Published OnlineFirst October 2, 2018; DOI: 10.1158/0008-5472.CAN-18-0696

Convergence and Technologies

Cancer Research

3D Deep Learning from CT Scans Predicts Tumor Invasiveness of Subcentimeter Pulmonary Adenocarcinomas

Wei Zhao[1,2], Jiancheng Yang[3,4,5], Yingli Sun[1], Cheng Li[1], Weilan Wu[1], Liang Jin[1], Zhiming Yang[1], Bingbing Ni[3,4], Pan Gao[1], Peijun Wang[6], Yanqing Hua[1], and Ming Li[1,2]

Abstract

Identification of early-stage pulmonary adenocarcinomas before surgery, especially in cases of subcentimeter cancers, would be clinically important and could provide guidance to clinical decision making. In this study, we developed a deep learning system based on 3D convolutional neural networks and multitask learning, which automatically predicts tumor invasiveness, together with 3D nodule segmentation masks. The system processes a 3D nodule-centered patch of preprocessed CT and learns a deep representation of a given nodule without the need for any additional information. A dataset of 651 nodules with manually segmented voxel-wise masks and pathological labels of atypical adenomatous hyperplasia (AAH), adenocarcinomas *in situ* (AIS), minimally invasive adenocarcinoma (MIA), and invasive pulmonary adenocarcinoma (IA) was used in this study. We trained and validated our deep learning system on 523 nodules and tested its performance on 128 nodules. An observer study with 2 groups of radiologists, 2 senior and 2 junior, was also investigated. We merged AAH and AIS into one single category AAH-AIS, comprising a 3-category classification in our study. The proposed deep learning system achieved better classification performance than the radiologists; in terms of 3-class weighted average F1 score, the model achieved 63.3% while the radiologists achieved 55.6%, 56.6%, 54.3%, and 51.0%, respectively. These results suggest that deep learning methods improve the yield of discriminative results and hold promise in the CADx application domain, which could help doctors work efficiently and facilitate the application of precision medicine.

Significance: Machine learning tools are beginning to be implemented for clinical applications. This study represents an important milestone for this emerging technology, which could improve therapy selection for patients with lung cancer.
Cancer Res; 78(24); 6881–9. ©2018 AACR.

Introduction

Lung cancer is the leading cause of cancer-related deaths in the world. The International Association for the Study of Lung Cancer (IASLC) International Staging Project confirms that a logical degradation of survival results, as tumor size increases (1), indicating that early detection and diagnosis is an effective and crucial way to decrease the mortality of patients with lung cancer. Lung cancer screening with low-dose computed tomography (LDCT) in high-risk patients (age >50 year, smoking history, family history lung cancer in first-degree relatives, etc.) has remarkably facilitated early stage pulmonary adenocarcinoma detection and diagnosis, especially for the nodules less than 1 cm in diameter (subcentimeter; refs. 2, 3). Previous screening programs have shown, 60% to 70% of detected lung cancers were in stage I, and 56% were subcentimeter lesions (1). However, the management of subcentimeter tumors encountered on screening CT images remains controversial. 10 mm diameter is used as a cutoff value to distinguish preinvasive (atypical adenomatous hyperplasia, AAH; adenocarcinoma in situ, AIS) and invasive lesions (minimally invasive adenocarcinoma, MIA; invasive pulmonary adenocarcinoma, IA) on CT images (4). However, previous studies reported that some subcentimeter ground-glass opacity nodules (GGN) may be MIA or IA (5, 6), and many of these are also recorded in the institution (see Fig. 1 for some examples). Prognosis varies widely among the different pathologic subtypes (7). Therefore, early identification of the invasive characteristics before surgery would be clinically important and could provide guidance to the clinical decision-making. However, subcentimeter GGNs presented on CT images make the differential diagnosis clinically difficult due to the absence of typical radiographic features (bubble lucency, pleural retraction, spiculated margin,

[1]Department of Radiology, Huadong Hospital Affiliated to Fudan University, Shanghai, P.R. China. [2]Diagnosis and Treatment Center of Small Lung Nodules of Huadong Hospital, Shanghai, P.R. China. [3]Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, P.R. China. [4]SJTU-UCLA Joint Center for Machine Perception and Inference, Shanghai Jiao Tong University, Shanghai, P.R. China. [5]Diannei Technology, Shanghai, P.R. China. [6]Department of Radiology, Tongji Hospital, School of Medicine, Tongji University, Shanghai, P.R. China.

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

W. Zhao and J. Yang contributed equally to this article.

Corresponding Authors: Ming Li, Huadong Hospital Affiliated to Fudan University, Shanghai 200040 P.R. China. Phone: 86-138-1662-0371; Fax: 86-21-5764-3271; E-mail: minli77@163.com; and Yanqing Hua, Department of Radiology, Huadong Hospital Affiliated to Fudan University, Shanghai 200040 P.R. China. E-mail: huayq007@163.com

doi: 10.1158/0008-5472.CAN-18-0696

Published OnlineFirst October 2, 2018; DOI: 10.1158/0008-5472.CAN-18-0696

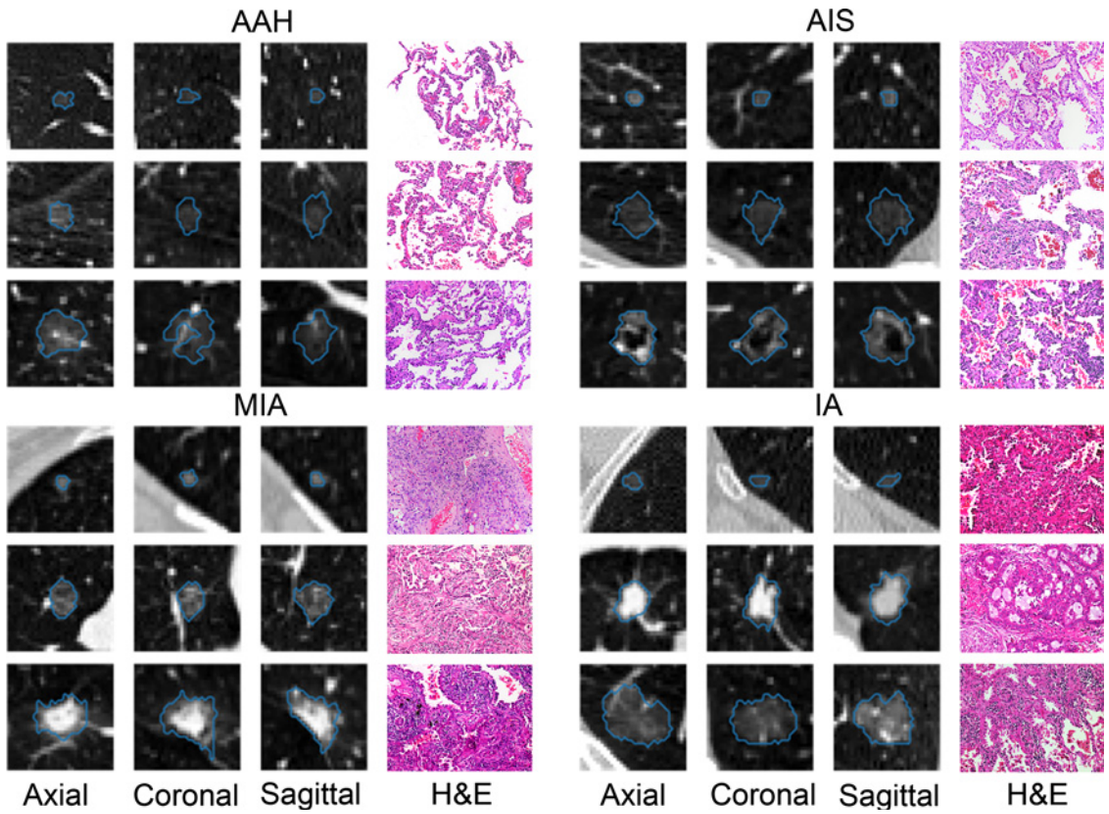


Figure 1.
Examples in the dataset of nodule patches in axial, coronal, and sagittal views. The type and volume in a cubic millimeter of nodules are shown in the subtitles. The blue contours represent the manually labeled boundary of the nodules. Each patch is depicted in a size of 32 mm. H&E, hematoxylin and eosin stain. Magnification, ×200.

etc.) of early cancers and may confuse clinical decision-making. In addition, evaluating a large number of detected GGNs by the experts or radiologists can still be time-consuming. In this context, computer-aided diagnosis (CADx), a much more efficient and effective way to evaluate the detected nodules, is expected to play an important role in the clinical evaluating task and is the hot spot of the current research.

In recent years, deep learning has become a powerful method of representation learning (8), reducing the necessities of hand-craft feature engineering. With the help of end-to-end deep supervised learning, especially convolutional neural networks, great progress have been made in natural image problems, such as image recognition (9), object detection (10), semantic segmentation (9). The effectiveness of deep learning has been proved in medical image analysis as well, such as recent progress in skin cancer classification (11), diabetic retinopathy detection (12), and pulmonary nodule detection in chest CT (13). With hierarchical representation learning in 3D views, deep neural networks can discover new patterns beyond the typical radiographic features, which may be invisible or subtle to human eyes and traditional CADx systems.

In this article, a deep learning system is designed to address the problem of automatically predicting the tumor invasiveness of subcentimeter pulmonary adenocarcinomas from CT scans. The main contributions are 3-fold: First, we proposed an automatic framework to predict the tumor invasiveness, trained with preprocessed chest CTs and the corresponding pathological labels. The proposed method does not require the nodule segmentation, estimate of nodule size, or other predefined features in the inference stage. To the best of our knowledge, this is the first automatic learning system for this problem. The neural networks are trained and validated using a dataset of 523 nodules, and a hold-out test set of 128 nodules is used to fairly evaluate our system. Instead of classifying the 4 categories (AAH, AIS, MIA, and IA), the problem is formalized as 3-category classification (AAH and AIS into a single class) due to some technical limitations (will be explained in "Materials and Methods"). Second, to make full use of the datasets, we proposed a "bottom-up and top-down" multitask learning architecture to predict the nodule invasiveness together with segmentation mask. Through joint training of the neural networks to solve these two related tasks, the model is able to attend to the areas that deserve more attention. In practice, this multitask learning approach is effective in both accuracy and training convergence, and makes the system less prone to be overfitting. Finally, to fairly compare the performance with human, an observer study with 2 groups of 4 radiologists, 2 experienced and 2 junior doctors is conducted, to classify the 128 hold-out nodules. It is shown that the proposed deep learning system achieves better classification performance than radiologists in the observer study.

Published OnlineFirst October 2, 2018; DOI: 10.1158/0008-5472.CAN-18-0696

Materials and Methods

Data collection

This retrospective study was approved by the institutional review board of Huadong Hospital affiliated to Fudan University (NO.2017K062), which waived the requirement for patients' written informed consent referring to the CIOMS guideline.

From October 2011 to October 2017, a search of the electronic medical records and the radiology information systems of the hospital for patients with subcentimeter pulmonary nodules identified on chest CT scans was performed by one author (Yingli Sun). A total of 651 subcentimeter nodules from 560 patients [mean age, 54.1 years ± 12.2 (SD); range, 16–82 years] were enrolled in the study. There are 182 men [mean age, 54.6 years ± 12.2 (SD); range, 26–82 years] and 378 women [mean age, 53.9 years ± 12.2 (SD); range, 16–80 years]. The unbalanced distribution of gender was determined by a unique characteristic of female predominance in this case (7, 14). The inclusion criteria are as follows:

1. The presence of thin-slice chest CT (1–1.25 mm) scan before surgical treatment.
2. Nodules noted on CT examination with a diameter $\leq$10 mm.
3. No treatment before surgical treatment.

Among the 651 subcentimeter nodules (see Table 1), 205 nodules were pathologically identified as preinvasion lesions (39 AAH and 166 AIS), where 446 nodules were invasive lesions (316 MIA and 130 IA). On preoperative CT evaluation, 21, 284, 346 of the 651 nodules were classified into solid, part-solid and pure GGNs, respectively.

Preoperative chest CT was performed by using the following four scanners: GE Discovery CT750 HD (143 nodules), 64-slice LightSpeed VCT (199 nodules; GE Medical Systems); Somatom Definition flash (150 nodules), Somatom Sensation-16 (159 nodules; Siemens Medical Solutions) with the following parameters: 120 kVp; 100–200 mAs; pitch, 0.75–1.5; and collimation, 1–1.25 mm, respectively. All imaging data were reconstructed by using a medium sharp reconstruction algorithm with a thickness of 1–1.25 mm. 259 of the 561 patients were then administered contrast material after non-contrast enhanced CT scan. In the case of contrast-enhanced CT, a bolus of 80–100 mL of IV contrast medium (350 mg I/mL; Optiray, Mallinckrodt) was administered at a rate of 3–4 mL/s with the use of a power injector via an 18- or 20-gauge cannula in an antecubital vein. The contrast-enhanced CT scan was acquired 60 seconds after the administration of contrast medium. In this study, only the unenhanced CT images of the latest CT examination before surgery were collected. In all patients, CT images were acquired in the supine position at full inspiration. The mean interval between the latest CT examination and surgery was 13 days (range, 1–132 days; median, 7 days).

Table 1. Number of nodules for training, validation, and testing

	Training and validation	Testing	Total
AAH	33	6	39
AIS	134	32	166
MIA	252	64	316
IA	104	26	130
AAH-AIS	167	38	205
Total	523	128	651

NOTE: To make full use of the training data, data augmentation was performed on the fly during the training process.

Nodule labeling and segmentation

A medical image processing and navigation software 3D Slicer (version 4.8.0, Brigham and Women's Hospital) was used to manually delineate the volume of interest (VOI) of the included 651 subcentimeter nodules at voxel level by one radiologist (Yingli Sun, with 5 years of experience in chest CT interpretation), then the VOI was confirmed by another radiologist Ming Li (with 12 years of experience in chest CT interpretation). Large vessels and bronchioles were excluded as much as possible from the volume of the nodule. The lung CT DICOM (Digital Imaging and Communications in Medicine) format images were imported into the software for delineating, and then the images with VOI information are extracted with NII format for next step analysis. Each segmented nodule was given a specific pathological label (AAH, AIS, MIA, IA), according to the detailed pathological report.

Dataset pretreatment

The data collected for this research is split into 5 parts: Subset 0, 1, 2, 3, and 4, each subset is selected by randomly choosing 20% of each of the 4 categories. Subset 4 is the hold-out test set and is never used before evaluation. Subset 0–3 is used for training and validation. See Table 1 for a detailed number of the nodules for training, validation, and testing. Hyperparameters are validated via cross-validation on Subset 0–3, and then the model with all of Subset 0–3 is trained with fixed hyperparameters.

However, the AAH samples are clearly too few for fairly training the deep neural networks. This is inherently destined because these particular lesions are usually considered as benign, and they rarely undergo surgical treatment unless obvious malignant signs are presented in the CT images. Therefore, pathologically identified AAH nodules are rare in practice. The study merged the samples labeled as AAH and AIS into a single class "AAH-AIS," to avoid the problem of shortage in training samples. Fortunately, it's still reasonable in the clinical context, as these two subtypes of lesions ($\leq$3 cm) are reported to have a 100% disease-specific survival if they are completely resected (7). In this way, the invasiveness prediction is treated as a 3-category classification problem in this work.

In the development of the deep learning system, each data sample is defined as:

1. A 3D patch of 32 mm × 32 mm × 32 mm, cropped from the CT scan at the mass center of a nodule.
2. The pathologically identified label of invasiveness, in one of AAH-AIS, MIA, and IA.
3. Manually labeled voxel-wise nodule mask.

For efficient training the networks, online data augmentation is performed. The details of hyper-parameter setting, generation of 3D patches, neural network design and training will be explained further.

Observer study

To compare the deep learning system with human performance, four radiologists (two senior radiologists, Ming Li, Weilan

Published OnlineFirst October 2, 2018; DOI: 10.1158/0008-5472.CAN-18-0696

Wu, with more than ten years of experience in chest CT interpretation; and two junior radiologists, Wei Zhao, Zhiming Yang, with more than 3 years of experience in chest CT interpretation) were enquired. They were blinded to the histopathologic results and clinical data independently to classify and diagnose all the test set nodules. Four chest radiologists classified the nodules on the basis of the new classification standard of lung adenocarcinoma published in 2011 (7).

Deep learning system

The input of the proposed model is a cubic patch of 32 mm × 32 mm × 32 mm, generated by a (preprocessed) chest CT scan and the position $c = [z, y, x]$, that is, the mass center (roughly) of the nodule, which can be marked manually or obtained by an automatic nodule detection system (13). The output of the model is the categorical probability for the 3 categories (AAH-AIS, MIA, IA), together with the model-generated mask of the nodule segmentation. The framework is based on the proposed 3D convolution neural networks (CNN), referred as *DenseSharp* Networks, which processes the input cubes via a "bottom-up and top-down" architecture: The classification head as bottom-up path can enforce the network to extract meaningful features for diagnosis; meanwhile, the segmentation head works as the top-down path, and is able to teach the network to attend the regions of interest (ROI). With multitask learning, DenseSharp Networks can learn the classification and segmentation tasks end-to-end efficiently.

Generation of 3D patches

The 3D patches are generated by cropping the preprocessed volumetric data into a size of 32 × 32 × 32 (voxels, 1 voxel denotes 1 mm). The preprocessing follows "standard" procedure for chest CT: the input CT scans are converted into Hounsfield units, followed by resizing of volumetric data into spacing of 1 mm × 1 mm × 1 mm by trilinear interpolation, clipping the voxel intensity into $I_{HU} \in [-1024, 400]$, quantifying the density into grayscale, and transforming the values to $I \in [-1, 1)$ by a mapping $I = \left[\frac{I_{HU} + 1024}{400 + 1024} \times 255\right]/128 - 1$.

The study uses many data augmentation techniques to increase the training data size, including:

1. Rotation by 90° increments
2. Left-right flipping
3. Transposition by small amounts in $[-3, 3]$ voxels in each axis
4. Reordering of axes
5. Zooming with a ratio in $[0.8, 1.15]$.

For the efficient use of the training data, data augmentation is performed on the fly during the training process, which acts as strong regularization for our models. Other sophisticated augmentation techniques like elastic transformation and salt-and-pepper noise are also tried, however, there seems no significant improvement.

The DenseSharp architecture

Because of the limited availability of data, the learning network should be very compact to make the training procedure relatively easy. DenseNets (15) have indicated compelling accuracy with more efficient use of parameters on natural image recognizing tasks; To leverage the power of dense connectivity, the study extended the 2D DenseNets into a 3D variant following the "bottleneck" and "compression" design (15), which naturally becomes the bottom-up classification head for predicting the invasiveness labels. Inspired by DeepMask (16) and SharpMask (17), a top-down segmentation head is used for predicting the nodule mask located near the center of patches, using shared features extracted by the same network. The study emphasizes, however, that the segmentation head is mainly used to teach the neural network to attend where it needs to pay more attention to, thus the segmentation head is designed to be lightweight, which consists of only transposed convolution without nonlinear activation.

The architecture of the proposed DenseSharp Networks is illustrated in Fig. 2. Specifically, it consists of stacked densely connected blocks (i.e., Dense Block), and each of the blocks consists of several convolutional modules (4 Dense Block Modules for this task). In each convolutional module, 1 × 1 × 1

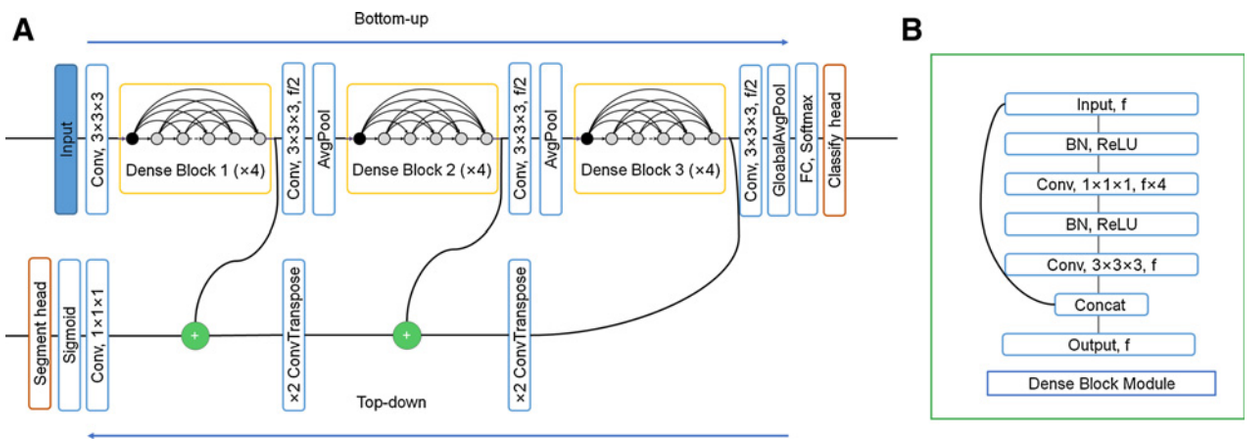


Figure 2.

The architecture of the proposed DenseSharp Network. **A,** The "classify head" gives the invasiveness labels, whereas the "segment head" gives the nodule mask. **B,** The illustration of the convolutional module in the Dense Block.

Published OnlineFirst October 2, 2018; DOI: 10.1158/0008-5472.CAN-18-0696

convolution kernels with 64 filters followed by $3 \times 3 \times 3$ convolution kernels with 16 filters (growth rate $k = 16$), the so-called "bottleneck" techniques (bottleneck $B = 4$), are used for efficient 3D representation learning. Batch normalization (18) layers are used for reducing internal covariance shift, and the rectified linear units, that is, $\mathrm{ReLU}(x) = \max(0, x)$, act as the nonlinear transform. The input features and the transformed features by the convolutional module are concatenated before sending to the next module, consequently, the subsequent layers receive feature maps from all their preceding layers. After an entire Dense Block, the feature maps will be "compressed" using $1 \times 1 \times 1$ convolution kernels with halved filters (compression $C = 2$), and then down-sampled by pooling. Finally, the last Dense Block and global pooling layer (19) get a representation of 120 channels, and fully connected layers with softmax activation as the classification head outputs the invasiveness labels.

The lightweight segmentation head outputs the nodule mask using the features extracted by the three Dense Blocks. Inspired by (19), rather than restoring the original resolution by a large up-sampling transposed convolutional layer directly, the feature maps are up-sampled gradually, and the high-resolution low-level features and the up-sampled high-level features are summed-up (shown in the top-down path). In this way, the low-level and high-level features are well combined to predict the segmentation masks, and the segmentation head becomes a top-down path for the entire network. The detailed architecture of the 3D DenseSharp Networks is shown in Supplementary Table S1.

The network was implemented using Python 3.6 based on TensorFlow 1.4.0 (20) and Keras 2.1.5 (21) deep learning library and trained the neural networks on a workstation with 2 NVIDIA TITAN X GPUs.

Code is open source at https://github.com/duducheng/DenseSharp/.

Training

The proposed DenseSharp Networks have two output heads, trained based on different loss. Stochastic gradient descent was used to minimize cross-entropy between the classification outputs and target labels for training the classification head. It was also used to maximize the Dice coefficient between the predicted masks and the real nodule masks for training the segmentation head. The two heads are trained jointly with a multitask loss l_{joint},

$$\ell_{joint} = \ell_{cls} + \lambda \ell_{seg}$$

$$\ell_{cls}(y_{cls}, t_{cls}) = -\frac{1}{n}\sum_{n}\sum_{c} t_{cls}^{c} \log y_{cls}^{c}$$

$$\ell_{seg}(y_{seg}, t_{seg}) = -\frac{1}{n}\sum_{n}\frac{2 \times \sum y_{seg} t_{seg}}{\sum y_{seg} + \sum t_{seg}}$$

l_{cls} shows the cross-entropy loss, y_{cls} is the output of classification head, t_{cls} is the invasiveness label. l_{seg} shows the dice loss, y_{seg} is the output of segmentation head, and t_{seg} is the manually labeled nodule mask. The study chose $\lambda = 0.2$ since segmentation works as an auxiliary supervised task.

To train the ConvNets, all the network parameters are well initialized using "he uniform" method (9). We have tried to pretrain the neural networks on LIDC-IDRI (22), a widely used lung nodule database, in a multitask learning scheme with radiological benign or malignant labels and nodule segmentation; However, it did not help in practice in terms of classification performance. During optimization, the study sampled the training data with a ratio of 1 : 1 : 1 for the 3 classes with a batch size of 24, and used Adam (23) optimizer with a fix learning rate of 10^{-4} to update the model parameters. We early stop the training after 60 epochs. No weight decay nor dropout (24) has been used in the network.

Prediction

Given an input 3D patch of 32 mm $\times$ 32 mm $\times$ 32 mm from CT scan, the trained network is able to predict the 3-class probability of invasiveness together with the nodule mask with the single forward pass. Because of the randomness in the neural network optimization process, the 15-run ensemble is constructed to reduce the error variance, which is an average result of 15 experiments with the same hyper-parameter setting. The predicted invasiveness labels are assigned by $y = \mathrm{argmax}_k \frac{1}{15}\sum_i^{15} proba_i$.

Results

Evaluation on three-category classification

After training, all nodules in the test set were processed by the proposed deep network, namely *DenseSharp* Network, a multitask architecture for classification and segmentation. As mentioned earlier, instead of classifying the 4 categories of AAH, AIS, MIA, and IA, the study merged AAH and AIS into one single category AAH-AIS, which makes a 3-category classification.

The study evaluated the classification performance of the best result on the test set, which is an ensemble of 15 experiments with the same setting to reduce the variance of neural network training. Because of the skewness of the distribution of the 3 nodule types, the classification accuracy [$\mathrm{Accuracy} = \frac{1}{n}\sum 1(y_i = t_i)$], per-class F1-score ($\mathrm{F1_{cls}} = \frac{2\mathrm{Precision_{cls}} \times Recall_{cls}}{\mathrm{Precision_{cls}} + Recall_{cls}}$) and weighted average F1-score ($\mathrm{F1_{cls}} = \frac{n_{AAH-AIS}F1_{AAH-AIS} + n_{MIA}F1_{MIA} + n_{IA}F1_{IA}}{n_{AAH-AIS} + n_{MIA} + n_{IA}}$) is compared with human radiologist performance. Besides, multi-class Matthews correlation coefficient (MCC; ref. 25), a metric less sensitive to class imbalance, is also used for evaluation. The results

Table 2. Three-category classification performance for nodule invasiveness, in terms of accuracy, per-class F1-score, weighted average F1-score, and MCC

	Accuracy	$F1_{AAH-AIS}$	$F1_{MIA}$	$F1_{IA}$	$F1_{AVG}$	MCC
3D DenseSharp Network	64.1%	55.7%	68.1%	62.7%	63.3%	0.407
3D DenseNet	59.4%	45.2%	62.9%	66.7%	58.4%	0.332
2D DenseNet	43.0%	54.9%	50.5%	59.3%	53.6%	0.293
Pretrained Inception-v3	35.9%	55.6%	34.8%	53.6%	44.9%	0.249
Senior 1	55.4%	50.0%	55.9%	63.0%	55.6%	0.304
Senior 2	56.3%	50.6%	57.9%	62.5%	56.6%	0.307
Junior 1	53.9%	49.4%	53.3%	63.8%	54.3%	0.271
Junior 2	50.8%	48.9%	59.6%	48.7%	51.0%	0.234

NOTE: "3D DenseSharp Network" denotes the results of our proposed network, and "3D DenseNet" denotes the performance without multitask learning. A 2D DenseNet with similar architecture and comparable parameters and an Inception-v3 pretrained on ImageNet database processing 2.5D (multiview) CT images are shown for comparison. Results for four observers (2 senior and 2 junior radiologists) are also reported. The higher is better.

Published OnlineFirst October 2, 2018; DOI: 10.1158/0008-5472.CAN-18-0696

of the comparison are reported in Table 2, the best model (3D DenseSharp Network) achieves better classification performance in terms of all metrics except for the minor disadvantage in $F1_{IA}$, even compared with senior radiologists, indicating the effectiveness of the proposed method. The proposed model without multitask learning (3D DenseNet) continues to achieve classification performance at a level matching or exceeding the observers.

Two 2D deep convolutional neural networks are used for comparison. As depicted in Table 2, the 2D (or 2.5D) CNNs work less well than the 3D ones. These 2D networks process 3-channel inputs of 2.5D CT images (on axial-coronal-sagittal views), see Supplementary Fig. S1 for illustration. The 2D DenseNets follow similar design pattern with our 3D DenseNets, using 2D convolutions instead of 3D convolutions; besides, we change the depth and filters of 2D DenseNets to keep the number of trainable parameters comparable with our 3D DenseNets. On the other hand, Inception-v3 Networks (26) have achieved a great success in both natural images and 2D medical images (11). We use an Inception-v3 network pretrained on ImageNet (27), and fine tune the network on the 2.5D CT images. See more in Supplementary "The details on the 2D CNNs processing the 2.5D CT images."

The 3-class confusion matrix is shown in Table 3. The model is not prone to make severe mistakes: Nodules labeled as AAH-AIS will not be predicted as IA, and those labeled as IA will not be predicted as AAH-AIS. It means the model implicitly learns the relationship of the 3 categories. However, the observers' results don't hold this property.

Evaluation on two subtasks of binary classification

To fairly analyze the classification performance of our model trained on three-category classification, we have considered two clinically important subtasks: (i) binary classification of invasive nodules (IA or MIA) and preinvasive nodules (AIS or AAH), and (ii) binary classification of IA nodules and non-IA nodules (MIA, AIS or AAH).

The subtask (a) is urgently needed in clinical practice. According to the recently proposed IASLC/ATS/ERS classification (7), the lesions correspond to preinvasive AAH or AIS sufficiently often warrant a conservative approach emphasizing long-term CT surveillance, whereas MIA and IA need elective or immediate surgery treatment due to a worse prognosis than preinvasive lesions. On the subtask (a), we merged the output score, by the trained model for three-category classification, for IA and MIA by addition, that is, $\gamma^{\text{invasive}} = \gamma^{\text{IA}} + \gamma^{\text{MIA}}$. In this way, our model achieved an area under receiver operating characteristic curve (AUC) of 0.788 on the subtask (a).

However, patients with MIA have a disease-free survival rate close to 100% if they are safely treated with limited resection (7), whereas those with IA have a disease-free survival rate of only 60% to 70% (28–30), indicating that more aggressive surgical treatment and subsequent treatment (e.g., chemotherapy) were needed. Besides, a few previous studies have merged AAH, AIS, and MIA into one category as well (31, 32). These are the reason why we addressed the subtask (b). We considered solely the output score of IA for the subtask (b), and our model achieved an AUC of 0.880.

As depicted in Fig. 3, the deep models trained with three-category classification approach produced competitive performances on two subtasks of binary classification, which were on par with, if not better than, the performances of the radiologists in our observer study. In fact, this strategy, that is, training the CNNs on finer disease partition, but running coarse inference, have been successfully applied in previous study (11). See Supplementary Tables S2 and S3 for detailed evaluation metrics of accuracy, weighted average F1-score, MCC and AUC on these two subtasks, and Supplementary Table S4 for the original experimental results.

Importance of multitask learning

The study argues that the top-down segmentation head is critical for training the bottom-up classification head, which teaches the network to attend the nodule part. The DenseSharp Network with multitask learning outperforms the one without multitask learning (see "3D DenseNet" in Table 2) in terms of classification performance on almost all metrics. Besides, it was found that the DenseSharp Networks are faster to train. The DenseSharp Networks achieve the best performance after training about 60 epochs, whereas the 3D DenseNets need 100 epochs of training. See Supplementary "Training and inference time cost" for more details.

Though segmentation works just as an auxiliary task, the DenseSharp Networks are able to predict fairly good nodule masks (see Fig. 4). On the test set, the study achieved an average Dice coefficient of 74.12% between the manually labeled and the model predicted masks.

Discussion

Automatic tumor invasiveness prediction from the CT scans can provide important medical insights. In the article, this task is tackled using novel 3D convolutional neural networks based on DenseNet. The approach with efficient multitask learning demonstrates promising accuracy with respect to this task and achieves better classification performance than the radiologists in the observer study.

Continuing to learn and discover hierarchical features invisible to the human eye is one of deep learning system's strengths, facilitating better performance to differentiate subtypes of lung cancer. On the contrary, radiologists diagnose the lesions mainly based on typically visible radiographic features (size, lesion margin, solid component, etc.), which might be less sensitive to some local evidence when compared with machine learning models. Moreover, substantial overlaps among radiographic

Table 3. Confusion matrix of the 3-category classification on the test set

	3D DenseSharp network			Radiologists (mean ± STD)		
Ground Truth	**AAH-AIS**	**MIA**	**IA**	**AAH-AIS**	**MIA**	**IA**
AAH-AIS	17	21	0	22.00 ± 0.71	14.25 ± 0.43	1.75 ± 0.83
MIA	6	49	9	26.00 ± 3.08	32.00 ± 2.55	6.00 ± 1.41
IA	0	10	16	2.5 ± 1.12	8.25 ± 1.30	15.25 ± 1.09

NOTE: 15-run ensemble results are listed in "3D DenseSharp Network," and the observers' (Radiologists) are also reported as the mean and standard variance of the four radiologists' results.

Published OnlineFirst October 2, 2018; DOI: 10.1158/0008-5472.CAN-18-0696

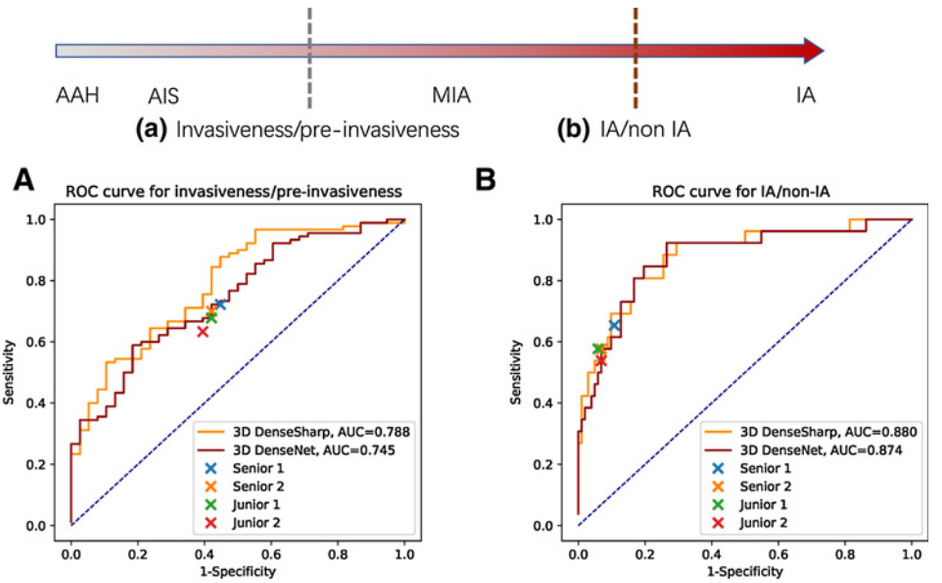


Figure 3.
ROC curves for two binary-classification subtasks. The performance of radiologists in the observer study are also depicted in the graph. **A,** Binary classification of invasive nodules (IA or MIA) and pre-invasive nodules (AIS or AAH). **B,** Binary classification of IA nodules and non-IA nodules (MIA, AIS, or AAH).

features of preinvasive lesions and invasive lesions make it very challenging for radiologists to correctly assess them. Experiences can help radiologists, as proved in this study (see Table 2), to improve the diagnostic accuracy on tumor invasiveness. However, the incremental progress is relatively limited, probably due to inadequate training for radiologists in subcentimeter GGNs interpretation. Therefore, when radiographic features suggesting malignancy are absent or not fully identified by a radiologist, an inappropriate diagnosis would appear, especially in early stage of lung cancers.

Although the proposed deep learning system shows some advantages over the radiologists on this problem, there are still a lot of limitations. Biased and insufficient data for training the neural networks could limit the performance. The model tends to predict AAH-AIS to be MIA, whereas the radiologists tend to label the MIA to be AAH-AIS (see Table 3), which may also indicate the data collection bias. Besides, the proposed deep learning system uses 32 mm $\times$ 32 mm $\times$ 32 mm patches of nodules to diagnose the tumor invasiveness, whereas ideally, radiologists can use the entire CT scan, together with other information (patient's age, smoking, medical history, etc.), to better estimate tumor invasiveness. Aggregating the global context of the lung and patient's information may boost the classification performance further.

To the best of our knowledge, the dataset is already the largest for this kind of research on automatic predicting tumor invasiveness for subcentimeter GGNs, it is, however, still insufficient. Pathological subtypes of lung cancer like mucinous AIS, non-lepidic predominant growth pattern lung adenocarcinomas (i.e., acinar, papillary, micropapillary, and/or solid; ref. 7), are rare and may not be fully learned by the deep neural networks. The training of deep neural networks should benefit from more data. Another limitation is the lack of external validation on an independent validation dataset from other institutions, regions, races. However, such a particular public dataset (subcentimeter lung adenocarcinoma, mainly presented as GGN, with new diagnosis standard of AAH, AIS, MIA, and IA) hardly exists to date. Transferring

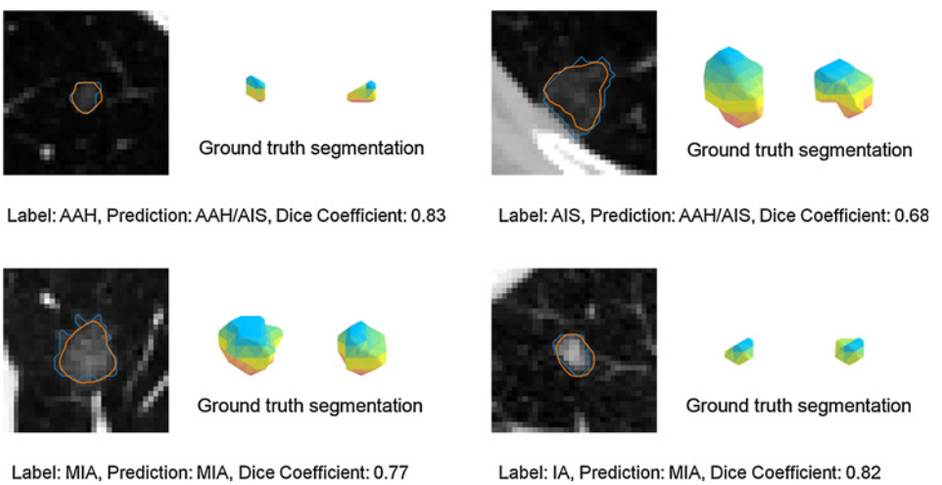


Figure 4.
Examples of the nodule segmentation predicted by the trained model. The blue contours show the manual segmentation, and the orange ones show the predicted segmentation at the center slice of the patches. The manually labeled masks (ground truth) and the predicted masks are also illustrated as the 3D contours. The color indicates the depth for each voxel in the coronal dimension. The well-trained neural network predicts the nodule mask and the invasiveness type in single forward computation.

Published OnlineFirst October 2, 2018; DOI: 10.1158/0008-5472.CAN-18-0696

neural network knowledge trained from larger databases for other related tasks (33), other than nodule segmentation and invasiveness classification, could also bring further improvements. Alternatively, pretraining on thousands of relatively cheap natural images is still worth more exploration. Inspired by recent advances in video analysis, it is feasible to convert 2D convolution kernels into 3D by "inflating" them (34). Plus, our 3D neural networks for medical image analysis may also benefit from large 3D neural networks pretrained on large-scale video dataset (35). We leave this as a future direction, which may boost the discriminative performance of our method further.

Another limitation for this study is the interpretability of the deep learning system. Though there have been great process on interpretability of machine learning system (36, 37)/deep learning system (38, 39), fully understanding the internal mechanism in deep neural networks is still a non-trivial task. Particularly, in biomedical analysis, we do want to understand how the imaging representations associate with specific molecular patterns, genotype (e.g., *EGFR*) and intratumoral microenvironment, which remains a rougher challenge at present. Such work, explaining the biological processes of the deep learning models was performed in our study, by investigating the association analysis between deep learned representations and *EGFR* mutations status. However, only 94 out of 651 nodules in this study were performed the *EGFR* mutation testing (see Supplementary Fig. S2), which are not enough for current deep learning methods to produce reasonable results. We will further address the interpretability for AI, especially in the medical context, by associating imaging information with genotypes and biomarkers, in a probabilistic deep learning framework. Moreover, combining deep learning with radiomics (33) may also help with robustness and interpretability.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: W. Zhao, J. Yang, Y. Hua, M. Li
Development of methodology: W. Zhao, J. Yang, Y. Sun, W. Wu, Z. Yang, B. Ni, Y. Hua, M. Li
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): W. Zhao, J. Yang, Y. Sun, L. Jin, Z. Yang, P. Gao, Y. Hua, M. Li
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): W. Zhao, J. Yang, Y. Sun, C. Li, W. Wu, L. Jin, Z. Yang, Y. Hua, M. Li
Writing, review, and/or revision of the manuscript: W. Zhao, J. Yang, Y. Sun, C. Li, W. Wu, L. Jin, Z. Yang, B. Ni, P. Wang, Y. Hua, M. Li
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): J. Yang, Z. Yang, P. Wang, Y. Hua, M. Li
Study supervision: W. Zhao, B. Ni, P. Wang, Y. Hua, M. Li
Other (algorithm and software development): J. Yang

Acknowledgments

We are grateful to Drs. Dexi Bi and Mengdi Xu for critically revising the manuscript. We thank Yuxiang Ye and Liang Ge in Diannei Technology for generous help in data and insightful discussion. We also express our sincere appreciation to all the kind and enthusiastic staffs from TCIA, NLST, and CDAS, who have helped us a lot to address the problem of public dataset validation. This study was supported by the Research Program of Shanghai Hospital Development Center SHDC22015025 (to M. Li), the National Key Research and Development Program of China 2017YFC0112800 (to P. Wang), 2017YFC0112905 (to M. Li), the National Science Foundation of China 61502301 (to B. Ni), and the Medical Imaging Key Program of Wise Information Technology of 120, Health Commission of Shanghai 2018ZHYL0103 (to M. Li). This study was supported by SJTU-UCLA Joint Center for Machine Perception and Inference (to B. Ni and J. Yang). The study was also partially supported by China's Thousand Youth Talents Plan, STCSM 17511105401, 18DZ2270700 (to B. Ni).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received March 5, 2018; revised July 3, 2018; accepted September 25, 2018; published first October 2, 2018.

References

1. Rami-Porta R, Bolejack V, Crowley J, Ball D, Kim J, Lyons G, et al. The IASLC lung cancer staging project: proposals for the revisions of the T descriptors in the forthcoming eighth edition of the TNM classification for lung cancer. J Thorac Oncol 2015;10:990–1003.
2. Kishi K, Homma S, Kurosaki A, Motoi N, Kohno T, Nakata K, et al. Small lung tumors with the size of 1 cm or less in diameter: clinical, radiological, and histopathological characteristics. Lung Cancer 2004;44:43–51.
3. Wood DE, Kazerooni EA, Baum SL, Eapen GA, Ettinger DS, Hou L, et al. Lung Cancer Screening, Version 3.2018, NCCN Clinical Practice Guidelines in Oncology. J Natl Compr Canc Netw 2018;16:412–41.
4. Lee SM, Park CM, Goo JM, Lee HJ, Wi JY, Kang CH. Invasive pulmonary adenocarcinomas versus preinvasive lesions appearing as ground-glass nodules: differentiation by using CT features. Radiology 2013;268:265–73.
5. Sakurai H, Nakagawa K, Watanabe S, Asamura H. Clinicopathologic features of resected subcentimeter lung cancer. Ann Thorac Surg 2015; 99:1731–8.
6. Wu F, Tian SP, Jin X, Jing R, Yang YQ, Jin M, et al. CT and histopathologic characteristics of lung adenocarcinoma with pure ground-glass nodules 10 mm or less in diameter. Eur Radiol 2017;27:4037–43.
7. Travis WD, Brambilla E, Noguchi M, Nicholson AG, Geisinger KR, Yatabe Y, et al. International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. J Thorac Oncol 2011;6:244–85.
8. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.
9. Jonathan L, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.
10. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 2017;39:1137–49.
11. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115–8.
12. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316: 2402–10.
13. Dou Q, Chen H, Jin Y, Lin H, Qin J. Automated Pulmonary Nodule Detection via 3D ConvNets with Online Sample Filtering and Hybrid-Loss Residual Learning. In:Intervention ICoMICaC-A, editor. Cham: Springer; 2017.
14. Jemal A, Miller KD, Ma J, Siegel RL, Fedewa SA, Islami F, et al. Higher lung cancer incidence in young women than young men in the United States. N Engl J Med 2018;378:1999–2009.
15. Huang G, Liu Z, Maaten LVD, Weinberger KQ. Densely connected convolutional networks. arXiv preprint arXiv 2016;1608: 06993.
16. Pinheiro PO, Collobert R, Dollár P. Learning to segment object candidates. Advances in Neural Information Processing Systems. 2015.

Published OnlineFirst October 2, 2018; DOI: 10.1158/0008-5472.CAN-18-0696

17. Pinheiro PO, Lin TY, Collobert R, Dollár P. Learning to refine object segments. European Conference on Computer Vision: Springer International Publishing; 2016. p. 75–91.
18. Sergey I, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift [abstract]. In: Proceedings of the International Conference on Machine Learning;2015.
19. Min L, Chen Q, Yan SC. Network in network. arXiv preprint arXiv 2013;1312:4400.
20. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv 2016:04467.
21. Chollet FK. GitHub repository. 2015.
22. Armato SG III, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys 2011;38:915–31.
23. Diederik K, Ba J. Adam: a method for stochastic optimization. arXiv preprint arXiv 2014;1412:6980.
24. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15:1929–58.
25. Contributors W. Matthews correlation coefficient. In Wikipedia, The Free Encyclopedia. Retrieved 09:56, June 26, 20182018;
26. Szegedy C. "Rethinking the inception architecture for computer vision" [abstract]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016: IEEE; 2016.
27. Deng J. Imagenet: A large-scale hierarchical image database" [abstract]. In: Proceedings of IEEE Conference onComputer Vision and Pattern Recognition, CVPR; 2009: IEEE; 2009.
28. Vazquez M, Carter D, Brambilla E, Gazdar A, Noguchi M, Travis WD, et al. Solitary and multiple resected adenocarcinomas after CT screening for lung cancer: histopathologic features and their prognostic implications. Lung Cancer 2009;64:148–54.
29. Borczuk AC, Qian F, Kazeros A, Eleazar J, Assaad A, Sonett JR, et al. Invasive size is an independent predictor of survival in pulmonary adenocarcinoma. Am J Surg Pathol 2009;33:462–9.
30. Yim J, Zhu LC, Chiriboga L, Watson HN, Goldberg JD, Moreira AL. Histologic features are important prognostic indicators in early stages lung adenocarcinomas. Mod Pathol 2007;20:233–41.
31. Son JY, Lee HY, Lee KS, Kim JH, Han J, Jeong JY, et al. Quantitative CT analysis of pulmonary ground-glass opacity nodules for the distinction of invasive adenocarcinoma from pre-invasive or minimally invasive adenocarcinoma. PLoS ONE 2014;9:e104066.
32. Lim HJ, Ahn S, Lee KS, Han J, Shim YM, Woo S, et al. Persistent pure ground-glass opacity lung nodules >/ = 10 mm in diameter at CT scan: histopathologic comparisons and prognostic implications. Chest 2013;144:1291–9.
33. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology 2016;278:563–77.
34. Joao C, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset [abstract]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017; 2017: IEEE; 2017.
35. Hara K, Hirokatsu K, Yutaka S. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet [abstract]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018; Salt Lake City, UT:IEEE;2018.
36. Ribeiro MT, Sameer S, Carlos G. "Why should i trust you?: Explaining the predictions of any classifier" [abstract]. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016: ACM; 2016.
37. Lundberg SM, L. S-I. "A unified approach to interpreting model predictions. "Advances in Neural Information Processing Systems. 2017.
38. Koh PW, Percy L. "Understanding black-box predictions via influence functions." arXiv preprint arXiv:1703.047302017.
39. Zhang QS, Song CZ. Visual interpretability for deep learning: a survey. Frontiers of Information Technology and Electronic Engineering 2018; 19:27–39.

Cancer Research
The Journal of Cancer Research (1916–1930) | The American Journal of Cancer (1931–1940)

AACR American Association for Cancer Research

3D Deep Learning from CT Scans Predicts Tumor Invasiveness of Subcentimeter Pulmonary Adenocarcinomas

Wei Zhao, Jiancheng Yang, Yingli Sun, et al.

Cancer Res 2018;78:6881-6889. Published OnlineFirst October 2, 2018.

Supplementary Methods

Training and inference time cost

The network was implemented using Python 3.6 based on TensorFlow 1.4.0 (1) and Keras 2.1.5 (2). The running time was averaged with 100 repeated experiments on a single Titan X GPU.

The training of the 3D DenseSharp Networks run at about 25.8s / epoch (523 samples), with a batch size of 24. For inference, inputs with size of $32 \times 32 \times 32$, run at about 86.7ms / batch (batch size=8) or 19.8ms / batch (batch size=1).

The training of the 3D DenseNets run at about 23.4s / epoch (523 samples), with a batch size of 24. For inference, input(s) with size of $32 \times 32 \times 32$, run at about 79.2ms / batch (batch size=8) or 16.3ms / batch (batch size=1).

Therefore, without considering the time cost of hyper-parameter tuning, a single run of the training for 3D DenseSharp Networks costs about 26min (for 60 epochs) for a good convergence. For a 15-run ensemble of 3D DenseSharp Networks, it costs 6.5h totally. As a comparison, a single run of the training for 3D DenseNets costs about 39min (for 100 epochs) for a good convergence. For a 15-run ensemble of 3D DenseNets, it costs 9.8h totally.

The details on the 2D CNNs processing the 2.5D CT images

The two 2D CNNs process 3-channel inputs of 2.5D CT images (on axial-coronal-sagittal views), see Supplementary Fig S1 for illustration. One can regard these 3 channels as the “RGB” channels for natural images.

The input size of the 2D DenseNets is also $32\text{mm} \times 32\text{mm} \times 32\text{mm}$, with the 2.5D representation, the input becomes $32 \times 32 \times 3$. The 2D DenseNets follow similar design pattern with our 3D DenseNets, using 2D convolutions instead of 3D convolutions; besides, we change the depth and filters of 2D DenseNets to keep the number of trainable parameters comparable with our 3D DenseNets. Specifically, as demonstrated in Supplementary Table S1, the 2D variants use growth rate $\mathrm{k} = 8$, bottleneck $\mathrm{B} = 4$, compression $\mathrm{C} = 2$ and structures $\mathrm{s} = [4,12,16,8]$.

The input size of the Inception-v3 Networks is $139\text{mm} \times 139\text{mm} \times 139\text{mm}$, with the 2.5D representation, the input becomes $139 \times 139 \times 3$, which is the minimal input size for Inception-v3 due to its network design. The voxel outside the raw CT is filled with -1. All the layers was initialized with the weight pretrained on the ImageNet database, except that the final layer of the pretrained Inception-v3 is replaced a "he-uniform" (3) initialized layer with 3-output softmax.

Hyper-parameter tuning

Deep learning is known to be requiring heavy tuning. There are algorithms on automatically tuning the hyper-parameters, such as Bayesian Optimization (4); however, they could still be too costly for deep learning models. Hyper-parameters for deep learning are not equivalently important in the tuning process (5, 6), which is the first principle for guiding our parameter tuning procedure. We describe the practice on tuning the hyper-parameters in our experiments.

Firstly, the most important "hyper-parameter" is the neural network architecture.

There are numbers of popular choices, e.g. VGG Networks (7), Inception Networks (8), ResNets (9) and their variants. We chose densely connected networks, DenseNets, (10), as they are shown to be parameter-efficient. Neural architecture search (11) is too computationally intensive, which is out of the scope of our study.

Secondly, we controlled the input size and DenseNet-specific hyper-parameters (growth rate, bottleneck, compression and block structures). The candidate input sizes were 16mm, 24mm, 32mm and 48mm. We fixed bottleneck $B = 4$ and compression $C = 2$ following the original paper. Then, we fixed growth rate $k = 16$, and found the largest depth (block structure) for each candidate input size, to fit the maximum memory of a single Titan X GPU (12GB), with a batch size of 24. We randomly selected 1/4 samples in the *training and validation* dataset as **development set (rather than cross validation)**, and the remaining 3/4 as training set. In this way, we selected 32mm as the best input size. Next, we varied growth rate $k \in \{8, 12, 16, 32\}$ with the corresponding maximum network depths, and determined the best growth rate $k = 16$ and structures $s = [4,4,4]$.

Thirdly, we controlled the balance of the classification and segmentation task (parameter λ). As segmentation worked as an auxiliary task, we set $\lambda \in \{0.05, 0.1, 0.2, 0.5, 1, 2\}$, and observed the classification performance on the development set. We chose λ=0.2, while there is no significant difference for choosing λ=0.1 or λ=0.5.

Finally, we tuned the optimization process. Adam (12) is one of the common choices for training deep CNNs. Thanks to the adaptive learning rate (with momentum),

Adam optimizer usually leads to fast and stable convergence. Learning rate and training epoch (to determine early stopping) are the most important hyper-parameters during the optimization process, which we tuned using **4-fold cross validation** on the *training and validation* dataset. In the end, we chose a fix learning rate scheme with $\text{learning rate} = 10^{-4}$, and early stopped the optimization process after $\text{epoch} = 60$.

References

1. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv. 2016:04467.
2. Chollet F. Keras. GitHub repository. 2015.
3. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Proceedings of the IEEE international conference on computer vision. 2015.
4. Snoek J, Hugo L, Ryan PA. Practical bayesian optimization of machine learning algorithms Advances in neural information processing systems. 2012.
5. Bergstra J, Yoshua B. Random search for hyper-parameter optimization. Journal of Machine Learning Research. 2012:281-305.
6. Greff K. LSTM: A search space odyssey. IEEE transactions on neural networks and learning systems 2810 2017:2222-32.
7. Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014)

8. Szegedy C. "Rethinking the inception architecture for computer vision." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition2016.

9. He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016

10. Huang G, Liu Z, Maaten LVD, Weinberger KQ. Densely connected convolutional networks. arXiv preprint arXiv. 2016;1608:06993

11. Zoph, Barret, and Quoc V. Le. "Neural architecture search with reinforcement learning." arXiv preprint arXiv:1611.01578(2016).

12. Diederik K, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv. 2014;1412:6980.

Table S1. The detailed architecture of 3D DenseSharp Networks, with growth rate $k = 16$, bottleneck $B = 4$, compression $C = 2$ and structures $s = [4,4,4]$.

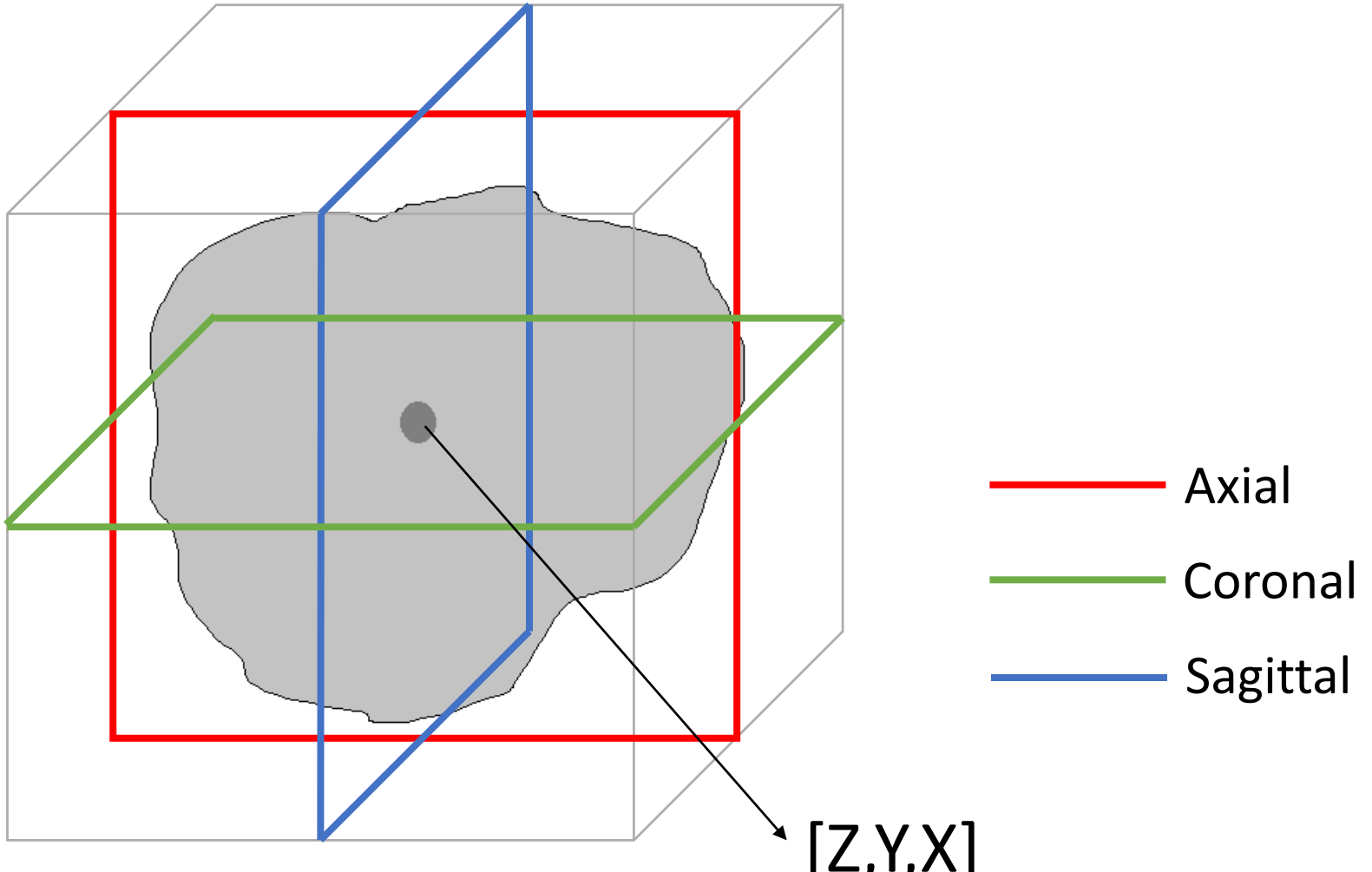
(input) – Layer – (output)	Tensor Size	Building Blocks
Input – (a)	$32 \times 32 \times 32 \times 1$	
(a) – Convolution – (b)	$32 \times 32 \times 32 \times 32$	$3 \times 3 \times 3$ conv
(b) – Dense Block 1 – (c)	$32 \times 32 \times 32 \times 96$	$\begin{bmatrix} \text{bn} - \text{relu} - 1 \times 1 \times 1 \text{ conv} \\ \text{bn} - \text{relu} - 3 \times 3 \times 3 \text{ conv} \end{bmatrix} \times 4$
(c) – Compression 1 – (d)	$16 \times 16 \times 16 \times 48$	$\begin{bmatrix} \text{bn} - \text{relu} - 1 \times 1 \times 1 \text{ conv} \\ 2 \times 2 \times 2 \text{ average pool} \end{bmatrix}$
(d) – Dense Block 2 – (e)	$16 \times 16 \times 16 \times 112$	$\begin{bmatrix} \text{bn} - \text{relu} - 1 \times 1 \times 1 \text{ conv} \\ \text{bn} - \text{relu} - 3 \times 3 \times 3 \text{ conv} \end{bmatrix} \times 4$
(e) – Compression 2 – (f)	$8 \times 8 \times 8 \times 56$	$\begin{bmatrix} \text{bn} - \text{relu} - 1 \times 1 \times 1 \text{ conv} \\ 2 \times 2 \times 2 \text{ average pool} \end{bmatrix}$
(f) – Dense Block 3 – (g)	$8 \times 8 \times 8 \times 120$	$\begin{bmatrix} \text{bn} - \text{relu} - 1 \times 1 \times 1 \text{ conv} \\ \text{bn} - \text{relu} - 3 \times 3 \times 3 \text{ conv} \end{bmatrix} \times 4$
(g) – Global Pooling – (h)	120	$8 \times 8 \times 8$ average pool
(h) – **Classification Output**	3	softmax
(g) – Up-sampling – (i)	$16 \times 16 \times 16 \times 112$	$2 \times 2 \times 2$ transpose conv
(i), (e) – Add – (j)	$16 \times 16 \times 16 \times 112$	add
(j) – Up-sampling – (k)	$32 \times 32 \times 32 \times 96$	$2 \times 2 \times 2$ transpose conv
(k), (c) – Add – (l)	$32 \times 32 \times 32 \times 96$	add
(l) – Convolution – (m)	$32 \times 32 \times 32 \times 1$	$1 \times 1 \times 1$ conv
(m) – **Segmentation Output**	$32 \times 32 \times 32 \times 1$	sigmoid

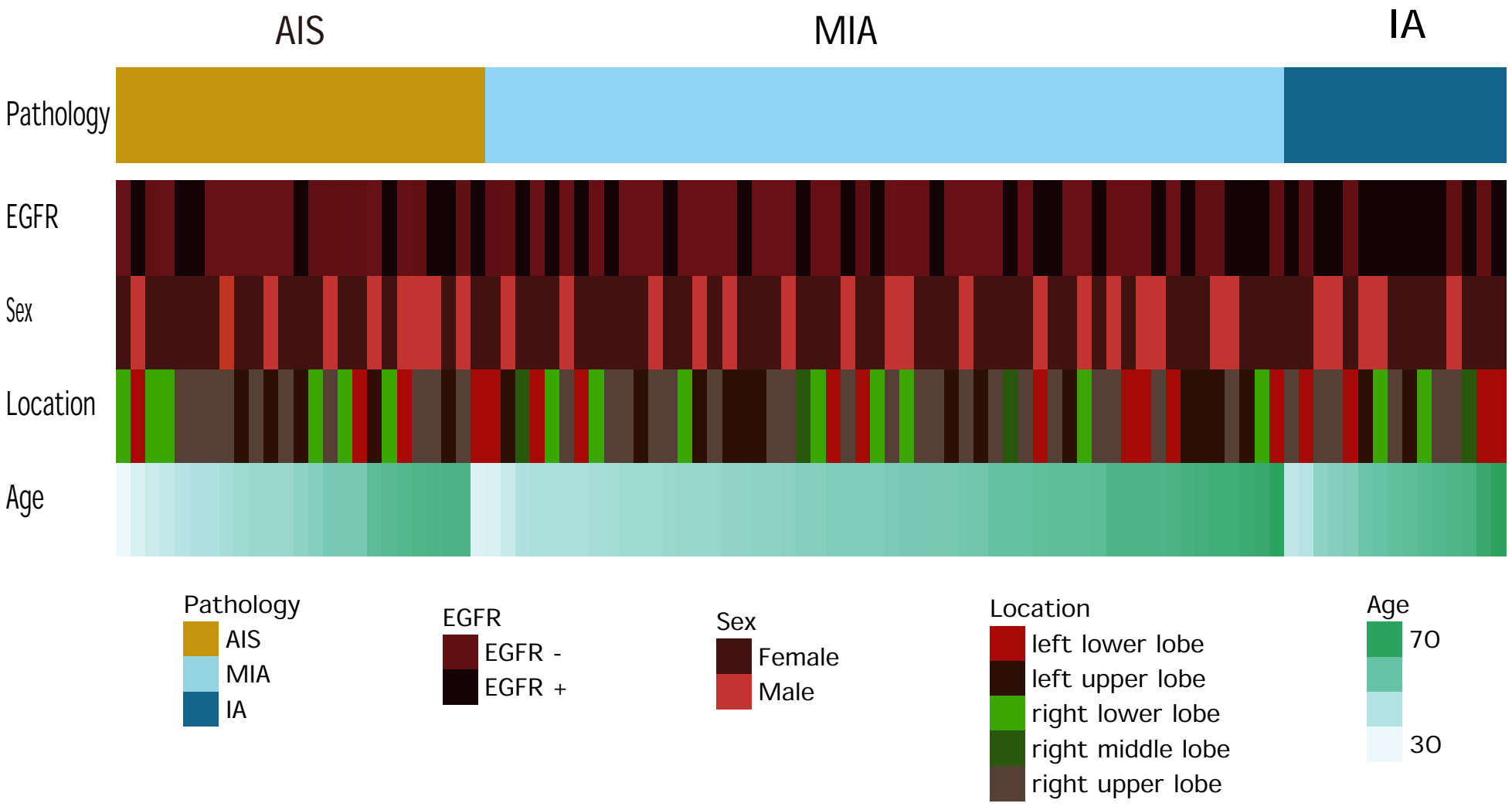
Table S2. Binary classification performance for differentiating invasive / pre-invasive nodules, in terms of accuracy, weighted average F1-score, MCC and AUC**.** "3D DenseSharp Network" denotes the results of our proposed network, and "3D DenseNet" denotes the performance without multi-task learning. Results for four observers (2 senior and 2 junior radiologists) are also reported. The higher is better.

	Accuracy	$F1_{AVG}$	MCC	AUC
3D DenseSharp Network	**86.7%**	85.9%	0.535	**0.880**
3D DenseNet	84.4%	84.1%	**0.578**	0.874
Senior 1	84.4%	64.6%	0.531	-
Senior 2	85.9%	85.5%	0.542	-
Junior 1	**86.7%**	**86.2%**	0.563	-
Junior 2	85.2%	84.5%	0.510	-

Table S3. Binary classification performance for differentiating IA / non-IA nodules, in terms of accuracy, weighted average F1-score, MCC and AUC**.** ”3D DenseSharp Network” denotes the results of our proposed network, and “3D DenseNet” denotes the performance without multi-task learning. Results for four observers (2 senior and 2 junior radiologists) are also reported. The higher is better.

	Accuracy	$F1_{AVG}$	MCC	AUC
3D DenseSharp Network	**80.5%**	**78.5%**	**0.453**	**0.788**
3D DenseNet	75.0%	70.7%	0.301	0.745
Senior 1	71.2%	68.0%	0.262	-
Senior 2	66.4%	67.4%	0.262	-
Junior 1	64.8%	66.0%	0.239	-
Junior 2	62.5%	64.0%	0.220	-


Axial
Coronal
Sagittal
[Z,Y,X]


AIS
MIA
IA
Pathology
EGFR
Sex
Location
Age
Pathology
AIS
MIA
IA
EGFR
EGFR -
EGFR +
Sex
Female
Male
Location
left lower lobe
left upper lobe
right lower lobe
right middle lobe
right upper lobe
Age
70
30

Supplementary Figure Legends

Figure S1. Illustration of the 2.5D representation of 3D voxels. The three views (axial, coronal and sagittal) are concatenated as 3 channels for inputs.

Figure S2. The distribution of clinical characteristics and EGFR mutation status of 94 patients from 651 patients included in our study. The distribution of EGFR mutation status, sex, location of the lesions and age of the 94 patients in three pathological labels is depicted. There is no AAH in these 94 cases. Note that EGFR mutation testing is not included in the initial work-up of patients with AAH, AIS, and even for patients with MIA and IA in our study due to the early stage of these lesions (T1a(mi) or T1a). Moreover, the lesions in this data set are subcentimeter in size, which determines that no enough residual specimen may be available for further EGFR mutation testing after histologic diagnosis. Only one malignant nodule was studied for each patient due to the availability of EGFR testing report. Finally, 94 nodules in our studied dataset were performed the EGFR mutation testing.