Received March 12, 2021, accepted March 25, 2021, date of publication April 2, 2021, date of current version April 22, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3070701

EGFR Assessment in Lung Cancer CT Images: Analysis of Local and Holistic Regions of Interest Using Deep Unsupervised Transfer Learning

FRANCISCO SILVA[1,2], TANIA PEREIRA[1], JOANA MORGADO[1,3], JULIETA FRADE[1,2], JOSÉ MENDES[1,2], CLÁUDIA FREITAS[4,5], EDUARDO NEGRÃO[4], BEATRIZ FLOR DE LIMA[4], MIGUEL CORREIA DA SILVA[4], ANTÓNIO J. MADUREIRA[4,5], ISABEL RAMOS[4,5], VENCESLAU HESPANHOL[4,5], JOSÉ LUÍS COSTA[6,7], ANTÓNIO CUNHA[1,8], AND HÉLDER P. OLIVEIRA[1,3], (Member, IEEE)

[1]Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), 4200-465 Porto, Portugal
[2]Faculty of Engineering, University of Porto (FEUP), 4200-465 Porto, Portugal
[3]Faculty of Science, University of Porto (FCUP), 4169-007 Porto, Portugal
[4]Centro Hospitalar e Universitário de São João (CHUSJ), 4200-319 Porto, Portugal
[5]Faculty of Medicine, University of Porto (FMUP), 4200-319 Porto, Portugal
[6]Instituto de Investigação e Inovação em Saúde (i3S), Universidade do Porto, 4200-135 Porto, Portugal
[7]Institute of Molecular Pathology and Immunology, University of Porto (IPATIMUP), 4200-135 Porto, Portugal
[8]University of Trás-os-Montes and Alto Douro (UTAD), 5001-801 Vila Real, Portugal

Corresponding author: Francisco Silva (francisco.c.silva@inesctec.pt)

This work was supported in part by the European Regional Development Fund (ERDF) through the Operational Program for Competitiveness and Internationalization—COMPETE 2020 Program, and in part by the National Funds through the Portuguese Funding Agency, Fundação para a Ciência e a Tecnologia (FCT), under Project POCI-01-0145-FEDER-030263.

ABSTRACT Statistics have demonstrated that one of the main factors responsible for the high mortality rate related to lung cancer is the late diagnosis. Precision medicine practices have shown advances in the individualized treatment according to the genetic profile of each patient, providing better control on cancer response. Medical imaging offers valuable information with an extensive perspective of the cancer, opening opportunities to explore the imaging manifestations associated with the tumor genotype in a non-invasive way. This work aims to study the relevance of physiological features captured from Computed Tomography images, using three different 2D regions of interest to assess the Epidermal growth factor receptor (*EGFR*) mutation status: nodule, lung containing the main nodule, and both lungs. A Convolutional Autoencoder was developed for the reconstruction of the input image. Thereafter, the encoder block was used as a feature extractor, stacking a classifier on top to assess the *EGFR* mutation status. Results showed that extending the analysis beyond the local nodule allowed the capture of more relevant information, suggesting the presence of useful biomarkers using the lung with nodule region of interest, which allowed to obtain the best prediction ability. This comparative study represents an innovative approach for gene mutations status assessment, contributing to the discussion on the extent of pathological phenomena associated with cancer development, and its contribution to more accurate Artificial Intelligence-based solutions, and constituting, to the best of our knowledge, the first deep learning approach that explores a comprehensive analysis for the *EGFR* mutation status classification.

INDEX TERMS Convolutional autoencoder, *EGFR* prediction, lung cancer, transfer learning, unsupervised feature learning.

I. INTRODUCTION

Lung cancer still presents high incidence and mortality rates [1]. Despite the relevant impact of early detection, the identification of the biomarkers associated with cancer development could define a personalized treatment plan based on target therapies, which would contribute to the survival rate increase [2]. Target therapies are drugs that have an effect on specific molecules responsible for growth, progression, and spread of cancer. These therapies act on

The associate editor coordinating the review of this manuscript and approving it for publication was Sotirios Goudos.

specific molecular targets increasing the success of the treatment response and decreasing the side effects, due to the specificity of their action. However, the development of these therapies requires the identification of the biomarkers [3]. In lung cancer, the two most relevant oncogenes are: Epidermal growth factor receptor (*EGFR*) and Kirsten rat sarcoma viral oncogene homolog (*KRAS*) [4]. *EGFR* is a predictive biomarker with clinically approved therapies [5]. *KRAS* has shown to be more difficult to target due to the biochemistry complexity and there are no *KRAS* inhibitors as an approved therapy [6], [7]. For this reason, the *EGFR* mutation status identification is the most important oncogene in the treatment-decision pathway. Traditionally, the oncogene mutation status is assessed by molecular testing using the tissues extracted during the biopsy. Recently, less invasive and more automatic techniques, such as computer-aided diagnosis (CAD) based on Computed Tomography (CT) analysis, have been developed, decreasing the risk for the patients and improving the accuracy of the diagnosis [8], [9].

Radiogenomics approaches, using CT images for lung cancer characterization, have recently been explored with a small number of publications and very limited by the small size of the available databases with those type of data (thoracic CT scans and molecular results for oncogene mutation status). Despite that, works based on the traditional statistical analysis to classify the *EGFR* mutation status for lung cancer patients [10], [11] showed that there are radiomic signatures in CT images that can be used to distinguish the *EGFR* mutated from the wild type. Semantic features annotated by radiologists from CT scans were used to feed a decision tree and achieved an Area Under the Curve (AUC) of 0.89 [12]. A more complex approach based on an ensemble of decision trees - XGBoost classifier - was used to predict the *EGFR* mutation status and obtained an AUC of 0.75 [13]. Exploratory studies that took into consideration semantic features from multiple lung structures, not focusing only on the nodule (traditional approach), showed the importance of including extra-tumoral features to obtain a successful *EGFR* mutation status classifier [10], [12], [13]. The use of more powerful Artificial Intelligence (AI)-based methods has shown to be able to automatically capture relevant information from CT images while avoiding *ad hoc* features extraction. A region of interest (ROI) containing the tumor from CT scans was used in a deep learning (DL) model similar to DenseNet, pre-trained on the ImageNet dataset [14], with Transfer Learning techniques for the *EGFR* classification, achieving an AUC of 0.85 [15]. A 3D DenseNet was developed to process 3D patches of lung nodules from CT data, and to learn representations with supervised end-to-end training, which combined with radiomic features, obtained an AUC of 0.76 for automatic *EGFR* prediction status [16]. Ensemble machine learning and DL models were proposed for a final decision and allowed the fusion of the radiomics-based model and the multi-level residual convolutional neural network (CNN) based model, which obtained an AUC value of 0.83 [17]. Using patches centered on the nodule, CNN-based features in conjunction with a Support Vector Machine (SVM) achieved an AUC of 0.83 [18].

The previous works have taken into consideration different regions for oncogene prediction, although usually centered on the nodule [14]–[16], [18]. The approaches based on the semantic features annotated by radiologists, which evaluate imagiological findings from the nodule and other lung structures [10], [12], [13], allow a more comprehensive analysis of the lung pathological processes associated with cancer development, which seems to indicate that cancer development is related to multiple physiological changes not restricted to the nodule region. However, the contribution of the amount of information of one region that is taken into consideration for *EGFR* status prediction was not studied. This work is focused on two main challenges in *EGFR* assessment: the study of the ROI that captures a more comprehensive analysis of the biological problem, and the development of an approach to overcome the lack of massive annotated datasets. This study proposes an innovative comparison of ROIs based on a binary classification to assess the *EGFR* mutation status using information not only from the nodule but also considering a larger ROI including the lung where the nodule is located or both lungs in the selected CT axial slice. The main motivation behind this evaluation relied on the hypothesis that it might be possible to find relevant information related to *EGFR* mutation status outside of the tumor ROI. On the other hand, the methodology proposed in this work was developed to allow the use of databases without molecular information. A Convolutional Autoencoder (CAE) was initially trained with unlabelled data, using the intrinsic mechanisms of reconstruction of the input image to train the encoder [19]. The encoder layers from the pre-trained CAE were used as feature extractor, stacking a classifier on top to be completely trained for the *EGFR* mutation status prediction. The proposed approach based on CAE pre-trained with unlabelled data allows to overcome one of the biggest limitations on the medical domain: the access of massive annotated datasets.

This work is organized as follows: the information on the datasets used, the proposed Transfer Learning methodology and the performed experiments inherent to each analysis are detailed in Section II; the results of the comparative study are presented and discussed in Section III; the conclusions are presented in Section IV.

II. MATERIALS AND METHODS

A. DATASETS

For this study, two databases that allow to achieve the defined objectives were identified: one containing thoracic CT scans with binary masks for pulmonary nodules used for the feature extraction phase, and another database with CT scans from lung cancer patients comprising the *EGFR* mutation status information.

TABLE 1. Summary of the number of images used regarding each task. The total number of images values are presented after the slice oversampling operations further detailed (see Figure 4).

Database	Included Patients	Total Number of Images	
		Nodule Analysis	Lung Analysis
LIDC-IDRI [20]	875	8007 (3 × 2669)	875
NSCLC-Radiogenomics [22]	116	Mutant: 217 (9 × 23) Wild type: 279 (3 × 93)	Mutant: 69 (3 × 23) Wild type: 93

1) LIDC-IDRI

The LIDC-IDRI [20], [21] is a lung cancer screening dataset that comprises thoracic CT scans for a total of 1010 patients, alongside with annotated nodules belonging to one of three classes: a) nodule $\geq$ 3 mm; b) nodule $<$ 3 mm or c) non-nodule $\geq$ 3 mm, made during a two-phase annotation process by four experienced radiologists. Regarding data acquisition, slice thickness ranged from 0.6 to 5.0 mm, with X-ray current from 40 to 627 mA (mean: 222.1 mA) at 120-140 kVp. From the 7371 detected lesions, only 2669 were classified as larger than 3 mm by at least one clinician. These were the examples included in this study given the availability of nodule contours marked by each radiologist (see Table 1).

2) NSCLC-RADIOGENOMICS

The NSCLC-Radiogenomics dataset [22] is a public available collection with CT images for a cohort of patients with non-small cell lung cancer (NSCLC), being the only public dataset that comprises paired information on lung cancer-related gene mutation status and CT data. Additionally, semantic tumor annotations are included in a controlled vocabulary as well as binary tumor masks, although not available for the entire set of subjects. This dataset includes CT scans obtained using different scanner models and scanning protocols, presenting variations in slice thickness from 0.625 to 3 mm (median: 1.5 mm) and X-ray tube current from 124 to 699 mA (mean 220 mA) at 80–140 kVp (mean 120 kVp) [22]. From the NSCLC-Radiogenomics data collection, despite including a cohort of 211 patients, only 116 were selected due to a required *EGFR* mutational test result of *Mutant* or *Wildtype* and the availability of tumor binary mask. From these 116 included patients, 23 patients (20%) belonged to the *Mutant* class, and 93 (80%) to the *Wildtype* class (see Table 1).

3) DATA PREPARATION

CT scans from both databases were resampled to standardize image representations. The pixel spacing was set to [1.00, 1.00, 1.00] mm and each CT dimensions were calculated to match this new spacing, obtaining the resampled image by interpolation [23]. Additionally, each pixel intensity value, measured in the Hounsfield Units (HU) scale, was normalized using the *min-max* normalization method, and values under −1000 HU, which corresponds to air's radiodensity value, were transformed into 0 and values above 400 HU, representing hard tissues like bones, were transformed into 1. A linear transformation was computed to map all values in the middle into the [0, 1] intended range. In the final result, each database was composed by images with size $N \times 512 \times 512$, with N representing the number of slices of the correspondent CT scan. A CT slice example from each one of the described databases is represented in Figure 1. In Table 1, a summary of the number of samples used regarding each task is presented.

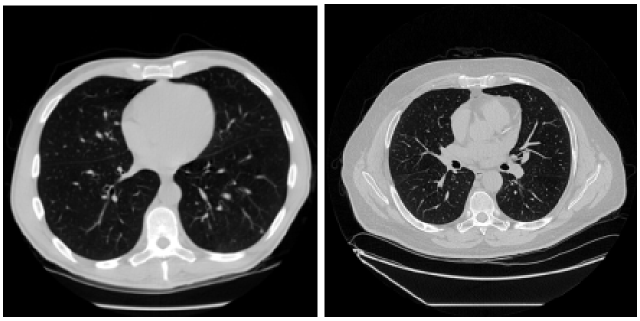

FIGURE 1. Illustrative examples of CT slices from an LIDC-IDRI (left) [20], and NSCLC-Radiogenomics patients (right) [22], resultant from the pre-processing procedure.

B. PREDICTIVE APPROACH

The proposed approach in this work is composed by two main phases: a feature learning task, where a Convolutional Autoencoder is implemented and trained on images in the lung cancer domain; and a second task, which consists of developing an end-to-end classification model based on Transfer Learning techniques, using the trained convolutional encoder and a classifier to predict the *EGFR* mutation status. The overview of the proposed approach is represented in Figure 2.

1) FEATURE LEARNING

Being widely explored to overcome the lack of publicly available data in the medical imaging field [24], Transfer Learning has proven to allow the use of deeper architectures by using pre-trained neural networks on massive datasets, like ImageNet [14], which significantly reduces the number of trainable parameters. In this work, given the scarce dataset size available to perform the target task, we considered an alternative approach, consisting of developing and training the feature extractor in the same domain of the final task [25]. This Transfer Learning strategy was chosen based on the intuition that the trained encoder would be capable of achieving the necessary general knowledge in the

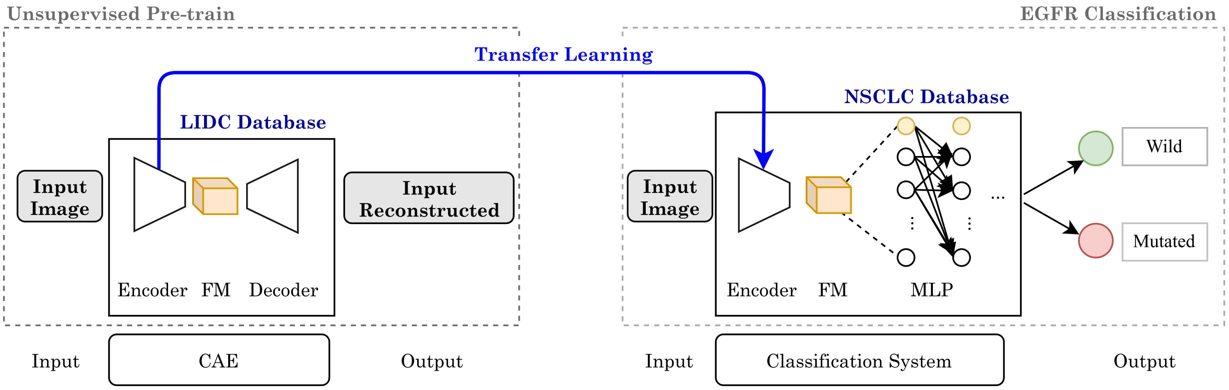

FIGURE 2. Overview of the proposed approach based on the unsupervised pre-training of the CAE to be used as feature extractor of the CT images, and an end-to-end classifier to predict the *EGFR* mutation status. Transfer Learning allows to reuse the encoder trained with unlabelled data (LIDC-IDRI [20] database) as a feature extractor for the *EGFR* classification (NSCLC-Radiogenomics [22] database).

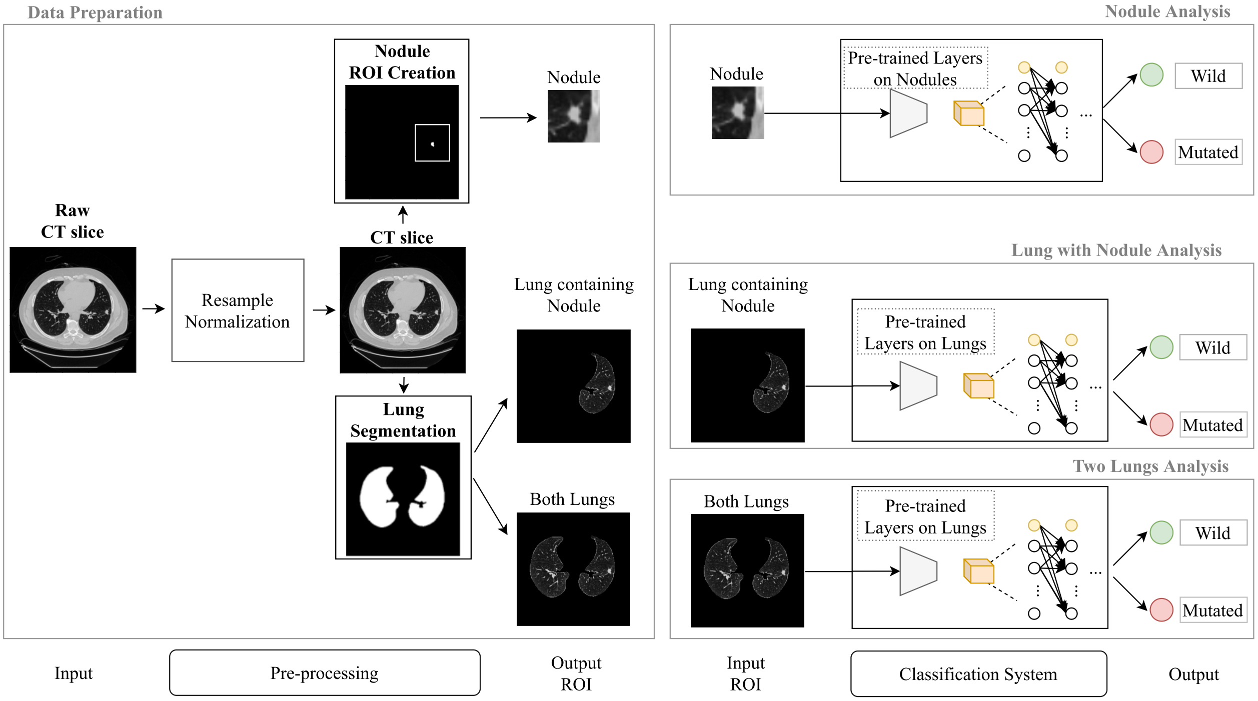

FIGURE 3. Overview of the three ROIs selected for *EGFR* mutation status prediction as wild or mutated: nodule, lung containing nodule and both lungs.

lung cancer domain, and intends to explore the relevance of the learned patterns while training an encoder-decoder based architecture for input reconstructions (see Figure 2). More specifically, it was investigated whether the knowledge achieved by the pre-trained encoder, in an unsupervised way, could be useful in the detection of relevant *EGFR*-related patterns. For the CAE development in this phase, the LIDC-IDRI [20] extracted samples were used in both nodule and lung analyses.

2) EGFR MUTATION STATUS CLASSIFICATION

To the end-to-end classification task, a multi-layer perceptron (MLP) was stacked on top of the pre-trained encoder (see Figure 2). Given the ability of this neural network based classifier to backpropagate the prediction error to the encoder layers, it was possible to fine-tune the higher-level layers of the convolutional feature extractor, which helped the model to learn to detect the most useful patterns for the *EGFR* mutation status assessment. Without this fine-tuning process, the general knowledge achieved by the pre-trained encoder would not be sufficient to extract such abstract and complex imaging manifestations. To perform the intended *EGFR* mutation status prediction, the dataset used in this task was based on the NSCLC-Radiogenomics [22] included examples.

C. EXPERIMENT DESIGN

Considering the experiments conducted, both the nodule and lung analysis consisted of a feature learning task for the CAE training, and a classification task, where by making advantage of Transfer Learning techniques, the *EGFR* mutation status was predicted on the correspondent ROI. Figure 3 shows the pipeline for the selection of the three considered ROIs used for the prediction: nodule, lung containing nodule and both lungs.

1) NODULE ANALYSIS

In this first experiment, the analysed ROI only contained the nodule region. An image with size 80 × 80 × 80 voxels was extracted for each considered example in this analysis. It was ensured that each nodule fit in the size chosen for the ROI. In an attempt to reduce the overfitting effect by increasing the number of training examples, middle slices from the axial, coronal and sagittal planes were sampled, alongside six more slices from the cube symmetry planes (Figure 4). During the training phase, this slice oversampling allowed not only to increase the training dataset size, but also to improve class balancing by sampling more slices for the examples of the minority class. Additionally, some data augmentation was also employed to decrease the chances of overfitting [26], consisting of horizontal and vertical flips, as well as random image rotations. Training, cross-validation and testing data combinations were made using a patient-level split.

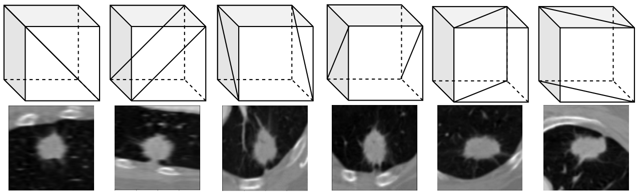

FIGURE 4. Slice extraction example using the cube symmetry planes for the nodule centered ROI.

The CAE proposed architecture consisted of four 3 × 3 convolutional layers, each one followed by a Rectified Linear Unit (ReLU) activation and a max-pooling layer to reduce the input by half. Passing through the encoder block, the resultant bottleneck is represented by a feature map (FM) with size $256 \times \frac{H}{8} \times \frac{W}{8}$, with H and W corresponding to the height and width dimensions of the original input tensor, respectively. To enable the image reconstruction mechanism, the decoder was implemented by mirroring the encoder part (see Figure 2). The CAE was trained to minimize the Mean Squared Error (MSE) value, which represented the averaged pixel-wise differences between the input and its reconstruction.

2) LUNG ANALYSIS

To investigate the correlation of other lung structures with the *EGFR* mutation status, two different analysis were conducted: using an axial section with both lungs and only the lung with main nodule (Figure 3).

For the feature learning task, it was necessary to create a dataset that contained lung segmented slices to study the importance of not eliminating non-lung structures. Thus, for each patient from the LIDC-IDRI data collection with a nodule classified as larger than 3 mm by at least one radiologist, one axial slice was extracted. This slice was selected based on the area of the nodule section, using the available nodule mask to find the slice with the largest nodule section. This selection criterion was used to keep the nodule as the main focus of the analysis. Thus, a 2D lung segmentation model based on U-Net architecture was implemented. The result was an axial lung slice generated from each patient scan, with only lung areas and background. Following this pre-processing phase, 875 lung axial slices from the LIDC-IDRI database were used to develop the new CAE. A similar encoder-decoder architecture was implemented, adding one extra strided convolutional layer for each part given the higher resolution of the input in this analysis.

For the experiments where only the lung with nodule was analysed, each considered slice was transformed as represented in Figure 5.

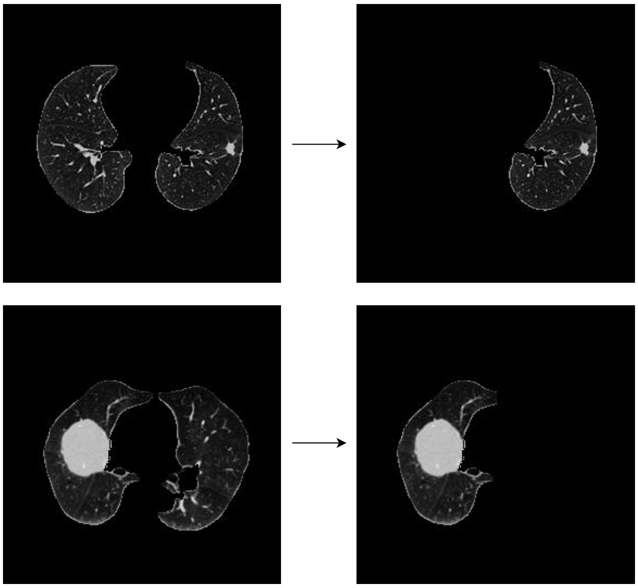

FIGURE 5. Examples of the selection of lung containing nodule based on the nodule position. Slices from NSCLC-Radiogenomics database patients [22].

A hyper-parameters search was adopted to find the best performance, with the range of values depicted in Table 2 regarding the CAE training, and Table 3 for the *EGFR* classification model, for the two experiments in the study: based on nodule and on the lung.

III. RESULTS AND DISCUSSION

The results for the approach optimization and the achieved classification performance are presented in this section, followed by the discussion and interpretation of the present results in comparison with the literature and the identified limitations of the study.

TABLE 2. Hyper-parameters range values considered in the search, regarding each CAE training.

Hyper-parameter	Nodule CAE	Lung CAE
Learning Rate	0.0001 - 0.001 - 0.01 - 0.1	
Optimizer	Adam - SGD[1]	
Momentum	0.1 - 0.5 - 0.9 - 0.99	
Batch-size	2 - 4 - 8 - 16	

[1] Stochastic Gradient Descent

TABLE 3. Hyper-parameters range values considered in the search for the *EGFR* classification model.

Hyper-parameter	Nodule ROI	Lung ROI
Learning Rate	0.0001 - 0.001 - 0.01	
Optimizer	Adam - SGD[1]	
Momentum	0.1 - 0.5 - 0.9	
Batch-size	8 - 16 - 32	8 - 16 -32 - 64
Dropout	0.25 - 0.5	
Hidden Layers	1 - 2	1 - 2 - 3
Hidden Neurons	32 - 64 - 128	32 - 64 -128 - 256
Weight Decay	0.0001 - 0.001 - 0.01	

[1] Stochastic Gradient Descent

A. HYPER-PARAMETERS SELECTION

The hyper-parameters selected for the CAE training on the nodule and lung reconstruction tasks that achieved best results are represented in Table 4. The best hyper-parameters for the end-to-end model classification for *EGFR* mutation status prediction are also presented in Table 5 for both nodule and lung ROI.

TABLE 4. Hyper-parameters values selected for each nodule and lung CAE training.

Hyper-parameter	Nodule CAE	Lung CAE
Learning Rate	0.01	
Optimizer	SGD[1]	
Momentum	0.9	
Batch-size	4	2

[1] Stochastic Gradient Descent

TABLE 5. Hyper-parameters values selected for *EGFR* mutation status prediction, using the nodule and lung ROI.

Hyper-parameter	Nodule ROI	Lung ROI
Learning Rate	0.001	
Optimizer	SGD[1]	
Momentum	0.9	
Batch-size	8	32
Dropout	0.25	0.5
Hidden Layers	1	
Hidden Neurons	64	
Weight Decay	0.0001	

[1] Stochastic Gradient Descent

B. EGFR CLASSIFICATION

Table 6 summarises the performance results obtained by those experiments. Mean and standard deviation values of AUC were computed for 20 random splits for training and testing sets.

TABLE 6. Classification results for lung axial slice *EGFR* mutation status prediction. Mean AUC values are depicted for each experiment: considering nodule, the lung containing nodule and both lungs in the CT slice.

Experiment	AUC[1]
Nodule	0.51 ± 0.06
Lung containing nodule	0.68 ± 0.08
Both lungs	0.60 ± 0.10

[1] AUC values are presented as mean and standard deviation.

This study addressed an analysis of three different ROIs for the lung tumor characterization by considering not only the nodule region but also the entire lung section in a 2D perspective using Transfer Learning techniques in CT images. Regarding the evaluation process, the AUC was the main performance metric used to assess the model ability to distinguish between the mutant and the wild classes. Giving the small size of the database selected, especially the small number of included *EGFR* mutant patients (23), the variance of the performance of the model was large, which can be confirmed on the high standard deviation values reported over the random data splits. This evaluation approach was followed to obtain a more realistic and overall perspective on the prediction ability of the methods, not dependent on some specific training data. For the same reason, giving the large cost of a misclassified mutant examples on the test set, decision metrics, such as Precision or Recall, were not considered useful for performance assessment. Considering the local nodule analysis, the characterization resulted in a poor classification model to assess the *EGFR* mutation status ($AUC = 0.51$), and showed that with the proposed approach is not possible to capture enough information to predict the *EGFR* mutation status. When extending the ROI to the entire lung axial section, the best classification performance ($AUC = 0.68$) was achieved when only including the lung that contained the nodule, showing a decrease in the prediction ability when both lungs were included in the analysed ROI ($AUC = 0.60$). All classification models were implemented with a feature extractor based on a trained Convolutional Autoencoder, reinforcing the relevance of the learned features when trying to reconstruct the input image. To the best of gathered knowledge, no other deep learning based work attempted to assess the *EGFR* mutation status using a lung holistic analysis. The results obtained in this work did not outperform the state-of-the-art but indicate a direction for future works dedicated to lung cancer characterization. Additionally, a direct comparison with related studies, in particular, the ones with higher ability to predict the *EGFR* mutation status would not bring a fair discussion point to this investigation, given the fact that the proposed approach addresses the lack of publicly available data to perform this task, which is not a factor that constrained the contribution of those studies. However, quantitative direct comparisons on prediction results are obviously crucial for a more clear understanding of the research evolution, increasing the need for more representative public databases to allow fair and useful comparisons. Performance comparisons between models trained and tested with different data do not allow clear and objective conclusions. Nevertheless, the current work represents a comparative study that contributes to the discussion about how complex and extensive are the biological changes associated with cancer development.

Considering the analysis proposed in the investigation of the three different ROIs, results confirm the hypothesis that a more extensive analysis of the lung structures combined with nodule information gives a more accurate prediction of the *EGFR* mutation status. The use of lung with tumor axial slice input provided better results in *EGFR* assessment, which emphasizes the idea behind the motivation of this study that complex transformations related to lung cancer might be present in other lung structures, and not only in the nodule region. The traditional approaches dedicated to lung cancer characterization were based only on the nodule features [27], assuming that all the physiological changes that can characterize the cancer development were based on the cluster of the tumor cells. Only few recent works showed the relevance of information from other parts of the lung to predict the *EGFR* mutation status associated with lung cancer [12], [13], [28], [29]. Our previous work [13] compared approaches using nodule radiomic and semantic features, and semantic information from other lung structures as well. Results pointed out the importance to use comprehensive approaches that take into consideration more elements to characterize these extremely complex physiological processes associated with lung cancer development. The current work still indicates that a comprehensive approach adds information that helps to characterize the *EGFR* mutations status. However, the experiments also showed a lower performance when analyzing both lungs compared to when the only lung with the tumor is considered, which indicates that the nodule and the near regions are the main contributions for the model decision. From the current results, it is considered of utmost importance to continue to investigate the relevance of comprehensive approaches, bringing a new perspective that might change the direction on this research topic, traditionally focused only on the nodule. However, to develop models capable of making a more comprehensive analysis with further potentially relevant information, more representative data of the population affected by lung cancer are needed to enable such abstract and complex transformations detection.

Although the presented study supported the idea that the general features learned while reconstructing images in the same domain can be transferred and be useful for several classification tasks, other strategies, which might present relevant benefits, were identified and should also be discussed: the use of a convolutional feature extractor trained for a related but simpler classification problem (e.g. lung abnormalities detection), in the same domain; the use of a discriminator-based feature extractor resultant from the development of Generative Adversarial Networks (GAN) models for the generation of synthetic samples (e.g. lung axial slices). In the proposed approach, the employed feature extractor was based on the development of a CAE for image reconstructions in the same domain of the final task, which is a Transfer Learning technique that already proved to be efficient in different biomedical imaging problems, especially when only small datasets are available [19], [30].

Having in mind other limitations that might have constrained the obtained results, the following deserve to be discussed and analysed: regarding the data used to develop the feature extractor in the nodule analysis, given the higher number of benign nodules, comparing with the malignant ones, it might have caused some lack on the ability to extract useful features related to malignant cases only, and not both; when conducting an investigation using a small size database, the impact of possible technical and annotation mistakes on the models is increased, and it should be considered in this analysis.

Considering the higher performance achieved in previous works with a mean AUC of 0.75 and 0.89 from [12], [13], respectively, where semantic features related to both the nodule and structures outside the nodule ROI were included, it is possible to suggest that, with the available datasets, working with qualitative assessed information might increase chances of better *EGFR* mutation status predictions. With deep learning based approaches, the larger set of deep features extracted from a larger ROI might not work well together with such a reduced dataset size, rising the idea that a hybrid approach where deep features combined with semantic *EGFR*-correlated information might gather the best of these two approaches.

Finally, an important limitation in this work that should also be noted relies on the number of studied genes. An investigation in imaging phenotypes of a more extensive list of lung cancer-related genes is necessary to obtain a more complete characterization, given the importance of other genes in targeted therapies development.

IV. CONCLUSION

This study proposed an approach based on a pre-trained encoder to work as a feature extractor, followed by an MLP for the final classification of the *EGFR* mutation status. Different regions of analysis were used in order to study the relevance of information from all the lung structures in this complex classification task. The results obtained showed that information from more extensive regions on the lung containing the nodule allow to capture information that might be relevant for lung cancer characterization, which emphasizes the importance of comprehensive approaches for an improved performance.

ACKNOWLEDGMENT

The authors acknowledge the National Cancer Institute and the Foundation for the National Institutes of Health for the free publicly available LIDC-IDRI Database used in this work. They also acknowledge The Cancer Imaging Archive (TCIA) for the open-access NSCLC-Radiogenomics dataset publicly available.

REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2017," *CA, Cancer J. Clinicians*, vol. 67, no. 1, pp. 7–30, Jan. 2017.

[2] M. Yuan, L.-L. Huang, J.-H. Chen, J. Wu, and Q. Xu, "The emerging treatment landscape of targeted therapy in non-small-cell lung cancer," *Signal Transduction Targeted Therapy*, vol. 4, no. 1, pp. 1–14, Dec. 2019.

[3] J. P. Sculier, T. Berghmans, and A. P. Meert, "Advances in target therapy in lung cancer," *Eur. Respiratory Rev.*, vol. 24, no. 135, pp. 23–29, 2015.

[4] S. E. D. C. Jorge, S. S. Kobayashi, and D. B. Costa, "Epidermal growth factor receptor (EGFR) mutations in lung cancer: Preclinical and clinical data," *Brazilian J. Med. Biol. Res.*, vol. 47, no. 11, pp. 929–939, Sep. 2014.

[5] X. Nan, C. Xie, X. Yu, and J. Liu, "EGFR TKI as first-line treatment for patients with advanced EGFR mutation-positive non-small-cell lung cancer," *Oncotarget*, vol. 8, no. 43, pp. 75712–75726, Sep. 2017.

[6] P. Tomasini, P. Walia, C. Labbe, K. Jao, and N. B. Leighl, "Targeting the KRAS pathway in non-small cell lung cancer," *Oncologist*, vol. 21, no. 12, p. 1450, 2016.

[7] S. Fang and Z. Wang, "EGFR mutations as a prognostic and predictive marker in non-small-cell lung cancer," *Drug Des., Develop. Therapy*, vol. 8, pp. 1595–1611, 2014.

[8] J. R. F. Junior, M. Koenigkam-Santos, F. E. G. Cipriano, A. T. Fabro, and P. M. D. Azevedo-Marques, "Radiomics-based features for pattern recognition of lung cancer histopathology and metastases," *Comput. Methods Programs Biomed.*, vol. 159, pp. 23–30, Jun. 2018.

[9] S. Chen, Y. Han, J. Lin, X. Zhao, and P. Kong, "Pulmonary nodule detection on chest radiographs using balanced convolutional neural network and classic candidate detection," *Artif. Intell. Med.*, vol. 107, Jul. 2020, Art. no. 101881.

[10] Y. Liu, J. Kim, Y. Balagurunathan, Q. Li, A. L. Garcia, O. Stringfield, Z. Ye, and R. J. Gillies, "Radiomic features are associated with EGFR mutation status in lung adenocarcinomas," *Clin. Lung Cancer*, vol. 17, no. 5, pp. 441.e6–448.e6, Sep. 2016.

[11] D. Hong, K. Xu, L. Zhang, X. Wan, and Y. Guo, "Radiomics signature as a predictive factor for EGFR mutations in advanced lung adenocarcinoma," *Frontiers Oncol.*, vol. 10, p. 28, Jan. 2020.

[12] O. Gevaert, S. Echegaray, A. Khuong, C. D. Hoang, J. B. Shrager, K. C. Jensen, G. J. Berry, H. H. Guo, C. Lau, S. K. Plevritis, D. L. Rubin, S. Napel, and A. N. Leung, "Predictive radiogenomics modeling of EGFR mutation status in lung cancer," *Sci. Rep.*, vol. 7, no. 1, pp. 1–8, Mar. 2017.

[13] 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," *Sci. Rep.*, vol. 10, no. 1, pp. 1–9, Dec. 2020.

[14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[15] S. Wang, J. Shi, Z. Ye, D. Dong, D. Yu, M. Zhou, Y. Liu, O. Gevaert, K. Wang, Y. Zhu, H. Zhou, Z. Liu, and J. Tian, "Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning," *Eur. Respiratory J.*, vol. 53, no. 3, Mar. 2019, Art. no. 1800986.

[16] W. Zhao, J. Yang, B. Ni, D. Bi, Y. Sun, M. Xu, X. Zhu, C. Li, L. Jin, P. Gao, P. Wang, Y. Hua, and M. Li, "Toward automatic prediction of EGFR mutation status in pulmonary adenocarcinoma with 3D deep learning," *Cancer Med.*, vol. 8, no. 7, pp. 3532–3543, Jul. 2019.

[17] X.-Y. Li, J.-F. Xiong, T.-Y. Jia, T.-L. Shen, R.-P. Hou, J. Zhao, and X.-L. Fu, "Detection of epithelial growth factor receptor (EGFR) mutations on CT images of patients with lung adenocarcinoma using radiomics and/or multi-level residual convolutionary neural networks," *J. Thoracic Disease*, vol. 10, no. 12, pp. 6624–6635, Dec. 2018.

[18] D. Yu, M. Zhou, F. Yang, D. Dong, O. Gevaert, Z. Liu, J. Shi, and J. Tian, "Convolutional neural networks for predicting molecular profiles of non-small cell lung cancer," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2017, pp. 569–572.

[19] F. Silva, T. Pereira, J. Frade, J. Mendes, C. Freitas, V. Hespanhol, J. L. Costa, A. Cunha, and H. P. Oliveira, "Pre-training autoencoder for lung nodule malignancy assessment using CT images," *Appl. Sci.*, vol. 10, no. 21, p. 7837, Nov. 2020.

[20] S. G. Armato, III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, and L. P. Clarke, "Data from LIDC-IDRI," *Cancer Imag. Arch.*, 2015.

[21] S. G. Armato *et al.*, "The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans," *Med. Phys.*, vol. 38, no. 2, pp. 915–931, Jan. 2011.

[22] 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, Oct. 2018, Art. no. 180202.

[23] E. J. Limkin, S. Reuzé, A. Carré, R. Sun, A. Schernberg, A. Alexis, E. Deutsch, C. Ferté, and C. Robert, "The complexity of tumor shape, spiculatedness, correlates with tumor radiomic shape features," *Sci. Rep.*, vol. 9, no. 1, pp. 1–12, Dec. 2019.

[24] M. A. Kassem, K. M. Hosny, and M. M. Fouad, "Skin lesions classification into eight classes for ISIC 2019 using deep convolutional neural network and transfer learning," *IEEE Access*, vol. 8, pp. 114822–114832, 2020.

[25] L. Alzubaidi, M. A. Fadhel, O. Al-Shamma, J. Zhang, J. Santamaría, Y. Duan, and S. R. Oleiwi, "Towards a better understanding of transfer learning for medical imaging: A case study," *Appl. Sci.*, vol. 10, no. 13, p. 4523, Jun. 2020.

[26] X. Ying, "An overview of overfitting and its solutions," *J. Phys., Conf. Ser.*, vol. 1168, Feb. 2019, Art. no. 022022.

[27] A. El-Baz, G. M. Beache, G. Gimel'farb, K. Suzuki, K. Okada, A. Elnakib, A. Soliman, and B. Abdollahi, "Computer-aided diagnosis systems for lung cancer: Challenges and methodologies," *Int. J. Biomed. Imag.*, vol. 2013, pp. 1–46, Jan. 2013.

[28] S. Rizzo, S. Raimondi, E. E. C. 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," *Eur. J. Radiol.*, vol. 110, pp. 148–155, Jan. 2019.

[29] J. Xiong, X. Li, L. Lu, L. H. Schwartz, X. Fu, J. Zhao, and B. Zhao, "Implementation strategy of a CNN model affects the performance of CT assessment of EGFR mutation status in lung cancer patients," *IEEE Access*, vol. 7, pp. 64583–64591, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8713576/

[30] K. Oh, Y.-C. Chung, K. W. Kim, W.-S. Kim, and I.-S. Oh, "Author correction: Classification and visualization of Alzheimer's disease using volumetric convolutional neural network and transfer learning," *Sci. Rep.*, vol. 10, no. 1, pp. 1–16, Dec. 2020.

FRANCISCO SILVA received the master's degree in electrical and computers engineering from the University of Porto, Portugal, in 2020. He is currently working as a Research Collaborator with the Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), Portugal. His research interests include biomedical image processing and deep learning.

TANIA PEREIRA received the master's and Ph.D. degrees in biomedical engineering from the University of Coimbra, Portugal, in 2009 and 2014, respectively. She completed her postdoctoral training at the University of Lleida, Spain, in 2017, and the University of California at San Francisco, USA, in 2019. She is currently an Assistant Researcher with the Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), Portugal, where she conducts research in the fields of image and signal processing and the applications of machine learning in biomedical research.

JOANA MORGADO received the M.Sc. degree in biomedical engineering from the Faculty of Sciences and Technology, University of Lisbon, Portugal, in 2020. She is currently pursuing the Ph.D. degree in computer science with the Faculty of Science, University of Porto.

Since 2020, she has been working as a Researcher with INESC TEC, a research and development institute affiliated to the University of Porto, in the Visual Computing and Machine Intelligence Group (VCMI). Her main research interests include computer vision, image processing, medical imaging, machine learning, and artificial intelligence.

JULIETA FRADE received the M.Sc. degree in informatics and computing engineering from the Faculty of Engineering, University of Porto, Portugal, in 2020. Her research interests include computer vision and deep learning.

JOSÉ MENDES received the M.Sc. degree in informatics and computing engineering from the Faculty of Engineering, University of Porto, Portugal, in 2020. His research interests include machine learning and deep learning.

CLÁUDIA FREITAS received the M.D. degree from the Instituto de Ciências Biomédicas de Abel Salazar, University of Porto (ICBAS-UP), in 2015. She has been a Pulmonology Resident with the Pulmonology Department, Centro Hospitalar e Universitário de São João, since 2017.

EDUARDO NEGRÃO received the M.D. degree from the Faculty of Medicine, University of Coimbra (FMUC), in 2014. He has been a Radiology Resident with the Radiology Department, Centro Hospitalar e Universitário de São João, since 2017.

BEATRIZ FLOR DE LIMA received the M.D. degree from the Faculty of Medicine, University of Coimbra (FMUC), in 2015. She has been a Radiology Resident with the Radiology Department, Centro Hospitalar e Universitário de São João, since 2017.

MIGUEL CORREIA DA SILVA received the M.D. degree from the Faculty of Medicine, University of Coimbra (FMUC), in 2016. He has been a Radiology Resident with the Radiology Department, Centro Hospitalar e Universitário de São João, since 2018.

ANTÓNIO J. MADUREIRA graduated from the Medical School, University of Porto (Portugal), in 1993. He has been a Staff Member with the Radiology Department, Centro Hospitalar e Universitário de São João, Porto, Portugal, since February 2001, and an Assistant of radiology, since April 2002. He is currently the Chairperson of the Radiology Department, Centro Hospitalar e Universitário de São João. He has published over 50 articles and written three book chapters. He has presented over 100 articles on national and international courses and congresses and has won eight prizes.

ISABEL RAMOS received the Ph.D. degree from Porto University, in 1989. Since 1989, she has been working with the Centro Hospitalar e Universitário de São João, as the Chairman of Radiology Department. From 1992 to 1996, she was the Vice President of the Portuguese Society of Radiology, where she was the President, from 1996 to 2000. From 1995 to 2010, she was a Portuguese Representative at UEMS–Radiology. Since 2017, she has been on the Board of Superior Council, Ordem dos Médicos, wrote more than 150 articles in Portuguese and international journals, nine chapters in radiologic books and has given more than 300 lectures. She received the fellowship from Yale University, USA.

VENCESLAU HESPANHOL received the M.D. degree in pulmonology and the Ph.D. degree in medicine from the Faculdade de Medicina, Universidade do Porto (FMUP), in 1981 and 1999, respectively. Since 2019, he has been the Director of pulmonology service with the Centro Hospitalar e Universitário de São João. He has a specialization domain in medicine-pulmonology, lung cancer (interventional pulmonology and thoracic oncology), and epidemiology and biostatistics. He performed research in clinical trials in oncology, and belongs to the Research Group-IPATIMUP–Genetic Dynamic Cancer Cells Group.

JOSÉ LUÍS COSTA received the Ph.D. degree in biology from Uppsala University, Sweden. Prior to joining the Institute of Molecular Pathology and Immunology, University of Porto (IPATIMUP), in 2006, he carried out Postdoctoral Research at Vrije University Medical Center, Amsterdam, The Netherlands. In 2014, he was appointed as an Affiliated Professor with the Faculty of Medicine, University of Porto. He is currently the Medical Affairs Director with Thermo Fisher Scientific, where he aims to promote and support the uptake of precision medicine for the benefit of patients. Scientifically, his career has focused on understanding the crucial events that are the basis for the development and progression of cancer.

ANTÓNIO CUNHA received the degree in 1993 and the Ph.D. degree in 2005, with a focus on computer vision related to control of automated guided vehicles. He did his master's thesis in 1998. He is currently an Assistant Professor with the University of Trás-os-Montes and Alto Douro (UTAD). He has been a member of the Centre for Biomedical Engineering Research (C-BER), INESC TEC, since 2015. He works in electrical engineering, electronics and computers, with a particular focus on computer vision, pattern recognition, biomedical image processing, and computer-aid diagnosis.

HÉLDER P. OLIVEIRA (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Porto, Portugal, in 2013. He is currently a Senior Researcher with the Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), Portugal, and the Manager of the Visual Computing and Machine Intelligence Area, where he conducts research in the fields of image and signal processing and the applications of machine learning in biomedical research.