Table S1. Locoregional or Systematic Recurrence

No. of cases	Locoregional or Systematic Recurrence
187	lung including ribs and mediastinal
30	lung
4	Liver, Bone, Brain, Vertebra
27	Brain
41	lung
29	Scapula, Ribs, pelvis, vertebra, skull
15	Pleura, Diaphragm
4	Adrenal gland
2	Mediastinum, left apical lesion abutting the subclavian vessels and mediastinum
11	Ribs, Lung, Diaphragm, Pancreas, spleen

Table S2. Radiomic features and the patterns that they capture.

Feature group	Quantity	Description	Rationale
Laws energy measures	25	Response to 5-pixel × 5-pixel filter targeting combination of specific textural enhancement patterns in the X and Y directions. Descriptors include all combinations of five 1D filters: level (L), edge (E), spot (S), wave (W), and ripple (R).	May possibly detect patterns of heterogeneous enhancement and abnormal structure
Laws Laplacian features	25	Laplacian pyramids allow for capture of multi- scale edge representations via a set of band pass filters. First, the original image is convolved with a Gaussian kernel. The Laplacian is then computed as the difference between the original image and the low-pass-filtered image. The resulting image is then sub-sampled by a factor of 2, and the filter subsample operation is applied recursively. This process is continued to obtain a set of band pass-filtered images. Laws Energy filters are then applied to the resulting images to obtain a set of 25 features.	Like Laws features can detect patterns of heterogeneous enhancement and abnormal structure
Gabor features	48	Detection of edges through response to Gabor wavelet features. Each descriptor quantifies response to a given Gabor filter at a specific frequency ($f \in \{0, 2, 4, 8, 16, \text{or } 32\}$) and orientation ($\theta \in \{0, \pi/8, \pi/4, 3\pi/8, \pi/2, 5\pi/8, 3\pi/4, 7\pi/8\}$).	May possibly capture changes in tumor microarchitecture or detect the presence of TILs
Haralick features	13	Quantify heterogeneity and entropy of local intensity texture as represented by the gray-level co-occurrence matrix within a 5-pixel × 5-pixel window.	Quantify heterogeneity and entropy of tumor texture
3D Shape Features	12	convexity, width, height, depth, perimeter, area, eccentricity, compactness, radial distance, roughness, elongation equivalent diameter and 3D-sphericity of the nodule.	irregularities in tumors shape can result from its internal heterogeneity and differences in the growth pattern

Table S3. Demographics and clinical characteristics for the 350 cases from D_1 and D_2.

Characteristics		***350 Patients from D_1 and D_2***
Sex	Male	136 (39%)
	Female	214 (61%)
Age	Median (range)	67 (22 – 87)
Race	White	287 (82%)
	Black	63 (18%)
Smoking	Never	49 (14%)
	Former or current	301 (86%)
Histology	Adenocarcinoma	276 (79%)
	Squamous cell carcinoma	53 (15%)
	Large cell carcinoma	21 (6%)
Cancer stage	I A, B	224 (64%)
	II A	126 (36%)
Surgery	Lobectomy	273 (78%)
	Pneumonectomy	14 (5%)
	Wedge Resection (Segmentectomy)	58 (17%)
Time to recurrence after surgery	Median (range)	17.5 (1.3 – 75.3)

1: High AUC (>0.7), Low PI (0.1<PI<0.25)

- ● Skewness of peri-tumoral (2, π/8)
- ● Kurtosis of peri-tumoral (2, π/8)
- ● Var of peri-tumoral (2, π/8)
- ● Mean of peri-tumoral (2, 3π 8)
- ● Median of peri-tumoral (2, 3π/8)
- ● SD of peri-tumoral (2, 3π/8)
- ● Kurtosis of peri-tumoral (2, π/2)
- ● Skewness of peri-tumoral (2, 7π/8
- ○ Mean of peri-tumoral inertia
- ○ Skewness of peri-tumoral sum_va

2: High AUC (>0.65), very Low PI (<0.1)

- ● SD of (0, π/8) Gabor
- ● Kurtosis of (0, π/8) Gabor
- ● Var of (0, π/8) Gabor
- ● Mean of (0, π/4) Gabor
- ● Median of (0, π/4) Gabor
- ● Kurtosis of (0, π/4) Gabor
- ● Var of (0, π/4) Gabor
- ● Mean of (0, 3π/8) Gabor
- ● Median of (0, 3π/8)
- ● Skewness of (0, 3π/8)
- ● SD of (0, π/2)
- ● Var of (0, π/2)
- ● Var of (0, 3π/4)
- ● Median of (0, 7π/8)
- ○ Var of peri-tumoral entropy
- ○ Mean of peri-tumoral idm
- ○ Median of peri-tumoral idm

3: Low AUC (<0.48), Low PI (<0.05)

- ● Kurtosis of L*E
- ● SD of L*R
- ● Kurtosis of L*E Laplace
- ● Mean of diff_av
- ● Mean of peri-tumoral (2, π/8)
- ● Mean of peri-tumoral W*E
- ● Skewness of peri-tumoral W*L Laplace
- ● Mean of peri-tumoral W*E Laplace

4: Low AUC (<0.5), High PI (>0.8)

- ● Kurtosis of (4, 0)
- ● Skewness of (4, π/8)
- ● Kurtosis of (4, π/8)
- ● Var of (4, π/8)
- ● Mean of (4, π/4)
- ● Kurtosis of (32, 0)
- ● Skewness of (32, π/8)
- ● Kurtosis of (32, π/8)
- ● Var of (32, π/8)
- ● Mean of (32, π/4)

Figure S1. 46 features found in ROIs of the PI-AUC space shown in Figure 2. The PI and AUC values were found by using all 175 patients in training set. PI is calculated across cancer recurrence and AUC is the mean value obtained across 100 iterations of 3-fold cross validation for the prognostic task.

Feature Stability Measures

Features may be affected by factors such as slice thickness, specimen preservation procedure, and scanner hardware. To evaluate the effect of these site-specific factors on the quantitative features, Leo et al. defined two measures of instability. The first, latent instability score (LI), describes the inherent randomness of a feature's distribution within a single preparation procedure. A feature's LI is calculated by repeatedly randomly dividing the patients of a single dataset in half and testing the feature's distributions in the two halves for a significant difference. A low LI indicates that there is a low probability of features being significantly different between two datasets from different sites due to effects such as a different set of patients. Secondly, they introduce a method involving cross-dataset comparisons for quantifying the frequency at which a feature is different between datasets from different sites. This is the preparation-induced instability score (PI). A high PI indicates that a feature is frequently different between sites, likely because it is affected by preparation procedures. In other words, the preparation instability (PI) is the rate that a feature has a significantly different distribution between sites. A feature with a low LI but high PI is likely to have been affected by site-induced variation. The low LI indicates that the feature is not unstable from differences in image content or patient population but the high PI indicates that the feature was frequently found to be different between sites. Thus, if $PI \gg LI$, differences between sites has greatly affected the features. Leo et al. found that LI was universally low in every feature of every dataset they tested, with LI being an order of magnitude less than PI, suggesting that LI can be neglected and cross-site feature instability can be almost entirely attributed to site-induced variation.

In this study, feature instability was calculated using the cancer-recurrence patients. Differences in the feature values of cancer recurrence patients was assumed to be a sign of site instability because it was considered unlikely that the texture would vary across sites without site-specific confounding effects. A low instability (PI) shows that features are not significantly different sites in the training set, whereas a high instability indicates that feature values have been greatly affected by site-specific factors.

Every feature which exhibited a PI above a predetermined threshold (PI = 0.1) was excluded from consideration. For this problem, PI=0.1 suggests that a radiomic feature associated with a nodule that resulted in cancer recurrence following surgical removal was significantly different between different sites for 10% of comparisons.

To determine the relationship between feature stability and discriminability, the area under the receiver operating characteristic (ROC) curve (AUC) for each feature was then calculated. Feature AUC was calculated with 100 iterations of 3-fold cross validation (CV) using twelve machine learning classifiers in the training sets. In this study, twelve machine-learning classifiers were used to evaluate the accuracy of features in the training set. A detailed description with regard to the machine learning classifiers can be found in section 2 of this supplementary. The mean AUC across twelve classifiers was then used to generate the final AUC for each feature. Every feature which exhibited an AUC under a predetermined threshold (0.67) was also excluded from further consideration. Each feature had two values, a PI and mean AUC. Thus each feature was defined by a unique position in the PI-AUC space. Those features identified simultaneously as stable and discriminating were then used to construct generalizable prognostic classifier.

Machine Learning Classifiers

In this study twelve machine-learning classifiers from nine classifier families were used.

Bagging (BAG): Bagging is a “bootstrap” ensemble method that creates individuals for its classes by training each classifier on a random redistribution of the training set. The training set is generated by randomly drawing with N replacement - where N is the size of the original training set.

Bayesian (BY): Bayesian is a probabilistic classifier that makes classifications using the Maximum a Posteriori decision rule in a Bayesian setting based on Bayes’ Theorem.

Boosting (BST): The term ‘Boosting’ refers to a family of algorithms which converts weak learner to strong learners. To find weak rule, the learning algorithms with a different distribution is applied and each time a new weak prediction rule is generated. After many iterations, the boosting algorithm combines the weak rules into a single strong prediction rule.

Decision Trees (DT): The decision trees are a type of supervised machine learning classifier where the data is continuously split according to a certain parameter (features). The tree is explained by two entities, decision nodes and leaves. The leaves are the decisions (outcomes) and the decision nodes are where the data is split. The classification is done based on the path that gives the highest information gain.

Linear Discriminant Analysis (LDA): The LDA classifier generates a linear class boundary while assuming that each class has normal distribution with a class-specific mean vector and a common variance.

Quadratic Discriminant Analysis (QDA): The QDA classifier generates nonlinear class boundaries (quadratic patterns) while assuming that the covariance of each class is not identical.

Nearest Neighbors (NN): In this classification, an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors

Neural Networks (Nnet): A neural network consists of units (neurons), arranged in layers, while each unit takes an input, applies a (often nonlinear) function to it and then passes the output on to the next layer. A unit feeds its output to all the units on the next layer, but there is no feedback to the previous layer. Weightings are applied to the signals passing from one unit to another, and it is these weightings which are tuned in the training phase.

Random Forests (RF): Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object.

Support Vector Machines (SVM): The classifier separates data points using a hyperplane with the largest amount of margin. A hyperplane is a decision plane which separates between a set of objects having different class memberships. The SVM algorithm is implemented in practice using a kernel. A kernel transforms an input data space into the required form. Based on which kernel is used to transform input data space into the required form, the classifier is divided into three subgroups. Linear Kernel, Polynomial Kernel and Radial Basis Function Kernel (RBF).