tumours at 1.5 T. Acta radiol. 1995, 36, 127–135.
7. Lambin, P.; Rios-Velazquez, E.; Leijenaar, R.; Carvalho, S.; van Stiphout, R.G.P.M.; Granton, P.; Zegers, C.M.L.; Gillies, R.; Boellard, R.; Dekker, A.; et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur. J. Cancer 2012, 48, 441–446, doi:10.1016/j.ejca.2011.11.036.
8. Gillies, R.J.; Kinahan, P.E.; Hricak, H. Radiomics: Images Are More than Pictures, They Are Data. Radiology 2016, 278, 563–577, doi:10.1148/radiol.2015151169.
9. Refaee, T.; Wu, G.; Ibrahim, A.; Halilaj, I.; Leijenaar, R.T.H.; Rogers, W.; Gietema, H.A.; Hendriks, L.E.L.; Lambin, P.; Woodruff, H.C. The Emerging Role of Radiomics in COPD and Lung Cancer. Respiration 2020, 99, 99–107, doi:10.1159/000505429.
10. Rogers, W.; Thulasi Seetha, S.; Refaee, T.A.G.; Lieverse, R.I.Y.; Granzier, R.W.Y.; Ibrahim, A.; Keek, S.A.; Sanduleanu, S.; Primakov, S.P.; Beuque, M.P.L.; et al. Radiomics: from qualitative to quantitative imaging. Br. J. Radiol. 2020, 93, 20190948, doi:10.1259/bjr.20190948.
11. Mackin, D.; Fave, X.; Zhang, L.; Fried, D.; Yang, J.; Taylor, B.; Rodriguez-Rivera, E.; Dodge, C.; Jones, A.K.; Court, L. Measuring Computed Tomography Scanner Variability of Radiomics Features. Invest. Radiol. 2015, 50, 757–765, doi:10.1097/RLI.0000000000000180.
12. Berenguer, R.; Pastor-Juan, M.D.R.; Canales-Vázquez, J.; Castro-García, M.; Villas, M.V.; Mansilla Legorburo, F.; Sabater, S. Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters. Radiology 2018, 288, 407–415, doi:10.1148/radiol.2018172361.
13. Strimbu, K.; Tavel, J.A. What are biomarkers? Curr. Opin. HIV AIDS 2010, 5, 463.

14. Davis, A.T.; Palmer, A.L.; Pani, S.; Nisbet, A. Assessment of the variation in CT scanner performance (image quality and Hounsfield units) with scan parameters, for image optimisation in radiotherapy treatment planning. Phys. Med. 2018, 45, 59–64, doi:10.1016/j.ejmp.2017.11.036.
15. Ibrahim, A.; Primakov, S.; Beuque, M.; Woodruff, H.C.; Halilaj, I.; Wu, G.; Refaee, T.; Granzier, R.; Widaatalla, Y.; Hustinx, R.; et al. Radiomics for precision medicine: current challenges,future prospects, and the proposal of a new framework. Methods 2020, doi:10.1016/j.ymeth.2020.05.022.
16. van Timmeren, J.E.; Leijenaar, R.T.H.; van Elmpt, W.; Wang, J.; Zhang, Z.; Dekker, A.; Lambin, P. Test-Retest Data for Radiomics Feature Stability Analysis: Generalizable or Study-Specific? Tomography 2016, 2, 361–365, doi:10.18383/j.tom.2016.00208.
17. Peerlings, J.; Woodruff, H.C.; Winfield, J.M.; Ibrahim, A.; Van Beers, B.E.; Heerschap, A.; Jackson, A.; Wildberger, J.E.; Mottaghy, F.M.; DeSouza, N.M.; et al. Stability of radiomics features in apparent diffusion coefficient maps from a multi-centre test-retest trial. Sci. Rep. 2019, 9, 4800, doi:10.1038/s41598-019-41344-5.
18. Zhovannik, I.; Bussink, J.; Traverso, A.; Shi, Z.; Kalendralis, P.; Wee, L.; Dekker, A.; Fijten, R.; Monshouwer, R. Learning from scanners: Bias reduction and feature correction in radiomics. Clin Transl Radiat Oncol 2019, 19, 33–38, doi:10.1016/j.ctro.2019.07.003.
19. Traverso, A.; Wee, L.; Dekker, A.; Gillies, R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. Int. J. Radiat. Oncol. Biol. Phys. 2018, 102, 1143–1158, doi:10.1016/j.ijrobp.2018.05.053.
20. Papanikolaou, N.; Matos, C.; Koh, D.M. How to develop a meaningful radiomic signature for clinical use in oncologic patients. Cancer Imaging 2020, 20, 33, doi:10.1186/s40644-020-00311-4.
21. Shafiq-Ul-Hassan, M.; Latifi, K.; Zhang, G.; Ullah, G.; Gillies, R.; Moros, E. Voxel size and gray level normalization of CT radiomic features in lung cancer. Sci. Rep. 2018, 8, 10545, doi:10.1038/s41598-018-28895-9.
22. Shafiq-ul-Hassan, M.; Zhang, G.G.; Latifi, K. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. Medical 2017.
23. Thévenaz, P.; Blu, T.; Unser, M. Image interpolation and resampling. of medical imaging, processing and analysis 2000.
24. Haddad, M.; Porenta, G. Impact of reorientation algorithms on quantitative myocardial SPECT perfusion imaging. J. Nucl. Med. 1998, 39, 1864–1869.
25. Menon, S.; Damian, A.; Hu, S.; Ravi, N.; Rudin, C. PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; openaccess.thecvf.com, 2020; pp. 2437–2445.
26. Parker, J.; Kenyon, R.V.; Troxel, D.E. Comparison of interpolating methods for image resampling. IEEE Trans. Med. Imaging 1983, 2, 31–39, doi:10.1109/TMI.1983.4307610.
27. Johnson, W.E.; Li, C.; Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 2007, 8, 118–127, doi:10.1093/biostatistics/kxj037.

28. Orlhac, F.; Frouin, F.; Nioche, C.; Ayache, N.; Buvat, I. Validation of a method to compensate multicenter effects affecting CT radiomic features. 2018.
29. Orlhac, F.; Boughdad, S.; Philippe, C.; Stalla-Bourdillon, H.; Nioche, C.; Champion, L.; Soussan, M.; Frouin, F.; Frouin, V.; Buvat, I. A Postreconstruction Harmonization Method for Multicenter Radiomic Studies in PET. J. Nucl. Med. 2018, 59, 1321–1328, doi:10.2967/jnumed.117.199935.
30. Mackin, D.; Fave, X.; Zhang, L.; Fried, D.; Yang, J.; Taylor, B.; Rodriguez-Rivera, E.; Dodge, C.; Jones, A.K.; and Court, L. Credence Cartridge Radiomics Phantom CT Scans - The Cancer Imaging Archive (TCIA) Public Access - Cancer Imaging Archive Wiki. Cancer Imaging Archive 2017.
31. Clark, K.; Vendt, B.; Smith, K.; Freymann, J.; Kirby, J.; Koppel, P.; Moore, S.; Phillips, S.; Maffitt, D.; Pringle, M.; et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J. Digit. Imaging 2013, 26, 1045–1057, doi:10.1007/s10278-013-9622-7.
32. van Griethuysen, J.J.M.; Fedorov, A.; Parmar, C.; Hosny, A.; Aucoin, N.; Narayan, V.; Beets-Tan, R.G.H.; Fillion-Robin, J.-C.; Pieper, S.; Aerts, H.J.W.L. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res. 2017, 77, e104–e107, doi:10.1158/0008-5472.CAN-17-0339.
33. Lowekamp, B.C.; Chen, D.T.; Ibáñez, L.; Blezek, D. The Design of SimpleITK. Front. Neuroinform. 2013, 7, 45, doi:10.3389/fninf.2013.00045.
34. Hsieh Hou; Andrews, H. Cubic splines for image interpolation and digital filtering. IEEE Trans. Acoust. 1978, 26, 508–517, doi:10.1109/TASSP.1978.1163154.
35. Meijering, E.H.; Niessen, W.J.; Viergever, M.A. Quantitative evaluation of convolution-based methods for medical image interpolation. Med. Image Anal. 2001, 5, 111–126, doi:10.1016/s1361-8415(00)00040-2.
36. Stevenson, M.; Stevenson, M.M.; BiasedUrn, I. Package “epiR.” 2020.
37. Team, R.C. R language definition. Vienna, Austria: R foundation for statistical computing 2000.
38. Gandrud, C. Reproducible Research with R and R Studio; CRC Press, 2013; ISBN 9781466572843.
39. Lin, L.I. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989, 45, 255–268.
40. McBride, G.B. A proposal for strength-of-agreement criteria for Lin’s concordance correlation coefficient. NIWA client report: HAM2005-062 2005, 62.
41. Zar, J.H. Spearman Rank Correlation. Encyclopedia of Biostatistics 2005.
42. Larue, R.T.H.M.; van Timmeren, J.E.; de Jong, E.E.C.; Feliciani, G.; Leijenaar, R.T.H.; Schreurs, W.M.J.; Sosef, M.N.; Raat, F.H.P.J.; van der Zande, F.H.R.; Das, M.; et al. Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice thicknesses: a comprehensive phantom study. Acta Oncol. 2017, 56, 1544–1553, doi:10.1080/0284186X.2017.1351624.
43. Ligero, M.; Jordi-Ollero, O.; Bernatowicz, K.; Garcia-Ruiz, A.; Delgado-Muñoz, E.; Leiva, D.; Mast, R.; Suarez, C.; Sala-Llonch, R.; Calvo, N.; et al. Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis. Eur. Radiol. 2021, 31, 1460–1470, doi:10.1007/s00330-020-07174-0.

44. Da-Ano, R.; Masson, I.; Lucia, F.; Doré, M.; Robin, P.; Alfieri, J.; Rousseau, C.; Mervoyer, A.; Reinhold, C.; Castelli, J.; et al. Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies. Sci. Rep. 2020, 10, 10248, doi:10.1038/s41598-020-66110-w.
45. Vetter, T.R.; Schober, P. Agreement Analysis: What He Said, She Said Versus You Said. Anesth. Analg. 2018, 126, 2123–2128, doi:10.1213/ANE.0000000000002924.

Supplementary Materials

The following are available online at www.mdpi.com/xxx/s1, Figure S1: The scanned CCR Phantom, Figure S2: The percentage of concordant features following resampling compared to no resampling with linear trendlines, LightSpeed Pro 32 model, Figure S3: The percentage of concordant features following resampling and ComBat harmonization compared to no resampling with linear trendlines, LightSpeed Pro 32 model, Table S1: Number of pair-wise concordant features with a CCC > 0.9 before resampling, LightSpeed Pro 32 model, Table S2: Number of pair-wise concordant features with a CCC > 0.9 after ComBat, LightSpeed Pro 32 model, Table S3: Number of pair-wise concordant features with a CCC > 0.9 after resampling* using CWS, LightSpeed Pro 32 model, Table S4: Number of pair-wise concordant features with a CCC > 0.9 after ComBat following resampling* using CWS, LightSpeed Pro 32 model, Table S5: Summary of the number of concordant features before and after resampling, Discovery STE model, Table S6: Summary of the number of concordant features before and after resampling, LightSpeed Pro 32 model, List S1: HRFs with CCC>0.9 across all pairs on Discovery STE model, List S2: HRFs with CCC>0.9 across all pairs on LightSpeed Pro 32 model, List S3: Non-highly correlated HRFs with CCC>0.9 across all pairs on Discovery STE model, List S4: Non-highly correlated HRFs with CCC>0.9 across all pairs on LighSpeed Pro 32 model.

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$minimum = \min(\mathbf{X})$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2$$

$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_s} \mathbf{P}(i,j)\right)^2}{N_p}$$

$$correlation = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) ij - \mu_x \mu_y}{\sigma_x(i) \sigma_y(j)}$$

$$MAD = \frac{1}{N_p} \sum_{i=1}^{N_p} |\mathbf{X}(i) - \bar{X}|$$

$$minimum = \min(\mathbf{X})$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_s} \mathbf{P}\right.}{N_z}$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2$$

Part III

$$variance = \frac{1}{N_p}\sum_{i=1}^{N_p}(\mathbf{X}(i)-\bar{X})^2$$

$$total\ energy = V_{voxel}\sum_{i=1}^{N_p}(\mathbf{X}(i)+c)^2$$

$$SDE = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}\frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$minimum = \min(\mathbf{X})$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2$$

$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_s} \mathbf{P}(i,j)\right)^2}{N_p}$$

$$correlation = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) ij - \mu_x \mu_y}{\sigma_x(i) \sigma_y(j)}$$

$$MAD = \frac{1}{N_p} \sum_{i=1}^{N_p} |\mathbf{X}(i) - \bar{X}|$$

$$minimum = \min(\mathbf{X})$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_s} \mathbf{P}\right.}{N_z}$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

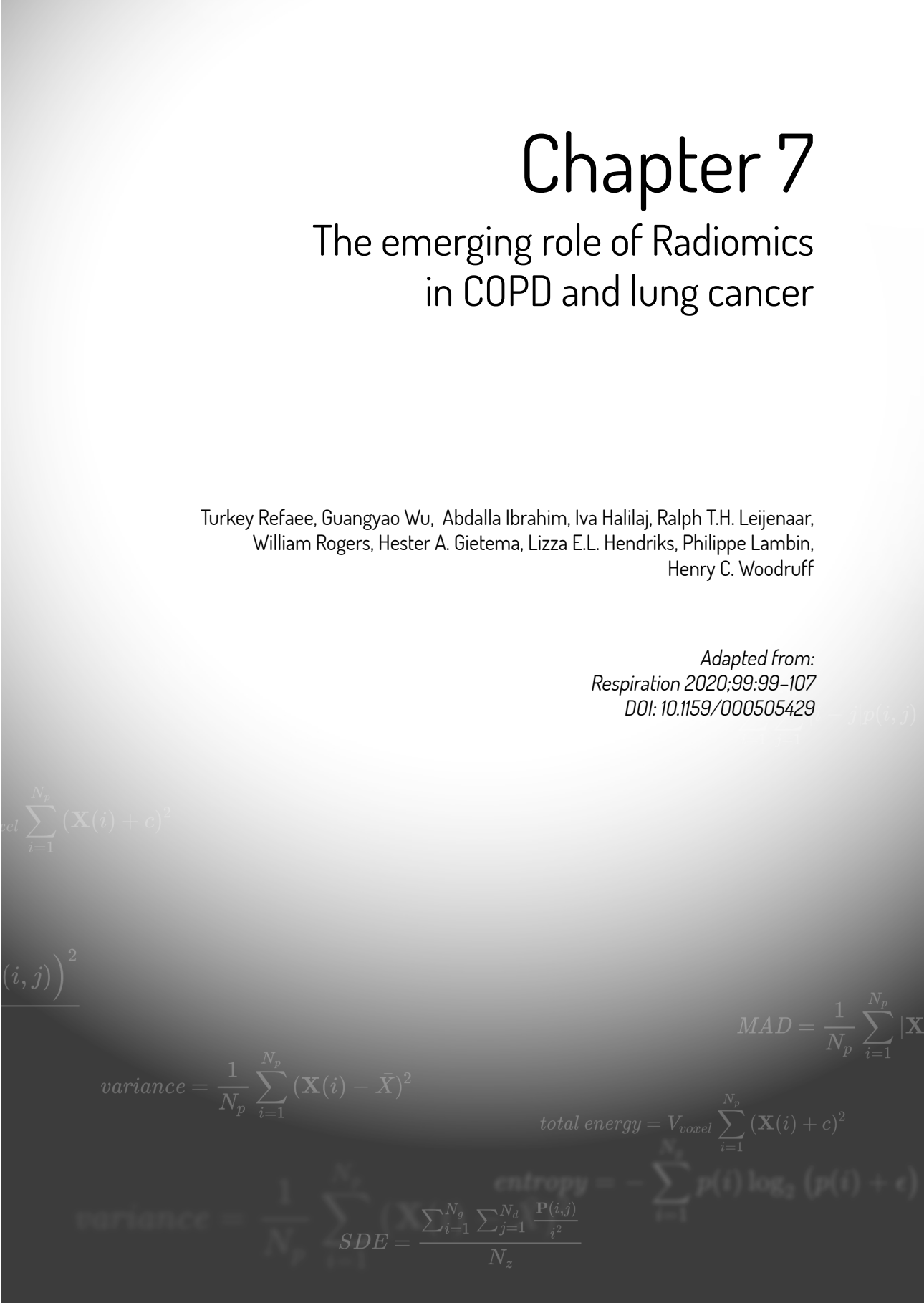
$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2$$

Chapter 7

The emerging role of Radiomics in COPD and lung cancer

Turkey Refaee, Guangyao Wu, Abdalla Ibrahim, Iva Halilaj, Ralph T.H. Leijenaar, William Rogers, Hester A. Gietema, Lizza E.L. Hendriks, Philippe Lambin, Henry C. Woodruff

Adapted from:
Respiration 2020;99:99–107
DOI: 10.1159/000505429



Abstract

Medical imaging plays a key role in evaluating and monitoring lung diseases such as chronic obstructive pulmonary disease (COPD) and lung cancer. The application of artificial intelligence (AI) in medical imaging has transformed medical images into mineable data, by extracting and correlating quantitative imaging features with patients' outcomes and tumor phenotype – a process termed radiomics. While this process has already been widely researched in lung oncology, the evaluation of COPD in this fashion remains in its infancy. Here we outline the main applications of radiomics in lung cancer and briefly review the workflow from image acquisition to the evaluation of model performance. Finally, we discuss the current assessments of COPD and the potential application of radiomics in COPD.

Introduction

Chronic obstructive pulmonary disease (COPD) is one of the most prevalent lung diseases, with an estimated 328 million people worldwide being affected, and in two decades it is expected to become the leading cause of death globally [1]. COPD is characterised by the limitation of airflow, which can be measured using spirometry. It is not completely reversible and is often caused by exposure to noxious particles or gas (e.g. cigarette smoking) which creates an inflammatory response in the lung. [2, 3]. COPD is a multicomponent disease comprising of a combination of bronchiolitis, emphysema and extrapulmonary effects [4]. While spirometry can measure airflow limitation, the contributions of large and small airway involvement and the extent and contribution of parenchyma destruction cannot be assessed [5]. Imaging by means of computed tomography (CT) has an increasing role in evaluation of COPD since CT-features can suggest the presence and severity of COPD. These features can be assessed visually [6] , but research is in advanced stages to automate the quantification of emphysema extent and distribution [7-10], airway wall thickness [11], and small airways disease [12].

Lung cancer is the other predominant lung disease, being one of the world's most prevalent cancers [13-16]. Globally, lung cancer is the most commonly diagnosed cancer (around 11% of all cancers in both sexes), and the world's leading cause of cancer related mortality (around 18% of total cancer related mortality) [17]. Lung cancers can be divided into two broad groups, small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC) [18]. NSCLC can be further divided into subgroups according to histopathology into squamous cell carcinoma (SCC) and adenocarcinoma (ADC) [19]. COPD has been shown to be a major additional risk factor for the development of lung cancer, specifically squamous cell carcinoma [20, 21]. Discovering the link between COPD and lung cancer has drawn significant attention in recent years [22]. It has been shown that COPD and lung cancer share similar pathological processes [23], while smoking cigarettes is one important common factor that causes both COPD and lung cancer [20], and patients with COPD and NSCLC have poor survival outcomes compared to NSCLC patients without COPD [24]. The link of pathophysiologic mechanisms between COPD and lung cancer is still not well understood (Fig. 1)[25].

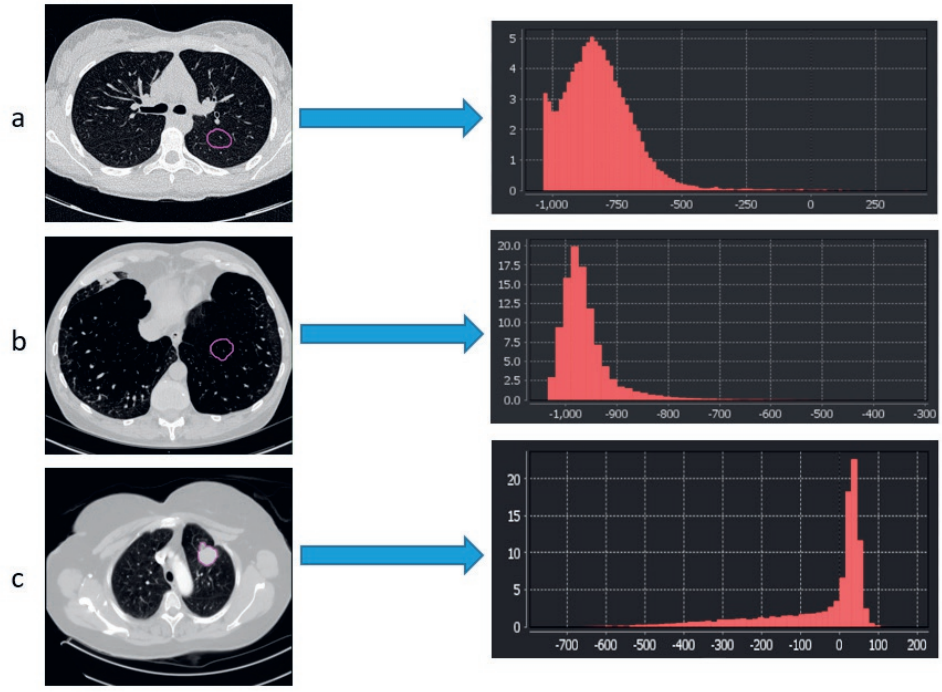


Figure 1. Different distributions of HU values extracted from the ROI (purple outline) for a) normal tissue, b) COPD tissue, and c) lung tumor.

The treatment of patients suffering from either disease would be greatly improved by personalised approaches, where patients are treated based on their and their diseases' individual characteristics rather than sub-population statistics gained from clinical trials. Which role artificial intelligence will play on the path to this paradigm shift towards individualised treatment selection is being extensively investigated [26]. For example, biopsies are used in clinical practice to phenotype the tumor, but the heterogeneous nature of cancer cells limits the biopsy's capacity to fully capture its condition [27, 28]. Medical imaging, on the other hand, has the potential to noninvasively asses the phenotypic differences of tumors in three dimensions [29] and has recently experienced great advances in the field of AI [30, 31]. In particular, radiomics, or quantitative image analysis (QIA) – the high-throughput extraction of quantitative features from medical images and their correlation with diagnostic and prognostic outcomes – has been researched to decode tumor phenotypes from a number of modalities such as computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET). Thousands of quantitative radiomic features can be extracted from each region of interest (ROI) and further analysed using machine learning tools to investigate correlations with biological and clinical endpoints [32-37]. Therefore, the application of radiomics to both COPD and lung cancer may improve the clinical workflow in diagnosing, managing, and following up the

patients. It can provide non-invasive, reliable and cost-effective clinical decision support systems, decreasing the need for invasive procedures.

The workflow of radiomics

The process of handcrafted radiomics consists of several steps (Fig 2): (1) collection of medical imaging (e.g CT, MR, PET/CT) for the target population; (2) segmentation of the region of interest (ROI) to be investigated; (3) extraction of radiomic features from the ROI; (4) the selection of radiomic features that best correlate with the outcome of interest ; (5) building the radiomics signature, and (6) evaluation of the model performance on various datasets using different metrics such as the receiver operating characteristic (ROC), area under the curve (AUC), and the precision-recall curve (PRC). The workflow has been previously described in detail [30, 37, 38].

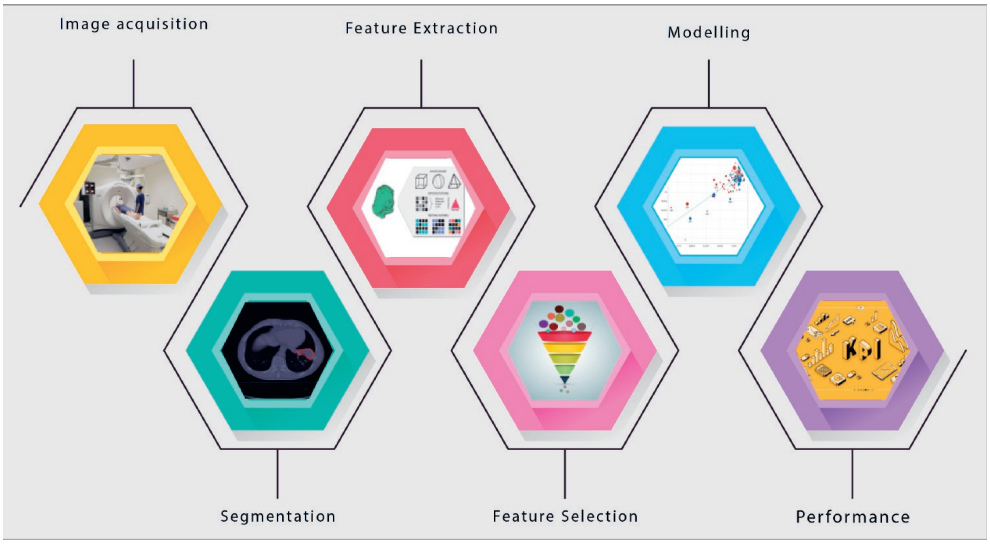


Figure 2. Graphic depiction of the radiomics workflow

Radiomics studies quality

Despite the potential of radiomics to facilitate precision medicine as highlighted in numerous publications, a number of obstacles still limits the generalizability of radiomics signatures, and thus their translation to clinical applications. The most important and widely known limitation is the lack of reproducibility for radiomics biomarkers [39-41]. Several studies have investigated the stability of radiomic features with test-retest experiments [42-44], and reported that a considered percentage of features is not reproducible in test-retest settings, i.e. using the same acquisition and reconstruction parameters on the same vendor for acquiring the scan. A study by Zhovannic et al [45] demonstrated that 62 of radiomic features are sensitive to differences in acquisition and reconstruction parameters using

the same imaging vendor. Other studies investigated the sensitivity of radiomic features to differences in segmentations, or what is known as inter-observer variability [46].

As such, efforts must be made to unify image acquisition and reconstruction across different centres to facilitate quantitative imaging analysis research, and integrate these methods into clinical decision support systems.

Several guidelines have been proposed to ensure that radiomic studies are methodologically sound and reproducible. Clear reporting in radiomics research is required to minimize bias and enhance the general application of prediction models. For instance, Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) initiative has established several recommendations in terms of reporting of the methodology of prediction models [47]. The Radiomics Quality Score (RQS) is, however, established specifically for radiomics research [38]. RQS is a checklist that contains sixteen elements to evaluate the design and reporting of a radiomics study. RQS guidelines include robust segmentation, the stability of test-retest, description of imaging protocol used, and internal/external validation. Due to the fast pace of advancement in this field, further improvement in the standardization of this score is required to ensure a high quality workflow. Furthermore, Image Biomarker Standardization Initiative (IBSI) is a newly formed guidelines to address the standardization of feature calculation and image pre-processing [48].

Role of radiomics in lung cancer

Diagnosis

Several studies have explored the use of radiomics in the screening of lung cancer. The advent of low-dose (LDCT) has altered the landscape of lung-cancer screening. Studies indicate that LDCT imaging, unlike molecular markers in blood, sputum, and bronchial brushings detects many tumors at early stages. For instance, The National Lung Screening Trial (NLST) in the United States demonstrated in a large population of 53.454 participants at a high risk for lung cancer, a 20% relative reduction in mortality when participants underwent three annual screening (LDCT) scans instead of single-view posterior- anterior chest radiography [49]. Kumar et al. used LIDC-IDRI dataset in order to differentiate between benign and malignant lesions, resulting in sensitivity and specificity of 79.06% and 76.11, respectively [50]. Other publications already shown promising results in the diagnosis of lung cancer [51-53].

Staging

Tumor node metastasis (TNM) staging of lung cancer is also important for cancer treatment. Several studies showed the added value of radiomic features in lung cancer staging. A study by Aerts et al. that included 1,019 patients to extract 440 CT radiomics per patient reported that radiomic features were associated with the overall stage (TNM) of lung cancer [54]. A

study by Wu et al. that used radiomic characteristics extracted from PET/CT to predict the early stage of distant metastasis (DM) in 101 early-stage NSCLC patients showed that PET radiomic features correlated with DM, and have added value in M staging [55]. Coroller et al. applied radiomics on 182 lung adenocarcinoma in order to predict (DM) showing that radiomics performed well on M staging [35].

Genetics and histopathology

Besides diagnosing and staging lung cancer, the use of radiomics has been extended to predict gene mutation or different pathology types of lung cancer. A study by Zhange et al. that included 298 patients found a correlation between EGFR mutation and CT radiomics features [56]. Liu et al. achieved the same results [57, 58]. Rios et al. developed a radiomic models that classifies mutations in patients with lung adenocarcinoma. The research found that radiomic signature based on CT images can predict EGFR status effectively [59]. Wu et al. used two NSCLC cohorts from Netherlands to predict the histologic types of lung cancer (ADC, SCC) [52].

Response to therapy

The use of radiomics signatures could be used to predict the response of patients to particular therapy. In a study by Aerts et al. it was reported that radiomics features obtained from CT images before treatment were able to predict the mutation status of EGFR in NSCLC and correlate with gefitinib response [60]. Coroller et al. showed that radiomic features based-CT images acquired prior to treatment could predict the pathological response to chemoradiation in NSCLC patients [61]. Mathhonen et al. predicted the recurrence of lung cancer following receiving Stereotactic Ablative Radiation Therapy (SART) using radiomics [62, 63]. Another study that utilized delta-radiomics, a method of analysing the difference of radiomic features obtained from longitudinal scans, in Stage III NSCLC patients to predict the outcome during radiation therapy, reported that the change in radiomic features values might be linked to the tumor response due to exposure to radiation [64]. Hao et al. used radiomic characteristics of peritumoral tissue derived from PET images to study its correlation with distant failure in NSCLC and cervical cancer (CC) [65]. The results showed a relationship between tumor boundaries and distant failure, suggesting that such an approach might be useful in predicting early response to radiotherapy in NSCLC and CC patients. In a recent study by Khorramin et al. CT-based radiomic features were extracted from peri- and intratumoral lung adenocarcinoma tissue and shown to have the potential to predict the response to chemotherapy, and correlated with both time to progression (TTP) and overall survival for patient with NSCLC [66]

Prognosis

Several studies investigated the prognosis of lung cancer using a radiomics approach. Coroller et al. found a prognostic relation between radiomics features and distant metastasis (DM) and survival in patients with lung cancer [67]. Aerts et al. found an association between the prognosis of lung cancer and radiomics features [54]. Balagurunathan et al. showed a correlation between the prognosis of lung cancer and radiomic features [42]. Song et al. showed a connection between features extracted from CT images and overall survival in NSCLC patients [68].

Potential translation of radiomics in COPD

The heterogeneous nature of COPD makes diagnosis challenging. However, it is crucial to unravel this variety of presentations to achieve an accurate diagnosis in early stages and help improve patients' outcomes [5]. Different COPD assessments are used in clinical practice, including pulmonary function test (PFTs) and quantitative CT (QCT). Pulmonary function test (PFTs) are essential to diagnose and classify COPD. A commonly used PTF is spirometry, which is used to measure the forced expiratory volume in 1 second (FEV1) and the forced vital capacity (FVC) as the primary parameters [69]. However, spirometry alone does not provide any locational information regarding emphysema [69]. Quantitative CT (QCT) is a promising approach that is able to quantify emphysema, airways abnormalities, and air trapping [5]. QCT has already demonstrated the capacity to evaluate the existence and degree of emphysema [70-76]. For example, CT densitometry parameters such as relative low-attenuation area [77-82] and percentile of the frequency – attenuation distribution [9, 83-85] are usually used to assess the degree of emphysema. Airways abnormalities are commonly measured by the calculation of the square root at an internal perimeter of 10mm (Pi10) using linear regression [86-89]. It is considered the gold standard tool and has already demonstrated significant correlation with the histological measurement of small airways [90]. Air trapping appears as decreased attenuation on expiratory CT images [91], making it the best way to evaluate air trapping in COPD [88]. The measurements of gas trapping using CT are highly correlated with PTF in COPD patients [92]

Despite the ability of QCT to quantify COPD, the interpretation of QCT is still time-consuming, qualitative, requires experts, and is prone to variability in the diagnosis between experts. CT image metric (radiomics) approach could potentially quantify COPD and uncover the disease's hidden mechanism and the link between lung cancer and COPD in more nuance and more powerful phenotypic classification. A radiomics signature would be easier to apply as a clinical decision support system (cDSS), and less time consuming compare to currently used QCT. Therefore, several potential applications for radiomic features in COPD are suggested. Texture analysis for example has shown its effectiveness in assessing the degree of emphysema. A study by Ginsburg et al. demonstrated the effectiveness of texture-based approach in classifying between the lungs of never-smokers, smokers without emphysema,

and smokers with emphysema, indicating that an early stage of smoking-related lung injury could potentially be identified before emphysema develops [93]. Another study by Castadi et al. used a local histogram-based technique to quantify distinct emphysema pattern using CT scans from 9,313 smoker subjects in the COPODGene study [94]. The results of the study suggests that information extracted from CT pattern of emphysema were more predictive than threshold-based emphysema measurements such as "low attenuation area less than -950" (LAA-950). As described above, the applications of radiomics in the screening of lung cancer showed interesting results. Automated screening of routine chest CT to diagnose COPD is therefore one possible use, with the ability to detect suspected sarcopenia not only in the lung but also in the muscle tissue. Detection and differentiation between COPD stages and phenotypes, especially in early stages, will allow for the early and suitable treatment for the patient. In a study by Lafata et al., the authors reported on the potential of radiomic features extracted from CT images to quantify the changes in lung function and associated with spirometry test [95]. The same approach using radiomics could be extended to investigate its relationship with other gold standard COPD markers such as waking exams, FEV/FVC ratio (Tiffeneau index) or to the frequency of exacerbations associated with COPD patients, enabling an accurate diagnose of COPD severity. In addition, the use of radiomics could improve the performance of the existing multifactorial models (nomograms) by adding radiomics features to existing clinical factors (age, sex, number of pack-years, current smoking, performance score, wheezing) as already shown in a previous publication [96]. Delta-radiomics has already demonstrated its ability to predict response to therapy in lung cancer. Therefore, such a technique could be used to identify quantitatively the evolution of the disease and the effect of (new) treatment. Additionally, delta-radiomics could be applied to assess the difference between inspiration and expiration scans and to explore hidden information that could help in evaluating the extent and severity of pulmonary emphysema, air trapping, and airway abnormalities. The use of radiomics potentially could be used to predict whether patient will respond to certain interventions, such as endoscopic lung volume reduction (ELVR), and inhalation steroids.

Conclusion

The field of radiomics is rapidly growing and has already shown its potential in assessing lung cancers in terms of detection, treatment response, and prognosis. Different QCT measurements have been used to quantify COPD abnormalities such as emphysema, air trapping, and airway remodelling. Applying radiomics in COPD has not been extensively investigated yet. We show examples of the potential use of radiomics in the diagnosis, treatment and the follow-up of COPD and future directions for further research.

References

1 Quaderi S, Hurst J: The unmet global burden of COPD. Global health, epidemiology and genomics 2018;3

2 Rabe KF, Hurd S, Anzueto A, et al: Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. Am J Respir Crit Care Med 2007;176:532-555.

3 Celli BR, MacNee W: Standards for the diagnosis and treatment of patients with COPD: a summary of the ATS/ERS position paper. Eur Respir J 2004;23:932-946.

4 Vogelmeier CF, Criner GJ, Martinez FJ, et al: Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease 2017 Report. GOLD Executive Summary. Am J Respir Crit Care Med 2017;195:557-582.

5 Mets O, De Jong P, Van Ginneken B, et al: Quantitative computed tomography in COPD: possibilities and limitations. Lung 2012;190:133-145.

6 Bodduluri S, Reinhardt JM, Hoffman EA, et al: Recent Advances in Computed Tomography Imaging in Chronic Obstructive Pulmonary Disease. Annals of the American Thoracic Society 2018;15:281-289.

7 Muller NL, Staples CA, Miller RR, et al: Density Mask - an Objective Method to Quantitate Emphysema Using Computed-Tomography. Chest 1988;94:782-787.

8 Gevenois PA, de Maertelaer V, De Vuyst P, et al: Comparison of computed density and macroscopic morphometry in pulmonary emphysema. Am J Respir Crit Care Med 1995;152:653-657.

9 Madani A, Zanen J, de Maertelaer V, et al: Pulmonary emphysema: objective quantification at multi-detector row CT--comparison with macroscopic and microscopic morphometry. Radiology 2006;238:1036-1043.

10 Gietema HA, Schilham AM, van Ginneken B, et al: Monitoring of smoking-induced emphysema with CT in a lung cancer screening setting: detection of real increase in extent of emphysema. Radiology 2007;244:890-897.

11 de Jong PA, Muller NL, Pare PD, et al: Computed tomographic imaging of the airways: relationship to structure and function. Eur Respir J 2005;26:140-152.

12 Mets OM, Zanen P, Lammers J-WJ, et al: Early identification of small airways disease on lung cancer screening CT: comparison of current air trapping measures. Lung 2012;190:629-633.

13 Fabre Ballalai Ferraz A, Rosim R, Anaya P: Standardization Process Of Raw Datasus And Consumption Analysis Of Oncology Therapies In The Brazil Public Health Care System: A Comparison Between Raw And Standardized Dataset In Colorectal And Lung Cancer. Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research 2015;18:A811.

14 Faris N, Yu X, Sareen S, et al: Preoperative Evaluation of Lung Cancer in a Community Health Care Setting. The Annals of thoracic surgery 2015;100:394-400.

15 Ryoo JJ, Malin JL, Ordin DL, et al: Facility characteristics and quality of lung cancer care in an integrated health care system. J Thorac Oncol 2014;9:447-455.
16 Sundaram B, Kazerooni EA: Preface. Lung cancer is an important public health care issue. Radiologic clinics of North America 2012;50:xi.
17 Bray F, Ferlay J, Soerjomataram I, et al: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians 2018;68:394-424.
18 Palma JF, Das P, Liesenfeld O: Lung cancer screening: utility of molecular applications in conjunction with low-dose computed tomography guidelines. Expert review of molecular diagnostics 2016;16:435-447.
19 Long F, Su J-H, Liang B, et al: Identification of gene biomarkers for distinguishing small-cell lung cancer from non-small-cell lung cancer using a network-based approach. BioMed research international 2015;2015
20 Durham A, Adcock I: The relationship between COPD and lung cancer. Lung cancer 2015;90:121-127.
21 Raviv S, Hawkins KA, DeCamp MM, Jr., et al: Lung cancer in chronic obstructive pulmonary disease: enhancing surgical options and outcomes. Am J Respir Crit Care Med 2011;183:1138-1146.
22 Houghton AM: Mechanistic links between COPD and lung cancer. Nature Reviews Cancer 2013;13:233.
23 Eapen MS, Hansbro PM, Larsson-Callerfelt A-K, et al: Chronic obstructive pulmonary disease and lung cancer: underlying pathophysiology and new therapeutic modalities. Drugs 2018;78:1717-1740.
24 Wang P, Zhu M, Zhang D, et al: The relationship between chronic obstructive pulmonary disease and non-small cell lung cancer in the elderly. Cancer medicine 2019
25 Chalela R, Gea J, Barreiro E: Immune phenotypes in lung cancer patients with COPD: potential implications for immunotherapy. J Thorac Dis 2018;10:S2186-s2189.
26 Jackson SE, Chester JD: Personalised cancer medicine. International journal of cancer 2015;137:262-266.
27 Gerlinger M, Rowan AJ, Horswell S, et al: Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. New England journal of medicine 2012;366:883-892.
28 Gerlinger M, Swanton C: How Darwinian models inform therapeutic failure initiated by clonal heterogeneity in cancer medicine. British journal of cancer 2010;103:1139.
29 Gillies RJ, Kinahan PE, Hricak H: Radiomics: images are more than pictures, they are data. Radiology 2015;278:563-577.
30 Ibrahim A, Vallières M, Woodruff H, et al: Radiomics Analysis for Clinical Decision Support in Nuclear Medicine: Seminars in Nuclear Medicine, Elsevier, 2019,
31 Walsh S, de Jong EE, van Timmeren JE, et al: Decision support systems in oncology. JCO clinical cancer informatics 2019;3:1-9.

32 Carvalho S, Leijenaar RT, Velazquez ER, et al: Prognostic value of metabolic metrics extracted from baseline positron emission tomography images in non-small cell lung cancer. Acta Oncol 2013;52:1398-1404.
33 Cook GJ, O'Brien ME, Siddique M, et al: Non-Small Cell Lung Cancer Treated with Erlotinib: Heterogeneity of (18)F-FDG Uptake at PET-Association with Treatment Response and Prognosis. Radiology 2015;276:883-893.
34 Cook GJ, Yip C, Siddique M, et al: Are pretreatment 18F-FDG PET tumor textural features in non-small cell lung cancer associated with response and survival after chemoradiotherapy? J Nucl Med 2013;54:19-26.
35 Coroller TP, Grossmann P, Hou Y, et al: CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. Radiother Oncol 2015;114:345-350.
36 Fried DV, Tucker SL, Zhou S, et al: Prognostic value and reproducibility of pretreatment CT texture features in stage III non-small cell lung cancer. Int J Radiat Oncol Biol Phys 2014;90:834-842.
37 Lambin P, Rios-Velazquez E, Leijenaar R, et al: Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 2012;48:441-446.
38 Lambin P, Leijenaar RTH, Deist TM, et al: Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 2017;14:749-762.
39 Leijenaar RT, Nalbantov G, Carvalho S, et al: The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis. Scientific reports 2015;5:11075.
40 van Velden FH, Kramer GM, Frings V, et al: Repeatability of radiomic features in non-small-cell lung cancer [18 F] FDG-PET/CT studies: impact of reconstruction and delineation. Molecular imaging and biology 2016;18:788-795.
41 Zhao B, Tan Y, Tsai W-Y, et al: Reproducibility of radiomics for deciphering tumor phenotype with imaging. Scientific reports 2016;6:23428.
42 Balagurunathan Y, Gu Y, Wang H, et al: Reproducibility and prognosis of quantitative features extracted from CT images. Translational oncology 2014;7:72-87.
43 Balagurunathan Y, Kumar V, Gu Y, et al: Test–retest reproducibility analysis of lung CT image features. Journal of digital imaging 2014;27:805-823.
44 Tixier F, Hatt M, Le Rest CC, et al: Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET. Journal of Nuclear Medicine 2012;53:693-700.
45 Zhovannik I, Bussink J, Traverso A, et al: Learning from scanners: Bias reduction and feature correction in radiomics. Clinical and translational radiation oncology 2019;19:33-38.
46 Yip SS, Aerts HJ: Applications and limitations of radiomics. Physics in Medicine & Biology 2016;61:R150.
47 Collins GS, Reitsma JB, Altman DG, et al: Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015;162:55-63.

48 Zwanenburg A, Leger S, Vallières M, et al: Image biomarker standardisation initiative. arXiv preprint arXiv:161207003 2016
49 Aberle DR, Adams AM, Berg CD, et al: Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011;365:395-409.
50 Kumar D, Shafiee MJ, G. Chung A, et al: Discovery Radiomics for Computed Tomography Cancer Detection. 2015
51 Liu Y, Balagurunathan Y, Atwater T, et al: Radiological Image Traits Predictive of Cancer Status in Pulmonary Nodules. Clin Cancer Res 2017;23:1442-1449.
52 Wu W, Parmar C, Grossmann P, et al: Exploratory study to identify radiomics classifiers for lung cancer histology. Frontiers in oncology 2016;6:71.
53 Maldonado F, Boland JM, Raghunath S, et al: Noninvasive characterization of the histopathologic features of pulmonary nodules of the lung adenocarcinoma spectrum using computer-aided nodule assessment and risk yield (CANARY)—a pilot study. Journal of Thoracic Oncology 2013;8:452-460.
54 Aerts HJ, Velazquez ER, Leijenaar RT, et al: Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature communications 2014;5:4006.
55 Wu J, Aguilera T, Shultz D, et al: Early-stage non–small cell lung cancer: quantitative imaging characteristics of 18F fluorodeoxyglucose PET/CT allow prediction of distant metastasis. Radiology 2016;281:270-278.
56 Zhang L, Chen B, Liu X, et al: Quantitative biomarkers for prediction of epidermal growth factor receptor mutation in non-small cell lung Cancer. Translational oncology 2018;11:94-101.
57 Liu Y, Kim J, Balagurunathan Y, et al: Radiomic Features Are Associated With EGFR Mutation Status in Lung Adenocarcinomas. Clin Lung Cancer 2016;17:441-448.e446.
58 Liu Y, Kim J, Qu F, et al: CT features associated with epidermal growth factor receptor mutation status in patients with lung adenocarcinoma. Radiology 2016;280:271-280.
59 Rios Velazquez E, Parmar C, Liu Y, et al: Somatic Mutations Drive Distinct Imaging Phenotypes in Lung Cancer. Cancer Res 2017;77:3922-3930.
60 Aerts HJ, Grossmann P, Tan Y, et al: Defining a radiomic response phenotype: a pilot study using targeted therapy in NSCLC. Scientific reports 2016;6:33860.
61 Coroller TP, Agrawal V, Narayan V, et al: Radiomic phenotype features predict pathological response in non-small cell lung cancer. Radiotherapy and Oncology 2016;119:480-486.
62 Mattonen S, Tetar S, Palma D, et al: Automated texture analysis for prediction of recurrence after stereotactic ablative radiation therapy for lung cancer. International Journal of Radiation Oncology• Biology• Physics 2015;93:S5-S6.
63 Mattonen SA, Palma DA, Haasbeek CJ, et al: Early prediction of tumor recurrence based on CT texture changes after stereotactic ablative radiotherapy (SABR) for lung cancer. Medical physics 2014;41:033502.
64 Fave X, Zhang L, Yang J, et al: Delta-radiomics features for the prediction of patient outcomes in non–small cell lung cancer. Scientific reports 2017;7:588.

65 Hao H, Zhou Z, Li S, et al: Shell feature: a new radiomics descriptor for predicting distant failure after radiotherapy in non-small cell lung cancer and cervix cancer. Phys Med Biol 2018;63:095007.
66 Khorrami M, Khunger M, Zagouras A, et al: Combination of Peri-and Intratumoral Radiomic Features on Baseline CT Scans Predicts Response to Chemotherapy in Lung Adenocarcinoma. Radiology: Artificial Intelligence 2019;1:180012.
67 Coroller TP, Grossmann P, Hou Y, et al: CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. Radiotherapy and Oncology 2015;114:345-350.
68 Song J, Liu Z, Zhong W, et al: Non-small cell lung cancer: quantitative phenotypic analysis of CT images as a potential marker of prognosis. Scientific reports 2016;6:38282.
69 Matsuoka S, Yamashiro T, Washko GR, et al: Quantitative CT assessment of chronic obstructive pulmonary disease. Radiographics 2010;30:55-66.
70 Bergin C, Müller N, Nichols DM, et al: The diagnosis of emphysema: a computed tomographic-pathologic correlation. American Review of Respiratory Disease 1986;133:541-546.
71 Coddington R, Mera S, Goddard P, et al: Pathological evaluation of computed tomography images of lungs. Journal of clinical pathology 1982;35:536-540.
72 Foster Jr W, Pratt P, Roggli V, et al: Centrilobular emphysema: CT-pathologic correlation. Radiology 1986;159:27-32.
73 Goddard PR, Nicholson E, Laszlo G, et al: Computed tomography in pulmonary emphysema. Clinical radiology 1982;33:379-387.
74 Kreel L: Computed tomography of the thorax. Medical imaging 1979:132-139.
75 Morgan M, Strickland B: Computed tomography in the assessment of bullous lung disease. British journal of diseases of the chest 1984;78:10-25.
76 Rosenblum LJ, Mauceri RA, Wellenstein DE, et al: Computed tomography of the lung. Radiology 1978;129:521-524.
77 Bankier AA, De Maertelaer V, Keyzer C, et al: Pulmonary emphysema: subjective visual grading versus objective quantification with macroscopic morphometry and thin-section CT densitometry. Radiology 1999;211:851-858.
78 Gevenois PA, De Vuyst P, de Maertelaer V, et al: Comparison of computed density and microscopic morphometry in pulmonary emphysema. Am J Respir Crit Care Med 1996;154:187-192.
79 Gurney JW, Jones KK, Robbins RA, et al: Regional distribution of emphysema: correlation of high-resolution CT with pulmonary function tests in unselected smokers. Radiology 1992;183:457-463.
80 Kinsella M, Müller NL, Abboud RT, et al: Quantitation of emphysema by computed tomography using a "density mask" program and correlation with pulmonary function tests. Chest 1990;97:315-321.
81 Knudson RJ, Standen JR, Kaltenborn WT, et al: Expiratory computed tomography for assessment of suspected pulmonary emphysema. Chest 1991;99:1357-1366.

82 Lucidarme O, Coche E, Cluzel P, et al: Expiratory CT scans for chronic airway disease: correlation with pulmonary function test results. AJR American journal of roentgenology 1998;170:301-307.
83 Dirksen A, Dijkman JH, Madsen F, et al: A randomized clinical trial of α1-antitrypsin augmentation therapy. Am J Resp Crit Care 1999;160:1468-1472.
84 Dirksen A, Friis M, Olesen K, et al: Progress of emphysema in severe α1-antitrypsin deficiency as assessed by annual CT. Acta Radiologica 1997;38:826-832.
85 Gould G, MacNee W, McLean A, et al: CT measurements of lung density in life can quantitate distal airspace enlargement—an essential defining feature of human emphysema. American Review of Respiratory Disease 1988;137:380-392.
86 Grydeland TB, Dirksen A, Coxson HO, et al: Quantitative computed tomography: emphysema and airway wall thickness by sex, age and smoking. Eur Respir J 2009;34:858-865.
87 Patel BD, Coxson HO, Pillai SG, et al: Airway wall thickening and emphysema show independent familial aggregation in chronic obstructive pulmonary disease. Am J Respir Crit Care Med 2008;178:500-505.
88 Smith BM, Hoffman EA, Rabinowitz D, et al: Comparison of spatially matched airways reveals thinner airway walls in COPD. The Multi-Ethnic Study of Atherosclerosis (MESA) COPD Study and the Subpopulations and Intermediate Outcomes in COPD Study (SPIROMICS). Thorax 2014;69:987-996.
89 Woodruff PG, Couper D, Han MK: Symptoms in Smokers with Preserved Pulmonary Function. N Engl J Med 2016;375:896-897.
90 Nakano Y, Wong JC, de Jong PA, et al: The prediction of small airway dimensions using computed tomography. Am J Respir Crit Care Med 2005;171:142-146.
91 Matsuoka S, Kurihara Y, Yagihashi K, et al: Quantitative assessment of air trapping in chronic obstructive pulmonary disease using inspiratory and expiratory volumetric MDCT. AJR Am J Roentgenol 2008;190:762-769.
92 Schroeder JD, McKenzie AS, Zach JA, et al: Relationships between airflow obstruction and quantitative CT measurements of emphysema, air trapping, and airways in subjects with and without chronic obstructive pulmonary disease. AJR Am J Roentgenol 2013;201:W460-470.
93 Ginsburg SB, Lynch DA, Bowler RP, et al: Automated texture-based quantification of centrilobular nodularity and centrilobular emphysema in chest CT images. Academic radiology 2012;19:1241-1251.
94 Castaldi PJ, San Jose Estepar R, Mendoza CS, et al: Distinct quantitative computed tomography emphysema patterns are associated with physiology and function in smokers. Am J Respir Crit Care Med 2013;188:1083-1090.
95 Lafata KJ, Zhou Z, Liu J-G, et al: An exploratory Radiomics Approach to Quantifying pulmonary function in ct images. Scientific reports 2019;9:1-9.
96 Huang YQ, Liang CH, He L, et al: Development and Validation of a Radiomics Nomogram for Preoperative Prediction of Lymph Node Metastasis in Colorectal Cancer. J Clin Oncol 2016;34:2157-2164.

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$minimum = \min(\mathbf{X})$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2$$

$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_s} \mathbf{P}(i,j)\right)^2}{N_p}$$

$$correlation = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) ij - \mu_x \mu_y}{\sigma_x(i) \sigma_y(j)}$$

$$MAD = \frac{1}{N_p} \sum_{i=1}^{N_p} |\mathbf{X}(i) - \bar{X}|$$

$$minimum = \min(\mathbf{X})$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_s} \mathbf{P}\right.}{N_z}$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2$$

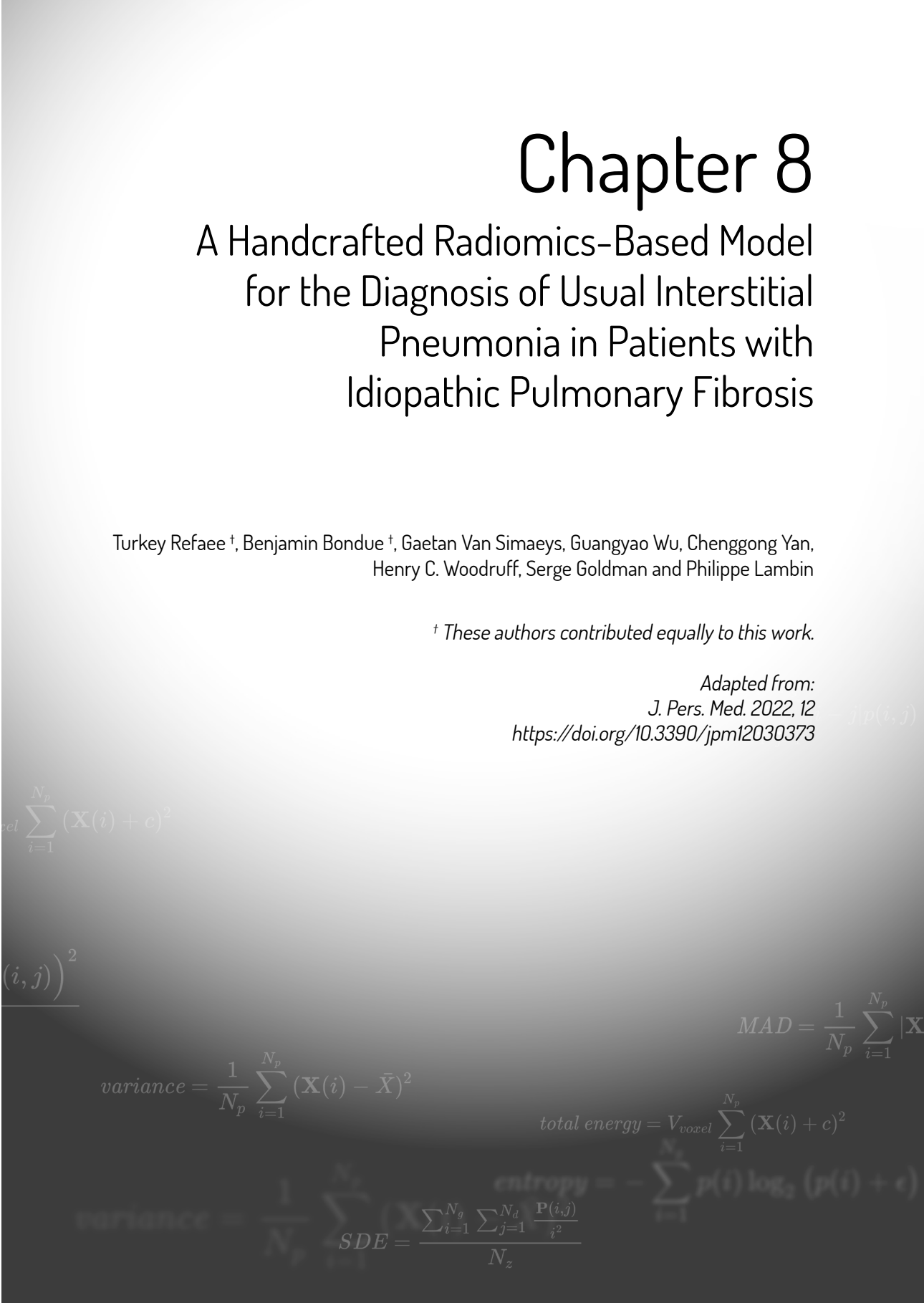
Chapter 8

A Handcrafted Radiomics-Based Model for the Diagnosis of Usual Interstitial Pneumonia in Patients with Idiopathic Pulmonary Fibrosis

Turkey Refaee †, Benjamin Bondue †, Gaetan Van Simaeys, Guangyao Wu, Chenggong Yan, Henry C. Woodruff, Serge Goldman and Philippe Lambin

† *These authors contributed equally to this work.*

Adapted from:
J. Pers. Med. 2022, 12
https://doi.org/10.3390/jpm12030373



Abstract

The most common idiopathic interstitial lung disease (ILD) is idiopathic pulmonary fibrosis (IPF). It can be identified by the presence of usual interstitial pneumonia (UIP) via high-resolution computed tomography (HRCT) or with the use of a lung biopsy. We hypothesized that a CT-based approach using handcrafted radiomics might be able to identify IPF patients with a radiological or histological UIP pattern from those with an ILD or normal lungs. A total of 328 patients from one center and two databases participated in this study. Each participant had their lungs automatically contoured and sectorized. The best radiomic features were selected for the random forest classifier and performance was assessed using the area under the receiver operator characteristics curve (AUC). A significant difference in the volume of the trachea was seen between a normal state, IPF, and non-IPF ILD. Between normal and fibrotic lungs, the AUC of the classification model was 1.0 in validation. When classifying between IPF with a typical HRCT UIP pattern and non-IPF ILD the AUC was 0.96 in validation. When classifying between IPF with UIP (radiological or biopsy-proved) and non-IPF ILD, an AUC of 0.66 was achieved in the testing dataset. Classification between normal, IPF/UIP, and other ILDs using radiomics could help discriminate between different types of ILDs via HRCT, which are hardly recognizable with visual assessments. Radiomic features could become a valuable tool for computer-aided decision-making in imaging, and reduce the need for unnecessary biopsies.

Keywords: handcrafted radiomics; interstitial lung diseases; usual interstitial pneumonia; machine learning

Introduction

Idiopathic pulmonary fibrosis (IPF) is the most common progressive form of interstitial lung disease (ILD) with an unknown etiology, usually impacting older adults [1,2]. In 2011, four societies—the American Thoracic Society, the European Respiratory Society, the Japanese Respiratory Society, and the Latin American Thoracic Association—came together to issue an evidence-based statement, which provided recommendations for both the diagnosis and management of IPF [3]. According to these recommendations, high-resolution computed tomography (HRCT) can play a crucial role in the diagnosis of fibrotic lung diseases and has a significant impact on medical decision-making.

Diagnosing IPF comes about using a multidisciplinary discussion (MDD) of the clinical, radiological, and, if available, pathological data showing a usual interstitial pneumonia (UIP) pattern which is the most common histopathological form of diffuse lung fibrosis [3,4]. The diagnostic radiological characteristic of UIP necessitates honeycombing with a basal and subpleural predominance. The upper lobes are less affected, and traction bronchiectasis may be present [5]. An IPF diagnosis requires a multidisciplinary discussion (MDD) and the exclusion of known causes of ILD, in addition to the presence of a UIP-specific pattern on thin-section CT, or a specific combination of HRCT patterns and histopathological UIP patterns in patients subjected to lung tissue sampling [3]. It is also worth noting that, in 2018, the Fleischner Society expanded on these recommendations for diagnosing IPF to include the appearance of probable UIP in HRCTs, if the clinical context was consistent with an IPF [6].

Surgical lung biopsy (SLB), which is recommended when no UIP pattern is present on the HRCT [3,7], is an invasive procedure that requires pleural drainage and is associated with a mortality rate ranging from 2.0% to 3.6% [8–13]. Moreover, in a recent study that included a cohort of patients with pathologically-proven UIP patterns, radiologists only identified a UIP pattern on thin-section CT with a sensitivity of 34% [14], according to the recent ATS-ERS guidelines [15]. Furthermore, the radiological assessment of fibrotic lung diseases is still challenging and often varies between experts [16–19]. Consequently, an automated approach that assists radiologists (especially less experienced ones) could be very useful in avoiding unnecessary biopsies in a context of a multidisciplinary discussion.

The interest in radiomics, pioneered in 2012, has increased in recent years [20]. The field of handcrafted radiomics, briefly stated, extracting a large number of mineable quantitative data from medical images using predetermined formulas, has developed rapidly in recent times [20]. The term radiomics (handcrafted radiomics and deep learning) refers to the high-throughput extraction of numeric features from medical imaging modalities, providing high-dimensional data that could be used to identify patterns relating to the pathophysiology

of a disease. These data could then be merged with the characteristics of each patient to aid clinical decision-making [20,21]. Different studies have shown that radiomics has the potential to complement clinical decision support systems, for example, for cancer diagnosis and prognosis [20,22–24]. These studies have shown some potential to function as imaging biomarkers and to predict clinical outcomes and drug responses [20,25–27]. While the potential of radiomics has mainly been investigated in oncology, it can also be applied to many other diseases, including ILDs and chronic obstructive pulmonary disease (COPD) [28–30].

We hypothesize that radiomic features are able to decode biological information from specified regions of interest within the lung that can be used to diagnose IPF with UIP pattern. The aims of this study are two-fold: (1) to evaluate the use of radiomics, to differentiate between normal lung tissue and ILDs; (2) to evaluate the use of radiomics to distinguish IPF with a typical or less typical (biopsy-proven) UIP pattern related to IPF from HRCT patterns related to non-IPF ILDs. We also conjecture, based on the literature [31], that tracheal enlargement and tracheal shape would significantly complement handcrafted radiomic features that would help in the classification of different types of ILDs.

Materials and Methods

Study Population

The study protocol was registered on clinicaltrials.gov (identifier: NCT04430491), approved by the ethics committee of the Erasme University hospital (ref: P2017/411). The electronic medical records at Erasme University hospital (center i) were searched between 2011 and 2018 for patients diagnosed with ILD. The inclusion criteria were: (i) the availability of HRCT with slices of less than 1.5 mm; (ii) the availability of a high-confidence diagnosis (MDD diagnosis of IPF with a typical UIP pattern; MDD diagnosis of IPF with a biopsy-proven UIP pattern; or MDD diagnosis of non-IPF ILD, validated by a lung biopsy showing a pattern other than UIP). The exclusion criteria were (i) the use of contrast enhancements in HRCT; (ii) images containing metal or motion artifacts; and (iii) images reconstructed with a slice thickness larger than 1.5 mm (Figure 1). At least 1 chest physician, 1 pathologist, 1 thoracic radiologist, 1 specialist in internal medicine or rheumatology participated in the MDD. For external validation (database A), we used the group of patients diagnosed with interstitial lung diseases from the publicly available Lung Tissue Research Consortium (LTRC, https://ltrcpublic.com/ (accessed on 19 September 2018)). Images from patients with ostensibly healthy lungs (database B) were collected from the publicly available Radiomics Imaging Archive (RIA, https://www.radiomicsimagingarchive.eu/ (accessed on 24 October 2021)) (G4). Information was also gathered from patients, such as the demographic (age, gender) and clinical data (body mass index—BMI), as well as the measurements of pulmonary function tests (PFT) (forced expiratory volume in 1s (FEV1), forced vital capacity (FVC), and

diffusion capacity of carbon monoxide (DLCO). The so-called gender, age, and pulmonary function (GAP) score and staging system that was developed by Ley et al. in 2012 [32] was calculated for each patient and the value was recorded.

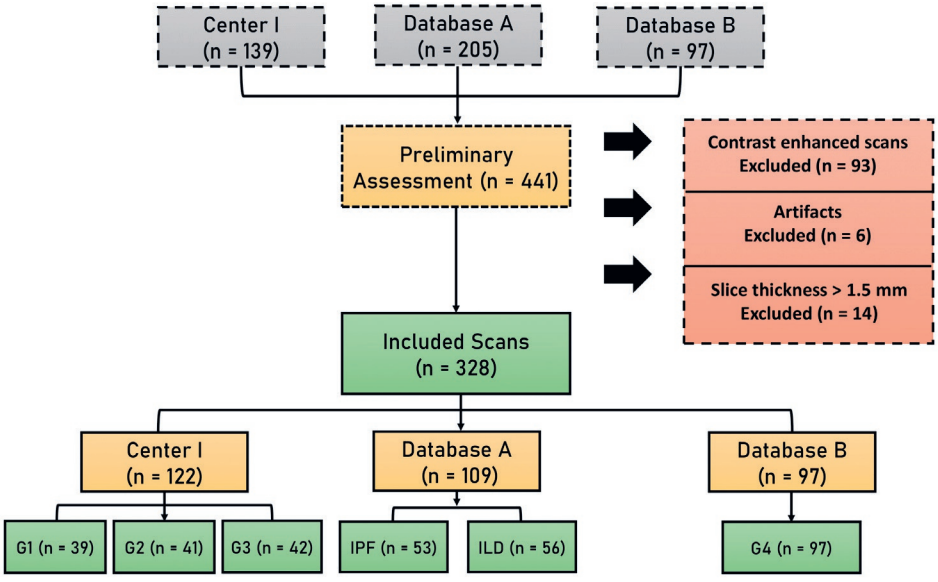


Figure 1. A flowchart diagram shows the patient selection process. (G1) patients with final MDD diagnosis of IPF with typical UIP pattern in HRCT and no lung biopsy; (G2) patients with a final MDD diagnosis of IPF confirmed by Surgical Lung Biopsy (SLB) (less typical HRCT pattern); (G3) patients with ILDs other than IPF with lung biopsy confirming a non-UIP pattern; (G4) patients with apparently healthy lungs.

High-Resolution CT (HRCT) Scanning

For center i, the HRCTs were acquired on a 64- or 128-detector row CT system (Somatom, Definition, Siemens Healthineers, Erlangen, Germany). For database A, HRCT images were acquired using 4 different CT vendors (Siemens, Erlangen, Germany), (GE, Waukesha, USA), (Philips, Amsterdam, the Netherlands), and (Toshiba, Tochigi-ken, Japan). For database B, all scans were acquired from the same scanner (GE Medical Systems, Waukesha, USA). The slice thickness of all scans varied between 0.5 and 1.5 mm.

Segmentation

The process of delineating a region of interest (ROI) that will be utilized to extract handcrafted radiomic features is known as segmentation. A workflow for radiomics from segmentation to data analysis is depicted in Figure 2. Segmentation of the lungs and sectors, as well as the tracheobronchial tree, were performed automatically using an automated workflow created with MIM software (MIM Software Inc., Cleveland, OH, USA). Sectorized lung segmentation was performed to account for the differences in the spatial distribution of the lesions between UIP and non-UIP patterns. Each sector was defined as a (ROI). As shown in the left

part of Figure 2, sectors 1 and 2 represent the upper section of the lung, sector 3 represents the middle section, and sector 4 represents the basal section.

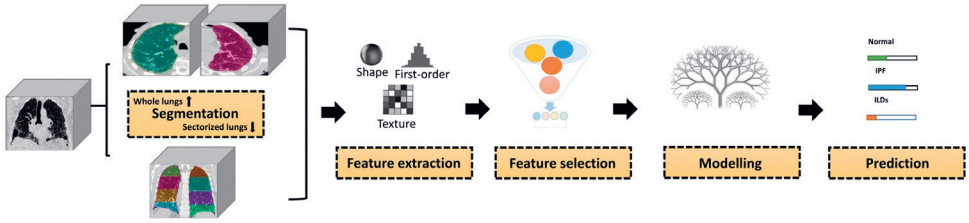


Figure 2. Radiomics Pipeline for lung fibrosis classification from CT images. First, the region of interest (ROI) was delineated. Second, handcrafted radiomic features were extracted from both ROIs. Third, feature selection methods were applied to select the most informative set of features. Fourth, the selected set of features were train the Random Forest classifier to arrive at a prediction.

Radiomic Features Extraction

To minimize the effects of the variations in image voxel size, all HRCT images were resampled into $1 \times 1 \times 1\ mm^3$ voxel size, using linear interpolation to address the disparate reconstruction settings found in the datasets [33]. $1 \times 1 \times 1\ mm^3$ was the maximum voxel size available in the dataset [34]. Radiomic features, except for the trachea volume, were extracted from the ROIs of the lung and sectors within the HRCT images, using the RadiomiX Discovery Toolbox (version, October 2019; https://www.radiomics.bio (accessed on 23 June 2020)), which calculates radiomics features in compliance with the Imaging Biomarkers Standardization Initiative (IBSI) [35]. Voxel intensities were aggregated into bins of 25 Hounsfield Units (HUs)—for nonfiltered features, excluding first-order statistics features—to reduce noise and interscanner variability [36]. The extracted features describe the fractal dimension, intensity histogram, first-order statistics, texture, and shape. Mathematical definitions and descriptions of the features mentioned can be found in other studies [21].

Data Splitting

For the first aim, i.e., normal vs. ILDs (G4 vs. G1,2,3), the data from center (i) and database B was combined and split into training and validation datasets, with a ratio of 0.8:0.2. For the second aim, i.e., IPF/UIP vs. non-IPF ILDs (G1 and 2 vs. G3), datasets from center (i) were randomly divided into training and validation dataset, using a ratio of 0.8:0.2, and data from database A was used as an external validation dataset.

Feature Selection and Modeling

To avoid any information leaking, all of the feature selection and model training was conducted in the training dataset alone. In order to reduce feature dimensionality, several steps were applied. Firstly, features with (near) zero variance (i.e., features that have the same value in ≥95% of the data points) were excluded. Next, feature pairs with Spearman correlation ($r \geq 0.90$) were considered to be highly correlated, and the feature with the

highest average correlation with all other features was removed. Then, the remaining features were fed into the Boruta dimension-reduction and feature-elimination algorithm, with the maximal number of important sources, runs set to 1000. The Boruta algorithm is a wrapper method based on random forest classification [37]. Afterward, a random forest model was trained with the remaining features and the top-10 features with the highest mean decrease in Gini were retained for the final random forest model. Five models were trained: 1 model was trained to classify between normal and ILDs, while the rest were used to classify between IPF with different UIP pattern appearances (i.e.**,** UIP on HRCT or UIP not on HRCT but confirmed with a lung biopsy) and non-IPF ILDs with no UIP pattern and confirmed by a lung biopsy.

Statistical Analysis

All statistical analyses were performed using R on RStudio (version 4.0.2; https://www.R-project.org/ (accessed on 10 January 2022)). Comparisons between datasets were summarized using a Wilcoxon rank-sum test for the continuous variables and an X^2 Fisher exact test for categorical variables. A Spearman correlation was used to evaluate the correlation between radiomic features.

To assess the model's level of performance, the area under the curve (AUC) from the receiver operating characteristic (ROC) analysis was used and a 95% confidence interval (CI) was reported. To estimate the goodness-of-fit of the models, the Hosmer–Lemeshow test was used, and calibration plots were generated to visualize the consistency of models. This study was assessed using a Radiomics Quality Score [21] that consists of 16 items with different scores that sum up to 36 points and was designed specifically for radiomic studies.

Results

Patients Characteristics

A total of 328 patients were included in the study after the application of the exclusion criteria (Figure 1). A group of 122 patients from the center (i) was included. These patients were divided into three groups: (G1) patients with a final diagnosis of IPF and with typical UIP pattern in HRCT (n = 39); (G2) patients with non-typical UIP pattern and a final MDD diagnosis of IPF confirmed by SLB (n = 41); (G3) patients non-IPF ILD diagnosis confirmed by SLB (n = 42). From database (A), a total of 109 patients were included and divided into two groups: (1) IPF with UIP pattern patients (n = 53) and (2) non-IPF ILD with no UIP pattern (n = 56). From database (B) (G4), 97 healthy patients were included. A comparison between patients with a final diagnosis of IPF\UIP, non-IPF ILD, and healthy patients was performed and summarized in Table 1. As expected, there was a higher percentage of males among IPF patients (79% vs. 51%, $p < 0.001$), whereas no significant differences were noticed regarding age ($p = 0.06$), and lung function tests (FEV1, $p = 0.8$; FVC, $p = 0.18$; DLCO, $p = 0.23$; BMI, $p = 0.34$).

Table 1. Demographic and clinical characteristics of patients with IPF, non-IPF ILD, and healthy groups. IQR: interquartile range; SD: standard deviation.

Variable	IPF\UIP (HRCT & Biopsy)	Non-IPF ILD (Biopsy)	Normal	*p*-Value
Age (median (IQR)	65 (60, 71)	63 (57, 72)	62 (56, 67)	0.06
Sex = M (%)	104 (78.8)	51 (51.5)	56 (57.7)	<0.001
FEV1 (mean (SD))	71.08 (18.34)	71.77 (21.94)	-	0.8
FVC (mean (SD))	67.39 (19.53)	71.07 (22.17)	-	0.18
DLCO (mean (SD))	38.92 (11.62)	36.73 (16.12)	-	0.23
BMI (mean (SD))	28.06 (4.42)	28.69 (5.59)	-	0.34

Feature Extraction and Feature Selection

Original features were extracted (n = 170) for the whole and sectorized lung. Shape features and features with little or zero variance were excluded (n = 33). A list of the selected features after removing the highly correlated features, applying the Boruta algorithm, and Gini decrease can be found in Appendix A, Table A1. Feature selection methods yielded ten radiomics features as inputs for the group comparisons.

Performance of the Models

The volume of the trachea was observed to differ significantly ($p < 0.001$) between the control, IPF/UIP, and ILDs other than IPF patients (49.23 ±1 2.96, 73.40 ± 22.01, and 61.67 ± 18.81 cm3, respectively, mean ± SD), and also between IPF/ UIP and ILD (non-IPF) ($p < 0.001$) (Figure 3). In addition, no association was detected between tracheal volume and either lung function (FVC% predicted, r = −0.03, $p = 0.59$), or the GAP index (r = 0.17, $p = 0.01$). Following the feature selection, the volume of the trachea was selected as an important feature for all models, except for the classification between normal and ILDs.

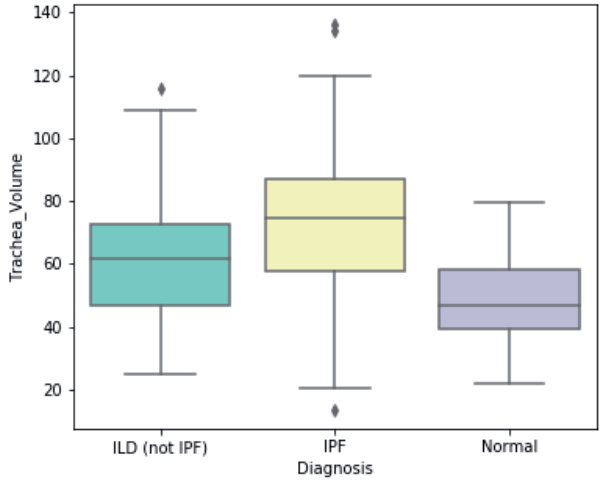


Figure 3. The difference in the volume of the trachea between IPF, non-IPF ILD, and normal, $p < 0.001$.

When classifying between a normal lung (G4, database B) and a lung with ILDs (G1 + G2 + G3) from center (i), an AUC of 1.0 (CI: 1.0–0.1) was achieved in validation (M1) (Figure 4). For the classification between G1 and G3 (center i), significant results were obtained using whole lungs with an AUC of 0.96 (95% CI: 0.90–1.0) in validation (M2). For the classification between G2 and G3 (center i), significant results were achieved using sector 1 (upper zone of the lung) with an AUC of 0.87 (95% CI: 0.74–1.0) in validation (M3).

When combining G1 and G2 to distinguish the results from G3 (center (i)), an AUC of 0.82 (95% CI: 0.68–0.95, M4) and 0.66 (95% CI: 0.59–0.73, M4.1) in validation and test dataset (database A) were achieved using whole lungs respectively. When 40% of the test dataset (from database A) is introduced to the training dataset, and retaining the remaining 60% as testing, an AUC of 0.77 (95% CI: 0.69–0.85) was achieved (M5).

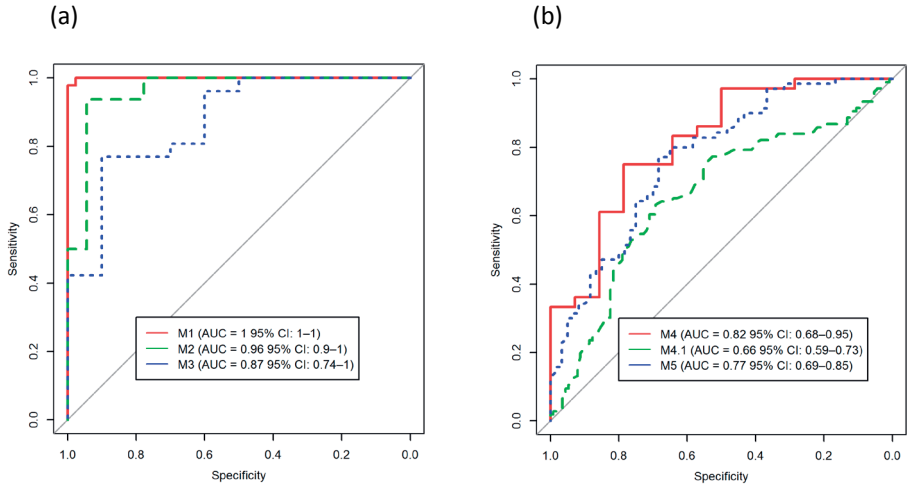


Figure 4. The graph shows the area under the receiver operating characteristic (AUC) curve of different models in the validation (a)\test (b) dataset. (M1) normal lungs vs. ILD; (M2) IPF\UIP on HRCT (G1) vs. non-IPF ILD (biopsy-proven) (G3); (M3) IPF\UIP pattern proven by biopsy (G2) vs. non-IPF ILD (biopsy-proven) (G3); (M4) IPF with UIP (G1 + G2) vs. non-IPF ILD (biopsy-proven) (G3); M4.1) IPF with UIP (G1 + G2) vs. non-IPF ILD (G3) vs. non-IPF ILD (biopsy-proven)(G3) in testing; (M5) IPF with UIP (G1 + G2) vs. non-IPF ILD (biopsy-proven) (G3) mixed with 40% of the testing dataset.

The detailed sensitivity and specificity of the models for validation/testing dataset are summarized in Table 2. To gauge the presence of overfitting when retraining all the models with randomized outcomes, no single feature was chosen as significant when the Boruta algorithm was applied and the workflow had to be halted.

Table 2. Detailed predictive and diagnostic values among various models studied, using the validation/testing dataset.

Model (M)	AUC (95% CI)	Accuracy %	Sensitivity %	Specificity %
M1	1.0 (1.0–1.0)	99	98	98
M2	0.96 (0.90–1.0)	91	88	94
M3	0.87 (0.74–1.0)	72	65	90
M4	0.82 (0.68–0.95)	70	66	79
M4.1	0.66 (0.59–0.73)	65	60	69
M5	0.77 (0.69–0.85	69	64	75

Among all models, M1, M2, and M4 showed proper calibration with p = 0.68, 0.32, and 0.07, respectively (Figure 5). The radiomics quality score of this study was 64% (23 of 36).

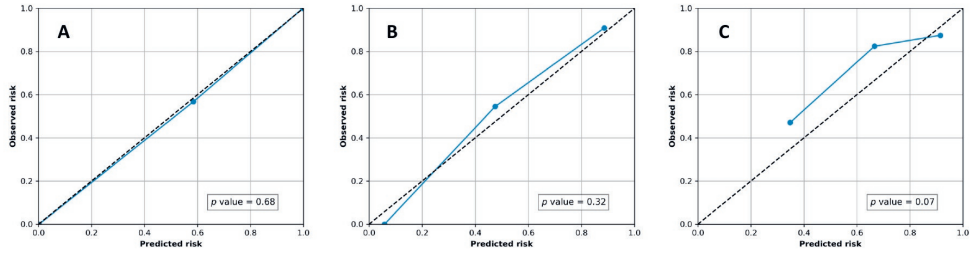


Figure 5. Calibration plots of radiomics models on the validation/testing dataset. (**A**) Normal vs. ILD (M1); (**B**) IPF\UIP vs. non-IPF ILD (M2); (**C**) IPF with UIP (G1 + G2) vs. non-IPF ILD (biopsy-proven) (M4).

Discussion

In this study, we developed a quantitative signature (radiomics) extracted from HRCT to classify fibrotic lung disease. A random forest classifier was used to differentiate between (1) normal lungs and interstitial lung diseases (ILDs); (2) idiopathic pulmonary fibrosis (IPF) (with typical or less typical usual interstitial pneumonia (UIP) radiological presentation), and non-IPF ILDs (other than IPF as proven by the absence of UIP in a surgical biopsy). Briefly stated, we were able to demonstrate that radiomic features derived from HRCT images can be used to distinguish between a normal state and ILDs, as well as between IPF with a UIP pattern and ILDs with no UIP pattern verified by surgical biopsy. The inclusion of biopsy-proven non-IPF ILDs patients strengthens the study, as well as making it unique (Appendix A, Table A2).

Differentiating between normal and ILD lung tissues might seem a trivial task. However, it is a time-consuming process since the clinician has to go through all the scans. Developing an automated approach that differentiates between normal and abnormal lungs would decrease the amount of time a clinician needs to assess images on a daily basis. A previous study presented a novel texture analysis method that incorporates texture matching with histogram features analysis [38]. This study reported that their method achieved a sensitivity of 92.96% and a specificity of 93.78% in differentiating between normal and abnormal lungs.

The study made use of a part of the handcrafted radiomic features used in our analysis. Using all-handcrafted radiomic features, we achieved a sensitivity of 98% and a specificity of 98% to identify an ILD.

Many ILDs have characteristics and changes in the lungs similar to those of IPF/UIP on HRCT, making the diagnosis very difficult—even for experienced radiologists [39]. Visual assessments of ILDs while using HRCT can be very subjective due to the high variability in the knowledge of inter-readers [16–18]. Therefore, providing automated diagnostic assistance in this setting would be highly beneficial, especially for less experienced radiologists. Texture image analysis is not new in fibrotic lung diseases and has been researched to automatically analyze ILDs on CT images [38,40–46]. However, most of the existing studies have focused on prognostic questions rather than providing diagnostic support. Maldonado et al. showed that short-term reticular changes evaluated by CALIPER (Computer-Aided Lung Informatics for Pathology Evaluation and Rating) correlated with physiological parameters and were predictive of survival in IPF patients [41]. Humphries et al. concluded that the use of Data-driven Texture Analysis (DTA) for IPF patients correlates with both pulmonary function tests and visual assessment on CT images at baseline [45]. However, a more thorough classification of phenotypes can be provided by applying radiomic data stratification. Walsh et al. used a deep learning approach for automated classification of fibrotic lung disease, according to the 2011 ATS/ERS/JRS/ALAT idiopathic pulmonary fibrosis diagnostic guidelines on a dataset of 1157 HRCT scans. The algorithm performance was compared to that of 91 radiologists and showed an accuracy of 73.3%, compared to the median accuracy of the radiologists, 70.7% [47]. To the best of our knowledge, no study has investigated the potential of handcrafted radiomics for differentiation between IPF/UIP and other ILDs.

By assessing the potential of handcrafted radiomics to differentiate between IPF with typical UIP presentation on HRCT and ILDs other than IPF, we discovered another benefit of automation similar to that achieved by differentiating between normal and abnormal lung tissue. It could serve mainly as a decision-aiding tool that would increase the diagnostic accuracy of the disease, reduce the need for invasive lung biopsies, and decrease the time needed to conduct routine scans.

IPF is also associated with wide parenchymal and airway conditions, such as those found in the trachea wall, which leads to pathological changes [48]. Ratwani et al. studied the correlation between the change of tracheobronchial tree size and the disease severity of IPF [31]. Our study found a significant difference in the volume of trachea between normal, IPF/UIP and, ILDs patients. Furthermore, it was found that the volume of the trachea was higher for IPF subjects compared to normal and ILDs other than IPF (Figure 3). No correlation was seen between the volume of the trachea and %FVC predicted. This conclusion may be consistent with the findings of Ratwani et al. [31], who found that there was no association

between %FVC predicted and growing tracheobronchial tree size, indicating that tracheal expansion is not only due to fibrosis and that other variables may be at play. Such findings suggest that the increase of the volume of the trachea might be a good new handcrafted radiomic feature to serve as a promising tool in the diagnosis of IPF.

The decrease in model performance in the test dataset might be explained by the presence of variation in acquisition and reconstruction parameters. When the random forest algorithm learned part of the test dataset in the training dataset (M4.1), the model AUC increased from 0.66 to 0.77. Such findings indicate the need for addressing the challenges associated with differences in imaging parameters.

This study has some limitations. Firstly, we did have the additional categories of UIP patterns (definite, probable, indeterminate, or alternative) in the training dataset but not in the test dataset. Therefore, we only used the test dataset when we combined G1 and G2. Secondly, the healthy CT scans (G4) were obtained only from one center (center iii). Thirdly, the CT acquisition parameters of HRCT varied between and within the centers, and radiomic features are known to be influenced by different CT acquisition and reconstruction parameters [34,49,50]. Furthermore, we could not assess the reproducibility of features due to the lack of anthropomorphic phantom or test-retest scans acquired with settings similar to the scans used in this study. Henceforth, future studies must employ reproducibility studies to ensure the generalizability of the developed models. The application of radiomics to IPF may be broadened to include treatment decision aids. Further research should be undertaken to investigate the progression of IPF/UIP at baseline and follow up to evaluate the effectiveness of the antifibrotic treatment. In addition, a combination of deep learning and handcrafted radiomics with the addition of blood or genetic biomarkers would be a powerful tool in the classification of ILDs.

Conclusions

At present, there is minimal radiomics research on ILDs. Our findings are, nonetheless, promising and underline the strong potential of HRCT-based radiomics for the identification of ILDs. The classification between IPF/UIP and other ILDs using radiomics might capture features indicating different types of ILDs in HRCT, which are hardly recognizable via visual assessment. The radiomic features extracted from HRCT, along with clinical features, might aid in the assessment of ILDs and be used as a valuable tool for computer-aided decision-making in imaging.

Appendix A

Table A1. Features name for each model.

Model	Features Name
M1	GLSZM_SZNN, GLDZM_LISDE, GLSZM_HISAE, GLSZM_HILAE, GLCM_diffVar, GLRLM_GLV, GLCM_ infoCorr2, GLSZM_LILAE, IH_medianD, GLDZM_LILDE
M2	NGLDM_LGSDE, GLDZM_DZN, GLDZM_LISDE, Trachea_Volume, NGLDM_HGLDE, GLRLM_GLV, GLCM_clusShade, IH_qcod, GLDZM_HILDE, GLCM_contrast
M3	GLCM_infoCorr2, Fractal_sd, Trachea_Volume, GLCM_maxCorr, GLDZM_SDE, GLRLM_GLV, IH_ energy, GLDZM_LISDE, NGLDM_DV, Stats_kurtosis
M4 M4.1	Trachea_Volume, GLDZM_DZN, NGLDM_LGSDE, GLCM_infoCorr2, GLDZM_SDE, GLCM_sumVar, NGTDM_strength, NGLDM_HGLDE, GLDZM_LISDE, GLCM_maxCorr
M5	Trachea_Volume, GLRLM_GLV, GLCM_diffVar, GLSZM_HILAE, NGLDM_LGSDE, GLSZM_SAE, IH_qcod, GLSZM_ZE, GLSZM_IV, Stats_kurtosis

Table A2. List of ILDs included in the study.

ILD Names
Hypersensitivity pneumonitis (HP)
Nonspecific interstitial pneumonia (NSIP)
Connective tissue disease-associated interstitial lung disease (other than systemic sclerosis (SSc-ILD)) (CTD-ILD)
Lymphoid interstitial pneumonia (LIP)
Unclassifiable ILD
Idiopathic pulmonary fibrosis (IPF)
Pleuro-parenchymal fibroelastosis
Desquamative interstitial pneumonia (DIP)
Eosinophilic pneumonia
systemic sclerosis SSc-ILD
Respiratory bronchiolitis (RB-ILD)

References

1. Kishaba, T. Evaluation and Management of Idiopathic Pulmonary Fibrosis. *Respir. Investig.* **2019**, *57*, 300–311.
2. Sgalla, G.; Biffi, A.; Richeldi, L. Idiopathic Pulmonary Fibrosis: Diagnosis, Epidemiology and Natural History. *Respirology* **2016**, *21*, 427–437.
3. Raghu, G.; Collard, H.R.; Egan, J.J.; Martinez, F.J.; Behr, J.; Brown, K.K.; Colby, T.V.; Cordier, J.-F.; Flaherty, K.R.; Lasky, J.A.; et al. An Official ATS/ERS/JRS/ALAT Statement: Idiopathic Pulmonary Fibrosis: Evidence-Based Guidelines for Diagnosis and Management. *Am. J. Respir. Crit. Care Med.* **2011**, *183*, 788–824.
4. Lederer, D.J.; Martinez, F.J. Idiopathic Pulmonary Fibrosis. *N. Engl. J. Med.* **2018**, *378*, 1811–1823.
5. Mueller-Mang, C.; Grosse, C.; Schmid, K.; Stiebellehner, L.; Bankier, A.A. What Every Radiologist Should Know about Idiopathic Interstitial Pneumonias. *Radiographics* **2007**, *27*, 595–615.
6. Lynch, D.A.; Sverzellati, N.; Travis, W.D.; Brown, K.K.; Colby, T.V.; Galvin, J.R.; Goldin, J.G.; Hansell, D.M.; Inoue, Y.; Johkoh, T.; et al. Diagnostic Criteria for Idiopathic Pulmonary Fibrosis: A Fleischner Society White Paper. *Lancet Respir. Med.* **2018**, *6*, 138–153.
7. Travis, W.D.; Costabel, U.; Hansell, D.M.; King, T.E., Jr; Lynch, D.A.; Nicholson, A.G.; Ryerson, C.J.; Ryu, J.H.; Selman, M.; Wells, A.U.; et al. An Official American Thoracic Society/European Respiratory Society Statement: Update of the International Multidisciplinary Classification of the Idiopathic Interstitial Pneumonias. *Am. J. Respir. Crit. Care Med.* **2013**, *188*, 733–748.
8. Han, Q.; Luo, Q.; Xie, J.-X.; Wu, L.-L.; Liao, L.-Y.; Zhang, X.-X.; Chen, R.-C. Diagnostic Yield and Postoperative Mortality Associated with Surgical Lung Biopsy for Evaluation of Interstitial Lung Diseases: A Systematic Review and Meta-Analysis. *J. Thorac. Cardiovasc. Surg.* **2015**, *149*, 1394.e1–1401.e1.
9. Morris, D.; Zamvar, V. The Efficacy of Video-Assisted Thoracoscopic Surgery Lung Biopsies in Patients with Interstitial Lung Disease: A Retrospective Study of 66 Patients. *J. Cardiothorac. Surg.* **2014**, *9*, 45.
10. Sonobe, M.; Handa, T.; Tanizawa, K.; Sato, M.; Sato, T.; Chen, F.; Omasa, M.; Bando, T.; Date, H.; Mishima, M. Videothoracoscopy-Assisted Surgical Lung Biopsy for Interstitial Lung Diseases. *Gen. Thorac. Cardiovasc. Surg.* **2014**, *62*, 376–382.
11. Ravaglia, C.; Bonifazi, M.; Wells, A.U.; Tomassetti, S.; Gurioli, C.; Piciucchi, S.; Dubini, A.; Tantalocco, P.; Sanna, S.; Negri, E.; et al. Safety and Diagnostic Yield of Transbronchial Lung Cryobiopsy in Diffuse Parenchymal Lung Diseases: A Comparative Study versus Video-Assisted Thoracoscopic Lung Biopsy and a Systematic Review of the Literature. *Respiration* **2016**, *91*, 215–227.
12. Hutchinson, J.; McKeever, T.; Fogarty, A.; Navaratnam, V.; Hubbard, R. Surgical Lung Biopsy for the Diagnosis of Interstitial Lung Disease in England: 1997–2008. *Eur. Respir. J.* **2016**, *48*, 1453–1461
13. Durheim, M.T.; Kim, S.; Gulack, B.C.; Burfeind, W.R.; Gaissert, H.A.; Kosinski, A.S.; Hartwig, M.G. Mortality and Respiratory Failure After Thoracoscopic Lung Biopsy for Interstitial Lung Disease. *Ann. Thorac. Surg.* **2017**, *104*, 465–470.

14. Richeldi, L.; Scholand, M.B.; Lynch, D.A.; Colby, T.V.; Myers, J.L.; Groshong, S.D.; Chung, J.H.; Benzaquen, S.; Nathan, S.D.; Davis, J.R.; et al. Utility of a Molecular Classifier as a Complement to High-Resolution Computed Tomography to Identify Usual Interstitial Pneumonia. *Am. J. Respir. Crit. Care Med.* **2021**, *203*, 211–220.
15. Raghu, G.; Remy-Jardin, M.; Myers, J.L.; Richeldi, L.; Ryerson, C.J.; Lederer, D.J.; Behr, J.; Cottin, V.; Danoff, S.K.; Morell, F.; et al. Diagnosis of Idiopathic Pulmonary Fibrosis. An Official ATS/ERS/JRS/ALAT Clinical Practice Guideline. *Am. J. Respir. Crit. Care Med.* **2018**, *198*, e44–e68.
16. Gruden, J.F. CT in Idiopathic Pulmonary Fibrosis: Diagnosis and Beyond. *AJR Am. J. Roentgenol.* **2016**, *206*, 495–507.
17. Tominaga, J.; Sakai, F.; Johkoh, T.; Noma, S.; Akira, M.; Fujimoto, K.; Colby, T.V.; Ogura, T.; Inoue, Y.; Taniguchi, H.; et al. Diagnostic Certainty of Idiopathic Pulmonary Fibrosis/usual Interstitial Pneumonia: The Effect of the Integrated Clinico-Radiological Assessment. *Eur. J. Radiol.* **2015**, *84*, 2640–2645.
18. Walsh, S.L.F.; Calandriello, L.; Sverzellati, N.; Wells, A.U.; Hansell, D.M. UIP Observer Consort Interobserver Agreement for the ATS/ERS/JRS/ALAT Criteria for a UIP Pattern on CT. *Thorax* **2016**, *71*, 45–51.
19. Walsh, S.L.F.; Wells, A.U.; Desai, S.R.; Poletti, V.; Piciucchi, S.; Dubini, A.; Nunes, H.; Valeyre, D.; Brillet, P.Y.; Kambouchner, M.; et al. Multicentre Evaluation of Multidisciplinary Team Meeting Agreement on Diagnosis in Diffuse Parenchymal Lung Disease: A Case-Cohort Study. *Lancet Respir. Med.* **2016**, *4*, 557–565.
20. Lambin, P.; Rios-Velazquez, E.; Leijenaar, R.; Carvalho, S.; van Stiphout, R.G.P.M.; Granton, P.; Zegers, C.M.L.; Gillies, R.; Boellard, R.; Dekker, A.; et al. Radiomics: Extracting More Information from Medical Images Using Advanced Feature Analysis. *Eur. J. Cancer* **2012**, *48*, 441–446.
21. Lambin, P.; Leijenaar, R.T.H.; Deist, T.M.; Peerlings, J.; de Jong, E.E.C.; van Timmeren, J.; Sanduleanu, S.; Larue, R.T.H.M.; Even, A.J.G.; Jochems, A.; et al. Radiomics: The Bridge between Medical Imaging and Personalized Medicine. *Nat. Rev. Clin. Oncol.* **2017**, *14*, 749–762.
22. Aerts, H.J.W.L.; Velazquez, E.R.; Leijenaar, R.T.H.; Parmar, C.; Grossmann, P.; Carvalho, S.; Bussink, J.; Monshouwer, R.; Haibe-Kains, B.; Rietveld, D.; et al. Decoding Tumour Phenotype by Noninvasive Imaging Using a Quantitative Radiomics Approach. *Nat. Commun.* **2014**, *5*, 4006.
23. Scrivener, M.; de Jong, E.E.C.; van Timmeren, J.E.; Pieters, T.; Ghaye, B.; Geets, X. Radiomics Applied to Lung Cancer: A Review. *Transl. Cancer Res.* **2016**, *5*, 398–409.
24. Bogowicz, M.; Vuong, D.; Huellner, M.W.; Pavic, M.; Andratschke, N.; Gabrys, H.S.; Guckenberger, M.; Tanadini-Lang, S. CT Radiomics and PET Radiomics: Ready for Clinical Implementation? *Q. J. Nucl. Med. Mol. Imaging* **2019**, *63*, 355–370.
25. Bogowicz, M.; Tanadini-Lang, S.; Guckenberger, M.; Riesterer, O. Combined CT Radiomics of Primary Tumor and Metastatic Lymph Nodes Improves Prediction of Loco-Regional Control in Head and Neck Cancer. *Sci. Rep.* **2019**, *9*, 15198.
26. Parmar, C.; Leijenaar, R.T.H.; Grossmann, P.; Rios Velazquez, E.; Bussink, J.; Rietveld, D.; Rietbergen, M.M.; Haibe-Kains, B.; Lambin, P.; Aerts, H.J.W.L. Radiomic Feature Clusters and Prognostic Signatures Specific for Lung and Head & Neck Cancer. *Sci. Rep.* **2015**, *5*, 11044.

27. Aerts, H.J.W.L.; Grossmann, P.; Tan, Y.; Oxnard, G.R.; Rizvi, N.; Schwartz, L.H.; Zhao, B. Corrigendum: Defining a Radiomic Response Phenotype: A Pilot Study Using Targeted Therapy in NSCLC. *Sci. Rep.* **2017**, *7*, 41197.
28. Martini, K.; Baessler, B.; Bogowicz, M.; Blüthgen, C.; Mannil, M.; Tanadini-Lang, S.; Schniering, J.; Maurer, B.; Frauenfelder, T. Applicability of Radiomics in Interstitial Lung Disease Associated with Systemic Sclerosis: Proof of Concept. *Eur. Radiol.* **2020**, *31*, 1987–1998. https://doi.org/10.1007/s00330-020-07293-8.
29. Refaee, T.; Wu, G.; Ibrahim, A.; Halilaj, I.; Leijenaar, R.T.H.; Rogers, W.; Gietema, H.A.; Hendriks, L.E.L.; Lambin, P.; Woodruff, H.C. The Emerging Role of Radiomics in COPD and Lung Cancer. *Respiration* **2020**, *99*, 99–107.
30. Occhipinti, M.; Paoletti, M.; Bartholmai, B.J.; Rajagopalan, S.; Karwoski, R.A.; Nardi, C.; Inchingolo, R.; Larici, A.R.; Camiciottoli, G.; Lavorini, F.; et al. Spirometric Assessment of Emphysema Presence and Severity as Measured by Quantitative CT and CT-Based Radiomics in COPD. *Respir. Res.* **2019**, *20*, 101.
31. Ratwani, A.; King, C.; Brown, W.; Shlobin, O.; Weir, N.; Nathan, S. Tracheobronchial Tree Size as a Predictor of Disease Severity and Outcomes in Idiopathic Pulmonary Fibrosis. *Chest* **2017**, *152*, A487.
32. Ley, B.; Ryerson, C.J.; Vittinghoff, E.; Ryu, J.H.; Tomassetti, S.; Lee, J.S.; Poletti, V.; Buccioli, M.; Elicker, B.M.; Jones, K.D.; et al. A Multidimensional Index and Staging System for Idiopathic Pulmonary Fibrosis. *Ann. Intern. Med.* **2012**, *156*, 684–691.
33. Shafiq-ul-Hassan, M.; Zhang, G.G.; Latifi, K.; Ullah, G.; Hunt, D.C.; Balagurunathan, Y.; Abdalah, M.A.; Schabath, M.B.; Goldgof, D.G.; Mackin, D.; et al. Intrinsic Dependencies of CT Radiomic Features on Voxel Size and Number of Gray Levels. *Med. Phys.* **2017**, *44*, 1050–1062.
34. Ibrahim, A.; Refaee, T.; Primakov, S.; Barufaldi, B.; Acciavatti, R.J.; Granzier, R.W.Y.; Hustinx, R.; Mottaghy, F.M.; Woodruff, H.C.; Wildberger, J.E.; et al. The Effects of In-Plane Spatial Resolution on CT-Based Radiomic Features' Stability with and without ComBat Harmonization. *Cancers* **2021**, *13*, 1848. https://doi.org/10.3390/cancers13081848.
35. Zwanenburg, A.; Vallières, M.; Abdalah, M.A.; Aerts, H.J.W.L.; Andrearczyk, V.; Apte, A.; Ashrafinia, S.; Bakas, S.; Beukinga, R.J.; Boellaard, R.; et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-Based Phenotyping. *Radiology* **2020**, *295*, 328–338.
36. Mackin, D.; Fave, X.; Zhang, L.; Fried, D.; Yang, J.; Taylor, B.; Rodriguez-Rivera, E.; Dodge, C.; Jones, A.K.; Court, L. Measuring Computed Tomography Scanner Variability of Radiomics Features. *Invest. Radiol.* **2015**, *50*, 757–765.
37. Kursa, M.B.; Rudnicki, W.R.; Others Feature Selection with the Boruta Package. *J. Stat. Softw.* **2010**, *36*, 1–13.
38. Zavaletta, V.A.; Bartholmai, B.J.; Robb, R.A. High Resolution Multidetector CT-Aided Tissue Analysis and Quantification of Lung Fibrosis. *Acad. Radiol.* **2007**, *14*, 772–787.
39. Hochhegger, B.; Marchiori, E.; Zanon, M.; Rubin, A.S.; Fragomeni, R.; Altmayer, S.; Carvalho, C.R.R.; Baldi, B.G. Imaging in Idiopathic Pulmonary Fibrosis: Diagnosis and Mimics. *Clinics* **2019**, *74*, e225.

40. Maldonado, F.; Moua, T.; Rajagopalan, S.; Karwoski, R.A.; Raghunath, S.; Decker, P.A.; Hartman, T.E.; Bartholmai, B.J.; Robb, R.A.; Ryu, J.H. Automated Quantification of Radiological Patterns Predicts Survival in Idiopathic Pulmonary Fibrosis. *Eur. Respir. J.* **2014**, *43*, 204–212.
41. Jacob, J.; Bartholmai, B.J.; Rajagopalan, S.; Kokosi, M.; Nair, A.; Karwoski, R.; Walsh, S.L.F.; Wells, A.U.; Hansell, D.M. Mortality Prediction in Idiopathic Pulmonary Fibrosis: Evaluation of Computer-Based CT Analysis with Conventional Severity Measures. *Eur. Respir. J.* **2017**, *49*, 1601011. https://doi.org/10.1183/13993003.01011-2016.
42. Uppaluri, R.; Hoffman, E.A.; Sonka, M.; Hunninghake, G.W.; McLennan, G. Interstitial Lung Disease: A Quantitative Study Using the Adaptive Multiple Feature Method. *Am. J. Respir. Crit. Care Med.* **1999**, *159*, 519–525.
43. Delorme, S.; Keller-Reichenbecher, M.A.; Zuna, I.; Schlegel, W.; Van Kaick, G. Usual Interstitial Pneumonia. Quantitative Assessment of High-Resolution Computed Tomography Findings by Computer-Assisted Texture-Based Image Analysis. *Invest. Radiol.* **1997**, *32*, 566–574.
44. Rodriguez, L.H.; Vargas, P.F.; Raff, U.; Lynch, D.A.; Rojas, G.M.; Moxley, D.M.; Newell, J.D. Automated Discrimination and Quantification of Idiopathic Pulmonary Fibrosis from Normal Lung Parenchyma Using Generalized Fractal Dimensions in High-Resolution Computed Tomography Images. *Acad. Radiol.* **1995**, *2*, 10–18.
45. Humphries, S.M.; Yagihashi, K.; Huckleberry, J.; Rho, B.-H.; Schroeder, J.D.; Strand, M.; Schwarz, M.I.; Flaherty, K.R.; Kazerooni, E.A.; van Beek, E.J.R.; et al. Idiopathic Pulmonary Fibrosis: Data-Driven Textural Analysis of Extent of Fibrosis at Baseline and 15-Month Follow-Up. *Radiology* **2017**, *285*, 270–278.
46. Kim, H.J.; Brown, M.S.; Chong, D.; Gjertson, D.W.; Lu, P.; Kim, H.J.; Coy, H.; Goldin, J.G. Comparison of the Quantitative CT Imaging Biomarkers of Idiopathic Pulmonary Fibrosis at Baseline and Early Change with an Interval of 7 Months. *Acad. Radiol.* **2015**, *22*, 70–80.
47. Walsh, S.L.F.; Calandriello, L.; Silva, M.; Sverzellati, N. Deep Learning for Classifying Fibrotic Lung Disease on High-Resolution Computed Tomography: A Case-Cohort Study. *Lancet Respir. Med.* **2018**, *6*, 837–845.
48. Abumossalam, A.M.; Elshafeey, M.M.; Abdelsalam, E.M. Tracheoechography versus CT Tracheography for Assessment of Idiopathic Pulmonary Fibrosis Related Tracheopathy. *Egypt. J. Chest Dis. Tuberc.* **2015**, *64*, 459–464.
49. Ibrahim, A.; Refaee, T.; Leijenaar, R.T.H.; Primakov, S.; Hustinx, R.; Mottaghy, F.M.; Woodruff, H.C.; Maidment, A.D.A.; Lambin, P. The Application of a Workflow Integrating the Variable Reproducibility and Harmonizability of Radiomic Features on a Phantom Dataset. *PLoS ONE* **2021**, *16*, e0251147.
50. Ibrahim, A.; Refaee, T.; Primakov, S.; Barufaldi, B.; Acciavatti, R.J.; Granzier, R.W.Y.; Hustinx, R.;; Mottaghy, F.M.; Woodruff, H.C.; Wildberger, J.E.; et al. Reply to Orlhac, F.; Buvat, I. Comment on "Ibrahim et Al. The Effects of In-Plane Spatial Resolution on CT-Based Radiomic Features' Stability with and without ComBat Harmonization. Cancers 2021, 13, 1848". Cancers **2021**, *13*, 3080. https://doi.org/10.3390/cancers13123080.

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$minimum = \min(\mathbf{X})$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2$$

$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_s} \mathbf{P}(i,j)\right)^2}{N_p}$$

$$correlation = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) ij - \mu_x \mu_y}{\sigma_x(i) \sigma_y(j)}$$

$$MAD = \frac{1}{N_p} \sum_{i=1}^{N_p} |\mathbf{X}(i) - \bar{X}|$$

$$minimum = \min(\mathbf{X})$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_s} \mathbf{P}\right.}{N_z}$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2$$

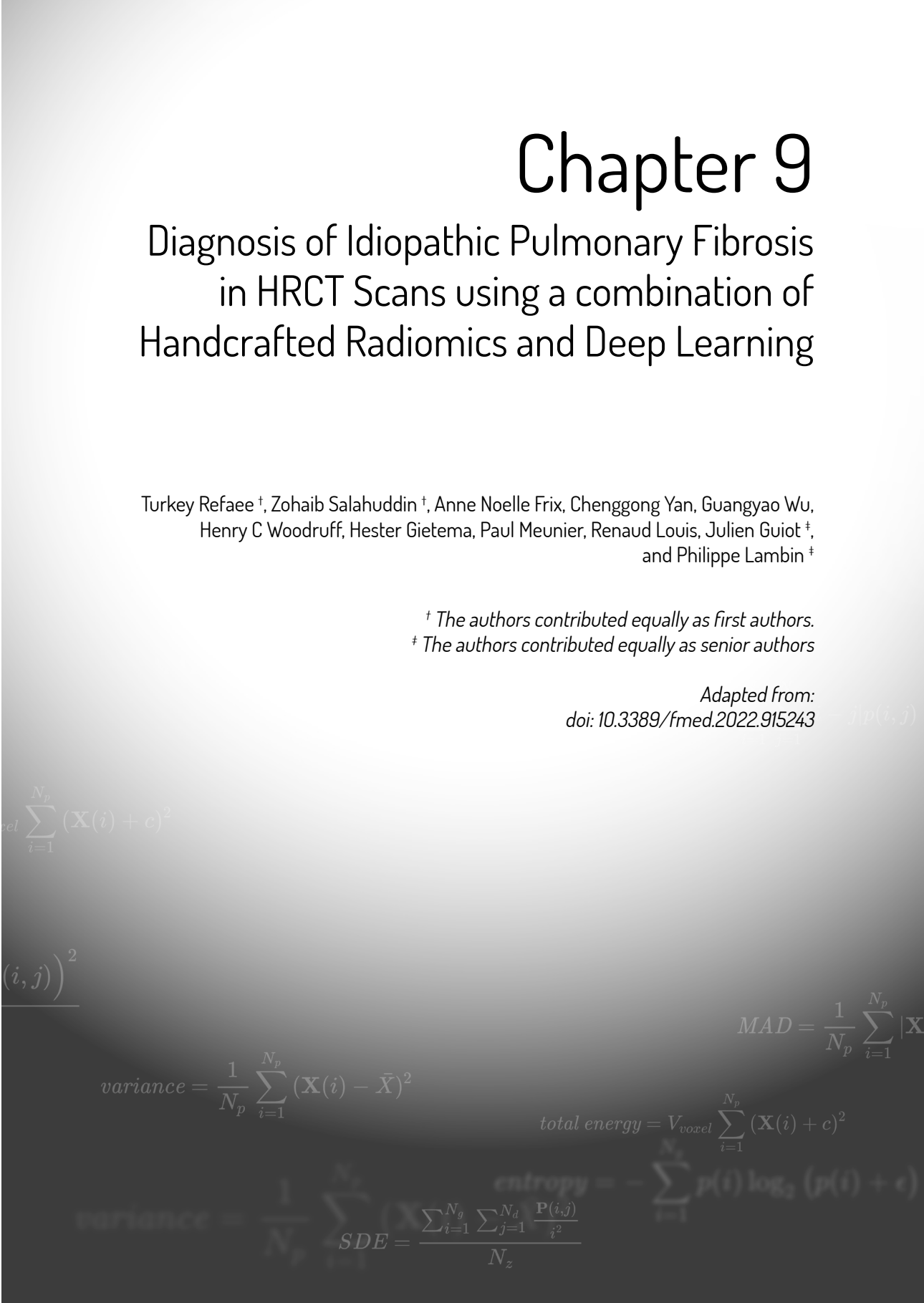
Chapter 9

Diagnosis of Idiopathic Pulmonary Fibrosis in HRCT Scans using a combination of Handcrafted Radiomics and Deep Learning

Turkey Refaee [†], Zohaib Salahuddin [†], Anne Noelle Frix, Chenggong Yan, Guangyao Wu, Henry C Woodruff, Hester Gietema, Paul Meunier, Renaud Louis, Julien Guiot [‡], and Philippe Lambin [‡]

[†] *The authors contributed equally as first authors.*
[‡] *The authors contributed equally as senior authors*

Adapted from:
doi: 10.3389/fmed.2022.915243



Abstract

Purpose: To develop handcrafted radiomics (HCR) and deep learning (DL) based automated diagnostic tools that can differentiate between idiopathic pulmonary fibrosis (IPF) and non-IPF interstitial lung diseases (ILDs) in patients using high-resolution computed tomography (HRCT) scans.

Material and Methods: In this retrospective study, 474 HRCT scans were included (mean age, 64.10 years ± 9.57 [SD]). Five-fold cross-validation was performed on 365 HRCT scans. Furthermore, an external dataset comprising 109 patients was used as a test set. An HCR model, a DL model, and an ensemble of HCR and DL model were developed. A virtual *in-silico* trial was conducted with two radiologists and one pulmonologist on the same external test set for performance comparison. The performance was compared using DeLong method and McNemar test. Shapley Additive exPlanations (SHAP) plots and Grad-CAM heatmaps were used for the post-hoc interpretability of HCR and DL models, respectively.

Results: In five-fold cross-validation, the HCR model, DL model, and the ensemble of HCR and DL models achieved accuracies of 76.2±6.8%, 77.9±4.6%, and 85.2±2.7%, respectively. For the diagnosis of IPF and non-IPF ILDs on the external test set, the HCR, DL, and the ensemble of HCR and DL models achieved accuracies of 76.1%, 77.9%, and 85.3%, respectively. The ensemble model outperformed the diagnostic performance of clinicians who achieved a mean accuracy of 66.3±6.7% ($p < 0.05$) during the *in-silico* trial. The area under the receiver operating characteristic curve (AUC) for the ensemble model on the test set was 0.917 which was significantly higher than the HCR model (0.817, $p = 0.02$) and the DL model (0.823, $p = 0.005$). The agreement between HCR and DL models was 61.4%, and the accuracy and specificity for the predictions when both the models agree were 93% and 97%, respectively. SHAP analysis showed the texture features as the most important features for IPF diagnosis and Grad-CAM showed that the model focused on the clinically relevant part of the image.

Conclusion: DL and HCR models can complement each other and serve as useful clinical aids for the diagnosis of IPF and non-IPF ILDs.

Keywords: Artificial Intelligence, Radiomics, Computed Tomography, Interpretability, Idiopathic Pulmonary Fibrosis, Interstitial Lung Disease.

Introduction

Interstitial lung disorders (ILDs) are a diverse group of ailments with an estimated 200 distinct entities and are linked with high morbidity and death (1). Many different parenchymal lung disorders have similar clinical signs and patterns of lung injury. Several disorders, including idiopathic pulmonary fibrosis (IPF), have unknown etiology and are labeled idiopathic or cryptogenic, while the rest are linked to other diseases, particularly connective tissue diseases, or to environmental exposures (2–6). One of the most common types of ILDs is IPF, a progressive illness marked by decreased lung function (7). IPF has an estimated incidence rate between 2.8 and 18 cases per 100,000 per year in Europe and North America (8). The median survival rate of patients with IPF is between two to four years from diagnosis (9). A prompt diagnosis and management are crucial for slowing down the progression of these lung disorders.

Medical imaging is becoming increasingly crucial for disease diagnosis, prognosis, and treatment planning in precision medicine (10). Computed tomography (CT) provides visual data that may be used to enhance decision-making (4,11). However, qualitative CT evaluation remains challenging and frequently varies amongst experts (12). The diagnosis of idiopathic pulmonary fibrosis using high-resolution computed tomography (HRCT) is a difficult task and high inter-observer variability is associated with it even with experienced radiologists (13). Consequently, there is a need for an automated clinical tool that can aid clinicians for accurate and timely diagnosis.

Artificial intelligence is becoming increasingly popular due to the increasing amount of imaging data and available computational resources (14). The use of quantitative imaging techniques in medical imaging has grown at an exponential rate (15). Handcrafted radiomics (HCR) is a quantitative approach that measures and extracts high-dimensional imaging characteristics to aid clinical decision-making (15,16). Deep learning (DL) methods learn different features and representations from the image data without the need for explicit feature engineering (17). Convolutional neural networks (CNNs) have shown remarkable results on numerous diagnostic tasks using medical image data including the diagnosis of fibrotic lung disease (18).

Despite promising results demonstrated by HCR and DL models for various medical imaging tasks, the clinical utility of such models is limited due to their lack of interpretability (19). Shapley Additive exPlanations (SHAP) (20) and Gradient-weighted class activation maps (Grad-CAM) (21) are post-hoc interpretability methods that are useful for understanding the decision-making process of HCR and DL models respectively.

In this paper, we propose a machine learning-based HCR pipeline and a DL pipeline for the automated diagnosis of IPF, non-IPF ILDs patients. We also perform an in-silico trial with experienced radiologists to compare the performance of HCR and DL on a test dataset.

Furthermore, we use post-hoc interpretability methods to aid the incorporation of these automated diagnostic tools in the clinical workflow.

Material and methods

Patients

A total of 652 HRCT scans were obtained from Site 1 (University Liege hospital) and 205 HRCT scans were obtained from database A (The Lung tissue research consortium database (LTCR)). The inclusion criteria were: the availability of non-contrast enhanced HRCT and the availability of HRCT with slices thickness of less than 1.5 mm. The exclusion criteria were: the use of contrast enhancement, images containing metal or motion artifacts, and images reconstructed with a slice thickness larger than 1.5 mm. All diagnoses were confirmed by the Multidisciplinary discussion (MDD) that included a histopathologist, pulmonologist, thoracic radiologist, and rheumatologist. Lung biopsy is only required in case of ILD inconsistent with IPF. Figure 1 shows the patient selection process. Demographic data, clinical data, and measurements of pulmonary function tests (PFT) were acquired for each patient. Demographic and clinical data include age, gender, body mass index (BMI), forced edxpiratory volume in 1s (FEV1), forced vital capacity (FVC), and diffusion capacity of the lungs for carbon monoxide (DLCO).

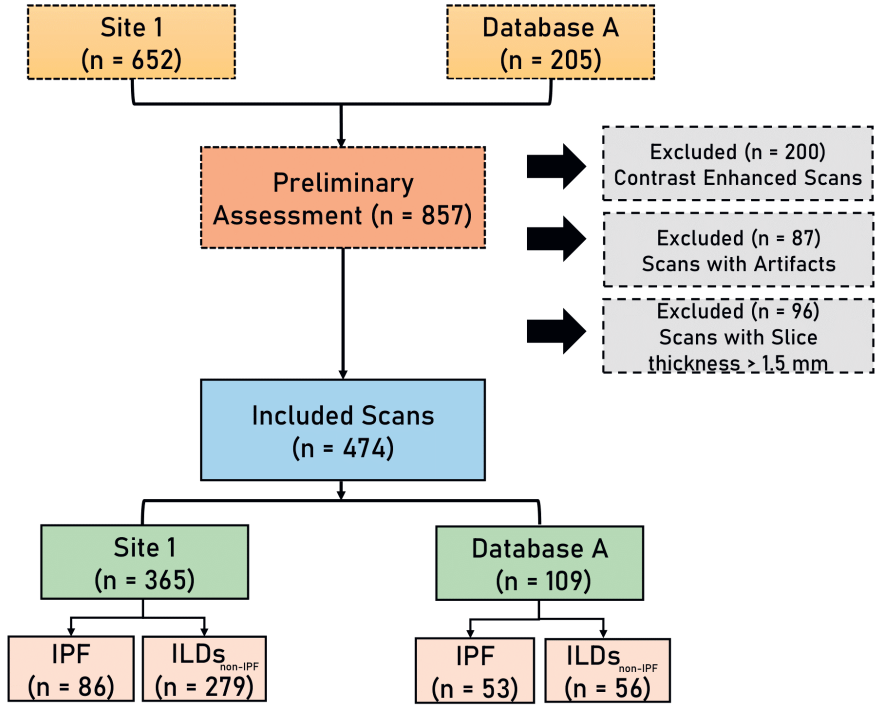


Figure 1: The flowchart diagram shows the patient selection process. IPF = Idiopathic pulmonary fibrosis, $ILDs_{non\text{-}IPF}$ = non-IPF Interstitial lung diseases.

Imaging acquisition and segmentation

The HRCT scans at site 1 were acquired at the same hospital using two different vendors (Siemens and GE). The scans acquired from database A were acquired using four different CT vendors (Siemens, GE, Philips, and Toshiba). The slice thickness of the scans varied between 0.5 mm and 1.5 mm. A further detailed description of the CT acquisition parameters can be found in Supplementary (Table E1). Whole lung segmentation was performed using an automated workflow created in MIM software (MIM Software Inc., Cleveland, OH).

Data Split

Five-fold cross-validation was performed on data from Site 1 consisting of 365 HRCT scans containing 279 non-IPF ILDs, and 86 IPF patients. External data from database A, comprising 53 IPF patients and 56 non-IPF ILDs patients was used to benchmark the performance of the proposed AI tools along with the *in-silico* trial.

Handcrafted Radiomics (HCR)

Handcrafted radiomics feature extraction

To minimize the effect of the variations in image voxel size, all CT images were resampled to a $1 \times 1 \times 1$ mm^3. Radiomics features were extracted from the HRCT images using the RadiomiX Discovery Toolbox (https://www.radiomics.bio) which calculates HCR features compliant with the Imaging Biomarkers Standardization Initiative (IBSI) (22). Voxel intensities were aggregated into 25 bins of Hounsfield Units to reduce noise and inter-scanner variability. The extracted features describe fractal dimension, intensity histogram, first-order statistics, texture, and shape. A workflow for handcrafted radiomics from segmentation to data analysis is illustrated in Figure 2.

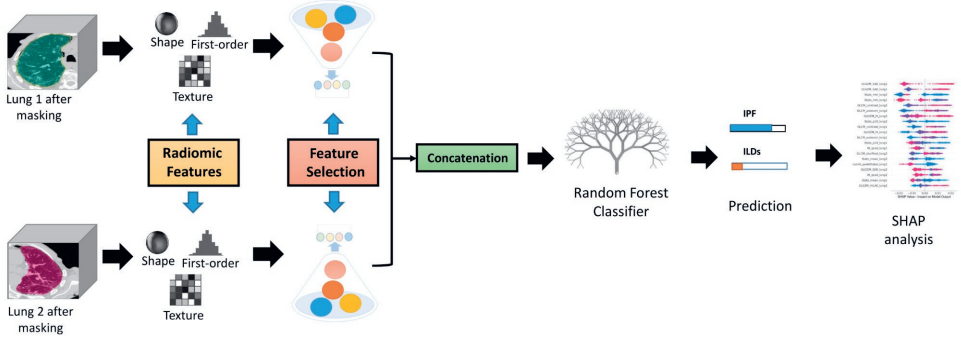


Figure 2: Radiomics Pipeline for Lung disease classification from CT images. The same 12 radiomics features from both lungs after feature selection are concatenated and fed to the Random Forest classifier. Post-hoc SHAP analysis is performed for interpretability.

Features selection and modeling

Features with near-zero variance (i.e. features that have the same value in ≥ 95 % of the data points) were excluded. Then, a correlation matrix was created between all HCR features and populated using Spearman's correlation coefficient (r). Feature pairs with $|r| \geq 0.90$ were considered to be highly correlated, and the feature with the highest average correlation with all other features was removed. Furthermore, a Recursive feature elimination (RFE) using a random forest classifier was performed on the subset of features that were selected after applying Spearman's correlation coefficient. RFE was applied with cross-validation in order to determine the accuracy of the classification and the top 12 features with the highest accuracy were selected for the final model. The same 12 features were extracted for each lung and concatenated to give a final feature vector consisting of 24 HCR features. A list of the names of the features along with their abbreviations that were used in the model can be found in Supplementary (Table E2). A random forest classifier was used to construct the HCR model to predict the probability of IPF in patients using HRCTs. Random forest classifier has proven to be effective for lungs CT-based radiomics problems in recent research findings (23-25). The random forest classifier was trained with class weights of 1 for non-IPF ILDs and 3 for IPF patients to compensate for the class imbalance. Five-fold cross-validation was used for hyper-parameter tuning.

Post-hoc Interpretability

SHapley Additive exPlanations (SHAP) analysis is based on co-operative game theory (20). SHAP analysis is a post-hoc interpretability method that quantifies the impact of each feature on the model prediction in terms of SHAP value. SHAP summary plots provide global explanations by highlighting the effect of features on the prediction in terms of SHAP value and help in recognizing the trends. These plots show whether a high or low feature value affects the model output positively or negatively. SHAP dependence plots highlight the relationship between the model output in terms of SHAP values and the corresponding feature values. These dependence plots can be useful for quantifying the trend of model output with respect to the feature values as well as understanding the interaction effects between a pair of features.

Deep learning (DL)

All the scans were resampled to an isotropic resolution of 1 x 1 x 1 mm^3. Min-max normalization was applied to the area within the lung mask. Two patches containing one lung each of size $240 \times 240 \times 240$ voxels were extracted using the lungs masks. Both lungs were randomly flipped for augmentation and concatenated along the z-axis. The image was then downsampled by taking every sixth slice along the z-axis. The start index was randomly chosen in the range of 1 to 6. This resulted in additional augmentation and reduction of the input image size. A Densenet-121 (26,27) classifier with 3D convolutional layers was used with weighted binary cross-entropy loss (non-IPF ILDs: 1, IPF: 3) in order

to minimize the effects of data imbalance. Adam optimizer with a learning rate of 1 e^{-5} and ReduceLROnPlateau scheduler was employed. The batch size was set at 16 and the network was trained for 50 epochs. Figure 3 shows the different steps involved in training the DL model for lung disease classification in CT images.

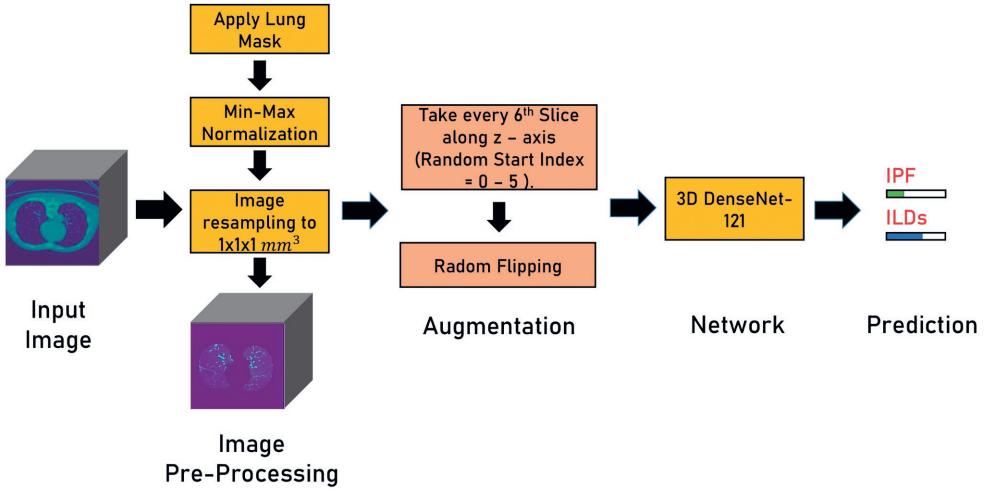


Figure 3: Figure shows different steps in the deep learning pipeline for the prediction of lung diseases in CT scans.

During prediction, six input images from the test image were extracted by setting the start slice index in the range from 1 to 6 and taking every sixth consecutive slice. These six test samples are fed to the trained 3D Densenet-121 model. The final prediction is the average of the prediction of these six test samples. Heatmaps highlight the regions of the input image that the model considers important for prediction. We utilized Grad-CAM (21) heatmaps for the post-hoc interpretability of the Densenet-121 model.

Ensemble Model

The ensemble methods utilize multiple machine learning methods in an effort to achieve better predictive performance as compared to the performance obtained by the constituent machine learning methods alone. We constructed an ensemble model from HCR and DL models by taking an average of the probabilities predicted by the two models.

***In-silico* Clinical Trial**

An application that allows the construction of a reference performance point by gathering medical imaging expert comments based on the visual assessment of HRCT images was created. The application allows displaying the CT images one at a time with the option of different planes (Axial, Coronal, or Sagittal), and the application also allows scrolling through the CT scan slices. The graphical user interface (GUI) of the application can be found in Supplementary (Figure E1). The radiologist can select one of the two classes (IPF or ILDs other

than IPF). The diagnostic performance of two radiologists (6 and 23 years of experience) and one pulmonologist (12 years of experience) was recorded for the same test dataset (n=109) to perform a comparison with the machine learning-based HCR, DL, and ensemble models.

Statistical analysis

Statistical analysis was performed in Python (version: 3.6). Wilcoxon rank-sum test was used for the continuous variables to test the group differences and Fisher exact test for categorical variables. To assess the model's performance, the areas under the curves (AUCs) for receiver operating characteristic (ROC) curves were compared using the DeLong test. The thresholds for each model were set at the highest Youden's index in the training set. The performance was evaluated using accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). For five-fold cross-validation, we also report the standard deviation (SD). The performance of the models on the test set was compared with the performance of clinicians using McNemar test. This study followed the Standard for Reporting Diagnostic accuracy studies (STRAD) (28) and was assessed using the Radiomics Quality Score (RQS) (29). The detailed description about RQS can be found in supplementary table E3.

Results

Patients Characteristics

A total of 474 patients, 335 of whom were diagnosed with non-IPF ILDs, and 139 with IPF, were included after the application of exclusion criteria (Figure 1). The demographic characteristics of the included patients can be found in Table 1.

Table 1. Demographic and clinical information of the study participants.

Variables	Site 1	Database A	P-value (*p*)
n	365	109	-
Age (mean(SD))	64.10 (9.57)	63.61 (14.17)	0.8
Sex = M (%)	213 (87	74 (67.9)	0.09
FEV1 (mean (SD))	80.42 (21.47)	69.60 (20.67)	< 0.001
FVC (mean(SD))	80.52 (21.25)	67.35 (21.37)	< 0.001
DLCO (mean(SD))	51.32 (24.99)	29.84 (5.36)	< 0.001
BMI (mean(SD))	25.48 (6.45)	29.55 (5.21)	< 0.001

Body mass index (BMI), forced expiratory volume (FEV), Forced vital capacity (FVC), and diffusion capacity of the lungs for carbon monoxide (DLCO) are shown in the table for different patients along with their mean and standard deviation (SD).

Handcrafted Radiomics

The HCR model achieved an AUC of 0.85 (95% CI: 0.771 – 0.924) in the validation set in five-fold cross-validation (Figure 4 (a)). The threshold of 0.51 was fixed based on Youden's index in the training set. An accuracy, sensitivity, and specificity of 0.762±0.068, 0.816±0.094, and

0.745±0.065 were obtained in five-fold cross-validation, respectively. In the external test set, the HCR model achieved an AUC, accuracy, sensitivity, and specificity of 0.817, 0.761, 0.698, and 0.821, respectively. Tables 2 and 3 show the performance metrics for the HCR model during five-fold cross-validation and external validation, respectively. Figure 4 (b) shows the test performance for the HCR model on the external dataset. The Radiomics Quality Score (RQS) achieved for this study is 52.78 % (19 of 36).

Table 2. Precision and recall metrics for five-fold cross-validation using handcrafted radiomics (HCR), deep learning (DL), and an ensemble of HCR and DL models.

Model	Accuracy	Sensitivity	Specificity	Positive Predictive Value (PPV)	Negative Predictive Value (NPV)
Handcrafted Radiomics (HCR)	0.762 ± 0.068	0.816 ± 0.094	0.745 ± 0.065	0.506 ± 0.084	0.923 ± 0.040
Deep Learning (DL)	0.779 ± 0.046	0.711 ± 0.10	0.800 ± 0.075	0.541 ± 0.074	0.901 ± 0.025
Ensemble (HCR + DL)	**0.852 ± 0.027**	**0.827 ± 0.005**	**0.860 ± 0.035**	**0.65 ± 0.063**	**0.94 ± 0.003**

Table 3. Comparison of diagnostic performance on the external test dataset for HCR, DL, an ensemble of HCR and DL, and *in-silico* trial with clinicians.

Model	Accuracy	Sensitivity	Specificity	Positive Predictive Value (PPV)	Negative Predictive Value (NPV)
Handcrafted Radiomics (HCR)	0.761	0.698	0.821	0.787	0.741
Deep Learning (DL)	0.779	0.792	0.768	0.763	0.796
Ensemble (HCR + DL)	**0.853**	**0.886**	**0.821**	**0.825**	**0.885**
In-silico trial with clinicians	0.66 ± 0.067	0.572 ± 0.186	0.750 ± 0.0525	0.680 ± 0.042	0.669 ± 0.100

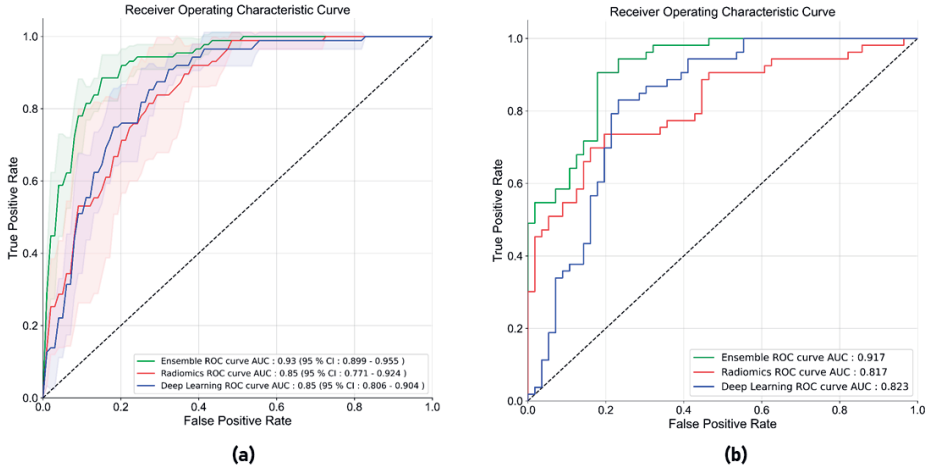


Figure 4: Receiver operating characteristics (ROC) curves for five-fold cross-validation (a) and external test dataset (b) for the classification of IPF and non-IPF ILDs using handcrafted radiomics (HCR), deep learning (DL), and ensemble (HCR + DL) models.

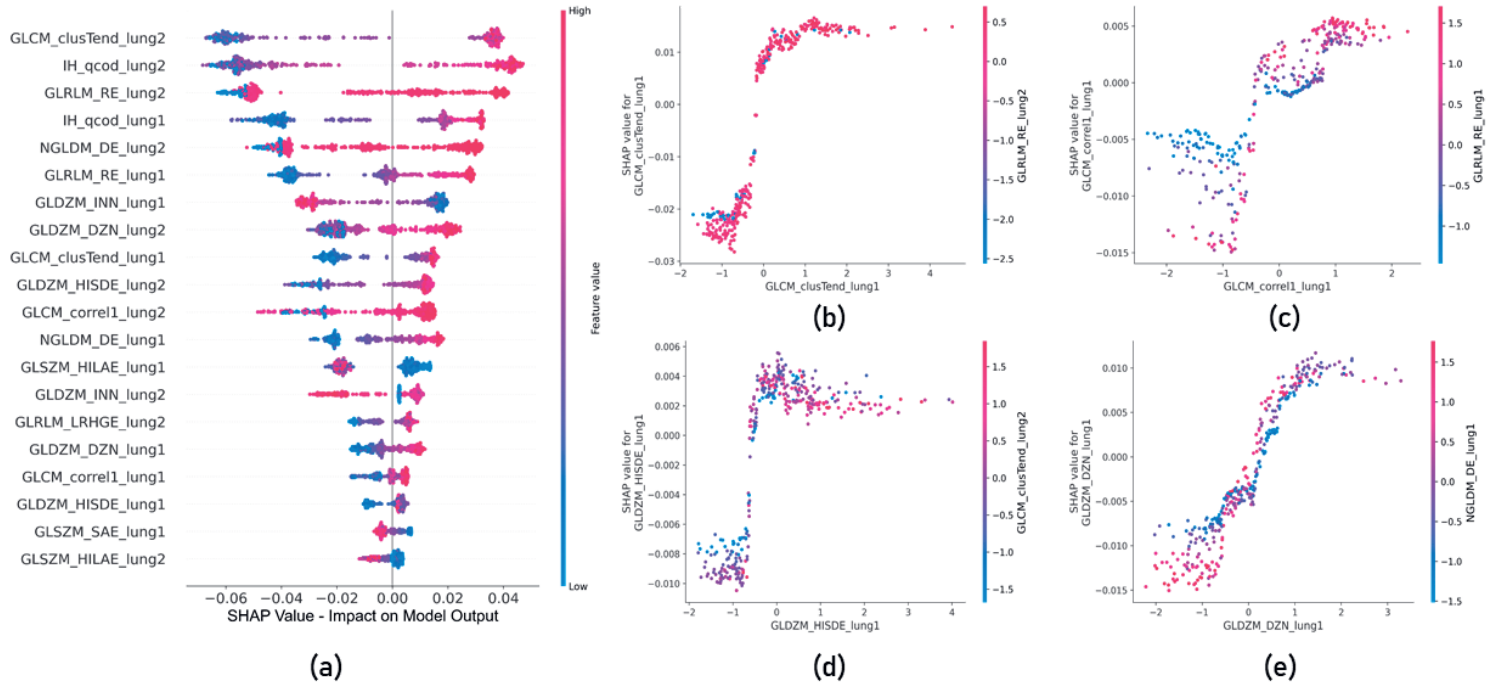


Figure 5: Global SHAP summary plots (a) demonstrate the impact of the top 20 features on the model output in terms of SHAP values and the corresponding feature values. SHAP dependence plots (b),(c),(d), and (e) show the effect of a particular feature value on the SHAP value and its interaction with another feature.

The global SHAP summary plots in Figure 5 (a) demonstrate that the same features extracted from each lung separately affect the model's prediction for IPF diagnosis in a similar way. A high feature value with a positive SHAP value forces the model's probability to be higher. The IH_qcod feature values extracted from lung1 and lung2 demonstrate a similar trend that a high feature value results in a positive SHAP value. However, there are some outliers in the trend that can seen be in features such as GLCM_correl1_lung and GLDZM_INN_lung. Similarly, the GLDZM_INN feature values extracted from lung1 and lung2 show a negative trend that a high feature value results in a negative SHAP value. Figure 5 (b,c,d,e) show the dependence plots of GLCM_clusTend, GLCM_correl1, GLDZM_HISDE, and GLDZM_DZN features, respectively. In Figure 5 (c), when the feature value of GLDZM_HISDE is low, high feature values of GLCM_clusTend result in a lower SHAP value. A similar effect can be seen in Figure 5 (d) between features GLDZM_DZN and NGLDM_DE.

Deep learning

The DL model achieved an AUC of 0.85 (95% CI: 0.806 – 0.904) in the validation set in five-fold cross-validation (Figure 4 (a)). The threshold of 0.45 was fixed based on Youden's index in the training set. An accuracy, sensitivity, and specificity of 0.779±0.046, 0.711±0.10, and 0.800±0.075 was achieved during five-fold cross-validation, respectively. In the external test set, the DL model achieved an AUC, accuracy, sensitivity, and specificity of 0.823, 0.853, 0.886, and 0.821, respectively. Tables 2 and 3 show the performance metrics for the HCR model during five-fold cross-validation and external validation, respectively. Figure 4 (b) shows the test performance for the DL model on the external dataset.

Figure 6 shows Grad-CAM overlayed on CT image slices obtained from HRCT scans from IPF and non-IPF ILDs patients. The overlayed heatmap shows the regions of the input image that the model considers important for prediction. The Grad-CAM focuses on the tissue pattern in the patient with IPF. However, no information is provided on how these areas contribute to the final model prediction.

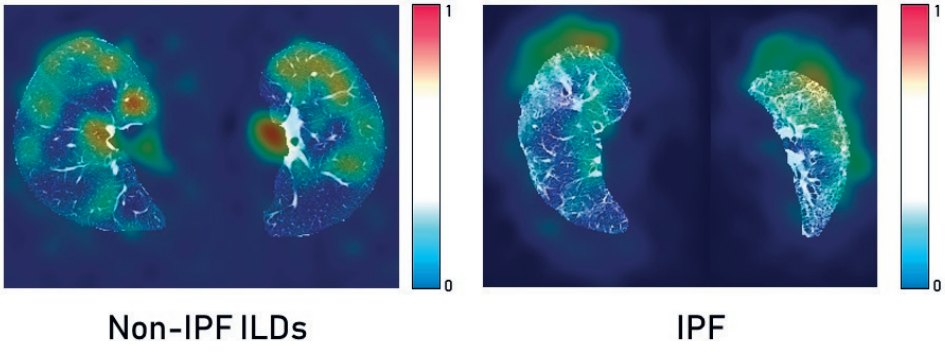


Figure 6: GradCAM heatmaps for post-hoc interpretability of IPF and non-IPF ILDs HRCT scans to understand the predictions made by the Densenet-121.

Ensemble

The ensemble model achieved an AUC of 0.93 (95% CI: 0.899 – 0.955) in the validation set during five-fold cross-validation (Figure 4 (a)). The threshold of 0.49 was fixed based on Youden's index in the training set. An accuracy, sensitivity, and specificity of 0.852±0.027, 0.827±0.005, and 0.860±0.035 was obtained during five-fold cross-validation, respectively. In the external test set, the DL model achieved an AUC, accuracy, sensitivity, and specificity of 0.917, 0.853, 0.886, and 0.821, respectively. Tables 2 and 3 show the performance metrics for the HCR model during five-fold cross-validation and external validation, respectively. The agreement between the predictions of HCR and DL models is 61.4%. The accuracy and specificity for the predictions when both the models agree were 93% and 97%, respectively. There was a statistically significant difference between the ROC curves for the ensemble model and HCR model (p = 0.02), and the ensemble model and the DL model (p = 0.005).

***In-silico* Clinical Trials**

Two radiologists and one pulmonologist achieved accuracies of 58.7%, 65.1%, and 75.2% with a mean of 66.3±6.7% for the diagnosis of IPF and non-IPF ILDs on the external test dataset. There was a statistically significant difference between performance of the ensemble model, and that of radiologists and pulmonologists (P < 0.05).

Discussion

In this study, we investigated the potential of HCR and DL to differentiate between different lung disorders i.e. IPF and non-IPF ILDs patients on HRCT scans. We also used post-hoc interpretability methods to explain the predictions of HCR and DL models. Moreover, we compare the performance of the proposed models to the diagnostic performance of radiologists using an *in-silico* trial on an external test set. Our results show that HCR and DL have a great potential to be used as an aid for clinical decision-making, which could minimize the time needed by radiologists, and increase diagnostic accuracy. The superior performance of an ensemble of DL and HCR models also demonstrates that these approaches can complement each other for lung disease diagnosis.

HCR and DL models achieved an accuracy of 76.2±6.8% and 77.9± 4.6% during five-fold cross-validation, respectively. In the external test set, HCR and DL models demonstrated a similar accuracy of 76.1% and 77.9%, respectively. There was no statistically significant difference between the ROC curves for HCR and DL models. The ensemble of HCR and DL models demonstrated the best accuracy of 85.2±2.7% and 85.3% for five-fold cross-validation and external test set, respectively. There was a statistically significant difference between the ROC curves for the ensemble model and HCR model (p = 0.023), and the ensemble model and DL model (p = 0.005). The HCR and DL models show an agreement of 61.4% for the predictions on the external test set. A sensitivity and specificity of 93% and 97% were

obtained when both the models agreed on the prediction. Hence, HCR and DL models add complementary value to each other resulting in a boost in performance.

We compared the performance of the developed models against the performance of the radiologists using a virtual clinical trial setting. The performance of HCR (76.1%), DL (77.9%), and ensemble (85.3 %) models were better than the performance of two radiologists and one pulmonologist (66 $\pm$ 7%) in discriminating IPF from non-IPF ILDs on the external test set. There was a statistically significant difference ($p < 0.05$) between the predictions of the ensemble model, and the two radiologists and one pulmonologist. There was a significant difference ($p < 0.001$) in the BMI, FEV, FVC, and DLCO values between site 1 and database A. The models demonstrated similar performance on the external database A despite the variability, showing that the trained models are robust and generalize well.

The clinical translation of HCR and DL is limited due to the "black-box" nature of the underlying complex classifiers. It is difficult for clinicians to understand the underlying mechanisms that govern the decision-making process of these complex classifiers. SHAP post-hoc explanations discover the patterns of the complex classifiers and increase transparency. SHAP global summary plots showed that Gray-level Co-occurrence Matrix Cluster Tendency and Intensity Histogram quartile coefficient of dispersion are the most important features for IPF diagnosis. These plots also showed that the same features extracted from different lungs demonstrate a similar trend in SHAP impact value. SHAP dependence plots demonstrated the effect of a single feature value and the interaction between a pair of features on the model output. Grad-CAM heatmaps highlight the area that the DL model considers important for the final prediction. These heatmaps can reinforce the trust in the model predictions if the model is focusing on the area relevant to the clinical task. However, Grad-CAM heatmaps do not offer any explanation of how the highlighted area contributes to the final prediction. Although DL demonstrates good performance, it is more opaque in nature due to its complexity that might hinder its clinical adoption.

Some studies previously investigated the potential of HCR and DL algorithms to classify lung disorders. Walsh el al. (18) employed a DL algorithm on a dataset of 1157 HRCT images for the diagnosis of fibrotic lung disease. The algorithm performance was compared to that of 91 radiologists and revealed an accuracy of 73.3 %, compared to the radiologist's median accuracy of 70.7 %. When compared to Walsh et al. (18), our study demonstrated greater accuracy using HCR (76.1%), DL (77.9%), and an ensemble of HCR and DL (85.3%). Christe et al. (30) conducted another study in which they employed a computer-aided diagnostic (CAD) system (INTACT system) to diagnose IPF cases based on HRCT images and compared the performance of the CAD system to the performance of radiologists. Their findings showed that the two radiologists and the CAD system obtained an accuracy of 60 %, 54 %, and 56 % respectively. Mean RQS score of 20.4%, 26.1%, and 27.4% were obtained after recent

analyses of papers reporting radiomics studies (31–33). This shows that RQS is a stringent and demanding criterion (34-36) that aims to encourage the best scientific practice. An RQS of 52.78% shows that this study tries to adhere to the best scientific practices and reporting guidelines.

This study has some limitations. The datasets utilized for this study contain HRCT scans acquired with different CT acquisition and reconstruction settings that can influence HCR feature values (37). Hence, phantom studies to evaluate the reproducibility of the HCR features or harmonization investigations need to be carried out to make a more robust HCR pipeline (38). Grad-CAMs only highlight the region of the input image that the model considers important for the decision-making process. There is a need to utilize interpretability methods that give an insight into how the relevant region contributes to the decision-making process (19). The high performance of an ensemble of HCR and DL model shows that these two approaches add complementary values. It may be useful to employ an interpretability method such as concept attribution that will investigate the HCR features that the DL model considers important for classification (39). A prospective virtual *in-silico* trial in a real-world environment where the predictions of DL/HCR model and post-hoc interpretability plots are made available to the doctors during diagnosis should be carried out to confirm the clinical utility of the proposed methods. The quality of lung segmentation can affect the performance of the models. Therefore, it is important to ensure the quality of the automatic segmentation in the presence of variability such as noise and artifacts

At the moment, there is little research on the diagnosis of ILDs using HCR and DL. The reported results are encouraging and highlight the significant potential of HCR and DL methods for the diagnosis of IPF. In the future, HCR and DL approaches may be expanded to include treatment decisions. More studies should be conducted to explore the development of IPF at baseline and follow-up, as well as to assess the efficacy of anti-fibrotic treatment.

Conclusion

In this study, we developed handcrafted radiomics and deep learning models for the classification of IPF and non-IPF ILDs using HRCTs. In addition, we compared the performance of both models to radiologists on an external test dataset. HCR, DL, and ensemble models demonstrated better accuracy than radiologists in a virtual *in-silico* clinical trial setting. An ensemble of HCR and DL models demonstrated the best performance highlighting the complementary value of the two quantitative approaches for lung disease diagnosis. SHAP and GRAD-CAM post-hoc interpretability methods are useful for explaining the predictions made by radiomics and DL models respectively. These automated diagnostic tools can serve as a useful clinical aid for diagnosing different lung diseases.

References

1. Coultas DB, Zumwalt RE, Black WC, Sobonya RE. The epidemiology of interstitial lung diseases. *Am J Respir Crit Care Med* (1994) **150**:967–972. doi: 10.1164/ajrccm.150.4.7921471
2. Cottin V. Pulmonary fibrosis: "idiopathic" is not "cryptogenic." *Eur Respir J* (2019) **53**: doi: 10.1183/13993003.02314-2018
3. Travis WD, Costabel U, Hansell DM, King TE Jr, Lynch DA, Nicholson AG, Ryerson CJ, Ryu JH, Selman M, Wells AU, et al. An official American Thoracic Society/European Respiratory Society statement: Update of the international multidisciplinary classification of the idiopathic interstitial pneumonias. *Am J Respir Crit Care Med* (2013) **188**:733–748. doi: 10.1164/rccm.201308-1483ST
4. Raghu G, Remy-Jardin M, Myers JL, Richeldi L, Ryerson CJ, Lederer DJ, Behr J, Cottin V, Danoff SK, Morell F, et al. Diagnosis of Idiopathic Pulmonary Fibrosis. An Official ATS/ERS/JRS/ALAT Clinical Practice Guideline. *Am J Respir Crit Care Med* (2018) **198**:e44–e68. doi: 10.1164/rccm.201807-1255ST
5. Fischer A, du Bois R. Interstitial lung disease in connective tissue disorders. *Lancet* (2012) **380**:689–698. doi: 10.1016/S0140-6736(12)61079-4
6. Fernández Pérez ER, Swigris JJ, Forssén AV, Tourin O, Solomon JJ, Huie TJ, Olson AL, Brown KK. Identifying an inciting antigen is associated with improved survival in patients with chronic hypersensitivity pneumonitis. *Chest* (2013) **144**:1644–1651. doi: 10.1378/chest.12-2685
7. Raghu G, Collard HR, Egan JJ, Martinez FJ, Behr J, Brown KK, Colby TV, Cordier J-F, Flaherty KR, Lasky JA, et al. An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. *Am J Respir Crit Care Med* (2011) **183**:788–824. doi: 10.1164/rccm.2009-040GL
8. Richeldi L, Collard HR, Jones MG. Idiopathic pulmonary fibrosis. *Lancet* (2017) **389**:1941–1952. doi: 10.1016/S0140-6736(17)30866-8
9. Ley B, Collard HR, King TE Jr. Clinical course and prediction of survival in idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* (2011) **183**:431–440. doi: 10.1164/rccm.201006-0894CI
10. Guiot J, Vaidyanathan A, Deprez L, Zerka F, Danthine D, Frix A-N, Lambin P, Bottari F, Tsoutzidis N, Miraglio B, et al. A review in radiomics: Making personalized medicine a reality via routine imaging. *Med Res Rev* (2022) **42**:426–440. doi: 10.1002/med.21846
11. Cho YH, Seo JB, Lee SM, Lee SM, Choe J, Lee D, Kim N. Quantitative CT imaging in chronic obstructive pulmonary disease: Review of current status and future challenges. *Korean J Radiol* (2018) **78**:1. doi: 10.3348/jksr.2018.78.1.1
12. Hochhegger B, Marchiori E, Zanon M, Rubin AS, Fragomeni R, Altmayer S, Carvalho CRR, Baldi BG. Imaging in idiopathic pulmonary fibrosis: diagnosis and mimics. *Clinics* (2019) **74**:e225. doi: 10.6061/clinics/2019/e225
13. Tominaga J, Sakai F, Johkoh T, Noma S, Akira M, Fujimoto K, Colby TV, Ogura T, Inoue Y, Taniguchi H, et al. Diagnostic certainty of idiopathic pulmonary fibrosis/usual interstitial pneumonia: The effect of the integrated clinico-radiological assessment. *Eur J Radiol* (2015) **84**:2640–2645. doi: 10.1016/j.ejrad.2015.08.016

14. Walsh S, de Jong EEC, van Timmeren JE, Ibrahim A, Compter I, Peerlings J, Sanduleanu S, Refaee T, Keek S, Larue RTHM, et al. Decision Support Systems in Oncology. *JCO Clin Cancer Inform* (2019) **3**:1–9. doi: 10.1200/CCI.18.00001
15. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RGPM, Granton P, Zegers CML, Gillies R, Boellard R, Dekker A, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer* (2012) **48**:441–446. doi: 10.1016/j.ejca.2011.11.036
16. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology* (2016) **278**:563–577. doi: 10.1148/radiol.2015151169
17. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* (2015) **521**:436–444. doi: 10.1038/nature14539
18. Walsh SLF, Calandriello L, Silva M, Sverzellati N. Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort study. *Lancet Respir Med* (2018) **6**:837–845. doi: 10.1016/S2213-2600(18)30286-8
19. Salahuddin Z, Woodruff HC, Chatterjee A, Lambin P. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Comput Biol Med* (2021) **140**:105111. doi: 10.1016/j.compbiomed.2021.105111
20. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Proceedings of the 31st international conference on neural information processing systems*. (2017). p. 4768–4777 http://papers.nips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
21. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE international conference on computer vision*. (2017). p. 618–626 http://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html
22. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, Ashrafinia S, Bakas S, Beukinga RJ, Boellaard R, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology* (2020) **295**:328–338. doi: 10.1148/radiol.2020191145
23. Jia T-Y, Xiong J-F, Li X-Y, Yu W, Xu Z-Y, Cai X-W, Ma J-C, Ren Y-C, Larsson R, Zhang J, et al. Identifying EGFR mutations in lung adenocarcinoma by noninvasive imaging using radiomics features and random forest modeling. Eur Radiol (2019) 29:4742–4750.
24. Bashir U, Kawa B, Siddique M, Mak SM, Nair A, Mclean E, Bille A, Goh V, Cook G. Non-invasive classification of non-small cell lung cancer: a comparison between random forest models utilising radiomic and semantic features. Br J Radiol (2019) 92:20190159.
25. Jiang C, Luo Y, Yuan J, You S, Chen Z, Wu M, Wang G, Gong J. CT-based radiomics and machine learning to predict spread through air space in lung adenocarcinoma. Eur Radiol (2020) 30:4050–4057.
26. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*. (2017). p. 4700–4708 http://openaccess.thecvf.com/content_cvpr_2017/html/Huang_Densely_Connected_Convolutional_CVPR_2017_paper.html

27. MONAI Consortium. *MONAI: Medical Open Network for AI*. (2020). doi: 10.5281/zenodo.5728262
28. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, Lijmer JG, Moher D, Rennie D, de Vet HCW, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ* (2015) **351**:h5527. doi: 10.1136/bmj.h5527
29. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, Sanduleanu S, Larue RTH, Even AJG, Jochems A, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology* (2017) **14**:749–762. doi: 10.1038/nrclinonc.2017.141
30. Christe A, Peters AA, Drakopoulos D, Heverhagen JT, Geiser T, Stathopoulou T, Christodoulidis S, Anthimopoulos M, Mougiakakou SG, Ebner L. Computer-Aided Diagnosis of Pulmonary Fibrosis Using Deep Learning and CT Images. *Invest Radiol* (2019) **54**:627–632. doi: 10.1097/RLI.0000000000000574
31. Park JE, Kim D, Kim HS, Park SY, Kim JY, Cho SJ, Shin JH, Kim JH. Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. Eur Radiol (2020) 30:523–536.
32. Lee S, Han K, Suh YJ. Quality assessment of radiomics research in cardiac CT: a systematic review. Eur Radiol (2022) doi: 10.1007/s00330-021-08429-0
33. Stanzione A, Gambardella M, Cuocolo R, Ponsiglione A, Romeo V, Imbriaco M. Prostate MRI radiomics: A systematic review and radiomic quality score assessment. Eur J Radiol (2020) 129:109095.
34. Spadarella G, Calareso G, Garanzini E, Ugga L, Cuocolo A, Cuocolo R. MRI based radiomics in nasopharyngeal cancer: Systematic review and perspectives using radiomic quality score (RQS) assessment. Eur J Radiol (2021) 140:109744.
35. Won SY, Park YW, Ahn SS, Moon JH, Kim EH, Kang S-G, Chang JH, Kim SH, Lee S-K. Quality assessment of meningioma radiomics studies: Bridging the gap between exploratory research and clinical applications. Eur J Radiol (2021) 138:109673.
36. Park JE, Kim HS, Kim D, Park SY, Kim JY, Cho SJ, Kim JH. A systematic review reporting quality of radiomics research in neuro-oncology: toward clinical utility and quality improvement using high-dimensional imaging features. BMC Cancer (2020) 20:29.
37. Ibrahim A, Primakov S, Beuque M, Woodruff HC, Halilaj I, Wu G, Refaee T, Granzier R, Widaatalla Y, Hustinx R, et al. Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework. *Methods* (2021) **188**:20–29. doi: 10.1016/j.ymeth.2020.05.022
38. Mali SA, Ibrahim A, Woodruff HC, Andrearczyk V, Müller H, Primakov S, Salahuddin Z, Chatterjee A, Lambin P. Making Radiomics More Reproducible across Scanner and Imaging Protocol Variations: A Review of Harmonization Methods. *J Pers Med* (2021) **11**: doi: 10.3390/jpm11090842
39. Graziani M, Andrearczyk V, Müller H. Regression Concept Vectors for Bidirectional Explanations in Histopathology. *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*. Springer International Publishing (2018). p. 124–132 doi: 10.1007/978-3-030-02628-8_14

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$minimum = \min(\mathbf{X})$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2$$

$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_s} \mathbf{P}(i,j)\right)^2}{N_p}$$

$$correlation = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) ij - \mu_x \mu_y}{\sigma_x(i) \sigma_y(j)}$$

$$MAD = \frac{1}{N_p} \sum_{i=1}^{N_p} |\mathbf{X}(i) - \bar{X}|$$

$$minimum = \min(\mathbf{X})$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_s} \mathbf{P}\right.}{N_z}$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2$$

Part IV

$$variance = \frac{1}{N_p}\sum_{i=1}^{N_p}(\mathbf{X}(i)-\bar{X})^2$$

$$total\ energy = V_{voxel}\sum_{i=1}^{N_p}(\mathbf{X}(i)+c)^2$$

$$SDE = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}\frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$minimum = \min(\mathbf{X})$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2$$

$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_s} \mathbf{P}(i,j)\right)^2}{N_p}$$

$$correlation = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) ij - \mu_x \mu_y}{\sigma_x(i) \sigma_y(j)}$$

$$MAD = \frac{1}{N_p} \sum_{i=1}^{N_p} |\mathbf{X}(i) - \bar{X}|$$

$$minimum = \min(\mathbf{X})$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_s} \mathbf{P}\right.}{N_z}$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2$$

Chapter 10

General discussion and future perspective

The application of artificial intelligence (AI) in diagnostic medical imaging is increasingly a topic of many different research projects. A great deal of the research makes use of handcrafted radiomics or deep learning algorithms to complete various tasks in a range of different medical imaging modalities (1,2) (Figure 1). AI has demonstrated outstanding levels of accuracy and sensitivity in the identification of imaging abnormalities, and it has the potential to improve tissue-based detection and characterization (3–5). To guarantee successful and safe inclusion of AI-assisted diagnostic imaging in clinical practices, the medical community must anticipate possible unknowns underlying these technologies already at the start of the AI-assisted diagnostic imaging revolution. A careful assessment of AI's possible risks in the context of its unique abilities is critical when establishing its place in clinical medicine. Though it should be pointed out that straddling the line between better detection and overdiagnosis will be difficult. When establishing this assessment, the regular use of out-of-sample external validation and well-defined cohorts to improve the quality and interpretability of AI studies will be of critical importance (6).

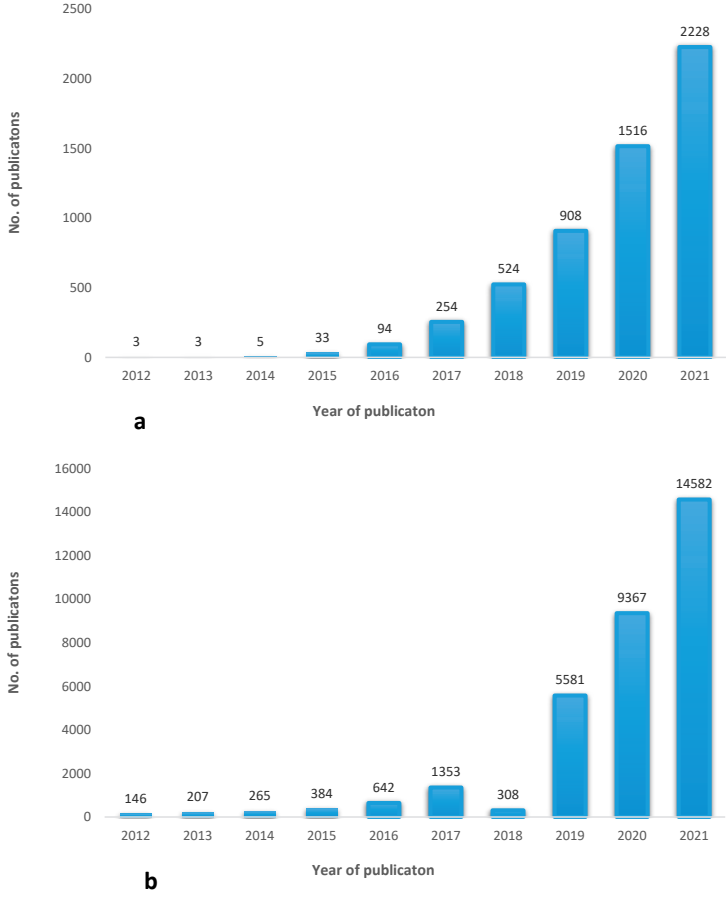


Figure 1. Number of publications on PubMed search; a) radiomics; b) deep learning.

This thesis provided two AI methods: a) handcrafted radiomics and b) deep learning (DL). As described in this thesis, the overall goals of the study for handcrafted radiomics were (i) to acquire better insights into their reproducibility (Figure 2) and (ii) to evaluate their potential in the categorization of various types of lung disorders. The primary goal for DL is to examine its capacity to classify various types of lung diseases. This chapter provides an in-depth discussion of the work completed in this thesis as well as future perspectives.

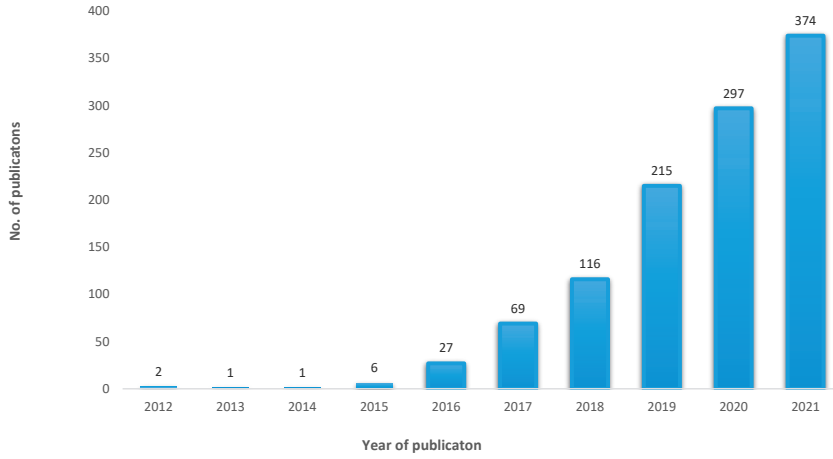


Figure 2. Number of publications on the reproducibility of handcrafted radiomics in the period between 2012 and 2021, based on PubMed research.

Reproducibility of handcrafted radiomics (HRFs)

In **chapter 3**, we investigated the robustness of HRFs on a dataset consisting of 13 phantom CT scans. The scans were obtained from different vendors, with different CT parameters. After the extraction of HRFs from the 13 scans, we assessed their reproducibility using the concordance correlation coefficient (CCC). The study's findings indicated that only a small percentage of HRFs were robust to differences in the imaging settings examined. The majority of the HRFs were reliant on imaging parameter changes. Furthermore, when applying ComBat harmonization to phantom scans, the findings demonstrated that ComBat's capacity to harmonize HRFs depends on variations in imaging parameters. However, the performance of ComBat harmonization may suffer as a result of treating each scan as a unique batch effect, despite the fact that variations between pair batches are not similar.

The reproducibility of hepatocellular carcinoma (HCC) HRFs, generated from various phases of contrast-enhanced CT images (CECT), was evaluated in **chapter 4**. For this study, HCC patients' arterial and venous CT scans were made accessible. To ensure that the same region of interest (ROI) was placed in the right position in both phases, the segmentation of ROIs

was performed on one phase and then replicated in the other. The finding of the presented study showed that, when no image settings were changed, a subset of HRFs were shown to be reproducible in both phases. In addition, the use of the ComBat harmonization approach resulted in an increase in reproducible HRFs by 1% across phases. This study also found that a number of HRFs may be utilized interchangeably across arterial and venous phase CT scans and that combining these scans might enhance the information gathered from HCC lesions. However, we speculate that the subgroup of reproducible HRFs identified in our study is confined to the HCC lesions derived from scans collected in a manner similar to our dataset. Furthermore, the reproducibility of the discovered HRFs must be tested using different acquisition and reconstruction conditions, which was not achievable due to a lack of data.

In **chapter 5,** we investigated the use of Reconstruction Kernel Normalization (RKN) and ComBat harmonization to improve the reproducibility of HRFs across scans acquired with different reconstruction kernels. A sample of 28 phantom scans collected on five distinct scanner types was evaluated. HRFs were derived from the original scans, and scans were harmonized using the RKN approach. ComBat harmonization was applied on both set of HRFs. Concordance correlation coefficient (CCC) was used to assess the reproducibility of HRFs. McNemar's test was used to determine the difference in the number of reproducible HRFs in each scenario. The results of the study showed that the majority of HRFs were found to be sensitive to variations in the reconstruction kernels, and only six HRFs were found to be robust with respect to variations in reconstruction kernels. Furthermore, combining RKN and ComBat harmonization led in considerable increases in reproducible HRFs as compared to HRFs derived from original images. For future radiomic studies, we suggest the systematic use of pre- and post- processing approaches in images collected with similar image acquisition and reconstruction parameters, except for the reconstruction kernels.

In **chapter 6**, we used a phantom dataset (n = 14) collected on two scanner types, the Discovery STE and the LightSpeed Pro 32, to examine the impact of changes in in-plane spatial resolution (IPR) on HRFs. All other imaging parameters were kept constant. Ten ROIs were performed for each scan, and HRFs were extracted from each ROI. CCC was used to evaluate HRF reproducibility across pairs of phantom CT images. Moreover, we looked at how ten various image resampling techniques (IR), as well as ComBat harmonization, affected the HRFs. According to the findings of this study, certain HRFs are immune to changes in pixel spacing; however, the reproducibility of the remaining HRFs depends on the degree of variation in pixel spacing. Furthermore, compared to the other IR techniques, scans resampled using cosine windowed sinc interpolation exhibited the largest number of concordant HRFs among the types of IR. The impacts of IR and ComBat harmonization on the reproducibility of HRFs, on the other hand, were shown to rely significantly on the variances in the scans being evaluated.

HRFs in lung disorders

In **chapter 2**, the current state of play of handcrafted radiomics and deep learning was evaluated with the use of a literature review. In this review, we provided a broad overview and update on the rapidly expanding field of quantitative imaging research, focusing on the two arms "handcrafted radiomics and deep learning." The chapter describes some of its limitations and provides examples of emerging clinical implementation, which are the stepping stones toward precision medicine.

In **chapter 7**, we provided an overview of available literature concerning the use of handcrafted radiomics in lung cancer – in terms of detection, treatment response, and prognosis. While the research on applying handcrafted radiomics in lung cancer has been increasing in recent times, the application of handcrafted radiomics on chronic obstructive pulmonary disease is still limited. The use of quantitative CT (QCT) has been shown to be able to quantify emphysema, airway abnormalities, and air trapping. However, the interpretation of QCT is still time-consuming, requires experts, and is prone to variability in the diagnosis between experts. The use of CT image metrics (radiomics) could be able to quantify COPD and identify the disease's underlying mechanism, as well as the relationship between lung cancer and COPD, in a more nuanced and stronger form of phenotypic categorization. Potentially, handcrafted radiomics might be useful in detecting and classifying between COPD stages and phenotypes, allowing for the early treatment for the patient.

In **chapter 8,** we investigated the application of handcrafted radiomics on interstitial lung disease (ILDs). The data used in this study was collected from one center and two databases. Four groups were included in the study, namely: a) IPF with UIP pattern presentation on HRCT, b) IPF with UIP presentation confirmed by surgical lung biopsy, b) non-IPF ILDs with surgical lung biopsy confirming the absence of a UIP pattern, and c) healthy lung subjects. Two lung segmentations were performed, one with whole lungs and the other with sectorized lungs. Briefly stated, we were able to demonstrate that radiomic features derived from HRCT images can be used to distinguish between a normal state and ILDs, as well as between IPF with a UIP pattern and ILDs with no UIP pattern verified by surgical biopsy. In addition, our investigation revealed a substantial variation in tracheal volume between normal, IPF/UIP, and non-IPF ILDs patients. The volume of the trachea was shown to be greater in IPF participants compared to normal and non-IPF ILDs. In addition, the performance on the external dataset was decreased. This decline in the performance might be explained by the fact that the computation of HRFs is highly dependent on the variation in acquisition and reconstruction parameters. For this reason, the need to assess the reproducibility of HRFs is of great importance. Nevertheless, it is not currently possible to perform a reproducibility study due to the lack of anthropomorphic phantom or test-retest scans acquired with settings similar to the scans used in this study.

In **chapter 9,** a similar analysis to that found in chapter 8 was performed on classifying different parenchymal lung diseases. Data was collected from one center and one databases. Whole lung segmentation was performed for each scan and HRFs were extracted from each lung. The models were trained on center 1, and validated on database A. The finding of the study showed the ability HRFs have in terms of the classification of different types of lung disorders, namely IPF, and non- IPF ILDs lung. The model's performance in the external validation dataset was better than that seen in chapter 8, with the same external validation dataset. The reason for this is might be that, in chapter 8, the training dataset was homogenous; however, in chapter 9, the training dataset was heterogeneous, and the machine learning algorithm most likely learned some of the differences in the training dataset that may already be present in the validation dataset.

DL in lung disorders

In addition to the application of HRFs, **chapter 9** outlines the development of a DL algorithm that might be used to identify various lung disorders. The identical training and validation data split utilized in chapter 9 for HRF models was employed for DL algorithms. Two patches with one lung – each with a size of $240 \times 240 \times 240$ voxles – were extracted using the lung mask. To reduce the impacts of data imbalance, a Densenet-121 classifier with 3D convolutional layers and weighted binary cross entropy loss was utilized. It was found that DL findings were similar to HRFs findings. However, due to its complexity, DL is less transparent in nature, which may impede its clinical adaptation.

Ensemble learning

Ensemble learning combines many different machine learning algorithms to obtain higher prediction performance than any single learning method alone (7). In **chapter 9**, an ensemble model of HRFs and DL was developed by taking the average of both models' performance in external validation, resulting in a greater level of accuracy when identifying IPF and non-IPF ILDs than either approach alone. HRFs and DL models both provide complimentary value to one another, resulting in improved overall performance.

Interpretability of HRFs and DL

One significant limitation of both handcrafted radiomics and DL is the absence of clinical routine interpretability (8). In **chapter 9**, a post-hoc interpretability approach, based on a SHAP analysis, was used to interpret HRF models, allowing us to visualize the influence of feature values on the model output of each class in terms of SHAP values. We evaluated the most important HRFs related to each class using SHAP analysis (IPF and non-IPF ILDs). For DL, Gradient-weighted Class Activation Mapping (Grad-CAM) was used to interpret

the performance of DL models. Both SHAP analysis for HRF models and Grad-CAM for DL models provided an insight into the reasoning process behind these models. However, Grad-CAM does not offer an explanation as to how the highlighted area contributes to the final prediction. A well-defined mathematical method is used to calculate HRFs, which makes them more understandable. On the other hand, with deep learning, the process from input images to prediction is less transparent , which may be detrimental in understanding good or bad model performance.

In-silico clinical trial

ISCT – also known as virtual clinical trials or virtual imaging trials – is increasingly playing a role in ascertaining and qualifying the effectiveness of medical imaging technologies or AI algorithms, as evidenced in a few recent FDA approvals based on ISCT (9). We therefore embarked on an ISCT to evaluate the performance of both handcrafted radiomics and DL tools, compared to the evaluation of medical doctors in **chapter 9**. The decision of two radiologists and one pulmonologist on the diagnosis was collected in the test dataset (n=109) for the same number of cases, in order to equate their results with the performance of the models. The findings showed that both HRFs, DL, and ensemble models had higher levels of accuracy than the doctors' mean accuracy in classifying IPF and non-IPF ILDs. Such findings point to the necessity for image-based categorization approaches to be combined with clinician input in order to obtain the most accurate diagnosis.

Future perspective

This thesis has made significant achievements in exploring the reproducibility of handcrafted radiomics and unraveling the challenges impeding the full potential of the field from being utilized. Such challenges include the reproducibility and repeatability of image-based features, the interpretability of signatures, and the need for big(ger) data. In fact, the majority of the work in this thesis was devoted to comprehending and overcoming the limitations of HRFs.

Several concerns must be resolved before HRFs may be used in real clinical practices. Future studies on the reproducibility of HRFs across multiple imaging settings should cover a broader range of imaging parameters. A bigger dataset with more variations might also increase our knowledge of the cumulative impact of the variances found in imaging parameters on HRF reproducibility, and therefore our capacity to improve and establish robust radiomic signatures, eventually leading to more personalized medicine and better patient outcomes (**chapter 3-6**). In addition, due to the fact that variations in imaging parameters can significantly alter the majority of HRFs, it is important to develop a method (or methods) of harmonization that takes imaging parameter differences into account. One newly proposed

method involves employing deep networks such as convolutional neural networks (CNN) or generative adversarial networks (GAN) to synthesize pictures with increasingly comparable features, aiming at multicenter harmonization (10). It is also necessary to examine HRFs' repeatability as well as their vulnerability to inter-reader variability.

The work described in **chapters 8 and 9** in this thesis only takes into account a single time-point. The method of delta-radiomics has previously shown the capacity to predict treatment responses in lung cancer (11). As a result, such a technique might be used to quantify the progression of the disease and the impact of (new) treatments. Regarding IPF patients, it would be very important for future research to include delta-radiomics, in order to investigate the efficacy of treatment in different time-points (12). In addition, delta-radiomics might be used to examine the difference between inspiration and expiration scans and to uncover hidden information that could aid in determining the extent and severity of pulmonary emphysema. Furthermore, future research will focus on determining the prognostic or predictive significance of these features, as well as developing appropriate modeling tools that allow for meaningful inclusion in longitudinal data.

In **chapter 9**, deep learning models were built on images that have already been segmented. Future work will involve using deep learning to segment lungs with different types of lung disorders. Furthermore, deep learning may automatically uncover visual features that are suitable for a certain purpose through an optimization process – including features with varying levels of complexity – without the need for human intervention (13,14). Although deep learning is a promising advancement, one significant difficulty is the need for enormous amounts of data. Nonetheless, the additional benefit of HRFs to deep learning should not be neglected, as it may be more practical and effective than significantly increasing the number of samples used to train a deep learning model (15,16). Furthermore, the process from input images to prediction is less transparent with DL, which may be detrimental to understanding model performance. In the future, HRFs may be used to explain the ambiguity of DL models in an attempt to make them more understandable (Figure 3).

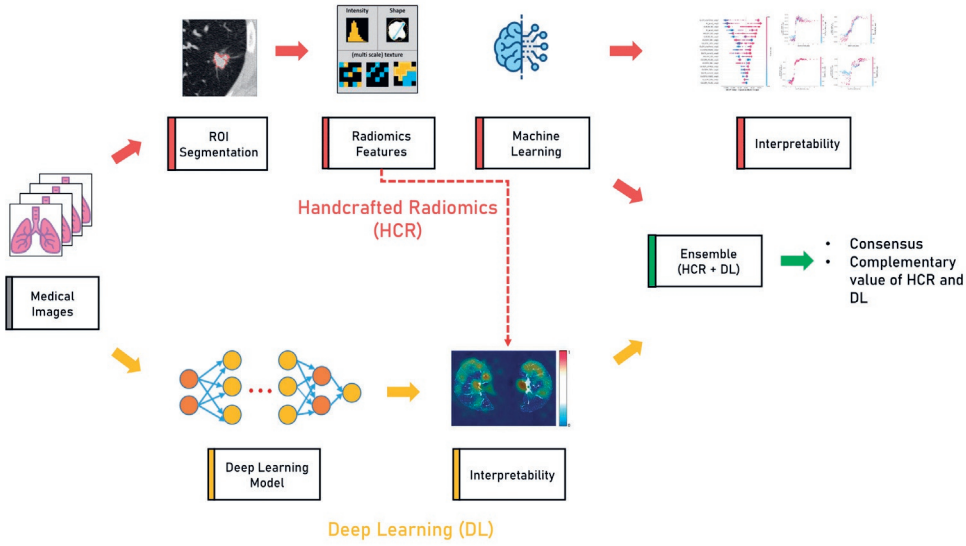


Figure 3. Overview of the process of both HCR and DL and the ensemble of both methods.

Conclusion

This thesis was divided into two parts: (i) investigating the reproducibility of HRFs (**chapters 3-6)**, and (ii) evaluating specific applications of HRFs and DL (**chapters 8-9**). Numerous studies focus on the impact of various acquisition and reconstruction parameters on the reproducibility of HRFs (17–19). The application of HRFs in differentiating between types of lung disorders in this thesis is promising, showing their potential to be applied in clinical practices. However, future work on investigating the reproducibility of those models is crucial and should not be ignored. The DL algorithm demonstrated its capacity to execute several tasks in medical image analysis, indicating its potential for supporting clinical decisions. Both handcrafted radiomics and DL have the potential to greatly contribute to clinical decision-making in the future, which together will enhance patient outcomes. Nevertheless, the challenge has yet to be fully solved.

References

1. Rogers W, Thulasi Seetha S, Refaee TAG, et al. Radiomics: from qualitative to quantitative imaging. Br J Radiol. 2020;93(1108):20190948.
2. Morin O, Vallières M, Jochems A, et al. A Deep Look Into the Future of Quantitative Imaging in Oncology: A Statement of Working Principles and Proposal for Change. Int J Radiat Oncol Biol Phys. 2018;102(4):1074–1082.
3. Kim H-E, Kim HH, Han B-K, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. Lancet Digit Health. 2020;2(3):e138–e148.
4. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nature Reviews Clinical Oncology. 2017. p. 749–762. doi: 10.1038/nrclinonc.2017.141.
5. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology. 2016;278(2):563–577.
6. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health. 2019;1(6):e271–e297.
7. Dietterich TG. Ensemble Methods in Machine Learning. Multiple Classifier Systems. Springer Berlin Heidelberg; 2000. p. 1–15.
8. Salahuddin Z, Woodruff HC, Chatterjee A, Lambin P. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. Comput Biol Med. 2021;140:105111.
9. Abadi E, Segars WP, Tsui BMW, et al. Virtual clinical trials in medical imaging: a review. J Med Imaging (Bellingham). 2020;7(4):042805.
10. Hognon C, Tixier F, Gallinato O, Colin T, Visvikis D, Jaouen V. Standardization of multicentric image datasets with generative adversarial networks. IEEE Nuclear Science Symposium and Medical Imaging Conference 2019. 2019. https://hal.archives-ouvertes.fr/hal-02447807/.
11. Fave X, Zhang L, Yang J, et al. Delta-radiomics features for the prediction of patient outcomes in non-small cell lung cancer. Sci Rep. 2017;7(1):588.
12. Humphries SM, Yagihashi K, Huckleberry J, et al. Idiopathic Pulmonary Fibrosis: Data-driven Textural Analysis of Extent of Fibrosis at Baseline and 15-Month Follow-up. Radiology. 2017;285(1):270–278.
13. Carneiro G, Oakden-Rayner L, Bradley AP, Nascimento J, Palmer L. Automated 5-year mortality prediction using deep learning and radiomics features from chest computed tomography. 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). 2017. p. 130–134.
14. Oakden-Rayner L, Carneiro G, Bessen T, Nascimento JC, Bradley AP, Palmer LJ. Precision Radiology: Predicting longevity using feature engineering and deep learning methods in a radiomics framework. Sci Rep. 2017;7(1):1648.

15. Paul R, Hawkins SH, Balagurunathan Y, et al. Deep Feature Transfer Learning in Combination with Traditional Features Predicts Survival Among Patients with Lung Adenocarcinoma. Tomography. 2016;2(4):388–395.
16. Kooi T, Litjens G, van Ginneken B, et al. Large scale deep learning for computer aided detection of mammographic lesions. Med Image Anal. 2017;35:303–312.
17. Zhao B, Tan Y, Tsai W-Y, et al. Reproducibility of radiomics for deciphering tumor phenotype with imaging. Sci Rep. 2016;6:23428.
18. Balagurunathan Y, Kumar V, Gu Y, et al. Test-retest reproducibility analysis of lung CT image features. J Digit Imaging. 2014;27(6):805–823.
19. Hu P, Wang J, Zhong H, et al. Reproducibility with repeat CT in radiomics study for rectal cancer. Oncotarget. 2016;7(44):71440–71446.

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$minimum = \min(\mathbf{X})$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2$$

$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_s} \mathbf{P}(i,j)\right)^2}{N_p}$$

$$correlation = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) ij - \mu_x \mu_y}{\sigma_x(i) \sigma_y(j)}$$

$$MAD = \frac{1}{N_p} \sum_{i=1}^{N_p} |\mathbf{X}(i) - \bar{X}|$$

$$minimum = \min(\mathbf{X})$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_s} \mathbf{P}\right.}{N_z}$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2$$

Part V

$$variance = \frac{1}{N_p}\sum_{i=1}^{N_p}(\mathbf{X}(i)-\bar{X})^2$$

$$total\ energy = V_{voxel}\sum_{i=1}^{N_p}(\mathbf{X}(i)+c)^2$$

$$SDE = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}\frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$minimum = \min(\mathbf{X})$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2$$

$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_s} \mathbf{P}(i,j)\right)^2}{N_p}$$

$$correlation = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) ij - \mu_x \mu_y}{\sigma_x(i) \sigma_y(j)}$$

$$MAD = \frac{1}{N_p} \sum_{i=1}^{N_p} |\mathbf{X}(i) - \bar{X}|$$

$$minimum = \min(\mathbf{X})$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \left(p(i) + \epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_s} \mathbf{P}\right.}{N_z}$$

$$total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2$$

Appendices

Impact Paragraph

The rise of artificial intelligence (AI) in medicine has been aided by the development of computer sciences, the prevalence of large quantities of data, and advancements in evidence-based clinical care. While prospects for AI and machine learning applications are expanding across different specialties and clinical services, radiology has led the way, with AI algorithms employed for various tasks going from scanning procedures and disease identification, prognostication, predictive biomarkers to referral systems and workflow optimization. It can be argued that AI's main objective is to deliver rapid, accurate, and cost-effective tools to help physicians make personalized decisions in much less time. The types of AI used in this thesis were: handcrafted radiomics and deep learning used separately or together. The primary goals for handcrafted radiomics were to study the influence of imaging parameter changes on the reproducibility of handcrafted radiomic features (HRFs) and to investigate its potential for discriminating between different forms of lung disease. In regard to deep learning, its potential applications for classifying different lung disorders were investigated.

Scientific impacts

1. Most of the studies in this thesis are published or under review in well-cited open access scientific journals (e.g., Cancers, Respiration, BJR, Journal of personalized medicine, Plos One, and Frontiers in Medicine), which will facilitate dissemination in academic communities. In addition, other groups world-wide will be able to reuse the methodology utilized in this thesis.

2. The experiments in Chapters 4 employed patient data to investigate the effect of different imaging phases (arterial and portal venous) on the reproducibility of HRFs. This knowledge can be reused for future studies where HRFs can be interchangeably used between arterial and portal venous phases, and these can be used to increase data points in retrospective imaging studies.

3. Chapters 3,5, and 6 are phantom investigations that aimed to improve knowledge of how changes in imaging parameters impact HRF reproducibility and how harmonization approaches, such as image resampling, Reconstruction Kernel Normalization (RKN), and ComBat harmonization, work in different contexts. Until now most of the groups, including ours, were using Combat harmonization alone we hope that this paper will convince group to use both approaches and that will lead to better results.

4. Chapters 2 and 7 cover the existing state of research, challenges, and future prospects of radiomic research and deep learning in various diseases. This knowledge dissemination may serve as a basis for future research and to write grants trying to fill knowledge gaps.

5. Chapter 8 examines the potential use of HRFs to differentiate between various interstitial lung diseases (ILDs), as well as the use of trachea volume as a novel HRF to categorize ILDs. Trachea volume is a new feature very explainable that should be used more systematically in the future chronic lung diseases.

6. In chapter 9, the potential application of HRFs and deep learning in classifying different lung disorders, including idiopathic pulmonary fibrosis (IPF), interstitial lung diseases (ILD) other than IPF subjects. This signature could be taken over by companies working of AI-based diagnostic clinical grade software. This could be particularly useful in understaffed department or areas in the world without radiologists to make a first screening of the patients needed immediate attention.

7. The combined model (ensemble learning), comprising both HRFs and deep learning, achieved the highest accuracy and precision for five-fold cross-validation and external test sets. Consequently, HRFs and deep learning models complement each other, resulting in improved performance. We hope that this combined approach will become the new standard: using several AI algorithm. The Department of Precision Medicine intend to revisit some of their published papers with this new approach.

Social impacts

1. Radiomics has the ability to speed up clinical work, reduce the workload of clinicians, and making healthcare more cost-effective.

2. Diagnostic radiomics signatures could be used in understaffed radiology department or in remote areas of the world without radiologists.

3. Diagnostic radiomics signatures could be used to support training of young radiologists.

4. The standardization of handcrafted radiomic features will aid in the generalization of radiomic signatures across institutions.

5. The development of generalizable and robust radiomic signatures will facilitate their inclusion into clinical decision-support systems.

6. Radiomics offers the potential to enhance patient care by directing personalized management rather than a one-size-fits-all approach. This can lead to less invasive methods, such as reducing the need for surgical autopsies.

7. Personalized clinical decisions are able to maximize public medical resources while lowering patient expenditures.

8. Accurate classification of interstitial lung diseases can reduce the mortality rate by allowing an earlier diagnosis for example in small center with limited experience with this rather rare diseases and aid in finding the right treatments.

Target groups

This dissertation seeks to extend and enhance our understanding of handcrafted radiomics and deep learning applied to medical imaging and potential applications. The main potential target groups are:

1. The scientists who are conducting handcrafted radiomics experiments in order to increase the awareness of the limitations associated with the field. Moreover, we anticipated that the results of our work would be useful as a reference for future researchers using handcrafted radiomics and/or deep learning.

2. The radiologists who is specializing in the thoracic imaging. The diagnosis of idiopathic pulmonary fibrosis using HRCT is a difficult task with considerable inter-observer variability even among experienced radiologists. Therefore, such methods might help the radiologist to achieve an accurate diagnosis.

3. The companies selling AI to deliver technological solutions and services for healthcare organizations and practitioners, diagnostic, and research centers.

4. The medical insurance can benefit from the use of AI and machine learning. It has the potential to detect at-risk individuals while also reducing growing healthcare expenses. In addition, the crucial aspect of a successful AI and machine learning system is its ability to develop efficient reasoning and intuitively read and understand trends.

5. Better treatment personalization will have the greatest impact on patients since they will be provided the best possible treatment to maintain a high quality of life, as well as facilitating consistent and rapid stratification of patients in drug trials.

6. The medical communities in poorer countries where thoracic imaging expertise is unavailable.

Summary

Medical imaging has the capacity to non invasively analyse the phenotypic differences of tumors in three dimensions, and lately it has seen significant improvements due to advancements in the field of artificial intelligence. For example, radiomics, or quantitative image analysis – the high-throughput extraction of quantitative features from medical images and their correlation with diagnostic and prognostic outcomes – has been studied in particular to decode tumor phenotypes from a variety of modalities, including CT, magnetic resonance imaging, and positron emission tomography (PET). Thousands of quantitative radiomic characteristics may be retrieved from each area of interest (ROI) and examined further using machine learning algorithms to look for connections with biological and clinical end objectives.

In this thesis, our objectives are; 1) to evaluate the reproducibility of radiomic features extracted from the same scanner, or from different scanners with different CT acquisition parameters ; 2) to explore how the power of AI can be harnessed for the classification between different ILDs, potentially overcoming some of the current difficulties in the decision-making surrounding lung diseases. The thesis is divided into four parts:

Part 1: General introduction and outline of the thesis.
Part 2: Challenges in handcrafted radiomics.
Part 3: Application of handcrafted radiomics and deep learning on lung disease.
Part 4: General discussion and future perspective of the thesis.

In part 1, **chapter 2** provides a literature review to assess the present state of play in handcrafted radiomics and deep learning. We presented a thorough overview and update on the rapidly increasing field of quantitative imaging research in this review, with an emphasis on the two arms “handcrafted radiomics and deep learning.” The chapter discusses some of its shortcomings as well as instances of developing clinical implementations that serve as stepping stones toward precision medicine.

In part 2, several studies have been conducted to investigate the potential of handcrafted radiomics (HRFs). Nonetheless, a number of barriers to clinical integration of radiomics signatures have been discovered. Numerous research studies have been published on the sensitivity of HRFs to inter-reader variability, test-retest, and variations in imagining parameters. In this thesis (**chapters 3-6**), we showed that HRFs are sensitive to imagine variations using phantom and patient reproducibility studies. In addition, we examined the use of different harmonization methods on reducing the effect of different variations in imagining parameters.

In chapters 3-6, we assess the reproducibility of HRFs to the variations in CT parameters and the role of harmonization methods to address those variations. **Chapter 3** investigated the robustness of HRFs on a dataset consisting of 13 phantom CT scans. The scans were obtained from different vendors, with different CT parameters. The study's findings indicated that only a small percentage of handcrafted (HRFs) radiomics were robust to differences in the imaging settings examined. We also found that the performance of ComBat harmonization depends on the variations in imaging parameters.

Chapter 4 assess the reproducibility of hepatocellular carcinoma (HCC) HRFs, generated from various phases of contrast-enhanced CT images (CECT). For this study, HCC patients' arterial and venous CT scans were made accessible. The finding of the presented study showed that, when no image settings were changed, a subset of HRFs were shown to be reproducible in both phases. Moreover, the application of ComBat harmonization increased the number of reproducible features by 1% across phases.

In chapter 5, we investigated the use of Reconstruction Kernel Normalization (RKN) and ComBat harmonization to improve the reproducibility of HRFs across scans acquired with different reconstruction kernels. A total of 28 phantom scans collected on five distinct scanners types were assessed. The HRFs were extracted from the original scans and scans that were harmonized using the RKN method. Moreover, ComBat harmonization method was applied on both set of HRFs. The finding of this study showed that the majority of HRFs were found to be sensitive to the variations in the reconstruction kernels. Furthermore, the use of both RKN and ComBat harmonization methods significantly increased the number of reproducible HRFs compared to HRFs extracted from original scans.

In chapter 6, we also investigated the impact of changes in the in-plane spatial resolution (IPR) on the reproducibility of HRFs extracted from phantom scans (n=14) while all other imaging parameters were the same. We also examine the impact of ComBat harmonization on HRFs. The finding of this study revealed that the reproducibly of HRFs depends on the degree of the variations in pixel spacing.

Part 3 in this thesis is related to the application of radiomics and deep learning in different lung disorders. **In chapter 7,** we presented a summary of the existing researches on the use of handcrafted radiomics in lung cancer diagnosis, treatment response, and prognosis. In addition, applying HRFs in chronic obstructive pulmonary disease (COPD) has not been extensively investigated yet. We show examples of the potential use of HRFs in the diagnosis, treatment, and follow-up of COPD and future direction.

In chapter 8, the approach of HRFs was studied in order to predict different interstitial lung diseases (ILDs). The data for this study came from one center and two databases. The study

comprised four groups: 1) IPF with UIP pattern on HRCT, 2) IPF with UIP pattern confirmed by surgical lung biopsy, 3) non-IPF ILDs with surgical lung biopsy confirming the absence of a UIP pattern, and 4) healthy lung patients. To summarize, we were able to show that radiomic characteristics generated from HRCT images may be utilized to differentiate between a normal state and ILDs, as well as between IPF with a UIP pattern and ILDs with no UIP pattern as confirmed by surgical biopsy. Furthermore, our study found a significant difference in tracheal volume between individuals with normal, IPF/UIP, and non-IPF ILDs. The trachea volume was shown to be larger in IPF participants compared to normal and non-IPF ILDs.

In chapter 9, the use of both HRFs and DL was explored in this thesis to differentiate between different lung disorders – namely, IPF, and non-IPF ILDs subjects. In addition, in order to interpret the performance of HRFs and DL, interpretability methods were used. We also made use of ensemble learning methods to improve the performance of both HRFs and DL. In silico clinical trials were also used to compare the performance of medical experts with AI. Our results showcased the utility of HRFs and DL algorithms as a tool to support clinical decisions.

Finally, in part 4 (chapter 10) we extensively discussed the results of this thesis and the future perspective of both HRFs and deep learning.

Overall, this thesis verified a number of hypotheses concerning the uses of handcrafted radiomics and deep learning in medical image analysis. For handcrafted radiomics, we assessed the robustness of handcrafted radiomics analyses, which will aid in the development of generalizable radiomics signatures, and provided unique quantitative methods to measure the reproducibility of HRFs among scans obtained differently. For deep learning, we evaluated and demonstrated the potential of automated algorithms to improve clinical decision making. More specifically, a deep learning algorithm was developed that performed very well and has the potential to be used in clinical settings.

List of publications

1. **Refaee, T**., Wu, G., Ibrahim, A., Halilaj, I., Leijenaar, R.T.H., Rogers, W., Gietema, H.A., Hendriks, L.E.L., Lambin, P. and Woodruff, H.C., 2020. The Emerging Role of Radiomics in COPD and Lung Cancer. *Respiration; International Review of Thoracic Diseases*, *99*(2), pp.99-107.
2. Ibrahim, A*., Widaatalla, Y*., **Refaee, T***., Primakov, S., Miclea, R.L., Öcal, O., Fabritius, M.P., Ingrisch, M., Ricke, J. and Hustinx, R., 2018. Reproducibility of CT-Based Hepatocellular Carcinoma Radiomic Features across Different Contrast Imaging Phases: A Proof of Concept on SORAMIC Trial Data. *Imaging*, *6*, pp.379-391.
3. **Refaee Turkey***, Benjamin Bondue*, Gaetan Van Simaeys, Guangyao Wu, Chenggong Yan, Henry C. Woodruff, Serge Goldman, and Philippe Lambin. 2022. "A Handcrafted Radiomics-Based Model for the Diagnosis of Usual Interstitial Pneumonia in Patients with Idiopathic Pulmonary Fibrosis" Journal of Personalized Medicine 12, no. 3: 373.
4. **Turkey Refaee**, Yousif Widaatalla, Razvan L. Miclea, Spencer Behr, Sergey Primakov, Elizaveta Lavrova, Henry C Woodruff, Tom Marcelissen, Olivier Morin, Abdalla Ibrahim†, Philippe Lambin†, 2022. Effects of variations in imaging phase on CT-based renal cyst's handcrafted radiomic features: A reproducibility study. [in preparation].
5. Abdalla Ibrahim, Bruno Barufaldi*, **Turkey Refaee***, Telmo M. Silva Filho, Raymond J. Acciavatti, Zohaib Salahuddin, Roland Hustinx, Felix M. Mottaghy, Andrew D.A. Maidment‡ and Philippe Lambin‡, 2022. MaasPenn radiomics reproducibility score: a novel quantitative measure for evaluating the reproducibility of CT-based handcrafted radiomic features. *Cancers 2022, 14, 1599.*
6. **Turkey Refaee*** , Zohaib Salahuddin* , Anne Noelle Frix , Chenggong Yan, Guangyao Wu, Henry C Woodruff, Hester Gietema , Paul Meunier , Renaud Louis , Julien Guiot** and Philippe Lambin**, 2022. Diagnosis of Idiopathic Pulmonary Fibrosis in HRCT Scans using a combination of Handcrafted Radiomics and Deep Learning. *Frontiers in Medicine,[Submitted].*
7. **Refaee, T**.; Salahuddin, Z.; Widaatalla, Y.; Primakov, S.; Woodruff, H.C.; Hustinx, R.; Mottaghy, F.M.; Ibrahim, A.; Lambin, P. CT Reconstruction Kernels and the Effect of Pre- and Post-Processing on the Reproducibility of Handcrafted Radiomic Features. J. Pers. Med. 2022, 12, 553
8. Ibrahim, A., Primakov, S., Beuque, M., Woodruff, H.C., Halilaj, I., Wu, G., **Refaee, T**., Granzier, R., Widaatalla, Y., Hustinx, R. and Mottaghy, F.M., 2021. Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework. *Methods*, *188*, pp.20-29.
9. Rogers, W., Thulasi Seetha, S., **Refaee, T.A**., Lieverse, R.I., Granzier, R.W., Ibrahim, A., Keek, S.A., Sanduleanu, S., Primakov, S.P., Beuque, M.P. and Marcus, D., 2020. Radiomics: from qualitative to quantitative imaging. *The British journal of radiology*, *93*(1108), p.20190948.

10. Wu, G., Woodruff, H.C., Shen, J., **Refaee, T**., Sanduleanu, S., Ibrahim, A., Leijenaar, R.T., Wang, R., Xiong, J., Bian, J. and Wu, J., 2020. Diagnosis of invasive lung adenocarcinoma based on chest CT radiomic features of part-solid pulmonary nodules: a multicenter study. *Radiology*, *297*(2), pp.451-458.
11. Ibrahim, A., **Refaee, T**., Primakov, S., Barufaldi, B., Acciavatti, R.J., Granzier, R.W.Y., Hustinx, R., Mottaghy, F.M., Woodruff, H.C., Wildberger, J.E. and Lambin, P., 2021. The Effects of In-Plane Spatial Resolution on CT-Based Radiomic Features' Stability with and without ComBat Harmonization. *Cancers*, *13*(8).
12. Wu, G., Jochems, A*., **Refaee, T***., Ibrahim, A., Yan, C., Sanduleanu, S., Woodruff, H.C. and Lambin, P., 2021. Structural and functional radiomics for lung cancer. *European Journal of Nuclear Medicine and Molecular Imaging*, *48*(12), pp.3961-3974.
13. Kalmet, P.H., Sanduleanu, S., Primakov, S., Wu, G., Jochems, A., **Refaee, T**., Ibrahim, A., Hulst, L.V., Lambin, P. and Poeze, M., 2020. Deep learning in fracture detection: a narrative review. *Acta orthopaedica*, *91*(2), pp.215-220.
14. Walsh, S., de Jong, E.E., van Timmeren, J.E., Ibrahim, A., Compter, I., Peerlings, J., Sanduleanu, S., **Refaee, T**., Keek, S., Larue, R.T. and van Wijk, Y., 2019. Decision support systems in oncology. *JCO clinical cancer informatics*, *3*, pp.1-9.
15. Ibrahim, A., Vallières, M., Woodruff, H., Primakov, S., Beheshti, M., Keek, S., **Refaee, T**., Sanduleanu, S., Walsh, S., Morin, O. and Lambin, P., 2019, June. Radiomics Analysis for Clinical Decision Support in Nuclear Medicine. In *Seminars in Nuclear Medicine* (Vol. 49, No. 5, pp. 438-449).
16. Wu, G., Woodruff, H.C., Sebastian, S., **Turkey, R**., Arthur, J., Ralph, L., Hester, G., Shen, J., Wang, R., Jingtong, X. and Jie, B., 2020. Preoperative CT-based radiomics combined with intraoperative frozen section is predictive of invasive adenocarcinoma in pulmonary nodules: a multicenter study. *European Radiology*, *30*(5), pp.2680-2691.
17. Ibrahim, A., **Refaee, T**., Leijenaar, R.T., Primakov, S., Hustinx, R., Mottaghy, F.M., Woodruff, H.C., Maidment, A.D. and Lambin, P., 2021. The application of a workflow integrating the variable reproducibility and harmonizability of radiomic features on a phantom dataset. *Plos one*, *16*(5), p.e0251147.
18. Frix, A.N., Cousin, F., **Refaee, T**., Bottari, F., Vaidyanathan, A., Desir, C., Vos, W., Walsh, S., Occhipinti, M., Lovinfosse, P. and Leijenaar, R.T.H., 2021. Radiomics in Lung Diseases Imaging: State-of-the-Art for Clinicians. *Journal of Personalized Medicine*, *11*(7).

Acknowledgements

Throughout my PhD journey, I have been fortunate to have a lot of help and support. As a result, I would like to convey my heartfelt appreciation to everyone who contributed to the successful completion of my thesis.

First and foremost, I would like to thank **Almighty Allah** for providing me with the opportunity, motivation and strength to complete this journey. Also, for surrounding me with beautiful individuals from the time I committed to this adventure, to the moment I was awarded the degree. This accomplishment may not have been possible if not for his kindness.

To my promotor, **Professor Philippe Lambin,** who has been a constant source of support and inspiration to me. I would like to express my deep gratitude to **Prof. Lambin,** for his continuous and enthusiastic encouragement, constructive critiques, and patient guidance throughout my research work. The willingness to give his time so generously has been very much appreciated. Thank you genuinely for your tremendous effort, the opportunities you opened up, and the direction you gave me over the years, all of which culminated in the accomplishment of this work.

To my co-promotor, Dr. **Julien Guiot**, thank you for your support. I am very grateful for your help with my final project. Your ideas and insights made it much easier to proceed in the right direction and complete the project successfully.

To Dr. **Benjamin Bondue**, it was a great privilege working with you. Thank you for all the meetings to discuss how to proceed with our project. We finally got it published. I hope very much that we collaborate again in the future.

I am grateful to Jazan University for funding my PhD studies. A particular thanks to the university's president, **Dr. Mar'ei bin Hussein Al-Qahtani**, for his unwavering support. My deepest gratitude also goes to **Dr. Aymen Madkhali**, Dean of the Faculty of Applied Medical Sciences, for his constant support and encouragement. My sincere thanks also go to **Dr. Nasser Shubayr**, my brother, friend, colleague, and chairman of the Medical Research Center, for his encouragement and endless support.

I would like to thank my PhD colleagues and office members: **Guangyao**, **Yvonka**, **Simon**, **Sergey**, **Manon**, **Iva**, **Liza**, **Shruti**, **Monideepa**, **William**, **Fadila**, **Relinde**, **Renee**, **Yi**, **Xian**, **Ralph**, **Evelyn**, **Janita**, **Aniek**, **Ruben**, **Jurgen**, **Timo**, **Mark**, **Joes**, **Cecile**, **Brent**, **Rianne**, **Floor**, **Cary**, and **Henry**, and anybody else I may have forgotten.

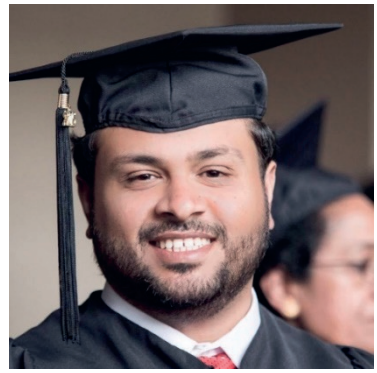
To my family, for being my greatest source of support and motivation in all aspects of my life. My mother, **Shareefa**, your unwavering support means so much to me that I do not know how to repay you. My words and gestures of gratitude will never be adequate to express my appreciation for the part you have played all along. My father, **Abdullah**, if not for you, I would not be where I am today. You have always shown me great care and have supported me like a pillar through everything. To my wife, **Samaher**, saying thank you is not enough. I am grateful for the sacrifices you have made for me. I could not have done any of this without your incredible support, drive, love, and encouragement. To my sweet daughter, **Rogeen**, having you in my life has been the greatest joy I have ever experienced. It is a blessing for which I am eternally grateful to Allah. You are the spirit image of me. I could never ask for a better child and I believe you will also become my dearest friend. I love you Rogeen, may Allah always light your way to happiness and success. Thank you also to my dear brothers and sisters, **Amena**, **Mohammed**, **Najah**, **Roqaia**, **Fatima**, **Khaled**, **Faisal**, **Kholod**, and **Abdullelah**, for your continuous support and encouragement throughout. A special thank you to my aunt **Fatima** for constantly checking in on me during my PhD journey.

A special thank you to my other family who I met in the Netherlands. To my brother from another mother **Abdalla Khalil**, saying thank you does not seem enough after your continued support during this adventure. You have been on my side from the moment we met. Thank you for all the help and support during my PhD. Your intellectual discipline and process of actively and skillfully conceptualizing and analyzing information never failed to amaze me. Abdalla, I will treasure your friendship until the end. To **Yousif**, another brother from another mother. Thank you for the amazing time we spent together. I never knew how much Sudanese culture had in common with my own (Jazani culture) until I met you. I will not say goodbye, but see you soon (probably in Jazan). To **Zohaib**, the polite and knowledgeable man. Thank you for the great time when we worked together. Your coding skills are remarkable. I am sure we will continue to work together in the future. So, until we meet again, brother. To my big brother, **Mohammed Al-Ahmari**. I really do not know where to start, so I will keep it brief. I remember the first time I met you in 2016, when you picked me up to help another friend. At that moment, I knew we would be good friends. Thank you for the constant support and encouragement. I am grateful for the time we talked, laughed, ate and travelled. Just to let you know, I will visit you once I get back to our country. I promise you that. To my younger brother, **Faris Al-Musabi.** Every time I visit you in the lab, I learn more about the brain and how it works. I also learned a lot from you about dopamine and serotonin and their functions. Thank you for the support and encouragement. To my big brother **Majed Aldehri**. Thank you for gathering us every weekend to eat dinner and talk at your place.

I would also like to thank my Saudi friends here in Maastricht for their continuous support: **Mohammed Jafer, Nader Kameli, Sultan Mashnafi, Abdullah Aseeri, Faisal Klufah, Ghaleb**

Moubarki, Hassan Moafa, Jehad Al-Tyyb, Jobran Meshi, Ghazi Al-Jouf, Faisal Al-Osaimi, Othman Al-shihri, and Khaled Al-Ameer. I also would like to thank my brother and dearest friend **Hussain Khairi** for his continuous support and encouragement.

Curriculum Vitae



Turkey Refaee is from Jazan city in the Kingdom of Saudi Arabia. He completed his 5 years bachelor degree on diagnostic Radiology at King Khaled University, Abha. After that, he started his job as teacher assistant at the department of diagnostic radiology at Jazan University. In 2011, he received a scholarship from Jazan University to purse his master's study in the United State. In 2014, he passed the board exam of the American Registry of Radiologic Technologist (ARRT) in Nuclear Medicine Technology. In 2015, he completed his master's degree in Positron Emission Tomography and Computed Tomography (PET/CT) at Thomas Jefferson University in Philadelphia, PA. Afterward, he moved back to Saudi Arabia and started working as a lecturer at Jazan University in the department of Diagnostic Radiology in the faculty of applied medical sciences. In 2016, he carried out his PhD trajectory in the department of Precision Medicine within the School of Oncology and Developmental Biology (GROW) at the faculty of Health, Medicine, and Life Science at Maastricht University, the Netherlands. He worked under the supervision of Prof. Philippe Lambin. For the project entitled "Quantitative Imaging Analysis: Challenges and Potentials".

نبذه مختصرة عن الرسالة

التصوير الطبي الإشعاعي لديه القدرة على تحليل الاختلافات النمطية للأورام في صورة ثلاثية الأبعاد، والتي بدورها شهدت تقدما ملحوظا بفضل التطور في مجال الذكاء الاصطناعي. من الأمثلة على ذلك: علم الأشعة المهتم بتحليل البيانات الكمية للصورة, الذي يربط بين الخصائص الكمية الهائلة الممكن استخراجها من صور الأشعه وما يمكن رؤيته أو تشخيصه من تغيرات غير طبيعية طارئة على العضو المراد تصويره. بل تجاوز الأمر ذلك إلى أن هذا العلم استطاع تحليل البيانات الكمية للصوره ومن ثم توقع نوع المرض الذي سيصاب به المريض قبل أن يرى بالعين المجردة من قبل الطبيب. اهتم هذا المجال من علم الأشعة تحديدا بدراسة وتحليل أنماط أورام عديدة باستخدام صور الأشعه المقطعية والمغناطيسية والأشعة المقطعيه البوزيترونية، بعد تحديد جزء الصورة المراد تحليله, يمكن استراج آلاف الخصائص النمطية لهذا الجزء تحديدا باستخدام خوارزميات تعلم الآلة وربطها مع التغيرات البيولوجية والاكلينيكية.

قامت العديد من الدراسات بقياس مدى إمكانية تطبيق علم (الريديومكس) و الاستفادة منه في التشخيص المبكر للأورام. هذه الدراسات بينت أن هناك عدد من العوائق التي تحول دون التطبيقات الإكلينيكية لهذا العلم. هذه العوائق تتمثل في حساسية ال (الريديومكس) مع وجود بعض العوامل كاختلاف نوع القراء ، وإعادة الاختبار ، والاختلافات في عوامل التصوير الإشعاعي. وباستخدام بيانات مبنية على دراسات قائمة على مرضى وأخرى قائمة على دمى تحاكي البشر (فانتومز).

أثبتت فصول هذه الأطروحة (من الفصل الثالث إلى السادس) أن ال (الريديومكس) يتأثر بمجرد تغير العوامل في التصوير الإشعاعي. بالإضافة إلى ذلك، قامت هذه الأطروحة باختبار استخدام طرق الإنسجام و الاندماج في علم تعلم الآلة والذكاء الاصطناعي للتقليل من تأثير هذه الاختلاف في عوامل التصوير الإشعاعي.

إضافة إلى ما سبق، قدمت هذه الأطروحة مراجعة منهجية مفصلة لما وصل إليه التعلم العميق بالنسبة ل (الريديومكس) وتطبيقاتها المحتملة. كذلك تم تقديم دراسة خاصة في هذه الأطروحة حول مجال ال (الريديومكس) من أجل التنبؤ بإحتمالية الإصابة بأمراض أنسجة الرئة من عدمها. قامت هذه الأطروحة بتقديم فصل آخر يهتم بدراسة استخدام حجم القصبة الهوائية كعامل تنبؤ للتفريق بين المرضى المصابين بال (التليف الرئوي مجهول السبب) مقابل المرضى المصابين ب (أمراض الرئة الخلالية) و الأشخاص السليمين.

كما قامت هذه الأطروحة كذلك بالإفادة من الطرق المعتمدة في إثبات فهم النتائج من أجل التثبت مما توصلنا إليه من نتائج باستخدام ال (الريديومكس) والتعلم العميق. ومن أجل تحسين أداء ال (الريديومكس) والتعلم العميق في إظهار نتائج أفضل. تم كذلك الإفاده مما يسمى بالتعلم (انسيمل أو خاصية الجمع). وأخيرا, قامت هذه الأطروحة بعمل ما يسمى بالتجارب السريرية الافتراضية لمقارنة أداء الأطباء أصحاب الخبرة في قراءة صور أشعة الرئة مع أداء خوارزميات الذكاء الاصطناعي والتي أثبتت بأن كلا من (الريديومكس) والتعلم العميق يمكن استخدامهما كأداة مساعدة للأطباء في اتخاذ تشخيصاتهم وقراراتهم الخاصة بالمريض.

بشكل عام, هذه الرسالة أكدت عددا من الفرضيات المتعلقة باستخدام ما يسمى بالريديومكس والتعلم العميق في تحليل صور الأشعة. أثبتنا كذلك متانة هذا العلم وأنه يمكن الاستمرار في تحسين ما توصلنا إليه من استخدام للخوارزميات الآلية والتعلم العميق لتكون أداة أساسية يستفيد منها الأطباء في تحليل وتشخيص صور الأشعه في وقت وجيز وبمستوى عال من الدقة لا تقل عن تلك التي يقوم بها خبراء أطباء الأشعة.

شكر وعرفان

سبحانك اللهم خير معلم... علمت بالقلم القرون الأولى
أخرجت هذا العقل من ظلماته ... وهديته النهج القويم سبيلا

الحمد لله كما ينبغي لجلال وجهه وعظيم سلطانه. لك ربي الثناء والفضل على توفيقك بإتمام هذه الرسالة، فلولاك ربي لم تر النور

إلى أعظم شخصين في حياتي أمي وأبي، أنتم من يستحق الشكر بعد الله سبحانه وتعالى، فدعمكم المتواصل ودعاءكم كان من أهم أسباب توفيقي. شكرا لكما بحجم الكون

إلى زوجتي ورفيقة دربي، شكرا من القلب لمرافقتك إياي في بلد الغربة، ودعمك المتواصل طيلة سنوات الدراسة

إلى نبض فؤادي وشمعة حياتي بنتي (روجين)، كم مريتُ بأيام صعبة في دراستي وكم عانيتُ من الهم ، لكن رؤيتك كانت تزيل عني كل هم وتعب. أنت نعمة من الله. أسال الله لك الصلاح

إلى إخواني وأخواتي، أنتم أكبر داعم في مسيرة حياتي. شكرا لكم جميعا على وقوفكم معي ومساندتي

إلى وطني الغالي، لك وافر الشكر على الدعم المتواصل طيلة دراستي

إلى عميد كلية العلوم الطبية التطبيقية، شكرا بحجم السماء على دعمكم

إلى رئيس قسمي، شكرا من أعماق قلبي على كل مساعدة قدمتها لي
إلى كل صديق وحبيب ، شكرا على الدعم والسؤال

شكرا شكرا شكرا من أعماق قلبي

تحليل التصوير الكمي: التحديات والإمكانيات

تركي رفاعي

$$correlation = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)ij - \mu_x\mu_y}{\sigma_x(i)\sigma_y(j)}$$

$$entropy = -\sum_{i=1}^{N_g} p(i)\log_2\left(p(i)+\epsilon\right)$$

$$SDE = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}\frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

$$total\ energy = V_{voxel}\sum_{i=1}^{N_p}(\mathbf{X}(i)+c)^2$$

$$variance = \frac{1}{N_p}\sum_{i=1}^{N_p}(\mathbf{X}(i)-\bar{X})^2$$