This is a reprint of material published in Proceedings of SPIE 10950 (Medical Imaging 2019: Computer-Aided Diagnosis), 1095046, Mar. 2019.

https://doi.org/10.1117/12.2511856

General Purpose Radiomics for Multi-Modal Clinical Research

Michael Wels, Félix Lades, Alexander Muehlberg, Michael Suehling

Siemens Healthineers, Forchheim, Germany

ABSTRACT

In this paper we present an integrated software solution* targeting clinical researchers for discovering relevant radiomic biomarkers covering the entire value chain of clinical radiomics research. Its intention is to make this kind of research possible even for less experienced scientists. The solution provides means to create, collect, manage, and statistically analyze patient cohorts consisting of potentially multimodal 3D medical imaging data, associated volume of interest annotations, and radiomic features. Volumes of interest can be created by an extensive set of semi-automatic segmentation tools. Radiomic feature computation relies on the de facto standard library PyRadiomics and ensures comparability and reproducibility of carried out studies. Tabular cohort studies containing the radiomics of the volumes of interest can be managed directly within the software solution. The integrated statistical analysis capabilities introduce an additional layer of abstraction allowing non-experts to benefit from radiomics research as well. There are ready-to-use methods for clustering, uni- and multivariate statistics, and machine learning to be applied to the collected cohorts. They are validated in two case studies: for one thing, on a subset of the publicly available NSCLC-Radiomics data collection containing pretreatment CT scans of 317 non-small cell lung cancer (NSCLC) patients and for another, on the Lung Image Database Consortium imaging study with diagnostic and lung cancer screening CT scans including 2,753 distinct lesions from 870 patients. Integrated software solutions with optimized workflows like the one presented and further developments thereof may play an important role in making precision medicine come to life in clinical environments.

Keywords: radiomics, precision medicine, segmentation, population studies, statistics, machine learning

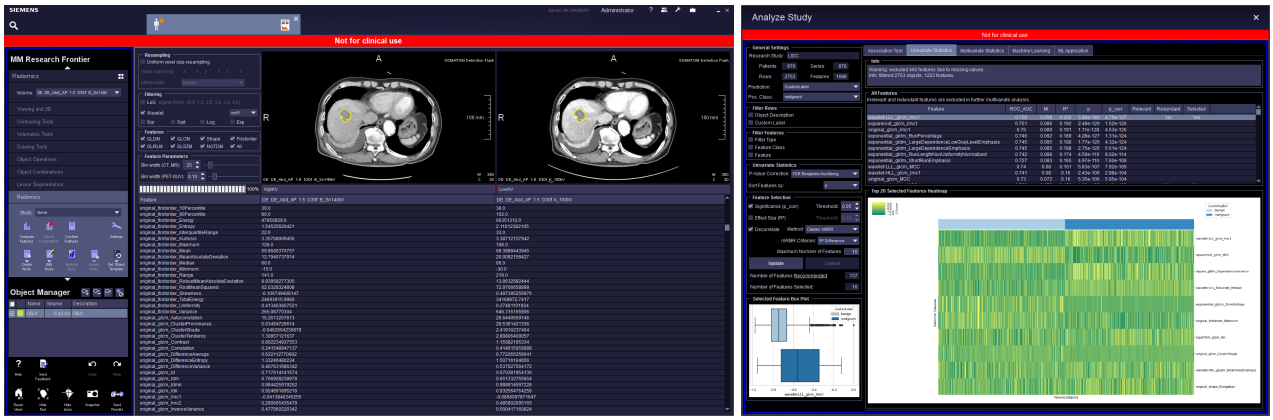
1. PURPOSE

Radiomics is still an emerging field in the domain of medical image analysis. When striving for precision medicine, therapy personalization may depend on subtle morphological differences in disease manifestation. Modern medical imaging modalities like CT or MR in many cases offer sufficient temporal and spatial resolution to capture these differences. However, they still may not necessarily be observable by the human eye. Radiomics therefore tries to represent morphology of volumes of interests (VOIs) in medical imaging data by a multitude of quantitative features encoding different morphological aspects of interest as for example texture or gross shape. The resulting representations for a cohort of patients can then be subject to data mining and predictive modeling.

The lack of convenient tools however prevents clinical researchers without the necessary scientific and technical background from carrying out such studies. Many available (research) tools are customized to a specific medical question causing significant development effort in case adaptations to other questions are necessary.

We therefore present a software solution (see Fig. 1) that supports carrying out generic radiomics research studies, which targets clinicians rather than computer scientists or statisticians. It aims to enable clinical researchers to easily prepare their data for cohort analysis related to clinical research hypotheses from various fields of medical imaging (oncology, screening, etc.). It provides universal tools for VOI definition by means of object segmentation and 3D surface creation in multiple scans sharing the same frame of reference. Furthermore, automatically created segmentations, e.g., from other tools, can be loaded as VOIs, too. The tool helps to create and manage cohort studies combining radiomic features from several patients with several VOIs and provides tools to statistically analyze these studies. The choice of statistical analysis tools has been driven by the endeavor to make the interaction as intuitive as possible and to avoid exposing too many parameters to the user. Where possible appropriate defaults have been set.

*A MeVisLab-based [1] implementation of our solution will be available as part of the Siemens Healthineers *syngo*.via Frontier research extension.



(a) (b)

Figure 1. The user interface of our software solution for radiomics computation from VOIs with the groups of segmentation tools to the left in the control area (a) and the user interface for cohort study analysis (b).

2. METHODS

2.1 General Radiomics Workflow

Our method and the associated software solution support the following general radiomics workflow (see Fig. 2(a)): i) loading of one or more image series sharing the same frame of reference, ii) intuitive creation and editing of VOIs by multiple semi-automatic tools for pathologic/anatomical image entities that are of interest for radiomic analysis, iii) radiomic feature computation for all series and all VOIs, iv) attachment of computed radiomics to existing or newly created cohort studies, and v) statistical and machine learning-based analysis of cohort studies including the preparation of result tables and plots.

2.2 Segmentation Tools

The following segmentation tools can be used for cohort study collection: contouring tools (Livewire [2], spline-based, and polygonal) with smart or linear 3D interpolation, volumetric tools (thresholding, region growing, and punching in VRT), drawing tools (nudging and brushing) for local segmentation refinement, object operations (amongst, morphological operations, inversion, and hole filling) for global segmentation refinement, object combinations (fusion, subtraction, and intersection), and Random Walker-based [3] lesion segmentation for generic lesions with adapted versions for solid and subsolid lung lesions, liver lesions, and lymph nodes [4].

2.3 Radiomics

For computation of radiomic features from VOIs and image series our prototype interfaces with the PyRadiomics library [5] in a similar manner to 3D Slicer's [6] Radiomics plugin. The library offers a variety of options to customize image pre-processing immediately before feature extraction (Laplacian of Gaussian filtering, wavelet filtering, and non-linear intensity transforms) and feature extraction itself. The supported feature classes contain first order, shape, and a multitude of texture encoding features. By relying on a public domain library an important step towards standardization of radiomics and towards interoperability and comparability of features and results has been taken.

In the case of multi-series cohort studies there is the possibility to concatenate features originating from different image series but describing the same VOI to a single VOI-specific feature vector. This is accomplished by assigning series data roles while collecting a cohort.

2.4 Hierarchical Clustering

In the context of unsupervised machine learning our solution offers z-score-based hierarchical clustering with conjoined dendrogram creation. Clustering can be customized with regards to the clustering method and the associated distance metric. [7, 8] The significance of the association between the clustering results and a user-selected label per radiomic specimen is assessed by a χ^2 test.

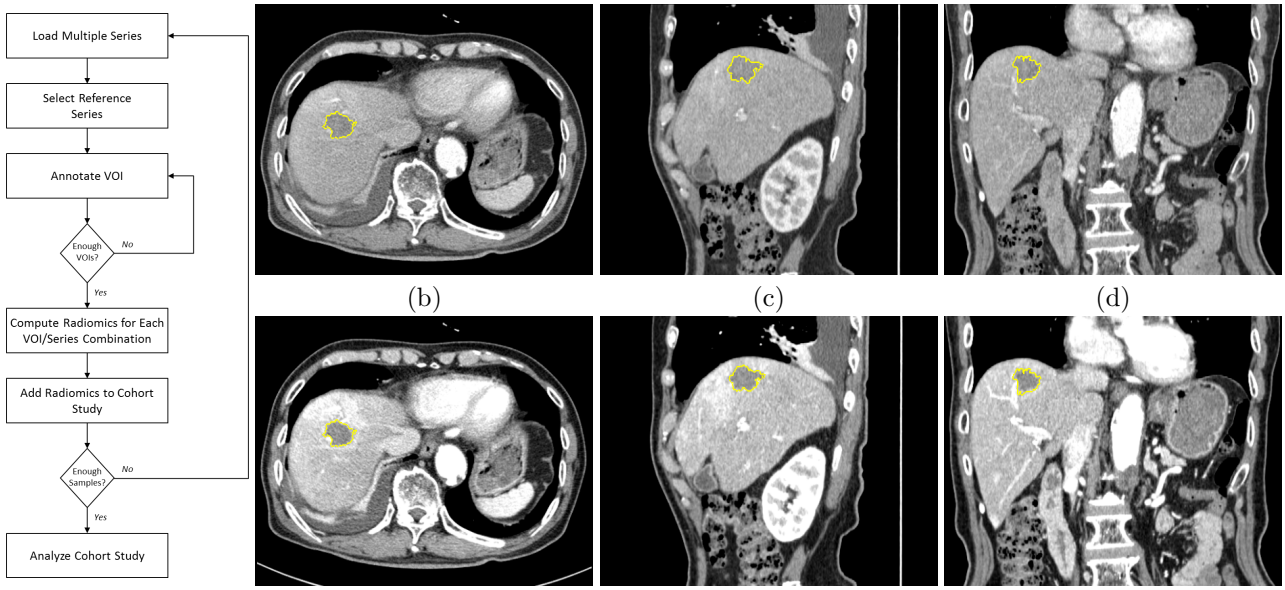


Figure 2. The general workflow of the proposed solution, which allows VOIs and radiomics to be successively defined and computed in order to create larger cohort studies for further analysis (a), and a semi-automatically created liver lesion segmentation overlaid to the anatomical reference planes of a 140 kV (b–d) and a 100 kV (e–g) series of an arterial phase liver dual-energy CT scan.

2.5 Univariate Statistics and Feature Selection

Using univariate statistics, all considered radiomic features can be ranked according to their explanatory power with regards to a user-selected label. The used sorting criterion is configurable. Amongst other things, the p-value or the R^2 value can be chosen for this purpose. If desired p-value correction to cope with the multiple comparison problem [9] can be applied. Besides, other indices like for example if a feature can be considered relevant and not redundant are computed as well.

Feature selection for subsequent multivariate statistics is embedded in the processing: appropriate thresholds for feature relevance can be defined and all features satisfying these relevance criteria will either be immediately forwarded to multivariate analysis or considered for subsequent optional decorrelation. In case of decorrelation there are two different methods: in the first one the maximum number of features to be forwarded can be defined. Then a fast minimum redundancy maximum relevance (mRMR) algorithm [10] is applied to select the proper number of features. In the second group-wise method a threshold for redundancy can be provided. In a subsequent greedy search features will be considered non-redundant and selected if they are relevant and their mutual information with the previously selected feature is lower than the provided redundancy threshold. As a best practice it is recommended to obey the one in ten rule [11] and to set the described parameters of feature selection appropriately such that only a limited amount of features is selected for multivariate analysis. This will reduce the risk of generating overfitted models.

2.6 Multivariate Statistics

Multivariate statistics in the context of our solution refers to the assessment of multivariate linear predictive models derived from a cohort study for a user-selected (binary) label. During computation multiple subsets of up to a configurable number of the forwarded features are simultaneously assessed. The best model is found by best subset forward selection: starting with the best feature and iterating over all remaining features exactly that feature is added to the model, which optimizes the chosen optimization criterion (Bayesian information criterion, Akaike information criterion, or adjusted R^2 value). This process is repeated until the predefined maximum number of features is reached. The resulting subsets and their models are ranked and the best ranking subset

and its associated multivariate linear model is further detailed by a multitude of characterizing indices, e.g., R^2 value, F-statistics, Log-likelihood, and odds ratios, and by its receiver operating characteristic (ROC) curve.

2.7 Machine Learning

Last but not least a machine learning-based model in terms of a random forest (RF) can be built from the cohort study. The method of RFs has been chosen due its repeatedly proven applicability for biomarker discovery. [12, 13] The generalization capability of the model is inherently evaluated by cross-validation with a configurable number of folds. The reported metrics and the computed ROC curve resulting from these experiments therefore provide a model assessment with respect to unseen data. In addition, RF-related feature importance is reported, too.

In principle, when an RF model is built the complete radiomic feature pool of potentially weakly relevant or even irrelevant features is considered. Even if the learning algorithm used is relatively robust to the presence of irrelevant features, processing these requires additional computational resources and may often even cause a decrease in the resulting predictive performance. Besides, high dimensional feature vectors generally imply several other undesirable effects (curse of dimensionality), and with a fixed number of training samples the predictive power even tends to reduce as the dimensionality increases (Hughes phenomenon). [14] We therefore provide the option to use the above-mentioned decorrelation-based feature selection methods within each cross-validation training/testing iteration, too. This way the inherent randomization in the RF algorithm is less likely to ineffectively fade away.

3. VALIDATION

3.1 Case Studies

The NSCLC-Radiomics collection [15] is publicly available from The Cancer Imaging Archive (TCIA) [16] and contains pretreatment CT scans of 422 non-small cell lung cancer (NSCLC) patients. There are manual 3D delineations of a radiation oncologist for parts of the collection and information about clinical outcomes. The data corresponds to the "Lung1" dataset described and analyzed in reference [17]. We use this data collection to validate the integrated cluster analysis mechanism of our solution in the context of the primary tumor stage. The data is further used to assess our solution's multivariate and machine learning-based modeling capabilities with respect to predicting the patients' two year survival.

The second collection used is an imaging study by the Lung Image Database Consortium (LIDC) [18,19] with diagnostic and lung cancer screening CT scans. It is also available on TCIA [16]. The data collection includes volumetric lesion annotations and a 5-stage malignancy rating (from benign to malignant) per lesion. We included 2,753 distinct lesions from 870 patients in our case study. With regards to binarization of the malignancy rating and annotator choice we followed the practice by Griethuysen et al. [5] but included considerably more lesions in our analysis. We use the LIDC data collection to validate our clustering approach, the univariate and multivariate statistics approaches, and the machine learning-based analysis in the use case of predicting tumor malignancy.

We have chosen to use these two data collections for validation purposes as they can be considered representative of data collections that may be subject to integrated radiomic analysis in our software solution.

3.2 Case Study: NSCLC-Radiomics

Clustering (see Fig. 3(a)) on the NSCLC-Radiomics collection reveals significant association between 3 identified radiomic phenotype subgroups in the set of 317 patients with tumor outlines and the primary tumor stage of the patients (T-stage; $p \approx 2.98 \cdot 10^{-10}$). The result points in the same direction as the findings of Aerts et al. [17] who analyzed all 422 patients with a PyRadiomics predecessor yielding 440 features per patient. With a comparable choice of feature classes (first order, shape, and texture) and identical pre-processing (wavelet decomposition) PyRadiomcs produces 529 features per patient in our case.

On the subject of two year survival our processing pipeline (with fast mRMR decorrelation) created a multivariate statistical model consisting of one texture feature and one first order feature from different wavelet filter passes of the original imaging data (wavelet-LHL_glcm_Idmn and wavelet-HLL_firstorder_TotalEnergy). As

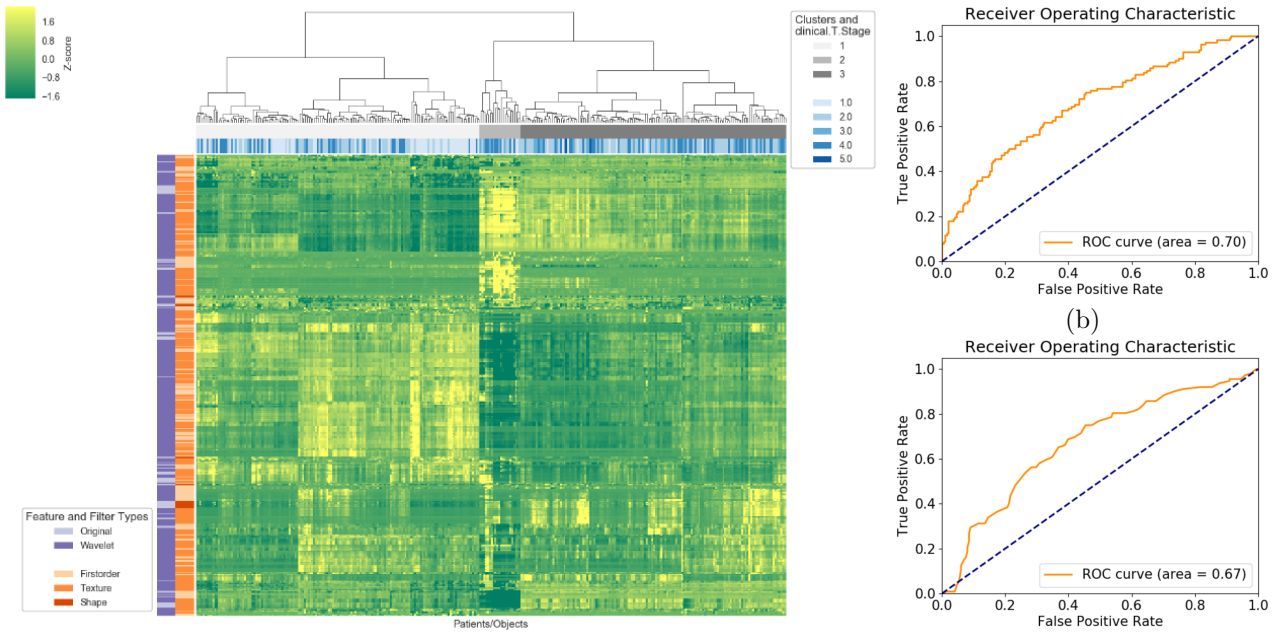


Figure 3. Clustering results for 317 patients of the NSCLC-Radiomics collection with 529 features (a) and the ROC curves of a multivariate linear prediction model (b) and a machine learning-based prediction model (c) for two year survival from the NSCLC-Radiomics data collection (b).

depicted in Fig. 3(b) it achieves an ROC area under the curve (AUC) of 0.70. Its machine learning-based counterpart, a 10-fold cross-validated RF consisting of 100 trees measuring split quality by the Gini impurity, reaches 0.67 (see Fig. 3(c)).

3.3 Case Study: LIDC

In the case of the LIDC data collection four phenotype subgroups could be found by clustering, which are significantly associated with the lesions' malignancy rating ($p \approx 8.47 \cdot 10^{-96}$)). From Fig. 4(a) it appears as if clusters 1 and 2 well separate benign and malignant lesions while clusters 3 and 4 are each dominated by one of the possible classifications. This result may motivate a 4-stage likelihood rating of lesion malignancy based on cluster affiliation in clinical practice.

Feature selection and univariate statistics on the LIDC data collection (with settings producing 1,223 radiomic features) yield that some texture-based features where wavelet_LLL_glcm_Imc1 ranks first with $R^2 = 0.206$ carry more discriminative information about malignancy than the shape feature that can be considered equivalent with the well-known RECIST criterion (original_shape_Maximum2DDiameterSlice; $R^2 = 0.133$).

Multivariate linear predictive modeling (with fast mRMR decorrelation) yields a model consisting of seven radiomic features that reaches $R^2 = 0.260$ and an ROC AUC of 0.794 for predicting malignancy (see Fig. 4(b)). An RF classifier with settings as above achieves an ROC AUC of 0.804 for the same purpose.

4. CONTRIBUTIONS

We presented an integrated software solution for creating, collecting, managing, and analyzing radiomic patient cohorts from multimodal 3D medical imaging data that may pave the way for radiomics-related precision medicine to clinical routine. Our solution offers means to semi-automatically define VOIs with a multitude of tools easing radiomic case preparation. Cohort studies can be collected little by little and immediately analyzed from within the software solution bypassing the need for additionally statistics tools. Our solution is generic in nature and not tied to any particular medical question and applicable for multiple use cases. It interfaces with a standard

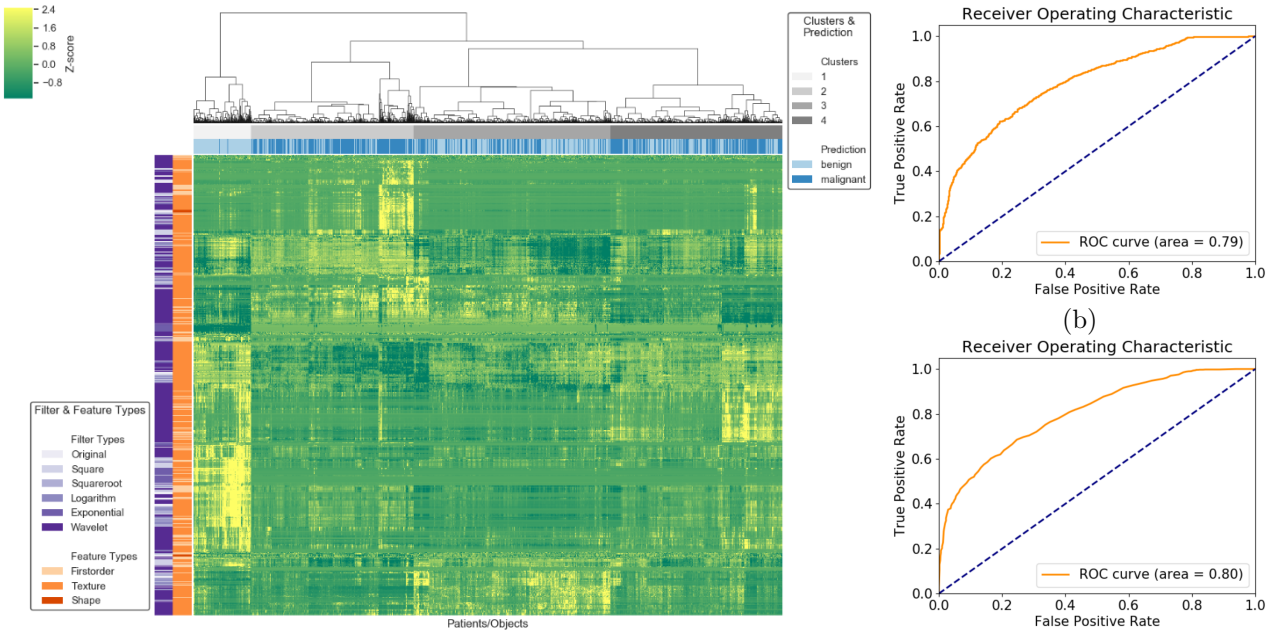


Figure 4. Clustering results for the 2,753 lesions of the LIDC collection with 1,223 features (a) and the ROC curve of a multivariate linear prediction model (b) and a machine learning-based prediction model (c) for malignancy from the LIDC data collection (b).

library for radiomic feature computation in order to make scientists achieve comparable and reproducible results. We have shown in real-world case studies that our software solution is capable of generating meaningful results in the context of radiomics-based biomarker research. We strongly believe that its integrated character will considerably lower the burden for interested clinical researchers to contribute to the exciting field of radiomics.

5. LIMITATIONS AND CONCLUSIONS

Radiomic features, particularly first order and texture features, may be very sensitive to technical variation in image acquisition. To better focus on biological variation associated parameters have to be either kept consistent during cohort study collection or the data needs to be standardized by appropriate techniques that are outside of the scope of this paper.

When trying to apply radiomics-based techniques to clinical use cases it will be advantageous if the imaging data contains visible evidence of the classification target. It may be of qualitative nature only in the first place but has the chance to be quantified and validated by the radiomics approach.

As mentioned above we tried to introduce another layer of abstraction encapsulating the diversity of underlying statistical and machine learning-based methods. As both disciplines are complex fields of science on their own this is compulsorily accompanied by shortcomings in this or that aspect: in our solution multi-class classification targets are for example only supported via the detour over one-hot encoding and regression is not supported at all. Design choices were driven by spread, availability, and reasonable computational complexity of methods, which may have excluded cutting-edge approaches.

Despite these limitations we are confident that our solution helps promoting radiomics-based research in further clinical use cases like for instance lung COPD [20], liver cirrhosis [21] or degenerated muscle tissue [22].

6. ACKNOWLEDGEMENT

The authors acknowledge the National Cancer Institute and the Foundation for the National Institutes of Health, and their critical role in the creation of the free publicly available LIDC/IDRI Database used in this study.

REFERENCES

[1] J. Rexilius, W. Spindler, J. Jomier, M. Koenig, H. Hahn, F. Link, and H. O. Peitgen, "A framework for algorithm evaluation and clinical application prototyping using ITK," in *The Insight Journal - 2005 MICCAI Open-Source Workshop*, (Palm Springs, CA, USA), Oct 2005.

[2] A. X. Falcão, J. K. Udupa, S. Samarasekera, S. Sharma, B. E. Hirsch, and R. de A. Lotufo, "User-steered image segmentation paradigms: Live wire and live lane," *Graphical Models and Image Processing* **60**, pp. 233–260, Aug 1998.

[3] L. Grady, "Random walks for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**, pp. 1768–1783, Nov 2006.

[4] J. H. Moltz, L. Bornemann, J. M. Kuhnigk, V. Dicken, E. Peitgen, S. Meier, H. Bolte, M. Fabel, H. C. Bauknecht, M. Hittinger, A. Kießling, M. Pusken, and H. O. Peitgen, "Advanced segmentation techniques for lung nodules, liver metastases, and enlarged lymph nodes in CT scans," *IEEE Journal of Selected Topics in Signal Processing* **3**, pp. 122–134, Feb 2009.

[5] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. W. L. Aerts, "Computational radiomics system to decode the radiographic phenotype," *Cancer Research* **77**, pp. e104–e107, Nov 2017.

[6] R. Kikinis, S. D. Pieper, and K. G. Vosburgh, *3D Slicer: A Platform for Subject-Specific Image Analysis, Visualization, and Clinical Support*, pp. 277–289. Springer, New York, NY, Jan 2014.

[7] D. Müllner, "fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python," *Journal of Statistical Software* **53**, pp. 1–18, May 2013.

[8] Z. Bar-Joseph, D. K. Gifford, and T. S. Jaakkola, "Fast optimal leaf ordering for hierarchical clustering," *Bioinformatics* **17**, pp. S22–S29, June 2001.

[9] Y. Benjamini, "Simultaneous and selective inference: Current successes and future challenges," *Biometrical Journal* **52**, pp. 708–721, Nov 2010.

[10] S. Ramírez-Gallego, I. Lastra, D. Martínez-Rego, V. Bolón-Canedo, J. M. Benítez, F. Herrera, and A. Alonso-Betanzos, "Fast-mRMR: Fast minimum redundancy maximum relevance algorithm for high-dimensional big data," *International Journal of Intelligent Systems* **32**, pp. 134–152, Feb 2017.

[11] F. E. Harrell, K. L. Lee, R. M. Califf, D. B. Pryor, and R. A. Rosati, "Regression modelling strategies for improved prognostic prediction," *Statistics in Medicine* **3**, pp. 143–152, Apr 1984.

[12] Y. Fan, T. Murphy, J. Byrne, L. Brennan, J. Fitzpatrick, and R. Watson, "Applying random forests to identify biomarker panels in serum 2D-DIGE data for the detection and staging of prostate cancer.," *Journal of proteome research* **10**, pp. 1361–73, Mar 2011.

[13] Y. Qi, *Random Forest for Bioinformatics*, pp. 307–323. Springer, New York, NY, Jan 2012.

[14] M. Wels, B. M. Kelm, A. Tsymbal, M. Hammon, G. Soza, M. Sühling, A. Cavallaro, and D. Comaniciu, "Multi-stage osteolytic spinal bone lesion detection from CT data with internal sensitivity control," in *SPIE Medical Imaging: Computer-Aided Diagnosis*, **8315**, p. 831513, (San Diego, CA, USA), Feb 2012.

[15] H. J. W. L. Aerts, E. R. Velazquez, R. T. H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, ..., and P. Lambin, "Data from NSCLC-Radiomics." `http://doi.org/10.7937/K9/TCIA.2015.PF0M9REI`.

[16] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior, "The cancer imaging archive (TCIA): Maintaining and operating a public information repository," *Journal of Digital Imaging* **26**, pp. 1045–1057, Dec 2013.

[17] H. J. W. L. Aerts, E. R. Velazquez, R. T. H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M. M. Rietbergen, C. R. Leemans, A. Dekker, J. Quackenbush, R. J. Gillies, and P. Lambin, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature Communications* **5**, Aug 2014.

[18] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. A. Kazerooni, H. MacMahon, E. J. R. van Beek, D. Yankelevitz, *et al.*, "The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans," *Medical Physics* **38**, pp. 915–931, Feb 2011.

[19] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, ..., and L. P. Clarke, "Data from LIDC-IDRI." `http://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX`.

[20] E. A. Regan, J. E. Hokanson, J. R. Murphy, B. Make, D. A. Lynch, T. H. Beaty, D. Curran-Everett, E. K. Silverman, and J. D. Crapo, "Genetic epidemiology of COPD (COPDgene) study design," *COPD: Journal of Chronic Obstructive Pulmonary Disease* **7**, pp. 32–43, Feb 2010.

[21] M. Hayashi, O. Matsui, K. Ueda, Y. Kawamori, M. Kadoya, J. Yoshikawa, T. Gabata, T. Takashima, A. Nonomura, and Y. Nakanuma, "Correlation between the blood supply and grade of malignancy of hepatocellular nodules associated with liver cirrhosis: evaluation by CT during intraarterial injection of contrast medium," *American Journal of Roentgenology* **172**, pp. 969–976, Apr 1999.

[22] A. Mühlberg, O. Museyko, V. Bousson, P. Pottecher, J.-D. Laredo, and K. Engelke, "Three-dimensional distribution of muscle and adipose tissue of the thigh at CT: Association with acute hip fracture," *Radiology* **Epub ahead of print**, p. 181112, Nov 2018.