SCIENTIFIC REPORTS

OPEN

Revealing Tumor Habitats from Texture Heterogeneity Analysis for Classification of Lung Cancer Malignancy and Aggressiveness

Received: 17 May 2018
Accepted: 8 January 2019
Published online: 14 March 2019

Dmitry Cherezov[1], Dmitry Goldgof[1], Lawrence Hall[1], Robert Gillies[2], Matthew Schabath[3], Henning Müller[4,5] & Adrien Depeursinge[4,6]

We propose an approach for characterizing structural heterogeneity of lung cancer nodules using Computed Tomography Texture Analysis (CTTA). Measures of heterogeneity were used to test the hypothesis that heterogeneity can be used as predictor of nodule malignancy and patient survival. To do this, we use the National Lung Screening Trial (NLST) dataset to determine if heterogeneity can represent differences between nodules in lung cancer and nodules in non-lung cancer patients. 253 participants are in the training set and 207 participants in the test set. To discriminate cancerous from non-cancerous nodules at the time of diagnosis, a combination of heterogeneity and radiomic features were evaluated to produce the best area under receiver operating characteristic curve (AUROC) of 0.85 and accuracy 81.64%. Second, we tested the hypothesis that heterogeneity can predict patient survival. We analyzed 40 patients diagnosed with lung adenocarcinoma (20 short-term and 20 long-term survival patients) using a leave-one-out cross validation approach for performance evaluation. A combination of heterogeneity features and radiomic features produce an AUROC of 0.9 and an accuracy of 85% to discriminate long- and short-term survivors.

Computed Tomography (CT) is widely used in early detection, diagnosis and treatment planning of lung cancer[1,2]. Using standard-of-care CT images, quantitative image features such as location, spiculation, size, calcification, density (intensity), necrosis and texture of a nodule can be extracted. Radiomics is the conversion of images to structured data and the resulting quantitative features can be used in mathematical models, often learned, for finding a dependence or inter-relationships between features and a medical question such as nodule malignancy, tumor aggressiveness and prediction of treatment response[3–5]. The second role of radiomics is the extraction of features that represent information that is not typically found from CT images by the human eye alone[6–9] or that cannot easily be quantified.

One of the well-known characteristics of cancer is tumor heterogeneity. Hence, small biopsy specimens may not be representative of a whole tumor. Moreover, tumor histology often changes over time. This makes habitat detection a subtle process. Up-to-date habitat detection using radiomic methods can be divided into two categories.

Multi-parametric or multi-modality methods such as T_1, T_2, Flair MRI imaging[10–13] or PET/CT imaging[14–17] provide enough data for the detection of physiologically similar sub-regions ("habitats") within a nodule or a tumor. Single-modality imaging provides less information. In this case radiomic texture features associated with heterogeneity of a nodule are used[18–25].

Features associated with heterogeneity of a nodule have one common characteristic: they compute texture signatures across the entire nodule (see Fig. 1). Knowing that cancer is heterogeneous and assuming that CT texture

[1]Department of Computer Sciences and Engineering, University of South Florida, Tampa, Florida, USA. [2]Department of Cancer Physiology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida, USA. [3]Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida, USA. [4]Institute of Information Systems, University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland. [5]University of Geneva, Geneva, Switzerland. [6]Biomedical Imaging Group, École polytechnique fédérale de Lausanne (EPFL), Lausanne, Switzerland. Correspondence and requests for materials should be addressed to D.C. (email: cherezov@mail.usf.edu)

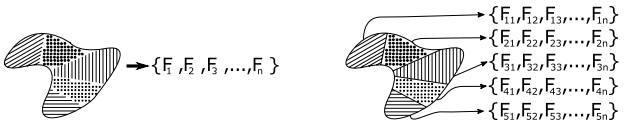

Figure 1. Schematic representation of a feature computation result in a workflow where a nodule is considered as a homogeneous object (on the left) and a feature computation result when heterogeneity is used for division of a nodule into habitats (on the right).

represents tissue types with different histology subtypes, we can conclude that computation of texture signatures in this case is averaging texture features of nodule regions with possibly different histology. As a result, the averaged texture features may not represent any individual habitat within a nodule correctly. In this case, the texture of each nodule is considered as a unique pattern, which makes the classification process more complicated[26]. For example, consider four habitats that have unique texture signatures: A, B, C and D. Each nodule contains two habitats: AA, AB, AC, etc. If we compute the texture signature of an AB nodule, the result will be different from the individual A and B habitat texture signatures. This difference is the result of averaging texture signatures. Thus, each unique combination of habitats in our case has a unique texture signature.

In this work, we present an approach where we compute circular harmonic wavelets for small patches within a nodule, cluster patches in order to define sub-regions of a nodule with similar patterns (habitats) and use information about the clusters and their texture signatures to describe a nodule (Fig. 1). This approach was used to classify nodules into benign and malignant. In addition, we used a dataset with 40 patients diagnosed with adenocarcinoma to evaluate how effective the approach is for classification of tumor aggressiveness.

Materials and Methods

Datasets. In this paper, we experimented with two datasets. First, we used the National Lung Screening Trial (NLST) dataset to evaluate how heterogeneity differentiates nodules from lung cancer patients from nodules among non-lung cancer patients. Second, we used patients from the H. Lee Moffitt Cancer Center & Research Institute diagnosed with lung adenocarcinoma for training and testing to predict patient survival time. These two datasets were chosen to allow comparisons with earlier work[3,25] and to show that the approaches generalize to different datasets. There are labels available for the nodules and tumor images used in the study, but the pixel level ground truth is not available.

National Lung Screening Trial. NLST was a randomized trial of 53,439 patients that compared LDCT (Low Dose CT) vs. standard chest x-rays. After an initial screening (T0), follow-up screenings (T1 and/or T2) were conducted in intervals of approximately one-year. If at T1 a patient was diagnosed with cancer, he/she started treatment and did not have a follow-up screening. According to the screening protocol, a screen was considered positive if a non-calcified nodule (NCN) had the longest diameter (LD) larger than 4 mm. For positive screenings, radiologists provided a clinical description such as location, margin etc.

We extracted two cohorts from NLST[27]. Cancer patients in the training cohort had a positive (non-cancer) screening at time 0 and were diagnosed with cancer on the first follow-up (N = 104). Cancer patients in the test cohort had a positive, so non-cancer screening at time 0 and time 1. They were diagnosed with cancer at time 2. For each cancer patient, two non-cancer subjects were selected by demographic criteria: the same age, sex, and other available criteria. Finally, we excluded cases with technical problems or other challenges that prevented analysis of nodules. When removing a cancer patient from the dataset the corresponding non-cancer patients remained. There are 253 patients in the training cohort (83 cancer and 170 non-cancer patients) and 207 patients in the test cohort (73 cancer and 135 non-cancer patients).

Labels for the dataset represent patient diagnosis during the trial.

Lung Adenocarcinoma Dataset. At the H. Lee Moffitt Cancer Center & Research Institute, 276 patients with Non-Small Lung Cancer were selected. Inclusion/exclusion criteria were: (1) Diagnosed with Lung cancer; (2) Pre-surgery contrast-enhanced CT imaging performed at the H. Lee Moffitt Cancer Center & Research Institute; (3) At least 2 years of follow-up information is known; (4) Patients with all TNM stages accepted; (5) No mix of cancer types for a patient.

Out of 276 patient 86 were diagnosed with Adenocarcinoma. From the Adenocarcinoma subset, two quartiles were selected to represent distinct phenotypes: aggressive phenotype associated with short term survival patients and a non-aggressive phenotype associated with long term survival patients. It was recognized that without a time gap around the class cutoff, it was likely that significant confusion would occur near any cutoff. For short term survival group selection criteria was survival time less than 500 days. For the long term survival group selection criteria was survival time greater than 1000 days. Survival time was computed as the difference between the day of pre-surgery imaging and the last day of contact or the day of patient's death.

Among 86 patients with adenocarcinoma 20 patients survived from 103 to 498 days. Mean survival time was 288 days. These patients were labeled short-term survivors. 20 patients survived from 1351 to 2163 days with the mean survival time of 1569 days. These patients were labeled long-term survivors.

Overall, 40 patients were used for classification of long term survivors and short term survivors. Demographic information for the patients is shown in Table 1.

Characteristics	Short Survival Class	Long survival class	P Value
Age, mean (SD)	69 (8.07)	64.45 (9.75)	0.1161 (Unpaired student t-test)
Sex, N (%)			0.2049 (Fisher exact test)
Male	12 (60%)	7 (35%)	
Female	8 (40%)	13 (64%)	
Race			1 (Fisher exact test)
White	20 (100%)	20 (100%)	
Black, Asian, and Others	0 (0%)	0 (0%)	
Ethnicity, N (%)			1 (Fisher exact test)
Hispanic or Latino	1 (5%)	0 (%)	
Neither Hispanic/Latino and unknown	19 (95%)	20 (100%)	
Histology, N (%)			
Adenocarcinoma	20 (100%)	20 (100%)	
Squamous cell carcinoma	0 (100%)	0 (100%)	
Other, NOS, unknown	0 (100%)	0 (100%)	
Stage, N (%)			0.07346 (Mann-Whitney U test)
I	4 (20%)	10 (50%)	
II	5 (25%)	5 (25%)	
III	10 (50%)	3 (15%)	
IV	1 (5%)	2 (10%)	
Carcinoid, unkown	0 (0%)	0 (0%)	
Tobacco Use, N (%)			
Moderate (1–2 PPD)	4 (20%)	4 (20%)	
Light (<1PPD)	0 (0%)	1 (5%)	
HIST	12 (60%)	12 (60%)	
None	0 (0%)	3 (15%)	
Cigarettes Nos	4 (20%)	0 (0%)	

Table 1. Demographic Summary of Patients in the Adenocarcinoma Data Set.

Pre-processing. For both datasets, segmentations of tumors were obtained at the H. Lee Moffitt Cancer Center where a qualified radiologist applied a semi-automated 3D segmentation algorithm[28].

2D wavelet features were used in the work of this paper. Thus, for each primary nodule from both datasets, we extracted one slice where the segmentation area was the largest. We re-sampled each selected slice such that the XY spacing became equal to 0.5 mm. In the case that, a nodule segmentation area is less than a single patch area, the original slice segmentation was used as a patch and we considered that a nodule had only one habitat. For re-sampling, we used the bicubic interpolation algorithm implemented in Matlab R2016b[29].

Methods

In this section, we describe the proposed workflow with dedicated subsections for each step. Figure 2 shows the workflow as a diagram. First, we describe texture feature computation, patch extraction within a nodule and computation of the convolution response for a patch. Second, we explain how we defined the number of habitats for each nodule, the habitats of multiple patches and the habitat texture response. We assume that malignancy of habitats within a nodule can vary. Finally, we explain how we used texture responses of habitats to evaluate their malignancy for training and test cohorts in detail.

In order to classify patients, we extracted quantitative features that described the heterogeneity of a nodule. These features were used in the patient classification experiments. Performance of the features is shown in the Results section.

Circular Harmonic Wavelet Features. We chose to use Circular Harmonic Wavelets (CHW) to characterize local texture properties of a tumor image $f(x, y)$[30]. CHWs quantify the amount of local circular frequencies, similar to local binary patterns (LBP)[31]. An interesting property of CHWs is their ability to characterize image directions in a rotation-invariant fashion at a very low computational price[32]. This allows quantification of benign or malignant tissue structures independently of their local orientations. CHWs of order n are constructed in the Fourier domain as

$$\hat{\phi}^{(n,s)}(\rho,\ \theta) = \hat{h}(2^s\rho)\mathrm{e}^{\mathrm{j}n\theta}, \tag{1}$$

where (ρ, θ) denotes the polar coordinates in the Fourier domain and $\hat{h}(\rho)$ is a purely radial bandpass function controlling the wavelet analysis at scale s. Simoncelli's isotropic collection of wavelets was used for $\hat{h}(\rho)$[33], which proved to work well for analyzing lung tissue in CT[34]. At a given position (x_0, y_0), the representation obtained from the collection of the complex magnitudes of the scalar products $|\langle f,\ \phi^{(n,i)}\rangle|$ characterizes the local circular frequencies in f of order $n = -N{:}N$ at a scale $s = 1{:}S$ and is locally rotation invariant[35]. This yields a collection of

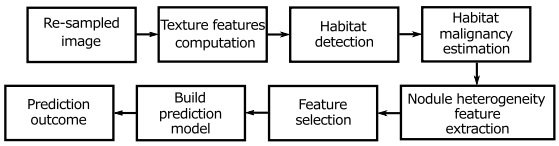

Figure 2. Suggested workflow for heterogeneity estimation.

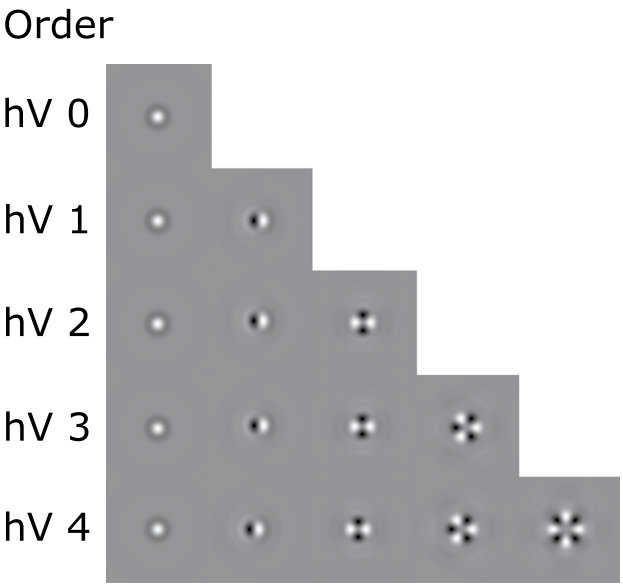

Figure 3. Set of Circular Harmonics filters used for texture signature computation (hV–Harmonic Vectors).

positive response maps having the same dimension as the domain of *f*. Features are obtained by averaging each response map over patches.

We consider five collections of circular harmonics. Figure 3 shows their impulse responses. For each frequency, in a collection, we consider three scales ($S = 3$).

For each pixel within the segmentation we get a convolution response. In order to detect texture patterns, we divide nodules into circular patches with a radius of 3 mm (6 pixels), a shift of 1.5 mm (3 pixels) and we average the absolute values of the wavelet responses within a patch. The choice of the radius of 3 mm was based on Lung-RADS (Reporting and Data System) categories. The procedure was repeated for each set of harmonic vectors (hV) individually and as a result, we obtained five sets of patches for each nodule where each patch has a different number of texture features. The number of texture features extracted from each collection shown in Fig. 3 is equal to 3, 9, 15, 21 and 27 respectively.

Habitat Detection. After computing a set of wavelet features (for each set of harmonic vectors) within each patch the k-means++ algorithm[36,37] was applied to identify regions with a similar texture. The number of texture features for each set is provided above. The number of clusters was estimated with the gap criterion clustering evaluation method[38]. The maximum number of possible clusters is limited to 15. As a result of the habitat detection step, we obtained five sets of habitats for each nodule with respect to five sets of harmonic vectors.

Habitat Malignancy Estimation. After habitat detection, each nodule is represented as a set of texture signatures for habitats (Fig. 1). In this work we assume that a difference in habitat histology can be described with texture patterns. To estimate the probability that a particular habitat belongs to a given class, we applied a leave-one-out cross-validation (LOOCV) on the training cohort.

We excluded from the training cohort one patient. Texture signatures from all the other patients with the corresponding labels were used for training. After training, the excluded patient's texture signatures were used for testing. The classifier produced a pseudo-probability for each signature (Fig. 4). The procedure was repeated for each patient in the training cohort. We collected these pseudo probabilities to describe a nodule and see if this information can be used for the computation of quantitative features and a nodule level classification task.

Several studies showed that Random Forests outperform other classifiers in Radiomics experiments[3,39,40]. Thus we chose random forests for classification, where the fraction of decision trees that voted that a habitat is malignant/aggressive to the total number of decision trees is recorded. This fraction is considered as a pseudo-probability of malignancy/aggressiveness.

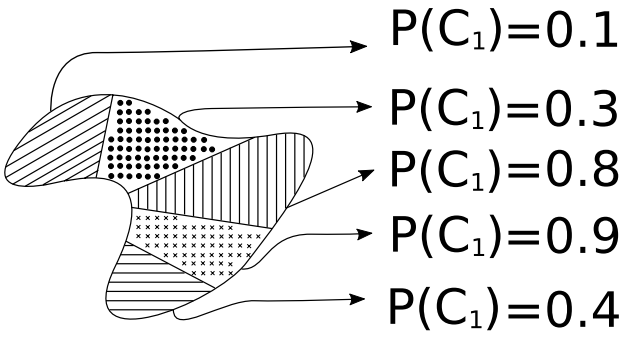

Figure 4. Example of the malignancy/aggressiveness probability assignment for habitats in a nodule.

To evaluate malignancy/aggressiveness of habitats in the test cohorts we used all signatures from the training cohort for training with the assigned labels of the corresponding patient. After this step, for each habitat in the training and the test cohorts, we estimate the probability of it being malignant or aggressive. Habitat malignancy/aggressiveness estimation for the training and test cohorts are repeated for each set of habitats.

Nodule Heterogeneity Feature Extraction. The detection of habitats within a nodule provides much information about its heterogeneity. Nevertheless, it makes it impossible to compare the texture of each nodule directly because the number of habitats differs and because we do not know the relationship between the habitat area, the level of malignancy/aggressiveness of a habitat and the malignancy of the nodule itself.

We produced 15 quantitative radiomics features. These features are statistical information of habitat area, habitat pseudo probability of malignancy/aggressiveness and variety in habitat texture signatures of the nodule. Table 2 shows the names of the features and the corresponding description. Heterogeneity features were extracted from all patients and were used for nodule classification.

Results

As a baseline for all experiments we use the results of the classification with 3D radiomics features extracted using the Definiens application and consisting of size, location, intensity and texture features (219 features) computed on the entire 3D ROI[41]. Balagurunathan *et al.* described all categories and features in detail in their work[42]. During feature computation, CTs were not resampled to have uniform spacing.

Heterogeneity features are 2D features and only texture information was used to define them. We combined 2D heterogeneity features and 3D radiomics features to test if such a fusion improves classification performance on the NLST dataset.

For experiments with only the heterogeneity features or their combination with 3D Definiens features, we tested heterogeneity features computed by each frequency collection individually (Fig. 3).

To reduce the number of features we applied the ReliefF feature selector[43–45] to select the best 5, 10 features and the minimum redundancy maximum relevance feature selector[46]. Experiments with no feature selection were used as a baseline. With the exception of Naive Bayes, all classifiers used do implicit feature selection. Reducing the number of features with a feature selector resulted in better performance. As a subset of the Definiens features we consider features that were shown to be stable on the RIDER dataset and separately features stable on the training cohort[41]. As the set of classifiers, we selected Naïve Bayes[47], J48[48], JRIP[49], Random Forests[50] and SVMs with a linear and a radial basis function kernel[51]. All the experiments were executed in Weka, version 3.8.1.

Lung Cancer vs. Non-Lung Cancer in NLST. Three experiments were based on the NLST dataset. First, we used patient screening results at the time of diagnosis: the training cohort at time 1 vs. test cohort at time 2. Second, we used patient CT screening one year ahead of diagnosis to evaluate the heterogeneity of malignant nodules before they were marked as cancer: the training cohort at time 0 vs. test cohort at time 1. Finally, we used the training cohort at time 0 vs. test cohort at time 0, which means that for training we used CT screenings of nodules a year ahead of diagnosis and for testing we use CT screenings of nodules two years ahead of diagnosis. The AUROC of classifiers was considered as the primary performance measure. Heterogeneity helped when combined with 3-D features in all cases except predicting two years in advance. Table 3 shows feature sets, feature selectors and classifiers producing the best AUROCs. The NLST combined model AUROC was compared to the Definiens model performance in R by using the pROC library[52] and the comparison algorithm of DeLong *et al.*[53]. The P-value of AUROC differences is 0.2215.

Survival time prediction in the Adenocarcinoma dataset. Patient stage information from Table 1 was used to evaluate baseline performance using clinical data due to the significant difference in patient survival time. If a patient stage is used as a quantitative feature to differentiate long/short term survival the AUROC is equal to 0.67 (Supplementary Fig. S1). If stages I/II are considered as early stages and stages III/IV are considered as late stages, then the split produces a confusion matrix which leads to an accuracy of 65% (Supplementary Table S1).

Feature name	Feature description
min P	Minimum value of the malignancy pseudo probability of a habitat in a nodule.
max P	Maximum value of the malignancy pseudo probability of a habitat in a nodule.
mean P	Mean value of the malignancy pseudo probability of a habitat in a nodule.
median P	Median value of the malignancy pseudo probability of a habitat in a nodule.
min A ratio	Minimum value of a habitat area in a nodule.
max A ratio	Maximum value of a habitat area in a nodule.
mean A ratio	Mean value of a habitat area in a nodule.
median A ratio	Median value of a habitat area in a nodule.
min disjoint A ratio	Minimum value of a habitat area in a nodule in the case of disjoint parts of a habitats being considered as different habitats.
max disjoint A ratio	Maximum value of a habitat area in a nodule in the case of disjoint parts of a habitats being considered as different habitats.
mean disjoint A ratio	Mean value of a habitat area in a nodule in the case of disjoint parts of a habitats being considered as different habitats.
median disjoint A ratio	Median value of a habitat area in a nodule in the case of disjoint parts of a habitats being considered as different habitats.
number of clusters	Total number of habitats in a nodule.
mean centroids dist	Computing mean value of habitat texture signatures for a nodule–nodule texture signature. The result is mean Euclidean distance from the nodule texture signature to its habitats texture signatures.
dist std centroids	Standard deviation of habitat texture signatures.

Table 2. Heterogeneity features description.

Screening time	Feature type	Feature set	Feature Selector	Classifier	AUROC	Accuracy (%)
Training Set Diagnosis Test Set Diagnosis	Heterogeneity	hV_4	RfF 10	RFs	0.77	72.95
	Definiens	All 219 features	mRMR 17*	RFs	0.83	78.77
	Combined	Training st. $+hV_4$	none	RFs	0.85	81.64
Training set Diagnosis - 1 year Training set Diagnosis - 1 year	Heterogeneity	hV_3	none	SVM_{lin}	0.69	74.88
	Definiens	RIDER st.	RfF 10	RFs	0.79	75
	Combined	All 219 $+hV_2$	mRMR 25*	RFs	0.79	74.4
Training set Diagnosis - 1 year Training set Diagnosis - 2 years	Heterogeneity	h_{V1}	RfF 10	RFs	0.67	65.7
	Definiens	RIDER st.	RfF 5	RFs	0.78	74.06
	Combined	RIDER st. $+h_{V0}$	RfF 10	RFs	0.78	70.53

Table 3. Overview of classification models that produce the best AUROC. The first column defines the time point of the CT screening that was used in the training and test cohorts. The second column defines which feature set was extracted for a given CT screening. Heterogeneity refers to only 15 texture heterogeneity features, Definiens refers to 219 features extracted in Definiens. Combined refers to the fusion of Definiens and heterogeneity features. The feature subset column defines the order of the Circular Harmonic vectors that were used to extract texture features or a subset of the Definiens features claimed to be stable on the RIDER or training datasets. The feature selector column defines one of the feature selectors that produces the best performance. There can be no feature selector, ReliefF (RfF) with top 10 or 5 ranked features or the minimum redundancy maximum relevance (mRMR) feature selector. The classifier column defines, which of the tested classifiers performed the best. From the table we can see that most of the time random forests (RFs) outperformed other classifiers. Finally, the last two columns refer to AUROC and accuracy of the corresponding model. *Weka v.3.8.1 provides mRMR algorithm whose implementation defines the optimal number of features for a particular dataset in terms of redundancy and relevance. As a result, the selected number of features varies.

Due to the small number of patients in the adenocarcinoma dataset, we applied cross-validation for the performance evaluation. Because habitat aggressiveness estimation was performed with leave-one-patient-out, the same concept was applied when testing nodules.

Cross-validation for this dataset with the Definiens features showed an AUROC = 0.71 and an accuracy = 77.5%[25]. Using heterogeneity features alone provides the best AUROC of 0.80 with the corresponding accuracy of 85%. The enhanced contrast likely enabled better habitat definition. Again, the combination of features provided the best results.

The combined model with the best performance from the Lung Adenocarcinoma Dataset is based on a subset of Definiens features that are shown to be stable and reproducible by test-retest analysis on the RIDER dataset[41]. There are 23 features in the RIDER subset of features. In addition to the 23 RIDER features, there are 15 heterogeneity features. Out of 38 total features, the ReliefF algorithm selected the top 5 predictive features. Random Forests for a classification task with N features used the square root of N features for building a decision tree. This means that a random set of sqrt(N) features are chosen with the best of them selected for the test at an internal node. This is an implicit form of feature selection.

Feature type	Feature set	Feature selector	Classifier	AUROC	Acc. (%)
Staging	NA	NA	NA	0.67	65
Heterogeneity	hV_3	mRMR 1*	J48	0.80	85
Definiens	all 219 features	RfF 5	J48	0.71	77.5
Combined	RIDER $+hV_3$	RfF 5	RFs	0.90	85

Table 4. Comparison of Adenocarcinoma aggressiveness estimation results using the heterogeneity and Definiens features. *Weka v.3.8.1 provides the mRMR algorithm, which defines the optimal number of features for a particular dataset in terms of redundancy and relevance. As a result, the selected number of features varies.

Table 4 shows the classification models that perform best for particular feature sets with the corresponding AUROC and accuracy. The Adenocarcinoma Combined model AUROC is compared to Definiens model performance in R by using the pROC library[52] and the DeLong *et al.*[53] comparison algorithm. The P-value of AUROC differences was 0.04924.

Discussion

The hypothesis of this work was that CT screening data can be used not only for the description of a nodule but it is a source of information to define habitats within nodules and that the level of heterogeneity is able to help identify malignancy and aggressiveness of tumors.

As can be seen from Table 3, the results of the three experiments performed on NLST for all features highlight that the highest performance was achieved when using images at the time of diagnosis (training set at T1 vs. test set at T2). The next best performance was reached where both cohorts used data one year ahead of diagnosis (training set at T0 vs. test set at T1). Finally, the worst result was obtained when we used as for training the CT screenings of nodules one year ahead of diagnosis and for testing the CT screenings of nodules that were taken two years ahead of diagnosis. Clinically, the complexity of these questions is on the same order, so from the simplest to the most complicated.

Heterogeneity helps performance most at the time of diagnosis. In addition, technically, it is hard to evaluate heterogeneity of small nodules due to the limited spatial resolution of CT volumes. Most of the nodules at the time of the initial CT screening have a longest diameter of less than 15 mm. In this work, we used patches with a radius of 6 mm. Thus, with very few patches we cover entire nodules and thus we may miss all the information about local habitats. At the time of diagnosis, nodules have grown and thus the performance of habitat detection increases.

In this regard, the Adenocarcinoma dataset is a better choice, as the nodules are larger and contrast is stronger. Thus, the proposed method can better leverage habitat heterogeneity information. Longest diameter mean and standard deviation values for the NLST dataset were 11.06 mm and 7.62 mm respectively. For Adenocarcinoma these parameters were 34.09 mm and 17.4 mm respectively. In addition, CT imaging of Adenocarcinoma dataset patients was performed with use of a contrast agent, which highlights CT texture inside a tumor.

As we can see from Table 4, in the experiment where we used only heterogeneity features the mRMR feature selector selected only one feature. The feature name is "Min P" which is explained in Table 2. After splitting nodules into habitats and computing pseudo-probability of aggressiveness for habitats with Random Forests we selected the minimum value. Just by using this value for nodule aggressiveness classification we got an AUROC of 0.8 and an Accuracy of 85%. This may be the result of the fact that we computed texture signatures for habitats individually.

Conclusions

In this paper, we propose a method for revealing tumor habitats from texture heterogeneity. We use this heterogeneity to classify lung cancer malignancy and aggressiveness. We analyze classification abilities of heterogeneity on two datasets and compared it with 3D features from Definiens based on the entire tumor volume (i.e., not considering tumor habitats). First, we applied heterogeneity for classifying cancer and non-cancer patients in the NLST screening dataset. The best results were obtained when using CT images at the time of diagnosis. When using Definiens features only (219 features), the best AUROC is 0.83. When using the proposed 15 2D heterogeneity features, the best AUROC is 0.79. Combining the two sets of features achieved the top AUROC of 0.85. This small gain suggests that NLST nodules are relatively small to fully benefit from the proposed heterogeneity analysis but it does add important information. To this end, we evaluated heterogeneity in adenocarcinoma patient survival time prediction, where nodules are much larger. In this dataset, the best AUROC was obtained when the model was based on the heterogeneity features (AUROC = 0.8), whereas the global Definiens features were mixing distinct habitats and only achieved the highest AUROC of 0.71. Combining heterogeneity features and the Rider subset of features resulted in a statistically significant improved AUROC of 0.9.

Data Availability

National Lung Screening Trial dataset[54] and Adenocarcinoma dataset[55,56] are available at The Cancer Imaging Archive[57]. The Matlab code for heterogeneity detection described in the section on Circular Harmonic Wavelet Features and the section on Habitat Detection is available on the GitHub server[58].

References

1. Bach, P. B. *et al.* Benefits and harms of CT screening for lung cancer: a systematic review. *JAMA: J. Am. Med. Assoc.* **307**, 2418–2429 (2012).
2. NLSTRT. Reduced lung-cancer mortality with low-dose computed tomographic screening. The New Engl. *J. Medicine* **2011**, 395–409 (2011).
3. Hawkins, S. *et al.* Predicting malignant nodules from screening ct scans. *J. Thorac. Oncol.* **11**, 2120–2128 (2016).

4. Harrell, F. E., Lee, K. L. & Mark, D. B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Medicine* **15**, 361–387 (1996).
5. Coroller, T. P. *et al*. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiother. Oncol.* **114**, 345–350 (2015).
6. Lambin, P. *et al*. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* **48**, 441–446 (2012).
7. Parmar, C., Grossmann, P., Bussink, J., Lambin, P. & Aerts, H. J.W. L. Machine learning methods for quantitative radiomic biomarkers. *Sci. Reports* **5** (2015).
8. Aerts, H. J. W. L. *et al*. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **5** (2014).
9. Parmar, C. *et al*. Robust radiomics feature quantification using semiautomatic volumetric segmentation. *PloS One* **9**, e102107 (2014).
10. Kermode, A. *et al*. Heterogeneity of blood-brain barrier changes in multiple sclerosis an mri study with gadolinium-dtpa enhancement. *Neurol.* **40**, 229–229 (1990).
11. Chaudhury, B. *et al*. Heterogeneity in intratumoral regions with rapid gadolinium washout correlates with estrogen receptor status and nodal metastasis. *J. Magn. Reson. Imaging* **42**, 1421–1430 (2015).
12. Sugahara, T. *et al*. Usefulness of diffusion-weighted mri with echo-planar technique in the evaluation of cellularity in gliomas. *J. Magn. Reson. Imaging* **9**, 53–60 (1999).
13. McVeigh, P. Z., Syed, A. M., Milosevic, M., Fyles, A. & Haider, M. A. Diffusion-weighted mri in cervical cancer. *Eur. radiology* **18**, 1058–1064 (2008).
14. Chicklore, S. *et al*. Quantifying tumour heterogeneity in 18 f-fdg pet/ct imaging by texture analysis. *Eur. journal nuclear medicine molecular imaging* **40**, 133–140 (2013).
15. Win, T. *et al*. Tumor heterogeneity and permeability as measured on the ct component of pet/ct predict survival in patients with non–small cell lung cancer. *Clin. Cancer Res.* **19**, 3591–3599 (2013).
16. Niekel, M. C., Bipat, S. & Stoker, J. Diagnostic imaging of colorectal liver metastases with ct, mr imaging, fdg pet, and/or fdg pet/ct: a meta-analysis of prospective studies including patients who have not previously undergone treatment. *Radiol.* **257**, 674–684 (2010).
17. Kwee, T. C. & Kwee, R. M. Combined fdg-pet/ct for the detection of unknown primary tumors: systematic review and meta-analysis. *Eur. radiology* **19**, 731–744 (2009).
18. Song, J., Dong, D., Huang, Y., Liu, Z. & Tian, J. Association between tumor heterogeneity and overall survival in patients with non-small cell lung cancer. In 2016 *IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, 1249–1252 (IEEE, 2016).
19. McNitt-Gray, M. F., Wyckoff, N., Sayre, J. W., Goldin, J. G. & Aberle, D. R. The effects of co-occurrence matrix based texture parameters on the classification of solitary pulmonary nodules imaged on computed tomography. *Comput. Med. Imaging Graph.* **23**, 339–348 (1999).
20. Bayanati, H. *et al*. Quantitative ct texture and shape analysis: Can it differentiate benign and malignant mediastinal lymph nodes in patients with primary lung cancer? Eur. *radiology* **25**, 480–487 (2015).
21. Ganeshan, B., Panayiotou, E., Burnand, K., Dizdarevic, S. & Miles, K. Tumour heterogeneity in non-small cell lung carcinoma assessed by ct texture analysis: a potential marker of survival. *Eur. Radiol.* **22**, 796–802 (2012).
22. Higashi, K. *et al*. FDG PET in the evaluation of the aggressiveness of pulmonary adenocarcinoma: correlation with histopathological features. *Nucl. Medicine Commun.* **21**, 707–714 (2000).
23. Depeursinge, A., Yanagawa, M., Leung, A. N. & Rubin, D. L. Predicting adenocarcinoma recurrence using computational texture models of nodule components in lung CT. *Med. physics* **42**, 2054–2063 (2015).
24. Pires, A. *et al*. Clustering of lung adenocarcinomas classes using automated texture analysis on CT images. In *Medical Imaging: Image Processing*, 866925 (2013).
25. Hawkins, S. H. *et al*. Predicting outcomes of non-small cell lung cancer using CT image features. *IEEE Access* **2**, 1418–1426 (2014).
26. Depeursinge, A. Multi-Scale and Multi-Directional Biomedical Texture Analysis: Finding the Needle in the Haystack. In *Biomedical Texture Analysis: Fundamentals, Applications and Tools, Elsevier-MICCAI Society Book series*, 29–53 (Elsevier, 2017).
27. Schabath, M. B. *et al*. Differences in Patient Outcomes of Prevalence, Interval, and Screen-Detected Lung Cancers in the CT Arm of the National Lung Screening Trial. *PloS One* **11**, e0159880 (2016).
28. Gu, Y. *et al*. Automated delineation of lung tumors from CT images using a single click ensemble segmentation approach. *Pattern Recognit.* **46**, 692–702 (2013).
29. Keys, R. Cubic convolution interpolation for digital image processing. *IEEE transactions on acoustics, speech, signal processing* **29**, 1153–1160 (1981).
30. Unser, M. & Chenouard, N. A Unifying Parametric Framework for 2D Steerable Wavelet Transforms. *SIAM J. on Imaging Sci.* **6**, 102–135 (2013).
31. Ojala, T., Pietikäinen, M. & Mäenpää, T. Multiresolution gray–scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis Mach. Intell.* **24**, 971–987 (2002).
32. Depeursinge, A. & Fageot, J. Biomedical texture operators and aggregation functions: A methodological review and user's guide. In *Biomedical Texture Analysis: Fundamentals, Applications and Tools*, Elsevier-MICCAI Society Book series, 55–94 (Elsevier, 2017).
33. Portilla, J. & Simoncelli, E. P. A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients. *Int. J. Comput. Vis.* **40**, 49–70 (2000).
34. Depeursinge, A. *et al*. Optimized steerable wavelets for texture analysis of lung tissue in 3-D CT: classification of usual interstitial pneumonia. In *IEEE 12th International Symposium on Biomedical Imaging*, ISBI 2015, 403–406 (IEEE, 2015).
35. Depeursinge, A., Püspöki, Z., Ward, J.-P. & Unser, M. Steerable Wavelet Machines (SWM): Learning Moving Frames for Texture Classification. *IEEE Transactions on Image Process.* **26**, 1626–1636 (2017).
36. Lloyd, S. Least squares quantization in PCM. *IEEE Transactions on Inf. Theory* **28**, 129–137 (1982).
37. Vassilvitskii, S. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 1027–1035 (Society for Industrial and Applied Mathematics, 2007).
38. Tibshirani, R., Walther, G. & Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *J. Royal Stat. Soc. Ser. B (Statistical Methodol.* **63**, 411–423 (2001).
39. Cherezov, D. *et al*. Improving malignancy prediction through feature selection informed by nodule size ranges in nlst. In *Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on*, 001939–001944 (IEEE, 2016).
40. Yan, Z., Li, J., Xiong, Y., Xu, W. & Zheng, G. Identification of candidate colon cancer biomarkers by applying a random forest approach on microarray data. *Oncol. reports* **28**, 1036–1042 (2012).
41. Balagurunathan, Y. *et al*. Reproducibility and prognosis of quantitative features extracted from CT images. *Transl. Oncol.* **7**, 72–87 (2014).
42. Balagurunathan, Y. *et al*. Test–retest reproducibility analysis of lung ct image features. *J. digital imaging* **27**, 805–823 (2014).
43. Kira, K. & Rendell, L. A. A practical approach to feature selection. In *Proceedings of the ninth international workshop on Machine learning*, 249–256 (1992).
44. Kononenko, I. Estimating attributes: analysis and extensions of RELIEF. In *European conference on machine learning*, 171–182 (Springer, 1994).

45. Robnik-˘ Sikonja, M. & Kononenko, I. An adaptation of Relief for attribute estimation in regression. In *Machine Learning: Proceedings of the Fourteenth International Conference (ICML'97)*, 296–304 (1997).
46. Peng, H., Long, F. & Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis Mach. Intell.* **27**, 1226–1238 (2005).
47. John, G. H. & Langley, P. Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 338–345 (Morgan Kaufmann Publishers Inc., 1995).
48. Quinlan, J. R. Decision trees and decision-making. IEEE Transactions on Syst. *Man, Cybern.* **20**, 339–346 (1990).
49. Cohen, W. W. Fast effective rule induction. In *Proceedings of the twelfth international conference on machine learning*, 115–123 (1995).
50. Breiman, L. Random forests. *Mach. learning* **45**, 5–32 (2001).
51. Chang, C.-C. & Lin, C.-J. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems technology (TIST)* **2**, 27 (2011).
52. Robin, X. *et al*. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC bioinformatics* **12**, 77 (2011).
53. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biom*. 837–845 (1988).
54. National Lung Screening Trial, https://wiki.cancerimagingarchive.net/display/NLST/National+Lung+Screening+Trial.
55. Long and Short Survival in Adenocarcinoma Lung CTs, https://wiki.cancerimagingarchive.net/display/DOI/Long+and+Short+Survival+in+Adenocarcinoma+Lung+CTs.
56. LungCT-Diagnosis, https://wiki.cancerimagingarchive.net/display/Public/LungCT-Diagnosis.
57. The Cancer Imaging Archive, http://www.cancerimagingarchive.net/.
58. Heterogeneity detection Matlab source code, https://github.com/VisionAI-USF/TextureHeterogeneityDetection.

Acknowledgements

This work was partly supported by the Swiss National Science Foundation with grant agreements PZ00P2_154891 and 205320_179069. This research was partially supported by the National Institutes of Health under grants 4KB17. This research was partially supported by the Florida Department of Health under grant (U01 CA143062) and (U24 CA180927). The authors thank the National Cancer Institute for access to NCI's data collected by the National Lung Screening Trial. The statements contained herein are solely those of the authors and do not represent or imply concurrence or endorsement by NCI.

Author Contributions

D.C., H.M. and A.D. conceived the presented idea. A.D. developed the theory. D.C. performed the computations. H.M., A.D, D.G. and L.H. verified the analytical methods from a machine learning perspective. R.G. and M.S. verified the analytical methods from a clinical prospective. All authors discussed the results and contributed to the final manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-38831-0.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019