Human-recognizable CT image features of subsolid lung nodules associated with diagnosis and classification by convolutional neural networks

Jiang, Beibei; Zhang, Yaping; Zhang, Lu; H. de Bock, Geertruida; Vliegenthart, Rozemarijn; Xie, Xueqian

Published in:
European Radiology

DOI:
10.1007/s00330-021-07901-1

European Radiology
https://doi.org/10.1007/s00330-021-07901-1

CHEST

Human-recognizable CT image features of subsolid lung nodules associated with diagnosis and classification by convolutional neural networks

Beibei Jiang[1] · Yaping Zhang[1] · Lu Zhang[1] · Geertruida H. de Bock[2] · Rozemarijn Vliegenthart[3] · Xueqian Xie[1]

Received: 20 November 2020 / Revised: 3 February 2021 / Accepted: 16 March 2021

Abstract

Objectives The interpretability of convolutional neural networks (CNNs) for classifying subsolid nodules (SSNs) is insufficient for clinicians. Our purpose was to develop CNN models to classify SSNs on CT images and to investigate image features associated with the CNN classification.

Methods CT images containing SSNs with a diameter of ≤ 3 cm were retrospectively collected. We trained and validated CNNs by a 5-fold cross-validation method for classifying SSNs into three categories (benign and preinvasive lesions [PL], minimally invasive adenocarcinoma [MIA], and invasive adenocarcinoma [IA]) that were histologically confirmed or followed up for 6.4 years. The mechanism of CNNs on human-recognizable CT image features was investigated and visualized by gradient-weighted class activation map (Grad-CAM), separated activation channels and areas, and DeepDream algorithm.

Results The accuracy was 93% for classifying 586 SSNs from 569 patients into three categories (346 benign and PL, 144 MIA, and 96 IA in 5-fold cross-validation). The Grad-CAM successfully located the entire region of image features that determined the final classification. Activated areas in the benign and PL group were primarily smooth margins ($p < 0.001$) and ground-glass components ($p = 0.033$), whereas in the IA group, the activated areas were mainly part-solid ($p < 0.001$) and solid components ($p < 0.001$), lobulated shapes ($p < 0.001$), and air bronchograms ($p < 0.001$). However, the activated areas for MIA were variable. The DeepDream algorithm showed the image features in a human-recognizable pattern that the CNN learned from a training dataset.

Conclusion This study provides medical evidence to interpret the mechanism of CNNs that helps support the clinical application of artificial intelligence.

Key Points

- *CNN achieved high accuracy (93%) in classifying subsolid nodules on CT images into three categories: benign and preinvasive lesions, MIA, and IA.*
- *The gradient-weighted class activation map (Grad-CAM) located the entire region of image features that determined the final classification, and the visualization of the separated activated areas was consistent with radiologists' expertise for diagnosing subsolid nodules.*
- *DeepDream showed the image features that CNN learned from a training dataset in a human-recognizable pattern.*

Keywords Artificial intelligence · Deep learning · Adenocarcinoma of lung · X-ray computed tomography

✉ Xueqian Xie
xiexueqian@hotmail.com

[1] Radiology Department, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Haining Rd.100, Shanghai 200080, China

[2] Department of Epidemiology, University Medical Center Groningen, University of Groningen, Hanzeplein 1, 9713 GZ Groningen, The Netherlands

[3] Department of Radiology, University Medical Center Groningen, University of Groningen, Hanzeplein 1, 9713 GZ Groningen, The Netherlands

Published online: 13 April 2021

Abbreviations

AAH	Atypical adenomatous hyperplasia
AIS	Adenocarcinoma in situ
AUC	Area under the ROC curve
BMI	Body mass index
CNN	Convolutional neural network
CT	Computed tomography
Grad-CAM	Gradient-weighted class activation map
IA	Invasive adenocarcinoma
MIA	Minimally invasive adenocarcinoma
PL	Preinvasive lesions
ROC	Receiver operating characteristic curve
SSN	Subsolid nodule

Introduction

Lung cancer is the leading cause of mortality among all malignancies. Recently, the 10-year follow-up result from the Dutch-Belgian lung cancer screening trial (NELSON) showed that CT screening reduced lung cancer-related mortality from 3.3 to 2.5 deaths per 1000 people per year [1]. Subsolid nodules (SSNs) are common findings in CT examinations. The prevalence of SSNs varies greatly among different countries/ethnicities [2–5]. Differential diagnosis of SSNs is important and challenging because early-stage lung adenocarcinoma often appears as an SSN. Malignant SSNs mainly include adenocarcinoma in situ (AIS), minimally invasive adenocarcinoma (MIA), and invasive adenocarcinoma (IA) [6]. Treatment-related decision-making depends on the presurgical evaluation of lesion invasiveness. Surgical resection is strongly recommended for invasive lesions (MIA and IA) [7].

Studies have shown that convolutional neural networks (CNNs) can classify the histological types of lung adenocarcinoma on CT images [8–10]. Zhao et al used a multi-task CNN to classify lung nodules into preinvasive lesions (PL), MIA, and IA with an accuracy of 63.3% [8]. Wang et al determined the invasiveness of SSNs using a CNN and achieved an accuracy of 89.2% [9]. Despite the good performance of CNNs, few studies have explained the classification mechanism of CNNs to clinicians. Current deep learning approaches are insufficient to meet the clinical requirements of explainability and interpretability if they only provide an inference result by expressing a probability. The clinical treatment decision for a lung nodule must come from evidence and confidence. First, a supervised and transparent approach is necessary since artificial intelligence (AI) can make mistakes. Second, a responsible doctor has to explain the treatment decision to the patient with solid evidence rather than only quoting a result from an AI algorithm.

Computer scientists have developed visualization techniques to interpret CNNs [11]. Zeiler et al visualized the feature channels of each CNN layer using deconvolution [12], which helps explain the transformation between input and continuous layers. Alexander et al developed the DeepDream algorithm [13], which maximizes the feature channel activation of CNNs and illustrates the image features that CNNs learned from the training dataset. Selvaraju et al generated class-activated thermograms for inputs image using a gradient-weighted class activation map (Grad-CAM) [14], which assists in the visualization of the prominent part of an image leading to the final classification. However, the Grad-CAM only indicates the overall region associated with the classification [15]. Detailed human-recognizable image features and explainable evidence contributing to CNN classification in medical imaging are still lacking.

Therefore, we used SSN classification as an example to investigate the internal mechanism of CNNs for classifying medical images. The preparation procedure involved collecting data, training CNN models to classify SSNs on CT images into three categories (benign and PL, MIA, and IA), and validating their performance. Subsequently, we applied three visualization methods to reveal the classification mechanism of CNNs. The first method was using a Grad-CAM to generate an overall feature map to observe whether the CNN identified the entire nodule. The second method was to visualize the separated activation channels and areas for locating the specific CT image features of an SSN that were associated with the classification. The last method involved using DeepDream to generate a high-resolution feature map to illustrate the image features that CNN had learned.

Methods

Study population

A retrospective search was conducted for patients entering the electronic health record system at our institute from December 2011 to September 2020, and finally identified 569 patients. The inclusion and exclusion flowchart is shown in Fig. 1. The inclusion criteria were as follows: (1) SSNs with a diameter of ≤ 3 cm in thin-section CT images, including pure ground-glass and part-solid nodules; (2) histological results obtained with immunohistochemical staining based on nodule resection within 1 month after the CT examination; and (3) no radiotherapy or chemotherapy performed before nodule resection. SSNs that were stable for at least 5 years were also considered benign if they did not grow or develop solid components because adenocarcinomas developed solid components during follow-up; also, the volume doubling time for SSNs that became adenocarcinomas ranged from 300 to 900 days [16, 17]. A stable size was defined as a volume alteration of ≤ 25% in follow-up scans, which is within the range of systematic error in lung cancer screening studies (the details of the evaluation of stable nodules are shown in the Appendix) [18, 19]. The exclusion criteria were as follows: (1) motion or respiratory artifacts leading to poor image quality; (2)

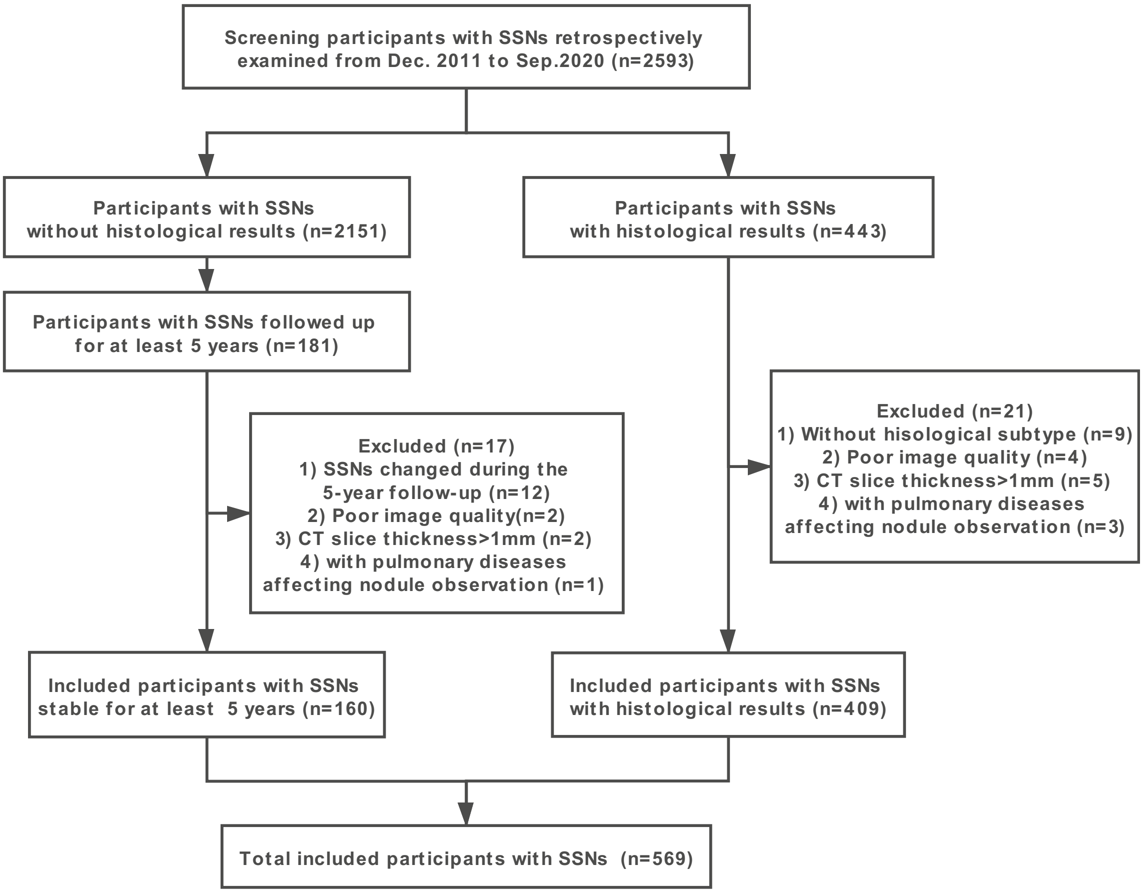

Screening participants with SSNs retrospectively examined from Dec. 2011 to Sep.2020 (n=2593)
Participants with SSNs without histological results (n=2151)
Participants with SSNs with histological results (n=443)
Participants with SSNs followed up for at least 5 years (n=181)
Excluded (n=17)
1) SSNs changed during the 5-year follow-up (n=12)
2) Poor image quality(n=2)
3) CT slice thickness>1mm (n=2)
4) with pulmonary diseases affecting nodule observation (n=1)
Excluded (n=21)
1) Without hisological subtype (n=9)
2) Poor image quality (n=4)
3) CT slice thickness>1mm (n=5)
4) with pulmonary diseases affecting nodule observation (n=3)
Included participants with SSNs stable for at least 5 years (n=160)
Included participants with SSNs with histological results (n=409)
Total included participants with SSNs (n=569)

Fig. 1 Inclusion and exclusion flowchart. *Some patients have ≥ 2 nodules. SSNs, subsolid nodules

CT slice thickness > 1 mm; and (3) with pulmonary diseases affecting nodule observation. The local Institutional Review Board approved this retrospective study (No. SGH-2018-56) and waived the need for written informed consent.

Image pre-processing

The CT acquisition protocol is shown in the Appendix. Square-patch images containing the entire nodule with marginal structures (such as peripheral vessels, pleura, and lung tissues) were obtained. These patch images were configured to the lung window setting, which is optimal for the evaluation of SSNs [20]. Then, the patch images were interpolated to 299×299 pixels to meet the input layer requirement of CNNs. If an SSN had a large number of images, we selected up to 20 characteristic images. In this way, 4–20 patch images were retained for each SSN.

Grouping outline

The dataset was divided into three categories: benign and PL (including histologically benign, atypical adenomatous hyperplasia [AAH], AIS, and stable SSNs), MIA, and IA. A five-fold cross-validation method was used to validate CNN performance and generalizability [21]. In each of the five consecutive deep learning sessions (folds 1 to 5), we divided the datasets into five non-overlapping splits, namely four (80%) as the training dataset and one (20%) as the validation dataset. In these 5 splits, the proportion of the three disease categories was similar. The overall accuracy of the CNN was the mean accuracy of the five validation sessions. Moreover, we selected the fold (training and validation dataset) with the highest accuracy in the 5-fold cross-validation for further visualization analysis.

Convolutional neural networks

The first CNN used in this study was GoogLeNet Inception v3, a directed acyclic graph (DAG) CNN; it was used to analyze CNN activation channels (Fig. 2a). We froze the first 18 layers and adjusted their parameters by backpropagation through our training dataset. The image augmentation methods were as follows: randomly flipping images along the horizontal or vertical axis, panning images by −10 to 10 pixels along the horizontal or vertical axis, and image scaling by 0.95 to 1.05. Training was performed for 5, 20, 50, 100, and 200 epochs. The minibatch size was 64, the learning rate was 0.0003, and the L2 regularization rate was 0.0001. The optimization algorithm was adaptive moment estimation

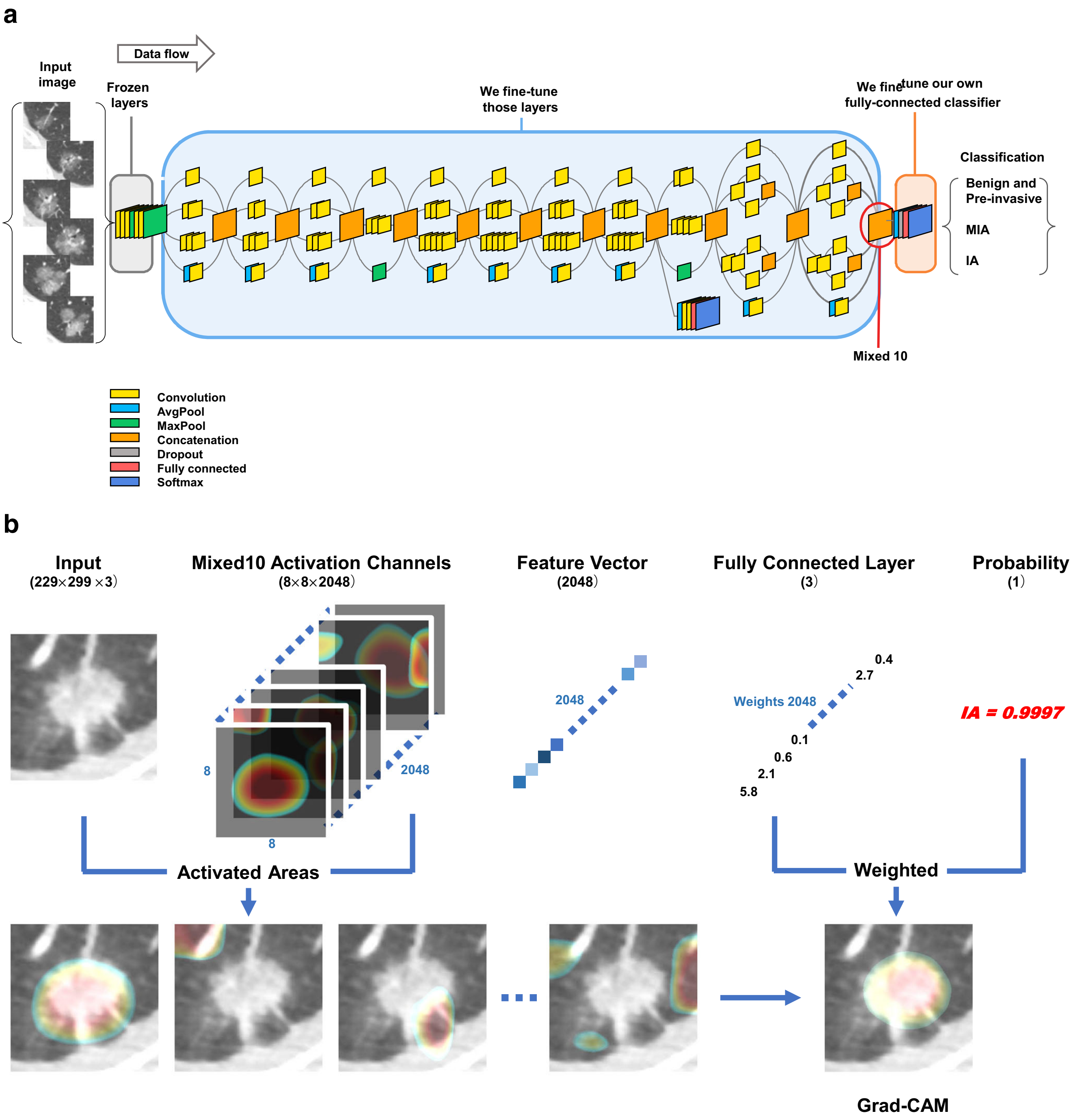


Fig. 2 **a** The Inception v3 convolutional neural network (CNN) and the visualized layer. **b** Visualized activation areas in layer mixed10 of Inception v3. By visualizing the activation of these 2D channels in layer mixed10 (the last feature layer) and superimposing the activated areas on the original images, the image features in the original image that were closely associated with the final classification result can be marked

(Adam). The second CNN was AlexNet, a serial CNN; it was used to generate DeepDream images. We froze the first 4 layers when training for 50 epochs. The other parameters were consistent with those of Inception v3. The inference method of the CNN is shown in the Appendix. A summary of the visualization techniques used in this study is shown in Table 1.

CNN activation

We studied the activation of this CNN by analyzing the distribution of parameters in the last 2D multi-channel layer (layer mixed 10) consisting of 2048 feature channels. This 2D layer transformed into a one-dimensional

Table 1 Summary of the visualization techniques applied in this study

Method	Scope	Purpose	Input and procedure	Output
Grad-CAM	Validation dataset	Demonstrating the region of image features for determining the final classification	Inputting a nodule image, showing an overall feature map generated by all feature channels in a layer with weights of its classification	One feature map for the whole activation, showing as one low-resolution thermograph per inputting image in this study
Channel activation	Training and validation dataset	Showing the influence from different channels on the final classification	Inputting a nodule image, showing the activated areas of all feature channels in a layer	Multiple thermographs for detailed activated areas, showing as 2,048 low-resolution thermographs per inputting image in this study
DeepDream	Training dataset	Illustrating the detailed image features that CNN learned	Starting from a blank image to maximize its activation value of a specified category by gradient descent	One high-resolution feature image per category, showing 3 images corresponding to 3 categories in this study

layer by a pooling algorithm and finally determined the classification layer through the fully connected layer and softmax layer [22]. The channels in the 2D layer with higher activation values directly affected the classification results.

Therefore, we illustrated the whole region associated with the final classification by a Grad-CAM, which synthetically generated an activation map based on the gradient of the classification score (i.e., weights of the fully connected layer) and the corresponding features of the layer mixed 10 [14]. Furthermore, we visualized separated activation channels and superimposed the activated areas into the original images; this was done to mark the image features in the original image that were closely associated with the final classification results of the CNN (Fig. 2b). Therefore, these separated activated channels allowed visualization of the specific image features that determined the classification results of the CNN. The details are shown in the Appendix.

CT features

We adopted the CT feature terms proposed by an evaluation panel consisting of 107 radiologists from 25 countries [23]. The features included shapes (oval, irregular, and lobulated), margins (smooth and spiculated), composition (ground-glass, part-solid, and solid), morphological features (vessel touching or passing through, air bronchograms and pleural tags), and non-nodule features (peripheral vessel, chest wall, or other features uncorrelated to a nodule). Based on this terminological system, two radiologists carefully reviewed the original CT images overlapping with activated areas and subsequently determined the CT features in the activated areas; differences were resolved by consensus.

DeepDream

DeepDream amplifies and displays the features learned in a channel or layer in a CNN [13]. DeepDream starts with a random noise image and then gradually adjusts this image so that the activation value of the channel or layer, which allows visualization of the image features learned by the CNN, gradually increases until it generates detailed image features on this image. We used DeepDream to visualize the fully connected layer of Inception v3 and AlexNet because this layer directly translates into the three-category output layer. The details are shown in the Appendix.

Statistics

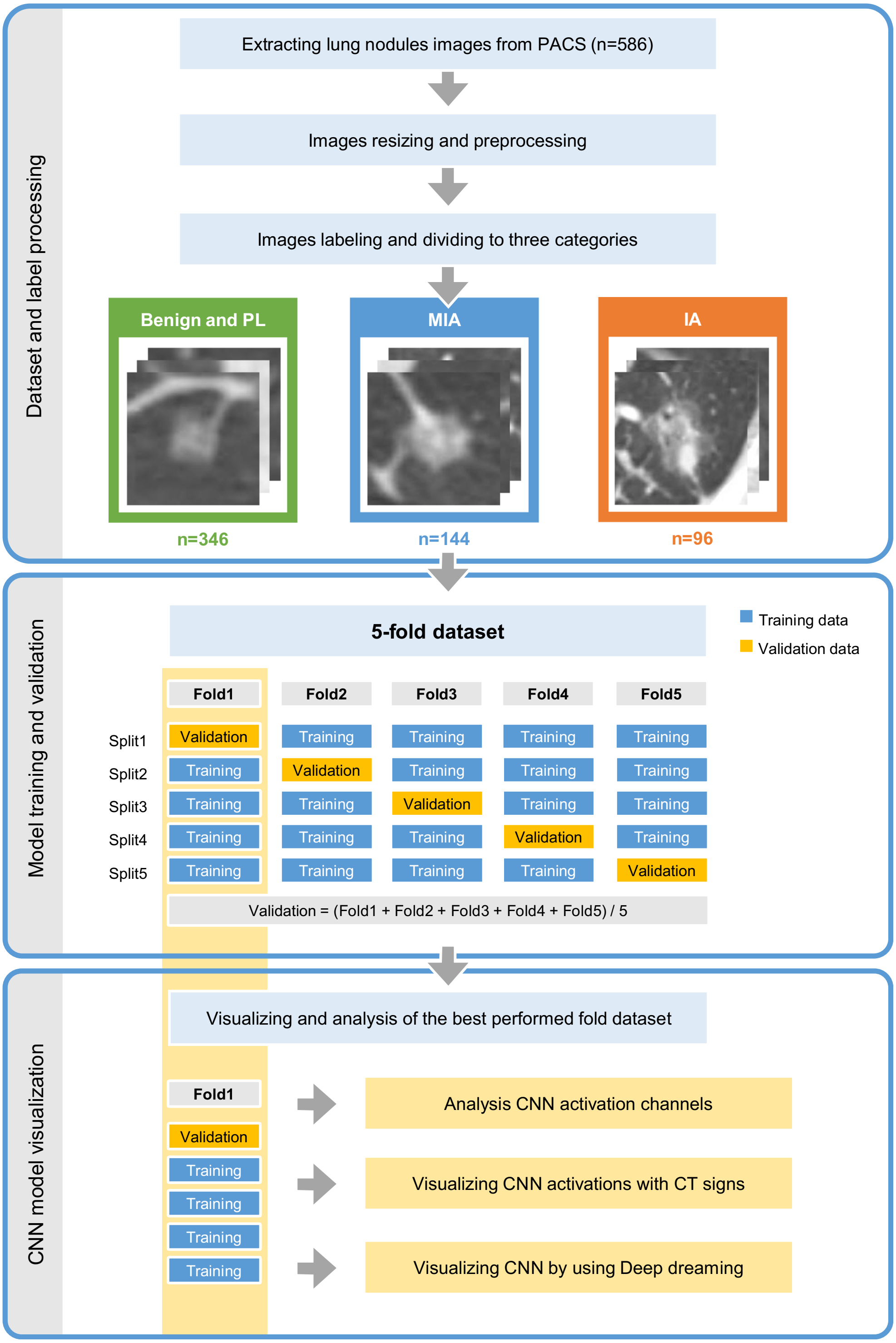
The association between the actual label and the classification determined by the CNN was assessed by a confusion matrix. Diagnostic accuracy, sensitivity (recall), specificity, F1 score, and area under the receiver operating characteristic curve (AUC) were evaluated to determine the three-way classification performance of CNNs for pulmonary nodules. The F1 score was the weighted harmonic average of recall and precision [24]. An independent-samples *t* test was used to compare age, and a chi-square test was used to compare sex. The activated areas of pulmonary nodules were compared among the three categories using the Kruskal-Wallis one-way ANOVA. A statistical package (SPSS Statistics 24, IBM) was used to analyze the data.

Results

Population

The study workflow is shown in Fig. 3. This study included 586 SSNs from 569 patients (mean age 59.6 ± 11.1 years) (Table 2). Of these nodules, 422 (72.0%) were surgically resected and immunohistochemically stained, whereas 164 (28.0%) were stable nodules (mean follow-up time 6.41 ± 1.48 years). There were 346 (59.0%) benign and PL nodules, i.e., 164 nodules that were stable for 6.4 years, 19 that were histologically benign, 60 that were AAH, and 103 that were

Fig. 3 The study workflow. PL, preinvasive lesions; MIA, minimally invasive adenocarcinoma; IA, invasive adenocarcinoma; CNN, convolutional neural network


Dataset and label processing
Extracting lung nodules images from PACS (n=586)
Images resizing and preprocessing
Images labeling and dividing to three categories
Benign and PL
MIA
IA
n=346
n=144
n=96
Model training and validation
5-fold dataset
Training data
Validation data
Validation = (Fold1 + Fold2 + Fold3 + Fold4 + Fold5) / 5
CNN model visualization
Visualizing and analysis of the best performed fold dataset
Fold1
Validation
Training
Training
Training
Training
Analysis CNN activation channels
Visualizing CNN activations with CT signs
Visualizing CNN by using Deep dreaming

	Fold1	Fold2	Fold3	Fold4	Fold5
Split1	Validation	Training	Training	Training	Training
Split2	Training	Validation	Training	Training	Training
Split3	Training	Training	Validation	Training	Training
Split4	Training	Training	Training	Validation	Training
Split5	Training	Training	Training	Training	Validation

AIS. There were 144 (24.6%) MIA and 96 (16.4%) IA cases, respectively. The long diameter of SSNs was 9.1 ± 5.1 mm; more specifically, the long diameter was 6.7 ± 2.8 mm, 9.2 ± 3.6 mm, and 16.6 ± 6.0 mm for benign and PL, MIA, and IA, respectively (detailed sizes per SNN subtype are shown in Supplementary Table S1).

Training and cross-validation

After 5, 20, 50, 100, and 200 training epochs of the Inception v3 CNN, the overall accuracy in the 5-fold cross-validation in each epoch was 0.925 ± 0.022, 0.933 ± 0.029, 0.921 ± 0.019, 0.927 ± 0.007, and 0.922 ± 0.017, respectively

Table 2 Patient characteristics in the training and validation datasets for visualization analysis

Variables	All	Training dataset	Validation dataset	*p* value
Patients, *n*	569	463	116	
Gender				0.453
Female, *n* (%)	395 (69.6%)	324 (70.0%)	77 (66.4%)	
Male, *n* (%)	174 (30.4%)	139 (30.0%)	39 (33.6%)	
Age (years), mean ± SD	59.6 ± 11.1	59.4 ± 11.1	60.0 ± 11.1	0.609
Age (years)				
< 60, *n* (%)	276 (48.5%)	219 (47.3%)	60 (51.7%)	
≥ 60, *n* (%)	293 (54.5%)	244 (52.7%)	56 (48.3%)	
Nodules, *n*	586	467	119	
Long diameter (mm), mean ± SD	9.1 ± 5.1	9.0 ± 5.0	9.5 ± 5.3	0.362
Histological type				0.999
Benign				
Histologically benign, *n* (%)	19 (3.2%)	15 (3.2%)	4 (3.4%)	
Stable for 6.4 years, *n* (%)	164 (28.0%)	131 (28.1%)	33 (27.7%)	
Preinvasive lesions				
Atypical adenomatous hyperplasia, *n* (%)	60 (10.2%)	48 (10.3%)	12 (10.1%)	
Adenocarcinoma *in situ*, *n* (%)	103 (17.6%)	82 (17.6%)	21 (17.6%)	
Minimally invasive adenocarcinoma, *n* (%)	144 (24.6%)	115 (24.6%)	29 (24.4%)	
Invasive adenocarcinoma, *n* (%)	96 (16.4%)	76 (16.3%)	20 (16.8%)	

*The training and validation datasets are from fold 1 in the 5-fold cross-validation procedure. Some patients have ≥ 2 nodules. The multiple nodules in one patient may be at a different histological condition

(Supplementary Figure S1). Among these five training epochs, the highest accuracy was 0.933 ± 0.029 at 20 epochs. The corresponding confusion matrix and performance metrics are shown in Supplementary Tables S2 and S3A, respectively. The sensitivity for benign and PL, MIA, and IA was 0.965, 0.875, and 0.905, respectively. The specificity for benign and PL, MIA, and IA was 0.937, 0.975, and 0.973, respectively. The F1 values for determining the three categories of SSNs were 0.961, 0.897, and 0.886. The AUCs were 0.979 (95% CI: 0.975–0.983), 0.955 (0.949–0.960), and 0.971 (0.967–0.976) for benign and PL, MIA, and IA, respectively (Fig. 4). In the 5-fold cross-validation at 20 epochs, fold 1 showed the highest accuracy of 0.958. Thus, we visualized the CNN model established by using the fold 1 dataset. The details of the fold 1 dataset are shown in Table 2. AlexNet was also trained with the fold 1 dataset, and the classification accuracy was 0.874. The AUCs were 0.955 (95% CI: 0.942–0.967), 0.847 (0.824–0.867), and 0.928 (0.911–0.942) for the three categories, respectively. The performance metrics are shown in Supplementary Table S3B.

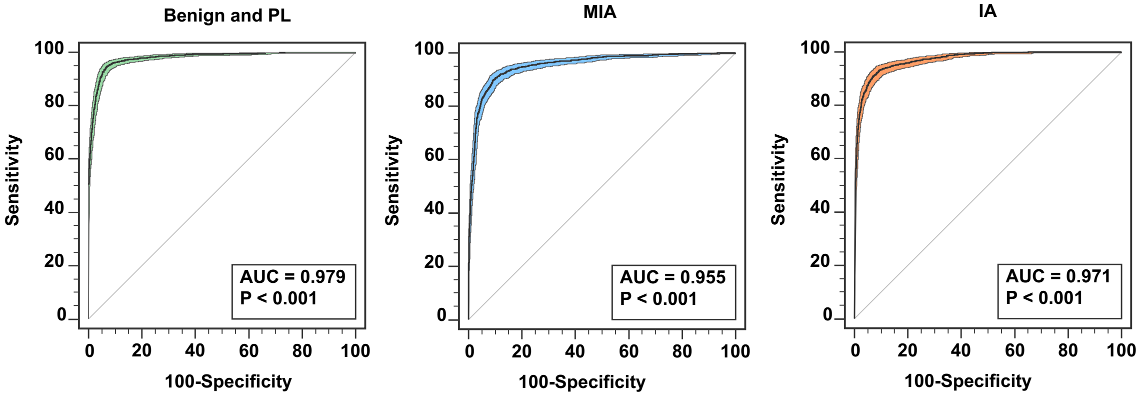


Fig. 4 The area under the receiver operating characteristic curve (AUC) of the 5-fold cross-validation based on the Inception v3 convolutional neural network. PL, preinvasive lesions; MIA, minimally invasive adenocarcinoma; IA, invasive adenocarcinoma

The training dataset of fold 1 comprised 467 pulmonary nodules. As the training epochs increased, validation accuracy decreased (Supplementary Figure S1). Also, the activated channels became increasingly sparse as training epochs increased (Supplementary Figure S2).

CNN activation and CT features

To visualize the image features that determine the classification results of the CNN, we analyzed 119 nodules in the validation dataset; of these 119 nodules, 114 (95.8%) were correctly classified, including 69 benign and PL nodules (92 slices, 1840 separated activated areas), 25 MIA nodules (50, 1000), and 20 IA nodules (53, 1060). For these 114 correctly classified nodules, 95.6% (109/114) of the Grad-CAM covered the entire nodule, and 4.3% (5/114) of the activation maps only involved the margin or part of the nodule. The separated activated areas of the top 20 most activated channels were associated with the CT features defined by radiologists' expertise (Fig. 5 and Supplementary Table S4). Several correctly classified nodules are shown in Fig. 6A.

For benign and PL nodules with 1840 separated activated areas, 10 image features were observed in the activated areas. Smooth margins were a significant feature (1044/1840

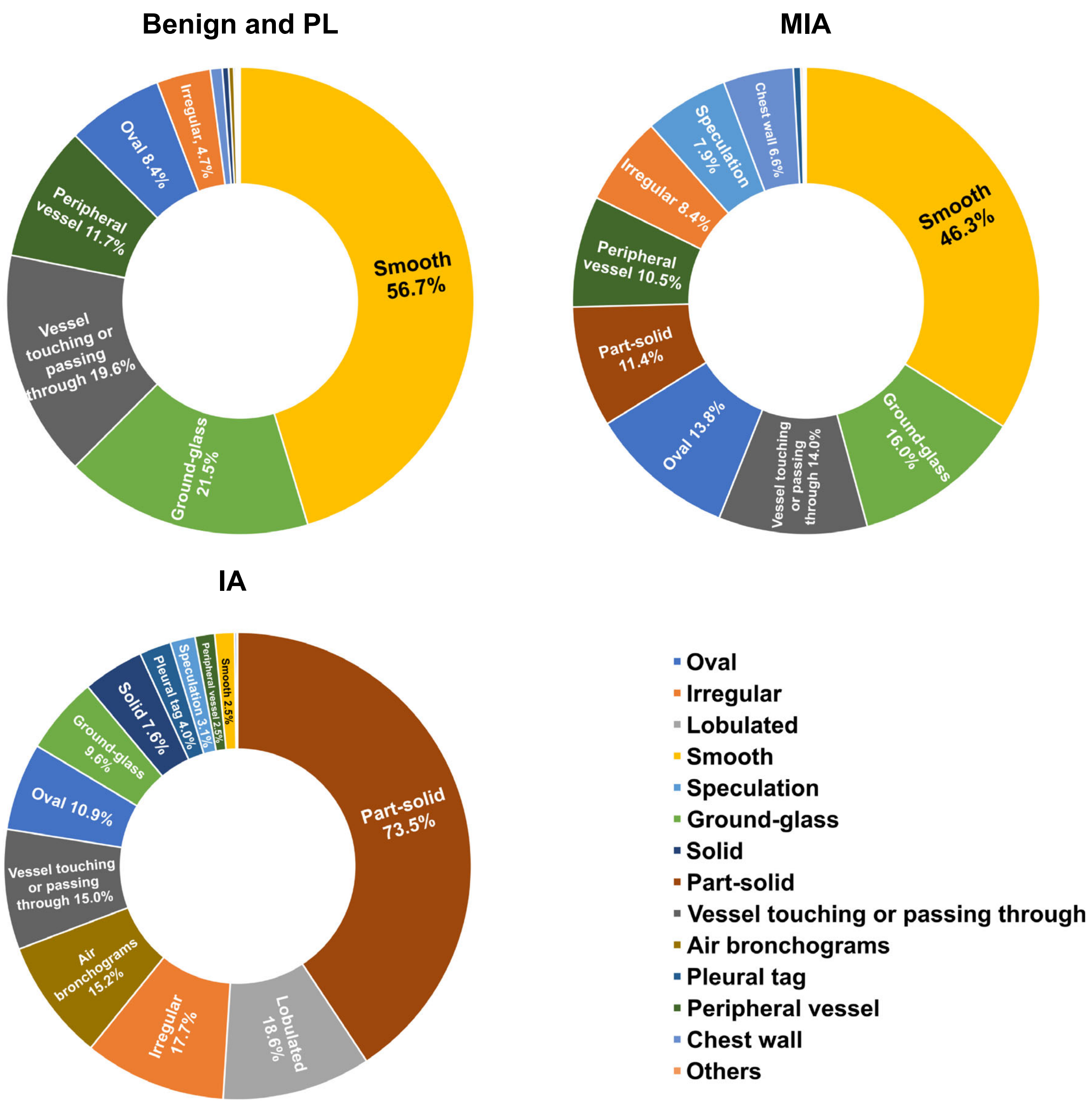


Fig. 5 The proportion of CT image features on the activated areas of the convolutional neural network. The percentage refers to the number of CT features focusing on the activated areas (numerator) divided by the total number (denominator) in each category. The denominators are 1840, 1000, and 1060 for benign and PL, MIA, and IA, respectively. Some of the activated areas focused on multiple kinds of CT features, so the sum was greater than 100%. The CT features are arranged clockwise by proportion. PL, preinvasive lesions; MIA, minimally invasive adenocarcinoma; IA, invasive adenocarcinoma

[56.7%]) compared with those in the other two categories ($p < 0.001$), followed by ground-glass components (395/1840 [21.5%], $p = 0.033$). Oval-shaped (154/1840 [8.4%], $p = 0.785$) and irregular-shaped (86/1840 [4.7%], $p = 0.074$) features were observed in this category, but they were not significant ($p > 0.05$).

For MIA cases with 1000 separated activated areas, 11 features were observed. Smooth margins were significant (463/1000 [46.3%], $p < 0.001$), followed by ground-glass components (160/1000 [16.0%], $p = 0.033$) and oval shapes (138/1000 [13.8%], $p = 0.785$), as well as part-solid components (114/1000 [11.4%], $p < 0.001$), irregular shapes (84/1000 [8.4%], $p = 0.074$), and spiculated margins (79/1000 [7.9%], $p = 0.003$). Other image features in the activated areas were unrelated to the nodules, such as chest wall (66/1000 [6.6%], $p = 0.087$).

For IA cases with 1060 separated activated areas, 13 features were observed. Part-solid components (779/1060 [73.5%]) were the most significant feature ($p < 0.001$), as well as lobulated shapes (197/1060 [18.6%], $p < 0.001$) and irregular shapes (188/1060 [17.7%], $p = 0.074$). Some CT features were also characterized, including oval shapes (116/1060 [10.9%], $p = 0.785$), ground-glass components (102/1060 [9.6%], $p < 0.001$), and solid components (81/1060 [7.6%], $p < 0.001$). Furthermore, some specific features were observed, such as air bronchograms (161/1060 [15.2%], $p < 0.001$) and pleural tags (42/1060 [4.0%], $p < 0.063$).

In addition, five nodules were misclassified (Fig. 6B), including one benign and PL nodule and four MIA nodules. Among them, the benign and PL nodule was misclassified as MIA, whereas two of the four MIA nodules were misclassified as benign and PL, and the other two were misclassified as IA. The Grad-CAM of 4 nodules was on the margin of or just part of the nodule. Only one Grad-CAM located the whole nodule.

DeepDream

The activation values of the three disease categories of AlexNet were 204.6, 535.0, and 1060.6 after 300 iterations. The DeepDream image of the benign and PL category (Fig. 7a) produced multiple oval nodule-like shapes with smooth margins, halo signs, and homogeneous composition. For MIA cases (Fig. 7b), there were multiple round and irregular nodule-like shapes with sharply smooth margins, uniform composition, and some nodule-like shapes with a linear shadow in the center. The images of IA nodules (Fig. 7c) showed similar oval nodule-like shapes with different size sand non-uniform composition, and there were blurred shadows around the margin. Most of them contained low-density shadows, similar to air bronchograms. There were high-density spots and scattered stripes, similar to blood vessel signs.

The DeepDream image of Inception v3 was full of noise because of low activation. The activation values of the three categories were 0.88, 0.35, and 3.26 for benign and PL, MIA, and IA, respectively, after 1000 iterations.

Discussion

In this study, CNNs reached a high accuracy of 93% in classifying SSNs into three categories. CNN classification was associated with morphological features, such as composition and margins, characterized by Grad-CAM and the separated activated areas. The DeepDream algorithm illustrated the human-readable image features that the CNN learned from the training dataset. The activated areas in the benign and PL group were primarily smooth margins and ground-glass components, whereas, in the IA group, the activated areas focused on the part-solid and solid components of the nodules, lobulated shapes, and air bronchograms. However, the activated areas for MIA cases were variable.

Several studies have investigated the two-way histological classification of lung adenocarcinoma, resulting in accuracy of up to 89% [8–10, 25]. The accuracy of three-way classification was 63% for nodules ≤ 1 cm in size [8, 26]. Our accuracy was 93% for nodules ≤ 3 cm. The reason for this difference in accuracy could be attributed to the larger nodules evaluated in our study, which provided more details and improved the learning effect of the CNN. Moreover, many researchers have trained CNNs based on malignant lesions that were unhelpful for the diagnosis of benign lesions. As the majority of SSNs found by CT screening or incidentally are benign, our model was developed with many benign nodules, which is clinically practical.

The activation area of Grad-CAM involved the whole region for the final classification. In classifying SSNs, Grad-CAM only located the whole nodule. However, it is important to indicate specific features. Therefore, we investigated and visualized separated activation channels and activated areas. We found that CNN classification was associated with the morphological image features of SSNs. The activated areas of the benign and PL were significantly associated with smooth margins and ground-glass components. Radiologists also diagnosed benign and PL mainly according to these signs, which are commonly considered non-malignant signs on CT images [27]. The image features of IA were associated with part-solid and solid components, lobulated shape, and air bronchograms. These are also important features for radiologists to evaluate when diagnosing malignant nodules [6]. However, the activated areas for MIA cases were variable and included smooth and spiculated margins and ground-glass and part-solid components. Because MIA characteristics are histologically between PL and IA features, they may share

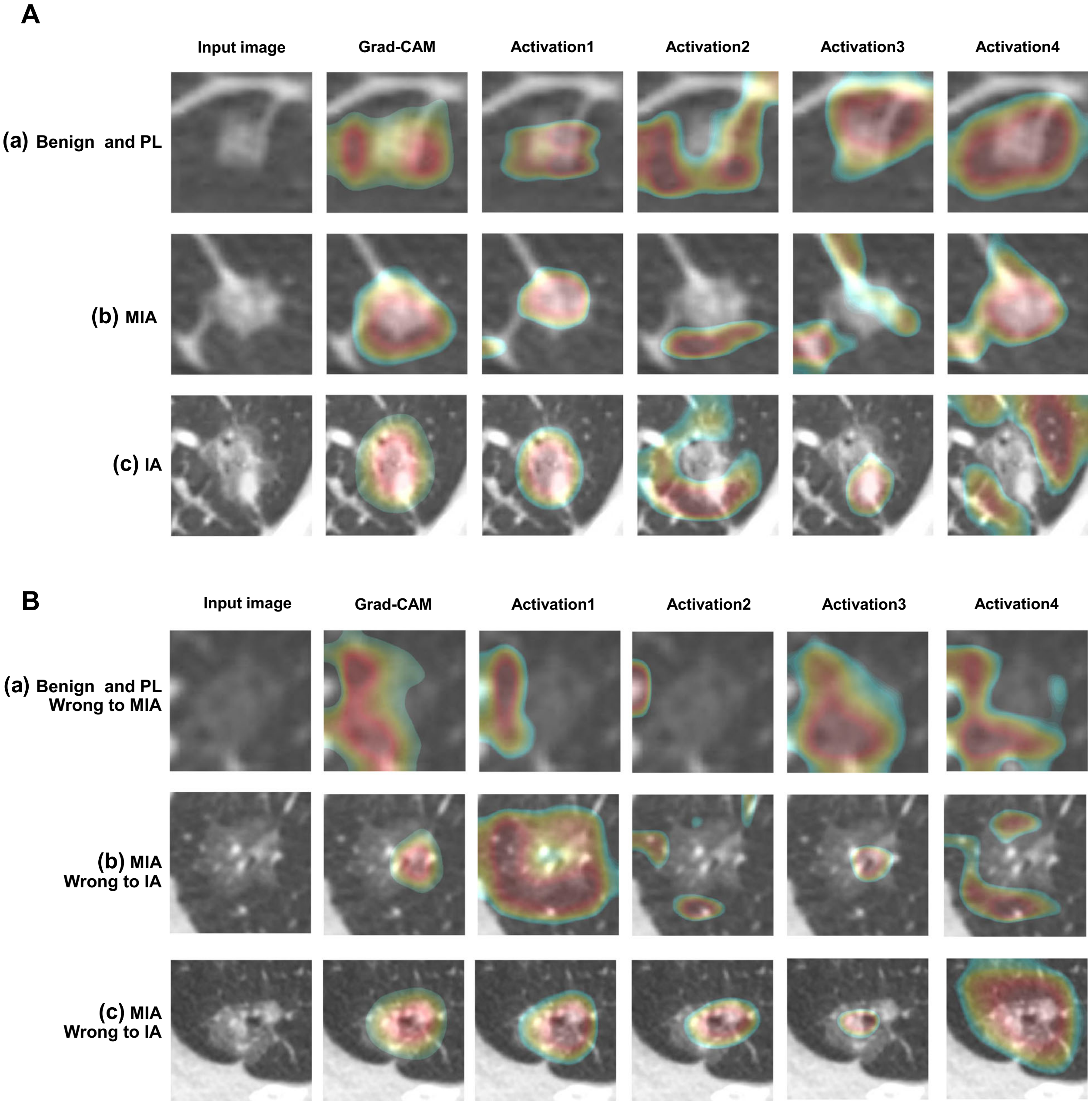


Fig. 6 **A.** Correctly classified representative cases with the Grad-CAM map and the top 4 activated areas of the Inception v3 convolutional neural network. **a** A 66-year-old female had a benign pure ground-glass nodule with a long diameter of 7mm (categorized as benign and PL). The Grad-CAM involved the entire nodule and focused more on its margin. The activated areas included a pure ground-glass component, a smooth margin, vessel touching or passing through, and the whole nodule with a smooth margin. **b** A 62-year-old female had a mixed ground-glass nodule with a long diameter of 7mm, proven as MIA. The Grad-CAM involved the entire nodule with a margin. The activated areas mainly included the part-solid component and an irregular shape, a mild spiculated margin, vessel touching or passing through, and the whole nodule with vessels. **c** A 44-year-old female has a mixed ground-glass nodule with a long diameter of 25mm, proven as IA. The Grad-CAM also involved the entire nodule. The activated areas mainly include the part-solid component and a lobulated shape, a spiculated margin, the solid component, and peripheral vessels. Grad-CAM, gradient-weighted class activation map; PL, preinvasive lesions; MIA, minimally invasive adenocarcinoma; IA, invasive adenocarcinoma. **B.** Incorrectly classified representative cases with the Grad-CAM and the top 4 activated areas of the Inception v3 convolutional neural network. **a** A 93-year-old male had a pure ground-glass nodule with a long diameter of 10mm (categorized as benign and PL), which was misclassified as MIA. The Grad-CAM only involved the margin of the nodule. The activations focused on a smooth margin, a peripheral vessel, the ground-glass component, and a smooth margin. **b** A 66-year-old female had a mixed ground-glass nodule with a long diameter of 13mm, proven as MIA but incorrectly classified as IA. The Grad-CAM only involved some part-solid component and an air bronchogram of the nodule. The activated areas included the part-solid component and an irregular shape, peripheral vessels, the air bronchogram, and smooth margins. **c** A 62-year-old female had a mixed ground-glass nodule with a long diameter of 15mm, proven as MIA but was incorrectly classified as IA. The Grad-CAM involved the entire nodule. The activated areas included the part-solid component and a lobulated shape, the air bronchogram, and the whole nodule. Grad-CAM, gradient-weighted class activation map; PL, preinvasive lesions; MIA, minimally invasive adenocarcinoma; IA, invasive adenocarcinoma

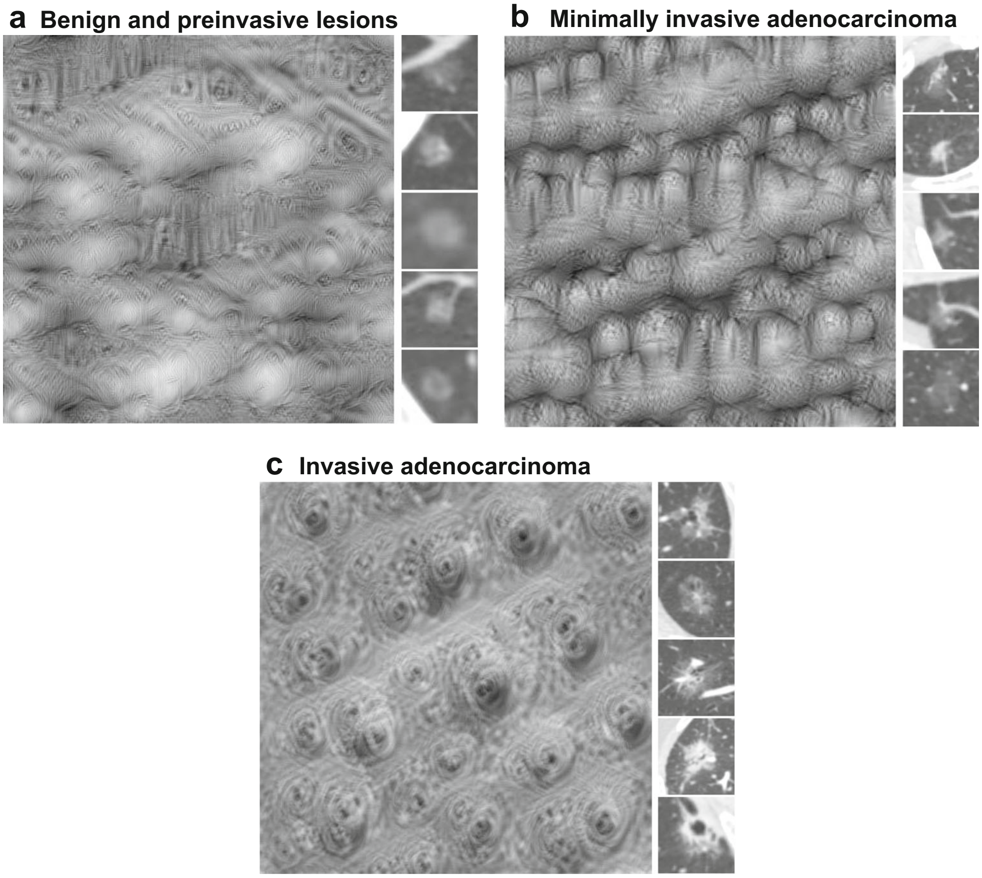

a Benign and preinvasive lesions
b Minimally invasive adenocarcinoma
c Invasive adenocarcinoma

Fig. 7 DeepDream illustrations of the learned image features of the AlexNet convolutional neural network. The DeepDream image is on the left. Five representative subsolid nodules in the corresponding category are on the right

the features of the two categories. Importantly, some activated areas of MIA fell outside the nodule, such as the features on the chest wall. Because CNNs sometimes use unreliable contexts for classification, which would cause mistakes to be made [28], unreliable contexts should be supervised in the diagnosis procedure.

Activation maximization techniques, such as DeepDream and Lucid [29], enlarge feature activation by gradient descent until the features are visible. In particular, we illustrated the high-level distinguishing features that the CNN learned. DeepDream generated nodule-like shapes from the benign and PL category, which showed uniform composition and smooth margins but no other sophisticated features. These findings are consistent with a human-readable diagnostic impression [27]. DeepDream generated nodule-like shapes with spiculated margins and uneven inner composition for IA. These features are also consistent with some malignant signs of IA on CT [6]. Regarding the use of the DeepDream algorithm, the developer addressed that this algorithm is suitable for shallow networks, such as AlexNet (25 layers) and VGG16 (41 layers), whose visualization performance was better than that of deep networks, such as Inception v3 (315 layers) [30–32]. Besides, Zeiler et al used AlexNet to represent CNN visualization [12]. This evidence strengthened our methodology that using AlexNet to illustrate the image features of subsolid nodules that CNN learned. Therefore, we used the shallow network AlexNet, whose activation values were 204.6, 535.0, and 1060.6 in this study, much higher than those of Inception v3 (0.88, 0.35, and 3.26 after 1000 iterations).

Previous studies have shown that thermography, such as Grad-CAM, can provide an intuitive understanding of CNN classification [33, 34]. The Grad-CAM superposes all the feature channels in a layer to generate an overall heatmap, which often contains the whole lesion. Therefore, we can determine whether the CNN accurately identifies the nodule and makes a diagnosis based on the entire nodule, rather than only identifying part of the nodule or interfering with the background. However, a Grad-CAM does not provide separated and detailed image features, so we further analyzed multiple independent channels and activated areas. This approach can help clinicians better understand the diagnostic criteria of CNNs and provide a medical imaging explanation for how CNNs classify SSNs.

There were several limitations in this study. First, this was a single-center study. Diagnostic accuracy may be variable for datasets in different institutes, but the internal mechanism of the CNN is consistent. Next, the sample size was not very large. Although using a larger dataset is commonly considered helpful to improve CNN performance, a recent study from OpenAI showed that accuracy would not continuously improve with increased sample size and more complex networks [35]. Nevertheless, the accuracy of 93% obtained in this study is sufficient to represent the CNN performance.

In summary, the CNN achieved high accuracy in classifying subsolid nodules on CT images into three histological categories. CNN classification was associated with CT features consistent with radiologist expertise for image features. The DeepDream algorithm illustrated the human-recognizable image features the CNN learned from the training dataset. Thus, this study provides medical imaging evidence to interpret the CNN classification for subsolid nodules, which helps to strengthen the application of deep learning in the diagnosis of subsolid nodules and can be seen as an example of CNN interpretability research for other imaging applications.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00330-021-07901-1.

Funding This study has received funding from the National Natural Science Foundation of China (project no. 81971612), Ministry of Science and Technology of China (2016YFE0103000), Shanghai Municipal Education Commission – Gaofeng Clinical Medicine Grant Support (20181814), Shanghai Jiao Tong University (ZH2018ZDB10), and Clinical Research Innovation Plan of Shanghai General Hospital (CTCCR-2018B04, CTCCR-2019D05). The funders played no role in the study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

Declarations

Guarantor The scientific guarantor of this publication is Xueqian Xie.

Conflict of interest The authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

Statistics and biometry One of the authors has significant statistical expertise.

Informed consent Written informed consent was waived by the Institutional Review Board.

Ethical approval Institutional Review Board approval was obtained (No. SGH-2018-56).

Study subjects or cohorts overlap No study subjects or cohorts have been previously reported.

Methodology
- Retrospective
- Diagnostic or prognostic study
- Performed at one institution

References

1. de Koning HJ, van der Aalst CM, de Jong PA et al (2020) Reduced lung-cancer mortality with volume CT screening in a randomized trial. N Engl J Med 382:503–513
2. Lee HW, Jin KN, Lee JK et al (2019) Long-term follow-up of ground-glass nodules after 5 years of stability. J Thorac Oncol 14:1370–1377
3. Silva M, Prokop M, Jacobs C et al (2018) Long-term active surveillance of screening detected subsolid nodules is a safe strategy to reduce overtreatment. J Thorac Oncol 13:1454–1463
4. McWilliams A, Tammemagi MC, Mayo JR et al (2013) Probability of cancer in pulmonary nodules detected on first screening CT. N Engl J Med 369:910–919
5. Henschke CI, Yip R, Yankelevitz DF et al (2013) Definition of a positive test result in computed tomography screening for lung cancer: a cohort study. Ann Intern Med 158:246–252
6. Travis WD, Brambilla E, Noguchi M et al (2011) International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. J Thorac Oncol 6:244–285
7. Naidich DP, Bankier AA, MacMahon H et al (2013) Recommendations for the management of subsolid pulmonary nodules detected at CT: a statement from the Fleischner Society. Radiology 266:304–317
8. Zhao W, Yang J, Sun Y et al (2018) 3D deep learning from CT scans predicts tumor invasiveness of subcentimeter pulmonary adenocarcinomas. Cancer Res 78:6881–6889
9. Wang S, Wang R, Zhang S et al (2018) 3D convolutional neural network for differentiating pre-invasive lesions from invasive adenocarcinomas appearing as ground-glass nodules with diameters ≤3 cm using HRCT. Quant Imaging Med Surg 8:491–499
10. Gong J, Liu J, Hao W et al (2020) A deep residual learning network for predicting lung adenocarcinoma manifesting as ground-glass nodule on CT images. Eur Radiol 30:1847–1855
11. Qin ZW, Yu FX, Liu CC, Chen X (2018) How convolutional neural networks see the world - a survey of convolutional neural network visualization methods. Mathematical Foundations of Computing 1:149–180
12. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. Computer Vision - ECCV 2014 8689:818–833
13. Alexander M, Christopher O, Tyka M (2015) Inceptionism: Going deeper into neural networks. Available via http://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html. Accessed 1 Feb 2021
14. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2016) Grad-CAM: Visual explanations from deep networks via gradient-based localization. Int J Comput Vis 2:336–359
15. Yamashita R, Nishio M, Do RKG, Togashi K (2018) Convolutional neural networks: an overview and application in radiology. Insights Imaging 9:611–629
16. Li JX, Xia TT, Yang XG et al (2018) Malignant solitary pulmonary nodules: assessment of mass growth rate and doubling time at follow-up CT. J Thorac Dis 10:S797–S806
17. Alpert JB, Ko JP (2018) Management of incidental lung nodules: Current strategy and rationale. Radiol Clin North Am 56:339–351
18. Xie X, Heuvelmans MA, van Ooijen PM, Oudkerk M, Vliegenthart R (2013) A practical approach to radiological evaluation of CT lung cancer screening examinations. Cancer Imaging 13:391–399
19. Horeweg N, Scholten ET, de Jong PA et al (2014) Detection of lung cancer through low-dose CT screening (NELSON): a prespecified analysis of screening test performance and interval cancers. Lancet Oncol 15:1342–1350

20. MacMahon H, Naidich DP, Goo JM et al (2017) Guidelines for management of incidental pulmonary nodules detected on ct images: from the Fleischner Society 2017. Radiology 284:228–243
21. Blagus R, Lusa L (2015) Joint use of over- and under-sampling techniques and cross-validation for the development and assessment of prediction models. BMC Bioinformatics 16:363
22. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE 2921–2929
23. Nair A, Bartlett EC, Walsh SLF et al (2018) Variable radiological lung nodule evaluation leads to divergent management recommendations. Eur Respir J 52:1801359
24. Goutte C, Gaussier E (2005) A probabilistic interpretation of precision, recall and F-score, with implication for evaluationProceedings of the 27th European conference on Advances in Information Retrieval Research 345–359
25. Son JY, Lee HY, Kim JH et al (2016) Quantitative CT analysis of pulmonary ground-glass opacity nodules for distinguishing invasive adenocarcinoma from non-invasive or minimally invasive adenocarcinoma: the added value of using iodine mapping. Eur Radiol 26:43–54
26. Yanagawa M, Niioka H, Hata A et al (2019) Application of deep learning (3-dimensional convolutional neural network) for the prediction of pathological invasiveness in lung adenocarcinoma: a preliminary study. Medicine (Baltimore) 98:e16119
27. Kim H, Park CM, Koh JM, Lee SM, Goo JM (2014) Pulmonary subsolid nodules: what radiologists need to know about the imaging features and management strategy. Diagn Interv Radiol 20:47–57
28. Zhang Q, Cao R, Shi F, Nian Wu Y, Zhu S-C (2017) Interpreting CNN knowledge via an explanatory graph. arXiv e-prints. Available via https://arxiv.org/abs/1708.01785. Accessed 1 Feb 2021
29. Olah C, Mordvintsev A, Schubert L (2017) Feature visualization. Distill. https://doi.org/10.23915/distill.00007
30. DeepDreaming with TensorFlow (2016) Available via https://colab.research.google.com/github/tensorflow/examples/blob/master/community/en/r1/deepdream.ipynb. Accessed 1 Feb 2021
31. Olah C, Satyanarayan A, Johnson I et al (2018) The building blocks of interpretability. Distill. https://doi.org/10.23915/distill.00010
32. Schubert L, Petrov M, Carter S, Cammarata N, Goh G, Olah C (2020) OpenAI microscope. OpenAI. Available via https://openai.com/blog/microscope/. Accessed 1 Feb 2021
33. Coudray N, Ocampo PS, Sakellaropoulos T et al (2018) Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med 24:1559–1567
34. Wang S, Shi J, Ye Z et al (2019) Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning. Eur Respir J 53:1800986
35. Nakkiran P, Kaplun G, Bansal Y, Yang T, Barak B, Sutskever I (2019) Deep double descent: where bigger models and more data hurt. arXiv e-prints. Available via https://arxiv.org/abs/1912.02292. Accessed 1 Feb 2021