The Image Biomarker Standardization Initiative: Standardized Convolutional Filters for Reproducible Radiomics and Enhanced Clinical Insights

Philip Whybra*, Alex Zwanenburg*, Vincent Andrearczyk, Roger Schaer, Aditya P Apte, Alexandre Ayotte, Bhakti Baheti, Spyridon Bakas, Andrea Bettinelli, Ronald Boellaard, Luca Boldrini, Irène Buvat, Gary J R Cook, Florian Dietsche, Nicola Dinapoli, Hubert S Gabryś, Vicky Goh, Matthias Guckenberger, Mathieu Hatt, Mahdi Hosseinzadeh, Aditi Iyer, Jacopo Lenkowicz, Mahdi A L Loutfi, Steffen Löck, Francesca Marturano, Olivier Morin, Christophe Nioche, Fanny Orlhac, Sarthak Pati, Arman Rahmim, Seyed Masoud Rezaeijo, Christopher G Rookyard, Mohammad R Salmanpour, Andreas Schindele, Isaac Shiri, Emiliano Spezi, Stephanie Tanadini-Lang, Florent Tixier, Taman Upadhaya, Vincenzo Valentini, Joost J M van Griethuysen, Fereshteh Yousefirizi, Habib Zaidi, Henning Müller, Martin Vallières, Adrien Depeursinge

* P.W. and A.Z. contributed equally to this work.

From the School of Engineering, Cardiff University, Cardiff, United Kingdom (P.W., E.S.); OncoRay – National Center for Radiation Research in Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Helmholtz-Zentrum Dresden - Rossendorf, Dresden, Germany (A.Z., S.L.); National Center for Tumor Diseases (NCT), Partner Site Dresden, Germany: German Cancer Research Center (DKFZ), Heidelberg, Germany, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany, and Helmholtz Association / Helmholtz-Zentrum Dresden - Rossendorf (HZDR), Dresden, Germany (A.Z.); Institute of Informatics, University of Applied Sciences and Arts Western Switzerland (HES-SO), Sierre, Switzerland (V.A., R.S., H.M., A.D.); Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY, USA (A.P.A., A.I.); Department of Computer Science, Université de Sherbrooke, Sherbrooke, QC, Canada (A.A., M.A.L.L., M.V.); Center for Artificial Intelligence and Data Science for Integrated Diagnostics (AI2D) and Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA, USA (B.B., S.B., S.P.); Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA (B.B., S.B., S.P.); Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA (B.B., S.B., S.P.); Division of Computational Pathology, Department of Pathology and Laboratory Medicine, Indiana University School of Medicine, Indianapolis, IN, USA (B.B., S.B., S.P.); Medical Physics Department, Veneto Institute of Oncology IOV - IRCCS, Padua, Italy (A.B., F.M.); Radiology and Nuclear Medicine, Amsterdam UMC, Amsterdam, the Netherlands (R.B.); Fondazione Policlinico Universitario “A. Gemelli” IRCCS, Rome, Italy (L.B., N.D., J.L.); Institut Curie, Université PSL, Inserm U1288, Laboratoire d’Imagerie Translationnelle en Oncologie, Orsay, France (I.B., C.N., F.O.); Cancer Imaging, School of Biomedical Engineering and Imaging Sciences, King's College London, London, United Kingdom (G.J.R.C., V.G., C.G.R.); Department of Radiation Oncology, University Hospital Zurich, University of Zurich, Zurich, Switzerland (F.D., H.S.G., M.G., S.T.); Department of Radiology, Guy's & St Thomas' NHS Foundation Trust, London, United Kingdom (V.G.); LaTIM, INSERM, UMR 1101, Université de Bretagne-Occidentale,

Brest, France (M.Ha., F.T.); Technological Virtual Collaboration (TECVICO Corp.), Vancouver, BC, Canada (M.Ho., M.R.S.); Department of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran (M.Ho.); Department of Radiation Oncology, University of California San Francisco, San Francisco, CA, USA (O.M., T.U.); Departments of Radiology and Physics, University of British Columbia, Vancouver, BC, Canada (A.R.); Department of Medical Physics, Faculty of Medicine, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran. (S.M.R.); Repository Unit, Cancer Research UK National Cancer Imaging Translational Accelerator, United Kingdom (C.G.R.); Department of Integrative Oncology, BC Cancer Research Institute, Vancouver, BC, Canada (M.R.S., F.Y.); Department of Nuclear Medicine, Universitätsklinikum Augsburg, Augsburg, Germany (A.S.); Division of Nuclear Medicine and Molecular Imaging, Geneva University Hospital, Geneva, Switzerland (I.S., H.Z.); Department of Cardiology, Inselspital, Bern University Hospital, University of Bern, Switzerland (I.S.); Dipartimento Radiodiagnostica, Radioterapia ed Ematologia, Fondazione Policlinico Universitario “A. Gemelli” IRCCS, Rome, Italy (V.V.); Professore ordinario di Radioterapia, Università Cattolica del Sacro Cuore - Milano, Milan, Italy (V.V.); Department of Radiology, The Netherlands Cancer Institute, Amsterdam, the Netherlands (J.J.M.v.G.); Department of Radiology, UMC Utrecht, Utrecht, the Netherlands (J.J.M.v.G.); Centre de recherche du Centre hospitalier universitaire de Sherbrooke (CHUS), Sherbrooke, QC, Canada (M.V.); and Department of Nuclear Medicine and Molecular Imaging, Lausanne University Hospital (CHUV), Lausanne, Switzerland (A.D.).

Summary Statement

Standardizing convolutional filters that enhance specific structures and patterns in medical imaging enables reproducible radiomics analyses, improving consistency and reliability for enhanced clinical insights.

Essentials

- Fifteen international teams that developed radiomics software defined and standardized eight convolutional filter types for radiomic analyses: mean, Laplacian-of-Gaussian, Laws and Gabor kernels, separable and non-separable wavelets (undecomposed, decomposed forms).
- Thirty-three reference filtered images and 323 reference feature values computed from filtered images were established to standardize radiomics analyses across various imaging modalities.
- A web-based tool is available for checking compliance of radiomics software.

Abstract

Filters are commonly used to enhance specific structures and patterns in images, such as vessels or peritumoral regions, to enable clinical insights beyond the visible image using radiomics. However, their current lack of standardization hampers reproducibility and clinical translation of radiomics decision support tools. Here, teams of researchers that developed radiomics software participated in a three-phase study (September 2020 to December 2022) to establish a standardized set of filters. The first two phases focused on finding reference filtered images and reference feature values for (commonly used) convolutional filters: mean, Laplacian-of-Gaussian, Laws and Gabor kernels, separable and non-separable wavelets (including decomposed forms), and Riesz transformations. In the first phase, 15 teams used digital phantoms to establish 33 reference filtered images out of 36 filter configurations. Then, 11 teams used a chest CT to derive reference values for 323 of 396 features computed from filtered images using 22 filter and image processing configurations. Reference filtered images and feature values for Riesz transformations were not established. Afterwards, reproducibility of standardized convolutional filters was validated on a public dataset of multi-modal imaging (CT, fluorodeoxyglucose-PET, T1-weighted MRI) from 51 patients with soft-tissue sarcoma. During validation, reproducibility of 486 features computed from filtered images using 27 filter and image processing configurations was assessed using the lower bounds of 95% confidence intervals of intraclass correlation coefficients (ICC-CI-low). 458 of 486 features were reproducible across 9 teams with ICC-CI-low>0.75. In conclusion, eight filter types were standardized with reference filtered images and reference feature values for verifying and calibrating radiomics software packages. A web-based tool is available for compliance-checking.

Introduction

Radiomics involves the high-throughput extraction of quantitative features from medical images to support clinical decision making (1,2). Relatively few radiomics decision support tools have entered the clinic, as their clinical translation is hampered by both the lack of standardization of the extraction process and by lack of quality clinical evidence for their efficacy (3). Focusing on software-related aspects of the extraction process, the Image Biomarker Standardization Initiative (IBSI) previously established modality-independent standards for digital image processing and computation of handcrafted, quantitative radiomic features (4). This improved reproducibility and interchangeability of IBSI-compliant radiomics software packages, provided that the extraction process is configured the same between packages (5,6).

Filters are frequently used in radiomics analyses to enhance and quantify potentially clinically relevant characteristics and textures in medical images, such as the peritumoral region, blood vessels, contrast agent uptake, degree of calcification, or fibrosis, among others (7) (Supplementary Note 1). For instance, Beuque et al. applied a Laplacian-of-Gaussian filter to contrast-enhanced mammography to classify breast lesions into benign and malignant cases (8). The Laplacian-of-Gaussian filter enhanced the regions with contrast uptake, amplifying the signal, and therefore was found to be highly important for classifying lesion malignancy. Many filters, including the Laplacian-of-Gaussian filter used by Beuque et al., rely on convolution. Convolution is a mathematical operation, where a filter (here an array of numbers) is systematically slid across the entire image, see Figure 1. This process yields a filtered image that enhances and spatially locates potentially relevant image characteristics such as those mentioned above. However, the computational implementation of these filters has not been standardized, and quantitative features extracted from regions of interest in the filtered images were found to be poorly reproducible between radiomics software packages (9), see Figure 2 for some examples. Consequently, radiomics decision support tools that incorporate features computed from regions of interest inside filtered images may be difficult to reproduce, validate and translate to the bedside.

Because convolutional filters are both important and commonly used, the IBSI aimed to improve reproducibility of radiomics decision support tools involving these filters and to facilitate their clinical translation through a modality-independent software standardization process, by: *(a)* establishing definitions for convolutional filters, including commonly used ones such as wavelets and Laplacian-of-Gaussian filters; *(b)* integrating convolutional filters into the previously established general radiomics image processing scheme (4); and *(c)* providing datasets, associated reference filtered images and reference feature values, as well as tools for verification and calibration of radiomics software packages.

Materials and Methods

Study design

This standardization effort was divided into three phases (Figure 3) and was conducted between September 2020 and December 2022. During the first two phases the

implementation and use of convolutional filters were standardized. Phase 1 concerned the creation of reference filtered images, i.e., the expected result of applying a convolutional filter with specific parameters to an image. In phase 2, convolutional filters were integrated into a radiomics workflow for the purpose of finding reference values for radiomic features computed from filtered images. In phase 3, we assessed whether standardization of convolutional filters resulted in reproducible feature values. A website (https://ibsi.radiomics.hevs.ch/; Supplementary Note 2) was created to coordinate the study.

Convolutional filters

Convolutional filters transform an image to a filtered image by convolution. These filters consist of numerical weights that are pre-defined or parameterized in the spatial domain or in the frequency (Fourier) domain. Several convolutional filters were assessed, i.e., mean filter, Laplacian-of-Gaussian filter, Laws kernels, Gabor kernels, separable and non-separable wavelets, and Riesz transformations of convolutional filters, see Figure 1. Further details are supplied in Supplementary Note 1 and in the reference manual (10).

Participating teams

Teams of radiomics researchers were invited to participate in this study. In addition to all teams that had previously participated in the IBSI (4), invitations were extended to any other team that indicated their desire to participate, e.g., through the main IBSI website (https://theibsi.github.io/) and through personal communication. Participation was voluntary and open for the duration of the study. Teams were eligible to participate if they *(a)* developed their own radiomics software, and *(b)* their software was compliant with the previous IBSI reference standard. Teams were not required to participate in all phases of the study.

Phase 1: Establishing reference filtered images

In phase 1, five digital three-dimensional phantoms were used (Supplementary Note 3), namely: 1) an orientation phantom to verify consistency of image orientation within the software of each team; 2) an impulse phantom with a single, central, active voxel; 3) a sphere phantom consisting of concentric spherical shells; 4) a phantom with a checkerboard pattern; and 5) a phantom with line patterns. Thirty-six convolutional filter configurations were defined to establish reference filtered images (Supplementary Note 4). Teams computed filtered images for each filter configuration and uploaded these to the study website.

The level of consensus for each filtered image was assessed using the same metrics as previously (4): *(a)* by the number of teams that matched the tentative reference filtered image (Supplementary Note 5), i.e. had filtered images with voxel-wise differences with the tentative reference filtered image that were less than 1% of the intensity range of the tentative reference filtered image for all voxels; and (*b*) the previous number divided by the number of teams that contributed a filtered image. Level of consensus was then: *none*, if the tentative reference filtered image was not produced by over 50% of contributing teams;

weak, match between fewer than three teams; *moderate*, three to five; *strong*, six to nine; *very strong*, ten or more.

Phase 2: Defining feature reference values

Convolutional filtering was integrated into the general radiomics image processing scheme (Figure 1). Image processing and convolutional filter configurations were then defined for each filter. Both 2D and 3D filter configurations were created, yielding twenty-two configurations in total (Supplementary Note 4). Teams computed a filtered image for each configuration from a publicly available chest CT image of a patient with lung cancer (11). Eighteen intensity-based features were computed from the gross tumor volume region of interest in each filtered image (Supplementary Note 6). Thus, a total of 396 features could be computed (eighteen features times twenty-two configurations). After computing feature values, teams uploaded their results to the study website. The level of consensus for feature values was assessed using the same metrics as in phase 1 by using contributed values for each feature as input and comparing matches within a tolerance margin (Supplementary Note 6).

Phase 3: Validation

After completing phases 1 and 2, teams were asked to compute intensity-based features from the gross tumor volume segmentation in filtered images of a multimodality imaging cohort (co-registered CT, fluorine 18 fluorodeoxyglucose PET, and T1-weighted MRI). This cohort consisted of 51 patients with soft-tissue sarcoma obtained from The Cancer Imaging Archive (12–14). PET and MRI were pre-processed to ensure that conversion of PET activity concentration to standardized uptake value and MR bias field correction and normalization could not affect validation results (Supplementary Note 4). Nine image processing and convolutional filter configurations were specified for each modality. Thus, a total of 486 features (eighteen features times nine configurations times three image modalities) could be computed. Teams were blinded to the results submitted by other teams. After submitting results, obvious configuration errors were reported back to the submitting team.

Statistical Analysis

Reproducibility of each of the 486 features computed in the validation phase was assessed using two-way random effects single-rater intraclass correlation coefficient (ICC) for absolute agreement between teams (15). Based on Koo and Li (16), reproducibility of each feature was assigned to one of the following categories, based on the lower bound of the 95% confidence interval of the ICC (17): poor, lower bound less than 0.50; moderate, between 0.50 and 0.75; good, between 0.75 and 0.90; and excellent, greater than 0.90. ICC and their confidence interval were computed in R version 4.2.1 (18).

Code

Analysis and results for phase 1 were scripted in MATLAB (The MathWorks Inc., version 2020b and later). Analysis and results for phases 2 and 3, the figures and tables pertaining

to the results, and the analysis presented in Supplementary Note 5 were scripted and created in R, version 4.2.1 (18), and later. All code is available here: https://github.com/theibsi/ibsi_2_data_analysis (commit fde70ca).

Results

Characteristics of Participating Teams

Fifteen teams from seven countries participated in the first phase, eleven teams in the second phase, and nine teams in the validation phase. Twelve teams had developed publicly available software: CaPTk, CERR, FAST, LIFEx, MIRAS, MIRP, moddicom, S-IBEX, SPAARC, ViSERA, and the McGill and Université de Sherbrooke teams (see Supplementary Note 7).

First Phase Results

Of the thirty-six filtered images that were assessed in the first phase, moderate or better consensus was found for seventeen (47%) at the initial timepoint (Figure 4). At the final time-point, moderate or better consensus was achieved for thirty-three (92%) configurations, of which twenty-four (67%) were very strong. Full consensus was reached for configurations corresponding to mean filters, Laplacian-of-Gaussian filters, Laws kernels, Gabor kernels, as well as separable and non-separable wavelets (including decomposed forms). No or only weak consensus was achieved for three (8%) configurations, corresponding to configurations involving Riesz transformations (Supplementary Figure 1).

Second Phase Results

At the initial time point of the second phase, moderate or better consensus was achieved for 198 (50%) of 396 features, aggregated over twenty-two different filter configurations (Figure 4). At the final time-point 323 (82%) features had at least moderate consensus. Again, full consensus was reached for features computed from filtered images of mean filters, Laplacian-of-Gaussian filters, Laws kernels, Gabor kernels, as well as separable and non-separable wavelets (including decomposed forms), except for the quantile coefficient of dispersion feature for three-dimensional non-separable wavelets. No consensus was established for features based on (steered) Riesz transformations (Supplementary Figure 2) because too few teams submitted values for these features.

Validation Results

In summary, eight types of convolutional filters were standardized in the first two phases. The reproducibility of features from filtered images created by these filters was assessed in the third phase. Here, 458 (92%) of 486 features were found to have good to excellent reproducibility (ICC 95% CI lower bound > 0.75; see Figure 4). Overall, nineteen (4%) features were poorly reproducible (ICC 95% CI lower bound < 0.50), and were found for

Laplacian-of-Gaussian, separable and non-separable wavelet filters. Most of these features were either coefficient of variation or quartile coefficient of dispersion features that represented eight and nine of nineteen features, respectively. A list of poorly reproducible features is provided in Supplementary Table 1. All ICC values and their 95% confidence intervals are listed in Supplementary Tables 2-10. No dependence on imaging modality could be observed.

Discussion

Convolutional filters enhance specific structures and patterns in medical images and are commonly used in radiomics analyses. However, due to lack of proper consensus-based reference implementations, features computed from filtered images provided by these filters were difficult to reproduce (9). In our study, fifteen teams from seven countries collaborated to remedy this situation by providing reference filtered images, reference feature values, and reference documentation. We were able to standardize and validate eight different filter types: mean, Laplacian-of-Gaussian, Laws and Gabor kernels, and separable and non-separable wavelet filters in both undecomposed and decomposed forms. Thirty-three reference filtered images and 323 reference feature values, computed from filtered images, were established to standardize radiomics analyses across various imaging modalities.

Our current results complement the previous results of the Image Biomarker Standardization Initiative (4). That work focused on standardizing both the image processing scheme for radiomics and a large set of radiomic features. It aimed to improve reproducibility of radiomics studies by mitigating the effect of using different radiomics software packages, and by providing a common framework for describing methodological details. Our current work adds to the previous by standardizing the use of convolutional filters frequently used in radiomics.

Despite the overall success of the standardization process, there were two instances in which we did not achieve the desired level of success. Firstly, we were unable to standardize Riesz transformations that, despite their attractive characteristics from a signal processing perspective, were not easy to implement. Thus, too few teams did so and we could not provide reference filtered images and reference values for Riesz transformations. As Riesz transformations are only rarely used in radiomics studies, the impact should be minimal. Secondly, several features could not always be computed in a reproducible manner, notably the coefficient of variation and quartile coefficient features in conjunction with high- and band-pass convolutional filters. Such filters are characterized by a filtered image with a mean intensity of zero. In the presence of high- and band-pass convolutional filters, the mathematical division operation present in both features led to otherwise negligible numeric differences between teams becoming relevant, resulting in poor reproducibility. Therefore, these features should not be used in combination with high- and band-pass filters.

Zooming out, our current work has several important implications: firstly, we found that reproducible implementation of most types of convolutional filters across different radiomics software is not straightforward, as evidenced by the initial lack of consensus on reference filtered images in phase 1 (see Supplementary Note 8 for lessons learned). Thus, we must assume that existing clinical or research radiomics software, that incorporates convolutional

filters in advanced image analysis workstations, may yield feature values that are not externally reproducible. This might impede external validation and subsequent clinical translation until the software is made compliant.

The second implication is that software labeled as “IBSI-compliant” is now expected to reproduce the reference filtered images and reference feature values found in our current study, insofar as convolutional filters are available in the software, in addition to the existing reference feature values (4). Developers of radiomics software supporting convolutional filters should endeavor to make their software compliant to improve reproducibility of radiomics analyses and allow for translation of enhanced clinical insights offered by convolutional filters. Developers should then clearly label their software as IBSI-compliant, to make it easier for users to identify and use their software for research and/or clinical purposes (with regulatory approval). Compliance may be checked using web-based tools (https://ibsi.radiomics.hevs.ch/), or by manually comparing the produced filtered images and feature values against the provided reference data. Compliant software is expected to produce filtered images where every voxel deviates from the reference filtered image by at most 1% of the range of intensity values of the reference filtered image (Supplementary Note 5). Similarly, feature values must fall within the specified tolerance margin around their reference feature values.

Thirdly, even though we contextualized our efforts within radiological imaging, our work is relevant for quantitative image analysis in general, including digital pathology. Like our previous study (4), the current work is anticipated to improve reproducibility of radiomics analyses beyond the modalities (digital phantoms, chest CT) and settings (non-small cell lung cancer) examined during the initial two phases of this study. To provide preliminary evidence supporting this notion, we conducted validation using a publicly available dataset comprising patients with soft-tissue sarcoma and multiple imaging modalities. The outcomes of the validation phase reinforce the potential applicability of our work in diverse settings.

Our current work had the following limitations. First, its scope is restricted. Compliance with IBSI reference values helps to improve reproducibility of radiomic features (5,6). Yet, the results of a radiomics analysis also depend on image acquisition, reconstruction, segmentation, and data analysis steps (19,20), which we did not address here or in our previous work. Differences in, for example, image acquisition protocols are known to affect the appearance of an image, and therefore also reproducibility of radiomic features(21). Such effects can be reduced by harmonization and cross-calibration of scanners and protocols (22) and post-hoc techniques such as perturbation (23,24), batch normalization (25), and other methods (26). Second, participation in the IBSI does not guarantee that a particular software package is compliant with the IBSI reference standard. Changes introduced in software (5), or design choices may limit compliance (27). Third, we standardized intensity-based statistical features computed from filtered images but no other types of features. Particularly, morphological features are mostly redundant as these are based on segmentation masks that are explicitly not altered by convolutional filtering. Most texture features, in our estimation, would be too abstract to allow for interpretation when computed from filtered images. Their use may add hundreds or thousands of features to a radiomics analysis, which complicates the process of creating generalizable and interpretable radiomics models in the typical setting where at most a few hundred images are available for analysis. Finally, the IBSI has so far focused on radiomics using

handcrafted features, and with this work offers a comprehensive reference standard for their computation. However, we recognize that there are more features and other filters than the ones we have standardized so far. These are not implemented often and will be hard to standardize for that reason.

In conclusion, we standardized eight types of convolutional filters for radiomics to ensure that the enhanced clinical insights that can be gained through their use can be validated and reproduced. Going forward, developers should ensure compliance of their software with the proposed reference standards, and users are encouraged to use compliant software. A web-based tool is available for compliance-checking. In the future, the IBSI will focus on deep learning applications of radiomics, with an aim to provide reference standards for image pre-processing.

Data sharing statement

Data generated by the authors or analyzed during the study are available at: https://github.com/theibsi/data_sets (imaging data) and https://github.com/theibsi/ibsi_2_reference_data (reference filtered images and reference feature values). Code used to analyze the data and obtain the results can be found here: https://github.com/theibsi/ibsi_2_data_analysis (commit fde70ca).

Funding information

The authors were supported by the National Cancer Institute grants P30CA008748 (A.P.A.), U01CA242871 (B.B., S.B.) and U24CA189523 (B.B., S.B.); UK Research & Innovation London Medical Imaging and Artificial Intelligence Centre (G.J.R.C.); UK Wellcome / Engineering and Physical Sciences Research Council Centre for Medical Engineering at King's College London (WT 203148/Z/16/Z) (G.J.R.C.); Cancer Research UK National Cancer Imaging Translational Accelerator awards C1519/A28682 (G.J.R.C., C.G.R.) and C4278/A27066 (V.G.); Swiss National Science Foundation grants 310030_170159 (H.S.G.), CRSII5_183478 (S.T.), 320030_176052 (H.Z.), 205320_179069 (A.D.), and 325230_197477 (A.D.); Natural Sciences and Engineering Research Council of Canada Discovery Grant (RGPIN-2019-06467) (A.R.); UK Engineering and Physical Sciences Research Council (EP/N509449/1) (E.S.); Canada CIFAR AI Chairs Program (M.V.); Swiss Personalized Health Network IMAGINE and QA4IQI projects (A.D.); and RCSO IsNET HECKTOR project (A.D.).

References

1. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology. 2016;278(2):563–577. doi: 10.1148/radiol.2015151169.

2. Tomaszewski MR, Gillies RJ. The Biological Meaning of Radiomic Features. Radiology. 2021;298(3):505–516. doi: 10.1148/radiol.2021202553.

3. Huang EP, O'Connor JPB, McShane LM, et al. Criteria for the translation of radiomics into clinically useful tests. Nat Rev Clin Oncol. 2022; doi: 10.1038/s41571-022-00707-0.

4. Zwanenburg A, Vallières M, Abdalah MA, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology. 2020;295(2):328–338. doi: 10.1148/radiol.2020191145.

5. Fornacon-Wood I, Mistry H, Ackermann CJ, et al. Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform. Eur Radiol. 2020;30(11):6241–6250. doi: 10.1007/s00330-020-06957-9.

6. Bettinelli A, Marturano F, Avanzo M, et al. A Novel Benchmarking Approach to Assess the Agreement among Radiomic Tools. Radiology. 2022;211604. doi: 10.1148/radiol.211604.

7. Depeursinge A, Al-Kadi OS, Ross Mitchell J. Biomedical Texture Analysis: Fundamentals, Tools and Challenges. Academic Press; 2017. doi: 10.1016/C2016-0-01903-4.

8. Beuque MPL, Lobbes MBI, van Wijk Y, et al. Combining Deep Learning and Handcrafted Radiomics for Classification of Suspicious Lesions on Contrast-enhanced Mammograms. Radiology. 2023;307(5):e221843. doi: 10.1148/radiol.221843.

9. Bogowicz M, Leijenaar RTH, Tanadini-Lang S, et al. Post-radiochemotherapy PET radiomics in head and neck cancer - The influence of radiomics implementation on the reproducibility of local control tumor models. Radiother Oncol. 2017;125(3):385–391. doi: 10.1016/j.radonc.2017.10.023.

10. Depeursinge A, Andrearczyk V, Whybra P, et al. Standardised convolutional filtering for radiomics. arXiv [eess.IV]. 2020. doi: 10.48550/arXiv.2006.05470.

11. Lambin P. Data from: Radiomics Digital Phantom. CancerData; 2016. doi: 10.17195/candat.2016.08.1.

12. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging. 2013;26(6):1045–1057. doi: 10.1007/s10278-013-9622-7.

13. Vallières M, Freeman CR, Skamene SR, El Naqa I. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. Phys Med Biol. 2015;60(14):5471–5496. doi: 10.1088/0031-9155/60/14/5471.

14. Vallières M, Freeman CR, Skamene SR, El Naqa I. Data from: A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. The Cancer Imaging Archive; 2015. doi: 10.7937/K9/TCIA.2015.7GO2GSKS.

15. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86(2):420–428. doi: 10.1037/0033-2909.86.2.420.

16. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155–163. doi: 10.1016/j.jcm.2016.02.012.

17. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods. 1996;1(1):30–46. doi: 10.1037//1082-989x.1.1.30.

18. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2022. https://www.R-project.org/.

19. Zwanenburg A. Radiomics in nuclear medicine: robustness, reproducibility, standardization, and how to avoid data analysis traps and replication crisis. Eur J Nucl Med Mol Imaging. 2019;46(13):2638–2655. doi: 10.1007/s00259-019-04391-8.

20. van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging-"how-to" guide and critical reflection. Insights Imaging. 2020;11(1):91. doi: 10.1186/s13244-020-00887-2.

21. Berenguer R, Pastor-Juan MDR, Canales-Vázquez J, et al. Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters. Radiology. 2018;288(2):407–415. doi: 10.1148/radiol.2018172361.

22. Sullivan DC, Obuchowski NA, Kessler LG, et al. Metrology Standards for Quantitative Imaging Biomarkers. Radiology. 2015;277(3):813–825. doi: 10.1148/radiol.2015142202.

23. Zwanenburg A, Leger S, Agolli L, et al. Assessing robustness of radiomic features by image perturbation. Sci Rep. 2019;9(1):614. doi: 10.1038/s41598-018-36938-4.

24. Teng X, Zhang J, Zwanenburg A, et al. Building reliable radiomic models using image perturbation. Sci Rep. 2022;12(1):1–10. doi: 10.1038/s41598-022-14178-x.

25. Orlhac F, Frouin F, Nioche C, Ayache N, Buvat I. Validation of A Method to Compensate Multicenter Effects Affecting CT Radiomics. Radiology. 2019;291(1):53–59. doi: 10.1148/radiol.2019182023.

26. Mali SA, Ibrahim A, Woodruff HC, et al. Making Radiomics More Reproducible across Scanner and Imaging Protocol Variations: A Review of Harmonization Methods. J Pers Med. 2021;11(9):842. doi: 10.3390/jpm11090842.

27. Wright DE, Cook C, Klug J, Korfiatis P, Kline TL. Reproducibility in medical image radiomic studies: contribution of dynamic histogram binning. arXiv [eess.IV]. 2022. doi: 10.48550/arXiv.2211.05241.

Figure Legends

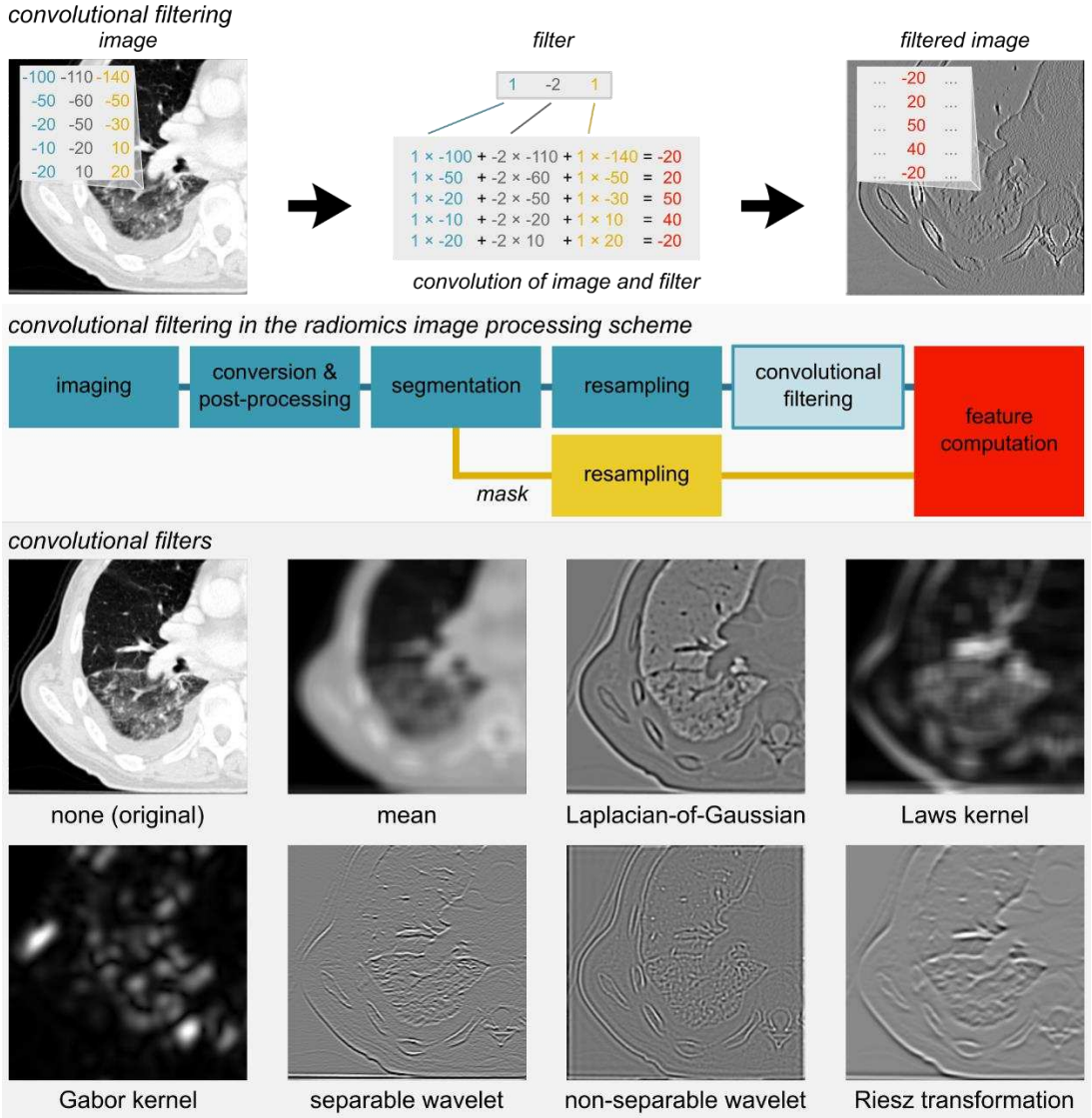

Figure 1: Overview of convolutional filters. An image is filtered using convolution to create a filtered image (*top panel*). Each image consists of values. Here, during convolution a filter with three weights (1.0, -2.0, 1.0) is slid across the image, and adjacent image values are multiplied with the corresponding filter values and summed to create a response value for each position in the image. Convolutional filtering is positioned after resampling in the overall radiomics image processing scheme (*center panel*). This workflow starts with an image that is obtained from a repository or archiving system in a digital format, such as DICOM. Then the image is optionally converted (e.g., from PET activity to standardized uptake values) and post-processed (e.g., MR bias-field correction). Segmentation masks are either loaded in a digital format, or automatically created. Both image and segmentation masks are then optionally resampled. Filtered images are created by filtering the image. Both filtered image and segmentation mask are then used to compute handcrafted radiomic features. This study attempts to standardize several types of convolution filters (*bottom panel*). The original CT image is shown for reference. Decomposition of separable and non-separable wavelets is not shown.

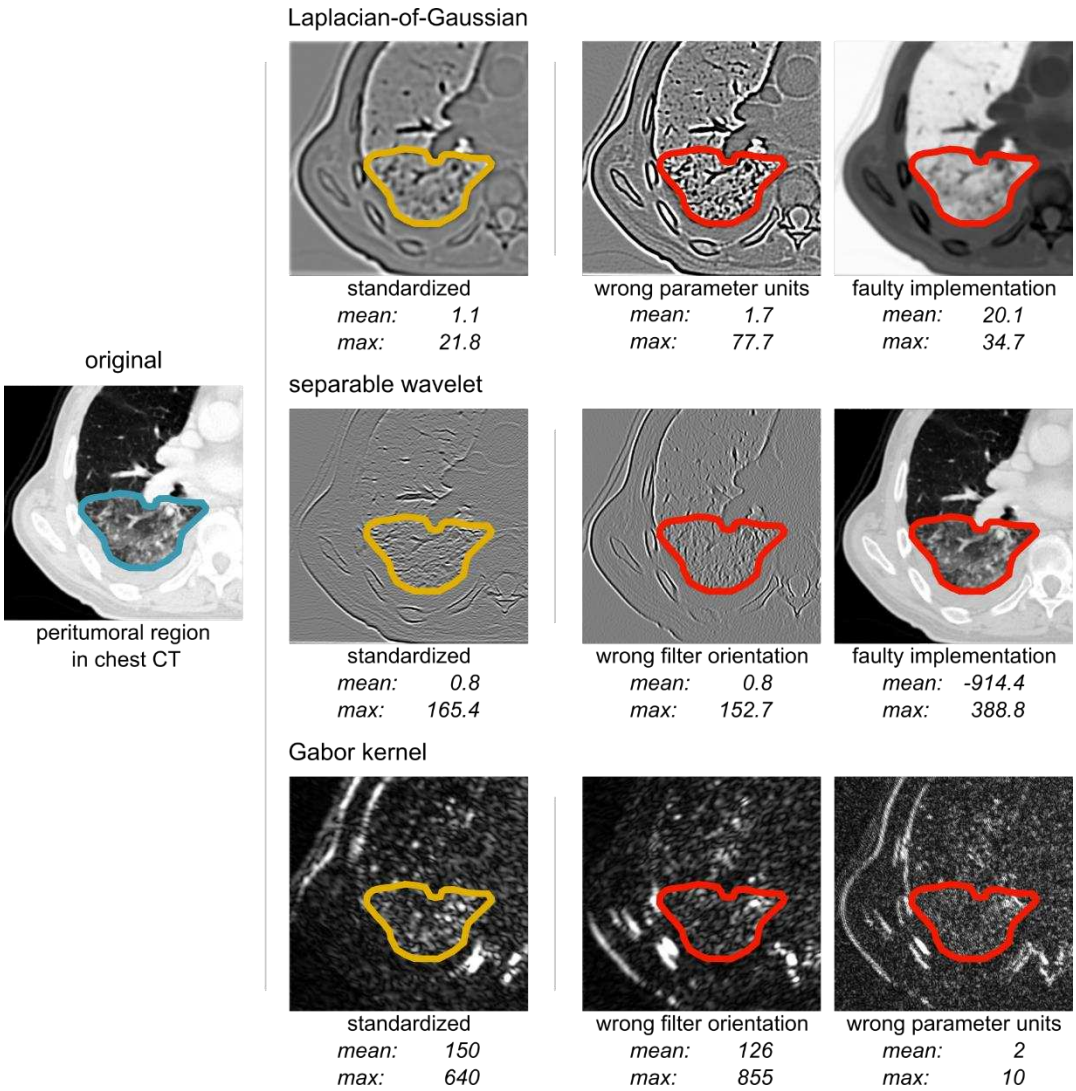

Figure 2: The need for standardization. Filters can enhance and quantify potentially clinically relevant characteristics and textures. Here three filters are used to quantify different characteristics of the peritumoral region in a chest CT, with the tumor being out-of-plane. However, filters are not trivial to implement, and their parameters may be ambiguous without standardization. For each filter, mean and maximum intensity are computed within the segmentation masks in three filtered images. The leftmost filtered image was created by applying a standardized filter to the original image. The other two filtered images resulted from filter implementations that were not standardized. The Laplacian-of-Gaussian filter is used to quantify the presence of edges and highlight fine details. The scale of the filter is 2.0 mm, and it is truncated at 8.0 mm. The non-standardized filters respectively use 2.0 voxels (not mm) and truncate at one filter scale (2.0 mm). Separable wavelets are designed to quantify image contents for different frequency bands, though in many radiomics analyses they are used to quantify edges. A pair of low-pass and high-pass wavelet kernels is used to filter the image, highlighting edges in the lateral direction. The non-standardized filters either use an incorrect orientation (i.e., low-pass and high-pass kernels were swapped) or are faulty because the first kernel is used for all directions (i.e., a pair of low-pass - low-pass wavelet kernels). Gabor filters are used to quantify directional structures (e.g., fibrosis and bronchi). The standardized filter used scale and wavelength parameters of 2.0 mm and was oriented under 30°. The non-standardized filters use an incorrect orientation, or express

parameters in 2.0 voxels (not mm), respectively. The lack of standardization leads to markedly different feature values, which prevents reproducing, validating and clinically translating decision support tools that use these features.

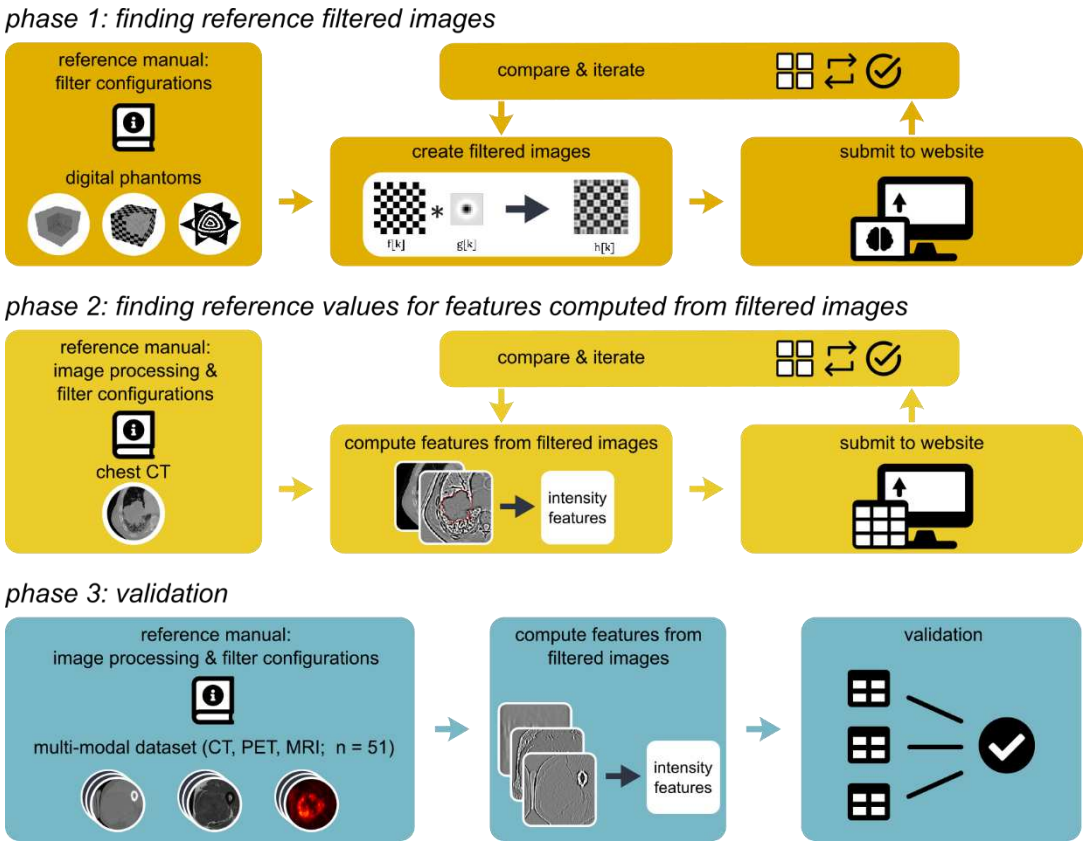

Figure 3: Study overview. The study is divided into three phases. In the first phase, convolutional filters were applied to digital phantoms to identify reference filtered images. In the second phase, reference values were identified for intensity-based features computed from filtered image of a chest CT image. In the third phase, the results of the first two phases were validated using a multi-modal dataset of soft tissue sarcoma patients. Unlike the first two phases, the validation phase was not iterative. Some figure elements were adapted from Depeursinge et al. (10).

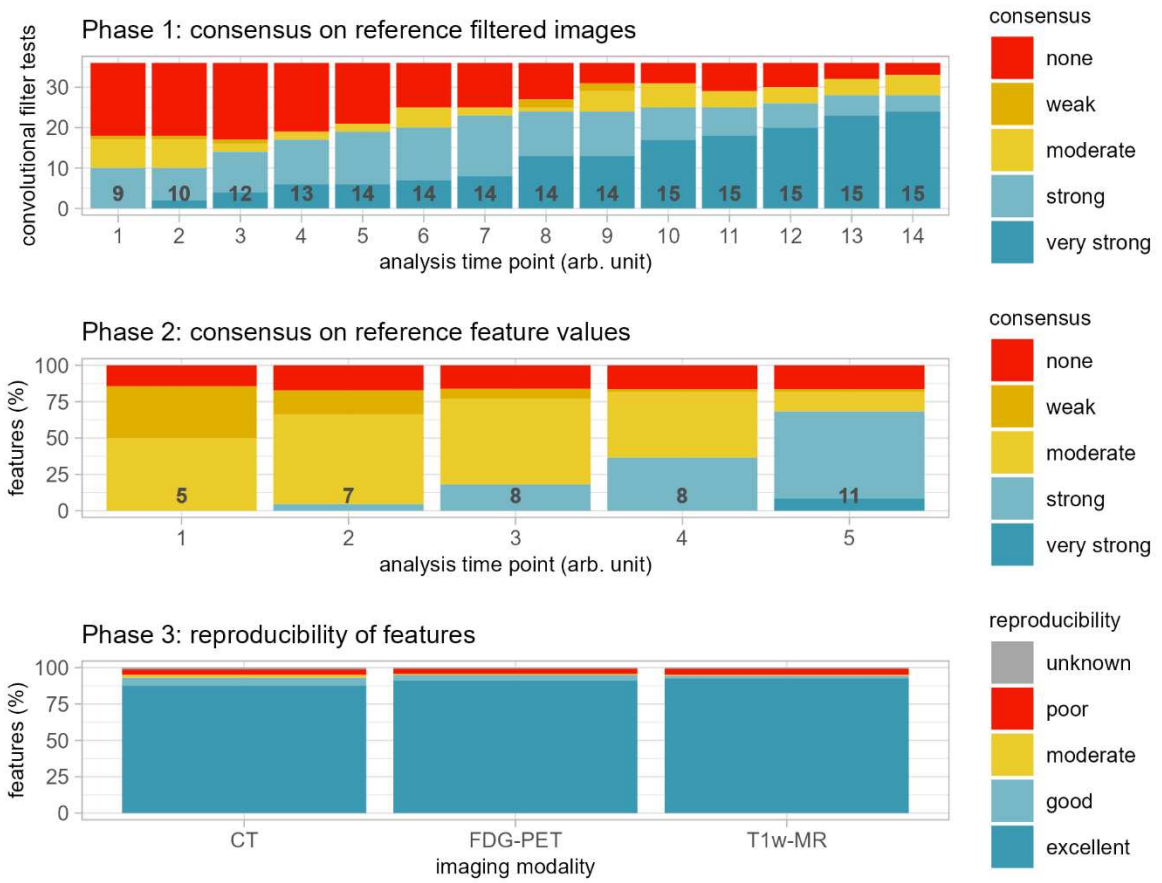

Figure 4: Results overview. In phase 1, participating teams computed thirty-six filtered images of convolutional filters according to predefined configurations. These filtered images were compared, and consensus was measured. Teams updated their implementations iteratively, which led to an improvement of consensus over time (arbitrary unit, the entire process took twenty-seven months). Consensus strength was based on matching the voxel-wise difference between filtered images and the tentative reference filtered image within a tolerance: weak, match between fewer than three teams; moderate, three to five; strong, six to nine; very strong, ten or more; none, 50% of the teams or more did not match. The number of participating teams at each timepoint is shown. In phase 2, participating teams computed 396 features from filtered images of convolutional filters according to predefined filter and image processing configurations. As in phase 1, teams updated their implementations iteratively. However, unlike phase 1, improvement in consensus was mostly due to more teams enrolling over time (arbitrary unit, the entire process took fifteen months). Consensus strength was based on the number of teams matching the tentative reference feature value within a tolerance and was assigned according to the same categories as in phase 1. In phase 3, reproducibility of features computed from filtered images was validated. Teams computed 486 features from a public dataset of fifty-one patients with soft-tissue sarcoma that were scanned using CT, fluorine-18 fluorodeoxyglucose (FDG)-PET, and T1w-MR imaging. Reproducibility was assessed using the lower bound of the 95% confidence interval of the intraclass correlation coefficient: poor, lower bound less than 0.50; moderate, between 0.50 and 0.75; good, between 0.75 and 0.90; excellent, greater than 0.90; and unknown, computed by fewer than two teams.

Tables

Table 1: Glossary of terms

standardization	the process of establishing uniform guidelines and protocols to ensure consistency and reproducibility.
convolutional filter	a filter consisting of fixed or parameterized numerical values, that is slid (convolved) over an image to enhance potentially relevant characteristics, such as normal tissue-tumor boundaries, blood vessels, texture, and fibrosis.
filtered image	the image produced by applying a (convolutional) filter to an image.
low-pass filter	a filter that suppresses noise and other sharp patterns in an image and enhances smooth aspects.
high-pass filter	a filter that suppresses smooth aspects of an image and enhances details and sharp image patterns.
band-pass filter	a filter that suppresses both smooth aspects of an image as well as sharp image patterns and enhances intermediate details.
reference filtered image	an established filtered image representing the expected output of a specific convolutional filter applied to a specific image, that serves as a benchmark for verification and calibration.
(radiomics) feature:	a quantitative measure that is computed from a region of interest in a (filtered) image. The computation of common features was previously standardized by the Image Biomarker Standardization Initiative (4).
reference feature value	an established expected value when computing a feature from a specific region of interest in a specific (filtered) image, that serves as a benchmark for verification and calibration.
radiomics software	a software package that (at least) processes medical imaging and computes radiomics features.
radiomics decision support tool	a computer application that provides clinical decision support based on radiomics features.
mean filter	a filter that computes the average value within a neighborhood of voxels.
Laplacian-of-Gaussian filter	a filter used to detect edges and highlight fine details in an image.
Laws kernels	sets of predefined filters used for highlighting various patterns in images, such as ripples.
Gabor kernels	filters used for detection of directional patterns.
wavelets	sets of filters used to decompose images into different spatial frequency ranges.
Riesz transformation	a mathematical operation on filters that enhances edges and directional patterns in an image.