SCIENTIFIC REPORTS
nature research

OPEN

Identifying relationships between imaging phenotypes and lung cancer-related mutation status: *EGFR* and *KRAS*

Gil Pinheiro[1,9], Tania Pereira[1,9*], Catarina Dias[1,2], Cláudia Freitas[3,4], Venceslau Hespanhol[3,4], José Luis Costa[4,5,6], António Cunha[1,7] & Hélder P. Oliveira[1,8]

EGFR **and** ***KRAS*** **are the most frequently mutated genes in lung cancer, being active research topics in targeted therapy. The biopsy is the traditional method to genetically characterise a tumour. However, it is a risky procedure, painful for the patient, and, occasionally, the tumour might be inaccessible. This work aims to study and debate the nature of the relationships between imaging phenotypes and lung cancer-related mutation status. Until now, the literature has failed to point to new research directions, mainly consisting of results-oriented works in a field where there is still not enough available data to train clinically viable models. We intend to open a discussion about critical points and to present new possibilities for future radiogenomics studies. We conducted high-dimensional data visualisation and developed classifiers, which allowed us to analyse the results for** ***EGFR*** **and** ***KRAS*** **biological markers according to different combinations of input features. We show that** ***EGFR*** **mutation status might be correlated to CT scans imaging phenotypes; however, the same does not seem to hold for** ***KRAS*** **mutation status. Also, the experiments suggest that the best way to approach this problem is by combining nodule-related features with features from other lung structures.**

Lung cancer is the cancer type leading the incidence and mortality rates[1,2]. This is linked to the fact that it is often diagnosed in an advanced stage, with 15% or less chance of a 5-year survival[3], which magnifies the importance of treatments for advanced-stage disease. In Non-small-cell lung cancer (NSCLC), which constitutes 85% of all cases of lung cancer, certain genomic biomarkers are now considered predictive biomarkers and critical for the prognostic[4]. Epidermal Growth Factor Receptor (*EGFR*) and Kristen Rat Sarcoma Viral Oncogene Homolog *KRAS* are the most frequently mutated gene in lung cancer[5]. *EGFR* mutated is present in 15 to 50% of NSCLC patients from never-smokers[5]. The two most common *EGFR* mutations (deletions in exon 19 and the single amino acid substitution L858R in exon 21) correspond to approximately 85% of the *EGFR* mutations in NSCLC. The other low frequency mutations include: point mutations, deletions, insertions, and duplications correspond to approximated 15% of *EGFR* mutations in NSCLC[6]. Unlike the previous marker, *KRAS* is associated with tobacco use, with only 5 to 10% of *KRAS*-mutant lung cancers arising in never or light smokers[5,7].

Surgically treated NSCLC patients with *EGFR* mutations showed better disease-free survival (DFS) and overall survival (OS) and the opposite was verified for *KRAS*, with worse DFS and OS[8]. For cytotoxic chemotherapy, the role of *EGFR* and *KRAS* as a predictive marker is still unclear[9]; however, it appears that mutant *KRAS* may predict a lack of response to chemotherapy[10]. Regarding target therapy, tumour driver mutations have reliable predictive value and, in fact, they guide treatment decision in clinics[11]. *EGFR* gene is a paradigmatic example, since its activating mutations, namely those located in exon 19 and 21, are associated with better response to target therapy, such as gefitinib and erlotinib[12–14]. Current molecularly-targeted therapies can effectively target specific

[1]INESC TEC - Institute for Systems and Computer Engineering, Technology and Science, Porto, Portugal. [2]Faculty of Engineering, University of Porto, Porto, Portugal. [3]Centro Hospitalar e Universitário de São João, Porto, Portugal. [4]Faculty of Medicine, University of Porto, Porto, Portugal. [5]i3S - Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal. [6]IPATIMUP - Institute of Molecular Pathology and Immunology of the University of Porto, Porto, Portugal. [7]University of Trás-os-Montes and Alto Douro, Vila Real, Portugal. [8]Faculty of Science, University of Porto, Porto, Portugal. [9]These authors contributed equally: Gil Pinheiro and Tania Pereira. *email: tania.pereira@inesctec.pt

biomarkers, decreasing multiple undesirable side effects associated with cancer treatment[15]. Several clinical trials have been performed to evaluate the efficacy and safety of treatments for lung cancer patients with *EGFR* mutations[16]. *EGFR* is a receptor tyrosine kinase that controls the growth and proliferation of cells. The target therapies of *EGFR* in lung cancer are based on small molecular tyrosine kinase inhibitors (EGFR-TKIs) that upon binding reduce intracellular signalling. *KRAS*, albeit the most common mutated oncogene, has been more difficult to target. The *KRAS* biochemistry complexity has hampered the development of direct *KRAS* inhibitors[17]. The current approaches are based on covalent binding of the inhibitors. In particular, the AMG 510 was the first *KRAS* inhibitor to demonstrate anti-tumour activity in clinical trials[18], including advanced NSCLC[19]. MRTX849 is also currently in phase I clinical trials and has the same target[20,21].

Tissue biopsy provides a detailed information of the tissue architecture and topography. However, these biopsies tend to increase medical complications, especially when repeated biopsies are needed. Alternative less-invasive clinical procedures for pathological and molecular classification of the tumour are cytological samples and liquid biopsy. Cytological sample can provide adequate cellular material for accurately diagnose and characterize lung cancer[22]. These can be obtained from broncho-alveolar lavage, bronchial washings, bronchial brush smears, pleural fluid, sputum and guided fine needle aspiration cytology of mediastinal and metastatic lymph nodes[23]. A limitation of this option is the limited number of tumour cells that is possible to collect. The liquid biopsy is even less-invasive and it allows the molecular classification through a simple blood sample, analysing the circulating tumour cells and/or circulating tumour DNA[24]. A limitation of this strategy is the high sensitive methodologies required to characterize the circulating genetic material. Thus, there is the urge to find a non-invasive ways to shape the treatment[25]. Medical image analysis can help to solve these issues in two ways: by providing tools capable of measuring characteristics of the lung and, more specifically, the tumour; and with models that use only image features to obtain results through automatic or semi-automatic processes. These models can either use qualitative features, obtained by semantic annotations from human observers, or use quantitative features, obtained through a radiomic approach, which extracts features directly from the image[26]. Radiogenomics, a specific field within radiomics, is defined by the correlation between quantitative features, directly extracted from radiological images (imaging phenotype), and genetic information (genotype)[27]. Studies in lung cancer have presented the association between *EGFR* mutation status and quantitative features extracted from computed tomography (CT) scans[27–30]. The most recent methods are based on convolutional neural networks, which are end-to-end approaches that allow to automatically learn the whole process, reducing the subjectivity and human effort[27,31]. Also, regarding qualitative features, recent works have shown that human semantic annotations of CT scans can be used to train a model to accurately predict *EGFR* mutation status, although the same was not verifiable for *KRAS*[32,33].

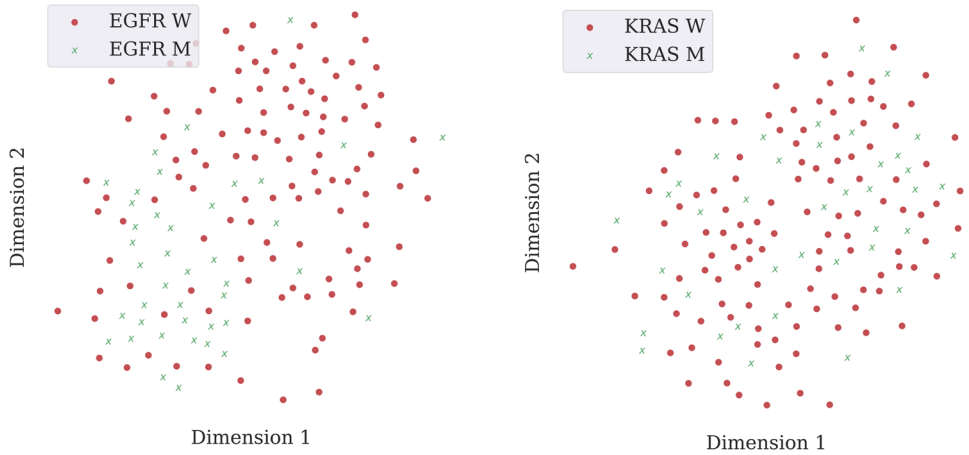
Our previous work was a preliminary study which used a public database[26,34,35] to create predictive models for *EGFR* and *KRAS*[33]. In the current work, we apply a more robust approach based on multiple splits to define the train and test sets, preventing an eventual bias from specific patients and effectively assessing the variance in the data. Additionally, this study aims to provide further advances and to open new discussions in the lung cancer radiogenomics field by exploring the data and building machine learning models, while considering different subsets of inputs. More specifically, predictive models for *EGFR* and *KRAS* mutation status in lung cancer were developed. Following the current direction in the literature, where the analysis only focuses on the nodule structure and texture[36,37], we started by using objective radiomic features directly extracted from nodules in CT scans. Then, semantic features, annotated during radiologist evaluation, were used as input. Unlike their radiomic counterpart, they comprise not only nodule characteristics, but also lung characteristics external to the tumour. Clinical features as patient's gender and smoking status were considered due to its significant association with mutation status prevalence, confirmed in recent studies[38–40]. Moreover, the comparison between its results and those found in the literature review is also presented and is suggested a new perspective about gene mutation status prediction based on image analysis.

Results

Data visualisation. When using Principal Component Analysis (PCA) followed by t-distributed Stochastic Neighbour Embedding (t-SNE) for dimensionality reduction, it is possible to conclude that the separation of classes between mutated and wild type *EGFR* gene status is better when using hybrid semantic features. However, the separation is not perfect, as there are samples outside their cluster, which illustrates the level of complexity faced in a classification process (Fig. 1a). Contrarily, for *KRAS*, there is no visible separation between classes with any type of input features (Fig. 1b). The remaining data visualisation images can be found in Supplementary Fig. 1.

Classification results. Mean values of Area Under the Curve (AUC) were reported for 100 random data splits, with a division of 80% and 20% for training and testing, respectively. Two main types of input features were considered: radiomic and semantic. The semantic were further divided into features that only describe the nodule, features that only describe structures external to nodule and a hybrid between the previous two. Radiomics were not further divided as they only describe the nodule.

Only the predictive models for *EGFR* showed relevant results, with a maximum mean AUC of 0.7458 ± 0.0877 using the hybrid semantic features (represented by the mean ROC curve in Fig. 2). The second best result was obtained using non-nodule semantic features. The worst results were obtained using features only from the nodule, using radiomic and semantic type of features. For *KRAS*, it was not possible to build any acceptable model. Table 1 shows the performance results obtained by each model trained with different groups of the features. Mean and standard deviation of AUC were determined for 100 of different splits for training and test. The performance results confirm the difficulty of gene mutation status classification, which is visible in the t-SNE projections, where there was not possible to achieve a clear separation between classes (Fig. 1).



(a) *EGFR* t-SNE using hybrid semantic features.

(b) *KRAS* t-SNE using hybrid semantic features.

Figure 1. Visualisation of sample distributions based on PCA and t-SNE. Each point is coloured according to its mutation status, with red dots and green crosses representing the wild type and mutated cases, respectively.

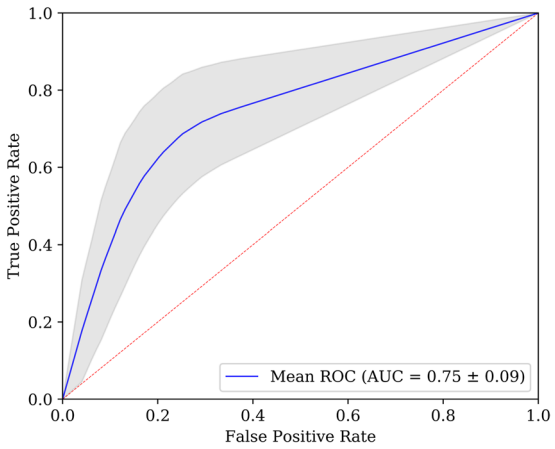


Figure 2. Averaged ROC curve obtained for *EGFR* predictive model based on semantic features. For each of the N = 100 runs, the ROC curve is calculated. The blue line depicts the arithmetic average ROC curve and the shading the standard deviation. The red dashed lines indicate ROC curves of at-chance classifiers.

Features	AUC (mean ± standard deviation)	
	EGFR Mutation Status	*KRAS* Mutation Status
Radiomic	0.5797 ± 0.1238	0.5087 ± 0.0104
Semantic Nodule	0.6542 ± 0.0953	0.4381 ± 0.0679
Semantic Non-Nodule	0.6831 ± 0.0890	0.4921 ± 0.0851
Semantic Hybrid	0.7458 ± 0.0877	0.5035 ± 0.0776

Table 1. Classification results for *EGFR* and *KRAS* mutation status predictive models considering different sets of input features.

Most relevant features. A subset of features, ranked by importance for the most successful model (*EGFR* mutation status prediction using hybrid semantic features), is presented in Fig. 3. They were selected using a minimum threshold of 0.02 and add up to cumulative importance of 0.92 out of 1. The complete list of features importance can be seen in the Supplementary Table 3.

Discussion and Future Work

The results of the present study suggest that even though *EGFR* mutation status is correlated to CT scans imaging phenotypes, the same does not hold true for *KRAS* mutation status. We hypothesise that this might be due to two reasons: mutated and wild type *KRAS* display identical imaging phenotypes, which is supported by the literature[32,41,42], or our number of samples was too small and unrepresentative to find a relevant pattern for such a complex problem.

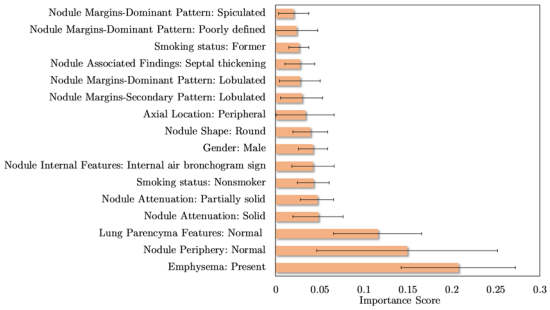


Figure 3. Top 16 semantic features based on the importance scores of features, measured via XGBoost, for predicting the *EGFR* mutation status. Were represented the features that have an average importance score greater than a 0.02. For each of the N = 100 runs, the importance score is determined and the average and standard deviation is displayed in the bar graph.

The outcomes of this work also indicate that general lung semantic features in conjunction with tumour specific semantic features should be used in order to obtain the best possible *EGFR* mutation status classification results. Only average results were obtained using semantic features that solely describe aspects external to the nodule. The worst performances come from models that only use tumour-describing set of features, either of the radiomic or semantic types. This, combined with the fact that the most relevant features (as determined by the classifier) were tumour external (Fig. 3), might hint towards the importance of a holistic lung analysis, instead of a local nodule analysis. Although no previous works profoundly discuss or highlight this particular implication of *EGFR* mutation in imaging phenotypes, there are experiments in the literature that agree with this statement. For example, previous works already showed the importance of extra-tumoral features to obtain a successful *EGFR* mutation status classifier[30,32]. The most recent review and meta-analysis of CT and clinical characteristics to predict the risk of *EGFR* mutation confirms that CT features with the highest correlation with *EGFR* mutation are from the nodule and other structures of the lung[43]. Also, another work based on deep learning techniques with an interpretable visual output, identified that the regions surrounding the nodule were the most relevant for the classification decision[31,44]. In our opinion, it is crucial to emphasise this characteristic, as it might change the direction and broaden the analysis spectrum of future radiogenomics studies, which until now have been mainly focusing on the nodule or in a region of interest (ROI) around it[16,45,46]. Lung cancer is the result of multiple and complex combinations of morphological, molecular and genetic alterations[47]. Since there is a large spectrum of clinicopathological processes that occur during the lung cancer development, it is only natural that important information for the predictive models can be obtained from a larger region of analysis that contains other structures from the lung.

The biggest limitation of this work is the reduced size of the used dataset, which hardly is a good representation of the population affected by lung cancer. In order to better understand the variance in the data and to ensure that the outcomes are not highly influenced by a small number of cases, we used 100 random combinations of cases as train and test sets, reporting the mean values and the standard deviations for the conducted experiences. This limitation is common to the studies in this field, since in general datasets are small and based on patients cohorts from only one medical centre. A reproducible and clinically viable predictive model needs a large and heterogeneous cohort of patients and methods capable of coping with the inherent data heterogeneity[27]. So, a reliable model would require a reliable dataset, collected from multiple centres in order to capture the heterogeneity of the population, but under an uniform protocol to avoid any inconsistency during data record. The access to the data and the uniform acquisition represent the main limitation to build a large dataset. Different protocols used for the data acquisition restrain the mixture of data from different clinical institutions. Additionally, because of privacy issues, the clinical and imaging data is extremely difficult to obtain and requires a large amount of time and effort to submit the protocol to the ethical committees and get approval, and there are also indirect barriers such as fees and data management requirements[48]. Another limitation of the current study is the number of genes taken into account. The two most frequent gene mutations with lung cancer were selected; however other genes were significantly frequent in this type of cancer, and their study could play an important role for novel target therapies, even more personalised and effective. Finally, the stratification of the dataset based on the different mutations is relevant as they may provide different clinical information. The recent approval for the KRAS G12C mutations is a clear example[18]. However, this rare publicly available dataset is small and so does not provide statistical power to address this.

In the future, new radiomic features should be extracted from radiological images and explored according to this study results. That is, features that reflect the state of pulmonary structures external to the tumour nodule combined with nodule features. This would allow us to have a more complete, objective and automatic outlook on the lung, probably delivering more accurate and robust classifiers for *EGFR* mutation status prediction. Another important future work is to build a large dataset more representative of the feature populations. A large dataset will allow us to build more robust models which can deal with heterogeneities of the population, also could allow us to study other types of gene mutation status, and the stratification of the population by different mutations.

Materials and Methods

Dataset. The NSCLC-Radiogenomics dataset[26,34,35] comprises data collected between 2008 and 2012 from a cohort of 211 patients with NSCLC referred for surgical treatment, being the only public dataset which comprehends information regarding the mutation status of lung cancer-related genes (*EGFR*, *KRAS* and *ALK*). It contains a set of CT images stored in DICOM format. Since the samples were retrospectively collected, the scanning protocol and scanning parameters were not standardised; thus slice thickness varied from 0.625 mm to 3 mm (median:1.5 mm) and the X-ray current from 124 mA to 699 mA (mean: 220 mA) at 80-140 kVp (mean: 120 kVp). The subjects were in the supine position with their arms to the side, while the scans were acquired from the top of the lung to the adrenal gland during a single breath[26]. The nodules segmentation masks are stored as DICOM Segmentation Objects[49] and are represented as 3D binary images, where voxels belonging to the tumour ROI are represented by the value 1 and voxels outside the tumour ROI are represented by 0. In the cases where the segmentation mask images did not have the same dimensions as their corresponding CT images, the appropriate number of slices was added to the segmentation mask.

Molecular data. Despite including a cohort of 211 NSCLC subjects, only 116 (wild type: 93, mutant: 23) were further considered in the presented radiomic study for *EGFR* mutation status prediction and 114 (wild type: 88, mutant: 26) for *KRAS* mutation status prediction. The scarce availability of tumour masks and target labels did not allow all subjects to be used. Also, Anaplastic lymphoma kinase (*ALK*), which is the third most frequent oncogene mutated in lung cancer[5], was not targeted by this study, as the prevalence of mutated cases was too small (wild type: 108, mutant: 2).

Patients were referred for surgical treatment, and the surgical samples were used to obtain molecular characterisation. Molecular data for EGFR, KRAS, ALK were obtained using gene expression microarrays, or RNA sequencing, or both[50]. SNaPshot technology based on dideoxy single-base extension of oligonucleotide primers after multiplex polymerase chain reaction (PCR) was used for single nucleotide mutation detection. For *EGFR* mutations the exons 18, 19, 20 and 21 were tested. For missense *KRAS* mutations the exon 2 positions 12 and 13 were tested[50].

Clinical features. Clinical features were added to the radiomic features as well as to the semantic features to build the predictive models. From now on, we only mention the name of the main group of features that contribute to the models, i.e., radiomic and semantic features. Supplementary Table 1 shows detailed information regarding the clinical data distribution and nomenclature.

Radiomic features. There are image properties, such as the distance between slices, which may differ from scan to scan, and consequently affect the features extracted and the learning ability of the algorithms. Therefore, before trying to extract patterns, the images went through a preprocessing step in order to standardise the scans across the whole dataset.

Firstly, the CT image values were converted to Hounsfield Unites (HU), which is a measure of radiodensity. The equation for computing the HU values based on radiodensity is shown in Eq. 1, where μ represents the original linear attenuation coefficient of substance, μ_{water} represents the linear attenuation coefficient of water and μ_{air} the linear attenuation of air[51].

$$HU = 1000 \times \frac{\mu - \mu_{water}}{\mu_{water} - \mu_{air}} \quad (1)$$

By default, the values returned by the CT scanner are not in this unit. In such manner, the radiodensity values were converted to HU units, by multiplying the voxel value by the slope and adding the intercept related to the linear transformation, which values are stored in the metadata of the scans. With the purpose of learning patterns from the data using an automatic analysis methodology, it is extremely important that a pixel is represented in the same way in the entire dataset. Having this in mind, the entire dataset (including the tumour masks) were resampled so that neighbour slices and adjacent pixels are separated by 1 mm in x, y and z directions. Images were normalised between −1000 HU and 400 HU, since −1000 HU is the radiodensity of air and values above 400 HU represent hard tissues, not relevant for the task at hand[52]. Values under −1000 HU or above 400 HU were defined as −1000 HU and 400 HU, respectively.

From the 3D images of the nodules of the pre-processed CT scans, a set of 1218 radiomic features were extracted using the open-source package *Pyradiomics*[53]. Features were computed both on the original image and on images obtained after application of wavelet and Laplacian of Gaussian (LoG) filters. A wavelet transform decouples textural information by decomposing the original image in low and high frequencies. A 3D undecimated wavelet transform was applied to each CT image, which decomposed the original image into 8 different images. Considering that L is a low-pass filter and H a high-pass filter, the original image X is decomposed into 8 new images after the wavelet decomposition: X_{LLL}, X_{LLH}, X_{LHH}, X_{LHL}, X_{HHH}, X_{HHL}, X_{HLL}, X_{HLH}. For instance, X_{LHL} is obtained after applying a low-pass filter along the x-dimension, a high-pass filter along the y-dimension and a low-pass filter along the z-dimension. The remaining images are built similarly, applying their respective sequence of low or high-pass filters in x, y and z-direction[54]. Concerning the LoG, five filters with different sigma values were applied (sigma = 1.0 mm, 2.0 mm, 3.0 mm, 4.0 mm, 5.0 mm), to improve texture analysis by detection of multi-scale edges and ridges[55]. In summary, considering the original image and the resulting images after filter application, there were 14 different images to extract features for each sample.

Six classes of features were extracted from the *Pyradiomics* package: shape-based features (14 features), first-order features (18 features), GLCM features (22 features), GLRLM features (16 features), GLSZM features

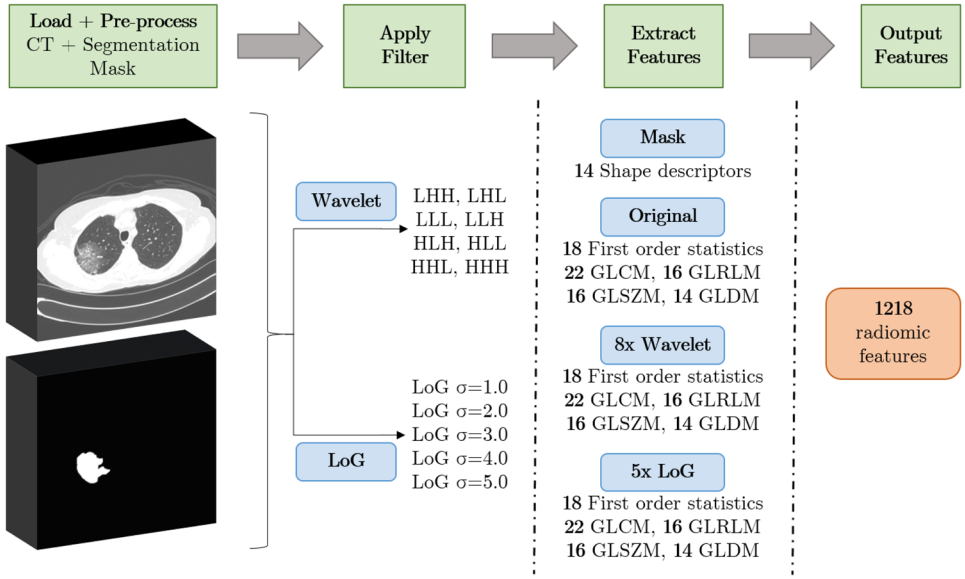


Figure 4. Overview of the process of feature extraction via *Pyradiomics*. First, medical images and segmentation masks are loaded into the software. This step allows to select the region of the tumour. Then, after filters have been applied to the original image, radiomic features are extracted from the ROI of the resultant images.

(16 features) and GLDM features (14 features). Shape features include different descriptors of the size and shape of the ROI. Having in mind that shape descriptors are independent of intensity values, only the tumour segmentation masks were required for its computation. First-order features are related to the voxel intensities within the ROI using basic metrics (e.g. mean and standard deviation). GLCM features describe the second-order joint probability function of the ROI. GLRLM features define the length of successive pixels that share the same grey level intensity. GLSZM features quantify the grey level zones in an image, which represent the number of connected voxels that have an equal grey level value. At last, GLDM features quantify grey level dependencies in an image. A straightforward overview of the steps involved with the feature extraction is presented in Fig. 4.

Semantic features. The dataset comprises a set of subjects whose tumour was analysed by radiologists using 30 nodule and parenchymal features, which describe nodule's geometry, location, internal features and other related findings. From these subjects, 158 are characterised in terms of *EGFR* mutation status and 157 subjects characterised in terms of *KRAS* mutation status, which were the samples selected for the presented semantic study.

The semantic terms that were used to characterise the patients are common in radiology clinical practice and derive from descriptions in the radiology literature[26]. Definitions of some of the terms used in this description can be found in[56]. The template of semantic terms was developed by two academic thoracic radiologists exclusively for tumours with identifiable nodules and excluded cases without this manifestation. More information about the semantic annotations protocol can be found at[26].

From the original set of semantic features, some were discarded due to a large number of not applicable values (e.g. the fibrosis type field in a patient that has fibrosis absent), thus, only 18 features were used in the final study. The final dataset comprises percentages of 26% and 25% mutated cases for *EGFR* and *KRAS*, respectively. Supplementary Table 2 shows detailed information regarding the semantic data distribution and nomenclature. Before feeding the data into the model, features were binarised following a one-hot encoding strategy. After that, the number of features increased from 20 to 73.

Feature engineering and selection. Considering semantic features, most *Lung Parenchyma* categories are underrepresented (see Supplementary Tables 1 and 2), with *Normal* and *Bronchial wall thickening* making up to 79.1% and 77.2% of the present *Lung Parenchyma* categories in the *EGFR* and *KRAS* datasets, respectively. To balance the occurrences, we binarise this feature, putting the *Normal* category in a group and the remaining in another, creating a new category titled *Not normal*.

Both semantic and radiomic features were submitted through a process of feature selection, where the correlation matrix was computed, and a correlation threshold of 0.95 between variables was set. Additionally, the least importance radiomic features were excluded. This was done by taking the feature importances from a gradient boosting machine algorithm and only keeping the ones necessary to achieve a cumulative importance of 0.95.

Dimensionality reduction. We use Principal Component Analysis (PCA)[57] followed by t-Distributed Stochastic Neighbour Embedding (t-SNE)[58] to reduce our high-dimensional data to a two-dimensional space, in order to investigate the existence of class separation between *EGFR* and *KRAS* wild type and mutated samples.

PCA allows to find the minimum number of variables that minimise information loss from the original data. This is done by creating new uncorrelated variables (principal components) that maximise variance, which comes down to solving an eigenvector problem. t-SNE is used to further reduce the data dimension to a 2D space. In order to reduce the data dimension, this method minimises the divergence between pairs of input samples (high-dimensional space) and pairs of the corresponding points in the embedding (low-dimensional space) using a cost function.

Balancing training set. In general, machine learning algorithms assume a similar distribution of classes. Here, *EGFR* wild type is over-represented, which could result in a model biased towards this class. However, in this study, the correct classification of both classes is equally important, as the classification of a patient with the wrong mutation status could lead to the administration of a less suitable treatment and, consequently, to shorter progression-free survival. To overcome this class imbalance, Synthetic Minority Over-sampling Technique - Nominal and Continuous (SMOTE-NC) was applied, an extended version of SMOTE generalised to handle data with continuous and nominal features[59]. This technique creates new random synthetic minority class instances between the lines that connect each one of the *n* nearest neighbours of each minority class sample. In comparison to traditional over-sampling, SMOTE-NC has the advantage of building a more general decision region of the minority class. After this algorithm is applied, the training set contains the same number of mutated and wild type samples.

Classification and feature importance. The classifier used in this work was Extreme Gradient Boosting (XGBoost), which is a scalable and accurate implementation of gradient boosted trees algorithms[60] that has been used for lung cancer related works[61,62]. A benefit of using gradient boosting is that after the boosted trees are constructed, it is possible to retrieve the importance scores for each feature, based on how useful or valuable each feature was in the construction of the boosted decision trees within the model. Their importance is calculated for a single decision tree by the amount that each attribute split point improves the performance measure, weighted by the number of observations the node is responsible. The feature importance is then averaged across all of the decision trees within the model.

Training and performance metrics. The training and testing processes were repeated for 100 random splits of the original dataset. Each split comprised a training and test sets consisting of 80% and 20% of the original data, respectively. The mean and standard deviation for all the 100 results were reported in favour of reliability and to demonstrate the variance in the data. The classifier hyper-parameters were tuned through a 5-fold cross-validated randomised search on the training data, maximising the models F-measure. The data is balanced individually for each fold using SMOTE-NC, avoiding data leakage. After parameter optimisation, probabilistic outputs of each model with optimal parameters were analysed using the AUC of Receiver Operating Characteristic (ROC). ROC is a probability curve, and AUC represents the degree or measure of separability, telling how much model is capable of distinguishing between classes. The ROC curve is plotted with True Positive Rate (TPR) against the False Positive Rate (FPR), usually with TPR on the y-axis and FPR on the x-axis.

Experimental design. We designed four experiments in order to test and compare which type of input features allow to achieve better performance in gene mutation status prediction. We first trained a model that took nodule-related radiomic features as input. Then, for direct comparison purposes and to allow a modular evaluation, we split the semantic data into three parts: *nodule*, *non-nodule* and *hybrid*. The first one contains only nodular information, the second one contains only information external to the nodule, and the third one is the combination of both. The split can be seen in detail in Table 2 of the supplementary material.

Accession codes. The developed code is available on GitLab (https://gitlab.inesctec.pt/ippr-pub/lucasradsemegfrkras).

Data availability

The data was obtained from the open-access NSCLC-Radiogenomics dataset publicly available at The Cancer Imaging Archive (TCIA) database[26,34,35]. Imaging and clinical data have been de-identified by TCIA and approved by the Institutional Review Board of the TCIA hosting institution. Ethical approval was reviewed and approved by the Washington University Institutional Review Board protocols. Informed consent was obtained from all individual participants included in the study.

Received: 21 October 2019; Accepted: 29 January 2020;
Published online: 27 February 2020

References

1. Ferlay, J. *et al*. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *International Journal of Cancer* (2015).
2. World Health Organisation. Latest global cancer data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018. *International Agency for Research on Cancer* (2018).
3. Janssen-Heijnen, M. L. & Coebergh, J.-W. W. Trends in incidence and prognosis of the histological subtypes of lung cancer in north america, australia, new zealand and europe. *Lung cancer* **31**, 123–137 (2001).
4. Rose-James, A. & Tt, S. Molecular Markers with Predictive and Prognostic Relevance in Lung Cancer. *Lung Cancer International* (2012).
5. Jorge, S. E., Kobayashi, S. S. & Costa, D. B. Epidermal growth factor receptor (EGFR) mutations in lung cancer: Preclinical and clinical data (2014).

6. Harrison, P. T., Vyse, S. & Huang, P. H. Rare epidermal growth factor receptor (EGFR) mutations in non-small cell lung cancer. *Seminars in Cancer Biology* 1-13 (2019).
7. Ferrer, I. *et al*. KRAS-Mutant non-small cell lung cancer: From biology to therapy (2018).
8. Zhang, S. M. *et al*. Prognostic value of EGFR and KRAS in resected non-small cell lung cancer: A systematic review and meta-analysis. *Cancer Management and Research* (2018).
9. Fang, S. & Wang, Z. EGFR mutations as a prognostic and predictive marker in non-small-cell lung cancer (2014).
10. Martin, P., Leighl, N. B., Tsao, M. S. & Shepherd, F. A. KRAS mutations as prognostic and predictive markers in non-small cell lung cancer (2013).
11. Planchard, D. *et al*. Metastatic non-small cell lung cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology* (2018).
12. Lynch, T. J. *et al*. Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non-Small-Cell Lung Cancer to Gefitinib. *New England Journal of Medicine* (2004).
13. Paez, J. G. *et al*. EGFR mutations in lung, cancer: Correlation with clinical response to gefitinib therapy. *Science* (2004).
14. Pao, W. *et al*. EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib. *Proceedings of the National Academy of Sciences of the United States of America* (2004).
15. Schrank, Z. *et al*. Current molecular-targeted therapies in NSCLC and their mechanism of resistance (2018).
16. Zhao, W. *et al*. Toward automatic prediction of EGFR mutation status in pulmonary adenocarcinoma with 3D deep learning. *Cancer Medicine* (2019).
17. Tomasini, P., Walia, P., Labbe, C., Jao, K. & Leighl, N. B. Targeting the KRAS Pathway in Non-Small Cell Lung Cancer. *The Oncologist* (2016).
18. Canon, J. *et al*. The clinical KRAS(G12C) inhibitor AMG 510 drives anti-tumour immunity. *Nature* (2019).
19. Fakih, M. *et al*. Phase 1 study evaluating the safety, tolerability, pharmacokinetics (PK), and efficacy of AMG 510, a novel small molecule KRAS G12C inhibitor, in advanced solid tumors. *Journal of Clinical Oncology* (2019).
20. Adderley, H., Blackhall, F. H. & Lindsay, C. R. KRAS-mutant non-small cell lung cancer: Converging small molecules and immune checkpoint inhibition. *EBioMedicine* **41**, P711–716 (2019).
21. Mullard, A. Cracking KRAS. *Nature Reviews Drug Discovery* (2019).
22. Folch, E., Costa, D. B., Wright, J. & VanderLaan, P. A. Lung cancer diagnosis and staging in the minimally invasive age with increasing demands for tissue analysis (2015).
23. Jain, E. & Roy-Chowdhuri, S. Molecular pathology of lung cancer cytology specimens a concise review (2018).
24. Cai, L. L. & Wang, J. Liquid biopsy for lung cancer immunotherapy (Review) (2019).
25. Rizzo, S. *et al*. CT Radiogenomic Characterization of EGFR, K-RAS, and ALK Mutations in Non-Small Cell Lung Cancer. *European Radiology* (2016).
26. Bakr, S. *et al*. A radiogenomic dataset of non-small cell lung cancer. *Scientific data* **5**, 180202 (2018).
27. Bodalal, Z., Trebeschi, S., Nguyen-Kim, T. D. L., Schats, W. & Beets-Tan, R. Radiogenomics: bridging imaging and genomics (2019).
28. Digumarthy, S. R., Padole, A. M., Gullo, R. L., Sequist, L. V. & Kalra, M. K. Can ct radiomic analysis in nsclc predict histology and egfr mutation status? *Medicine* **98** (2019).
29. Mei, D., Luo, Y., Wang, Y. & Gong, J. Ct texture analysis of lung adenocarcinoma: can radiomic features be surrogate biomarkers for egfr mutation statuses. *Cancer Imaging* **18**, 52 (2018).
30. Liu, Y. *et al*. Radiomic features are associated with egfr mutation status in lung adenocarcinomas. *Clinical lung cancer* **17**, 441–448 (2016).
31. Wang, S. *et al*. Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning. *European Respiratory Journal* (2019).
32. Gevaert, O. *et al*. Predictive radiogenomics modeling of egfr mutation status in lung cancer. *Scientific reports* **7**, 41674 (2017).
33. Dias, C., Pinheiro, G., Cunha, A. & Oliveira, H. P. Radiogenomics: Lung Cancer-Related Genes Mutation Status Prediction. In *IbPRIA 2019: 9th Iberian Conference on Pattern Recognition and Image Analysis* (2019).
34. Clark, K. *et al*. The cancer imaging archive (tcia): Maintaining and operating a public information repository. *Journal of Digital Imaging* **26**, 1045–1057 (2013).
35. Gevaert, O. *et al*. Non-small cell lung cancer: Identifying prognostic imaging biomarkers by leveraging public gene expression microarray data - Methods and preliminary results. *Radiology* (2012).
36. Shen, S., Han, S. X., Bui, A. A. & Hsu, W. An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. *Expert Systems with Applications* (2019).
37. Mei, D., Luo, Y., Wang, Y. & Gong, J. CT texture analysis of lung adenocarcinoma: Can Radiomic features be surrogate biomarkers for EGFR mutation statuses. *Cancer Imaging* (2018).
38. Papadopoulou, E. *et al*. Determination of egfr and kras mutational status in greek non-small-cell lung cancer patients. *Oncology letters* **10**, 2176–2184 (2015).
39. Varghese, A. M. *et al*. Lungs don't forget: comparison of the kras and egfr mutation profile and survival of collegiate smokers and never smokers with advanced lung cancers. *Journal of Thoracic Oncology* **8**, 123–125 (2013).
40. Dogan, S. *et al*. Molecular epidemiology of egfr and kras mutations in 3,026 lung adenocarcinomas: higher susceptibility of women to smoking-related kras-mutant cancers. *Clinical cancer research* **18**, 6169–6177 (2012).
41. Yip, S. S. *et al*. Associations between somatic mutations and metabolic imaging phenotypes in non-small cell lung cancer. *Journal of Nuclear Medicine* (2017).
42. Yip, S. S. *et al*. Impact of experimental design on PET radiomics in predicting somatic mutation status. *European Journal of Radiology* (2017).
43. Zhang, H., Cai, W., Wang, Y., Liao, M. & Tian, S. CT and clinical characteristics that predict risk of EGFR mutation in non-small cell lung cancer: a systematic review and meta-analysis. *International Journal of Clinical Oncology* (2019).
44. Hosny, A. *et al*. Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. *PLoS Medicine* (2018).
45. Wilson, R. & Devaraj, A. Radiomics of pulmonary nodules and lung cancer (2017).
46. Yamashita, R., Nishio, M., Kinh, R., Do, G. & Togashi, K. Convolutional neural networks : an overview and application in radiology. *Insights Imaging* **9**, 611–629 (2018).
47. Davidson, M. R., Gazdar, A. F. & Clarke, B. E. The pivotal role of pathology in the management of lung cancer (2013).
48. Doshi, J. A., Hendrick, F. B., Graff, J. S. & Stuart, B. C. Data, Data Everywhere, But Access Remains a Big Issue for Researchers: A Review of Access Policies for Publicly-Funded Patient-level Health Care Data in the United States. *eGEMs (Generating Evidence & Methods to improve patient outcomes)* (2016).
49. Kahn, C. E., Carrino, J. A., Flynn, M. J., Peck, D. J. & Horii, S. C. Dicom and radiology: past, present, and future. *Journal of the American College of Radiology* **4**, 652–657 (2007).
50. Bakr, S. *et al*. Data descriptor: A radiogenomic dataset of non-small cell lung cancer. *Scientific Data* (2018).
51. Kalra, A. Developing fe human models from medical images. In Yang, K.-H. (ed.) *Basic Finite Element Method as Applied to Injury Biomechanics* (2018).
52. Bolliger, S. A., Oesterhelweg, L., Spendlove, D., Ross, S. & Thali, M. J. Is differentiation of frequently encountered foreign bodies in corpses possible by hounsfield density measurement? *Journal of forensic sciences* **54**, 1119–1122 (2009).

53. Van Griethuysen, J. J. *et al*. Computational radiomics system to decode the radiographic phenotype. *Cancer research* **77**, e104–e107 (2017).
54. Prochazka, A., Grafova, L., Vyšata, O. & Caregroup, N. Three-dimensional wavelet transform in multi-dimensional biomedical volume processing. In *Proc. of the IASTED International Conference on Graphics and Virtual Reality, Cambridge*, 263–268 (2011).
55. Fotin, S. V., Reeves, A. P., Biancardi, A. M., Yankelevitz, D. F. & Henschke, C. I. A multiscale laplacian of gaussian filtering approach to automated pulmonary nodule detection from whole-lung low-dose ct scans. In *Medical Imaging 2009: Computer-Aided Diagnosis*, vol. 7260, 72601Q (International Society for Optics and Photonics, 2009).
56. Hansell, D. M. *et al*. Fleischner society: glossary of terms for thoracic imaging. *Radiology* **246**, 697–722 (2008).
57. Abdi, H. and Williams, L. J. Principal component analysis. In *Encyclopedia of Biometrics* (2009).
58. Maaten, L. v. d. & Hinton, G. Visualizing data using t-sne. *Journal of machine learning research* **9**, 2579–2605 (2008).
59. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research* **16**, 321–357 (2002).
60. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016).
61. Nishio, M. *et al*. Computer-aided diagnosis of lung nodule using gradient tree boosting and Bayesian optimization. *PLoS ONE* (2018).
62. Zhang, X. *et al*. Identification of Cancer-Related Long Non-Coding RNAs Using XGBoost With High Accuracy. *Front. Genet.* **10**, 1–14 (2019).

Author contributions

G.P., T.P., C.D., A.C. and H.O. conceived the experiments, G.P., T.P. and C.D. conducted the experiments, statistical analyses, and manuscript writing. G.P., T.P., J.C., C.F., V.H., A.C. and H.O. performed the clinical interpretation of the results. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-60202-3.

Correspondence and requests for materials should be addressed to T.P.

Reprints and permissions information is available at www.nature.com/reprints.

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