2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Ensemble Strategies for *EGFR* Mutation Status Prediction in Lung Cancer

Mafalda Malafaia, Tania Pereira, Francisco Silva, Joana Morgado, António Cunha and Hélder P. Oliveira *(Member, IEEE)*

***Abstract*—Lung cancer treatments that are accurate and effective are urgently needed. The diagnosis of advanced-stage patients accounts for the majority of the cases, being essential to provide a specialized course of treatment. One emerging course of treatment relies on target therapy through the testing of biomarkers, such as the Epidermal Growth Factor Receptor (*EGFR*) gene. Such testing can be obtained from invasive methods, namely through biopsy, which may be avoided by applying machine learning techniques to the imaging phenotypes extracted from Computerized Tomography (CT). This study aims to explore the contribution of ensemble methods when applied to the prediction of *EGFR* mutation status. The obtained results translate in a direct correlation between the semantic predictive model and the outcome of the combined ensemble methods, showing that the utilized features do not have a positive contribution to the predictive developed models.**

I. Introduction

Lung cancer is the leading oncologic disease both in incidence and mortality rates [1]. Non-Small Cell Lung Cancer (NSCLC) represents about 80 - 85% of all histological subtypes of lung cancer [2]. Approximately 50% – 60% of patients with NSCLC have at least one identifiable driver mutation, with one of the most common mutations being in the Epidermal Growth Factor Receptor (*EGFR*) gene [3]. The discovery of oncogenic driver mutations has led to new ways of classifying NSCLC and offered the opportunity to develop target therapies.

CT is a medical imaging technique greatly used for screening and analysis of lung cancer. Additionally, the identification of biomarkers translates into the detection of specific gene mutations, among them the *EGFR*, which provides a more accurate and personalized choice of adequate therapy. The detection of gene mutation status is achieved nowadays through invasive procedures such as biopsy and liquid biopsy, which may be avoided by applying machine learning techniques to CT characteristics [4], [5], [6]. Statistical methods were initially developed in order to identify radiomic features associations with both clinical and biological data, which were able to prove a relationship between the *EGFR* mutation status, clinical data, and automatically computed data from CT scans. On the other hand, semantic features are able to describe lung pathologies and are also useful in learning models to predict the mutation status of the *EGFR* gene [7]. Therefore, both radiomic and semantic data have shown the potential for the detection of the *EGFR* mutation status, leaving still room for improvement. With the intuition of improving the performance results whilst taking advantage of all the available data, the combination of the two different sets of features inherently arises, along with the study of the resemblance and complementarity of one another.

It is known that a clear association exists between radiomic and semantic data since both types of information are extracted from CT scans. In fact, there are some features of both groups that are highly correlated and, therefore, tend to provide similar contributions to the predictive model [8]. Radiomic Data represent a continuous range of values that are automatically computed and, therefore, yields a more detailed characterization that may not be perceived by radiologists at the naked eye. On the other hand, Semantic Data consist of binary features that otherwise could be poorly identified by continuous features. Despite having a partially significant correlation among them, radiomic and semantic data have complementary information with different potential to characterize the lung phenotype [8]. Therefore, the motivation to combine the two sets of features with different prediction potentials arises, as does the need to study and develop approaches that take advantage of both types of information that a CT can provide. Ensemble methods integrate multiple models and their prediction to compute a final classification decision that outperforms the otherwise weak learners.

This paper proposes ensemble learning techniques to combine radiomic and semantic features in order to predict the *EGFR* mutation status. This approach allows a more comprehensive analysis of lung cancer and as a consequence, it is expected an improvement of the performance results.

II. Materials and Methods

This section presents the dataset used in this work, the explored feature extraction techniques and the ensemble classification methods. The pipeline implemented to predict the *EGFR* mutation status is represented in Fig. 1.

A. Dataset

The NSCLC-Radiogenomics Dataset [9] comprises imaging and molecular information for a cohort of 211 patients with Non-Small Cell Lung Cancer (NSCLC). However, only patients with provided nodule segmentation masks and semantic annotations for EGFR mutation status were further considered. In order to fairly evaluate the contribution of each

M. Malafaia, T. Pereira and F. Silva are with the INESC TEC, Portugal.
A. Cunha is with the INESC TEC, Portugal and UTAD, Portugal.
J. Morgado and H. P. Oliveira are with the INESC TEC and FCUP, Portugal.

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project UIDB/50014/2020.

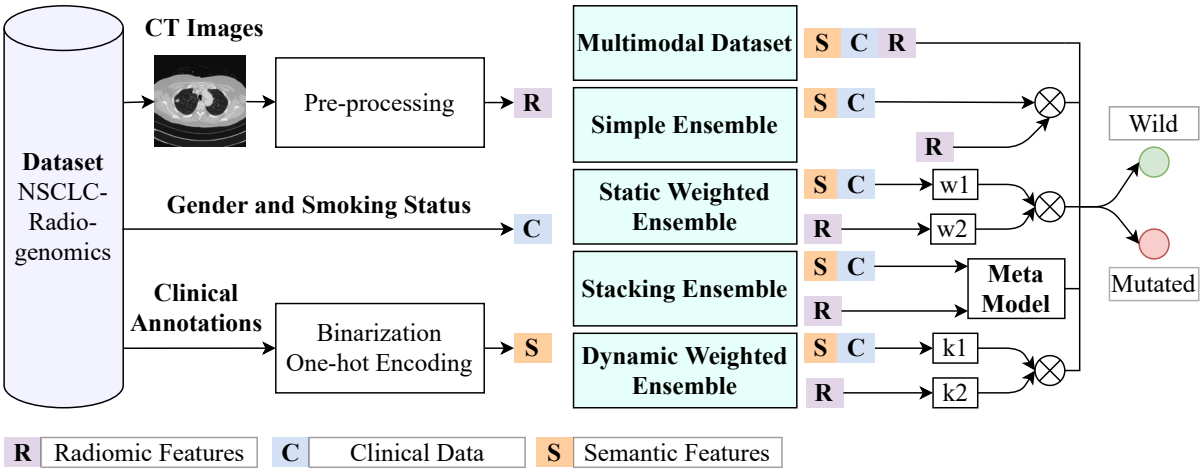

Fig. 1: Pipeline developed for *EGFR* mutation status classification based on ensemble methods to combine patient clinical data, radiomic and semantic features.

type of feature to the final classification, there was the need to reassure that the patients had both nodule segmentation for radiomic feature calculation, and semantic annotations. Hence, the number of samples was inherently reduced by cross eliminating the data that was not available for both groups, leading to a partial loss of data samples, resulting in a total of 116 patients (93 wildtype and 23 mutant cases). Regarding data acquisition protocol, slice thickness ranges from 0.625 mm to 3 mm, with median equal to 1.5 mm [9]. The database used in the experiments of this work ensures, on the correspondent cited description paper, that the necessary ethical approvals regarding data access were obtained.

B. Pre-processing

This step is a fundamental proceeding for the extraction of radiomic features, which are the quantitative characteristics from the CT scans (Fig. 1). For pre-processing of the obtained images, pixel representations were standardized by resampling, setting the pixel spacing in x, y and z directions to 1 mm. The pixel intensity value, measured in the Hounsfield Units (HU) scale, was normalized using the *min-max* normalization method, with the normalization window from -1000 to 400 HU. The same resampling operation was also applied to the provided tumor binary masks to match the correspondent CT dimensions. The segmentation of the region of interest (ROI) is a central process within radiomics pipeline since the extracted features depend on the segmented volumes. The provided tumor segmentation ground-truth were reviewed by thoracic experienced radiologists before being publicly available.

C. Features

Three types of features were used to better predict the *EGFR* mutation status: semantic, radiomics and clinical information (see Fig. 1). The semantic features were obtained from observation of CT images by experienced radiologists, to characterize the lung in clinical practice, and include 30 nodule and parenchymal features, which describe geometry, location, internal features of the nodule, and other related findings. Those features were binarised through one-hot encoding process, consequently increasing the number of features. Medical images contain a large number of features, which may be valuable for tumor characterization; however, most of them are not visible to the naked eye but can be capture by quantitative feature extraction from the images. Radiomics allow to transform radiographic medical images into a high-dimensional feature space by quantifying tumor phenotype characteristics. Radiomics offers a vast group of quantitative features that can be extracted from the ROI of the nodule. From the 3D tumor of the pre-processed CT images, a set of 1218 radiomic features were extracted using the open-source package *Pyradiomics* [10], and they can be divided into different classes (histogram-based, morphological, and texture features). After extracting the whole set of radiomic features, the latter were submitted to a feature selection process, where a correlation matrix was computed and a correlation threshold of 0.95 was applied, following a removal of the features with lower importance by the implementation of a gradient boosting machine to compute the cumulative feature importance. Beyond radiomic and semantic features, clinical data were also used in the study, which comprised gender and smoking status information.

D. Baseline Reference Results

In order to achieve a fair comparison between each type of features and the developed ensemble methods, the previously mentioned public database was utilized to create the baseline approach based in the predictive models from a previous work [7].

E. Ensemble Classification

In Machine Learning-based approaches, ensemble methods allow the combination of multiple predictions in order to develop a more reliable predictive model. Several popular

predictive ensemble models were applied in order to explore the pool of approaches that allows for a combination of two learners trained with the same samples but different types of features, being able to use all the available information in the dataset simultaneously:

1) Multimodal dataset trained with XGBoost;
2) Simple Ensemble: train Semantic model first and feed its prediction as features along with Radiomic features as final model;
3) Static Weighted Ensemble of Semantic and Radiomic Predictions;
4) Stacking Ensemble with Meta-Model;
5) Dynamic Weighted Ensemble of Semantic and Radiomic Predictions.

1) Multimodal dataset trained with XGBoost: A simple XGBoost model was trained using all data modalities. Semantic and Radiomic data were concatenated and used as features to the learning model.

2) Simple Ensemble: train Semantic and Radiomic features consecutively: A simple cascade model was implemented, in which the semantic data was trained firstly, following its predictions and the radiomic features as input for the final model. This design aims to improve the performance of the individual semantic model by joining the radiomic features with the previously trained semantic predictions.

3) Static Weighted Ensemble of Semantic and Radiomic Predictions: The output classification of the model in question results from a weighted average of predictions from both radiomic and semantic data. To do so, both data types were trained separately with the XGBoost model and their probability predictions were given different weights, in order to obtain the final classification. The goal of this design is to obtain a more reliable and accurate result than the individually trained models by averaging their predictions with different contributions to the final classification.

4) Stacking Ensemble with Meta-Model: A simple linear meta-model outputs the final predictions. Along with the weighted average approach, both feature sets were trained separately with a simple XGBoost model. To obtain the final classification, several meta-models were studied in order to use the one that showed better performance with the given dataset, such as Logistic Regression, Linear Regression and SVM with linear kernel. Each tested meta-model was given as input the decision probabilities of the previously learned models, returning a final probability decision.

5) Dynamic Weighted Ensemble of Semantic and Radiomic Predictions: A innovative approach is suggested in this work, developed with the intuition of dynamically assigning a different importance to each prediction of the trained sets of features, according to the confidence of each model in each different sample. To achieve that, a confidence measurement was defined as the distance of a sample on a specific predictive outcome to the value 0.50. The higher this distance, the higher the confidence that a prediction has on its decision, and, therefore, the higher the contribution of that submodel to the final decision. This confidence parameter was built in order to influence a weighted average of both predictions according to the ratio of confidences between both models in each sample.

Given the motivation, the developed simple algorithm follows Equation 1:

$$y_{pred} = \alpha \frac{d_{sem}}{d_{rad}} y_{sem} + (1-\alpha) \frac{d_{sem}}{d_{rad}} y_{rad} \quad (1)$$

d_{sem} and d_{rad} represent the confidence measurements of each sample for both attributes, being α a proportionality constant that is experimentally optimized with the training set. Both y_{sem} and y_{rad} are the output predictions of the previously trained XGBoost trained models with semantic and radiomic features, respectively. The weight of a given sample in each of the attributes increases with the confidence ratio of the value on the mentioned probability, whilst decreasing with the confidence ratio value of the remaining attribute. The final classification is computed according to the dynamically changed weight of each model for each patient.

F. Training

All the ensemble classification approaches described previously were trained in the same conditions, in order to fairly evaluate their outcomes and be able to draw some conclusions. All patient samples with semantic annotations and radiomic features were trained and tested over 100 random splits into 80% for training and 20% for testing. The performance of each experimental model was computed from the average and standard deviation of the total number of splits, obtaining the AUC of the Receiver Operating Characteristic (ROC) for each ensemble approach.

III. Results

The baseline models were trained in order to evaluate the performance obtained using only radiomic or semantic features. The results using the radiomic features showed a lower performance (AUC of 0.578±0.138) comparing with model based on the semantic features (AUC of 0.703±0.112). To study the combination between Radiomic and Semantic data and better understand the results, one can compare the individual performances and use them to justify the disparities of results when joining the features. All the baseline performance values are described in Table I.

Individual Data	AUC
Radiomic	0.578±0.138
Semantic	**0.703±0.118**

TABLE I: Performance results of the individual types of features for baseline purposes.

In Table II, the performance results were summarized. The highest performance was achieved by the Static Weighted Ensemble (AUC of 0.705±0.112), with a weight of 0.90 for the semantic model and 0.10 for the radiomic model. The performance in question was able to slightly outperform

Ensemble Method	AUC
Multimodal Dataset	0.682±0.122
Simple Ensemble	0.665±0.113
Stacking Ensemble*	0.643±0.131
Static Weighted Ensemble**	**0.705±0.119**
Dynamic Weighted Ensemble***	0.691±0.134

∗ The best meta model within the tested ones was the Logisic Regression.
∗∗ The optimal set weights consisted on 0.90 for the semantic predictions and 0.10 for the radiomic predictions, on both AUC calculation results.
∗ ∗ ∗ The optimized alfa constant was given the value of 1.2.

TABLE II: Performance results of each Ensemble Method.

the baseline reference results. The Dynamic Weighted model (AUC of 0.691±0.134) and the Multimodal Dataset model (AUC of 0.682±0.122) showed very similar results.

IV. Discussion

The results of the present work lean towards a significant correlation between the mutation status of the EGFR gene and CT scan imaging phenotypes, namely by combining both different type of features than can be extracted from the said medical images. However, when making a fair comparison between each used ensemble method and the baseline computed results, one can state that the combination of the semantic features with the radiomic features reveal a general decrease towards the semantic features whilst showing an increase regarding the radiomic features. It is possible to draw the hypothesis that the significant discrepancy between the two different groups of features rather lowers the performance of the best model with the semantic features, than the otherwise expected increase of performance when combining two different learners with different information. With the computed results, the radiomic features seem to have a significantly lower predictive capability, which can be suggestive of a need to look for different sets of features that may have a more interesting target potential towards the EGFR mutation status detection. The said new types of features may reveal more reliable results and with better performance, for instances by demonstrating a more objective characterization of the lung and, perhaps, leading to potentially more explainable results. Another possible improvement of the obtained performance in this work could result from the expansion of the area of the lung that is accessed for feature extraction, since previous works have shown that the use of general lung characteristics can improve the performance results [7]. Regarding each one of the attempted approaches, it is clear that the ensemble models with best performance are the ones that implement a weighted average on both prediction probabilities that are received as input. Being the Static Weighted Ensemble Approach directly related to each output decision of each set of features, there seems to be a pattern where the higher the semantic predictions weight, the better the final performance is. This can only confirm that the radiomic predictions appear to have a negative contribution to the final performance results, being the optimal model almost exclusively the semantic predictions. The Dynamic Weighted Ensemble, on the other hand, does not have such a direct comparison to either one of the submodels. Despite not showing a better performance than the baseline, the computed AUC was really similar to the baseline previously acquired performance, which may imply that a relatively more complex development of the used algorithm would be interesting to study and access its results. One can also notice that the multimodal dataset model, as the simplest non-ensemble addressed model in this study, revealed AUC values very close to the baseline ones, being almost as efficient as the best ensemble methods discussed.

V. Conclusions

In the present work an exploratory study was proposed on Ensemble Methods that allow for the combination of all the available features that are extracted from CT images. Ensemble methods provide a joint prediction with a contribution from semantic, radiomic and clinical features. The developed models were able to achieve the same performance than the semantic features trained individually, not being able to outperform the baseline result. Therefore, the need for the improvement of the radiomic available data once again is reinforced, being imperial to find innovative approaches that are able to increase the present performances, namely regarding more advanced feature extraction methods. The developed approach can potentially provide an accurate methodology to inform physicians.

References

[1] J. A. Barta, C. A. Powell, and J. P. Wisnivesky, "Global epidemiology of lung cancer," 2019.

[2] T. Sher, G. K. Dy, and A. A. Adjei, "Small cell lung cancer," *Mayo Clinic Proceedings*, vol. 83, no. 3, pp. 355–367, 2008.

[3] M. B. Antonoff, M. B. Antonoff, and J. D'Cunha, "Non-small cell lung cancer: The era of targeted therapy," pp. 31–41, 2012.

[4] Z. Cheng, F. Shan, Y. Yang, Y. Shi, and Z. Zhang, "CT characteristics of non-small cell lung cancer with epidermal growth factor receptor mutation: A systematic review and meta-analysis," *BMC Medical Imaging*, vol. 17, no. 1, 2017.

[5] J. Zou, T. Lv, S. Zhu, Z. Lu, Q. Shen, L. Xia, J. Wu, Y. Song, and H. Liu, "Computed tomography and clinical features associated with epidermal growth factor receptor mutation status in stage I/II lung adenocarcinoma," *Thoracic Cancer*, vol. 8, no. 3, pp. 260–270, 2017.

[6] S. Rizzo, S. Raimondi, E. E. de Jong, W. van Elmpt, F. De Piano, F. Petrella, V. Bagnardi, A. Jochems, M. Bellomi, A. M. Dingemans, and P. Lambin, "Genomics of non-small cell lung cancer (NSCLC): Association between CT-based imaging features and EGFR and K-RAS mutations in 122 patients—An external validation," *European Journal of Radiology*, vol. 110, pp. 148–155, 2019.

[7] G. Pinheiro, T. Pereira, C. Dias, C. Freitas, V. Hespanhol, J. L. Costa, A. Cunha, and H. P. Oliveira, "Identifying relationships between imaging phenotypes and lung cancer-related mutation status: EGFR and KRAS," *Scientific Reports*, 2020.

[8] S. S. Yip, Y. Liu, C. Parmar, Q. Li, S. Liu, F. Qu, Z. Ye, R. J. Gillies, and H. J. Aerts, "Associations between radiologist-defined semantic and automatically computed radiomic features in non-small cell lung cancer," *Scientific reports*, vol. 7, no. 1, pp. 1–11, 2017.

[9] S. Bakr, O. Gevaert, S. Echegaray, K. Ayers, M. Zhou, M. Shafiq, H. Zheng, J. A. Benson, W. Zhang, A. N. Leung, M. Kadoch, C. D. Hoang, J. Shrager, A. Quon, D. L. Rubin, S. K. Plevritis, and S. Napel, "Data descriptor: A radiogenomic dataset of non-small cell lung cancer," *Sci. Data*, vol. 5, 2018.

[10] J. J. Van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J. C. Fillion-Robin, S. Pieper, and H. J. Aerts, "Computational radiomics system to decode the radiographic phenotype," *Cancer Research*, 2017.