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.

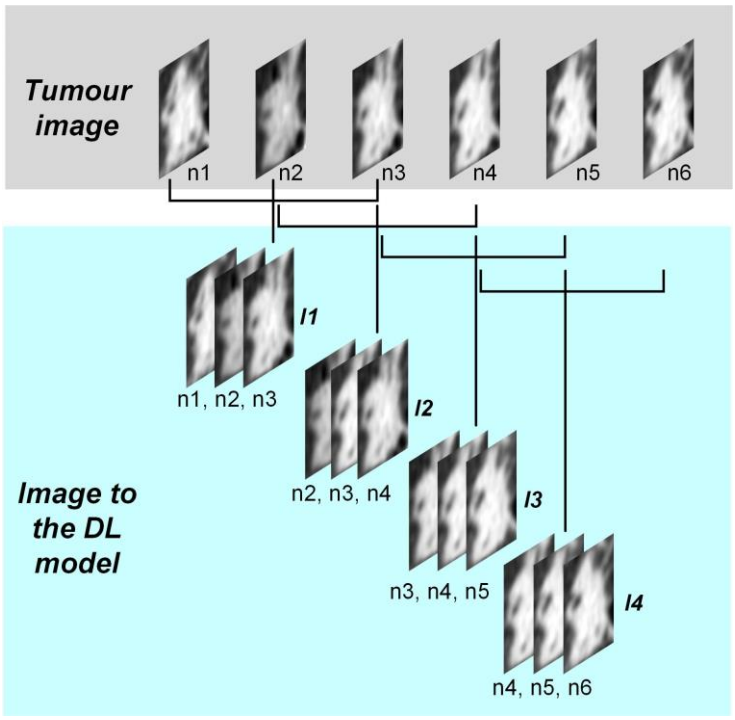

Figure S2. The process of generating input images to the DL model. All adjacent three image slices were combined as a three-channel image to the DL model. n1 to n6 represent the slice numbers of the axial CT images. I1 to I4 are the four input images to the DL model.

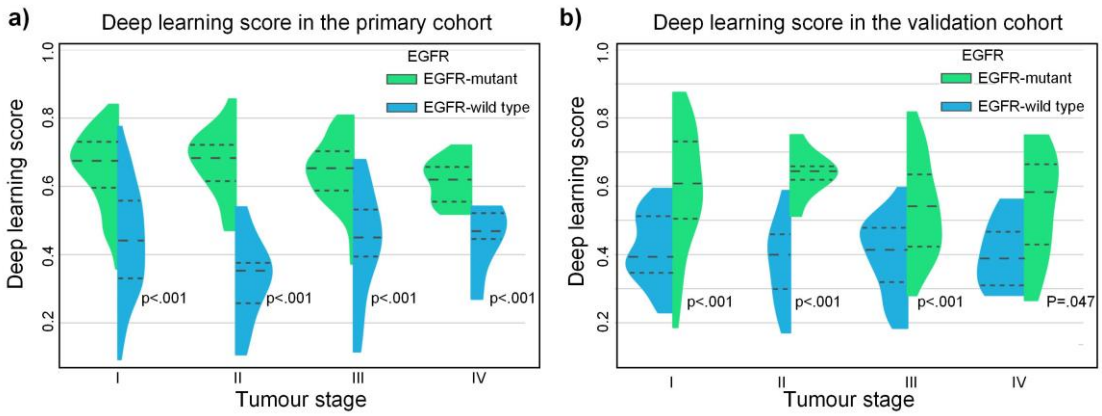

Figure S3. Deep learning score distribution in different tumour stages. The horizontal dash lines are the quartiles.

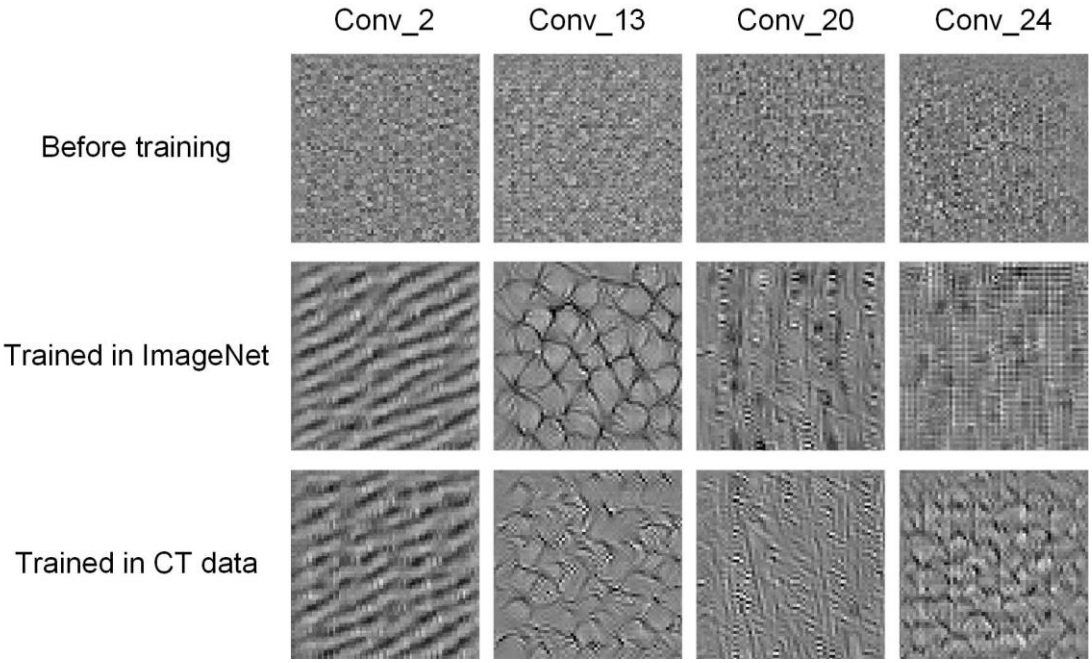

Figure S4. Convolutional filters trained in different datasets. Each column represents the same convolutional filter in different status (before training, trained in ImageNet, and trained in CT data).

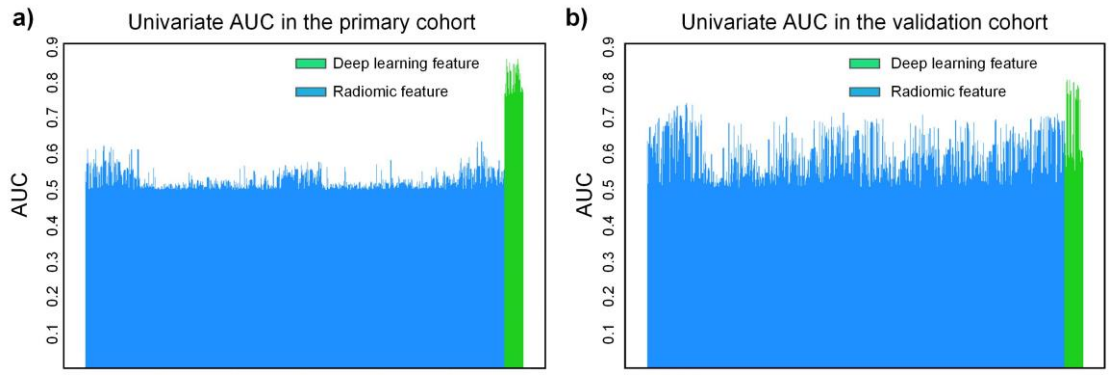

Figure S5. Univariate AUC testing for all the deep learning features from the *Conv_24* layer and radiomic features.

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.