ERJ Express. Published on January 11, 2019 as doi: 10.1183/13993003.00986-2018

EUROPEAN RESPIRATORY *journal*

FLAGSHIP SCIENTIFIC JOURNAL OF ERS

Early View

Original article

Predicting EGFR Mutation Status in Lung Adenocarcinoma on CT Image Using Deep Learning

Shuo Wang, Jingyun Shi, Zhaoxiang Ye, Di Dong, Dongdong Yu, Mu Zhou, Ying Liu, Olivier Gevaert, Kun Wang, Yongbei Zhu, Hongyu Zhou, Zhenyu Liu, Jie Tian

Please cite this article as: Wang S, Shi J, Ye Z, *et al*. Predicting EGFR Mutation Status in Lung Adenocarcinoma on CT Image Using Deep Learning. *Eur Respir J* 2019; in press (https://doi.org/10.1183/13993003.00986-2018).

Predicting EGFR Mutation Status in Lung Adenocarcinoma on CT Image Using Deep Learning

Authors: Shuo Wang[1,5], Jingyun Shi[3], Zhaoxiang Ye[4], Di Dong[1,5], Dongdong Yu[1,5], Mu Zhou[2], Ying Liu[4], Olivier Gevaert[2], Kun Wang[1], Yongbei Zhu[1], Hongyu Zhou[6], Zhenyu Liu[1], Jie Tian[1,5,7]

Affiliations:

1. CAS Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing, China.
2. The Stanford Center for Biomedical Informatics Research, Department of Medicine, Stanford University, California, USA.
3. Department of Respiratory Medicine, Shanghai Pulmonary Hospital, Tongji University School of Medicine, Shanghai, China.
4. Department of Radiology, Tianjin Medical University Cancer Institute and Hospital, Tianjin, China.
5. University of Chinese Academy of Sciences, Beijing, China.
6. Paul C. Lauterbur Research Center for Biomedical Imaging, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China.
7. Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, Beihang University, Beijing, 100191

S. Wang, J. Shi, Z. Ye, D. Dong, D. Yu, and M. Zhou contribute equally to this work.

Corresponding author:

Jie Tian, PhD

Fellow of ISMRM, IAMBE, AIMBE, IEEE, SPIE, OSA, IAPR

CAS Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China;

Phone: 86-010-82618465; Fax: 86-010-82618465; E-mail: jie.tian@ia.ac.cn

Abstract

Epidermal Growth Factor Receptor (EGFR) genotyping is critical for treatment guideline such as the use of tyrosine kinase inhibitors in lung adenocarcinoma (LA). Conventional identification of EGFR genotype requires biopsy and sequence testing that is invasive and may suffer from the difficulty in accessing tissue samples. Here, we proposed a deep learning (DL) model to predict the EGFR mutation status in LA by non-invasive computed tomography (CT).

We retrospectively collected 844 LA patients with preoperative CT image, EGFR mutation and clinical information from two hospitals. An end-to-end DL model was proposed to predict the EGFR mutation status by CT scanning.

By training in 14926 CT images, the DL model achieved encouraging predictive performance in both the primary cohort (n = 603; AUC = 0.85, 95% CI 0.83-0.88) and the independent validation cohort (n = 241; AUC = 0.81, 95% CI 0.79-0.83), which showed significant improvement than previous studies using hand-crafted CT features or clinical characteristics ($p < 0.001$). The deep learning score demonstrated significant difference in EGFR-mutant and EGFR-wild type tumours ($p < 0.001$).

Since CT is routinely used in lung cancer diagnosis, the DL model provides a non-invasive and easy-to-use method for EGFR mutation status prediction.

Keywords: Lung adenocarcinoma, epidermal growth factor receptor mutation, lung cancer, deep learning, artificial intelligence, health informatics

Take-home message: Deep learning provides a non-invasive method for EGFR mutation prediction (AUC=0.81) in lung adenocarcinoma, which shows significant improvement than using hand-crafted CT features or clinical characteristics.

Introduction

Lung adenocarcinoma is a common histological type of lung cancer and the discovery of epidermal growth factor receptor (EGFR) mutations has revolutionized treatment of lung adenocarcinoma [1, 2]. In the first-line treatment, detecting EGFR mutation is critical since EGFR tyrosine kinase inhibitors can target specific mutations within the EGFR gene, and have resulted in improved outcomes in EGFR-mutant lung adenocarcinoma patients [3, 4]. Mutational sequencing of biopsies has become gold standard for EGFR mutation detection. However, biopsy testing for measuring EGFR status likely suffers from locating tissue regions because of the extensive heterogeneity of lung tumour [5, 6]. In addition, biopsy testing raises a potential risk of cancer metastasis [7]. Furthermore, repeated tumour sampling, difficulty in accessing tissue samples, poor DNA quality [8] and the relative high costs can limit the applicability of mutational sequencing [9]. In these situations, a non-invasive and easy-to-use method for predicting EGFR mutation status is necessary.

Computed tomography (CT) as a routinely used technique in cancer diagnosis provides a non-invasive way to analyze lung cancer [10-12]. Recent studies revealed that features extracted from lung cancer CT images were related to gene-expression patterns [13-16] and showed predictive power on EGFR profiles [17-19]. Although image assessment cannot replace biopsies, image-driven studies can provide additional information that is complementary to biopsies [5, 9]. For example, CT imaging provides a complete scope of tumour and its microenvironment, enabling us to predict EGFR mutation status by considering intra-tumour heterogeneity. In addition, predicting EGFR-mutation status by CT imaging helps us to choose the highly suspicious tumour for biopsy if multiple tumours present in a patient.

Furthermore, CT imaging is non-invasive and easy to acquire throughout the course of treatment.

Early findings demonstrated that CT semantic features and quantitative 'radiomic' features showed predictive value to EGFR mutation status [9]. However, these methods can only reflect generalized image features which lack specificity to EGFR mutation. In addition, the radiomics methods based on feature engineering rely on precise tumour boundary annotation which requires human-labeling efforts. Since radiomic features are computed only inside tumour area, the microenvironment and tumour-attached tissues are ignored. By contrast, advanced artificial intelligence models can overcome these problems through a self-learning strategy such as deep learning methods [20-22]. Benefiting from the strong feature-learning ability, deep learning models have shown human expert-level performance in skin cancer classification [23], eye diseases diagnosis [24], and non-invasive liver fibrosis prediction [25]. Moreover, deep learning models also gained promising performance in assisting lung cancer analysis [26-29]. Compared with feature engineering-based radiomic methods, deep learning-based radiomics do not require precise tumour boundary annotation and learn features automatically from image data [30]. Furthermore, deep learning-based radiomics can extract features that are adaptive to specific clinical outcomes, while feature engineering-based radiomics can only describe general features that may lack specificity for outcome prediction.

In this study, we proposed a deep learning (DL) model to mine CT image information that is related to EGFR mutation status. Our method is an end-to-end pipeline that requires only the manually selected tumour region in CT image without precise tumour boundary segmentation or human-defined features, which is different with conventional radiomic methods based on feature engineering. The proposed

model can learn EGFR mutation-related features from CT images automatically and predicts the probability of the tumour being EGFR-mutant. Furthermore, the DL model can discover the suspicious tumour sub-regions that are strongly related to EGFR mutation status, aiming to rapidly facilitate clinicians' treatment decision for patients. To evaluate the performance of the DL model, we collected a large dataset from two independent hospitals (844 patients) and provided independent validation results of the proposed DL model.

Material and methods

Patients

The institutional review board of Tianjin Medical University and Shanghai Pulmonary Hospital approved this retrospective study and waived the need to obtain informed consent from the patients. Patients who meet the following inclusion criteria were collected into this study: (i) histologically confirmed primary lung adenocarcinoma; (ii) pathologic examination of tumour specimens been carried out with proven records of EGFR mutation status; (iii) preoperative contrast-enhanced CT data obtained. Patients were excluded if: (i) clinical data including age, gender, and stage was missing; (ii) preoperative treatment was received; (iii) the duration between CT examination and subsequent surgery exceeded one month. Finally, 844 patients from two hospitals were used for this study. We allocated the patients into a primary cohort and an independent validation cohort according to the hospital. The primary cohort included 603 patients from Shanghai Pulmonary Hospital between January 2013 to July 2014. The validation cohort included 241 patients from Tianjin Medical University between January 2013 to February 2014. The primary and validation

cohorts were used for developing and validating the DL model respectively. CT scanning parameters and detailed description about the datasets were available in supplementary methods.

In regards to molecular profiles, tumour specimens were obtained using surgical resection. EGFR mutations were identified on four tyrosine kinase domains (exons 18-21), which are frequently mutated in lung cancer. The mutation status was determined using an amplification refractory mutation system with human EGFR gene mutations detection kit (Beijing ACCB Biotech Ltd). If any exon mutation was detected, the tumour was identified as EGFR-mutant; otherwise, the tumour was identified as EGFR-wild type. In this study, we therefore focused on predicting these binary outcomes (EGFR-mutant and EGFR-wild type) for patients with lung adenocarcinoma.

Development of the deep learning model

Deep learning is a hierarchical neural network that aims at learning the abstract mapping between raw data to the desired label. The computational units in the DL model are defined as layers and they are integrated to simulate the analysis process of human brain. The main computational formulas are convolution, pooling, activation and batch normalization. Finally, supplementary methods define the terms of the computational process in building the DL model.

Figure 1 illustrated the pipeline of the EGFR mutation status prediction. For applying the DL model, a cubic ROI (region of interest) containing the entire tumour was manually selected (by J. Shi and Y. Liu) according to the following rule: the ROI should include the full tumour region including the edges of tumours. This rule is easy to use in practice since we do not require the tumour to be precisely in the center of

ROI (supplementary figure S1 illustrates several ROIs selected by users). Afterwards, the ROI was resized to 64×64 pixels by third-order spline interpolation in each CT slice, and fed into the DL model. Through a sequential activation of convolution and pooling layers, the DL model gave an EGFR-mutant probability for the image. To make a robust prediction, all the CT slices of the tumour were fed into the DL model, and the average probability is treated as the EGFR-mutant probability for the tumour. Specifically, all the adjacent three CT slices were combined as a three-channel image and were fed into the DL model for prediction (details in supplementary figure S2).

During model training, we used transfer learning to train the first 20 convolutional layers (sub-network 1 in figure 1) by 1.28 million natural images from the ImageNet dataset [31]. This transfer learning technique has shown good performance in disease diagnosis since it enlarged the training data [23, 32]. Afterwards, the last four convolutional layers (sub-network 2 in figure 1) were trained by 14926 CT images from lung adenocarcinoma tumours in the primary cohort. Details about the model building was presented in supplementary methods.

Given the CT image of tumour, the DL model predicts a probability of the tumour being EGFR-mutant directly without any pre- or post-processing or image segmentation. The DL model generated by using the primary cohort of this study is available at http://radiomics.net.cn/post/110. Part of the CT images from the validation cohort can also be downloaded as examples for testing the DL model.

Visualization of the deep learning model

Due to the end-to-end manner of deep learning, the inference process of the DL model is not intuitive for users. To further understand the prediction process of the DL model, we used visualization techniques to analyze features learned by the DL

model. The most important component of the DL model is convolutional layer. Therefore, we visualized convolutional layers from two perspectives to understand the inference process of the DL model: 1) visualizing the feature patterns extracted by convolutional layer; 2) visualizing the response of each convolutional layer to different tumours.

A convolutional layer consists of multiple convolutional filters where each convolutional filter extracts different features. Through a filter visualizing algorithm [33, 34], we can visualize the feature pattern extracted by a convolutional filter, and we define this feature pattern as a deep learning feature (details in supplementary methods).

To further explore the meaning of the deep learning features, we observed the response of each convolutional filter to different tumours. Given a tumour image, each convolutional filter in the DL model generates a response map indicating the corresponding feature patterns in the tumour. The average value of the response map is defined as response value. A good convolutional filter should have different response values between EGFR-mutant and EGFR-wild type tumours. Therefore, visualizing the response values for a convolutional filter in different tumour groups can help us evaluating the performance of the convolutional filter.

Statistical analysis

Statistical analysis was performed using SPSS Statistics 21. The independent samples *t* test was adopted to assess the significance of the mean value on ages between the patients in EGFR-mutant and EGFR-wild type groups. The same statistical analysis was performed to assess the difference of deep learning score between the EGFR-mutant and EGFR-wild type groups. The chi-squared test was

used to evaluate the difference of categorical variables such as gender and tumour stage in all the cohorts. In addition, we used DeLong test to evaluate the difference of the receiver operating characteristic (ROC) curves between various models. P-value<0.05 was treated as significant. Our implementation of the DL model used the Keras toolkit and Python 2.7.

Results

Clinical characteristics of patients

The clinical characteristics of patients were presented in table 1. There was no significant difference between the primary and validation cohorts in terms of age and gender (p=0.083 for age, p=0.321 for gender). The tumour stage showed statistical difference between the two cohorts probably because of the regional differences, since patients in the two cohorts are from two different cities in China. To eliminate this difference, we performed a stratified analysis in the two cohorts to validate the robustness of the DL model. Clinical characteristics such as age, gender and stage illustrated difference between EGFR-mutant and EGFR-wild type patients, therefore, these characteristics were used to build a clinical model for comparison to the DL model.

Diagnostic validation of the DL model

Table 2 listed the predictive performance of the DL model where we used area under the ROC curve (AUC), accuracy, sensitivity and specificity as main measurements. All the quantitative results were performed in tumour level which is also in subject level since each patient includes only one tumour. In the primary cohort, the DL model showed good predictive performance by 5-fold cross validation

(AUC = 0.85, 95% CI 0.83-0.88). This performance was further confirmed in the independent validation cohort (AUC=0.81, 95% CI 0.79-0.83). The close AUC between the primary and validation cohorts indicated that the DL model generalized well on predicting EGFR mutation status of unseen new patients. Benefiting from transfer learning with 1.28 million natural images, the DL model did not suffer from over-fitting. Meanwhile, we illustrated the ROC curves of the DL model in the two cohorts in figure 2a. Moreover, the deep learning score revealed a significant difference between EGFR-mutant and EGFR-wild type groups in the two cohorts ($p<0.001$ in both the primary and validation cohorts, figure 2b).

In addition, we performed a stratified analysis to validate the diagnostic performance of the DL model concerning tumour stage. Supplementary table S1 and supplementary figure S3 indicated that the DL model achieved good results in all the tumour stages. Moreover, the deep learning score showed significant difference between EGFR-mutant and EGFR-wild type groups regardless of tumour stages.

Figure 2c plotted the decision curve of the DL model. This curve showed that if the threshold probability of a patient or doctor is bigger than 10%, using the DL model to predict EGFR mutation status in lung adenocarcinoma adds more benefit than either the treat-all-patients scheme or the treat-none scheme [35]. This highlighted the clinical use of the DL model.

Comparison between the DL model and other methods

In early studies, clinical characteristics, semantic features [17] and quantitative 'radiomic' features [9] were used for EGFR mutation status prediction. Therefore, we built a clinical model, a semantic model and a radiomics model as comparison to the proposed DL model. The clinical model involved gender, stage and age as features,

and used support vector machine (SVM) with radius-basis kernel for EGFR mutation prediction. The semantic model used 16 semantic features reported in the previous study and a multivariate logistic regression (details in supplementary methods and supplementary table S4) [17]. The radiomics model extracted 1108 features by the pyradiomics toolkit [36] and selected 8 features using recursive feature elimination. Finally, a random forest containing 100 trees were built for EGFR mutation prediction in the radiomics model.

The quantitative performance in table 2 and the ROC curves in figure 2a indicated that the DL model gained better performance than the clinical model with significant difference (AUC = 0.66, 95% CI 0.62-0.70 in the primary cohort, $p < 0.0001$; AUC = 0.61, 95% CI 0.58-0.64 in the validation cohort, $p < 0.0001$). A significant improvement over semantic model was also observed in the two cohorts (AUC = 0.76, 95% CI 0.72-0.80 in the primary cohort, $p < 0.0001$; AUC = 0.64, 95% CI 0.61-0.67 in the validation cohort, $p < 0.0001$). Similar improvement over radiomics model was also confirmed in the two cohorts (AUC = 0.70, 95% CI 0.66-0.74 in the primary cohort, $p < 0.0001$; AUC = 0.64, 95% CI 0.61-0.67 in the validation cohort, $p = 0.0002$).

Suspicious tumour area discovery

Since deep learning is an end-to-end prediction model that learns abstract mappings between tumour image and EGFR mutation status directly, it is important to explain the predicting process such that we can estimate how reliable the prediction is. We used a deep learning visualization method [33, 34] to find the tumour region that was most related to EGFR mutation status (details in supplementary methods). This important region was defined as suspicious area in our study. When the DL model

predicts an EGFR mutation status, it tells clinicians which area draws attention of this model at the same time.

Figure 3 depicted the suspicious areas found by the DL model. For a lung adenocarcinoma tumour, the DL model generated an attention map indicating the importance of each part in the tumour; we used 0.5 as the cut-off value to reserve the high-response area (suspicious tumour area). These areas were more important than other regions of tumour since they drew the attention of the DL model. As shown in the bottom row in figure 3, the suspicious areas found by the DL model varied in different tumours. For example, the suspicious area in figure 3a was the tissue between tumour and pleura, whereas the suspicious area in figure 3b was the tumour edge. Based on these observations, the DL model interpreted these two tumours as EGFR-mutant. On the other hand, the deep learning model focused on the cavitary area in figure 3c and predicted it to be EGFR-wild type. Since the DL model required only raw CT image of tumour as input without any tumour segmentation, some normal tissues can be fed into the model. However, the model was capable of finding suspicious area inside tumour instead of being disturbed by normal tissues. Figure 3d illustrated a tumour adjacent to mediastinum. In this case, the ROI for the DL model included some normal tissues outside the tumour. However, the DL model found a suspicious area inside the tumour instead of the normal tissues. The suspicious tumour area was inferred to be strongly related to EGFR mutation status by the DL model. Therefore, it can potentially provide a biopsy position for clinicians to avoid false negative diagnosis caused by the intra-tumour hetrogeneity. The difference between the suspicious tumour area and other tumour areas may be further explained by combining PET-CT data.

Deep learning feature analysis

The advantage of deep learning mainly comes from its automatic feature-learning ability. By learning from 14926 tumour images, the DL model detects features that are strongly associated with EGFR mutation status.

For a better understanding about the deep learning feature, we visualized several convolutional filters in the DL model (figure 4a). The shallow convolutional layer learned low-level simple features such as horizontal and diagonal edges (*Conv_2*). A deeper convolutional layer learned more complex features such as tumour shape. For instance, the filters in layer *Conv_13* had strong response on circle or arch shapes because most tumours contain circle or arch structure. When going deeper, the features became more abstract and were gradually related to EGFR mutation status (*Conv_20*, *Conv_24*). In supplementary figure S4, we compared the convolutional filters before training and after transfer learning (trained in CT data). This figure indicated that the convolutional filters learned various feature patterns that are different with their initial status. Furthermore, transfer learning makes the filters more specific to CT data especially when the network layer going deeper.

To further demonstrate the association between the deep learning features with EGFR mutation status, we extracted two convolutional filters from the last convolutional layer (the positive and negative filters). These two filters captured different texture patterns (the first column in figure 4b) responding to EGFR-mutant and EGFR-wild type tumors respectively. When we fed EGFR-wild type tumours to the DL model, the negative filter generated strong response while the positive filter was nearly shut down. Similarly, when we fed EGFR-mutant tumours to the DL model, the negative filter was depressed but the positive filter was strongly activated.

As depicted in figure 4c, the response of the positive/negative filters on EGFR-mutant and EGFR-wild type tumours had significant difference in all the cohorts ($p<0.001$). In figure 4d, we illustrated the clustering map of deep learning features from the last convolutional layer (*Conv_24*) in the whole dataset (844 patients). The deep learning features showed obvious clusters that had different response to EGFR-mutant and EGFR-wild type patients. Meanwhile, tumours of different EGFR mutation status (EGFR-mutant/wild type) can be roughly separated (vertical axis in figure 4d).

To compare the importance of the deep learning features and the radiomic features, we combined the 32 deep learning features from the *Conv_24* layer with the 1108 radiomic features, and used RFE to select the important features. In this step, the RFE used linear SVM and 5 fold cross-validation to determine the optimal feature amount using the primary cohort, which is consistent with the RFE settings in building the radiomics model. Finally, 11 features were selected including 8 deep learning features and 3 radiomic features. This indicated that the deep learning features showed stronger association with EGFR mutation status in comparison to radiomic features. In addition, we calculated the univariate AUC for all the deep learning features and the radiomic features. As illustrated in supplementary figure S5, many of the deep learning features have higher AUC than radiomic features.

Discussion

In this study, we proposed a DL model using non-invasive CT image to predict EGFR mutation status for patients with lung adenocarcinoma. We trained the DL model in 14926 CT images from the primary cohort (603 patients), and validated its performance in an independent validation cohort from another hospital (241 patients). The DL model showed encouraging results in the primary cohort (AUC = 0.85, 95% CI 0.83-0.88) and achieved strong performance in the independent

validation cohort (AUC = 0.81, 95% CI 0.79-0.83). The DL model revealed that there was a significant association between high-dimensional CT image features and EGFR genotype. Our analysis provides an alternative method to non-invasively assess EGFR information for patients, and offers a great supplement to biopsy. Meanwhile, our model can discover the suspicious tumour area that dominates the prediction of EGFR mutation status. This analysis offered visual interpretation to clinicians about understanding the prediction outcomes in CT data. Moreover, the DL model requires only the raw tumour image as input and predicts the EGFR mutation status directly without further human assist, which is easy-to-use and very fast.

Previous studies used clinical factors [8] and radiomics based on feature engineering [9, 17, 18] to predict EGFR mutation status. For example, clinical factors such as age, gender, tumour stage and predominant subtype were used to build a nomogram for EGFR mutation status prediction [8]. In this study, the clinical factors achieved AUC=0.64 in a validation cohort including 464 Asian patients. Clinical model is interpretable since clinical factors are widely used and the nomogram represents an intuitive linear model. However, clinical features such as stage and predominant subtype require invasive biopsy. In addition, clinical features only reflect few tumour information in pathological level. By contrast, radiomic methods used CT image to quantify tumour information in macroscopic level, and built the relationship between tumour image and EGFR mutation status. Compared with clinical factors, radiomic analysis provides quantitative features to mine high-dimensional information that are associated with EGFR genotype. In a cohort including 353 patients, the radiomic method achieved AUC=0.69 by using hand-crafted CT image features [9]. Despite the advantage of radiomic method, the hand-crafted feature requires time-consuming tumour boundary segmentation and may lack specificity to EGFR

genotype. Consequently, we proposed a deep learning method to learn EGFR-related tumour features automatically and avoid complex tumour boundary segmentation. Furthermore, the deep learning method only requires a user-defined ROI of tumour instead of four complex procedures in radiomics based on feature engineering (e.g. tumour boundary segmentation, feature extraction, feature selection and model building).

Advantage of deep learning

Previous studies suggested that CT-based semantic features [18, 19] and quantitative radiomic features [9, 17] reflected EGFR mutation status. However, they can only reflect low-order visual features or simple high-order features. There are abstract features that can probably be associated with EGFR mutation status; however, they are difficult to be represented by hand-crafted feature engineering. In these situations, deep learning demonstrates its advantage since it can mine abstract features that are difficult to be formulized but are important for identifying EGFR mutation status.

Compared with previously reported hand-crafted features, the DL model has the following advantages: 1) through a hierarchical neural network structure, the DL model extracts multi-level features from visual characteristics to abstract mappings that are directly related to EGFR information. 2) The DL model does not require time-consuming tumour boundary annotation, which is a big advantage over hand-crafted feature engineering. Moreover, the microenvironment of tumour and the relationship between tumour and attached tissues (pleura traction, etc.) are also considered in the deep learning model. 3) The DL model is fast and easy to use, which requires only the raw CT image as input and predicts the EGFR mutation status directly without further human assist.

Clinical utility of the DL model

The DL model provides potential clinical utility from the following perspectives: 1) The proposed DL model provides a non-invasive method to predict EGFR mutation status, which can be easily used in routine CT diagnosis. 2) If the biopsy result of a tumour shows EGFR-wild type, the result may include false-negatives because of the intra-tumour heterogeneity. At this time, the DL model can be seen as an alternative validation tool. If the DL model predicts the tumour to be EGFR-mutant, clinicians may need to re-biopsy tissues [37]. 3) The DL model only requires routinely-used CT image without adding additional cost. Therefore, this model can be used multiple times throughout the course of treatment [9]. 4) Most importantly, although we studied only adenocarcinoma, the DL model also shows predictive value in other histological types. This enables the DL model to be used directly in CT scans of lung cancer without identifying histological types. To validate this hypothesis, we additionally collected 125 patients with other lung cancer histological types from Shanghai Pulmonary Hospital between January 2013 to July 2014 (clinical characteristics in supplementary table S2). Quantitative results in supplementary table S3 indicates that the DL model can achieve AUC=0.77 (95% CI 0.73-0.81) in other histological types of lung cancer. Consequently, even without knowing histological type of a lung cancer, the DL model can achieve AUC=0.81 in adenocarcinoma and AUC=0.77 in other histological types.

Despite the encouraging performance of the DL model, this study has several limitations. First, we only examined patients in Asian population. However, EGFR mutation rate can be affected by race. In the future work, population from multiple sources is necessary to test whether the DL model can be generalized to other populations. Second, although the DL model shows better performance than clinical,

semantic and radiomics models, the combination of these models is unclear. The predictive performance may be improved if we combine these models together. Third, our study only focused on EGFR mutation status. The relationship between EGFR mutation and other genetic mutations (e.g. ROS-1, ALK) can be explored in the future work.

Acknowledgments

This work was supported by the National Key R&D Program of China (2017YFA0205200, 2017YFC1308700, 2017YFC1308701, 2017YFC1309100, 2016YFC010380), National Natural Science Foundation of China (81227901, 81771924, 81501616, 61231004, 81671851, and 81527805), the Beijing Municipal Science and Technology Commission (Z171100000117023, Z161100002616022), the Instrument Developing Project of the Chinese Academy of Sciences (YZ201502), and the Youth Innovation Promotion Association CAS. O.G. was supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under Award Number R01EB020527.

Author contributions: D. Dong, J. Shi, Y. Liu, Z. Ye collected the clinical dataset. Z. Liu, K. Wang, Y. Zhu processed and analyzed the data. H. Zhou provided statistical analysis. S. Wang, D. Yu and M. Zhou built the deep learning model and wrote the paper. O. Gevaert and J. Tian conceived the project and edited the paper.

Competing interests: The authors declare that they have no competing interests.

References

1. Sequist LV, Yang JC, Yamamoto N, Obyrne K, Hirsh V, Mok T, Geater SL, Orlov S, Tsai CM, Boyer M. Phase III study of afatinib or cisplatin plus pemetrexed in patients with metastatic lung adenocarcinoma with EGFR mutations. *J Clin Oncol* 2013: 31(27): 3327-3334.

2. Maemondo M, Inoue A, Kobayashi K, Sugawara S, Oizumi S, Isobe H, Gemma A, Harada M, Yoshizawa H, Kinoshita I. Gefitinib or chemotherapy for non–small-cell lung cancer with mutated EGFR. *N Engl J Med* 2010: 362(25): 2380-2388.

3. Li T, Kung H-J, Mack PC, Gandara DR. Genotyping and genomic profiling of non–small-cell lung cancer: implications for current and future therapies. *J Clin Oncol* 2013: 31(8): 1039.

4. Zhou C, Wu Y-L, Chen G, Feng J, Liu X-Q, Wang C, Zhang S, Wang J, Zhou S, Ren S. Erlotinib versus chemotherapy as first-line treatment for patients with advanced EGFR mutation-positive non-small-cell lung cancer (OPTIMAL, CTONG-0802): a multicentre, open-label, randomised, phase 3 study. *Lancet Oncol* 2011: 12(8): 735-742.

5. Itakura H, Achrol AS, Mitchell LA, Loya JJ, Liu T, Westbroek EM, Feroze AH, Rodriguez S, Echegaray S, Azad TD, Yeom KW, Napel S, Rubin DL, Chang SD, Harsh GR, Gevaert O. Magnetic resonance image features identify glioblastoma phenotypic subtypes with distinct molecular pathway activities. *Sci Transl Med* 2015: 7(303): 303ra138-303ra138.

6. Sacher AG, Dahlberg SE, Heng J, Mach S, Jänne PA, Oxnard GR. Association between younger age and targetable genomic alterations and prognosis in non–small-cell lung cancer. *JAMA Oncol* 2016: 2(3): 313-320.

7. Loughran C, Keeling C. Seeding of tumour cells following breast biopsy: a literature review. *Br J Radiol* 2011: 84(1006): 869-874.

8. Girard N, Sima CS, Jackman DM, Sequist LV, Chen H, Yang JC-H, Ji H, Waltman B, Rosell R, Taron M, Zakowski MF, Ladanyi M, Riely G, Pao W. Nomogram to predict the presence of EGFR activating mutation in lung adenocarcinoma. *Eur Respir J* 2012: 39(2): 366-372.

9. Velazquez ER, Parmar C, Liu Y, Coroller TP, Cruz G, Stringfield O, Ye Z, Makrigiorgos M, Fennessy F, Mak RH. Somatic mutations drive distinct imaging phenotypes in lung cancer. *Cancer Res* 2017: 77(14): 3922-3930.

10. Lambin P, Leijenaar RT, Deist TM, Peerlings J, de Jong EE, van Timmeren J, Sanduleanu S, Larue RT, Even AJ, Jochems A. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol* 2017: 14(12): 749.

11. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology* 2015: 278(2): 563-577.

12. Kauczor H-U, Heussel CP, von Stackelberg O. Time to take CT screening to the next level? *Eur Respir J* 2017: 49(4).

13. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, Bussink J, Monshouwer R, Haibe-Kains B, Rietveld D. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun* 2014: 5: 4006.

14. Karlo CA, Di Paolo PL, Chaim J, Hakimi AA, Ostrovnaya I, Russo P, Hricak H, Motzer R, Hsieh JJ, Akin O. Radiogenomics of clear cell renal cell carcinoma: associations between CT imaging features and mutations. *Radiology* 2014: 270(2): 464-471.

15. Gevaert O, Xu J, Hoang CD, Leung AN, Xu Y, Quon A, Rubin DL, Napel S, Plevritis SK. Non–small cell lung cancer: identifying prognostic imaging biomarkers by leveraging public gene expression microarray data—methods and preliminary results. *Radiology* 2012: 264(2): 387-396.

16. Zhou M, Leung A, Echegaray S, Gentles A, Shrager JB, Jensen KC, Berry GJ, Plevritis SK, Rubin DL, Napel S. Non–small cell lung cancer radiogenomics map identifies relationships between molecular and imaging phenotypes with prognostic implications. *Radiology* 2017: 286(1): 307-315.

17. Liu Y, Kim J, Qu F, Liu S, Wang H, Balagurunathan Y, Ye Z, Gillies RJ. CT features associated with epidermal growth factor receptor mutation status in patients with lung adenocarcinoma. *Radiology* 2016: 280(1): 271-280.

18. Yano M, Sasaki H, Kobayashi Y, Yukiue H, Haneda H, Suzuki E, Endo K, Kawano O, Hara M, Fujii Y. Epidermal growth factor receptor gene mutation and computed tomographic findings in peripheral pulmonary adenocarcinoma. *J Thorac Oncol* 2006: 1(5): 413-416.

19. Zhou J, Zheng J, Yu Z, Xiao W, Zhao J, Sun K, Wang B, Chen X, Jiang L, Ding W. Comparative analysis of clinicoradiologic characteristics of lung adenocarcinomas with ALK rearrangements or EGFR mutations. *Eur Radiol* 2015: 25(5): 1257-1266.

20. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015: 521(7553): 436.

21. Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, Hubert T, Baker L, Lai M, Bolton A. Mastering the game of Go without human knowledge. *Nature* 2017: 550(7676): 354.

22. Hazlett HC, Gu H, Munsell BC, Kim SH, Styner M, Wolff JJ, Elison JT, Swanson MR, Zhu H, Botteron KN. Early brain development in infants at high risk for autism spectrum disorder. *Nature* 2017: 542(7641): 348.

23. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017: 542(7639): 115.

24. Ting DSW, Cheung CY-L, Lim G, Tan GSW, Quang ND, Gan A, Hamzah H, Garcia-Franco R, San Yeo IY, Lee SY. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA* 2017: 318(22): 2211-2223.

25. Wang K, Lu X, Zhou H, Gao Y, Zheng J, Tong M, Wu C, Liu C, Huang L, Meng F. Deep learning Radiomics of shear wave elastography significantly improved diagnostic performance for assessing liver fibrosis in chronic hepatitis B: a prospective multicentre study. *Gut* 2018: gutjnl-2018-316204.

26. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* 2017: 284(2): 574-582.

27. Wang S, Zhou M, Liu Z, Liu Z, Gu D, Zang Y, Dong D, Gevaert O, Tian J. Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation. *Med Image Anal* 2017: 40: 172-183.

28. Shen W, Zhou M, Yang F, Yu D, Dong D, Yang C, Zang Y, Tian J. Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. *Pattern Recognit* 2017: 61: 663-673.

29. Wang S, Liu Z, Chen X, Zhu Y, Zhou H, Tang Z, Wei W, Dong D, Wang M, Tian J. Unsupervised Deep Learning Features for Lung Cancer Overall Survival Analysis. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2018: IEEE; 2018. p. 2583-2586.

30. Wang S, Liu Z, Rong Y, Zhou B, Bai Y, Wei W, Wang M, Guo Y, Tian J. Deep learning provides a new computed tomography-based prognostic biomarker for recurrence prediction in high-grade serous ovarian cancer. *Radiother Oncol* 2018.

31. Huang G, Liu Z, Weinberger KQ, van der Maaten L. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017; 2017. p. 3.

32. Kermany DS, Goldbaum M, Cai W, Valentim CC, Liang H, Baxter SL, McKeown A, Yang G, Wu X, Yan F. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* 2018: 172(5): 1122-1131. e1129.

33. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: 2017 IEEE International Conference on Computer Vision (ICCV); 2017; 2017. p. 618-626.

34. Kotikalapudi Rac. keras-vis. GitHub, https://github.com/raghakot/keras-vis, 2017.

35. Huang Y-q, Liang C-h, He L, Tian J, Liang C-s, Chen X, Ma Z-l, Liu Z-y. Development and validation of a radiomics nomogram for preoperative prediction of lymph node metastasis in colorectal cancer. *J Clin Oncol* 2016: 34(18): 2157-2164.

36. van Griethuysen JJ, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, Beets-Tan RG, Fillion-Robin J-C, Pieper S, Aerts HJ. Computational radiomics system to decode the radiographic phenotype. *Cancer Res* 2017: 77(21): e104-e107.

37. Liu Y, Kim J, Balagurunathan Y, Li Q, Garcia AL, Stringfield O, Ye Z, Gillies RJ. Radiomic features are associated with EGFR mutation status in lung adenocarcinomas. *Clin Lung Cancer* 2016: 17(5): 441-448.e446.

Tables

Table 1. Clinical characteristics of patients in the primary and validation cohorts.

Characteristics	Primary cohort (n=603)		*P*	Validation cohort (n=241)		*P*
	EGFR-wild type	EGFR-mutant		EGFR-wild type	EGFR-mutant	
Age, mean (SD), years	59.50 (9.72)	61.36 (8.96)	0.016	59.59 (8.83)	59.21 (7.28)	0.716
Gender, No. (%)			<0.001			<0.001
Female	99 (39.76)	206 (58.19)		52 (42.62)	79 (66.39)	
Male	150 (60.24)	148 (41.81)		70 (57.38)	40 (33.61)	
Stage, No. (%)			0.047			0.017
I	181 (72.69)	240 (67.80)		50 (40.98)	65 (54.62)	
II	27 (10.84)	27 (7.63)		22 (18.03)	8 (6.72)	
III	36 (14.46)	69 (19.49)		43 (35.25)	35 (29.41)	
IV	5 (2.01)	18 (5.08)		7 (5.74)	11 (9.24)	
EGFR mutation, No. (%)	249 (41.29)	354 (58.71)	--	122 (50.62)	119 (49.38)	--

Table 2. Predictive performance of various methods in the primary and validation cohorts.

Methods	Cohorts	AUC	Accuracy (%)	Sensitivity (%)	Specificity (%)
Clinical model	Primary	0.66 (0.62-0.70)	61.60 (57.90-65.15)	64.39 (59.75-68.90)	56.75 (50.65- 62.68)
	Validation	0.61 (0.58-0.64)	61.83 (58.88-64.88)	56.30 (52.41-60.41)	67.21 (63.20-71.20)
Semantic model	Primary	0.76 (0.72-0.80)	64.77 (61.31-68.22)	71.49 (67.86-75.09)	61.22 (57.45-65.12)
	Validation	0.64 (0.61-0.67)	62.24 (59.94-64.72)	63.03 (59.61-66.60)	61.48 (58.22-64.92)
Radiomics model	Primary	0.70 (0.66-0.74)	66.27 (62.96-69.83)	**85.05** (81.81-88.46)	40.98 (35.82-46.34)
	Validation	0.64 (0.61-0.67)	61.47 (58.69-64.69)	64.04 (60.34-68.34)	58.97 (55.10-63.10)
DL model	Primary	**0.85** (0.83-0.88)	**77.02** (74.02-79.97)	76.83 (73.17-80.49)	**79.03** (74.26-83.61)
	Validation	**0.81** (0.79-0.83)	**73.86** (71.82-75.82)	**72.27** (69.27-75.27)	**75.41** (72.32-78.32)

AUC is area under the receiver operating characteristic curve.

Data in parentheses are the 95% confidence interval.

All the results in the primary cohort are evaluated by 5-fold cross validation.

The best performance is indicated in bold font.

Figure legends

Figure 1. Illustration of the deep learning model.

This model is composed of convolutional layers with kernel size 3×3 and 1×1, batch normalization and pooling layers. Sub-network 1 shares the same structure with the first 20 layers in DenseNet [31], which was pre-trained using 1.28 million natural images. Sub-network 2 was trained in the EGFR mutation dataset, aiming at capturing the association between image features to EGFR mutation labels. When we feed a tumour into the deep learning model, it predicts the probability of the tumour being EGFR-mutant.

Figure 2. Predictive performance of the deep learning model.

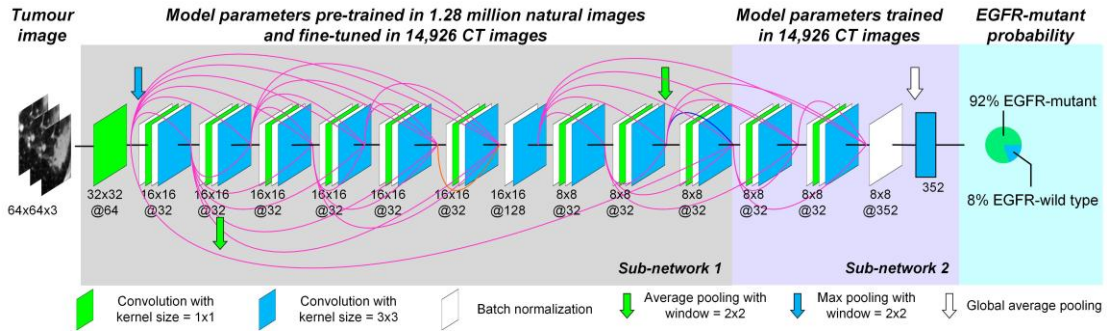
a). ROC curves of the deep learning (DL) model, radiomics model, semantic model and clinical model in the primary/validation cohorts. **b).** Deep learning score between EGFR-mutant and EGFR-wild type groups in the primary and validation cohorts. The horizontal dotted lines are the quartiles. **c)**. Decision curve of the DL model. The green line represents the benefit of treating all the patients as EGFR-wild type, and the blue line represents the benefit of treating all the patients as EGFR-mutant. The red line shows the benefit of using the DL model.

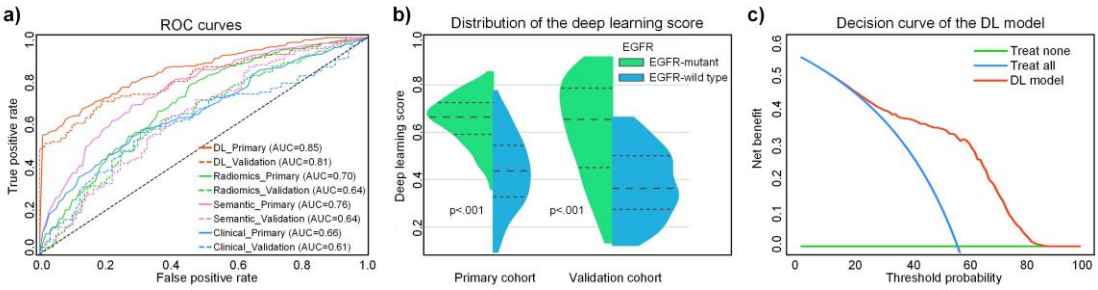
Figure 3. Suspicious tumour area discovery.

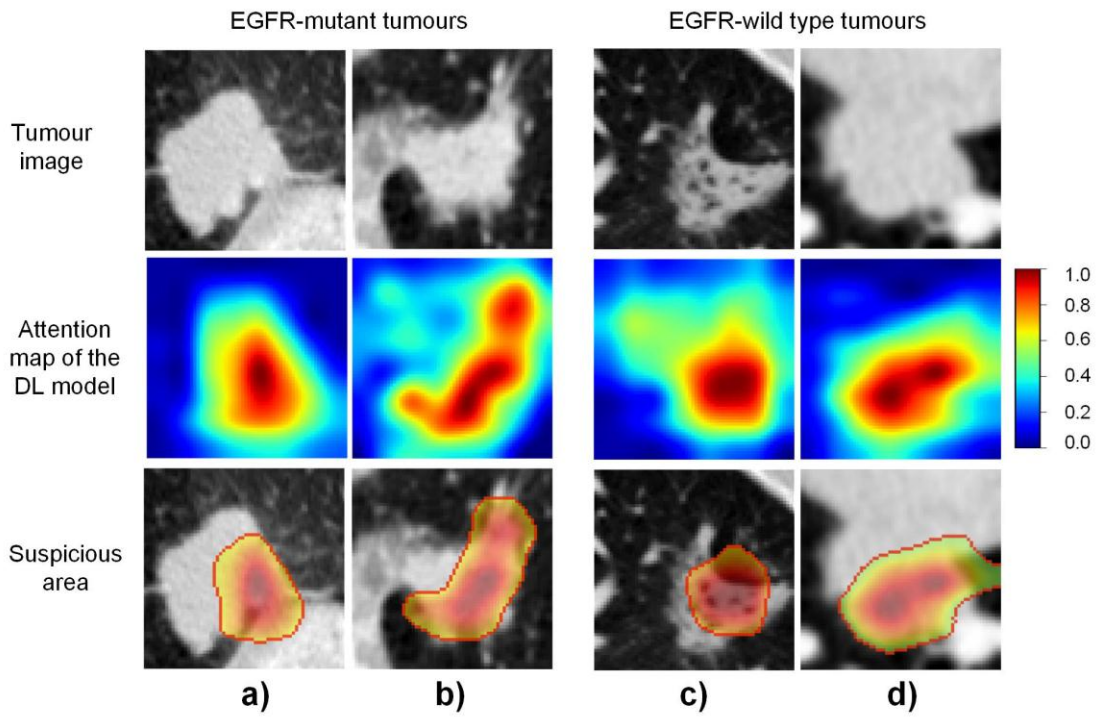
We used 0.5 as cut-off value to acquire the suspicious areas according to the attention map of the DL model.

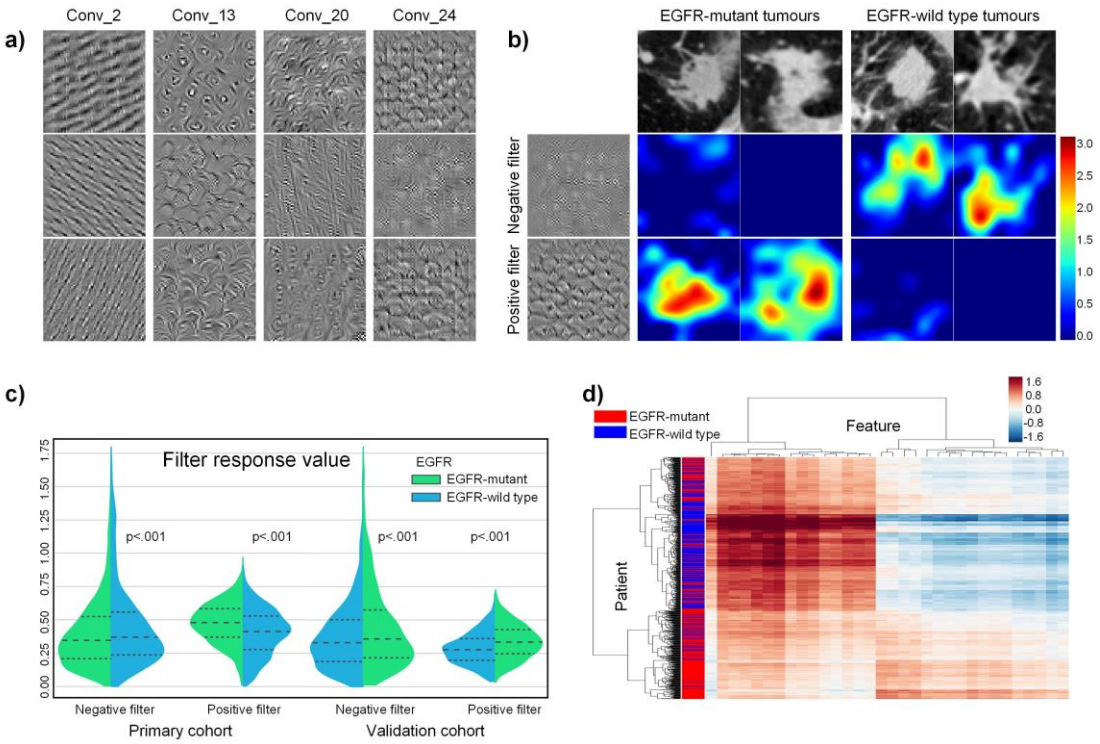
Figure 4. Deep learning feature analysis.

a). Convolutional filters from the 2nd, 13th, 20th, and 24th layers of the DL model. Each convolutional layer includes hundreds of filters, and we illustrate only the first three filters in each layer. **b).** Response of the negative filter and the positive filter in EGFR-mutant/-wild type tumours. The positive filter has strong response to EGFR-mutant tumors and the negative filter has strong response to EGFR-wild type tumors. All the tumour images are from the validation cohort. **c).** Response value of the positive and the negative filters in the two cohorts. The horizontal dotted lines are the quartiles. **d).** Unsupervised clustering of lung adenocarcinoma patients (n = 844) on the vertical axis and deep learning feature expression (feature dimension = 32, the *Conv_24* layer) on the horizontal axis.


Tumour image
Model parameters pre-trained in 1.28 million natural images and fine-tuned in 14,926 CT images
Model parameters trained in 14,926 CT images
EGFR-mutant probability
64x64x3
32x32 @64
16x16 @32
16x16 @32
16x16 @32
16x16 @32
16x16 @32
16x16 @32
16x16 @128
8x8 @32
8x8 @32
8x8 @32
8x8 @32
8x8 @32
8x8 @352
352
92% EGFR-mutant
8% EGFR-wild type
Sub-network 1
Sub-network 2
Convolution with kernel size = 1x1
Convolution with kernel size = 3x3
Batch normalization
Average pooling with window = 2x2
Max pooling with window = 2x2
Global average pooling


a)
ROC curves
True positive rate
False positive rate
DL_Primary (AUC=0.85)
DL_Validation (AUC=0.81)
Radiomics_Primary (AUC=0.70)
Radiomics_Validation (AUC=0.64)
Semantic_Primary (AUC=0.76)
Semantic_Validation (AUC=0.64)
Clinical_Primary (AUC=0.66)
Clinical_Validation (AUC=0.61)
b)
Distribution of the deep learning score
Deep learning score
EGFR
EGFR-mutant
EGFR-wild type
p<.001
p<.001
Primary cohort
Validation cohort
c)
Decision curve of the DL model
Net benefit
Threshold probability
Treat none
Treat all
DL model


EGFR-mutant tumours
EGFR-wild type tumours
Tumour image
Attention map of the DL model
Suspicious area
1.0
0.8
0.6
0.4
0.2
0.0
a)
b)
c)
d)


a)
Conv_2
Conv_13
Conv_20
Conv_24
b)
EGFR-mutant tumours
EGFR-wild type tumours
Negative filter
Positive filter
3.0
2.5
2.0
1.5
1.0
0.5
0.0
c)
Filter response value
EGFR
EGFR-mutant
EGFR-wild type
p<.001
p<.001
p<.001
p<.001
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
Negative filter
Positive filter
Negative filter
Positive filter
Primary cohort
Validation cohort
d)
EGFR-mutant
EGFR-wild type
Feature
1.6
0.8
0.0
-0.8
-1.6
Patient

Online Data Supplement

Supplementary methods

Histological evaluation, Pathological staging and CT imaging protocols

Lung tumours were classified histologically by using the 2015 World Health Organization (WHO) Classification of Tumours of the Lung classification system. For pathological staging, the TNM stage of tumours was determined according to the American Joint Committee on Cancer (AJCC), 7th edition. The scanner parameters from the two hospitals were as following:

Shanghai Pulmonary Hospital: Chest CT images of 603 patients were acquired on Philips Brilliance 40 and Siemens Defintion AS in Shanghai pulmonary hospital. The acquisition parameters of Philips Brilliance 40 were as following: tube voltage = 120 kV; tube current = 200 mA; rotation time = 0.75 s; detector collimation = 40 mm; field of view (FOV) = 30 × 30 cm; pixel matrix=512 × 512; Filter sharp (C) for CT reconstruction; reconstruction thickness=0.75 mm; reconstruction interval=0.75 mm. The Siemens Defination AS used the following acquisition parameters: tube voltage=120 kV; tube current = 130 mA; rotation time = 0.5 s; detector collimation = 40 mm; FOV = 30 × 30 cm; image matrix = 512 × 512; kernel B31f medium sharp+ for CT reconstruction; reconstruction thickness=1.0 mm; reconstruction interval=1.0 mm.

Ioversol (350 mg of iodine per millilitre; Jiangsu Hengrui Medicine, Jiangsu, China) was injected at a dose of 1.3-1.5 mL per kilogram of body weight at a rate of 2.5 mL/sec by using an automated injector.

Tianjin Medical University: In Tianjin medical university cancer institute and hospital, chest CT images of 241 patients were acquired using the three types of CT scanners: Somatom Sensation 64 (Siemens Medical Solutions, Forchheim, Germany), Light speed 16 (GE Medical Systems, Milwaukee, WI), and Discovery CT750 HD scanner (GE Medical Systems, Milwaukee, WI).

For the 64-detector scanner, scanning parameters were as following: 120 kV with tube current adjusted automatically; pitch of 0.969; reconstruction thickness=1.5 mm; reconstruction interval=1.5 mm; pixel matrix=512 × 512. For the 16-detector scanner and Discovery CT750 HD scanner, scanning parameters were as following: tube voltage=120 kV; tube current was 150-200 mA; beam pitch, 0.969; reconstruction thickness=1.25 mm; reconstruction interval=1.25 mm. FOV = 40 ×40 cm; rotation time=0.6s; detector collimation=40 mm; pixel matrix=512 × 512.

Non-ionic iodinated contrast material (300 mg of iodine per millilitre, Ul-travist; Bayer Pharma, Berlin, Germany) was injected at a dose of 1.3‑ 1.5 mL per kilogram of body weight at a rate of 2.5 mL/sec by using an automated injector. CT enhanced scanning was performed with a 70-second delay.

Mathematical description of the DL model

The computational units in the DL model are defined as layers, which include convolution, activation, pooling and batch normalization. The details are explained as following.

***Convolution*.** Convolution is used to extract features from tumour image. Different convolutional filters can extract different features to characterize the tumour. Assuming matrix $I = \begin{pmatrix} I_{11} & I_{12} & I_{13} \\ I_{21} & I_{22} & I_{23} \\ I_{31} & I_{32} & I_{33} \end{pmatrix}$ is the mathematical representation of the

tumour image, and matrix $K = \begin{pmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{pmatrix}$ is the convolutional filter. Then, the output of the convolution layer is $F = conv(I, K)$, where *conv* represents convolutional operation. This can be further understood as the following formula.

$$F = conv(I, K)$$
$$= \begin{pmatrix} I_{11} * k_{11} + I_{12} * k_{12} + I_{21} * k_{21} + I_{22} * k_{22} & I_{12} * k_{11} + I_{13} * k_{12} + I_{22} * k_{21} + I_{23} * k_{22} \\ I_{21} * k_{11} + I_{22} * k_{12} + I_{31} * k_{21} + I_{32} * k_{22} & I_{22} * k_{11} + I_{23} * k_{12} + I_{32} * k_{21} + I_{33} * k_{22} \end{pmatrix}$$

The output F is called feature map.

Activation. After the operation of convolution, the result (feature map) will be activated by an activation function to obtain non-linear features, here we adopt the "ReLU" function[1] $ReLU(x) = max(0, x)$. When the input x is negative, the output of the activation function will be zero, and when the input is positive, the result will be equal to the input.

Pooling. To select representative features that are strongly associated with EGFR mutation status, non-relevant and redundant features need to be eliminated. This is achieved by pooling operation. Assuming the feature map is $F = \begin{pmatrix} 1 & 5 & 2 & 8 \\ 3 & 9 & 7 & 8 \\ 1 & 0 & 2 & 6 \\ 8 & 5 & 3 & 2 \end{pmatrix}$, whose size is 4×4, and pooling window is 2×2 with stride 2. The pooling operation will divide the matrix F into four disjoint small matrixes of size 2×2, each maximum value of the small matrix will be extracted to form the result matrix $P = \begin{pmatrix} 9 & 8 \\ 8 & 6 \end{pmatrix}$.

Batch normalization. To accelerate the training process of the DL model, we use batch normalization [2] operation to normalize the feature maps from each convolutional layer. This strategy avoids gradient vanishing during training, and therefore accelerates the learning process of the DL model.

Details of the DL model

The DL model is similar to the DenseNet [3] but with several modifications. In this model, a stack of two convolutional layers and two batch normalization layers is defined as a group. The first 20 groups form the sub-network 1, where each group is connected to all the preceding groups (dense connection). Sub-network 1 shares the same structure with the first 20 layers in the DenseNet that was pre-trained using 1.28 million natural images. Layers in the sub-network 2 are freshly trained using images from EGFR mutation dataset aiming at capturing the map between image features to EGFR mutation labels. These freshly added convolutional layers are densely connected to the sub-network 1. Finally, this model predicts the probability of the tumour being EGFR-mutant.

Training process of the DL model

Model training aims at optimizing the parameters of the DL model to build the relationship between CT image and EGFR mutation status. The model training is an iterative process, which optimizes the model at each iteration until the model achieves the best predictive performance. At each iteration, we used cross entropy as cost function to measure the predictive performance of the DL model as following:

$$\mathrm{L}(w) = \frac{1}{N}\sum_{n=1}^{N}[y_n log p_n + (1-y_n) log(1-y_n)] + \lambda|w|$$

In this formula, *w* was the parameter of the model that needed to be trained; *N* was the training sample number; y_n represented the true EGFR mutation status (1 for EGFR-mutant, 0 for EGFR-wild type); p_n was the predicted EGFR-mutant probability. λ was the regularization term used to avoid over-fitting, which was set to 5×10^{-4}. If the

cost function $L(w)$ was not minimum, we used Adadelta algorithm [4] to update the parameters of the DL model and minimize the loss function.

Specifically, we froze the sub-network 1 first, and trained the sub-network 2 with a learning rate of 1×10^{-3}. This is necessary because the sub-network 2 was initialized randomly and therefore generated large gradient, which may disturb the transferred layers in sub-network 1. After training the model on 10 epochs, we trained the full network with a smaller learning rate (1×10^{-5}), and the model converged after 30 epochs of training.

To eliminate image intensity variance between different equipment, we standardized the tumour image by z-score normalization, which meant the tumour image was subtracted by the mean intensity value and divided by the standard deviation of the image intensity. In addition, all the tumour images were resized to the same size (64×64) using third-order spline interpolation for the DL model training. Our implementation of the deep learning model used the Keras toolkit and Python 2.7.

Details of deep learning model visualization

We used convolutional filter visualization technique to acquire the feature patterns extracted by convolutional layers [5, 6]. For each convolutional filter in the DL model, we input an image initialized with random white noise to observe the filter response. If the filter response reaches a maximum, the input image reveals the feature pattern extracted by the convolutional filter; otherwise, a back-propagation algorithm was involved to change the input image until the filter response reaches a maximum. Through this convolutional filter visualization method, we can understand the feature patterns extracted by each convolutional filter in the DL model.

Details of suspicious tumour area discovery

When the DL model is well trained, the network established thousands inference paths that work together for EGFR mutation status prediction. Given a tumour, we calculated the gradient of the predicted value with respect to the input image. This gradient told us how the predicted value changes with respect to a small change in tumour image voxels. Hence, visualizing these gradients helped us to find the attention of the DL model [5, 6].

Details of semantic model building

In previous study, 16 semantic features extracted from CT images (e.g., pleural retraction, lymphadenopathy, etc.) were reported to be significantly associated with EGFR mutation status in lung adenocarcinoma [7]. Therefore, we extracted these 16 semantic features in our dataset (definitions listed in Table S4). The semantic features were assessed by two radiologists (10+ years' experience) from the two hospitals. Afterwards, we used multivariate logistic regression to build a semantic model for EGFR mutation status prediction, which is consistent with the published study.

Supplementary Tables

Table S1. Predictive performance of the DL model in different tumour stages.

Stage	AUC	
	Primary cohort	Validation cohort
I	0.87 (0.86, 0.88)	0.81 (0.78, 0.84)
II	0.98 (0.97, 0.99)	0.98 (0.96, 1.00)
III	0.88 (0.84, 0.92)	0.76 (0.72, 0.80)
IV	0.95 (0.91, 0.99)	0.77 (0.68, 0.86)

AUC is area under the receiver operating characteristic curve.

Results in the primary cohort are evaluated in the full primary cohort.

Table S2. Clinical characteristics of patients (n = 125) with other histological types except for adenocarcinoma.

Characteristics	value
Age, mean (SD), years	63.86 (9.44)
Gender, No. (%)	
Female	12 (9.60)
Male	113 (90.40)
Histological type, No. (%)	
Squamous cell carcinoma	96 (76.80)
Large cell carcinoma	17 (13.60)
Sarcomatoid carcinoma	6 (4.80)
Adenosquamous carcinoma	5 (4.00)
Atypical carcinoid	1 (0.80)
Stage, No. (%)	
I	74 (59.20)
II	35 (28.00)
III	15 (12.00)
IV	1 (0.80)
EGFR mutation, No. (%)	
EGFR-mutant	15 (12.00)
EGFR-wild type	110 (88.00)

Table S3. Predictive performance of the DL model in other histological types of lung cancer.

Methods	AUC (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
DL model	0.77 (0.73-0.81)	73.60 (0.71-0.76)	80.00 (72.70-88.02)	72.73 (69.70-75.77)

AUC is area under the receiver operating characteristic curve.

Data in parentheses are the 95% confidence interval.

Table S4. Univariate predictive performance of the semantic features.

Semantic features	Definition	AUC		p-value	
		Primary cohort	Validation cohort	Primary cohort	Validation cohort
Pleural attachment	0-none; 1-tumor attaches to the pleura	0.537	0.422	<0.001	<0.001
Border definition	1-well defined; 3-poorly defined; 2-otherwise	0.346	0.474	<0.001	0.238
Spiculation	1-none; 2-fine spiculation; 3-coarse spiculation	0.502	0.608	<0.001	<0.001
Texture	1-pure GGO; 2-mixed GGO; 3-solid	0.433	0.360	<0.001	<0.001
Air bronchogram	0-none; 1-presence of air bronchogram	0.519	0.564	<0.001	<0.001
Bubblelike lucency	0-none; 1-presence of bubblelike lucency	0.531	0.518	<0.001	0.182
Enhancement heterogeneity	1-homogeneous; 2-slight or moderate heterogeneous; 3-marked heterogeneous	0.433	0.485	<0.001	0.002
Vascular convergence	0-none; 1-obvious convergence	0.489	0.692	<0.001	<0.001
Thickened adjacent bronchovascular bundles	0-none; 1-normally tapering bundle leading to the nodule was observed to be distinctly widened	0.484	0.679	<0.001	<0.001
Pleural retraction	0-none; 1-presence of pleural retraction	0.431	0.551	<0.001	0.017
Peripheral emphysema	1-none; 2-slight or moderate focal emphysema; 3-severe focal emphysema	0.484	0.411	<0.001	<0.001
Peripheral fibrosis	1-none; 2-slight or moderate focal fibrosis; 3-severe focal fibrosis	0.739	0.447	<0.001	0.002
Lymphadenopathy	1-Thoracic lymph nodes (hilar or mediastinal) with short-axis diameter greater than 1 cm; 0-otherwise	0.533	0.437	<0.001	0.004
Size category	1-diameter≤3 cm; 2-diameter>3 cm	0.486	0.329	<0.001	<0.001
Long-axis diameter	Longest diameter of the tumor (cm)	0.506	0.287	0.699	<0.001
Short-axis diameter	Shortest diameter of the tumor (cm)	0.464	0.306	0.254	<0.001

AUC is area under the receiver operating characteristic curve.
p-value is generated by independent samples t test for long-axis diameter and short-axis diameter, and chi-squared test for other categorical semantic features.

Supplementary Figures

Figure S1. The ROIs selected by users.

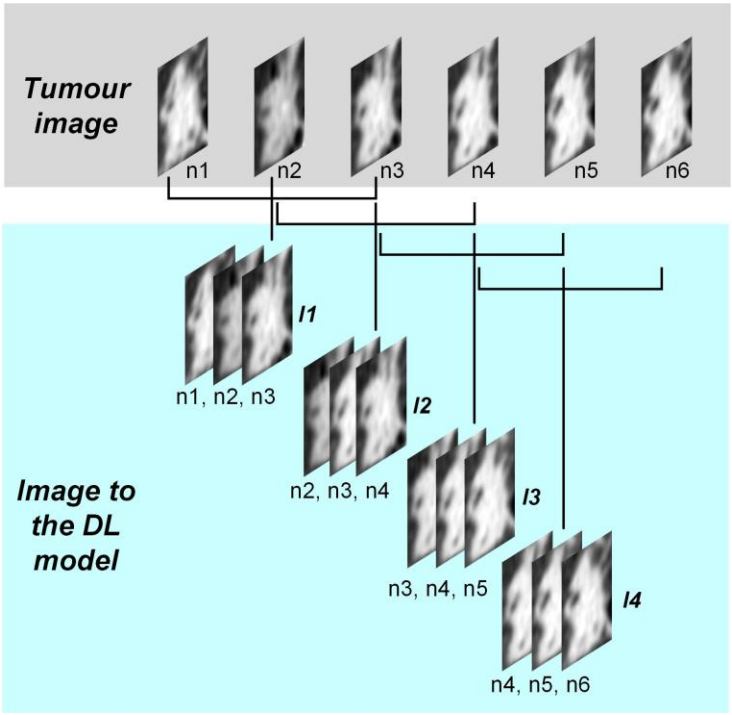


Figure S2. The process of generating input images to the DL model. All adjacent three image slices were combined as a three-channel image to the DL model. n1 to n6 represent the slice numbers of the axial CT images. I1 to I4 are the four input images to the DL model.

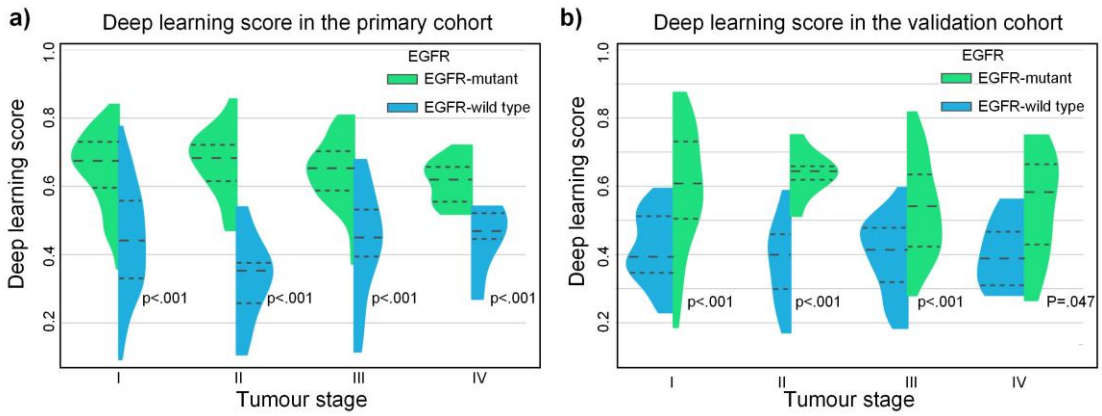


Figure S3. Deep learning score distribution in different tumour stages. The horizontal dash lines are the quartiles.

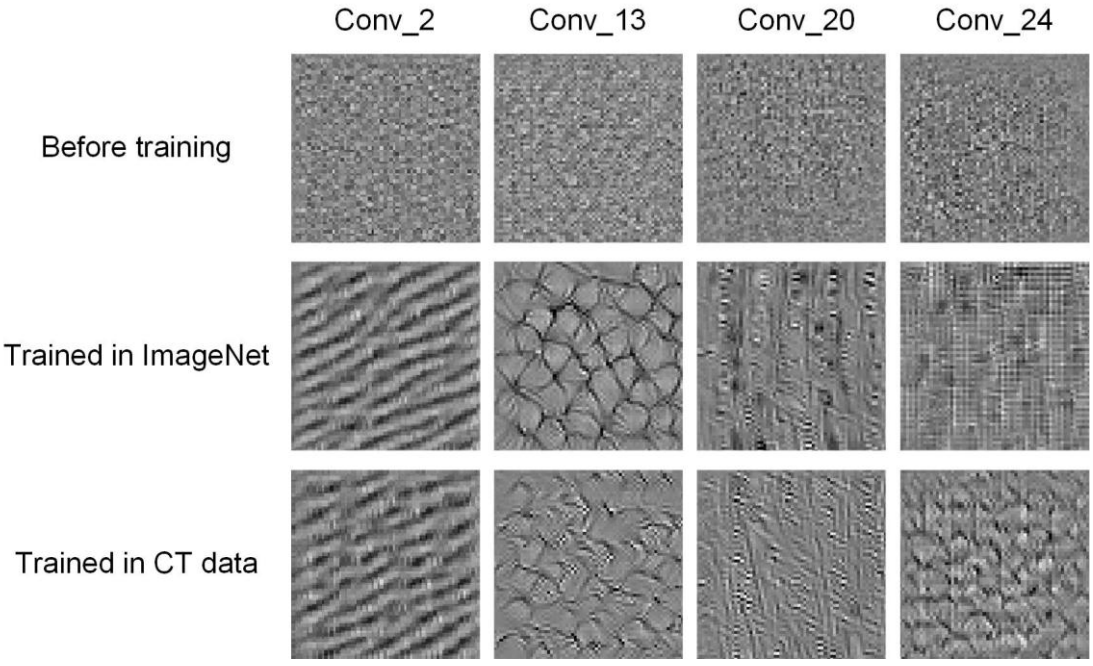


Figure S4. Convolutional filters trained in different datasets. Each column represents the same convolutional filter in different status (before training, trained in ImageNet, and trained in CT data).

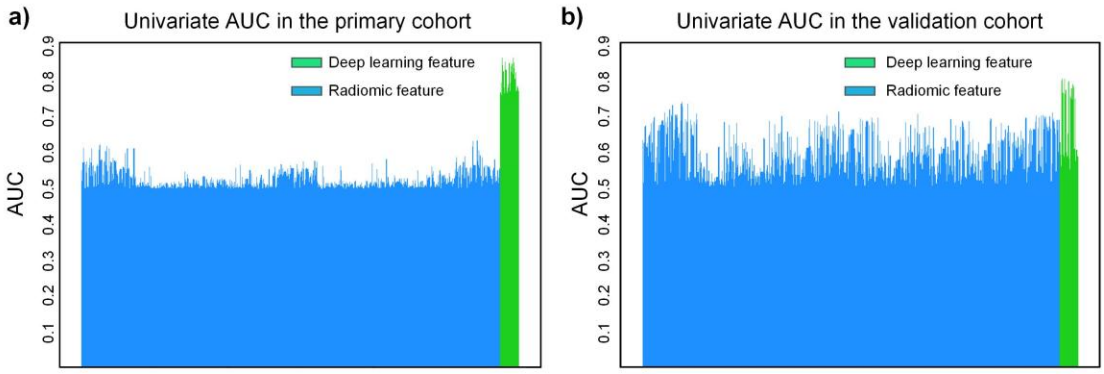


Figure S5. Univariate AUC testing for all the deep learning features from the *Conv_24* layer and radiomic features.

References

1. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems; 2012; 2012. p. 1097-1105.

2. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:150203167* 2015.

3. Huang G, Liu Z, Weinberger KQ, van der Maaten L. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017; 2017. p. 3.

4. Zeiler MD. ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:12125701* 2012.

5. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: 2017 IEEE International Conference on Computer Vision (ICCV); 2017; 2017. p. 618-626.

6. Kotikalapudi Rac. keras-vis. GitHub, https://github.com/raghakot/keras-vis, 2017.

7. Liu Y, Kim J, Qu F, Liu S, Wang H, Balagurunathan Y, Ye Z, Gillies RJ. CT features associated with epidermal growth factor receptor mutation status in patients with lung adenocarcinoma. *Radiology* 2016: 280(1): 271-280.