This is a pre-print of the paper accepted for publication at the European Workshop on Visual Information Processing (EUVIP) 2024

Comparing Stability and Discriminatory Power of Hand-crafted Versus Deep Radiomics: A 3D-Printed Anthropomorphic Phantom Study

Oscar Jimenez-del-Toro*, Christoph Aberle†, Roger Schaer‡, Michael Bach†, Kyriakos Flouris§, Ender Konukoglu§, Bram Stieltjes†, Markus M. Obmann†, André Anjos*, Henning Müller‡¶, Adrien Depeursinge‡‖

*Idiap Research Institute, Martigny, Switzerland
†Clinic of Radiology and Nuclear Medicine, University Hospital Basel, Basel, Switzerland
‡University of Applied Sciences Western Switzerland (HES-SO) Valais, Sierre, Switzerland
§Computer Vision Lab, ETH Zurich, Zurich, Switzerland
¶Faculty of Medicine, University of Geneva (UNIGE), Geneva, Switzerland
‖Dept. of Nuc. Med. and Molecular Imaging, Lausanne University Hospital, Lausanne, Switzerland

***Abstract*—Radiomics have the ability to comprehensively quantify human tissue characteristics in medical imaging studies. However, standard radiomic features are highly unstable due to their sensitivity to scanner and reconstruction settings. We present an evaluation framework for the extraction of 3D deep radiomics features using a pre-trained neural network on real computed tomography (CT) scans for tissue characterization. We compare both the stability and discriminative power of the proposed 3D deep learning radiomic features versus standard hand-crafted radiomic features using 8 image acquisition protocols with a 3D-printed anthropomorphic phantom containing 4 classes of liver lesions and normal tissue. Even when the deep learning model was trained on an external dataset and for a different tissue characterization task, the resulting generic deep radiomics are at least twice more stable on 8 CT parameter variations than any category of hand-crafted features. Moreover, the 3D deep radiomics were also discriminative for the tissue characterization between 4 classes of liver tissue and lesions, with an average discriminative power of 93.5%.**

***Index Terms*—Radiomics, deep learning, feature stability, biomedical texture, CT**

I. Introduction

Radiomics transform radiological studies into mineable quantitative or semi-quantitative data [1]. Combined with machine learning methods, radiomic features can enable the diagnostic and therapeutic decision-making process with objective measurements of high dimensional data [2]. Nevertheless, limited generalizability of findings from radiomic studies has been a major deterrent in the translation of radiomics into standard clinical practice [3]. Radiomics is a complex multi-step process with methodological challenges within each step to ensure robustness and reproducibility [4]. Variations in the acquisition and reconstruction processes, including protocol and scanner differences, can have a strong impact on radiomic features, creating bias and reducing significantly their stability [5]. Moreover, the required feature stability and discriminative power for a particular task is strongly linked to imaging modality, organ, disease, and other factors [6].

Retrospective and prospective validation of radiomic features and models is a key step towards clinical application [7], which are however are not systematically evaluated in most radiomic studies. In the context of computed tomography (CT) radiomics, test-retest studies imply radiation exposure entailing ethical concerns in the recruitment of a large cohort of patients and are thus limited to small sample sizes [8]. Physical phantoms can be exploited to study variations related to image acquisition by performing test-retest measurements in highly controlled settings [9]. Simulator environments have also been developed to perform this analysis [10]. There are discrepancies in methodology across the available studies on how to measure the generalizability of radiomic models [11]. As a general assumption, the differences in radiomic feature values obtained from different types of tissues should be greater than the intra-class differences resulting from parameter variations during the acquisition [12]. It is therefore essential to evaluate the scope of feature stability in conjunction with its discriminative power for a specific medical task.

Deep learning radiomics, also called discovery radiomics, use non-engineered radiomic features that can be learned from data [13]. Only a couple of studies have measured the stability of deep radiomics across multiple acquisition protocols [14], [15]. Features obtained with pre-trained 2D convolutional neural networks (CNN) were more stable against technical variations in CT when compared to hand-crafted radiomics features [15]. However, pre-trained deep learning models are usually trained on two-dimensional data from other domains, *i.e.,* photographs, while radiomic data is often three-dimensional (3D) [16]. Deep learning models used as 2D feature extractors lose the spatial context of 3D patient scans. Moreover, generic features for tissue characterization are thought to be more interpretable by a human reader [17]. To the best of our knowledge, no phantom study has evaluated the stability and discriminative power of pre-trained 3D networks for deep radiomic feature extraction.

In this pilot study we propose an evaluation framework for pre-trained 3D deep learning models developed for the extraction of deep radiomic features that could be used for tissue characterization on independent data sets (see Fig. 1). The feature stability and discriminative power are compared to standard hand-crafted features and to Riesz wavelet features [18], in a anthropomorphic phantom study.

II. Material and Methods

A. Anthropomorphic 3D-printed CT phantom

The iodine-ink based 3D-printed phantom described in Bach et al. was used to study the stability and discriminative power of hand-crafted, and generic deep radiomics features [9]. The used 3D-printing technique can produce high detail in both anatomical structure and realistic CT texture in the phantom [19]. In this study, the radiomic feature analysis was performed on the phantom section containing the printed dataset derived from an abdominal CT scan from an oncologic patient. The liver of this section contains various types of liver lesions including benign cysts, a hemangioma and a pathology proven cancer metastasis. Six 3D regions of interest (ROIs) of 4 classes of liver tissue and normal tissue were manually annotated by a board-certified radiologist in the liver of a thin-sliced phantom acquisition: two ROIs with normal liver tissue, two benign cysts, an hemangioma and a liver metastasis. The ROIs were not enhanced or modified to focus on image differences that are attributed to scanning protocols. The 3D ROI volumes were rigidly registered to the available CT phantom acquisitions used in this study using the Elastix toolbox [20].

B. Data acquisition

The phantom was imaged with a Siemens SOMATOM Definition Edge (Siemens Healthineers, Erlangen, Germany) CT scanner. No ethical approval was required. Eight groups of CT parameter variations were selected to study the stability of radiomic features as it was set up in the study by Jimenez-del-Toro et al. [12]. The following image reconstruction parameters were varied: reconstruction algorithm (iterative reconstruction, IR or filtered back projection, FBP), reconstruction kernel (2 standard soft tissue kernels per algorithm) and slice thickness in millimetres (1, 1.5, 2, 3), and slice spacing in millimetres (0.75, 1, and 2). Thirty repetition phantom scans with identical settings were acquired for each of the 8 parameter variation groups. The DICOM dataset is publicly available [1] [12].

C. Extraction of hand-crafted radiomic features

A total of 86 hand-crafted radiomic features were extracted from the 3D ROI volumes using the open-source PyRadiomics (version 3.0) toolkit [21]. The categories of hand-crafted features include: first-order statistics (N=18), gray level co-occurrence matrices (N=22), gray level dependence matrices (N=14), gray level run length matrices (N=16) and gray level size zone matrices (N=16). Feature parameters were set to their default value with a fixed bin width of 25 for the discretization of the gray-intensity levels in the CT scans, *i.e.,* Hounsfield units.

To further explore tissue texture properties and the characterization of local scales, 3D Riesz filterbanks were also extracted from the ROI volumes [18]. Twenty-seven Riesz wavelet features were extracted using the 2nd order of the Riesz transform at 3 scales, which provides a good trade-off between the dimensionality of the feature space and the wealth of the filter banks.

D. Extraction of deep radiomics features

A 3D deep learning model was trained on an external dataset, the VISCERAL dataset, for generic tissue characterization [22]. The pre-trained model was used as a 3D deep radiomics feature extractor for the phantom study. The VISCERAL project organized three benchmarks on the automated anatomy localization and segmentation of whole-body 3D volumes. The benchmark dataset included 30 contrast-enhanced CT scans (ceCT) of the trunk from patients with malignant lymphoma with a resolution of $0.604^2 - 0.793^2 \times 1.2 - 1.5$ mm^3. Twenty anatomical structures were manually annotated by physicians. An additional Silver Corpus of 65 ceCT scans was generated with the fusion of the output segmentations from the participants' algorithms [23]. For this work, we used 93 ceCT scans, containing 5 anatomical structures, from the VISCERAL Anatomy dataset and the Silver Corpus to train a CNN with the goal of classifying 3D tissue samples from these 5 structures: right lung, liver, spleen, right kidney and urinary bladder. The ceCT scans were resampled to isometric voxels of 1 mm. A single 3D block centered on the segmentation volume of each anatomical structure was used per patient scan. The size of each 3D block was $64 \times 64 \times 32$ voxels. A total of 465 3D blocks were randomly separated into an 80-20 distribution (375-90), keeping blocks from the same patient in the same set to avoid data leakage.

The self-designed 3D deep learning model comprises two convolutional layers and two fully connected (FC) layers. Max-pooling was performed after each of the convolutional layers. The convolutional layers had 32 channels, a $5 \times 5 \times 5$ kernel size, a stride of 1, and a padding of 1 for each layer. We used dropout regularization of 0.5 after the first FC layer of the model to improve the generalization performance. Cross-entropy loss with logits is computed to supervise the training with an Adam optimizer. Once the model has been pre-trained to classify the 5 anatomical structures, the weights were fixed and deep radiomics features were extracted on a 3D block of the same size, centered on the liver ROIs from the phantom. Thirty deep radiomics features were extracted from the last FC layer, after the rectified linear unit (ReLU) activation function.

E. Statistical analysis

Sets of univariate Wilcoxon signed rank tests were performed for each of the extracted radiomics features to assess their stability and discriminative power. Two-tailed tests with

[1] https://doi.org/10.7937/a1v1-rc66, as of July 2024.

p-value<0.05 were considered statistically significant. For the feature stability, the tissue class is fixed and the feature values from each of the 8 CT parameter variation groups (each group composed of 30 repetition scans) are compared in a pairwise approach. The hypothesis is that the values for each pairwise acquisition group comparison originate from the same distribution (intraclass variation), as they are obtained from the same ROI in the 3D-printed phantom. This process is repeated with the 4 tissue classes, resulting in 112 stability tests per radiomics feature: 28 unique correlations between the 8 CT parameter variations × 4 liver tissue classes. For the discriminative power, the CT parameters are fixed while feature values from different tissue classes are compared. The hypothesis is that feature values between two different liver tissue classes should not originate from the same distribution (interclass variation), to be useful for tissue characterization tasks. All 8 CT parameter variation groups are compared, resulting in 48 discriminative power tests per radiomics feature: 6 unique correlations between the 4 liver tissue classes × 8 CT parameter variations. The number of stability tests and discriminative power tests supporting the null hypotheses between pairwise group comparisons were translated into percentages for the radiomics analysis. Statistical significance was not corrected for multiple testing as the analysis is performed for individual features, *i.e.*, univariate performance.

III. Results

A. Pre-trained 3D deep learning model

The 3D deep learning model for tissue classification obtained a 0.93 accuracy on a balanced independent test set of 90 3D blocks from 5 anatomical structures segmented in the VISCERAL dataset. Thirty deep radiomics features were extracted post-ReLU by only forward propagation on all the liver ROIs from the available CT acquisitions of the 3D-printed phantom.

B. Radiomics analysis on 3D-printed phantom

Radiomics features were unstable in more than 25% of all the feature stability tests among the 8 CT parameter variations. Deep radiomics were overall more stable than both hand-crafted and Riesz wavelet features (see Fig. 2). Deep radiomics had an average stability of 9.58%, while the most stable category of hand-crafted features, *i.e.,* first-order features, had an average stability of 4.27% (see Tab. I). Multiple radiomics features had statistically significant differences in all of their test comparisons (100% discriminative power) among the four liver tissue classes. Riesz wavelet features were the most discriminant features with an average discriminative power of 95.76%, but had the lowest stability among all the categories, *i.e.,* 1.65%. The deep radiomics features had an average discriminative power of 93.47%, even though the pre-trained 3D deep learning model was developed with an external dataset, with different CT protocols and different anatomical structures used for tissue characterization.

In Fig. 3 the boxplot distributions of the features with highest score when adding the stability and discriminative power from each category of features, *i.e.* hand-crafted, Riesz wavelets and deep radiomics, are shown. It is possible to linearly separate all of the ROIs using only the obtained deep radiomics feature values with similar distributions across the 8 CT parameter variation groups (Fig. 3, right). This is not the case with the tissue characterization computed with the best performing hand-crafted features, particularly to identify differences between the hemangioma and the metastasis ROIs (Fig. 3, left). The stability of deep radiomics was two times larger than the most stable group of hand-crafted radiomics features, while also maintaining a high discriminative power. Additionally, the deep radiomics targeting the generic characterization of different human tissues could be more interpretable by a human reader than end-to-end deep learning models [17]. Hybrid radiomics models could improve over the hand-crafted and deep radiomcs models, particularly through the use of standardized feature selection based on feature stability and discriminative power.

There are a couple of limitations in the generalization of our results to other works. First, the discrimination task is overly simple as there is only a limited set of lesions present in the phantom. More challenging discriminative tasks are common in medical image analysis and require further research on the development of generic deep learning radiomics. Second, all the CT acquisitions were performed on a single scanner. This allowed to focus on reconstruction parameters alone while removing variations attributed to scanner manufacturers and models. However, larger studies involving multiple scanners, ideally from different manufacturers, are necessary to better understand the scope and limitations on the stability from deep radiomic features.

IV. Conclusion

In this work, generic 3D deep radiomics are presented as a more stable category of radiomics features for tissue characterization that could also be more interpretable than end-to-end deep learning radiomics. Additionally, they improve on the feature stability shown by standard hand-crafted radiomics features with different CT parameter variations, while also showing a promising discriminative power. They could be used in more generalizable deep and hybrid radiomics models.

Acknowledgments: This work was partly supported by the QA4IQI SPHN project as well as SNSF grants 205320_219430 and 325230_197477.

References

[1] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, 2016.

[2] A. Ibrahim, S. Primakov, M. Beuque, H. Woodruff, Halilaj *et al.*, "Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework," *Methods*, vol. 188, pp. 20–29, 2021.

[3] A. Traverso, L. Wee, A. Dekker, and R. Gillies, "Repeatability and reproducibility of radiomic features: a systematic review," *International Journal of Radiation Oncology* Biology* Physics*, vol. 102, no. 4, pp. 1143–1158, 2018.

[4] A. Stanzione, "Feasible does not mean useful: Do we always need radiomics?" *European Journal of Radiology*, vol. 156, 2022.

[5] D. Mackin, X. Fave, L. Zhang, D. Fried, J. Yang *et al.*, "Measuring computed tomography scanner variability of radiomics features," *Investigative radiology*, vol. 50, no. 11, pp. 757–765, 2015.

[6] S. S. Yip and H. J. Aerts, "Applications and limitations of radiomics," *Physics in Medicine & Biology*, vol. 61, no. 13, p. R150, 2016.

[7] R. Berenguer, M. D. R. Pastor-Juan, J. Canales-Vázquez, M. Castro-García, M. V. Villas *et al.*, "Radiomics of ct features may be nonreproducible and redundant: influence of ct acquisition parameters," *Radiology*, vol. 288, no. 2, pp. 407–415, 2018.

[8] F. Prayer, J. Hofmanninger, M. Weber, D. Kifjak, A. Willenpart, J. Pan, S. Röhrich, G. Langs, and H. Prosch, "Variability of computed tomography radiomics features of fibrosing interstitial lung disease: a test-retest study," *Methods*, vol. 188, pp. 98–104, 2021.

[9] M. Bach, C. Aberle, A. Depeursinge, O. Jimenez-del Toro, R. Schaer *et al.*, "3d-printed iodine-ink ct phantom for radiomics feature extraction-advantages and challenges," *Medical Physics*, 2023.

[10] K. Flouris, O. Jimenez-del Toro, C. Aberle, M. Bach *et al.*, "Assessing radiomics feature stability with simulated ct acquisitions," *Scientific reports*, vol. 12, no. 1, p. 4732, 2022.

[11] R. Reiazi, E. Abbas, P. Famiyeh, A. Rezaie, J. Y. Y. Kwan, T. Patel, S. V. Bratman, T. Tadic, F.-F. Liu, and B. Haibe-Kains, "The impact of the variation of imaging parameters on the robustness of Computed Tomography radiomic features: A review," *Computers in Biology and Medicine*, vol. 133, p. 104400, Jun. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482521001943

[12] O. Jimenez-del Toro, C. Aberle, M. Bach, R. Schaer, M. M. Obmann *et al.*, "The discriminative power and stability of radiomics features with computed tomography variations: task-based analysis in an anthropomorphic 3d-printed ct phantom," *Investigative radiology*, vol. 56, no. 12, pp. 820–825, 2021.

[13] M. Avanzo, L. Wei, J. Stancanello, M. Vallieres, A. Rao *et al.*, "Machine and deep learning methods for radiomics," *Medical physics*, vol. 47, no. 5, pp. e185–e202, 2020.

[14] R. Paul, M. S.-u. Hassan, E. G. Moros, R. J. Gillies, L. O. Hall *et al.*, "Deep feature stability analysis using ct images of a physical phantom across scanner manufacturers, cartridges, pixel sizes, and slice thickness," *Tomography*, vol. 6, no. 2, pp. 250–260, 2020.

[15] S. Ziegelmayer, S. Reischl, F. Harder, M. Makowski, R. Braren *et al.*, "Feature robustness and diagnostic capabilities of convolutional neural networks against radiomics features in computed tomography imaging," *Investigative Radiology*, vol. 57, no. 3, pp. 171–177, 2022.

[16] A. Demircioğlu, "Are deep models in radiomics performing better than generic models? a systematic review," *European Radiology Experimental*, vol. 7, no. 1, p. 11, 2023.

[17] A. A. Ardakani, N. J. Bureau, E. J. Ciaccio, and U. R. Acharya, "Interpretation of radiomics features–a pictorial review," *Computer Methods and Programs in Biomedicine*, vol. 215, p. 106609, 2022.

[18] Y. D. Cid, H. Müller, A. Platon, P.-A. Poletti, and A. Depeursinge, "3d solid texture classification using locally-oriented wavelet transforms," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1899–1910, 2017.

[19] P. Jahnke, F. R. Limberg, A. Gerbl, G. L. Ardila Pardo, Braun *et al.*, "Radiopaque three-dimensional printing: a method to create realistic ct phantoms," *Radiology*, vol. 282, no. 2, pp. 569–575, 2017.

[20] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. Pluim, "Elastix: a toolbox for intensity-based medical image registration," *IEEE transactions on medical imaging*, vol. 29, no. 1, pp. 196–205, 2009.

[21] J. J. Van Griethuysen, A. Fedorov, C. Parmar, A. Hosny *et al.*, "Computational radiomics system to decode the radiographic phenotype," *Cancer research*, vol. 77, no. 21, pp. e104–e107, 2017.

[22] O. Jimenez-del Toro, H. Müller, M. Krenn, K. Gruenberg *et al.*, "Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: Visceral anatomy benchmarks," *IEEE transactions on medical imaging*, vol. 35, no. 11, pp. 2459–2475, 2016.

[23] M. Krenn, M. Dorfer, O. A. Jiménez del Toro, H. Müller *et al.*, "Creating a large-scale silver corpus from multiple algorithmic segmentations," in *Medical Computer Vision: Algorithms for Big Data: International Workshop, MCV 2015, Held in Conjunction with MICCAI 2015, Munich, Germany, October 9, 2015, Revised Selected Papers 18*. Springer, 2016, pp. 103–115.

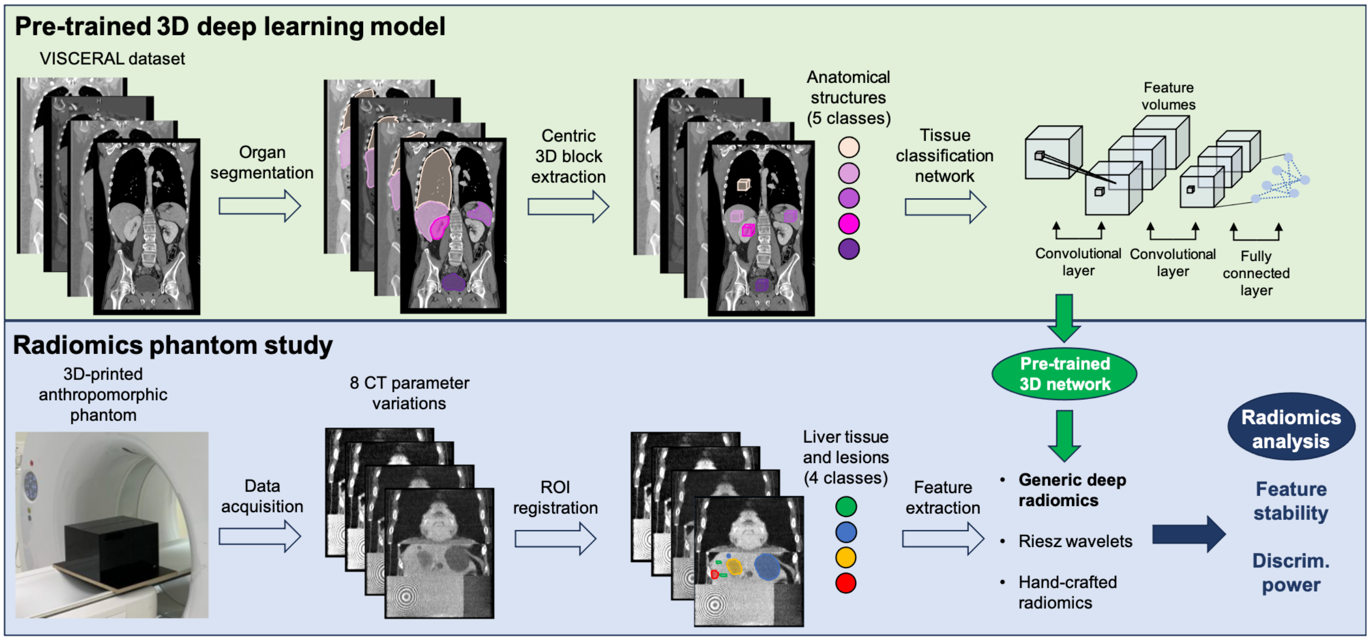

Fig. 1. Our proposed evaluation framework for creating generic 3D deep radiomics for tissue characterization. A 3D deep learning model was pre-trained and validated on an external dataset for the classification of 5 anatomical structures: right lung, liver, spleen, right kidney and urinary bladder. The pre-trained model is then used as a feature extractor of deep radiomics in a anthropomorphic phantom study. The feature stability and discriminative power of hand-crafted radiomics, Riesz wavelets, and deep radiomics is compared using 8 CT parameter variations, and four classes of liver tissue and lesions.

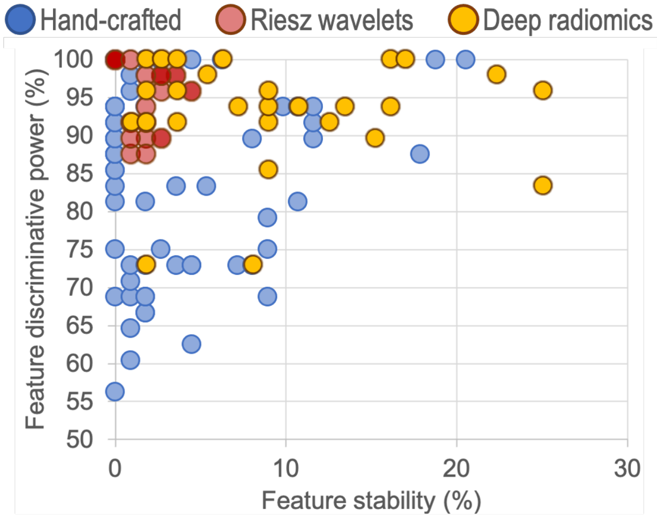

Fig. 2. Scatter plot of hand-crafted radiomics, deep and Riesz wavelet features in terms of percentage of successful stability (x axis) and discriminative power (y axis) pairwise tests. The univariate performance from each feature is measured among 8 CT parameter variations and 4 liver tissue classes.

TABLE I
AVERAGE FEATURE STABILITY AND DISCRIMINATIVE POWER PER CATEGORY OF RADIOMICS FEATURES.

Category	*Num.**	*Stability %*	*Discrim. Power %*
Hand-crafted	86	3.00	86.89
First-order	18	4.27	93.98
GLCM	22	2.19	81.82
GLDM	14	2.68	85.27
GLRLM	16	2.40	85.42
GLSZM	16	3.57	88.80
Riesz wavelets	27	1.65	**95.76**
Deep radiomics	30	**9.58**	93.47

**Num.*: Total number of individual features.

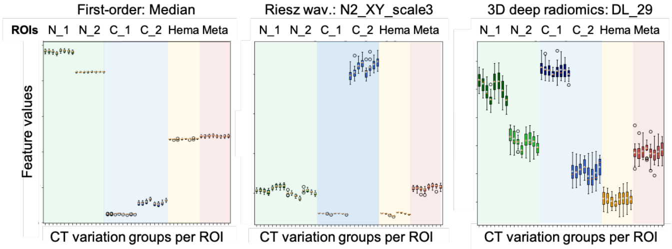

Fig. 3. Best performing radiomics features from each category: hand-crafted, Riesz wavelets and deep radiomics. The boxplots of the feature values obtained in each of the 6 phanthom ROIs containing 4 classes of liver tissue or lesion are shown with a different color, normal tissue (N_1 and N_2): green, cysts (C_1 and C_2): blue, hemangioma (Hema): yellow, and metastasis (Meta): red.