Selection of Radiomics Features based on their Reproducibility

Marta Ligero†, Guillermo Torres*, Carles Sanchez*, Katerine Diaz-Chito*, Raquel Perez†, and Debora Gil*‡
*Computer Vision Center and Computer Science Department at Universitat Autònoma de Barcelona (UAB).
† Vall d'Hebron Institute of Oncology (VHIO). Barcelona.
‡ Serra Húnter fellow.

***Abstract*—Dimensionality reduction is key to alleviate machine learning artifacts in clinical applications with Small Sample Size (SSS) unbalanced datasets. Existing methods rely on either the probabilistic distribution of training data or the discriminant power of the reduced space, disregarding the impact of repeatability and uncertainty in features.**

In the present study is proposed the use of reproducibility of radiomics features to select features with high inter-class correlation coefficient (ICC). The reproducibility includes the variability introduced in the image acquisition, like medical scans acquisition parameters and convolution kernels, that affects intensity-based features and tumor annotations made by physicians, that influences morphological descriptors of the lesion.

For the reproducibility of radiomics features three studies were conducted on cases collected at Vall Hebron Oncology Institute (VHIO) on responders to oncology treatment. The studies focused on the variability due to the convolution kernel, image acquisition parameters, and the inter-observer lesion identification. The features selected were those features with a ICC higher than 0.7 in the three studies.

The selected features based on reproducibility were evaluated for lesion malignancy classification using a different database. Results show better performance compared to several state-of-the-art methods including Principal Component Analysis (PCA), Kernel Discriminant Analysis via QR decomposition (KDAQR), LASSO, and an own built Convolutional Neural Network.

***Index Terms*—Feature Selection, Reproducibility, Radiomics**

I. Introduction

Radiomics is a recent discipline that uses sophisticated image analysis and machine learning tools to obtain quantitative image-based features (signatures) that correlate to final diagnosis and treatment outcome [1]. Like most clinical applications, radiomics must deal with Small Sample Size (SSS) and minority classes in possibly unbalanced settings. Most matching learning methods are ill-posed under such conditions and might drop their performance [2]. Dimensionality reduction and automatic feature selection tools have been crucial to mitigate the curse of dimensionality and SSS inherent to classification.

Principal Component Analysis (PCA) is an unsupervised method that uses a linear orthogonal transformation to project features into a dimensionally reduced set of uncorrelated variables called principal components. This technique transforms the data in a reduced dimension but does not perform feature selection. The main problem of this technique is the loss of interpretation of the variables [3]. Partial Least Square - Discriminant Analysis (PLS-DA) PLS-DA is a supervised classification method based on Partial Least Squares Regression and Linear Discriminant Analysis. In this case the technique performs both dimensionality reduction and classification. The dimensionality reduction is similar to PCA but the new components are created by projecting the variables and the outcome into a new space based on linear regression models. The variables of the new reduced subspace called latent variables can predict the outcome and, unlike PCA, there is an interpretation of the projected variables given that the importance of the original variables in the new subspace is quantified [4].

Kernel Trick (KT) has been widely used to extend the above linear methods to the nonlinear case. Kernel methods have aroused great interest in the last decade since they are universal nonlinear approximations and facilitate solving complex problems where the samples are not linearly separable as is the case of many machine learning and pattern recognition application. Kernel methods use nonlinear mapping to project samples from the original space to a feature space where the samples are expected to be easily separable using linear approaches [5].

A variety of subspace-based kernel methods have been proposed, including Kernel Principal Component Analysis (KPCA) and Kernel Discriminant Analysis (KDA) [6]. A Kernel-Independent Component Analysis (KICA) [7] by using the KT and the Infomax algorithm has also been proposed for enhancing classification, but its application is limited to classes statistically independent. KDA-based approaches are better suited for supervised classification applications since a similar supervision process is performed during the dimen-

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 712949 (TECNIOspring PLUS) and from the Agency for Business Competitiveness of the Government of Catalonia.

Work supported by Spanish Projects FIS-G64384969, Generalitat de Catalunya, 2017-SGR-1624 and CERCA-Programme. D.Gil is a Serra Húnter fellow.

In this research Guillermo Torres is supported by UAB.

The Titan X Pascal used for this research was donated by the NVIDIA Corporation.

sionality reduction, but they require solving an expensive optimization problem [8]. Efficient KDA approaches have been proposed as a solution such as the Kernel Discriminant Analysis via QR decomposition (KDAQR) [9], based on the QR decomposition to replace the costly eigen decomposition of the kernel matrix, and the Kernel Discriminant Analysis by using Spectral Regression (KDASR) [8] which combines spectral graph analysis and regularised regression. Finally, a Discriminative Common Vector with Kernel (KDCV) was originally proposed by Cevikalp et al. [10], [11] and extended (Kernel Generalized Discriminative Common Vectors, KGDCV) to manage large dimensional data in [5].

Methods for dimension reduction in classification problems rely on the probabilistic distribution of samples and, thus, might not be the best suited for SSS in unbalanced settings. Furthermore, the features projected in the reduced spaces are computed following probabilistic considerations and are not easy to be clinically interpreted. In the context of clinical applications (especially in radiomics for personalized medicine), feature selection methods are a preferred choice. Several methods for feature selection are applied in the field of predictive models for personalized medicine.

Random forest selects features according to the change in the classification error. Although it is accurate in case of highly uncorrelated data, in radiomic multi-view problems variables are highly correlated and their selection usually leads to a correlated subset of variables that can produce over-fitting [12]. Besides, the selection of the subsets is random and there is a lot of variability when applying the technique.

To avoid correlation and over-fitting, Minimum Redundancy Maximum Relevance (mRMR) algorithm [13] selects features according to the Mutual Information (MI) between the set of features and the class variable outcome. Features are selected by a threshold on MI which must be carefully adjusted to avoid inclusion of redundant variables or the elimination of clinical relevant ones.

The least absolute shrinkage and selection operator (LASSO) [14] uses a logistic regression model with a penalty term to select features according to their significance in class variable prediction. This method is quite popular for the definition of radiomic signatures [15] and malignancy classification [16]. However, there are some limitations like the low repeatability and reproducibility of textural features in the clinical setting and the limitation of the method to properly modelling SSS unbalanced problems. We consider this could be corrected by the introduction of uncertainty measures into predictive models and the filtering of most unstable data in the training stage.

None of the above methods considers feature reproducibility for their selection. In the process of clinical data collection, there are several factors prone to introduce variability in multi-view features. Among others, the main ones are medical scans acquisition parameters with an impact on intensity-based values and inter-observer variability in manual annotations required to identify tumors with an impact on shape and volumetric descriptors [17]. Such sources of variability introduce an uncertainty in models that should be considered to issue more reliable reproducible predictions, while avoiding over-fitting.

A recent work [15] adds scan acquisition parameters as fixed factors in a regression model for the development of a radiomic signature that predicts immunotherapy response. However, being unable to select which radiomic features were most affected, the signature reproducibility was low due to over fitting. In [17] it is reported that method reproducibility increases if features are selected based on their stability and reproducibility. Latest methods [18] based on fully connected convolutional networks use dropout as boosting method to define a measure of uncertainty in semantic classifiers output that it is used as post-segmentation filtering. However, up to our knowledge a selection of features based on reproducibility and uncertainty remains unexplored.

In this work we conduct a study with data from Vall Hebron Oncology Institute (VHIO) that analyzes the reproducibility of radiomics features against different image acquisition conditions and inter-observer variability in lesion identification. Our reproducibility study bases on the correlation of feature values obtained from data collected using different conditions and settings. In this study, features were selected as reproducible if they had high inter-class correlation coefficient (ICC) for all sources of variability. The performance of the selected features was compared to state-of-art methods on a different public data base. Results obtained for the classification of lesion malignancy show the better performance of our selection based on reproducibility.

II. Reproducibility of Radiomics Features

Based on the studies [17] that demonstrated the influence of image acquisition parameters (specially the convolution kernel) and inter-observer variability on radiomics features reproducibility, three test-retest reproducibility studies were done: 1) Identification of similar convolution kernels; 2) identification of robust radiomics features based on image acquisition parameters variability; and 3) identification of robust radiomics features based on inter-observer variability.

The selection of radiomics features based on their reproducibility was conducted on cases acquired at Vall Hebron Oncology Institute (VHIO) collected for a study on responders to oncologic treatments. Contrast-enhanced CT scans were acquired on 16- or 64-channel CT scanner during a clinical trial performed at Hospital Quironsalud Barcelona. Given the nature of the retrospective study, images were obtained from different scanners (SIEMENS, GE Systems and Philips) with different convolution kernels according to the manufacturer. CT scans of the same patient were acquired 2 months apart, from a population set of 32 patients, 17 (mixed population) of whom presented changes in convolution kernels and 15 of whom (control population) presented standardized image acquisition parameters. Table I summarizes the image acquisition parameters and number of cases of VHIO-Quironsalud database.

TABLE I
DESCRIPTION OF THE VHIO-QUIRONSALUD DATABASE USED FOR THE REPRODUCIBILITY STUDY.

	Image Acquisition Parameters		
	Manufacturer	**Convolutional Kernels**	**kVp**
Control population n=15	SIEMENS	B20f/B20s	120
Mixed population n=17	SIEMENS	B20f/B20s B30f/B30s B31s, B41s, I31s	100 110 120 130
	PHILIPS	B/IMR1	120
	GE Systems	STANDARD/ LUNG	120
Standardized population n=23	SIEMENS	B20f/B20s B30f/B30s B31s/B41s	100 110 120 130
	PHILIPS	B	120

For each scan, 105 radiomics features were extracted using PyRadiomics [19], an open-source python package for the extraction of Radiomics features from medical imaging volumes. PyRadiomics features include shape features, first order features, and textural features describing several aspects of the lesion. The lesion was defined by a volume mask manually delineated by experts. Shape features were computed to obtain morphological measures of the lesion mask. First order features were calculated to obtain a description of the distribution of voxel intensities of within the image region defined by the mask. Finally, textural features were computed to obtain 3D patterns of voxel intensity to describe the interrelationships of gray-levels values within image region constrained by the mask. PyRadiomics texture features include Gray Level Co-ocurrence Matrix (GLCM), Gray Level Size Zone (GLSZM), Gray Level Run Length Matrix (GLRLM)and Gray Level Dependency Matrix (GLDM) [20].

To conduct the first two studies related to scanner parameters, a solid organ not affected by the disease, the spleen, was segmented. To avoid inter-observer variability the whole spleen was segmented. To standardize the population based on convolution kernels, Bland Altman and ICC were performed on different pairs of convolution kernels to find those kernels with significant changes in radiomics features. Determining 0.7 as a cutoff for high correlation on interclass correlation coefficient (ICC), the control population showed that 62 features from 105 presented high correlation vs the non-standardized mixed population that only presented one feature. The difference in ICC between the two populations was substantial with inter quartile ranges equal to $[0.48, 0.88]$ for control population and $[0.18, 0.42]$ for the mixed one. Convolution kernels B from Philips and B20f/B20s, B30f/B30s, B41f and B31f from SIEMENS presented higher interclass correlation coefficient between them than with convolution kernels of I31f, B26f from SIEMENS and LUNG, STANDARD from GE Systems. The latter group presented significant differences after applying Wilcoxon test and those patients containing these convolution kernels were removed from the population set.

From the first reproducibility study, some convolution kernels (CK) were identified as comparable. In the second study, a standardized population of 23 patients was selected, 15 of which had two images with exact same image acquisition parameters (Identical CK population) and 8 of which had comparable image acquisition parameters (Similar CK population). Each radiomic feature was compared separately from the Identical CK population and Similar CK population. Therefore, two ICC were obtained for each radiomic feature. The values for the minimum of the two ICC was comparable to the ICC for the control population with an inter quartile range equal to $[0.43, 0.74]$. A cut-off 0.7 applied to this minimum ICC for the standardized population selected 29 features of 105.

The third reproducibility study was done with the tumor ROIs delineated by two experienced radiologist from a randomly selected population of 83 patients (inter-observer population) obtained using the comparable CKs selected in the first study. As in the previous studies the radiomics features from the ROI were extracted and compared among radiologists' segmentations using ICC score.

Finally, our reproducibility feature selection chose those features that present ICC higher than 0.7 in the three studies: comparable CKs, image acquisition parameters and inter-observer variability in lesion delineation. The 26 features selected by our reproducibility criteria together with the 3 ICC measures obtained for each study are reported in Table II.

III. EXPERIMENTS

The goal of our experiments is to validate the generalization capability of our selection based on reproducibility. We address to what extend our selection also provides: 1) the most informative features for radiomics issues; 2) features with high potential to perform independently of the data base.

To assess the above points, we have used the features in Table II to classify the malignancy of pulmonary nodules on cases selected from the Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI) public databases [21]. The LIDC/IDRI Database contains 1018 cases/patients, each of which includes images from a clinical thoracic CT scan and an associated XML file that records the results of a two-phase image annotation (including lesion delineation) process performed by four experienced thoracic radiologists. From these patients 7371 lesions were marked as "nodule". We chose a subset of 86 patients with unambiguous diagnosis available for some of their nodules along with patient final diagnosis (malign or benign). Since for patients with multiple nodules, there is not a correspondence between nodule diagnosis and experts' lesion delineation, data was split to ensure unambiguous labels for training and test. A total number of 52 patients with only one nodule were selected as training set, whereas the remaining 36 cases with more than one nodule were selected as test set for patient

TABLE II
FEATURES SELECTED ACCORDING TO REPRODUCIBILITY

Variable	ICC 1	ICC 2	ICC 3
Firstorder Entropy	0.898	0.760	0.726
Firstorder TotalEnergy	0.884	0.847	0.999
Firstorder Uniformity	0.893	0.762	0.852
GLCM Id	0.835	0.830	0.839
GLCM Idm	0.840	0.853	0.835
GLCM JointEnergy	0.878	0.745	0.931
GLCM JointEntropy	0.886	0.813	0.788
GLCM MaximumProbability	0.739	0.815	0.953
GLDM Dependence Non Uniformity Normalized	0.719	0.822	0.719
GLDM Dependence Variance	0.778	0.809	0.990
GLDM Large Dependence Emphasis	0.810	0.787	0.986
GLRLM Gray Level Non Uniformity Normalized	0.899	0.709	0.788
GLRLM Run Length Non Uniformity Normalized	0.830	0.846	0.922
GLRLM Run Percentage	0.808	0.825	0.963
GLRLM Short Run Emphasis	0.715	0.752	0.946
Shape Elongation	0.965	0.801	0.944
Shape Flatness	0.886	0.772	0.963
Shape Least Axis	0.948	0.921	0.997
Shape Major Axis	0.964	0.718	0.997
Shape Max 2D Diameter Column	0.843	0.725	0.991
Shape Max 2D Diameter Slice	0.967	0.973	0.998
Shape Minor Axis	0.980	0.872	0.997
Shape Sphericity	0.861	0.907	0.834
Shape Surface Area	0.969	0.816	0.989
Shape Surface Volume Ratio	0.943	0.830	0.982
Shape Volume	0.961	0.759	0.999

diagnosis. The classification metrics were precision, recall and AUC over nodule diagnosis for the training set and patient diagnosis for testing.

Our feature selection based on reproducibility was compared to unsupervised (PCA and KPCA) and supervised (LDA and its kernel variants) methods for dimensionality reduction, as well as, to LASSO feature selection. We also implemented our own CNN architecture to obtain a new and best representation of the original PyRadiomics features by using the generalization power of the CNN layers. Our CNN architecture is composed of 4 convolutional blocks -each one containing a convolution layer followed by a ReLU and a Max Pooling- and the last layer being a fully connection with features ranging from 2 to 10. In all cases, we used a 12 batch size and 20 epochs to train the models with dropout in all convolutional blocks. All these methods were trained on the 52 patients from LIDC/IDRI Database having only one nodule, while our approach used the 26 features learned from VHIO database.

TABLE III
CLASSIFICATION METRICS FOR PATIENT DIAGNOSIS

	Precision		Recall		AUC
	Malign	Benign	Malign	Benign	
ICC	81.8	100	100	33.3	0.80
LASSO	74.3	0	96.3	0	0.77
PCA	75	0	100	0	0.74
KDAQR	75	0	100	0	0.61
CNN	74.3	0	96.3	0	0.75

The reduced sub-space of features provided by each method was the input of a logistic regression model trained to classify between benign and malign nodules on the 52 cases having a single nodule. Taking into account that the test set contains more than one nodule per patient we chose the most pessimistic diagnosis as final patient diagnosis. This way a test patient was considered to have cancer (malign class) if one of its nodules was classified as malign by the model. Regarding the patient-sensitive probability required for AUC computation, we took the average probability for all patient nodules.

Table III reports results for the proposed method (labelled ICC) and best performers from the other methods considered. These methods, selected as the ones with higher AUC in training data, were the following: PCA, KDAQR, LASSO and CNN with 3 features (labelled CNN). Except the proposed ICC, all methods have 0 recall and precision for benign cases. This is due to the fact that their diagnosis was malign for all cases. It follows that their accuracy remains equal to the number of malign cases (75%) in contrast to ICC 84% accuracy. The proposed method had also the highest AUC, obtained from the ROC curves shown in figure 1.

Fig. 1 shows different ROC curves. The upper-left plot shows ROC curves on the test data for all methods. The remaining plots compare, for each of the methods, ROC curves between training and test to assess over-fitting. For these last plots, methods are grouped into supervised (LASSO, KDAQR and CNN) an unsupervised (PCA, ICC). Visual inspection of the plot comparing the performance in test data shows that, except for KDAQR, all methods perform similarly on test data. Comparison between training and test, indicates that supervised methods (LASSO and, especially, KDAQR) are prone to present over-fitting. Meanwhile, unsupervised methods and CNN present a comparable performance between training and test.

IV. CONCLUSIONS

This paper presents a feature selection strategy based on a reproducibility study conducted on an own database for prediction of cancer treatment outcome. Selection is exclusively based on the variability of features under different acquisition conditions. An interesting point of our approach is that, unlike existing methods, selection takes into account neither the distribution of training population features nor the outcome variable (treatment outcome).

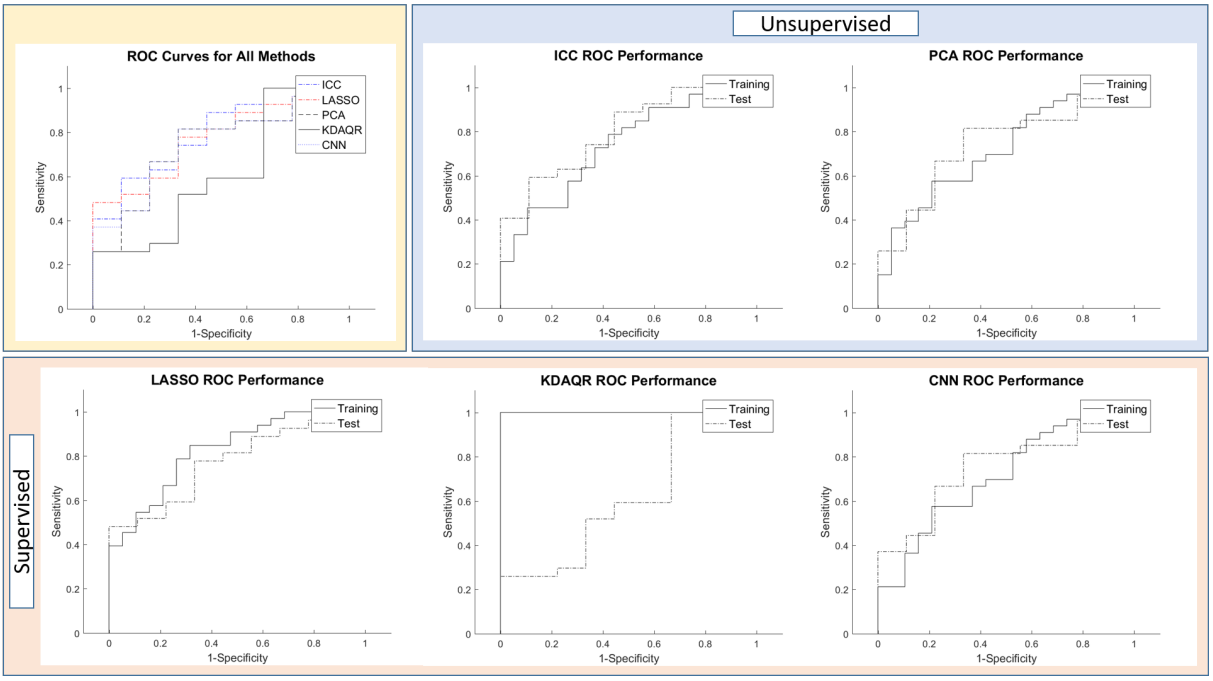

Fig. 1. ROC curves for all methods on test data (upper-left plot) and for each method on training and test.

In order to test the validity and repeatability of such a criterion, our selected features were used to classify the malignancy of pulmonary nodules on cases from LIDC/IDRI public data base. Results were compared to state-of-art methods for dimensionality reduction trained on the same cases. The proposed method is the only one presenting sensitivity to benign cases. In terms of generalization capability, comparison between training and test (ROC curves in fig.1) show that the proposed method, the unsupervised PCA and CNN are comparable and outperform supervised techniques. Since these techniques could be the ones with higher generalization power, we have, in particular, that features selected on the basis of their stability and reproducibility can also be the most informative ones.

However, it is important to remark that, unlike ICC, PCA and CNN selection was learned from LIDC/IDRI data. A main concern is whether the performance of PCA/CNN reduced set of features would be comparable applied to a different dataset and machine learning problem. We consider that our method could be a good alternative, since selection was performed on an independent data-set and, its mechanisms did not take into account neither population distribution nor any specific machine learning problem.

These promising results encourage further research to incorporate uncertainty into the training process to obtain, by their own definition, classifiers with a selection of reproducible features minimizing over-fitting. Given that CNNs provide a framework with high generalization power allowing the optimization of complex loss functions, we plan to incorporate the presented measure of repeatability into a multi-task deep learning approach.

REFERENCES

[1] P Lambin and et al, "Radiomics: the bridge between medical imaging and personalized medicine," *Nature Reviews*, vol. 12, pp. 749–53, 2017.

[2] J. P. Cohen, M. Luck, and S. Honari, "Distribution Matching Losses Can Hallucinate Features in Medical Image Translation," *arXiv preprint arXiv:1805.08841*, 2018.

[3] P. S. Gromski, H. Muhamadali, D. I. Ellis, Y. Xu, E. Correa, M. L. Turner, and R. Goodacre, "A tutorial review: Metabolomics and partial least squares-discriminant analysis–a marriage of convenience or a shotgun wedding," *Analytica chimica acta*, vol. 879, pp. 10–23, 2015.

[4] J. Lever, M. Krzywinski, and N. Altman, "Points of significance: Principal component analysis," pp. 641–642, 2017.

[5] K. Diaz-Chito, J. M. del Rincón, A. Hernández-Sabaté, M. Rusiñol, and F. J. Ferri, "Fast Kernel Generalized Discriminative Common Vectors for Feature Extraction," *Journal of Mathematical Imaging and Vision*, vol. 60, no. 4, pp. 512–524, 2018.

[6] M.-H. Yang, "Kernel eigenfaces vs. kernel fisherfaces: Face recognition using kernel methods," in *fgr*. IEEE, 2002, p. 0215.

[7] Q. Liu, J. Cheng, H. Lu, and S. Ma, "Modeling face appearance with nonlinear independent component analysis," in *null*. IEEE, 2004, p. 761.

[8] D. Cai, X. He, and J. Han, "Speed up kernel discriminant analysis," *The VLDB Journal—The International Journal on Very Large Data Bases*, vol. 20, no. 1, pp. 21–33, 2011.

[9] T. Xiong, J. Ye, Q. Li, R. Janardan, and V. Cherkassky, "Efficient kernel discriminant analysis via QR decomposition," in *Advances in neural information processing systems*, 2005, pp. 1529–1536.

[10] H. Cevikalp, M. Neamtu, and A. Barkana, "The kernel common vector method: A novel nonlinear subspace classifier for pattern recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 37, no. 4, pp. 937–951, 2007.

[11] H. Cevikalp, M. Neamtu, and M. Wilkes, "Discriminative common vector method with kernels," *IEEE Transactions on Neural Networks*, vol. 17, no. 6, pp. 1550–1565, 2006.

[12] Y. Li, F.-X. Wu, and A. Ngom, "A review on machine learning principles for multi-view biological data integration," *Briefings in bioinformatics*, vol. 19, no. 2, pp. 325–340, 2016.

[13] M. Radovic, M. Ghalwash, N. Filipovic, and Z. Obradovic, "Minimum redundancy maximum relevance feature selection approach for temporal gene expression data," *BMC Bioinformatics*, pp. 18–19, 2017.

[14] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society*, vol. 58(1), pp. 267–88, 1996.

[15] R. Sun, E. J. Limkin, M. Vakalopoulou, L. Dercle, S. Champiat, S. R. Han, L. Verlingue, D. Brandao, A. Lancia, S. Ammari *et al.*, "A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: an imaging biomarker, retrospective multicohort study," *The Lancet Oncology*, vol. 19, no. 9, pp. 1180–1191, 2018.

[16] P. T. et al, "Novel high-resolution computed tomography-based radiomic classifier for screen-identified pulmonary nodules in the National Lung Screening Trial," *PLOS ONE*, vol. 13 (10), 2018.

[17] H. J. Aerts, E. R. Velazquez, R. T. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld *et al.*, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature communications*, vol. 5, p. 4006, 2014.

[18] T. Nair, D. Precup, D. L. Arnold, and T. Arbel, "Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 655–663.

[19] J. J. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. Aerts, "Computational radiomics system to decode the radiographic phenotype," *Cancer research*, vol. 77, no. 21, pp. e104–e107, 2017.

[20] R. M. Haralick, K. Shanmugam *et al.*, "Textural features for image classification," *IEEE Transactions on systems, man, and cybernetics*, no. 6, pp. 610–621, 1973.

[21] 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 *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*, vol. 38, no. 2, pp. 915–931, 2011.