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

Tania Pereira(1), Gil Pinheiro(1), Catarina Dias(1,2), Antonio Cunha(1,3), Helder Oliveira(1,4)

1 - INESC TEC - Institute for Systems and Computer Engineering, Technology and Science, Portugal;; 2 - Faculty of Engineering, University of Porto, Portugal.; 3 - University of Trás-os-Montes and Alto Douro, Portugal; 4 - Faculty of Science, University of 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.
This study aimed to investigate in which extent features extracted from CT images are related to helpful genotype factors for tumor gene mutation status characterization. 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.

Introduction

EGFR and KRAS are the most relevant oncogenes for the management of lung cancer. Current molecularly-targeted therapies can effectively target specific biomarkers, decreasing multiple undesirable side effects associated with cancer treatment. Radiogenomics is defined by the correlation between quantitative features, directly extracted from radiological images (imaging phenotype), and genetic information (genotype). 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.

[1] - Radiomics and its emerging role in lung cancerresearch, imaging biomarkers and clinical management: state of the art.European journalof radiology, 86:297–307, 2017.

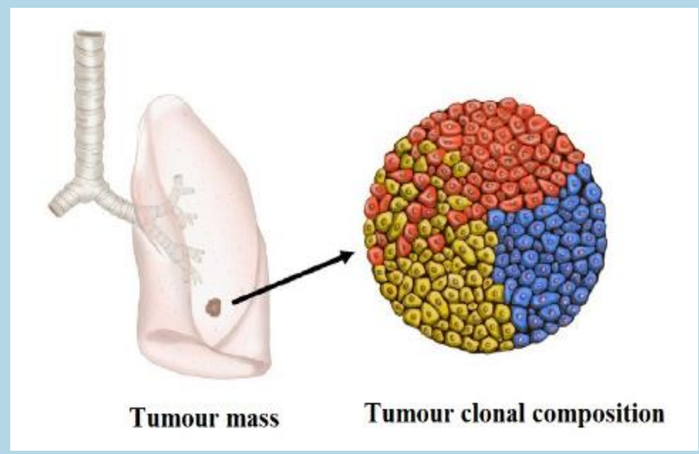

Figure 1. Tumour mass composition. Adapted from Lee et al. [1]

Methods

The NSCLC-Radiogenomics dataset from a cohort of 211 patients with Non-small-cell lung cancer (NSCLC) referred for surgical treatment, with information regarding the mutation status of lung cancer-related genes *EGFR* and *KRAS*. Only 116 patients were further considered in the presented radiomic study for *EGFR* mutation status prediction (wild type: 93, mutant: 23) and 114 for *KRAS* mutation status prediction (wild type: 88, mutant: 26). Radiomic features from the CT scans and semantic features annotated by the radiologists were used. The Extreme Gradient Boosting (XGBoost) classifier was used in this work and it allows to retrieve the importance scores for each feature.

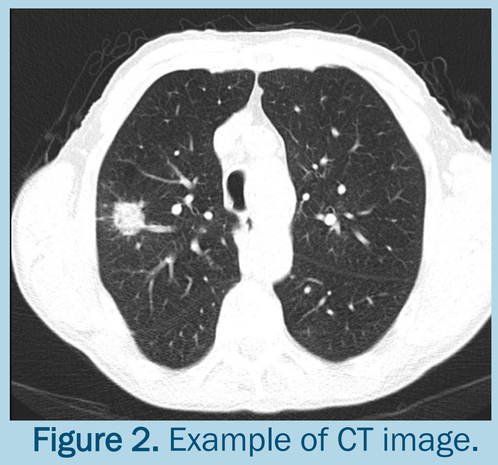

Figure 2. Example of CT image.

Results

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). A subset of features, ranked by importance for the most successful model (*EGFR* mutation status prediction using hybrid semantic features), is presented in Figure 3. 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.

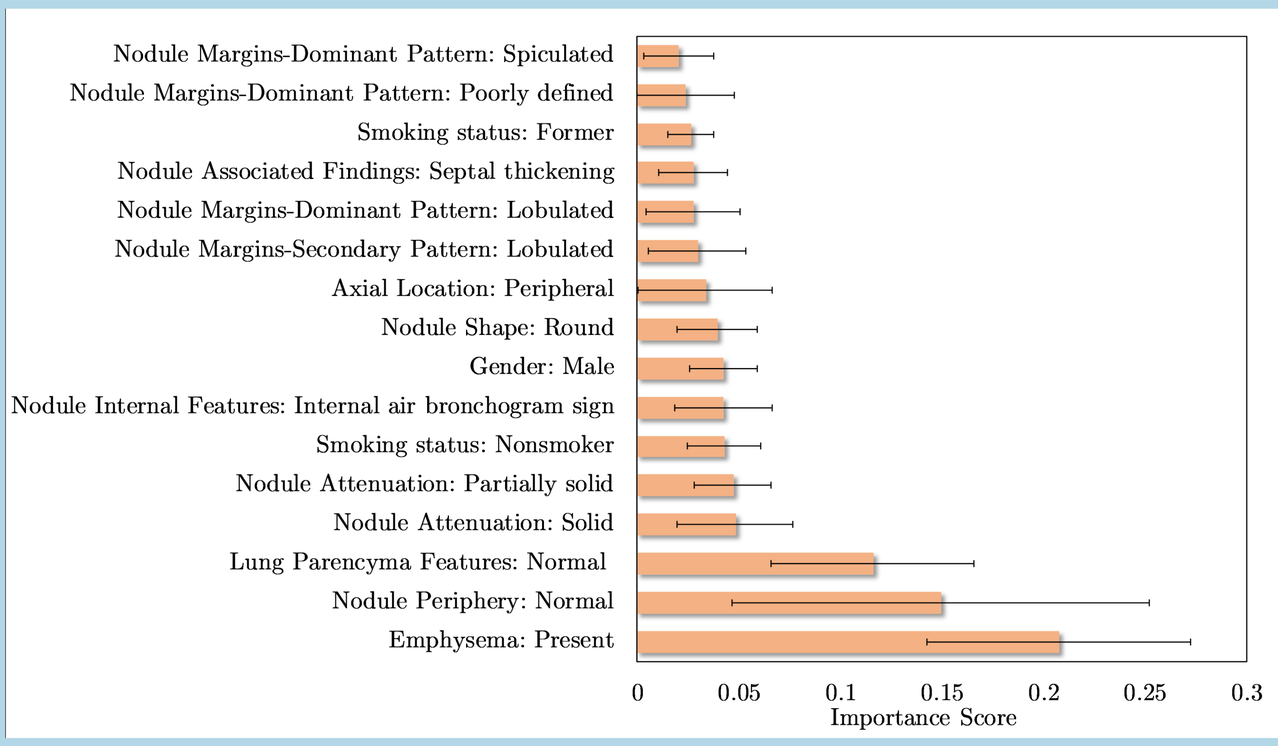

Figure 3. Features importance scores.

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. The outcomes of this work also indicate that general lung semantic features in conjunction with tumor specific semantic features should be used in order to obtain the best possible *EGFR* mutation status classification results.