Semantic *Vs* Radiomic Features from CT Images to Predict Gene Mutation Status in Lung Cancer

Tania Pereira[1]
tania.pereira@inesctec.pt
Gil Pinheiro[1]

Catarina Dias[12]

António Cunha[13]
acunha@utad.pt
Hélder P. Oliveira[14]
helder.f.oliveira@inesctec.pt

[1] INESC TEC - Institute for Systems and Computer Engineering, Technology and Science, Portugal
[2] FEUP - Faculty of Engineering, University of Porto, Porto, Portugal
[3] UTAD - University of Trás-os-Montes and Alto Douro, Vila Real, Portugal
[4] FCUP - Faculty of Science, University of Porto, Porto, Portugal

Abstract

In lung cancer, the biopsy is the traditional method to assess the mutation status of the most frequent and relevant oncogenes. Medical imaging, which is already a common source of information in clinical practice, is a potential alternative to the biopsy. It contains a large number of features that, although not visible to the naked eye, may be valuable for tumour characterisation. The recent field of radiomics allows new opportunities for the genomic analysis of a tumour, by extracting hundreds of quantitative features from medical images which, in a non-invasive way, provide a full state visualisation of a tumour at a macroscopic level. This study aimed to investigate in which extent features extracted from medical images are related to helpful genotype factors for tumour characterisation, in particular for *EGFR* and *KRAS* mutation status. Radiomic and semantic features were used for the prediction. The performance of the models demonstrated that *EGFR* (AUC=0.75) mutation status can be differentiated through medical images using semantic features. The experiments suggest that the best way to approach this problem is by combining nodule-related features with features from other lung structures.

1 Introduction

Lung cancer is the cancer type leading the incidence and mortality rates [5]. This is linked to the fact that it is often diagnosed in an advanced stage, which magnifies the importance of treatments for advanced-stage disease. Epidermal Growth Factor Receptor (*EGFR*) and Kristen Rat Sarcoma Viral Oncogene Homolog (*KRAS*) are the most frequently mutated gene in lung cancer [8]. Current molecularly-targeted therapies can effectively target specific biomarkers, decreasing multiple undesirable side effects associated with cancer treatment. 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). Studies in lung cancer have presented the association between *EGFR* mutation status and quantitative features extracted from computed tomography (CT) scans [1, 4].

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. The current paper is an adaptation of our previously published work [9].

2 Material and Methods

2.1 Dataset

The NSCLC-Radiogenomics dataset [7] comprises data collected between 2008 and 2012 from a cohort of 211 patients with Non-small-cell lung cancer (NSCLC) referred for surgical treatment, being the only public dataset which comprehends information regarding the mutation status of lung cancer-related genes (*EGFR* and *KRAS*). It contains a set of CT images stored in DICOM format.

2.1.1 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.

2.1.2 Clinical Features

Clinical features were added to the radiomic features as well as to the semantic features to build the predictive models.

2.1.3 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. The CT image values were converted to Hounsfield Unites (HU), which is a measure of radiodensity. 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* [10]. Features were computed both on the original image and on images obtained after application of wavelet and Laplacian of Gaussian (LoG) filters. 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 (16 features) and GLDM features (14 features).

2.1.4 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.

2.2 Balancing Training Set

In general, machine learning algorithms assume a similar distribution of classes. *EGFR* wild type is over-represented, which could result in a model biased towards this class. 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 [2].

2.3 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 [3]. 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.

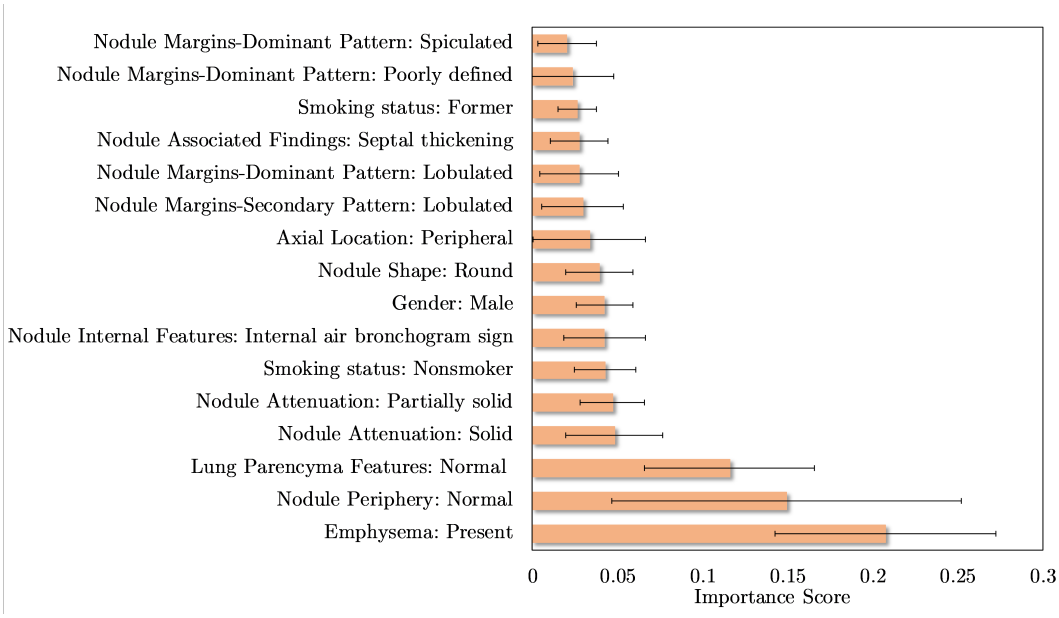

Figure 1: Top 16 semantic features based on the importance scores of features, measured via XGBoost, for predicting the *EGFR* mutation status.

3 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. We designed those four experiments in order to test and compare which type of input features allow to achieve better performance in gene mutation status prediction (Table 1). Only the predictive models for *EGFR* showed relevant results, with a maximum mean AUC of 0.7458 ± 0.0877 using the hybrid semantic features (Table 1). A subset of features, ranked by importance for the most successful model (*EGFR* mutation status prediction using hybrid semantic features), is presented in Figure 1. They were selected using a minimum threshold of 0.02 and add up to cumulative importance of 0.92 out of 1.

AUC	*EGFR* Prediction	*KRAS* Prediction
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.

4 Discussion and Conclusions

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 [6, 11, 12], or our number of samples was too small and unrepresentative to find a relevant pattern for such a complex problem.

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. This, combined with the fact that the most relevant features (as determined by the classifier) were tumour external, might hint towards the importance of a holistic lung analysis, instead of a local nodule analysis. 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 around it [13]. 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.

References

[1] Z. Bodalal, S. Trebeschi, T. D. L. Nguyen-Kim, W. Schats, and R. Beets-Tan. Radiogenomics: bridging imaging and genomics. *Abdominal Radiology*, 2019.

[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[3] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[4] S. R. Digumarthy, A. M. Padole, R. L. Gullo, L. V. Sequist, and M. K. Kalra. Can ct radiomic analysis in nsclc predict histology and egfr mutation status? *Medicine*, 98(1), 2019.

[5] J. Ferlay, I. Soerjomataram, R. Dikshit, et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *International Journal of Cancer*, 2015.

[6] O. Gevaert, S. Echegaray, A. Khuong, et al. Predictive radiogenomics modeling of egfr mutation status in lung cancer. *Scientific reports*, 7:41674, 2017.

[7] O. Gevaert, J. Xu, C. D. Hoang, et al. Non-small cell lung cancer: Identifying prognostic imaging biomarkers by leveraging public gene expression microarray data - Methods and preliminary results. *Radiology*, 2012.

[8] S. E. Jorge, S. S. Kobayashi, and D. B. Costa. Epidermal growth factor receptor (EGFR) mutations in lung cancer: Preclinical and clinical data, 2014.

[9] T. Pereira, G. Pinheiro, C. Dias, et al. Identifying relationships between imaging phenotypes and lung cancer-related mutation status: EGFR and KRAS. *Scientific Reports*, 10:3625, 2020.

[10] J. J. Van Griethuysen, A. Fedorov, C. Parmar, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107, 2017.

[11] S. S. Yip, J. Kim, T. P. Coroller, et al. Associations between somatic mutations and metabolic imaging phenotypes in non-small cell lung cancer. *Journal of Nuclear Medicine*, 2017.

[12] S. S. Yip, C. Parmar, J. Kim, et al. Impact of experimental design on PET radiomics in predicting somatic mutation status. *European Journal of Radiology*, 2017.

[13] W. Zhao, J. Yang, B. Ni, et al. Toward automatic prediction of EGFR mutation status in pulmonary adenocarcinoma with 3D deep learning. *Cancer Medicine*, 2019.

*/