Hybrid U-Net based deep-learning model for volume segmentation of lung nodules in CT images

Yifan Wang, Chuan Zhou*, Heang-Ping Chan, Lubomir M. Hadjiiski, Aamer Chughtai, Ella A. Kazerooni

(Department of Radiology, The University of Michigan, Ann Arbor, MI 48109-0904)

* Corresponding author:

Chuan Zhou chuan@umich.edu 734-647-8554

Author Manuscript

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/mp.15810.

This article is protected by copyright. All rights reserved.

Abstract

Objectives: Accurate segmentation of the lung nodule in CT images is a critical component of a computer-assisted lung cancer detection/diagnosis system. However, lung nodule segmentation is a challenging task due to the heterogeneity of nodules. This study is to develop a hybrid deep learning (H-DL) model for the segmentation of lung nodules with a wide variety of sizes, shapes, margins, and opacities.

Materials and methods: A data set collected from LIDC-IDRI containing 847 cases with lung nodules manually annotated by at least two radiologists with nodule diameter greater than 7 mm and less than 45 mm was randomly split into 683 training/validation and 164 independent test cases. The 50% consensus consolidation of radiologists' annotation was used as the reference standard for each nodule. We designed a new H-DL model combining two deep convolutional neural networks (DCNN) with different structures as the encoders to increase the learning capabilities for segmentation of complex lung nodules. Leveraging the basic symmetric U-shape architecture of a U-Net, we redesigned two new U-shape deep learning (U-DL) models that were expanded to 6 levels of convolutional layers. One U-DL model used a shallow DCNN structure containing 16 convolutional layers adapted from the VGG-19 as the encoder and the other used a deep DCNN structure containing 200 layers adapted from DenseNet-201 as the encoder, while the same decoder with only one convolutional layer at each level was used in both U-DL models, and we referred to them as the Shallow and Deep U-DL model, respectively. Finally, an ensemble layer was used to combine the two U-DL models into the H-DL model. We compared the effectiveness of the H-DL, the Shallow U-DL and the Deep U-DL models by deploying them separately to the test

This article is protected by copyright. All rights reserved.

Author Manuscript

set. The accuracy of volume segmentation for each nodule was evaluated by the 3D DICE coefficient and Jaccard index (JI) relative to the reference standard. For comparison, we calculated the median and minimum of the 3D DICE and JI over the individual radiologists who segmented each nodule, referred to as M-DICE, min-DICE, M-JI, and min-JI, respectively.

Results: For the 164 test cases with 327 nodules, our H-DL model achieved an average 3D DICE coefficient of 0.750±0.135 and the average JI of 0.617±0.159. The radiologists' average M-DICE was 0.778±0.102 and the average M-JI was 0.651±0.127, both were significantly higher than those achieved by the H-DL model ($P < 0.05$). The radiologists' average min-DICE (0.685±0.139) and the average min-JI (0.537±0.153) were significantly lower than those achieved by the H-DL model ($P<0.05$). The results indicated the H-DL model approached the average performance of radiologists and was superior to the radiologist whose manual segmentation had the min-DICE and min-JI. Moreover, the average DICE and average JI achieved by the H-DL model were significantly higher than those by the individual Shallow U-DL model (DICE of 0.745 ± 0.139, JI of 0.611 ± 0.161) ($P < 0.05$) or the individual Deep U-DL model alone (DICE of 0.739±0.145, JI of 0.604±0.163) ($P < 0.05$).

Conclusion: Our newly developed H-DL model outperformed the individual shallow or deep U-DL models. The hybrid deep learning method combining multi-level features learned by both the shallow and deep DCNNs could achieve segmentation accuracy comparable to radiologists' segmentation for nodules with wide ranges of image characteristics.

Author Manuscript

This article is protected by copyright. All rights reserved.

Keywords: Computer-aided diagnosis, lung nodule, Deep learning, nodule segmentation

Author Manuscript

INTRODUCTION

Lung cancer is one of the most common cancers and the leading cause of cancer-related death in men and women in the United States. According to the American Cancer Society, about 13% of all new cancers are lung cancers, with about 235,760 new cases (119,100 in men and 116,660 in women) and about 131,880 deaths from lung cancer (69,410 in men and 62,470 in women) in 2021.[1] The overall prognosis of lung cancer is poor, with the 5-year survival rate of only 21%.

Computed tomography (CT) has become a preferred method for detecting and diagnosing lung cancer. Accurate segmentation of the lung nodule in CT images not only provides an objective measurement of nodule size for clinical surveillance of nodule growth[2] but also constitutes a critical component for the development of a computer-assisted lung cancer detection/diagnosis system.

Despite the development of computerized methods over the years, lung nodule segmentation remains a difficult task because of the wide range of heterogeneity in lung nodule characteristics such as shape, size, and attenuation. The complexity of lung parenchyma surrounding the nodules further poses a challenge in developing robust

This article is protected by copyright. All rights reserved.

segmentation models[3]. As the examples shown in Fig.1, it is challenging to segment the nodules with heterogeneous intensity distribution characterized by a wide range of varied x-ray attenuation distributed within the nodules containing solid, sub-solid, non-solid ground glass opacity (GGO) or mixed components, the nodules with "irregular shapes" categorized as irregular or spiculated margins, and the juxtapleural or juxtavascular nodules attached to the chest wall, pleural surface or pulmonary vessels.

Conventional methods for automated lung nodule segmentation in CT images[4,5] commonly consist of two steps: the detection of nodule locations and then the segmentation of the detected nodules from the surrounding lung parenchyma[6]. The features characterizing nodule intensity, textures, and morphologies are usually extracted to differentiate nodules from other lung structures during nodule detection. Then these features are used to segment the nodules by various methods such as intensity-based methods with morphological operations[7,8], region growing methods[9], optimization methods with level set[10], graph cut[11], or reinforcement-learning techniques[12]. In an early study[13], we used 3D active contours guided by gradient and curvature energies for segmentation and extracted morphological and texture features to classify malignant and benign lung nodules. In our recent study[14], we developed a 3D adaptive multi-component EM analysis (3D-AMEA) method to segment the nodule volume including the solid and non-solid GGO components and the surrounding lung parenchyma region. Radiomic features were then extracted to characterize the CT attenuation distribution patterns of the nodule components. Our results demonstrated the feasibility of classifying pathologic invasive nodules, pre-invasive nodules, or benign nodules using the proposed method. Although a wide variety of methods have

This article is protected by copyright. All rights reserved.

Author Manuscript

been developed, the accuracy and robustness of the segmentation have yet to be further improved, especially for nodules with irregular shapes and heterogeneous intensity distribution within the nodules (e.g., partially solid and non-solid GGO nodules)[8].

Supervised deep learning methods are emerging technologies increasingly used in medical image analysis, shifting from the classical methods trained with handcrafted features to the training of deep learning models in which the features are learned automatically without manual extraction and selection. Deep convolutional neural network (DCNN) based deep learning methods have been used for learning discriminative features from the training data in various machine learning applications from image analysis to natural language processing. The DCNN models, such as VGG[15,16], DenseNet[17], Fast-CNN[18], and some much deeper CNNs have been successfully employed for a wide variety of tasks. The Mask R-CNN[19] represents one of the state-of-the-art DCNNs that use a Region Proposal Network (RPN) followed by a Region-based CNN and a semantic segmentation model to simultaneously perform the tasks of detection and segmentation. Different network structures have been developed specifically for the many types of lesions or organs to be segmented in various medical imaging modalities. The U-Net[20] model supplements the deeply supervised encoder sub-network by a decoder sub-network through simple skip connections that allow the network to propagate context information to higher resolution layers. The iW-Net[21] was composed of two U-Nets; the first performed automatic segmentation, while the second U-Net allowed user correction by marking 2 points along the nodule boundary to refine the segmentation result of the first U-Net. The PN-SAMP[22] first segmented a nodule using U-Net

This article is protected by copyright. All rights reserved.

Author Manuscript

and then used the feature maps from the encoder and the segmentation output from the decoder as inputs to a CNN to predict the malignancy of the nodule. However, these previous methods relied on user interaction to refine the segmentation such as iW-Net or were trained for specific types of nodules such as PN-SAMP. Continued effort is needed to develop new architectures to take advantage of the DCNN-based approach for the segmentation of heterogeneous lung nodule volumes.

The key feature of the current U-Net architecture and its variants for image segmentation is the use of two DCNN networks with a similar structure as the encoder and decoder for pixel-wise image segmentation. However, there are still many significant challenges in advancing U-Net segmentation approaches, including improving the learning capability of the encoder to discover enough useful hidden image patterns with large variations[23,24] so as to characterize the differences between lung cancer and lung parenchyma, and the better understanding of the intricate relationships between a large number of interdependent variables[25], especially for segmentation of complex objects in medical images, such as the highly heterogeneous lung nodules. Increasing the depth of a DCNN to generate deeper and diverse representations enables the network to progressively explore different levels of features with different sizes of the receptive field as it sequentially goes through each layer is a popular method in the previous years. However, an excessively deep network can result in saturation and cause degradation of performance[17], and a relatively small training set such as our lung cancer dataset is more prone to such risks.

Author Manuscript

This article is protected by copyright. All rights reserved.

Another approach to increasing the learning ability of a DCNN network is to combine multiple convolutional sequences that may better characterize complex image patterns. The convolutional sequences have different structures, depths, and receptive fields that may allow them to independently capture different features and focus on different kinds and levels of patterns. When these convolutional sequences were hybridized together, they can learn more complex patterns or capture a larger combination of patterns.

Following the second approach, in this study, we designed a new hybrid deep learning (H-DL) model for the segmentation of lung nodules of a wide range of heterogeneous characteristics. Compared with the conventional U-Net related methods, our H-DL model was an ensemble of two U-shaped architectures (U-DL), one had a shallow DCNN as encoder and the other had a deep DCNN as encoder with a different network structure, to increase the learning capability so that different levels of features can be explored to characterize the highly heterogeneous lung nodules in CT images. Moreover, unlike the simple long skip connections utilized in the conventional U-Net, we adapted a series of nested and dense skip structures[25] to provide alternative pathways to connect the encoder and the decoder in each U-DL network. These skip structures further alleviate the vanishing gradient problem that saturates gradient backpropagation in deeper networks. After a large number of patterns were captured independently by these two networks, an ensemble layer was used to hybridize the two different U-DL networks to the H-DL model to get the larger combinations of patterns. To increase the efficiency of our H-DL model, we used an asymmetric encoder and decoder path structure in which the same decoder with only one

This article is protected by copyright. All rights reserved.

Author Manuscript

convolutional layer at each level was used for both the shallow and deep U-DL models. The simplified decoder could not only reduce computational and memory costs but also provide flexibility to further expand the H-DL model by adding more DCNNs with different structures as encoders if needed.

To evaluate the effectiveness of our H-DL model in segmenting heterogeneous lung nodules, we deployed the H-DL and the individual U-DL models separately to an independent test set with a wide range of characteristics manifested in CT images and demonstrated the increased learning capability resulting from the ensemble of two different feature extraction (encoding) networks. We also compared with other methods reported in the literature for lung nodule segmentation. To evaluate the generalizability and robustness of our H-DL model, we deployed our H-DL model with the trained weights frozen to the test set of an offline challenge, the Sub-Challenge B (Nodule Segmentation) of the LNDb challenge[26], and demonstrated that our H-DL model can be directly deployed to an "external" data set and achieve high accuracies in lung nodule segmentation.

This paper is organized as follows. Section 2 introduces our H-DL model and the dataset used. Section 3 presents the results of our method. Section 4 provides the discussion and Section 5 concludes the paper.

Author Manuscript

This article is protected by copyright. All rights reserved.

Materials and methods

Dataset

From 1,010 patient cases publicly available through the Lung Image Database Consortium image collection (LIDC-IDRI) with manually annotated lung nodules[27], a data set of 847 cases containing the lung nodules marked by at least two radiologists with nodule diameter greater than or equal to 7 mm and less than 45 mm were selected in our study. The CT images were acquired with different CT scanners manufactured by GE, Philips, Siemens, and Toshiba. The tube current ranged from 40 to 627 mA (mean: 221.1 mA), the tube peak potential energies ranged from 120 kV to 140 kV, the slice thickness ranged from 0.6 mm to 5.0 mm with reconstruction interval from 0.45 mm to 5.0 mm, and the pixel size in the axial plane varied from 0.46 mm to 0.98 mm. We used 683 cases with 2,558 nodules for training and validation and the remaining 164 cases with 327 nodules for independent testing.

In the LIDC dataset, the lung nodules were marked and manually segmented by at least two radiologists. The LIDC radiologists also subjectively assessed the nodule characteristics by descriptors and providing ratings on a scale from 1 to 5 for each marked nodule including subtlety, spiculated margin, solid opacity, lobulated shape, and the likelihood of malignancy (e.g., nodule subtlety, 1 = extremely subtle, 5 = obvious). The typically used 50% consensus consolidation27 of radiologists' annotations for each nodule was calculated by the LIDC

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/mp.15810.

This article is protected by copyright. All rights reserved.

suggested python package "pylidc" and used as the reference standard for training and testing. That is, a voxel was labeled as within the nodule when at least 50% of radiologists' segmentation included that voxel. Fig. 2 shows examples of radiologists' variabilities in manual outlining of nodules of various sizes, shapes, and locations. The dashed contours with different colors (left) are the outlines by four different radiologists, and the red contour enclosed nodule (right) is the 50% consensus consolidation of radiologists' annotations for that nodule.

In general, lung nodules with smooth shapes and obvious margins resulted in more similar outlines by the radiologists as the boundaries were clearly recognizable. In contrast, lung nodules with irregular shapes and fuzzy margins caused higher variabilities because their boundaries were difficult to be clearly identified.

U-Net based deep learning (U-DL) model for Lung nodule segmentation

The U-Net neural network architecture[20] was initially developed for biomedical image segmentation and has been widely used in medical image analysis. Despite the outstanding overall performance, some studies suggested that the conventional U-Net architecture still has room for improvement due to its simple series of convolutional layers[28,29] , plain probable sematic gap[25,29], and relatively shallow network structures[30]. Based on these previous observations we made the following targeted innovation. Leveraging the basic U-shaped architecture of the encoder and decoder paths in a conventional U-Net, we redesigned the DCNN architectures of the encoder and decoder in two separate U-shape

This article is protected by copyright. All rights reserved.

Author Manuscript

networks and combined them to a hybrid deep learning (H-DL) model to improve the learning capabilities for segmentation of complex lung nodules that have large variations in the structural and attenuation characteristics. Each of the U-DL networks consisted of a contracting (encoder) and an expanding (decoder) path (Fig. 3a,b). Compare with the original U-Net that contained five levels of the same number of 9 convolutional layers, we expanded each of the U-DL networks to six levels in both paths. In one U-DL network, we used a relatively shallow DCNN structure containing 16 convolutional layers adapted from VGG-19 (the 3 fully connected layers in VGG-19 was not used). In the other U-DL network, we used a deep DCNN structure containing 200 layers adapted from DenseNet-201 (196 convolutional layers and 4 transitional layers in 5 dense blocks, while dropping the one fully connected layer in DenseNet-201). We referred to our two U-shaped DCNN backbone networks as Shallow U-DL (Fig. 3a) and Deep U-DL (Fig. 3b) model, respectively.

We also redesigned the decoder in the expanding path of the conventional U-Net. In our U-DL models, the expanding path consisted of only one convolutional layer at each level that was much fewer than those in the contracting path and was not symmetric to the contracting path. The simplified decoder can not only reduce computational and memory costs but also provide flexibility that allows further expanding the H-DL model by adding more DCNNs as encoders. Both the encoding and decoding paths used 3 x 3 padded convolution followed by a rectified linear unit (ReLU). A 2 x 2 max pooling operation with stride two was used for feature map downsampling in the contracting path. We experimentally chose the transpose convolution (inverse convolution) operations for the

This article is protected by copyright. All rights reserved.

Author Manuscript

upsampling operation in the expanding path by comparing it to the interpolation (nearest neighbor or bilinear).

In a conventional U-Net model, the supervised encoder and decoder sub-networks were connected through simple long skip connections, allowing the network to propagate contextual information to higher resolution layers. As we utilized deeper encoder network structures in our U-DL models, to further alleviate the vanishing gradient problem that saturates gradient backpropagation in deeper networks, we modified our U-DL models by adding a series of nested and dense skip structures to provide alternative pathways to connect the encoder and the decoder, as shown in Fig. 3. This skipping scheme was derived and modified from the U-Net++[25], a U-Net variant, that has the advantage of reducing the semantic gap between the feature maps of the encoder and the decoder.

In our hybrid model, the two U-DL models were separately trained with the training set and then the probabilities predicted by the two U-DL models were combined into the H-DL model through an ensemble layer to maximize the probabilities of the pixels belonging to a nodule as shown in Fig. 4. The ensemble layer concatenated the vectors output from the two U-DL models, followed by a 3 x 3 convolutional layer with the sigmoid activation function and output the multi-component vectors to a likelihood map indicating pixelwise the chance that a pixel being inside the lung nodule. A threshold of 0.5 likelihood value was

Author Manuscript

This article is protected by copyright. All rights reserved.

85 determined during the training process to segment the likelihood map to a binary image that labeled the interior and exterior regions of the nodule.

Data preparation

The Hounsfield Unit (HU) is commonly used in CT scans that measure the radiodensity to characterize the tissue property. As the 12- or 16-bit CT data read directly from the LIDC DICOM file is not the HU value, we first converted the data read from the DICOM file to the HU values by multiplying the pixel values with the rescaling slope and adding the intercept which is stored in the metadata of the DICOM header. The CT scans were originally acquired with a slice interval ranging from 0.45 mm to 5.0 mm and a pixel size in the axial plane varying from 0.46 mm to 0.98 mm. We resampled all CT scans to isotropic volumes with a voxel size of 0.5 mm x 0.5 mm x 0.5 mm using the 3D spline interpolation method. For each reference standard nodule marked by radiologists, a volume of interest (VOI) of 64 x 64 x 64 pixels in size centered at the center of the nodule was cropped. For each VOI, the voxel values were scaled as follow:

$$f'(x,y,z) = \frac{f(x,y,z) - Min}{Max - Min} \#(1.)$$

where $f(x,y,z)$ was the voxel value at (x,y,z) *Min* and *Max* were the minimum and maximum voxel values within the VOI, respectively.

Author Manuscript

This article is protected by copyright. All rights reserved.

Training of H-DL models

Our H-DL models were trained with the set of 683 LIDC cases containing 2,558 nodules marked by at least two radiologists. This data set was separated randomly by case with a ratio of 9:1 as the training set and the validation set during the training process. For each 64 x 64 x 64-pixel VOI, three 64 x 64-pixel 2D patches in the axial plane, with the central patch centered at the nodule center, were sampled with a 1.5 mm interval and treated as three different training samples. With an image patch as input, the H-DL model output the likelihood map, which indicates pixels the chance that a pixel being inside the nodule, in the same size as the input image patch (64 x 64 pixels).

Our H-DL model was trained with a mini-batch training for stochastic optimization and the Adam optimizer[31]. The DICE coefficient (DSC) combined with the binary cross-entropy was used as the loss function during training. A mini-batch size of 64 randomly divided from the training set was used in each training epoch. A normal distribution with a mean of 0 and a standard deviation of 0.02 was used to initialize the networks' weights. The learning rate was initially set to 0.001 as a compromised balance of slow progress (with lower learning rate) and undesirable divergences (with a larger learning rate) in the loss function and decreased by ten times when the loss did not continuously decrease on the validation set after ten consecutive epochs. The early stop strategy was used when the loss on the validation set did not decrease over 30 consecutive epochs.

This article is protected by copyright. All rights reserved.

Performance evaluation

The performances of our trained models in lung nodule segmentation were evaluated by comparing the segmentation results to the reference standard, defined as the 50% consensus consolidation of radiologists' annotations. Different from using only three image patches sampled from each VOI to train the models, the trained model was deployed to the entire 64 x 64 x 64 VOI slice by slice. For performance evaluation, the 3D Dice similarity coefficient (DICE) and Jaccard index (JI) applied to the 64 slices in each VOI were calculated as quantitative performance measures:

$$DICE = \frac{2(Obj \cap Ref)}{Obj + Ref} \#(2.)$$

$$JI = \frac{Obj \cap Ref}{Obj \cup Ref} \#(3)$$

where *Obj* was the segmented volume, *Ref* was the reference standard.

For comparison, we also calculated the 3D DICE coefficient and JI for each LIDC radiologist relative to the reference standard. Since the nodules were segmented by a different number of radiologists (N >= 2) in the LIDC data set, we calculated the median (M) and minimum (min) of the 3D DICE and JI over all radiologists who segmented a given nodule, referred to as M-DICE, min-DICE, M-JI, and min-JI, respectively. The averages of the above quantitative measures over the entire test set of nodules were compared with the average 3D DICE coefficient and JI of the two U-DL models and the combined H-DL model. The two-tailed paired t-test was used to compare the differences between our models and the radiologists' manual segmentations.

This article is protected by copyright. All rights reserved.

RESULTS

LIDC independent test set

Table 1 summarizes the results of our H-DL model, Shallow U-DL and Deep U-DL models, and radiologists' performance. Fig. 5 shows the box and whisker plots of the distributions of the 3D DICE coefficients for the nodule segmentation.

For 164 test cases with 327 nodules, our H-DL model achieved an average 3D DICE coefficient of 0.750±0.135 and an average JI of 0.617±0.159. The radiologists' average M-DICE was 0.778±0.102 and the average M-JI was 0.651±0.127; both were significantly higher than those achieved by the H-DL model ($P < 0.05$). On the other hand, both the average min-DICE (0.685±0.139) and the average min-JI (0.537±0.153) were significantly lower than the corresponding average DICE and average JI achieved by the H-DL model. The results indicated that the automated segmentation by the H-DL model approached the average performance of radiologists and was superior to the radiologist whose manual segmentation had the minimum DICE and JI among the radiologists in the group outlining the same nodule. Note that, as the task of nodule marking was randomly assigned to different radiologists for each nodule in the LIDC study, the minimum DICE and JI could come from any radiologists.

To assess the effectiveness of combining the Shallow U-DL and Deep U-DL models into the H-DL model, we deployed the separately trained Shallow U-DL and Deep U-DL models to the test set. The Shallow U-DL model achieved an average DICE of 0.745±0.139 and an average JI of 0.611±0.161. The corresponding average DICE and average JI achieved by the Deep U-DL model were 0.739±0.145 and 0.604±0.163, respectively. The average DICE and average JI achieved by our H-DL model were

This article is protected by copyright. All rights reserved.

significantly higher than those of the Deep U-DL and Shallow U-DL model ($P < 0.05$). Fig. 6 shows examples of nodules segmented by the shallow U-DL, the deep U-DL, and the final H-DL, in comparison to the radiologists' segmentation.

For the nodules with different radiologic characterizations assessed by LIDC radiologists, we used the median of radiologists' ratings as the final rating for each nodule and then separated the nodules into two groups using the median of the nodules' final ratings for each descriptor. Table I shows the analysis for the two groups of each descriptor. For example, a nodule with diameters < 10.02 mm was considered as a small nodule, the malignancy rating ≥ 3.0 and the margin ≥ 4.25 indicated the nodules had a higher likelihood of malignancy and sharp margins, respectively. For the nodules with diameters ≥ 10.02 mm or nodule subtlety rating ≥ 4.25, the average DICE coefficients achieved by the H-DL model were 0.782±0.128 and 0.798±0.113, respectively, compared to the radiologists' average M-DICE of 0.790±0.105 and 0.785±0.116. Their differences did not achieve statistical significance ($P > 0.05$), indicating the segmentation accuracies achieved by H-DL were comparable to those of radiologists for relatively large and obvious nodules. For the nodules with other radiologic characterizations described by LIDC, the H-DL model achieved an average DICE coefficient of 0.774±0.126, 0.776±0.120, 0.767±0.125, and 0.776±0.125 for nodules with sharp margins (≥ 4.25), lobulated (≥ 1.75), spiculated (≥ 1.50), and solid (≥ 5.00) nodules, respectively, which were comparable ($P > 0.05$) to radiologists' average M-DICE of 0.795±0.094, 0.775±0.113, 0.774±0.112, and 0.794±0.099, respectively. For malignant (≥ 3.00) nodules, the H-DL model achieved an average DICE coefficient of 0.783±0.123 compared to the average M-DICE of 0.785±0.105 achieved by

This article is protected by copyright. All rights reserved.

radiologists (P > 0.05). Similar results of the JI metrics were achieved by the H-DL model. The details of the comparisons were shown in Table 1.

To evaluate the robustness of the H-DL model against the variability of centering the VOI at the nodule, we also deployed our H-DL model to the VOI obtained by shifting the LIDC-defined nodule center with a random distance (up to 1/3 of the longest diameter of a nodule and at the same time keeping the nodule within the VOI) in the horizontal or vertical direction for each test nodule. This simulates the situation that the nodule candidate is detected automatically in a computer-aided diagnosis pipeline, where the centroid of the detected object may not be well-centered because the object boundary is unknown before segmentation but the centroid is still located within the object region. The results showed that the average 3D DICE coefficient of 0.741 ± 0.142 using our H-DL achieved at the shifted-VOI was not significantly different (p-value > 0.05 by paired t-test) from that of the VOIs centered at the LIDC-defined nodule centers.

We also compared the segmentation results of our H-DL model with four nodule segmentation methods reported in the literature, which were also tested by LIDC cases. Table 2 shows that our H-DL achieved higher DICE coefficients and JI (if reported in their studies).

LNDb: Grand Challenge on automatic lung cancer patient management

We deployed our H-DL model that has been trained with the LIDC dataset directly without retraining to the test set of LNDb Sub-Challenge B for lung nodule segmentation with CT images [26]. The test set of LNDb-challenge-B contained 58 CT scans. LNDb provided challenge participants the VOI findings

This article is protected by copyright. All rights reserved.

from an automated lung nodule detection method, in which each CT scan contained 50 VOIs of nodule candidates, resulting a total of 2900 VOIs. Among those 50 VOIs for each CT, only one or two were the true positive of nodule, others were false positives from the automated nodule detection, and LNDb only evaluated the segmentation performance for those true positive nodules which were manually outlined by LNDb but unknown to the challenge participants. During the LNDb-challenge-B, the participates trained their lung nodule segmentation methods with an LNDb training set that contained only the true nodules with manual outlines, then applied to the test set containing 2900 VOIs of true or false positives and submitted the segmentation results at the LNDb website for evaluation. Although the submission for the challenge ranking was closed, the submission still remains open for those who want to benchmark algorithms. Without re-training with the LNDb-provided training set, we directly deployed our H-DL model to the LNDb test set and submitted the segmentation results to the LNDb for evaluation. The LNDb evaluation results showed that our H-DL model achieved a Hausdorff distance (HD) of 3.05 mm and a JI of 0.468. Comparing with the participants in the leaderboard listed at the LNDb website, our H-DL model would be ranked at the 5th place in the total ranking leaderboard, while the teams of 1st-place achieved a HD of 2.028 mm and a JI of 0.522, and the original 5th-place achieved a HD of 4.406 mm and a JI of 0.403.

Discussion

In CT screening of lung cancer, the measurement of nodule size, especially the volume of a nodule, is a vital tool that can help differentiate malignant nodules from benign nodules by the nodule growth rate. The growth rate is estimated by monitoring the change in nodule volume in serial CT scans, such as between the baseline screening CT and the follow-up scans. CT volumetry also plays an important role in lung cancer treatment by providing size change information to assess treatment

This article is protected by copyright. All rights reserved.

response. Manual segmentation of lung nodules is a time-consuming and tedious task, and substantial inter-radiologist variability exists as evident in the LIDC study[33]. Automatic lung nodule segmentation can provide this valuable information without radiologists' effort or requiring only minimal effort in identifying the nodule of interest. Once developed and validated, computerized measurement can be more consistent and reproducible in segmenting the nodule boundaries and thus quantifying the volume changes without inter- and intra-radiologist variabilities. Automated nodule segmentation is also a fundamental step in computer-aided diagnosis that can assist radiologists in classifying lung nodules as malignant or benign by extracting radiomic features.

In the past decades, a large number of studies have investigated diverse methods, and most of the methods used a single model to segment lung nodules. Those developed models cannot properly represent decision bounds of a broad spectrum of nodules with high heterogeneity. In this study, we developed a hybrid model that combined two lung nodule segmentation models with different neural network architectures and demonstrated that combining multiple models may have the potential to better adapt to the heterogeneous data distribution in the lung nodule segmentation task. Although it has been shown that a feedforward neural network using only one single hidden layer that contains enough neurons can approximate any model[34], it is difficult to determine the number of nodes needed. As the number of neurons used in a multi-layer network could be quite large, a deeper neural network with multiple layers could be more efficient and flexible to accomplish the tasks[35]. With an increased number of network layers, expression and abstraction learning abilities will be increased in the network[36-38]. However, in practice, the deepness of architecture has a significant drawback because excessive depth may degrade the accuracy[36,39]. In general, each layer will produce a lossy-compression-like effect after passing through the convolution kernel. The deeper neural network with multiple levels of convolutions may inevitably

This article is protected by copyright. All rights reserved.

extract features of excessive abstraction and is more difficult to train than a shallow neural network. In this study, we balanced a deep network with a shallow network and leveraged state-of-the-art network structures such as VGG and DenseNet to increase the learning capabilities of our H-DL model and exploit different levels of image features from CT images (Figure 4). In figure 7, we showed some examples to demonstrate that a deep or shallow network works better or worse than each other for different kinds of nodules. In summary, a shallow encoding path will focus on more high-level features and our result demonstrated that it had better segmentation for nodules of large sizes with smooth or sharp margins. On the contrary, a deep encoding path will focus more on detail information and our result demonstrated that it had better segmentation result for small, no-solid nodules with poorly defined margin.

Our results showed that, although the segmentation results are different between Shallow U-DL and Deep U-DL when trained with the same training set, the segmentation accuracies achieved by the combined H-DL model were superior to either one alone for most nodules (Table 1). It indicated that the different network structures of the two U-DL models can extract features at different levels of abstraction and provide complementary information in the hybrid model.

In the LIDC data set, the ratings provided by radiologists for the descriptors of the different nodule characteristics as well as their manual outlines of nodule boundaries exhibited large variabilities. Table 3 shows the root-mean-square deviations (RMSD) of the ratings for individual nodules provided by radiologists, averaged over all nodules, for each of the descriptors. To evaluate the performance of our H-DL model and the separate Shallow U-DL and Deep U-DL, we separated the nodules into two groups by using the radiologists' ratings for each descriptor, as described in the

This article is protected by copyright. All rights reserved.

Results section. Although the segmentation accuracies achieved by our H-DL model and those by the radiologists showed a similar trend for most of the descriptors in each group (Table 1), there are exceptions. For example, both the H-DL model and the radiologists achieved higher segmentation accuracies for the nodules with sharp margins (≥4.25) than nodules with fuzzy margins. On the other hand, radiologists had lower accuracy in segmenting the spiculated nodules (spiculated ≥1.5) compared with non-spiculated nodules (M-DICE of 0.774 vs. 0.783), whereas the H-DL model achieved a better segmentation result for spiculated nodules than for non-spiculated nodules (DICE of 0.767 vs. 0.724). Similarly, radiologists achieved a lower accuracy in segmenting lobulated nodules (lobulated ≥1.75) than less lobulated nodules (M-DICE of 0.775 vs 0.782), whereas the H-DL model had much better performance with lobulated nodules than less lobulated nodules (DICE of 0.776 vs. 0.724). One reason could be the difficulty in visually judging the nodule boundaries in the presence of subtle spiculations and lobulations, and another reason could be that it was too time-consuming for radiologists to consistently trace the spiculations or lobulations. Because the degree of spiculation or lobulation of the nodule boundary is strongly correlated with the probability of malignancy, an automated segmentation tool that can segment the boundary of these nodules accurately, reproducibly, and efficiently will be helpful in the assessment of screen-detection lung nodules.

We have conducted a preliminary exploration of methods to hybridize the two U-DL networks. We compared our current method of using a trained ensemble convolutional layer to a simple voting method using a pre-defined voting threshold. We observed differences in the segmentation performance measures for the nodule subgroups of different characteristics; however, the latter method achieved an average 3D DICE coefficient of 0.749±0.141 and an average JI of 0.617±0.162,

This article is protected by copyright. All rights reserved.

which were similar to the former method. This indicated that the ensemble layer might have learned a similar strategy as voting but using more adaptive weighting for different nodule characteristics. Although the results were similar on average, a potential advantage is that the ensemble layer method may be more robust as it achieved smaller variances. Further studies are needed to evaluate different fusion methods.

To evaluate the performance of our H-DL model with the LNDb-challenge-B test set, we directly deployed our H-DL model to the LNDb test set that had several major differences with LIDC dataset, including, 1). Our LIDC training set only included nodules with size >= 7mm, the LNDb included many small nodules down to 3 mm. 2) The annotations of LNDb nodules were relatively rough and lacked details of the nodule boundary compared with LIDC annotation, especially for the irregular-shaped or spiculated nodules (Figure 8), 3) LNDb cropped the VOI to a size of 80 x 80 x 80 pixels with the image resolution normalized to 0.6375mm x 0.6375mm x 0.6375mm, that was different from our VOI of 64 x 64 x 64 pixels with resolution of 0.5 mm x 0.5 mm x 0.5 mm. Since we did not retrain the network and kept the input dimension as before, we automatically cut the VOI symmetrically from 80 x 80 x 80 to 64 x 64 x 64 and then padded zeros to the segmentation results' peripheral to recover 80 x 80 x 80 VOI.

Despite the differences, we were still able to achieve competitive performance without retraining with the LNDb training dataset, demonstrating the generalizability and robustness of our H-DL model in lung nodule segmentation. Among the LNDb-Challenge-B leaderboard, most participants used 3D U-Net structure[40-42] which may have inherent advantages for volume segmentation of nodules in 3D CT images. However, the 3D U-Net usually requires more training time and more GPU/CPU

This article is protected by copyright. All rights reserved.

memories, and the achieved accuracies by those 3D networks were widely ranked across the leaderboard, indicating that the 3D network architectures may not be the major reason to achieve better results. Additionally, as the LNDb allowed the participants to submit their results multiple times, the higher ranked models also used some pre- or post-processing methods to improve the test performance, such as the methods of attention mechanism CBAM and switchable normalization for fine tuning of the loss functions and hyper-parameters based on the feedbacks of submitted test results[43], or using self-supervised learning method[44] combined with their own dataset with LNDb training set to train the model. Among the participated methods, a DL model[44] used the similar training method as ours: trained a model with the LIDC dataset and then deployed to the LNDb test set. This model employed a conventional 3-D U-net structure and achieved a HD of 9.12 mm and JI of 0.22, which were significantly lower than those achieved by our H-DL model.

There are several limitations in this study. In our H-DL method, the two U-DL base models shared a similar U-shape structure that could limit the networks to explore more diverse features to better characterize lung nodules. We will study other state-of-the-art networks with different architectures such as Mask R-CNN[19] or YOLO[45] that can be adapted to our U-DL models to further improve the hybrid model for lung nodule segmentation. Another limitation is that we have not extensively optimized the fusion method and explored methods such as attention structures to hybridize the outputs from the U-DL models. These limitations will be addressed in future studies.

This article is protected by copyright. All rights reserved.

Conclusion

In this study, we developed a new hybrid deep learning (H-DL) method for volume segmentation of lung nodules with large variations in size, shape, margin, and opacity in CT scans. The H-DL model combined two asymmetric U-shaped network architectures, one with a 16-layer shallow DCNN and the other with a 200-layer deep DCNN as encoders for feature extraction. The results demonstrated that our H-DL model outperformed the individual shallow or deep U-DL models. The hybrid deep learning method combining multi-level features learned by both the shallow and deep DCNNs could achieve high segmentation accuracy comparable to radiologists' segmentation for nodules with wide ranges of image characteristics.

ACKNOWLEDGEMENT

This study is supported by NIH grant U01CA216459.

CONFLICT OF INTEREST

The authors have no conflicts to disclose.

DATA AVAILABILITY STATEMENT

The Lung Image Database Consortium image collection (LIDC-IDRI) is public available from website[27] refer to https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI

Grand Challenge on automatic lung cancer patient management dataset is public available from website[26] refer to https://lndb.grand-challenge.org

This article is protected by copyright. All rights reserved.

REFERENCES

1. *Key statistics for lung cancer*.

2. Han D, Marjolein A. Heuvelmans, and Matthijs Oudkerk. Volume versus diameter assessment of small pulmonary nodules in CT lung cancer screening. In. *Translational lung cancer research* 6, no. 1 (2017): 52.2017.

3. Wang S, Zhou, M., Liu Z., Liu, Z., Gu, D., Zang, Y., ... & Tian, J. Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation. In. Medical Image Analysis, 40, 172-183.2017.

4. Fetita CI, Preteux, F. Beigelman-Aubry, C., & Grenier, P. 3D automated lung nodule segmentation in HRCT. In. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 626-634). Springer, Berlin, Heidelberg.2003.

5. Nithila EE, & Kumar, S. S. Segmentation of lung nodule in CT data using active contour model and Fuzzy C-mean clustering. In. Alexandria Engineering Journal, 55(3), 2583-2588.2016.

6. Xie H, Yang, D., Sun, N. Chen, Z. and Zhang, Y. Automated pulmonary nodule detection in CT images using deep convolutional neural networks. In. Pattern Recognition, 85, pp.109-119.2019.

7. Diciotti S, Lombardo, S.,, Falchini M, Picozzi, G., & Mascalchi, M. Automated segmentation refinement of small lung nodules in CT scans by local shape analysis. In. IEEE Transactions on Biomedical Engineering, 58(12), 3418-3428.2011.

8. Messay T, Hardie, R. C. & Tuinstra, T. R. Segmentation of pulmonary nodules in computed tomography using a regression neural network approach and its application to the lung image database consortium and image database resource initiative dataset. In. Medical Image Analysis, 22(1), 48-62.2015.

9. Dehmeshki J, Amin, H. Valdivieso, M., & Ye, X. Segmentation of pulmonary nodules in thoracic CT scans: a region growing approach. In. IEEE Transactions on Medical Imaging, 27(4), 467-480.2008.

10. Farag AA, Abd El Munim H. E., Graham, J. H., & Farag, A. A. A novel approach for lung nodules segmentation in chest CT using level sets. In. IEEE Transactions on Image Processing, 22(12), 5202-5213.2013.

11. Ye X, Beddoe, G., & Slabaugh, G. Automatic graph cut segmentation of lesions in CT using mean shift superpixels. In. International Journal of Biomedical Imaging, 2010.2010.

12. Sahba F, Tizhoosh, H.R. and Salama, M.M., July. A reinforcement learning framework for medical image segmentation. In. In The 2006 IEEE International Joint Conference on Neural Network Proceedings (pp. 511-517). IEEE.(2006)2006.

13. Way TW, Hadjiiski, L. M., Sahiner, B., Chan, H.-P., Cascade, P. N., Kazerooni, E. A., ... & Zhou C. Computer-aided diagnosis of pulmonary nodules on CT scans: Segmentation and classification using 3D active contours. In. Medical Physics, 33(7Part1), 2323-2337.2006.

This article is protected by copyright. All rights reserved.

14. Zhou C, Chan, H.-P., Chughtai,, A. H, L. M., Kazerooni, E. A., & Wei, J. Pathologic categorization of lung nodules: Radiomic descriptors of CT attenuation distribution patterns of solid and subsolid nodules in low-dose CT. In. European Journal of Radiology, 129, 109106.2020.

15. Simonyan K, & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In. arXiv preprint arXiv:1409.1556.2014.

16. Hasan M, & Aleef, T. A. Automatic mass detection in breast using deep convolutional neural network and svm classifier. In. arXiv preprint arXiv:1907.04424.2019.

17. Huang G, Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. Densely connected convolutional networks. In. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4700-4708).2017.

18. Huang X, Sun, W., Tseng, T. L. B., Li, C., & Qian, W. Fast and fully-automated detection and segmentation of pulmonary nodules in thoracic CT scans using deep convolutional neural networks. In. Computerized Medical Imaging and Graphics, 74, 25-36.2019.

19. Kopelowitz E, & Engelhard, G. Lung Nodules Detection and Segmentation Using 3D Mask-RCNN. In. arXiv preprint arXiv:1907.07676.2019.

20. Ronneberger O, Fischer P., & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 234-241). Springer, Cham.2015.

21. Aresta G, Jacobs, C. Araújo, T., Cunha, A., Ramos, I., van Ginneken, B., & Campilho, A. iW-Net: an automatic and minimalistic interactive lung nodule segmentation deep network. In. Scientific Reports, 9(1), 1-9.2019.

22. Wu B, Zhou, Z., Wang, J. & Wang, Y. Joint learning for pulmonary nodule segmentation, attributes and malignancy prediction. In. In Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018) (pp.1109-1113).2018.

23. Szegedy C, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In. In P*roceedings of the IEEE conference on computer vision and pattern recognition* , pp. 1-9. 2015.2015.

24. Siddique N, Paheding Sidike, Colin Elkin, and Vijay Devabhaktuni. U-Net and its variants for medical image segmentation: theory and applications. In. *arXiv preprint arXiv:2011.01118* (2020).2020.

25. Zhou Z, Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In. Deep learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (pp.3-11). Springer, Cham.2018.

26. Pedrosa J, Guilherme Aresta, Carlos Ferreira, Márcio Rodrigues, Patrícia Leitão, André Silva Carvalho, João Rebelo et al. LNDb: a lung nodule database on computed tomography. In. *arXiv preprint arXiv:1911.08434* (2019).2019.

This article is protected by copyright. All rights reserved.

27. Armato III SGM, G., Bidaut, L., McNitt - Gray, M. F., Meyer, C. R., Reeves, A. P., ... & Clarke, L. P. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. In. Medical Physics, 38(2), 915-931.2011.

28. Christian Szegedy VV, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2818–2826, 2016.2018.

29. Ibtehaz N, and M. Sohel Rahman. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. In. Neural Networks 121 (2020): 74-87.2020.

30. Li R, W. Liu, L. Yang, S. Sun, W. Hu, F. Zhang, and W. DeepUNet Li. A Deep Fully Convolutional Network for Pixel-Level Sea-Land Segmentation. . In. arXiv." a*rXiv preprint arXiv:1709.00201* (2017)2017.

31. Kingma DP, & Ba, J. Adam: A method for stochastic optimization. In. arXiv preprint arXiv:1412.6980.2014.

32. Çiçek Ö, Abdulkadir, A. Lienkamp, S. S., Brox, T., & Ronneberger, O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In. In International Conference on Medical Image Computing and Computer-Assised Intervention (pp. 424-432). Springer, Cham.2016.

33. Armato SG, Roberts, R. Y., McNitt-Gray, M. F., Meyer, C. R., Reeves, A. P., McLennan, G., ... & Clarke, L. P. The Lung Image Database Consortium (LIDC): Ensuring the integrity of expert-defined “truth”. In: Academic Radiology, 14(12), 1455-1463.; 2007.

34. Goodfellow I, Bengio, Y. & Courville, A. Deep learning. Book in preparation for MIT Press. In. URL http://www.deeplearningbook. org, 1.2016.

35. Zhao ZQ, Zheng, P., Xu S. T., & Wu, X. Object detection with deep learning: A review. In. IEEE Transactions on Neural Networks and Learning Systems, 30(11), 3212-3232.2019.

36. He K, & Sun, J. Convolutional neural networks at constrained time cost. In. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5353-5360).2015.

37. Wang H, & Raj, B. On the origin of deep learning. In. arXiv preprint arXiv:1702.07800.2017.

38. Delalleau O, & Bengio, Y. Shallow vs. deep sum-product networks. In. Advances in Neural Information Processing Systems, 24, 666-674.2011.

39. He K, Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).2016.

40. Chen H, Qi Dou, Xi Wang, Jing Qin, Jack CY Cheng, and Pheng-Ann Heng. 3D fully convolutional networks for intervertebral disc localization and segmentation. In. *International Conference on Medical Imaging and Augmented Reality* , pp. 375-382. Springer, Cham, 2016.2016.

41. Baumgartner CF, Lisa M. Koch, Marc Pollefeys, and Ender Konukoglu. An exploration of 2D and 3D deep learning techniques for cardiac MR image segmentation. In. In I*nternational Workshop on Statistical Atlases and Computational Models of the Heart* , pp. 111-119. Springer, Cham, 2017.2017.

This article is protected by copyright. All rights reserved.

42. Zhou X, Ryosuke Takayama, Song Wang, Takeshi Hara, and Hiroshi Fujita. Deep learning of the sectional appearances of 3D CT images for anatomical structure segmentation based on an FCN voting method. In. *Medical physics* 44, no. 10 (2017): 5221-5233.2017.

43. Galdran A, and Hamid Bouchachia. Residual Networks for Pulmonary Nodule Segmentation and Texture Characterization. In. In I*nternational Conference on Image Analysis and Recognition* , pp. 396-405. Springer, Cham, 2020.2020.

44. Kaluva KC, Kiran Vaidhya, Abhijith Chunduru, Sambit Tarai, Sai Prasad Pranav Nadimpalli, and Suthirth Vaidya. An Automated Workflow for Lung Nodule Follow-Up Recommendation Using Deep Learning. In. In I*nternational Conference on Image Analysis and Recognition* , pp. 369-377. Springer, Cham, 2020.2020.

45. Redmon J, & Farhadi A. Yolov3: An incremental improvement. In. arXiv preprint arXiv:1804.02767.2018.

Table 1. Test results achieved by H-DL model, Shallow U-DL and Deep U-DL alone, and the performance of LIDC radiologists' manual segmentations relative to the reference standard. The *P*-values of the differences between H-DL and others were calculated by paired *t*-test. The *P*-value < 0.05 indicated that the difference was statistically significant. M-DICE and M-JI are the averages of the median, and m-DICE and m-JI are the averages of the minimum DICE and JI by radiologists, respectively.

		H-DL		Radiologist				Shallow U-DL		Deep U-DL	
Categories		DICE	JI	M-DICE	min-DICE	M-JI	min-JI	DICE	JI	DICE	JI
All nodules	N = 327	0.750 ± 0.135	0.617 ± 0.159	0.778 ± 0.102	0.685 ± 0.139	0.651 ± 0.127	0.537 ± 0.153	0.745 ± 0.139	0.611 ± 0.161	0.739 ± 0.145	0.604 ± 0.163
	P-value			**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**
Diameter (mm)	≥ 10.00 (N = 208)	0.782 ± 0.128	0.658 ± 0.152	0.790 ± 0.105	0.704 ± 0.135	0.667 ± 0.128	0.559 ± 0.150	0.777 ± 0.129	0.652 ± 0.154	0.766 ± 0.150	0.640 ± 0.166
	P-value			0.423	**< 0.05**	0.434	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**
	< 10.00 (N = 119)	0.695 ± 0.131	0.547 ± 0.145	0.758 ± 0.093	0.652 ± 0.139	0.623 ± 0.121	0.499 ± 0.153	0.688 ± 0.137	0.540 ± 0.149	0.691 ± 0.125	0.541 ± 0.137
	P-value			**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	0.651	0.498
Subtlety	≥ 4.25 (N = 163)	0.798 ± 0.113	0.677 ± 0.140	0.785 ± 0.116	0.701 ± 0.148	0.662 ± 0.139	0.558 ± 0.163	0.796 ± 0.111	0.674 ± 0.138	0.785 ± 0.125	0.662 ± 0.149
	P-value			0.210	**< 0.05**	0.222	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**

This article is protected by copyright. All rights reserved.

	< 4.25 (N = 164)	0.702 ± 0.139	0.557 ± 0.154	0.771 ± 0.086	0.670 ± 0.128	0.640 ± 0.112	0.517 ± 0.140	0.694 ± 0.145	0.549 ± 0.159	0.693 ± 0.150	0.548 ± 0.157
	P-value			**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**0.197**	**0.173**
Malignancy	≥ 3.00 (N = 203)	0.783 ± 0.123	0.658 ± 0.149	0.785 ± 0.105	0.700 ± 0.134	0.660 ± 0.127	0.553 ± 0.147	0.778 ± 0.124	0.652 ± 0.148	0.773 ± 0.135	0.647 ± 0.155
	P-value			0.842	**< 0.05**	0.855	**< 0.05**	**< 0.05**	**< 0.05**	0.078	0.058
	< 3.00 (N = 124)	0.697 ± 0.139	0.551 ± 0.157	0.767 ± 0.095	0.661 ± 0.145	0.636 ± 0.126	0.511 ± 0.160	0.690 ± 0.145	0.544 ± 0.16	0.683 ± 0.145	0.535 ± 0.152
	P-value			**< 0.05**	0.056	**< 0.05**	0.051	**< 0.05**	**< 0.05**	0.100	**< 0.05**
Margin	≥ 4.25 (N =175)	0.774 ± 0.126	0.647 ± 0.154	0.795 ± 0.094	0.710 ± 0.137	0.672 ± 0.123	0.567 ± 0.155	0.767 ± 0.135	0.640 ± 0.160	0.756 ± 0.149	0.627 ± 0.166
	P-value			0.070	**< 0.05**	0.091	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**
	< 4.25 (N =152)	0.723 ± 0.140	0.582 ± 0.157	0.759 ± 0.107	0.657 ± 0.136	0.627 ± 0.128	0.503 ± 0.144	0.719 ± 0.140	0.578 ± 0.156	0.720 ± 0.139	0.579 ± 0.156
	P-value			**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	0.667	0.584
Lobulated	≥ 1.75 (N=165)	0.776 ± 0.120	0.648 ± 0.147	0.775 ± 0.113	0.682 ± 0.151	0.648 ± 0.136	0.536 ± 0.162	0.771 ± 0.121	0.642 ± 0.148	0.767 ± 0.131	0.638 ± 0.154
	P-value			0.958	**< 0.05**	0.978	**< 0.05**	**< 0.05**	**< 0.05**	0.136	0.097
	< 1.75 (N=162)	0.724 ± 0.145	0.586 ± 0.164	0.782 ± 0.089	0.689 ± 0.126	0.655 ± 0.117	0.539 ± 0.144	0.718 ± 0.151	0.579 ± 0.168	0.710 ± 0.154	0.570 ± 0.166
	P-value			**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**0.058**	**< 0.05**
Spiculated	≥ 1.50 (N = 173)	0.767 ± 0.125	0.638 ± 0.150	0.774 ± 0.112	0.680± 0.142	0.648 ± 0.134	0.532 ± 0.153	0.764 ± 0.125	0.633 ± 0.149	0.755 ± 0.145	0.624 ± 0.162
	P-value			0.515	**< 0.05**	0.443	**< 0.05**	0.767	0.739	0.250	0.309
	< 1.50 (N = 154)	0.731 ± 0.143	0.596 ± 0.164	0.783 ± 0.089	0.691 ± 0.136	0.655 ± 0.119	0.543 ± 0.154	0.723 ± 0.151	0.586 ± 0.17	0.721 ± 0.145	0.582 ± 0.162
	P-value			**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**0.587**	**0.608**	**0.494**	**0.431**
Solid	≥ 5.00 (N = 167)	0.776 ± 0.125	0.649 ± 0.151	0.794 ± 0.099	0.709 ± 0.135	0.672 ± 0.126	0.564 ± 0.154	0.771 ± 0.131	0.644 ± 0.156	0.759 ± 0.150	0.631 ± 0.166
	P-value			0.092	**< 0.05**	0.084	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**
	< 5.00 (N = 160)	0.723 ± 0.141	0.584 ± 0.160	0.762 ± 0.103	0.661 ± 0.139	0.629 ± 0.125	0.509 ± 0.147	0.717 ± 0.142	0.577 ± 0.160	0.718 ± 0.138	0.577 ± 0.156
	P-value			**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	0.630	0.457	0.685	0.447
Sphericity	≥ 4.00 (N = 179)	0.760 ± 0.139	0.632 ± 0.162	0.796 ± 0.085	0.706 ± 0.129	0.673 ± 0.112	0.561 ± 0.148	0.756 ± 0.144	0.626 ± 0.166	0.754 ± 0.143	0.623 ± 0.163
	P-value			**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**0.212**	**0.129**

This article is protected by copyright. All rights reserved.

< 4.00 (N = 148)	0.737 ± 0.129	0.600± 0.153	0.757 ± 0.116	0.660 ± 0.146	0.624 ± 0.139	0.509 ± 0.155	0.732 ± 0.131	0.593 ± 0.154	0.721 ± 0.147	0.582 ± 0.162
P-value			0.144	**< 0.05**	0.058	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**

Table 2. The performance comparisons of our H-DL model with other deep learning methods in the segmentation of lung nodules.

	Dataset used	VOI size	DICE	JI
H-DL model	LIDC-IDRI	64 x 64 x 64	**0.750±0.135**	**0.617±0.159**
3D U-Net[32]	LIDC-IDRI	64 x 64 x 64	0.720±0.049	0.380±0.080
iW-Net[21]	LIDC-IDRI	64 x 64 x 64	-	0.550±0.140
PN-SAMP[22]	LIDC-IDRI	64 x 64 x 64	0.741±0.357	-

Table 3. The root-mean-square deviation (RMSD) of the descriptors of nodule characteristics was provided by different radiologists in the test set. The ratings of the descriptors were given on a 5-point scale except for the diameter.

	Diameter (mm)	Subtlety	Malignancy	Margin	Lobulated	Spiculated	Solid	Sphericity
RMSD	1.725	0.659	0.787	0.706	0.79	0.709	0.51	0.691

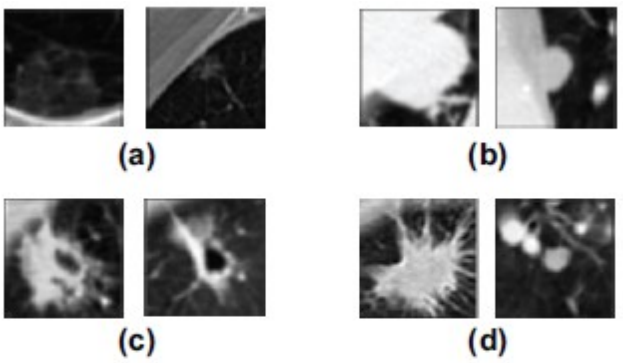

Figure 1. Examples of lung nodules from the LIDC dataset used in the test set of this study with different characteristics in CT images: (a) ground-glass opacity (GGO) nodule. (b) juxtapleural nodule. (c) cavitary nodule. (d) nodule with irregular margins.

This article is protected by copyright. All rights reserved.

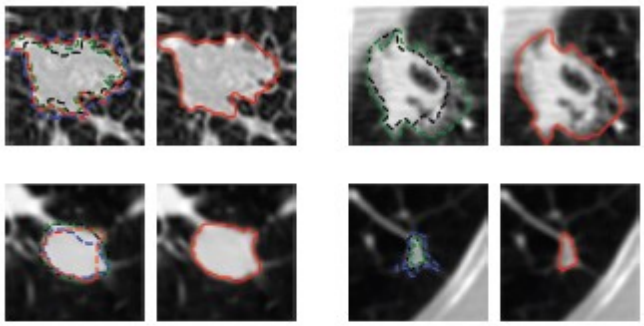

Figure 2. Examples of radiologists' annotations on four different nodules showing the variabilities among radiologists in manual outlining of nodules (left, dashed contours with different colors). The 50% consensus consolidation of radiologists' annotations for each nodule was used as the reference standard (right, red solid contour).

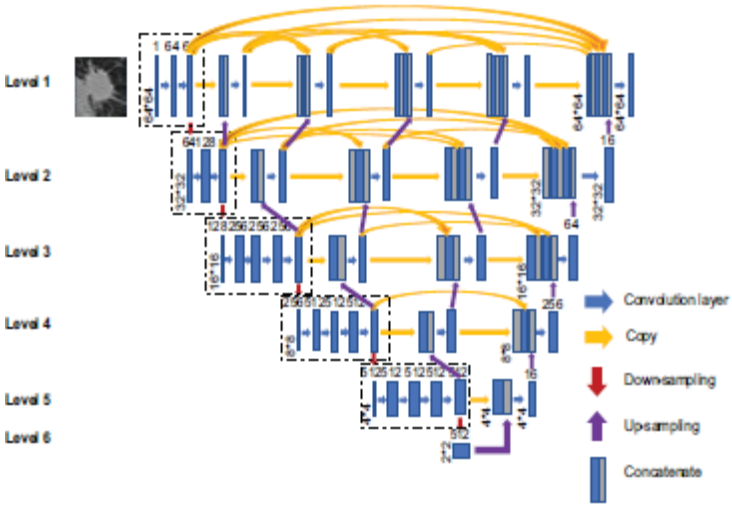

Figure 3. (a) The architecture of our asymmetric U-shaped deep learning model using VGG19-based encoder path (Shallow U-DL) for lung nodule segmentation in CT images. The size of each feature map is shown at the lower-left edge of the box. The arrows of different colors represent different operations.

This article is protected by copyright. All rights reserved.

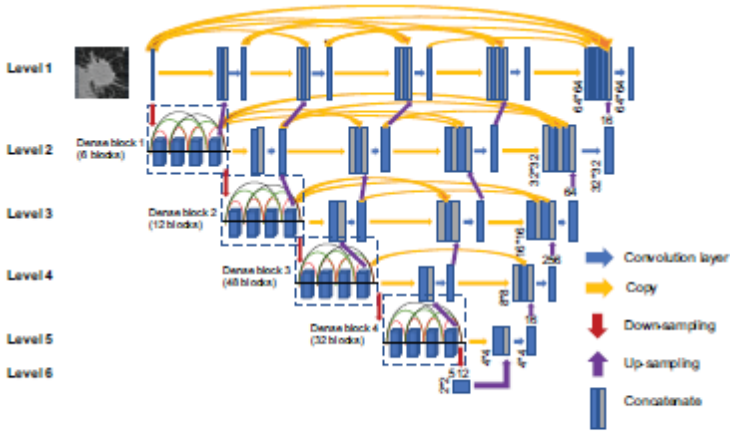

Figure 3. (b) The architecture of our asymmetric U-shaped deep learning model using deep DenseNet-based encoder path (Deep U-DL) for lung nodule segmentation in CT images. The size of each feature map is shown at the lower-left edge of the box. The arrows of different colors represent different operations.

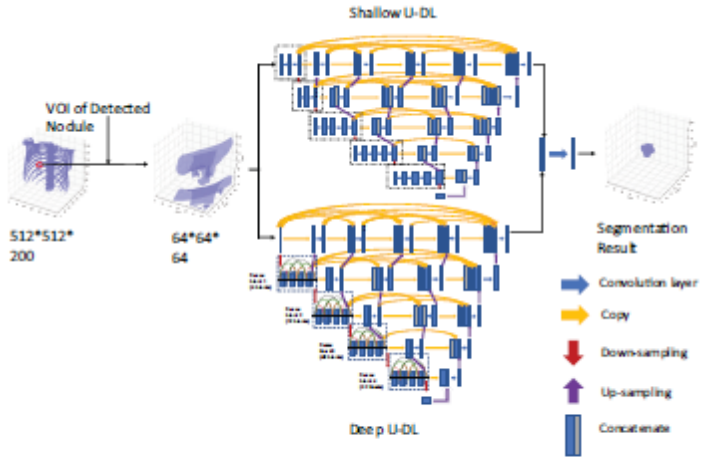

Figure 4: The overall framework of our Hybrid deep-learning (H-DL) model.

This article is protected by copyright. All rights reserved.

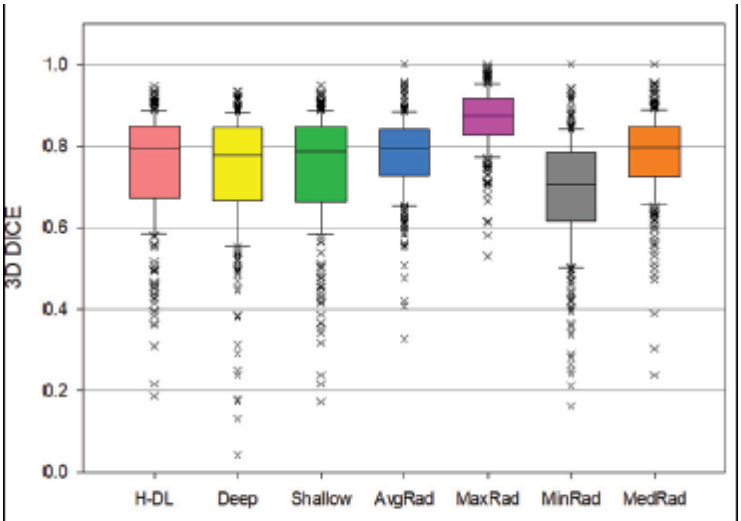

Figure 5. Box plots showing the distributions of the 3D DICE coefficients for the lung nodule segmentation achieved by the H-DL model, Deep U-DL model, Shallow U-DL model, (HDL, Deep, Shallow, respectively) and average, maximum, minimum and median of the radiologists' manual outlines (AvgRad, MaxRad, MinRad, MedRad, respectively) relative to the reference standard. In this plot, the horizontal line represents the median value, the top line of the box is the 25% quartile, and the bottom line of the box is the 75% quartile, the whiskers are 10th and 90th percentiles and the points are outliers.

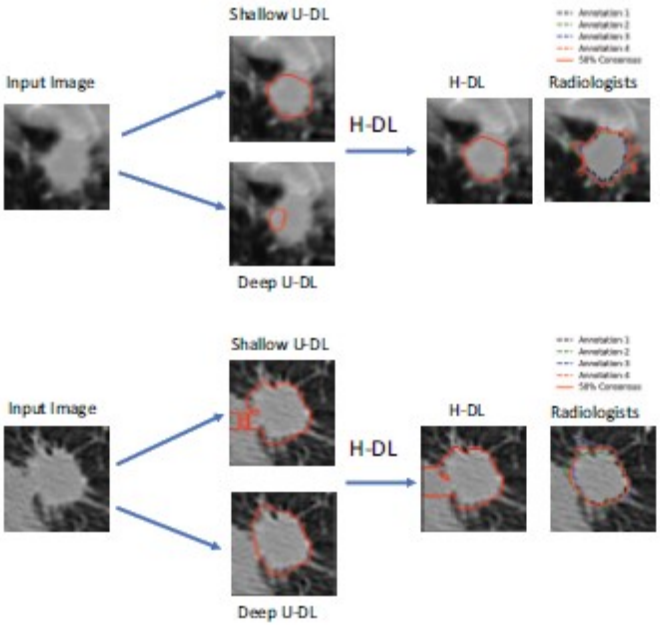

Figure 6: Examples showing the differences between the H-DL, Shallow U-DL, and Deep U-DL models.

This article is protected by copyright. All rights reserved.

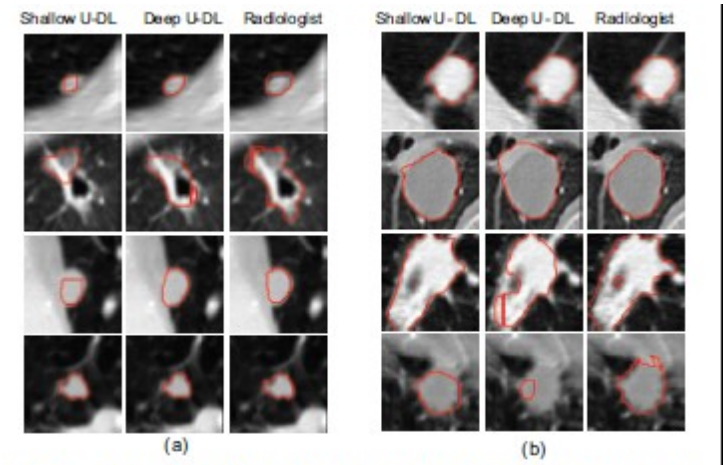

Figure 7. Examples showing the differences between the nodule segmentations by the shallow and deep U-DL model alone. (a) Deep U-DL outperformed shallow U-DL, (b) Shallow U-DL outperformed deep U-DL.

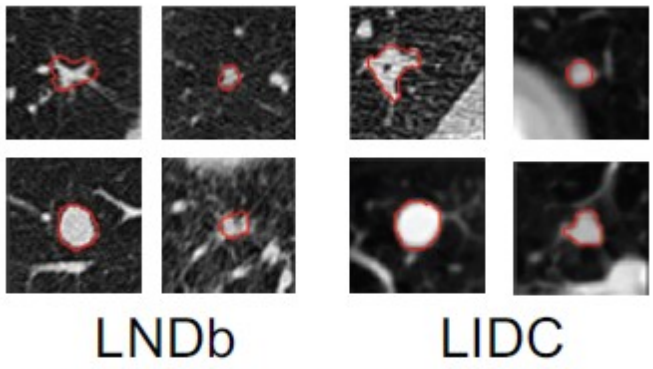

Figure 8. The annotations manually outlined by LNDb (left) and LIDC (right) readers, showing that the LIDC radiologist provided more refined outlines than those of LNDb, especially for the irregular-shaped or spiculated nodules.

Table 1. Test results achieved by H-DL model, Shallow U-DL and Deep U-DL alone, and the performance of LIDC radiologists' manual segmentations relative to the reference standard. The *P*-values of the differences between H-DL and others were calculated by paired *t*-test. The *P*-value < 0.05 indicated that the difference

This article is protected by copyright. All rights reserved.

was statistically significant. M-DICE and M-JI are the averages of the median, and m-DICE and m-JI are the averages of the minimum DICE and JI by radiologists, respectively.

		H-DL		Radiologist				Shallow U-DL		Deep U-DL	
Categories		DICE	JI	M-DICE	min-DICE	M-JI	min-JI	DICE	JI	DICE	JI
All nodules	N = 327	0.750 ± 0.135	0.617 ± 0.159	0.778 ± 0.102	0.685 ± 0.139	0.651 ± 0.127	0.537 ± 0.153	0.745 ± 0.139	0.611 ± 0.161	0.739 ± 0.145	0.604 ± 0.163
	P-value			**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**
Diameter (mm)	≥ 10.00 (N = 208)	0.782 ± 0.128	0.658 ± 0.152	0.790 ± 0.105	0.704 ± 0.135	0.667 ± 0.128	0.559 ± 0.150	0.777 ± 0.129	0.652 ± 0.154	0.766 ± 0.150	0.640 ± 0.166
	P-value			0.423	**< 0.05**	0.434	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**
	< 10.00 (N = 119)	0.695 ± 0.131	0.547 ± 0.145	0.758 ± 0.093	0.652 ± 0.139	0.623 ± 0.121	0.499 ± 0.153	0.688 ± 0.137	0.540 ± 0.149	0.691 ± 0.125	0.541 ± 0.137
	P-value			**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	0.651	0.498
Subtlety	≥ 4.25 (N = 163)	0.798 ± 0.113	0.677 ± 0.140	0.785 ± 0.116	0.701 ± 0.148	0.662 ± 0.139	0.558 ± 0.163	0.796 ± 0.111	0.674 ± 0.138	0.785 ± 0.125	0.662 ± 0.149
	P-value			0.210	**< 0.05**	0.222	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**
	< 4.25 (N = 164)	0.702 ± 0.139	0.557 ± 0.154	0.771 ± 0.086	0.670 ± 0.128	0.640 ± 0.112	0.517 ± 0.140	0.694 ± 0.145	0.549 ± 0.159	0.693 ± 0.150	0.548 ± 0.157
	P-value			**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**0.197**	**0.173**
Malignancy	≥ 3.00 (N = 203)	0.783 ± 0.123	0.658 ± 0.149	0.785 ± 0.105	0.700 ± 0.134	0.660 ± 0.127	0.553 ± 0.147	0.778 ± 0.124	0.652 ± 0.148	0.773 ± 0.135	0.647 ± 0.155
	P-value			0.842	**< 0.05**	0.855	**< 0.05**	**< 0.05**	**< 0.05**	0.078	0.058
	< 3.00 (N = 124)	0.697 ± 0.139	0.551 ± 0.157	0.767 ± 0.095	0.661 ± 0.145	0.636 ± 0.126	0.511 ± 0.160	0.690 ± 0.145	0.544 ± 0.16	0.683 ± 0.145	0.535 ± 0.152
	P-value			**< 0.05**	0.056	**< 0.05**	0.051	**< 0.05**	**< 0.05**	0.100	**< 0.05**
Margin	≥ 4.25 (N =175)	0.774 ± 0.126	0.647 ± 0.154	0.795 ± 0.094	0.710 ± 0.137	0.672 ± 0.123	0.567 ± 0.155	0.767 ± 0.135	0.640 ± 0.160	0.756 ± 0.149	0.627 ± 0.166
	P-value			0.070	**< 0.05**	0.091	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**
	< 4.25 (N =152)	0.723 ± 0.140	0.582 ± 0.157	0.759 ± 0.107	0.657 ± 0.136	0.627 ± 0.128	0.503 ± 0.144	0.719 ± 0.140	0.578 ± 0.156	0.720 ± 0.139	0.579 ± 0.156
	P-value			**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	0.667	0.584
Lobulated	≥ 1.75 (N=165)	0.776 ± 0.120	0.648 ± 0.147	0.775 ± 0.113	0.682 ± 0.151	0.648 ± 0.136	0.536 ± 0.162	0.771 ± 0.121	0.642 ± 0.148	0.767 ± 0.131	0.638 ± 0.154
	P-value			0.958	**< 0.05**	0.978	**< 0.05**	**< 0.05**	**< 0.05**	0.136	0.097

This article is protected by copyright. All rights reserved.

	< 1.75 (N=162)	0.724 ± 0.145	0.586 ± 0.164	0.782 ± 0.089	0.689 ± 0.126	0.655 ± 0.117	0.539 ± 0.144	0.718 ± 0.151	0.579 ± 0.168	0.710 ± 0.154	0.570 ± 0.166
	P-value			**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**0.058**	**< 0.05**
Spiculated	≥ 1.50 (N = 173)	0.767 ± 0.125	0.638 ± 0.150	0.774 ± 0.112	0.680± 0.142	0.648 ± 0.134	0.532 ± 0.153	0.764 ± 0.125	0.633 ± 0.149	0.755 ± 0.145	0.624 ± 0.162
	P-value			0.515	**< 0.05**	0.443	**< 0.05**	0.767	0.739	0.250	0.309
	< 1.50 (N = 154)	0.731 ± 0.143	0.596 ± 0.164	0.783 ± 0.089	0.691 ± 0.136	0.655 ± 0.119	0.543 ± 0.154	0.723 ± 0.151	0.586 ± 0.17	0.721 ± 0.145	0.582 ± 0.162
	P-value			**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**0.587**	**0.608**	**0.494**	**0.431**
Solid	≥ 5.00 (N = 167)	0.776 ± 0.125	0.649 ± 0.151	0.794 ± 0.099	0.709 ± 0.135	0.672 ± 0.126	0.564 ± 0.154	0.771 ± 0.131	0.644 ± 0.156	0.759 ± 0.150	0.631 ± 0.166
	P-value			0.092	**< 0.05**	0.084	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**
	< 5.00 (N = 160)	0.723 ± 0.141	0.584 ± 0.160	0.762 ± 0.103	0.661 ± 0.139	0.629 ± 0.125	0.509 ± 0.147	0.717 ± 0.142	0.577 ± 0.160	0.718 ± 0.138	0.577 ± 0.156
	P-value			**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	0.630	0.457	0.685	0.447
Sphericity	≥ 4.00 (N = 179)	0.760 ± 0.139	0.632 ± 0.162	0.796 ± 0.085	0.706 ± 0.129	0.673 ± 0.112	0.561 ± 0.148	0.756 ± 0.144	0.626 ± 0.166	0.754 ± 0.143	0.623 ± 0.163
	P-value			**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**0.212**	**0.129**
	< 4.00 (N = 148)	0.737 ± 0.129	0.600± 0.153	0.757 ± 0.116	0.660 ± 0.146	0.624 ± 0.139	0.509 ± 0.155	0.732 ± 0.131	0.593 ± 0.154	0.721 ± 0.147	0.582 ± 0.162
	P-value			0.144	**< 0.05**	0.058	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**

Table 2. The performance comparisons of our H-DL model with other deep learning methods in the segmentation of lung nodules.

	Dataset used	VOI size	DICE	JI
H-DL model	LIDC-IDRI	64 x 64 x 64	**0.750±0.135**	**0.617±0.159**
3D U-Net[32]	LIDC-IDRI	64 x 64 x 64	0.720±0.049	0.380±0.080
iW-Net[21]	LIDC-IDRI	64 x 64 x 64	-	0.550±0.140
PN-SAMP[22]	LIDC-IDRI	64 x 64 x 64	0.741±0.357	-

This article is protected by copyright. All rights reserved.

Table 3. The root-mean-square deviation (RMSD) of the descriptors of nodule characteristics was provided by different radiologists in the test set. The ratings of the descriptors were given on a 5-point scale except for the diameter.

	Diameter (mm)	Subtlety	Malignancy	Margin	Lobulated	Spiculated	Solid	Sphericity
RMSD	1.725	0.659	0.787	0.706	0.79	0.709	0.51	0.691

Table 1. Test results achieved by H-DL model, Shallow U-DL and Deep U-DL alone, and the performance of LIDC radiologists' manual segmentations relative to the reference standard. The *P*-values of the differences between H-DL and others were calculated by paired *t*-test. The *P*-value < 0.05 indicated that the difference was statistically significant. M-DICE and M-JI are the averages of the median, and m-DICE and m-JI are the averages of the minimum DICE and JI by radiologists, respectively.

		H-DL		Radiologist				Shallow U-DL		Deep U-DL	
Categories		DICE	JI	M-DICE	min-DICE	M-JI	min-JI	DICE	JI	DICE	JI
All nodules	N = 327	0.750 ± 0.135	0.617 ± 0.159	0.778 ± 0.102	0.685 ± 0.139	0.651 ± 0.127	0.537 ± 0.153	0.745 ± 0.139	0.611 ± 0.161	0.739 ± 0.145	0.604 ± 0.163
	P-value			**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**
Diameter (mm)	≥ 10.00 (N = 208)	0.782 ± 0.128	0.658 ± 0.152	0.790 ± 0.105	0.704 ± 0.135	0.667 ± 0.128	0.559 ± 0.150	0.777 ± 0.129	0.652 ± 0.154	0.766 ± 0.150	0.640 ± 0.166
	P-value			0.423	**< 0.05**	0.434	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**
	< 10.00 (N = 119)	0.695 ± 0.131	0.547 ± 0.145	0.758 ± 0.093	0.652 ± 0.139	0.623 ± 0.121	0.499 ± 0.153	0.688 ± 0.137	0.540 ± 0.149	0.691 ± 0.125	0.541 ± 0.137
	P-value			**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	0.651	0.498
Subtlety	≥ 4.25 (N = 163)	0.798 ± 0.113	0.677 ± 0.140	0.785 ± 0.116	0.701 ± 0.148	0.662 ± 0.139	0.558 ± 0.163	0.796 ± 0.111	0.674 ± 0.138	0.785 ± 0.125	0.662 ± 0.149
	P-value			0.210	**< 0.05**	0.222	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**
	< 4.25 (N = 164)	0.702 ± 0.139	0.557 ± 0.154	0.771 ± 0.086	0.670 ± 0.128	0.640 ± 0.112	0.517 ± 0.140	0.694 ± 0.145	0.549 ± 0.159	0.693 ± 0.150	0.548 ± 0.157
	P-value			**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**0.197**	**0.173**
Malignancy	≥ 3.00 (N = 203)	0.783 ± 0.123	0.658 ± 0.149	0.785 ± 0.105	0.700 ± 0.134	0.660 ± 0.127	0.553 ± 0.147	0.778 ± 0.124	0.652 ± 0.148	0.773 ± 0.135	0.647 ± 0.155
	P-value			0.842	**< 0.05**	0.855	**< 0.05**	**< 0.05**	**< 0.05**	0.078	0.058

This article is protected by copyright. All rights reserved.

	< 3.00 (N = 124)	0.697 ± 0.139	0.551 ± 0.157	0.767 ± 0.095	0.661 ± 0.145	0.636 ± 0.126	0.511 ± 0.160	0.690 ± 0.145	0.544 ± 0.16	0.683 ± 0.145	0.535 ± 0.152
	P-value			**< 0.05**	0.056	**< 0.05**	0.051	**< 0.05**	**< 0.05**	0.100	**< 0.05**
Margin	≥ 4.25 (N =175)	0.774 ± 0.126	0.647 ± 0.154	0.795 ± 0.094	0.710 ± 0.137	0.672 ± 0.123	0.567 ± 0.155	0.767 ± 0.135	0.640 ± 0.160	0.756 ± 0.149	0.627 ± 0.166
	P-value			0.070	**< 0.05**	0.091	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**
	< 4.25 (N =152)	0.723 ± 0.140	0.582 ± 0.157	0.759 ± 0.107	0.657 ± 0.136	0.627 ± 0.128	0.503 ± 0.144	0.719 ± 0.140	0.578 ± 0.156	0.720 ± 0.139	0.579 ± 0.156
	P-value			**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	0.667	0.584
Lobulated	≥ 1.75 (N=165)	0.776 ± 0.120	0.648 ± 0.147	0.775 ± 0.113	0.682 ± 0.151	0.648 ± 0.136	0.536 ± 0.162	0.771 ± 0.121	0.642 ± 0.148	0.767 ± 0.131	0.638 ± 0.154
	P-value			0.958	**< 0.05**	0.978	**< 0.05**	**< 0.05**	**< 0.05**	0.136	0.097
	< 1.75 (N=162)	0.724 ± 0.145	0.586 ± 0.164	0.782 ± 0.089	0.689 ± 0.126	0.655 ± 0.117	0.539 ± 0.144	0.718 ± 0.151	0.579 ± 0.168	0.710 ± 0.154	0.570 ± 0.166
	P-value			**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**0.058**	**< 0.05**
Spiculated	≥ 1.50 (N = 173)	0.767 ± 0.125	0.638 ± 0.150	0.774 ± 0.112	0.680± 0.142	0.648 ± 0.134	0.532 ± 0.153	0.764 ± 0.125	0.633 ± 0.149	0.755 ± 0.145	0.624 ± 0.162
	P-value			0.515	**< 0.05**	0.443	**< 0.05**	0.767	0.739	0.250	0.309
	< 1.50 (N = 154)	0.731 ± 0.143	0.596 ± 0.164	0.783 ± 0.089	0.691 ± 0.136	0.655 ± 0.119	0.543 ± 0.154	0.723 ± 0.151	0.586 ± 0.17	0.721 ± 0.145	0.582 ± 0.162
	P-value			**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**0.587**	**0.608**	**0.494**	**0.431**
Solid	≥ 5.00 (N = 167)	0.776 ± 0.125	0.649 ± 0.151	0.794 ± 0.099	0.709 ± 0.135	0.672 ± 0.126	0.564 ± 0.154	0.771 ± 0.131	0.644 ± 0.156	0.759 ± 0.150	0.631 ± 0.166
	P-value			0.092	**< 0.05**	0.084	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**
	< 5.00 (N = 160)	0.723 ± 0.141	0.584 ± 0.160	0.762 ± 0.103	0.661 ± 0.139	0.629 ± 0.125	0.509 ± 0.147	0.717 ± 0.142	0.577 ± 0.160	0.718 ± 0.138	0.577 ± 0.156
	P-value			**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	0.630	0.457	0.685	0.447
Sphericity	≥ 4.00 (N = 179)	0.760 ± 0.139	0.632 ± 0.162	0.796 ± 0.085	0.706 ± 0.129	0.673 ± 0.112	0.561 ± 0.148	0.756 ± 0.144	0.626 ± 0.166	0.754 ± 0.143	0.623 ± 0.163
	P-value			**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**0.212**	**0.129**
	< 4.00 (N = 148)	0.737 ± 0.129	0.600± 0.153	0.757 ± 0.116	0.660 ± 0.146	0.624 ± 0.139	0.509 ± 0.155	0.732 ± 0.131	0.593 ± 0.154	0.721 ± 0.147	0.582 ± 0.162
	P-value			0.144	**< 0.05**	0.058	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**	**< 0.05**

This article is protected by copyright. All rights reserved.

Table 2. The performance comparisons of our H-DL model with other deep learning methods in the segmentation of lung nodules.

	Dataset used	VOI size	DICE	JI
H-DL model	LIDC-IDRI	64 x 64 x 64	**0.750±0.135**	**0.617±0.159**
3D U-Net[32]	LIDC-IDRI	64 x 64 x 64	0.720±0.049	0.380±0.080
iW-Net[21]	LIDC-IDRI	64 x 64 x 64	-	0.550±0.140
PN-SAMP[22]	LIDC-IDRI	64 x 64 x 64	0.741±0.357	-

Table 3. The root-mean-square deviation (RMSD) of the descriptors of nodule characteristics was provided by different radiologists in the test set. The ratings of the descriptors were given on a 5-point scale except for the diameter.

	Diameter (mm)	Subtlety	Malignancy	Margin	Lobulated	Spiculated	Solid	Sphericity
RMSD	1.725	0.659	0.787	0.706	0.79	0.709	0.51	0.691

This article is protected by copyright. All rights reserved.