Supplementary Information

Contents

Section S1. Supplementary Methods

S1A. Datasets From TCIA

NSCLC Radiogenomics

This dataset, produced by Bakr et al. (51), consists of 211 NSCLC cases which were collected from the Stanford University School of Medicine and the Palo Alto Veteran Affairs Healthcare System. Subjects were selected from a pool of early stage NSCLC patients, prior to surgical procedure. Samples of excised tissue from a later surgical procedure were then used to obtain mutation data. The CT images we selected for our experiment were obtained from a variety of scanners, protocols, and scanning parameters: slice thickness of 0.625–3mm (median 1.5mm) and an X-ray tube current of 124–699 mA (median 220 mA) at 80–140 kVp (median 120 kVp). Scans were acquired with subjects in supine position with arms at sides, from the apex of the lung to the adrenal gland within a single breath hold.

Segmentation was provided for 144 subjects and were initially obtained using an unpublished automatic segmentation algorithm. These were edited as necessary, and reviewed by two thoracic radiologists.

CT images and segmentation were provided as Digital Imaging and Communications in Medicine (DICOM) and DICOM segmentation objects respectively, on TCIA.

TCGA LUAD and TCGA LUSC

The Cancer Genome Atlas Lung Adenocarcinoma (TCGA-LUAD) (52) and Lung Squamous Cell Carcinoma (TCGA-LUSC) (53) data collections provide clinical images to matched subjects in the Cancer Genome Atlas. The imaging dataset consisted of 69 cases for the TCGA-LUAD set and 37 for the TCGA-LUSC dataset, for a total of 106 cases. Imaging data for TCGA was collected from many sites all over the world and is extremely heterogenous in terms of scanner modalities, manufacturers, and acquisition protocols. The CT images we used in our experiment had a variety of scanning parameters: slice thickness of 1–8mm (median 3.75mm), and an X-ray tube current of 40–651 mA (median 289.5 mA) at 120–140 kVp (median 120 kVp). CT images were provided in DICOM format on the TCIA.

S1B. Open-Source Radiomics Feature Extractors

Pyradiomics

This software can be found at http://www.radiomics.io/pyradiomics.html and further documentation and feature definitions can be found at http://pyradiomics.readthedocs.io/en/latest/features.html.

All available 3D features were calculated using the same image pre-processing parameters as published in the original case study (36). The categories of features included shape, size, intensity histogram, gray-level co-occurrence matrix, gray-level size zone matrix, gray-level run-length matrix, neighboring gray-tone difference matrix, and gray-level dependence matrix. Image pre-processing filters included Laplacian of Gaussian (LoG) filter at sigma values of 1.0, 2.0, 3.0, 4.0, 5.0 mm, and a wavelet filter, using a high band-pass or low band-pass filter in x, y, and z directions, yielding 8 different combinations of decompositions. In addition, bin width was set to 25 and resampled pixel spacing to [1, 1, 1]. In total 1319 features were extracted from each segmented tumor using Pyradiomics.

Pyradiomics does not accept DICOM or DICOM RTSTRUCT as importable files and requires a third-party program to convert files to a SimpleITK accepted format. The program that is recommended on the pyradiomics documentation is plastimatch or dcm2niix. We used plastimatch (74) to convert DICOM and DICOM RTSTRUCT files to NRRD and MHA files respectively.

Imaging Biomarker Explorer (IBEX)

All available features were extracted without image pre-processing filters. There was no single best recommendation for image pre-processing filters mentioned in the original paper introducing IBEX (47), or the guidelines paper later published (75). Although Fave et. al showed that in general, Butterworth smoothing filter, either on its own or in conjunction with 8-bit depth resampling, resulted in the ability to extract statistically significant features, specific trends were noted to be feature-dependent (76). As a result, feature-specific image preprocessing may be required to maximize the usefulness of each radiomics feature. In order to minimize the amount of end-user modification and to maintain some degree of comparability with the other packages, which may or may not have these specific preprocessing filters, we decided to use IBEX in its default setting, without any image pre-processing filters. There are 134 original features available for extraction in the categories of intensity direct, intensity histogram, shape, and texture. In total, 1767 features were extracted from each segmented tumor using IBEX.

IBEX accepts DICOM and RTSTRUCT file formats for images and segmentation masks, respectively.

Columbia Image Feature Extractor (CIFE)

The 1126 radiomics features were derived from 15 feature classes developed in a previous lung radiomics study (32) by expanding the value range of feature parameters for the sake of capturing as much image information as possible. The definitions of these features and relevant references are provided in the supplementary file of reference 12 and 48. The default image pre-processing for this package is voxel resampling at 0.5 x 0.5 x 0.5 mm^3 in order to acquire uniform volumetric spacing.

Section S2. Supplementary Results

S2A. Open-Source Radiomics Feature definitions

Table S1. Definitions of selected features from Pyradiomics, IBEX and CIFE

Group	Feature	Definition
IBEX	1GaussAmplitude	-Description: Amplitude of each gaussian curve -Reference: http://www.mathworks.com/help/curvefit/gaussian.html
	LocalRangeStd	-Description: 1. First, at each voxel, compute range value(MaxValue-MinValue) in its neighborhood region. 2. Then, compute the standard deviation among all the voxel's range value calculated from 1. -Parameters: 1. NHood: Size of the neighborhood
	135-1Correlation	-Description: For the feature description, refer to the documentation on MATLAB function "graycoprops". -Reference: 1. Haralick, R.M., K. Shanmugan, and I. Dinstein, "Textural Features for Image Classification", IEEE Transactions on Systems, Man, and Cybernetics, Vol. SMC-3, 1973, pp. 610-621. 2. Haralick, R.M., and L.G. Shapiro. Computer and Robot Vision: Vol. 1, Addison-Wesley, 1992, p. 459.
	-333-4ClusterShade	-Reference: 1. L. Soh and C. Tsatsoulis. Texture analysis of sar sea ice imagery using gray level co-occurances matrices. IEEE Trans. on Geoscience and Remote Sensing, 37(2):780–795, 1999
	VoxelSize	The physical voxel size.
Pyradiomics	log-sigma-2-0-mm-3D_firstorder_Minimum	minimum = min(X), X is a set of voxels included in the ROI. Calculated under the LOG filter with sigma set to 2mm.
	log-sigma-2-0-mm-3D_glszm_SizeZoneNonUniformityNormalized	$SZNN = \frac{\sum_{j=1}^{N_s}\left(\sum_{i=1}^{N_g}\mathbf{P}(i,j)\right)^2}{N_z^2}$ SZNN measures the variability of size zone volumes throughout the image, with a lower value indicating more homogeneity among zone size volumes in the image. This is the normalized version of the SZN formula
	log-sigma-2-0-mm-3D_glcm_InverseVariance	$inverse\ variance = \sum_{k=1}^{N_g-1}\frac{p_{x-y}(k)}{k^2}$
	wavelet-HHH_glszm_SmallAreaEmphasis	$SAE = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}\frac{\mathbf{P}(i,j)}{j^2}}{N_z}$ SAE is a measure of the distribution of small size zones, with a greater value indicative of more smaller size zones and more fine textures. Calculated under the wavelet filter HHH.

	wavelet-LHL_firstorder_Skewness	$$skewness = \frac{\mu_3}{\sigma^3} = \frac{\frac{1}{N_p}\sum_{i=1}^{N_p}(\mathbf{X}(i)-\bar{X})^3}{\left(\sqrt{\frac{1}{N_p}\sum_{i=1}^{N_p}(\mathbf{X}(i)-\bar{X})^2}\right)^3}$$ Skewness measures the asymmetry of the distribution of values about the Mean value. Depending on where the tail is elongated and the mass of the distribution is concentrated, this value can be positive or negative. Calculated under the wavelet filter LHL
	wavelet-LHH_firstorder_Skewness	Same as above. Calculated under the wavelet filter LHH
CIFE	DWF_Z_H	DWF_H is the discrete wavelet filter in the high frequency domain. Z means the use of CT images reconstructed in the axial direction.
	Intensity_Skewness	$$\text{Skewness} = \frac{\frac{1}{N}\sum_{i=1}^{N}(X(i)-\bar{X})^3}{\left(\sqrt{\frac{1}{N}\sum_{i=1}^{N}(X(i)-\bar{X})^2}\right)^3}$$ Skewness is a measure of the degree of distribution symmetry. A value of zero (0) indicates that the distribution has a normal shape;
	Gabor_Max_Z	Gabor filters are linear filters designed for edge detection, which are used in image processing for feature extraction and texture analysis. Z means the use of CT images reconstructed in the axial direction.
	Intensity_Minimum	The minimum intensity value within the segmented tumor.

Feature definitions of IBEX are obtained from within the help section of each feature from the GUI. Feature definitions of Pyradiomics are obtained from the online documentation.

S2B. Distribution of feature values between EGFR Wildtype and Mutant subgroups

A depiction of the feature value distributions for a select feature from each extractor. Each boxplot represents the distribution of feature values for that subset of a cohort, e.g. the wildtype cases of the training cohort. Comparisons are made using the Wilcoxon Rank Sum or Mann-Whitney test. The bottom row of p-values on each figure indicate the comparison between mutant and wildtype subsets of the same cohort, in which we set a $p < 0.05$ to indicate that these distributions are unequal and distinguishable.

IBEX features

Figure S2.1 IBEX: Gauss Amplitude

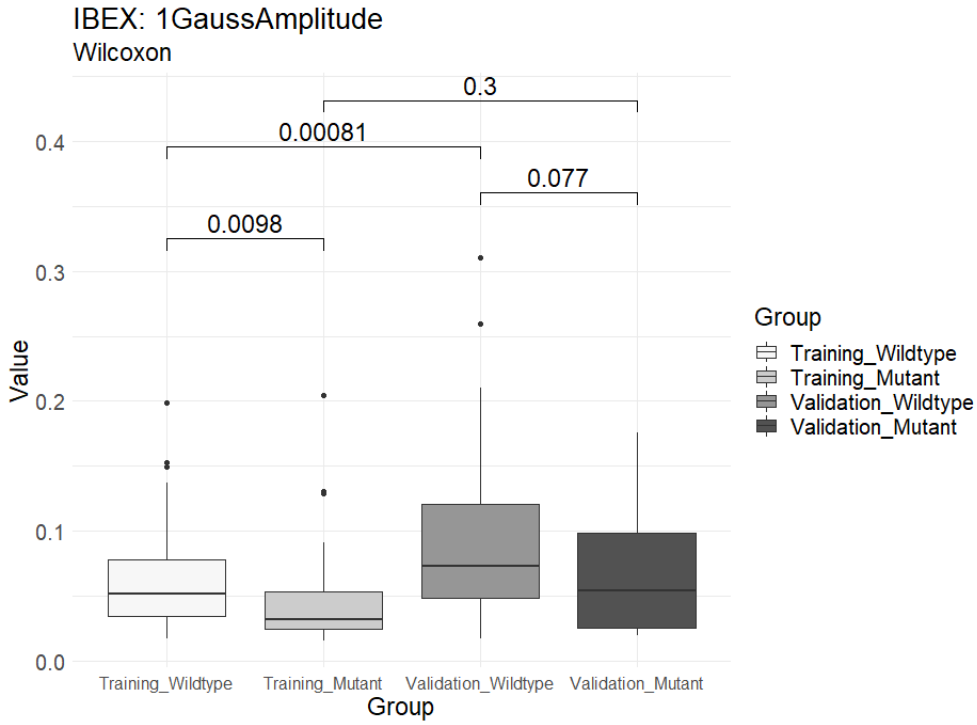


Figure S2.2 IBEX: LocalRangeStd

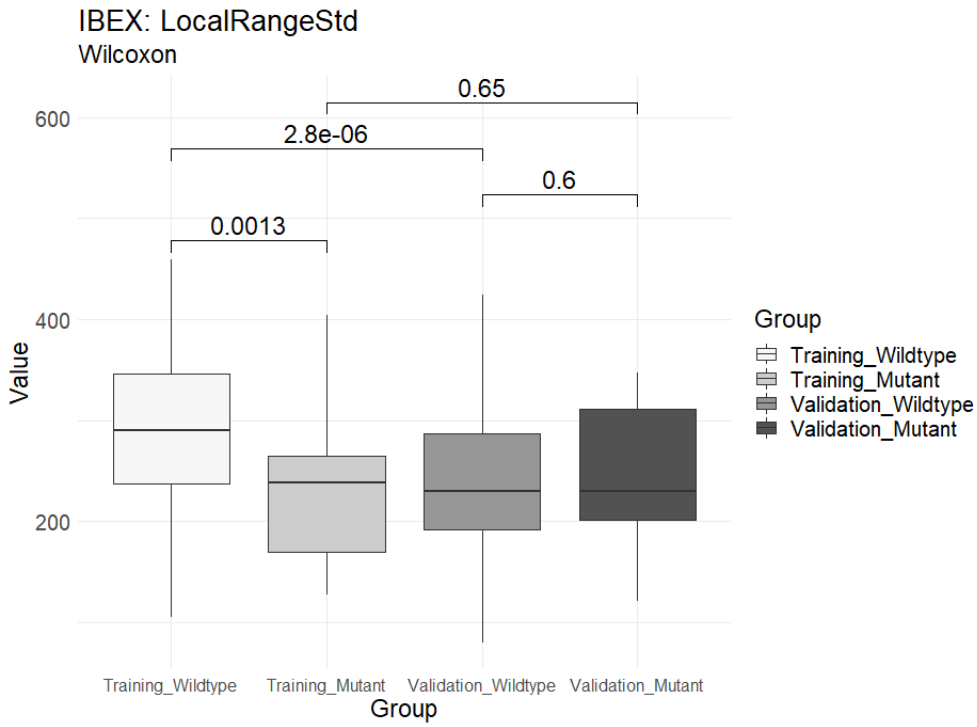


Figure S2.3 IBEX: ClusterShade

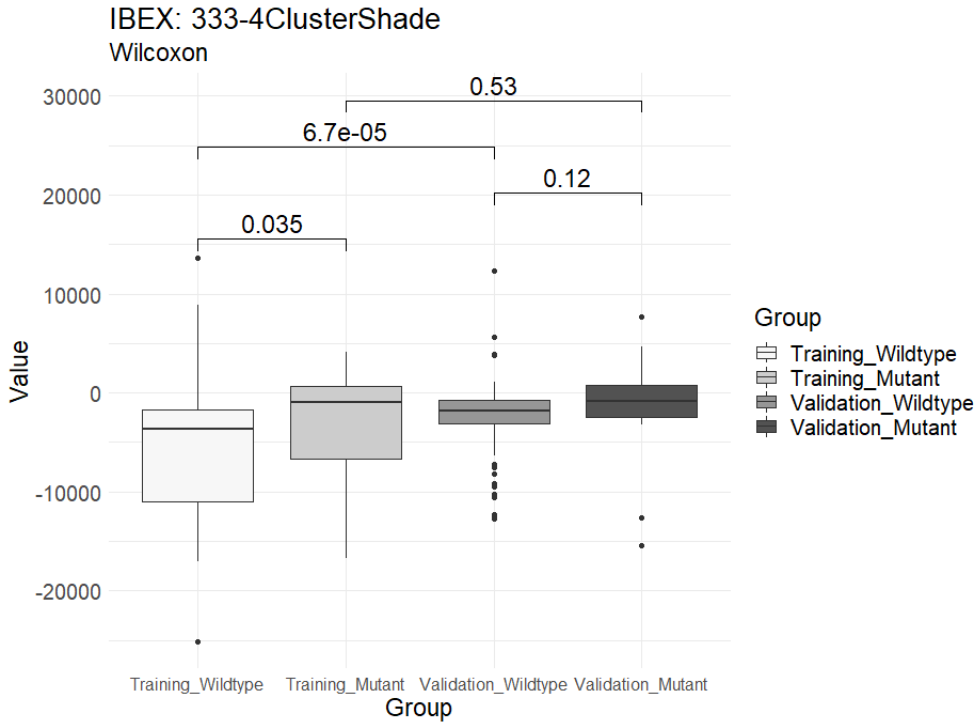


Figure S2.4 IBEX: Correlation

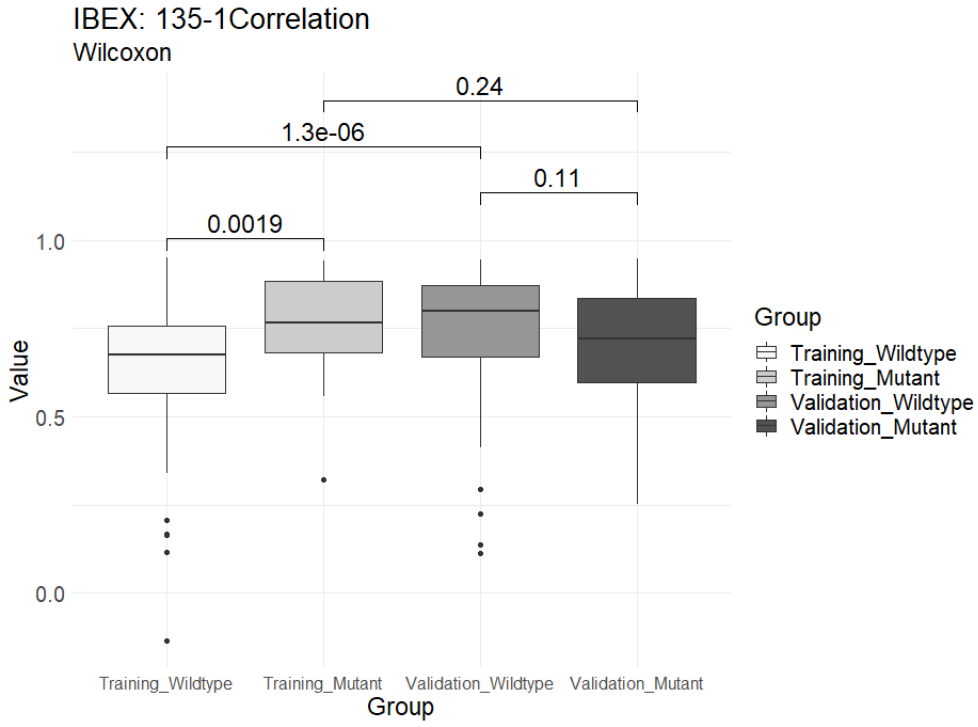
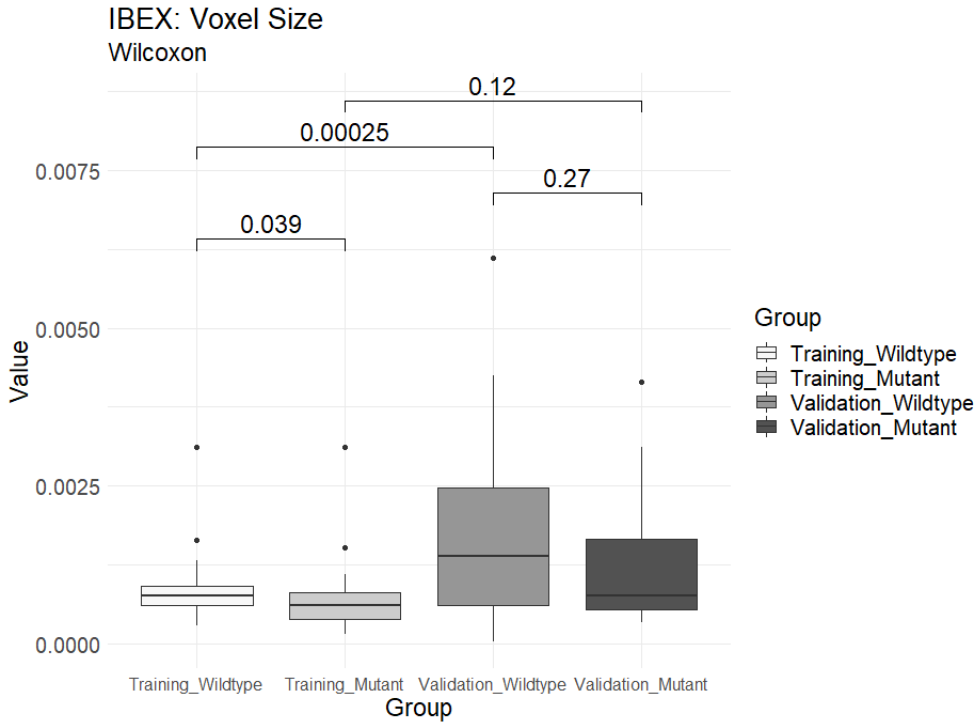


Figure S2.5 IBEX: Voxel Size



Pyradiomics features

Figure S3.1 Pyradiomics: Minimum

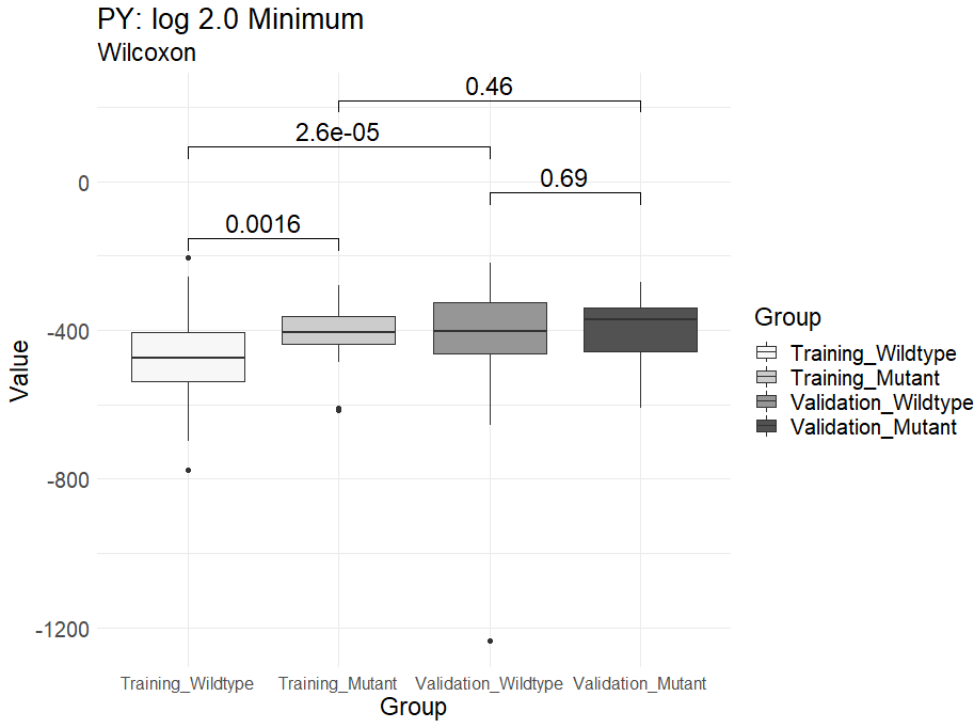


Figure S3.2 Pyradiomics: Inverse Variance

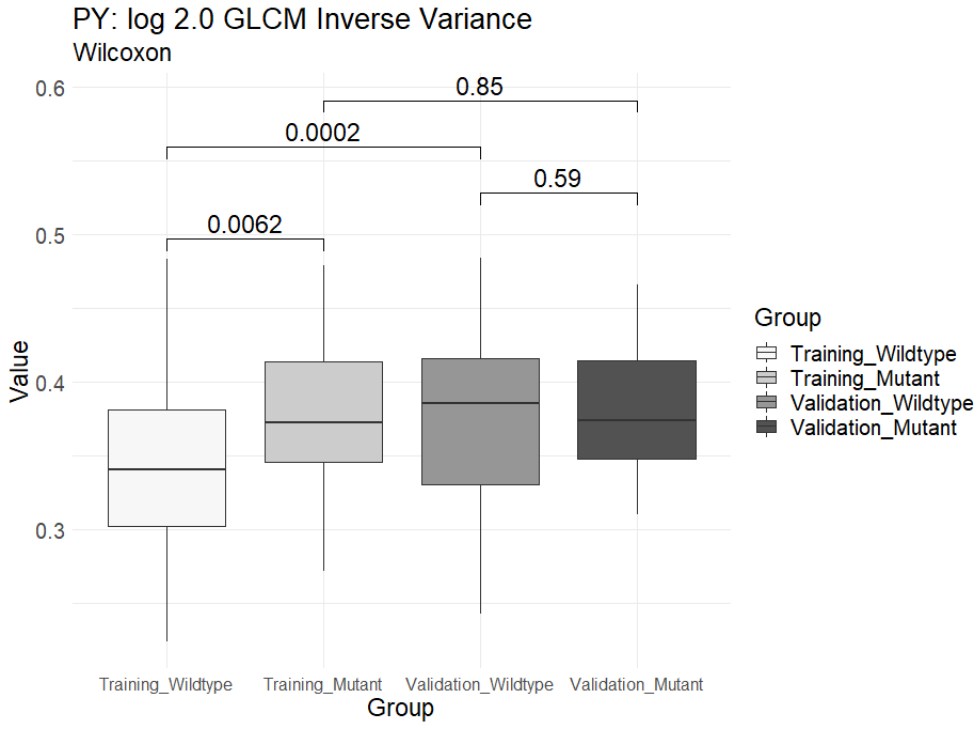


Figure S3.3 Pyradiomics: Size Zone NonUniformity Normalized

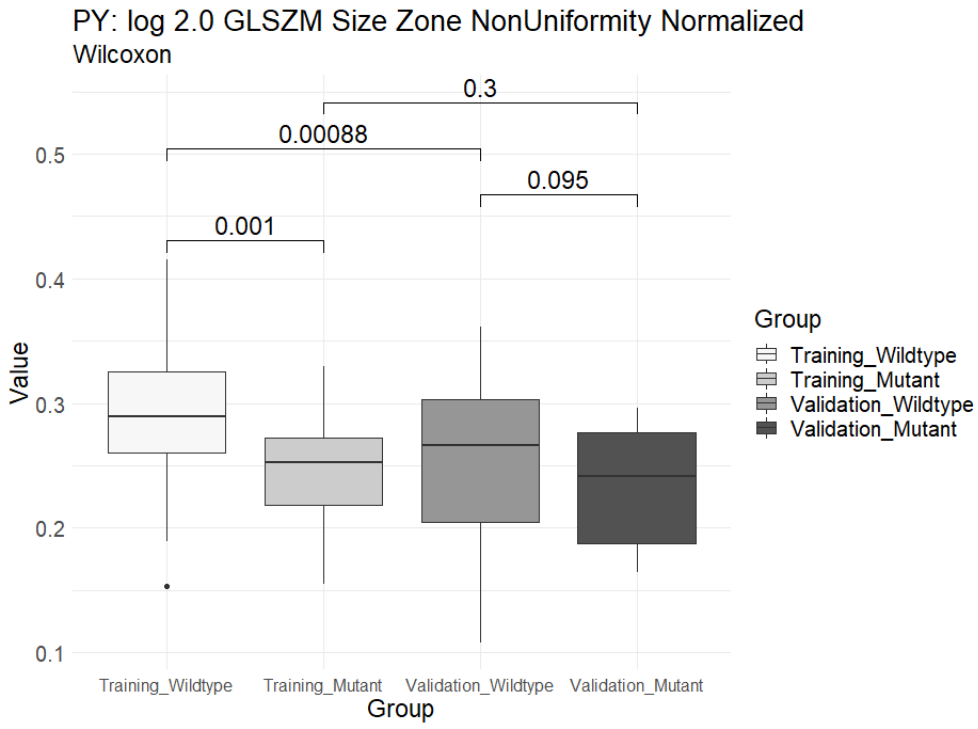


Figure S3.4 Pyradiomics: Small Area Emphasis

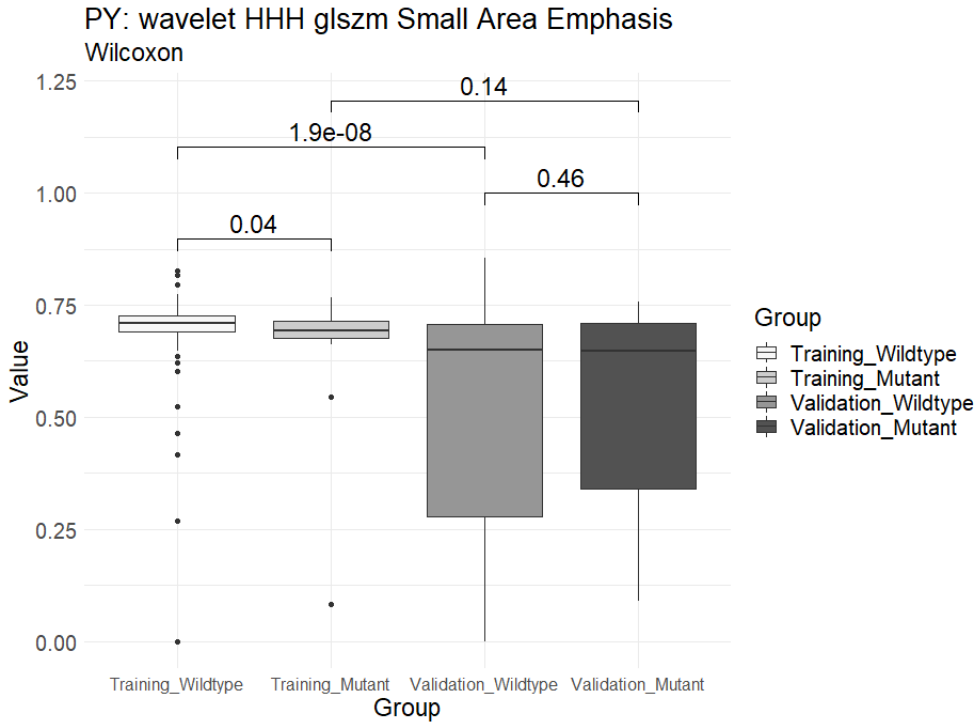


Figure S3.5 Pyradiomics: Skewness wavelet LHH

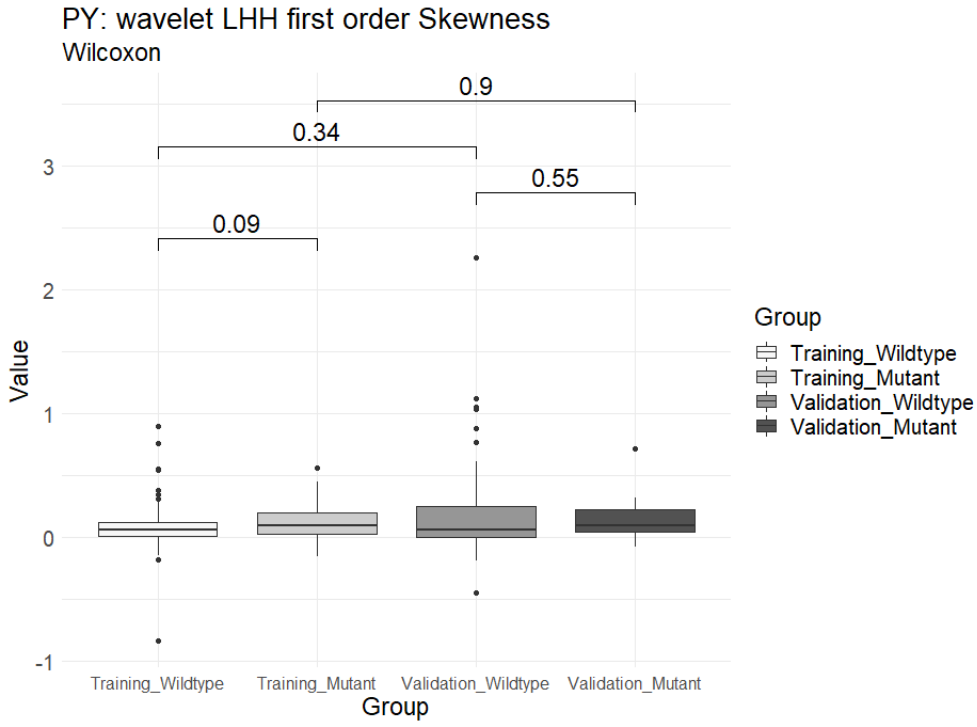
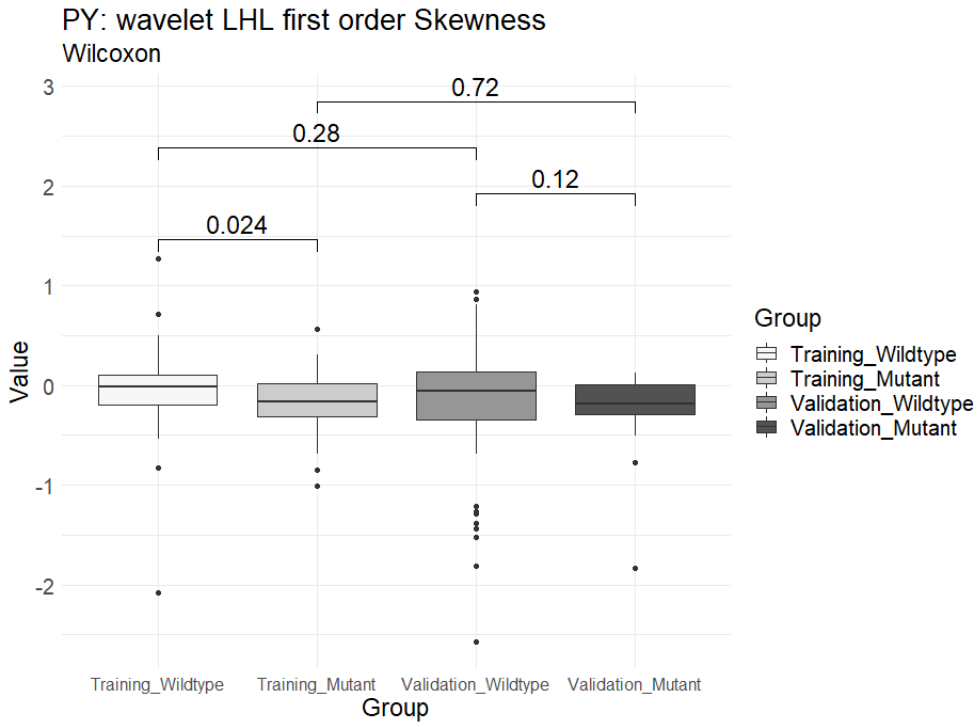


Figure S3.6 Pyradiomics: Skewness, wavelet LHL



CIFE features

Figure S4.1 CIFE: Intensity Minimum

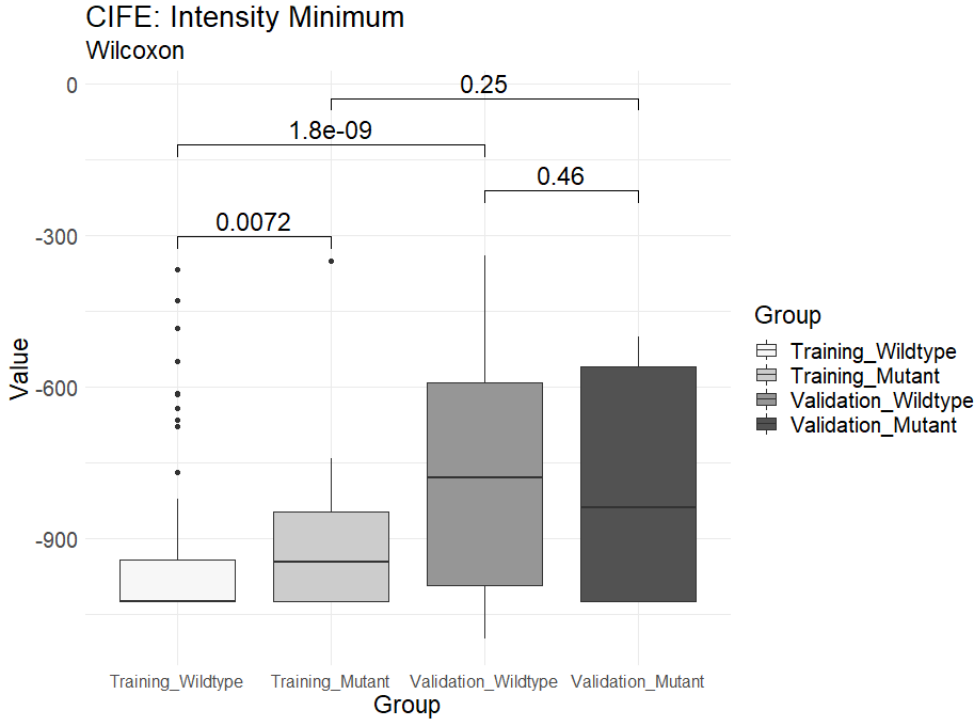


Figure S4.2 CIFE: Gabor Max Z

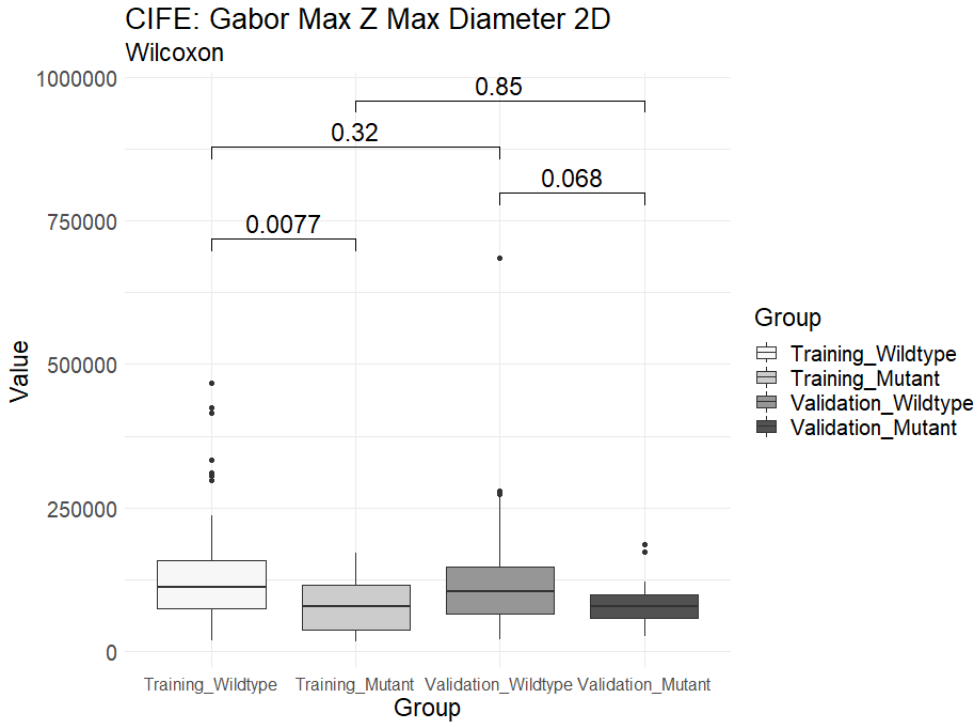


Figure S4.3 CIFE: DWF_Z_H

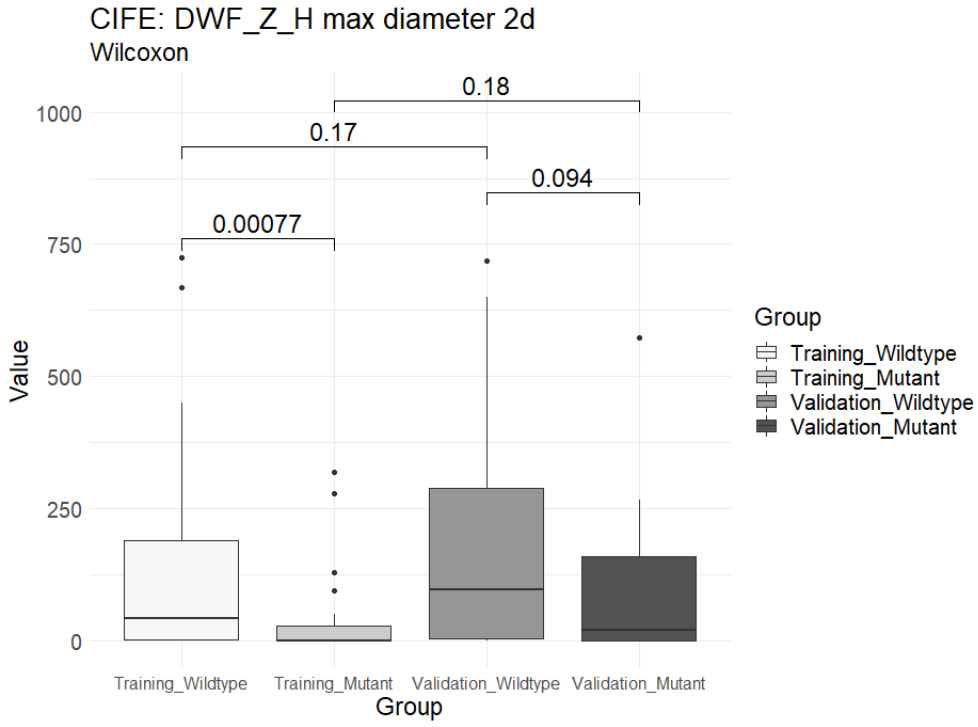
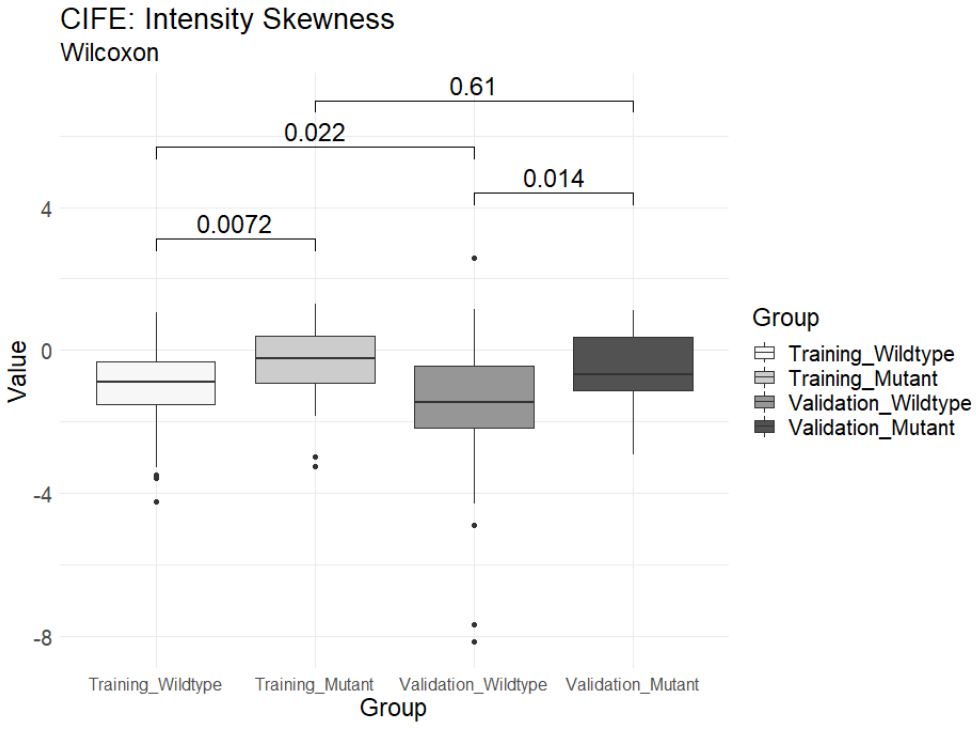
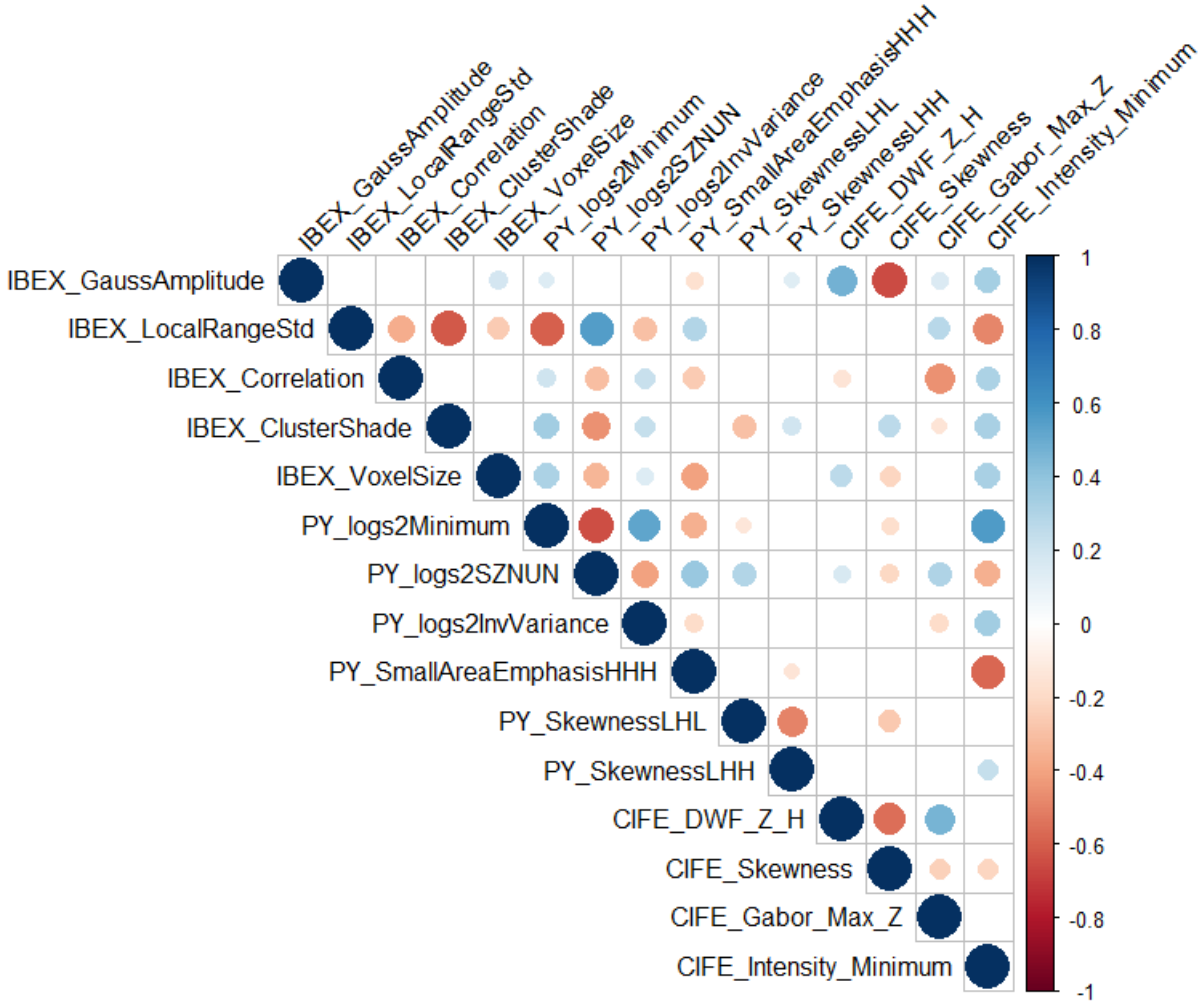


Figure S4.4 CIFE: Intensity Skewness



S2C. Correlation between candidate features

Figure S5. Correlogram of all selected features.



S2D. Comparison of optimal multivariate models based Pyradiomics, IBEX and CIFE

Table S2. Comparison of best multivariate models.

A) Pairwise comparison between packages

Bootstrap comparison	P-value
IBEX ~ PY	0.1883
CIFE ~ PY	2.016e-10
CIFE ~ IBEX	1.536e-14

B) Comparison of training vs validation performance

Bootstrap comparison	P-value
IBEX	< 2.2e-16
PY	< 2.2e-16
CIFE	0.005365

P-values are generated using the Wilcoxon Mann Whitney test on the AUC distributions from the bootstrap approach. The left table (A) details a pairwise comparison of the best predictive models (Random Forests) from each feature extractor. The right table (B) details a comparison of the best predictive models on performance in training versus validation datasets.

Section S3. Overall Experience in using Public Datasets and Open-source Feature Extractors

The number of datasets is already extensive on TCIA and promises to only grow further in the future. We were able to find 3 separate public datasets with lung CT images and EGFR mutation status information on TCIA. Granted, there is always room for further improvement in the procurement of missing data, but realistically some degree of incomplete information will be present in especially larger public datasets.

The NSCLC Radiogenomics dataset is well documented by Bakr et al. (51), and the original publication mentions in detail the number of cases that have full clinical data, histopathological grading, pathologic TNM staging, CT tumor segmentations, RNA-seq, and more. During our collection, we found only 1 case that did not have a chest CT scan. In addition, for the subgroup of cases for which segmentation was not provided, we found 1 case with no clear lung lesion and 7 cases which had multiple lung lesions. These were ultimately excluded from the study.

The TCGA LUAD and LUSC datasets are detailed on both the Genomic Data Commons (GDC) and the TCIA databases, but it was overall more difficult to acquire the information we required for our study. Out of the 585 cases in the TCGA LUAD project, only 69 cases have imaging available on the TCIA. The clinical data presented on TCIA for TCGA LUAD includes only information for 26 out of 69 cases. Furthermore, only 1 out of 26 cases had an EGFR mutation result, whereas the rest were not available. After searching the GDC database, we were able to locate the clinical information for 5 more cases, and 38 cases were found to never have had clinical information submitted. We found 567 out of 585 TCGA LUAD cases to have EGFR mutation data on the GDC and ultimately excluded 6 out of 69 cases with imaging for not having EGFR mutation data. The TCGA LUSC dataset was found in a similar condition: imaging data was available for 37 of 505 total cases, and TCIA provided clinical information for 36 out of 37. EGFR mutation status was available for none of these, and had to be found on the GDC. Ultimately 1 case out of 37 had to be excluded because EGFR mutation status was never found. We contacted the TCIA help desk during this process and received response in a timely manner.

Open-source feature extraction software packages have become more common and are seeing increased use from researchers all over the world (8, 15, 36-41, 59-68). We selected Pyradiomics and IBEX for investigation because of their detailed documentation and extensive use by other researchers (28-36). Although overall the experience was straightforward, each program required its fair share of user experimentation in order to be run successfully.

For Pyradiomics, image and segmentation files can only be read in formats supported by SimpleITK. Plastimatch is recommended to convert RT STRUCT files into 3D volumes, but there is an issue we encountered, also noted by other users on the Github (https://github.com/Radiomics/pyradiomics/issues/423), involving an Image and Mask geometry mismatch. Ultimately, we were able to find a solution using the –output-prefix option, details of which is included in the supplementary file. In addition, 3 cases failed the feature extraction from Pyradiomics with the following error: "No label object with label 1". As we were unable to find a solution, we excluded these 3 cases from our experiment. Pyradiomics is only available in a source code format, and may require some knowledge of the Python language to adjust the settings. However, there are clear examples in the documentation of how to run a feature extraction, and settings to use, which makes it more accessible.

For IBEX, we used the standalone version, which did not require any coding knowledge, but had more user input requirements. After uploading a batch folder, individual cases need to be selected and added into a data set. Any errors would also stop the feature extraction process and would require addressing before continuing. We found 10 cases that would encounter an error during the feature extraction process and were excluded from the experiment. A common error encountered was that "X, Y and WEIGHTS cannot have NaN values". A user guide is provided in a published format by Ger et al. titled "Guidelines and Experience Using Imaging Biomarker Explorer (IBEX) for Radiomics" (75). However, the google group of users (77) did not seem to be accessible.

Overall our experience with public datasets and open source feature extraction has been quite smooth. The majority of data cases fulfilled our inclusion criteria for our experiment and is easily accessible and ready for use. Features were able to be extracted for the majority of cases with all programs and had clear documentation to facilitate use by a beginner.