EDICNet: An end-to-end detection and interpretable malignancy classification network for pulmonary nodules in computed tomography

Yannan Lin [1,2], Leihao Wei [2,4], Simon X. Han [1,2], Denise R. Aberle [1,2,3], William Hsu [1,2,3]

[1] Department of Bioengineering, University of California, Los Angeles, CA, USA
[2] Medical & Imaging Informatics Group, Department of Radiological Sciences, University of California, Los Angeles, CA, USA
[3] Department of Radiological Sciences, David Geffen School of Medicine at University of California, Los Angeles, CA, USA
[4] Department of Electrical and Computer Engineering, University of California, Los Angeles, CA, USA

ABSTRACT

We present an interpretable end-to-end computer-aided detection and diagnosis tool for pulmonary nodules on computed tomography (CT) using deep learning-based methods. The proposed network consists of a nodule detector and a nodule malignancy classifier. We used RetinaNet to train a nodule detector using 7,607 slices containing 4,234 nodule annotations and validated it using 2,323 slices containing 1,454 nodule annotations drawn from the LIDC-IDRI dataset. The average precision for the nodule class in the validation set reached 0.24 at an intersection over union (IoU) of 0.5. The trained nodule detector was externally validated using a UCLA dataset. We then used a hierarchical semantic convolutional neural network (HSCNN) to classify whether a nodule was benign or malignant and generate semantic (radiologist-interpretable) features (e.g., mean diameter, consistency, margin), training the model on 149 cases with diagnostic CTs collected from the same UCLA dataset. A total of 149 nodule-centered patches from the UCLA dataset were used to train the HSCNN. Using 5-fold cross validation and data augmentation, the mean AUC and mean accuracy in the validation set for predicting nodule malignancy achieved 0.89 and 0.74, respectively. Meanwhile, the mean accuracy for predicting nodule mean diameter, consistency, and margin were 0.59, 0.74, and 0.75, respectively. We have developed an initial end-to-end pipeline that automatically detects nodules ≥ 5 mm on CT studies and labels identified nodules with radiologist-interpreted features automatically.

Keywords: computer-aided diagnosis; computed tomography; pulmonary nodule detection; pulmonary nodule classification; deep learning.

1. INTRODUCTION

Lung cancer is the leading cause of cancer-related death in men and the second leading cause of cancer-related death among women globally [1]. Computed tomography (CT) is an essential clinical imaging procedure for lung cancer detection and diagnosis. The National Lung Screening Trial (NLST) demonstrated that screening with low-dose CT (LDCT) reduced lung cancer mortality [2], leading to the United States Preventive Services Task Force (USPSTF) recommendation that high-risk populations be routinely screened [3]. While patients will benefit from earlier diagnosis of lung cancer through screening, the potential harms cannot be neglected [4]. One notable downside of screening is the high number of false positives, such as when a pulmonary nodule detected on LDCT turns out to be a benign lesion upon biopsy. Moreover, many cancerous nodules are also detected incidentally on diagnostic CTs in a non-screening setting. To help reduce unnecessary invasive procedures, we aim to build a computer-aided detection and diagnosis tool for pulmonary nodules in CTs.

We propose an End-to-end Detection and Interpretable Classification Network (EDICNet) for pulmonary nodules. A similar end-to-end network has been proposed recently, achieving performance that rivals six radiologists [5]; however, their classification network focused on moderately homogeneous low-dose CT scans and did not incorporate clinical information such as the semantic features of pulmonary nodules. Semantic features are important because radiologists often use them to help determine a nodule's suspicion level for malignancy, informing follow-up recommendations based on guidelines such as Fleischner [6] and Lung-RADS [7]. We extend our prior work on the hierarchical semantic convolutional neural network (HSCNN) [8] to interpret semantic features from deep features in the hopes of making the network more human understandable.

2. METHODS

2.1 Overall pipeline

EDICNet is comprised of two modules: 1) a nodule detection module and 2) a nodule malignancy classification module (the first row of **Figure 1**). Two independent datasets were used to build EDICNet, the Lung Image Database Consortium image collection (LIDC-IDRI) [9] and an annotated diagnostic CT dataset from our institution. The LIDC-IDRI dataset has pixel-level annotations for all nodules, whereas the UCLA dataset provides pixel-level annotations for one primary lesion in each CT scan, clinical semantic labels of the primary lesion, as well as the diagnosis of the primary lesion (benign or malignant). We utilized these two datasets to build our two-stage pipeline. First, CT slices with pixel-level annotations of nodules from the LIDC-IDRI dataset were used to train the nodule detection model. Second, a set of nodule-centered patches together with semantic and diagnostic labels from the UCLA dataset were used to train the nodule malignancy classification model. During the inference phase (the second row of **Figure 1**), the nodule detector generates regions of interest (ROIs) with the highest probability to be a pulmonary nodule given the full CT volume as input. For each proposed ROI, the nodule malignancy classification model is used to generate predictions of cancer/non-cancer along with a set of semantic features.

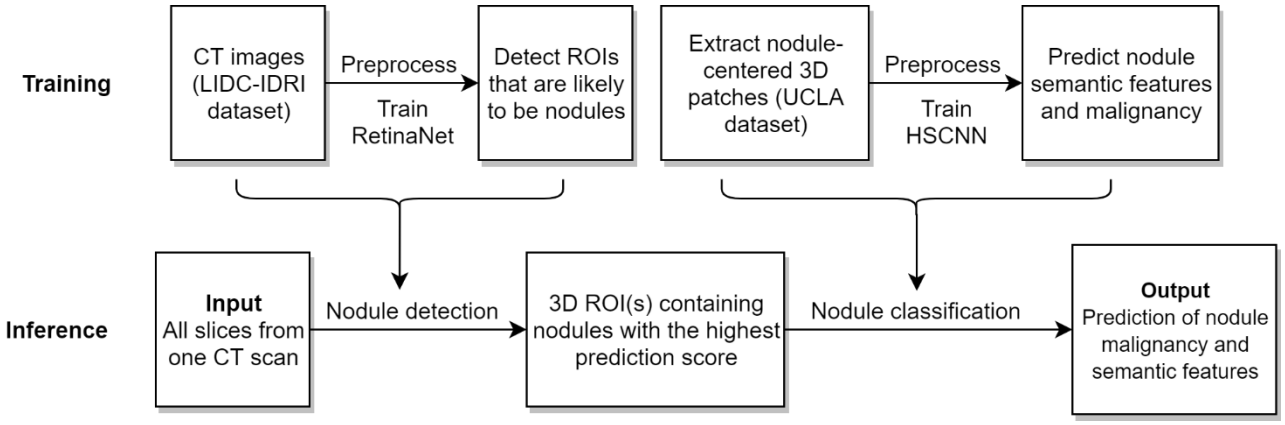

Figure 1. The training and inference pipelines for End-to-end Detection and Interpretable Classification Network (EDICNet). CT: computed tomography; 3D: 3-dimensional; HSCNN: hierarchical semantic convolutional neural network; ROI: region of interest.

2.2 Datasets

2.2.1 LIDC-IDRI dataset

The LIDC-IDRI dataset is a publicly available dataset that consists of 1,018 diagnostic and lung cancer screening thoracic CT scans with marked-up annotations created by up to four radiologists per lesion [9]. Among the 1,018 scans, we excluded those with a slice thickness greater than or equal to 3 mm, resulting in 897 CT scans. Among the annotated nodules from the 897 scans, we included nodules that were annotated by at least three radiologists. We used the union of the contours made by all the annotators as the final contour for a nodule, thus in our study each nodule on a specific slice was only included once. If a nodule was identified on multiple slices, all annotations on these slices were included. For one scan,

slices that contained at least one qualified nodule (i.e., slice thickness < 3 mm and annotated by at least three radiologists) were included, and the largest dimension of the nodule on the slice must be equal to or greater than 5 mm (e.g. a nodule was detected on six slices and on two of the slices the largest dimension was less than 5 mm, so the two slices were excluded). In total, we had 5,445 slices with nodules. To balance the dataset with a comparable number of negative samples, for each scan, we randomly sampled five slices without nodules, adding another 4,485 slices. In total, 9,930 CT slices were used to train and validate the nodule detector. **Figure 2** shows the inclusion of CT slices from the LIDC-IDRI dataset.

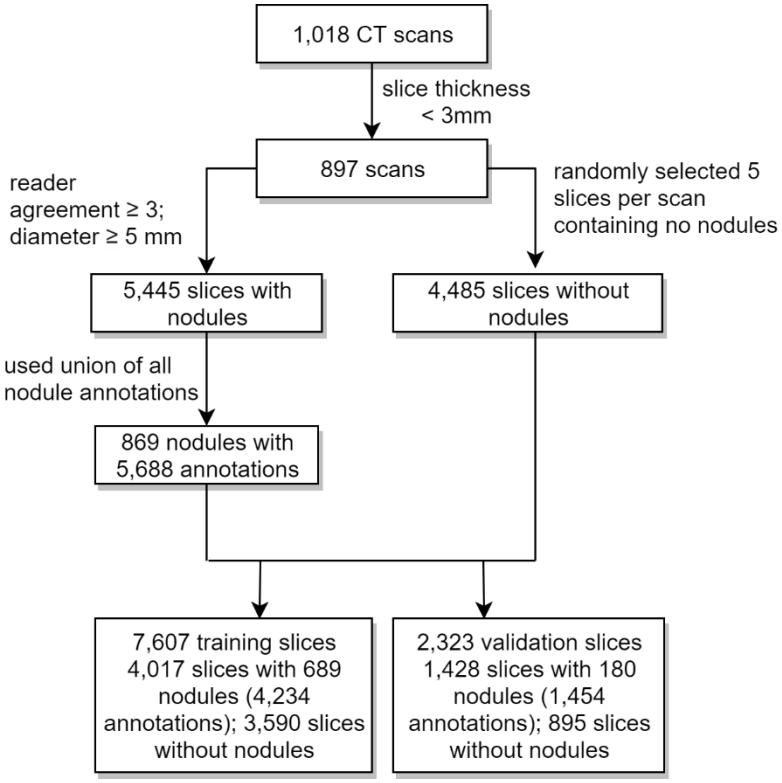

Figure 2. Inclusion of CT slices from the LIDC-IDRI dataset. CT: computed tomography.

2.2.2 UCLA dataset

The UCLA dataset contains 149 diagnostic CT scans collected between 2003 and 2014. There was only one predetermined primary lesion for each scan as identified by a thoracic radiologist (DRA), and this lesion was determined as the most critical lesion to a patient's disease progress. The mask of this lesion was provided by a board-certified thoracic radiologist with more than 10 years of experience and verified by a board-certified thoracic radiologist (DRA) with more than 30 years of experience. The demographic and clinical information of the included scans is shown in **Table 1**. Three semantic features (mean diameter, consistency, and margin) and diagnosis were used as labels for classification. The semantic features were also annotated and validated by the aforementioned two board-certified thoracic radiologists.

2.3 Data preprocessing

2.3.1 RetinaNet

Prior to training, all slices were normalized from the Hounsfield (HU) scale of (-1000 HU, 500 HU) to a range of (0, 1). CT slices were input into RetinaNet using their original dimensions (i.e., 512 by 512 pixels). The coordinates of the upper left and bottom right vertices of the bounding box for a nodule were required inputs for training the nodule detector. We

used the minimum and maximum values on x- and y-axes from the union of a nodule's annotation mentioned in 2.2.1 as the two index vertices for generating the bounding box around the nodule (i.e., upper left was [x_min, y_min] and bottom right was [x_max, y_max]).

Table 1. Demographic and clinical information of the UCLA cohort, 2003-2014 (N=149).

Variable	Category	No. (%)
Age at scan (years)	<50	4 (3)
	50-70	84 (56)
	>70	61 (41)
Sex	Female	90 (60)
	Male	59 (40)
Smoking status	Ever smoker	118 (79)
	Never smoker	31 (21)
Lesion mean diameter (mm)	≤10	53 (36)
	>10 and ≤20	72 (48)
	>20	24 (16)
Lesion consistency	Solid	97 (65)
	Part-solid	38 (26)
	Pure ground-glass (non-solid)	14 (9)
Lesion margin	Smooth	36 (24)
	Lobulated	49 (33)
	Serrated/Spiculated	47 (32)
	Poorly defined	17 (11)
Lesion diagnosis	Benign	52 (35)
	Malignant	97 (65)
Mode of diagnosis	Clinical	31 (21)
	FNA	74 (49)
	Surgery	42 (28)
	FOB	1 (1)
	Pathology	1 (1)

FNA: fine needle aspiration, FOB: fiberoptic bronchoscopy.

2.3.2 HSCNN

The same data preprocessing method was applied to the raw DICOM images in the UCLA dataset before extracting the nodule-centered patches, including HU transformation and normalization (mentioned in 2.3.1). We resampled x, y spacing and slice thickness to 1 mm for all CT scans. After that, we applied a mask onto the nodule in order to separate it from the background, which could be part of the chest wall, great vessels, or the heart. Then, the non-nodule regions were assigned a value of zero. We posited that this process would eliminate potential noise from the background, forcing the network to learn features only from the nodule. A 52 x 52 x 52 nodule-centered patch was generated from each CT scan. We augmented the data through flipping along the x-, y-, and z-axes before training.

The coding of the malignancy and semantic feature prediction tasks are shown in **Table 2**. Given our small sample size, we chose to binarize consistency and margin. For consistency, solid nodules formed one group whereas part-solid and pure ground-glass nodules were grouped together. For margin, smooth margins were in a single group whereas lobulated, serrated/spiculated, and poorly defined margins were categorized as a separate group.

Table 2. Summary of coding for HSCNN.

Label	Category	Coding
Mean diameter (mm)	≤10	0
	> 10 and ≤ 20	1
	> 20	2
Consistency	Solid	1
	Part-solid	0
	Pure ground-glass (non-solid)	0
Margin	Smooth	0
	Lobulated	1
	Serrated/Spiculated	1
	Poorly defined	1
Diagnosis	Benign	0
	Malignant	1

HSCNN: hierarchical semantic convolutional neural network.

2.4 Modeling approaches

2.4.1 Nodule detection

We used a 2D RetinaNet [10] with a ResNet 34 [11] backbone for the nodule detection model. Grayscale 2-dimensional CT slices were converted to RGB images by copying the same image across all three RGB channels. Then the slices were rescaled from 512 x 512 pixels to 608 x 608 pixels and padded to 640 x 640 pixels with zeros. Thus, the final input dimensions of the slices were 640 x 640 pixels. The model training was initialized using the pre-trained weights from the ImageNet [12]. Real-time data augmentation was used, an input slice might be flipped along the y-axis depending on a random process. We did not modify the overall structure of the RetinaNet (**Figure 3**) and we used the Adam optimizer. The batch size was 2 during training and 1 during validation. The learning rate was set to 1e-5.

2.4.2 Nodule malignancy classification

Figure 4 shows the network structure for our multi-task nodule malignancy classification model. The high-level task was to determine whether the target lesion is benign or malignant. The low-level tasks were to predict the three semantic features. We used the same parameters described in the HSCNN paper [8] but instead of assigning varying weighting hyperparameters to low-level tasks (i.e., in the HSCNN paper the weighting hyperparameters was 0.1 for calcification, margin, and texture, and 0.2 for sphericity and subtlety), we applied roughly equal weighting hyperparameters to the three semantic features when calculating the total loss (0.33, 0.34, and 0.33). We set the dropout rate to be 0.8 instead of 0.6 comparing to the original HSCNN model to further reduce overfitting. The batch size was 6 during training and 1 during validation. The learning rate was 1e-3.

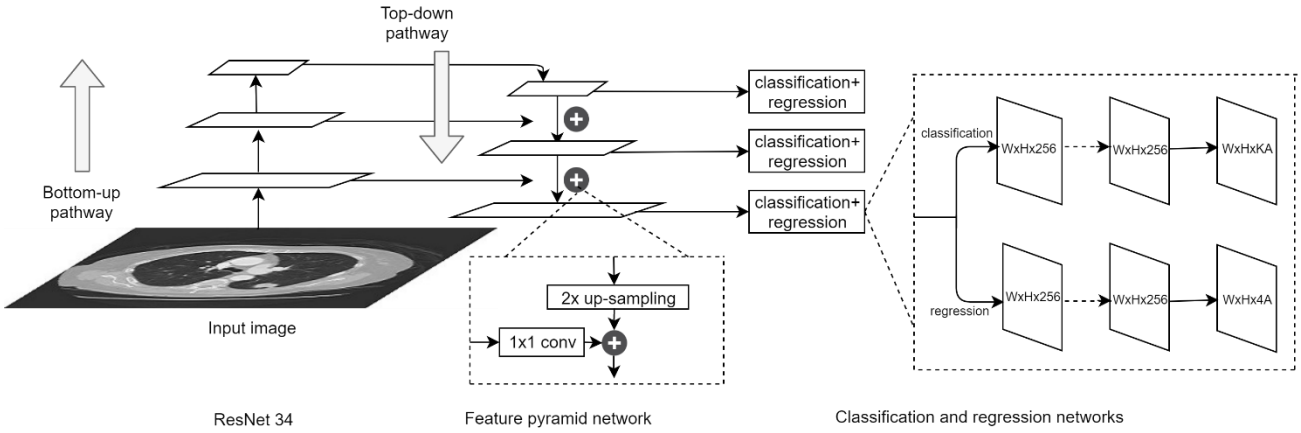

Figure 3. The RetinaNet network architecture. ResNet: residual network.

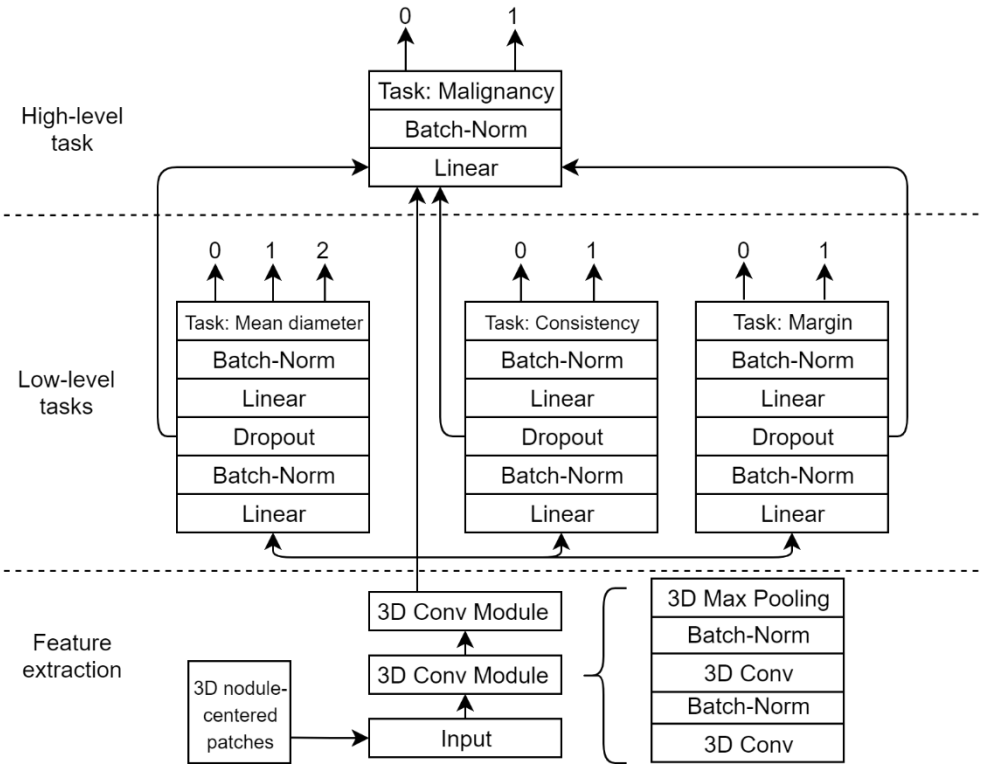

Figure 4. The network structure of the hierarchical semantic convolutional neural network (HSCNN).

2.5 Evaluation metrics

2.5.1 RetinaNet

Average precision (AP), which is the area under the precision-recall curve, is the standard metric for evaluating the performance of an object detection model [13]. In our case, the nodule detector only labels one class (‘nodule’) per image, (hence, we use AP rather than mAP). During validation, the model outputs a set of bounding boxes that are called anchors and their corresponding scores. The bounding box is an ROI that the model has identified as containing a nodule. The score associated with each bounding box, ranging from 0 to 1, conveys the probability that the ROI contains a nodule.

The reference bounding box is defined based on radiologist-provided annotations, as described in 2.2.1. The intersection over union (IoU) is defined as the intersection between the reference bounding box and an output bounding box divided by the union of the two. We set a threshold for IoU (range of the threshold (0,1]). If the IoU between the reference bounding box and the output bounding box is greater or equal to the IoU threshold, then this detection is a true positive (TP) detection; otherwise it is considered as a false positive (FP) detection (i.e., IoU < 0.5, no overlap, duplicated output bounding box on one reference bounding box). Therefore, we are able to know whether each detection is a TP or FP. Precision is calculated through TP/(TP+FP) where TP+FP equals the total number of output bounding boxes. Recall is calculated through TP/total number of reference bounding boxes.

After examining all images in the validation set, all output bounding boxes will be ranked from high to low based on their values of the scores. Since TP or FP is known, precisions and recalls can be calculated for the ranked list of detections, later being used to draw a precision-recall curve (PR curve). AP is estimated by calculating the area under the PR curve. In our experiments, we obtained AP at a predefined IoU threshold, 0.5.

2.5.2 HSCNN

Five-fold cross validation was used when training the HSCNN. The data was split into five folds based on the diagnosis label, preserving the percentage of samples for the benign nodule class and the malignant nodule class (stratified 5-fold). For each training, one out of the five folds was the validation set while the remaining four folds served as the training set. Additionally, the preprocessed augmented patches (flipped along three axes) were added to the training set. Mean area under the receiver operating characteristic curve (AUC) and mean accuracy were calculated for each task.

2.6 External evaluation for RetinaNet

Since the UCLA dataset and the LIDC-IDRI dataset are separate datasets, we used the UCLA dataset to externally test the performance of the RetinaNet. However, given that only a single primary nodule was annotated on each scan in the UCLA dataset, we could not try to detect all nodules and calculate AP because we would not have the reference bounding boxes for all nodules. Therefore, we were only able to select one primary detected bounding box for each scan to make comparisons. For each of the 149 diagnostic CTs, the nodule detector performed detection on every slice and only the ROI (or ROIs) with the highest prediction score was (were) retained. Each retained ROI was manually compared to the primary lesion identified by the radiologist. This manual process was done by a research assistant with three years of experience in identifying pulmonary nodules in CT images (YL). The total number of identified nodules with the highest prediction score and the number of those matched with the primary lesions are reported in section 3.2.

2.7 Implementation details

The EDICNet was implemented in Python 3.5 using PyTorch (version 0.4.1) [14]. The RetinaNet model was trained on an Amazon Web Service (AWS) g3s.xlarge instance with an NVIDIA Tesla M60 GPU with 8GB onboard memory. Training 50 epochs took 20 hours. The HSCNN model was trained on the same AWS server. The training took approximately 2 hours for 200 epochs for a single fold. We used a single NVIDIA Tesla V100 GPU from an NVIDIA DGX-1 server with 32GB onboard GPU memory to test the nodule detector using the UCLA dataset.

3. RESULTS

3.1 Model training

When training the RetinaNet, a total of 7,607 slices (containing 689 nodules with 4,234 nodule annotations) were used for training and the rest of the 2,323 slices (containing 180 nodules with 1,454 nodule annotations) for validation. The AP was 0.24 after 50 epochs (190,150 iterations). **Figure 5** provides examples of detection results from the validation set.

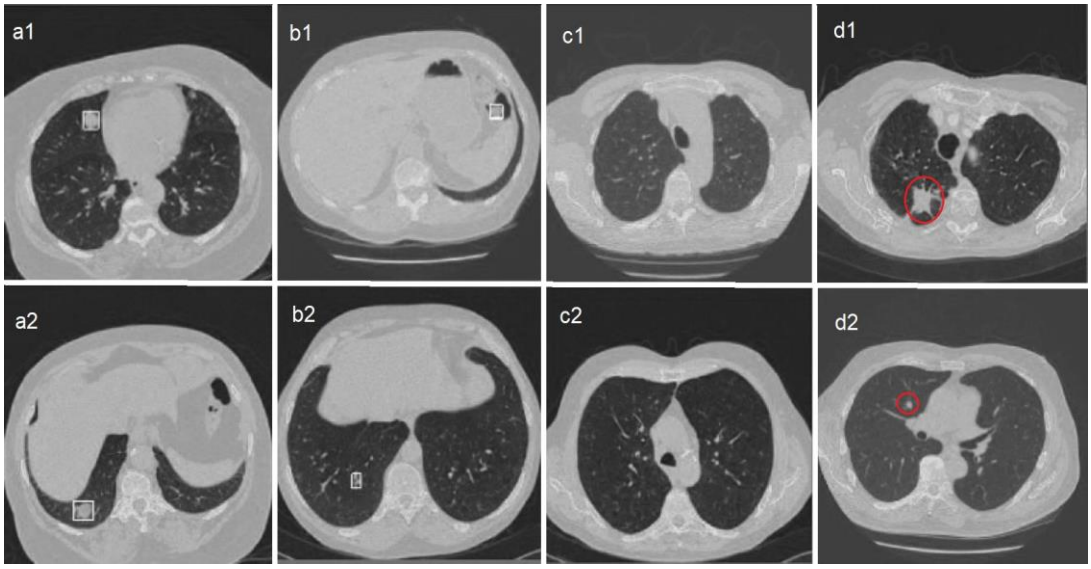

Figure 5. Visualization of nodule detection results in the validation set of LIDC-IDRI with a prediction score greater than 0.5. A positive slice is a slice of CT image with at least one nodule whereas a negative slice is a slice without nodule. TP is defined as the network correctly detecting the nodule on a positive slice; FP is defined as the network detecting a non-nodule structure on a positive or negative slice; TN is defined as the network detecting nothing on a negative slice; FN is defined as the network failing to detect any nodule on a positive slice. The first (a1 and a2), second (b1 and b2), third (c1 and c2), and fourth (d1 and d2) columns show TP, FP, TN, and FN examples, respectively. The white rectangles are bounding boxes generated by the nodule detector and the red circles denote nodules that the nodule detector failed to detect. Note that the definitions for TP, FP, TN, and FN used here are different from the definitions used to calculate the AP. TP: true positive, FP: false positive, TN: true negative, FN: false negative; AP: average precision.

For the HSCNN, we used a total of 149 nodule-centered patches with data augmentation. We trained 200 epochs for each fold. The mean AUC for the high-level task achieved 0.89. The mean accuracy (prediction threshold at 0.5) for the high-level task achieved 0.74. For the three semantic low-level tasks, the mean accuracy for mean diameter, consistency, and margin were 0.59, 0.74, and 0.75 respectively. **Table 3** summarizes details of AUCs and accuracies in each fold.

Table 3. Summary of AUCs and accuracies in each fold for all tasks.

Label	Metric	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean
Mean diameter	Accuracy	0.77	0.71	0.76	0.41	0.31	0.59
Consistency	Accuracy	0.84	0.74	0.83	0.76	0.52	0.74
Margin	Accuracy	0.74	0.81	0.72	0.69	0.79	0.75
Diagnosis	Accuracy	0.74	0.68	0.76	0.76	0.76	0.74
	AUC	0.88	0.98	0.85	0.82	0.89	0.89

AUC: area under the receiver operating characteristic curve.

3.2 External validation for RetinaNet

The trained nodule detector was validated on 149 UCLA scans. As described in section 2.6, for each scan only detections with the highest prediction score were included. Among the 149 scans, the nodule detector detected one ROI for 148 scans and did not detect any ROI for one scan. Of the 148 detected ROIs, 59 were primary nodules identified by the radiologist, 15 were other nodules that were not identified by the radiologist, and the remaining 74 ROIs were not pulmonary nodules. Most of the non-nodule detections were either bright bone structures with higher intensity than the background pixels or airway and vessel structures that looked like nodules on a single slice. **Figure 6** shows some of the detections from the UCLA dataset.

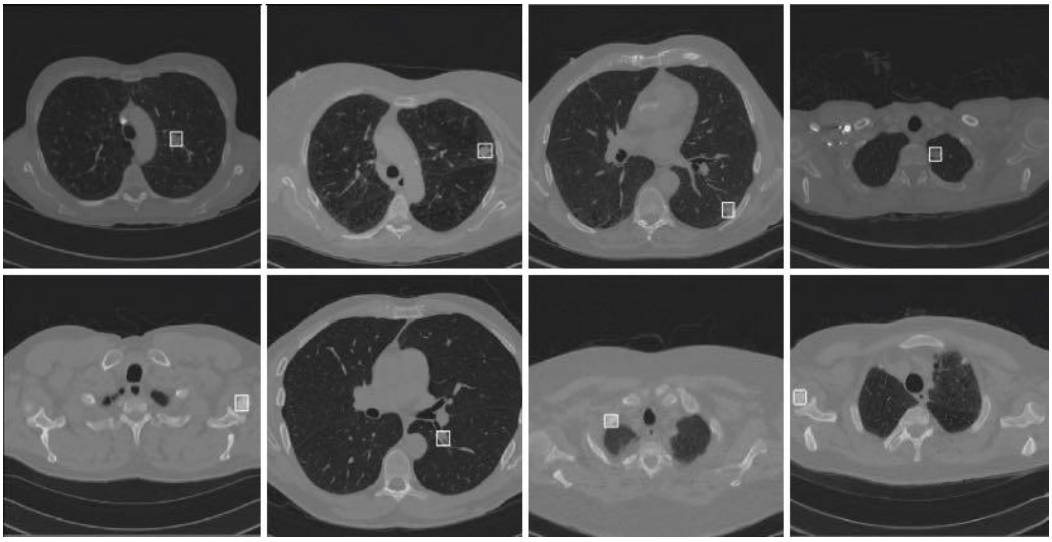

Figure 6. Visualization of detected ROIs in the UCLA dataset using the trained nodule detector. The bounding boxes in the first row show four pulmonary nodules in CT scans, including primary nodules identified by the radiologist and nodules that were not annotated by the radiologist. The bottom row shows four ROIs that are not pulmonary nodules. They are either bright bone structures or airway/vessel structures that looked like nodules. ROI: region of interest, CT: computed tomography.

4. DISCUSSION

We presented an end-to-end pipeline consisting of two modules: a nodule detection module based on RetinaNet and a malignancy classification module based on HSCNN. Our work has the following contributions:

1) RetinaNet can be adapted for pulmonary nodule detection in CT but requires refinement (i.e., lung segmentation) to improve true positive detections.
2) Along with pulmonary nodule malignancy classification result, semantic features can also be generated simultaneously to help radiologists better understand the model's prediction.
3) Our pipeline provides a tool for radiologists to automate the identification of the most important lesion in a CT scan and generates semantic features that could be used as the basis of a radiology report.

Pulmonary nodule detection and classification have been widely investigated given the availability of publicly available large datasets (i.e., LIDC-IDRI, NLST). With the application of convolutional neural networks (CNNs) to medical images, faster networks for nodule detection have been proposed and refined resulting in higher sensitivity and lower false positive rate per slice [5, 15-19]. A prior study that utilized RetinaNet [15] using patches from the LIDC-IDRI dataset reported AP at an IoU of 0.1. They argued that setting the threshold at 0.1 "respects the clinical need for coarse localization". We also investigated the impact of IoU on AP by setting the IoU threshold to 0.1. In our experiment, AP increased from 0.24 (IoU=0.5) to 0.27 after 25 epochs (95,075 iterations). But for our nodule detector module, we used a higher IoU threshold at 0.5 because we would like to be more stringent about the definition of true positives. Moreover, while the nodule detector is able to detect the primary nodule on roughly 40% (59/149) of the scans in the external validation set, the detected ROIs on almost half (74/149) of the CT scans are non-nodule structures. This suggests that without lung segmentation the model is likely to be affected by bright bone structures, and without learning from consecutive slices the model is unable to differentiate nodule-like structures from real nodules.

Malignancy classification task of pulmonary nodules has also benefited from the expansion of deep learning-based methods and has demonstrated promising results [5, 8, 19-24]. However, one critical issue of the nodule classification task is the difficulty to obtain datasets with biopsy- or clinically-proven diagnostic labels for lesions. The majority of the published studies used public datasets that lack appropriate diagnostic labels. The LIDC-IDRI dataset only provides the

likelihood of malignancy serving as the proxy for truth, whereas the NLST dataset only provides a patient-level diagnosis. The UCLA dataset used in our study has lesion-level biopsy- or clinically-proven diagnosis related to lung cancer, thus, our model can show more clinical relevance compared to other models that used datasets with weaker diagnostic labels.

We also investigated the effect of resampling the CT slices along the x- and y-axes prior to training the nodule detector, making both x and y spacing 1 mm. However, the performance of the detector model with resampled training data did not perform as well as without resampling. One possible explanation is that the nodule became smaller on resampled slices as the dimensions of the slices ranged from 236 x 236 pixels to 500 x 500 pixels after resampling (originally 512 x 512 pixels), which increased the difficulty for the detector to locate nodules in general. Nevertheless, we resampled the CT scans from the UCLA dataset to [1, 1, 1] mm to train the HSCNN model because when the HSCNN model was interpreting both high-level (malignant/benign) and low-level (semantic features) tasks, we wanted to provide the network with as much information about the nodule as possible. The semantic labels in the UCLA dataset were multi-class labels when they were annotated. We grouped some of them to avoid sparse data issue in some subclasses; for instance, there was only 14 out of 149 pure ground-glass nodules. The performance of the mean diameter task fell behind other tasks, which suggests that the 3D volume of a nodule could not be well represented by the nodule mean diameter obtained from a 2D CT image plane.

Our approach has limitations that are being addressed in ongoing work. First, the nodule detector was not trained under the assumption that the same nodule appearing on multiple slices was one nodule. For radiologists, nodule detection is a 3D task. We intend to modify RetinaNet to support 3D volumes as inputs, incorporating nodule information between slices to further reduce false positives. Second, we did not perform lung segmentation prior to inputting the CT image into the nodule detector. Error analysis of our current implementation revealed a number of false positive findings (non-lung structures) such as bright bone structures in the chest wall that were detected as nodules. We believe these false positives will be reduced with appropriate preprocessing. Third, our external validation of the nodule detector was limited to a dataset in which only a single primary nodule was annotated by the radiologist per scan. We assumed that the detected ROI from one scan with the highest prediction score was the one that the radiologist deemed as the most critical pulmonary lesion for a patient. Fourth, the nodule detector was trained on both low-dose and diagnostic CT scans and was externally validated using diagnostic CT scans. Evaluating the nodule detector on LDCT scans are necessary for future work and may demonstrate improved performance. Finally, given our interest in incorporating semantic features into our model, we trained the HSCNN with a small dataset from a single institution. These features were only annotated by a single radiologist, and we have yet to perform an external validation.

5. CONCLUSIONS

EDICNet is an end-to-end approach for localizing pulmonary nodules on CT studies, predicting whether the nodule is benign or malignant, and generating semantic labels for each nodule that help radiologists interpret the results. Our objective is to develop not only a reliable approach for nodule detection and classification but also provide insights about why the model believes a nodule is benign or malignant through semantic features. While the performance of our model has underperformed other reported models, we believe that additional improvements to our approach and the use of larger, more diverse datasets could overcome some of the reported issues.

6. ACKNOWLEDGEMENTS

This work was supported in part by the National Institutes of Health [National Cancer Institute under awards **R01CA210360** to YL, DRA, WH and **R01CA226079** to DRA, WH, SXH]; the National Science Foundation [#**1722516** to WH, LW]; Amazon Web Services and UCLA Department of Computational Medicine partnership to WH; and the Integrated Diagnostics Program, jointly funded by the Department of Radiological Sciences and Pathology & Laboratory Medicine to DRA and WH.

7. CODE AVAILABILITY

The code for RetinaNet was adapted from https://github.com/yhenon/pytorch-retinanet. The HSCNN was reimplemented in PyTorch based on the Keras code provided by the original paper [8].

8. REFERENCES

[1] Torre LA, Siegel RL, Jemal A. Lung cancer statistics. *Adv Exp Med Biol*. 2016;893:1-19.

[2] Team, N. L. S. T. R. , et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med*. 2011;365(5):395-409.

[3] Humphrey L, Deffebach M, Pappas M et al. Screening for lung cancer with low-dose computed tomography: A Systematic Review to Update the U.S. Preventive Services Task Force Recommendation. *Ann Intern Med*. 2013;159(6):411.

[4] Bach P, Mirkin J, Oliver T et al. Benefits and harms of CT screening for lung cancer. *JAMA*. 2012;307(22):2418.

[5] Ardila D, Kiraly A, Bharadwaj S et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat Med*. 2019;25(6):954-961.

[6] MacMahon H, Naidich D, Goo J et al. Guidelines for management of incidental pulmonary nodules detected on CT images: from the Fleischner society 2017. *Radiology*. 2017;284(1):228-243.

[7] LungRADS v 1.0 04 28 14.xlsx. https://www.acr.org/-/media/ACR/Files/RADS/Lung RADS/LungRADS_AssessmentCategories.pdf. Published 2019. Accessed August 22, 2019.

[8] Shen S, Han S, Aberle DR et al. An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. *Expert Syst Appl*. 2019;128:84-95.

[9] Armato S, McLennan G, McNitt-Gray M et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed public database of CT scans for lung nodule analysis. *Med Phys*. 2010;37(6Part6):3416-3417.

[10] Lin T, Goyal P, Girshick R et al. Focal loss for dense object detection. *IEEE ICC.* 2017.

[11] He K, Zhang X, Ren S et al. Deep residual learning for image recognition. In *CVPR*. 2016.

[12] Deng J, Dong W, Socher R et al. ImageNet: A large-scale hierarchical image database. In *CVPR*. 2009.

[13] Oksuz K, Cam BC, Akbas E et al. Localization recall precision (LRP): A new performance metric for object detection. In *ECCV*. 2018.

[14] Paszke A, Gross S, Massa F et al. PyTorch: An imperative style, high-performance deep learning library. *Neural Inf. Process. Syst.*. 2019:8024-8035.

[15] Jaeger PF, Kohl SAA, Bickelhaupt S et al. Retina u-net: Embarrassingly simple exploitation of segmentation supervision for medical object detection. *arXiv:1811.08661*. 2018.

[16] Hamidian S, Sahiner B, Petrick N et al. 3d convolutional neural network for automatic detection of lung nodules in chest CT. *SPIE Medical Imaging*. 2017.

[17] Setio AAA, Ciompi F, Geert Litjens et al. Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks. *IEEE Trans. Med. Imaging*. 2016;35:1160-1169.

[18] Dou Q, Chen H, Yu L, et al. Multilevel contextual 3-D CNNs for false positive reduction in pulmonary nodule detection. *IEEE Trans. Biomed. Eng*. 2017;64(7):1558-1567.

[19] Zhu W, Liu C, Fan W et al. DeepLung: Deep 3D Dual Path Nets for Automated Pulmonary Nodule Detection and Classification. In *IEEE WACV*. 2018.

[20] Onishi Y, Teramoto A, Tsujimoto M et al. Multiplanar analysis for pulmonary nodule classification in CT images using deep convolutional neural network and generative adversarial networks. *Int J Comput Assist Radiol Surg*. 2019;15(1):173-178.

[21] Nibali A, He Z, Wollersheim D. Pulmonary nodule classification with deep residual networks. *Int J Comput Assist Radiol Surg*. 2017;12:1799-1808.

[22] Liu S, Xie Y, Jirapatnakul A et al. Pulmonary nodule classification in lung cancer screening with three-dimensional convolutional neural networks. *J. of Medical Imaging*. 2017;4(4).

[23] Hua KL, Hsu HC, Hidayati SC et al. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. *Onco Targets Ther*. 2015;8:2015-2022.

[24] Dey R, Lu Z ; Hong Y. Diagnostic classification of lung nodules using 3D neural networks. *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*. 2018;774-778.