CONVOLUTIONAL NEURAL NETWORKS FOR PREDICTING MOLECULAR PROFILES OF NON-SMALL CELL LUNG CANCER

Dongdong Yu[1], Mu Zhou[2], Feng Yang[3], Di Dong[1], Olivier Gevaert[2], Zaiyi Liu[4], Jingyun Shi[5,], and Jie Tian[1,*]*

[1]The Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences
[2]The Stanford Center for Biomedical Informatics Research, Stanford University
[3]School of Computer and Information Technology, Beijing Jiaotong University
[4]Department of Radiology, Guangdong Academy of Medical Sciences, Guangdong General Hospital
[5]Department of Radiology, Shanghai Pulmonary Hospital, Tongji University School of Medicine

ABSTRACT

Quantitative imaging biomarkers identification has become a powerful tool for predictive diagnosis given increasingly available clinical imaging data. In parallel, molecular profiles have been well documented in non-small cell lung cancers (NSCLCs). However, there has been limited studies on leveraging the two major sources for improving lung cancer computer-aided diagnosis. In this paper, we investigate the problem of predicting molecular profiles with CT imaging arrays in NSCLC. In particular, we formulate a discriminative convolutional neural network to learn deep features for predicting epidermal growth factor receptor (EGFR) mutation states that are associated with cancer cell growth. We evaluated our approach on two independent datasets including a discovery set with 595 patients (Datset1) and a validation set with 89 patients (Dataset2). Extensive experimental results demonstrated that the learned CNN-based features are effective in predicting EGFR mutation states (AUC=0.828, ACC=76.16%) on Dataset1, and it further demonstrated generalized predictive performance (AUC=0.668, ACC=67.55%) on Dataset2.

Index Terms— Non-Small Cell Lung Carcinoma, Convolutional neural networks, Computed tomography, Computed-aided diagnosis

1. INTRODUCTION

Non-small cell lung cancer (NSCLC) is a lethal disease accounting for about 85% of all lung cancers with a dismal 5-year survival rate of 15.9% [1]. Molecular profiles of NSCLC, like epidermal growth factor receptor (EGFR), have been well documented over the past decade to suggest targeted treatment [2]; Computed tomography (CT), on the other hand, has been a major imaging modality for early cancer detection in NSCLC [3]. A promising yet challenging task is to infer the diagnostic value from CT images, such as the identification of discriminative image features that are able to predict molecular signatures. Since image traits would create a unique avenue to non-invasively assess molecular events [4], it will offer an opportunity to discern early indicators of targeted treatment. A majority of image-based studies have proposed to estimate nodule malignancy likelihood [5-6]. However, the association between image signatures and molecular profiling information, particularly, predicting gene-mutation types (e.g., EGFR mutation) from computational image features, has not been explicitly addressed.

In this paper, we focus on developing computational CT image features for predicting EGFR presence (EGFR+) and absence (EGFR-). In particular, we introduce a data-driven framework utilizing a convolutional neural network (CNN) for the EGFR mutation prediction. By the use of the deep learning paradigm, we sought to leverage the predictive power of image signatures in reflecting EGFR mutation states. The proposed approach consists of a six-layer CNN for learning deep image features and a Support Vector Machine (SVM) classifier for prediction. Our goal is towards developing an end-to-end framework that automatically learns the gene mutation-sensitive information from CT imaging, thus the presented work is largely opposed to conventional image evaluation heavily relying on radiologists' inputs. Such as in [7], image characteristics including air bronchogram and small lesion size were found to associate with EGFR mutation in 280 patients. Additionally, standardized uptake value (SUV) has been reported to surrogate EGFR mutation with 100 patients [8]. However, the lack of external validation with cross-sectional imaging impedes the translational value of the detected image biomarkers. We here present evaluation of the learned CNN features for EGFR mutation states on two independent sets.

More specifically, our contribution is three-fold: i) We formulate a discriminative CNN framework that leads to improved prediction performance of EGFR mutation states in NSCLC; ii) Because external validation is a crucial step in finding predictive biomarkers, we report results of the CNN-based features on two independent datasets, including

978-1-5090-1172-8/17/$31.00 ©2017 IEEE

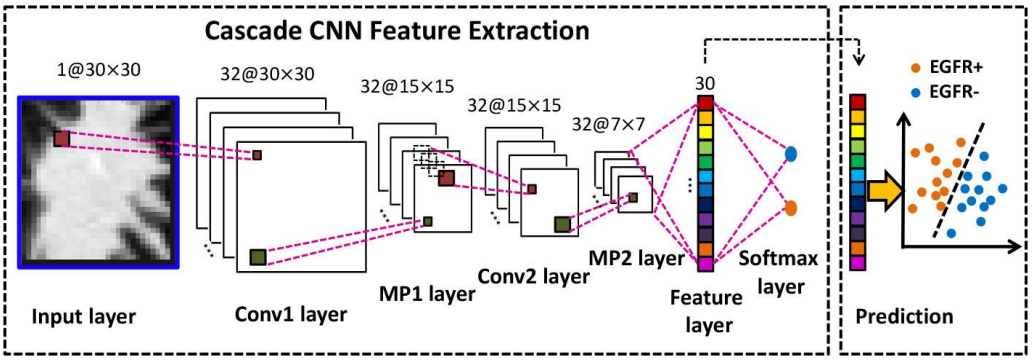

Fig. 1. Illustration of the proposed Convolutional Neural Network in conjunction with the Support Vector Machine. The input 2-D nodule patches centering around nodule shapes were fed into the concatenated Conv1+MP1 layers with each convolutional layer contains 32 convolutional kernels. The outputs from the feature layer are the learned deep features, which are applied to the SVM classifier for predicting EGFR mutation states. In notation, 32@30×30 indicates 32 feature maps with size 30 × 30.

Dataset1 (discovery set, 595 patients) and Dataset2 (validation set, 89 patients), revealing outperformed performance as opposed to traditional CT texture, gabor filter features, and statistical features; iii) The proposed data-driven CNN framework emphasizes computation on nodule image patches centered around most relevant nodule information, which is highly appealing when dealing with growing volumes of imaging data. Thus, it holds promise to accelerate the process by removing nodule boundary delineation and hand-crafted feature engineering.

2. LEARNING FEATURES FOR EGFR MUTATION PREDICTION

We explore deep features by incorporating nodule CT patches into a convolutional neural network (CNN). An overview of our proposed approach is shown in Figure 1. Building upon a hierarchical structure, our convolutional neural network contains two convolutional layers, both of which are followed by a max-pooling layer, and a fully connected layer which represents the final output feature, and a softmax layer. By taking use of the alternated layers of convolution and max-pooling, extracted feature dimensions continue to decrease along the network hierarchy, resulting in highly compact deep features as outcomes in the feature layer. Next, the SVM classifier is used for the binary molecular class prediction (i.e., EGFR+, EGFR-).

The proposed CNN architecture starts from a convolution layer (i.e., Conv1 layer), in which the input 2-D nodule patch is convolved by the convolution kernels to generate new feature maps as shown in Fig.1. A convolution operation is formulated as: $y = R(\sum_i x_i \otimes k_i + b)$, where y represents the new feature map. The x_i and k_i respectively indicate the *ith* slice of the input feature maps and that of the convolution kernels. b is the bias scalar of the convolution kernels. The $\otimes$ denotes the convolution operation. $R(u) = \max(0, u)$ is the rectified linear unit function.

Following the convolutional layer, a max-pooling layer (i.e., MP1 layer) is used to achieve the feature reduction by subsampling the feature maps. Given the pooling window size be s × s, the max-pooling operation is defined as: $f(i,j) = \max_{0 \le l,m \le s} \{y_{i*s+l,j*s+m}\}$, where y is the output of the convolution layer and f is the new feature map. The (i,j) is the position of the feature node on the new feature map.

After two cascade convolutional layers and max-pooling layers, the final feature layer is followed by a softmax layer as $p_j = \frac{\exp(y_j)}{\exp(y_0) + \exp(y_1)}, j = 0,1.$ where $y_i = Wv + b$ is the linear combination of the obtained deep feature v. W is the weight matrix and b is the bias term.

The loss function is the softmax loss that calculates the loss between the prediction and the molecular labels: $Loss = -\sum_i (q_i \log p_{0,i} + (1 - q_i) \log p_{1,i})$, where i is the ith nodule patch, $p_{0,i}$, $p_{1,i}$ is the prediction probability calcualted by the softmax layer p_j, and q_i is the clinical label of EGFR mutation states. The CNN is learned by minimizing the loss function using the stochastic gradient descent.

The feature layer is used to obtain the deep features from the input nodule patch. Next, the learned deep features are used to apply the SVM classifier for the EGFR mutation prediction.

3. EXPERIMENTS AND DISCUSSION

We first introduce two independent CT datasets of lung nodules with EGFR mutation profiles. We then present experimental results on our CNN features (i.e., termed as Deep30) in predicting EGFR mutation states on the two datasets separately, competing with hand-crafted features including Statistical features, Texture features, and Gabor filter features (termed as the STG feature set).

Dataset1: it consists of 595 patients including 316 EGFR- and 279 EGFR+ patients. Both non-enhanced and

Table 1. Prediction performance on Dataset1 (mean ± std) and Dataset2, respectively. AUC, ACC, SEN, and SPE are the area under the ROC curve, accuracy, sensitivity and specificity, respectively.

	Cross validation on Dataset1				Testing on Dataset2			
	AUC	ACC(%)	SEN(%)	SPE(%)	AUC	ACC(%)	SEN(%)	SPE(%)
STG30	0.726 ± 0.006	68.55 ± 0.80	68.49 ± 1.22	68.61 ± 1.06	0.494	58.55	25.93	72.76
STG60	0.734 ± 0.007	70.27 ± 0.79	69.07 ± 1.16	71.33 ± 1.07	0.655	63.66	55.56	67.19
STG100	0.721 ± 0.007	69.22 ± 0.87	68.50 ± 1.36	69.84 ± 1.14	0.636	55.85	76.65	46.80
STG150	0.722 ± 0.007	68.71 ± 0.85	67.79 ± 1.38	69.50 ± 1.24	0.609	50.51	**77.78**	38.64
Deep30	$\mathbf{0.828 \pm 0.005}$	$\mathbf{76.16 \pm 0.84}$	$\mathbf{73.80 \pm 1.15}$	$\mathbf{78.24 \pm 1.08}$	**0.668**	**67.55**	48.59	**75.81**

contrast-enhanced chest CT images are acquired on Philips Brilliance 40 and Siemens Defintion AS.

Dataset2: it consists of 89 patients with 59 EGFR- and 30 EGFR+. Both non-enhanced and contrast-enhanced chest CT data are acquired by the two multi-detector row CT systems (GE Lightspeed Ultra 8, GE Healthcare). All the nodule images from both datasets are resampled and set the resolution to a fixed 0.8 mm/pixel along all three axes. When performing comparison with hand-crafted features, the Toboggan Based Growing Automatic (TBGA) segmentation [9] is used to first identify nodule boundaries.

4.1. Implementation Details

For Dataset1, we evaluate the proposed CNN model with a 5-fold cross-validation. For each patient's CT scan, a nodule patch is fed into the CNN model. The nodule patch is defined by cropping from the CT image based on the nodule centers marked by 2 radiologists. The size of the nodule patch is 30×30 pixels. The size of convolutional kernel in our CNN model is $3 \times 3 \times 32$, and the kernel size of max-pooling is 3×3 with the stride sets as 2. The number of feature layer and output layer are 30 and 2 respectively. In Dataset1, during each round of cross-validation, there are originally 475 nodules (222 EGFR+ nodules and 253 EGFR- nodules) in the training set and 119 nodules (56 EGFR+ nodules and 63 EGFR- nodules) in the testing set. To enlarge the training samples to train the CNN, we augment both EGFR+ nodules and EGFR- nodules by translating the nodule patches along two axes with ± 2 and ± 1 pixels. Thus, each patch is translated 8 times. Such setting helps capture a range of translation invariant features. Once the trained CNN network is built, the Deep30 feature set can be extracted from the well-trained CNN model.

In regards to the competing approaches, we extract totally 592 dimensional features (i.e., STG feature set) for each segmented nodule image including: 7-dimensional statistical features capturing characteristics of nodule shape and volumes; 45-dimensional texture features including grey-level frequency, Gray-Level Co-occurence Matrix (GLCM), Gray-Level Size Zone Matrix (GLSZM), Gray-Level Run-Length Matrix (GLRLM), and Neighborhood Gray-Tone Difference Matrix (NGTDM); and 540-dimensional Gabor wavelet features with 4 scales and 8 directions. Given the extracted 592-dimensional features, we use the MRMR feature selector to achieve the top30 (STG30), top60 (STG60), top100 (STG100), and top150 (STG150) feature sets.

Next, the Deep30 and the defined 4 competing feature sets are trained with the SVM classifier with the 5-fold cross-validation on Dataset1. We report average accuracy (ACC), sensitivity (SEN), specificity (SPE) and the average area under the ROC curve (AUC) values of experiment results on Dataset1 from 200 times 5-fold cross-validation. In addition, we extend to report additional results on by directly using the well-trained model from Dataset1 and predict all data samples from Dataset2, where the ACC, SEN, SPE and AUC values are presented.

4.2. EGFR Mutation Prediction

Table 1 summarized the results of our approach (Deep30) and the four compared models (STG30, STG60, STG100, and STG150) on the two datasets. The proposed Deep30 achieved promising performance on both datasets, all outperforming conventional hand-crafted features in a variety of feature parameters. The accuracy of our CNN feature based prediction outperforms the traditional hand-crafted feature based classification by at least 5.89% and 3.89% on Dataset1 and Dataset2, respectively. The superiority of the Deep30 suggests that the CNN structure is able to preserve class-specific information, while reducing noisy, irrelevant features in a form of feature reduction.

When comparing our Deep30 and the STG30 where both feature sets are in the same feature dimension, the proposed Deep30 performed significantly higher results, revealing the power of feature reduction of our CNN model. It also confirmed that conventional STG sets were parameter sensitive, likely leading to inferior outcome especially when feature numbers are limited. Even increasing dimensions of STG feature sets did not reveal improved performance. The fact may be ascribed to that conventional feature extraction contains highly-correlated features with different feature extraction parameters, which simply do not contribute to the discriminative performance. We also reported ROC curves in Fig. 2 to observe the prediction outcomes.

In regards to results on Dataset2, as introduced in the experimental details, it is a complete, external validation. The new data samples from Dataset 2 are fed into the CNN model which has been well trained on Dataset1. Despite the fact that different imaging parameters involved in the two Datasets, which can lead to varying performance, the results

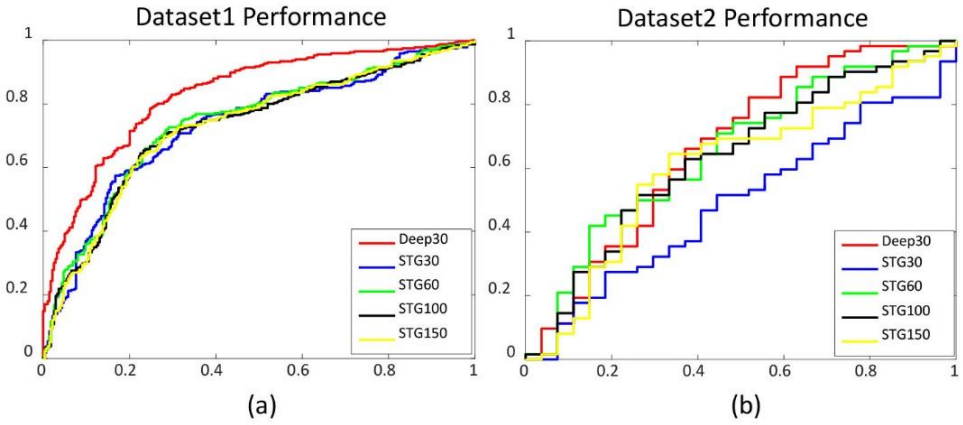

Fig.2. ROC curves of EGFR mutation prediction using different types of feature sets based on Dataset1 (a) and Dataset2 (b).

from both Table 1 and Fig. 2(b) revealed that the proposed approach has the potential to generalize well on the external datasets. Notably, it may not be surprising that the overall results dropped which could be ascribed to the testing on limited samples from Dataset2. Since our approach is in favor of data-driven scheme as already detailed in Dataset1 with over 500 samples, with growing number of testing samples, we would expect to further boost prediction performance on EGFR mutation states.

4. CONCLUSION

In this paper, we investigate the association between CT imaging and molecular profiles for patients suffering from NSCLC. The proposed data-driven CNN framework presented encouraging results in predicting EGFR mutation states. Extensive experiments on two independent clinic datasets repeatedly revealed positive outcomes of our approach, outperforming conventional hand-crafted features. In the future, we will collect growing number of samples and continue to leverage the structure of fine-tuning CNNs that would allow improved performance on prediction.

5. ACKKNOWLEDEMENT

This paper is supported by the National Natural Science Foundation of China under Grant No. 81501616, 81227901, 61231004, Science and Technology Service Network Initiative Program of Chinese Academy of Science under Grant NO. KFJ-SW-STS-160, the Strategic Priority Research Program from Chinese Academy of Sciences under Grant NO. XDB02060010, the Instrument Developing Project of Chinese Academy of Sciences under Grant No. YZ201502. O.G. is supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under Award Number R01EB020527.

6. REFERENCES

[1] S. Kligerman and C. White, "Epidemiology of lung cancer in women: risk factors, survival, and screening.," *AJR. Am. J. Roentgenol.*, vol. 196, no. 2, pp. 287–95, 2011.

[2] R. Rosell, E. Carcereny, et.al, "Erlotinib versus standard chemotherapy as first-line treatment for European patients with advanced EGFR mutation-positive non-small-cell lung cancer (EURTAC): A multicentre, open-label, randomised phase 3 trial," *Lancet Oncol.*, vol. 13, no. 3, pp. 239–246, Mar. 2012.

[3] C. I. Henschke, D. F. Yankelevitz, et al, "CT screening for lung cancer: Frequency and significance of part-solid and nonsolid nodules," *Am. J. Roentgenol.*, vol. 178, no. 5, pp. 1053–1057, 2002.

[4] O. Gevaert, J. Xu, et.al, "Non-small cell lung cancer: identifying prognostic imaging biomarkers by leveraging public gene expression microarray data--methods and preliminary results.," *Radiology*, vol. 264, no. 2, pp. 387–96, Aug. 2012.

[5] W. Shen, M. Zhou, et al, "Multi-scale convolutional neural networks for lung nodule classification," in *International Conference on Information Processing in Medical Imaging*, 2015, vol. 9123, pp. 588–599.

[6] A. El-Baz, M. Nitzken, et al, "3D shape analysis for early diagnosis of malignant lung nodules.," *Med. Image Comput. Comput. Assist. Interv.*, vol. 14, no. Pt 3, pp. 175–82, 2011.

[7] S. Rizzo, F. Petrella, et al, "CT Radiogenomic Characterization of EGFR, K-RAS, and ALK Mutations in Non-Small Cell Lung Cancer," *Eur. Radiol.*, vol. 26, no. 1, pp. 32–42, 2016.

[8] I. Il Na, B. H. Byun, et al, "18F-FDG uptake and EGFR mutations in patients with non-small cell lung cancer: A single-institution retrospective analysis," *Lung Cancer*, vol. 67, no. 1, pp. 76–80, 2010.

[9] J. Song, C. Yang, et al, "Lung lesion extraction using a toboggan based growing automatic segmentation approach," *IEEE Trans. Med. Imaging*, vol. 35, no. 1, pp. 337–353, Jan. 2016.