Original Article

An uncertainty-aware self-training framework with consistency regularization for the multilabel classification of common computed tomography signs in lung nodules

Ketian Zhan[1], Yunpeng Wang[2], Yaoyao Zhuo[3], Yi Zhan[4], Qinqin Yan[4], Fei Shan[4], Lingxiao Zhou[5], Xinrong Chen[1], Lei Liu[1,6,7]

[1]Academy for Engineering & Technology, Fudan University, Shanghai, China; [2]Institutes of Biomedical Sciences, Fudan University, Shanghai, China; [3]Department of Radiology, Zhongshan Hospital, Fudan University, Shanghai, China; [4]Department of Radiology, Shanghai Public Health Clinical Center, Fudan University, Shanghai, China; [5]Institute of Microscale Optoelectronics, Shenzhen University, Shenzhen, China; [6]Intelligent Medicine Institute, Fudan University, Shanghai, China; [7]Shanghai Institute of Stem Cell Research and Clinical Translation, Shanghai, China

Contributions: (I) Conception and design: All authors; (II) Administrative support: All authors; (III) Provision of study materials or patients: Y Zhuo, Y Zhan, Q Yan; (IV) Collection and assembly of data: K Zhan; (V) Data analysis and interpretation: K Zhan; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Lei Liu, PhD. Academy for Engineering & Technology, Fudan University, Shanghai, China; Intelligent Medicine Institute, Fudan University, Shanghai, China; Shanghai Institute of Stem Cell Research and Clinical Translation, 138 Yixueyuan Rd., Shanghai 200032, China. Email: liulei_sibs@163.com; Xinrong Chen, PhD. Academy for Engineering & Technology, Fudan University, 220 Handan Rd., Shanghai 200433, China. Email: chenxinrong@fudan.edu.cn; Lingxiao Zhou, PhD. Institute of Microscale Optoelectronics, Shenzhen University, 3688 Nanhai Avenue, Shenzhen 518000, China. Email: lingxiaoz@szu.edu.cn.

Background: Computed tomography (CT) signs of lung nodules play an important role in indicating lung nodules' malignancy, and accurate automatic classification of these signs can help doctors improve their diagnostic efficiency. However, few relevant studies targeting multilabel classification (MLC) of nodule signs have been conducted. Moreover, difficulty in obtaining labeled data also restricts this avenue of research to a large extent. To address these problems, a multilabel automatic classification system for nodule signs is proposed, which consists of a 3-dimensional (3D) convolutional neural network (CNN) and an efficient new semi-supervised learning (SSL) framework.

Methods: Two datasets were used in our experiments: Lung Nodule Analysis 16 (LUNA16), a public dataset for lung nodule classification, and a private dataset of nodule signs. The private dataset contains 641 nodules, 454 of which were annotated with 6 important signs by radiologists. Our classification system consists of 2 main parts: a 3D CNN model and an SSL method called uncertainty-aware self-training framework with consistency regularization (USC). In the system, supervised training is performed with labeled data, and simultaneously, an uncertainty-and-confidence-based strategy is used to select pseudo-labeled samples for unsupervised training, thus jointly realizing the optimization of the model.

Results: For the MLC of nodule signs, our proposed 3D CNN achieved satisfactory results with a mean average precision (mAP) of 0.870 and a mean area under the curve (AUC) of 0.782. In semi-supervised experiments, compared with supervised learning, our proposed SSL method could increase the mAP by 7.6% (from 0.730 to 0.806) and the mean AUC by 8.1% (from 0.631 to 0.712); it thus efficiently utilized the unlabeled data and achieved superior performance improvement compared to the recently advanced methods.

Conclusions: We realized the optimal MLC of lung nodule signs with our proposed 3D CNN. Our proposed SSL method can also offer an efficient solution for the insufficiency of labeled data that may exist in the MLC tasks of 3D medical images.

Keywords: Computed tomography signs; multilabel classification (MLC); semi-supervised learning (SSL); computed tomography (CT); 3D convolutional neural networks

Submitted Jan 19, 2023. Accepted for publication Jun 08, 2023. Published online Jul 03, 2023.
doi: 10.21037/qims-23-40
View this article at: https://dx.doi.org/10.21037/qims-23-40

Introduction

Lung cancer is currently the second most common cancer in the world and is also the leading cause of cancer death (1). For all stages combined, the 5-year relative survival rate in lung cancer is relatively low, at approximately 22% (2). Therefore, the early screening of lung nodules is of great significance. From computed tomography (CT) images, radiologists can form a reliable, preliminary diagnosis of nodules' malignance according to the size and shape of nodules, which greatly improves the early detection rate of lung cancer. In the process of diagnosis, common lung CT signs provide important evidence for radiologists to diagnose diseases in patients. CT signs are strongly associated with specific thoracic disorders, which can increase diagnostic specificity (3) and are often used as the qualitative features to predict the invasiveness of lung cancer (4). As shown in *Figure 1*, recognition of these signs is an important basis for radiologists to differentiate lung nodules in CT. Therefore, accurate automatic classification of nodule signs can help doctors improve diagnostic efficiency. Moreover, there is often a certain degree of subjectivity in judging signs, and thus automatic classification can reduce errors caused by radiologists' subjective judgments. Additionally, end-to-end diagnostic models are being increasingly applied in nodule classification tasks (5,6), but it is not clear how the models actually work (7). This black-box diagnostic mode is often questioned in clinical applications. Doctors' trust in artificial intelligence will be greatly reduced under external time pressure. Thus, providing doctors with more explainable evidence is more important than providing an unexplainable answer (8), and nodule signs classification can meet this need well. Overall, building an automatic classification system of nodule signs would be a highly worthwhile endeavor.

CT sign classification models can be divided into 3 types: binary classification models, multiclassification models, and multilabel classification (MLC) models. First, the binary classification model is typically built for a certain sign. Through use of hybrid resampling and feature fusion, 2 classification models for ground-glass opacity and the cavity in nodules, respectively, have been designed (9,10). In one study, the snake model (active contour model) was applied to extract the precise contour information of a nodule and to help build a spiculation recognition model (11). The binary classification model for a certain CT sign often makes full use of the shape and texture characteristics of this sign and tends to have good performance in the classification. However, when the model is used for other signs, its performance will drop sharply. If one model is designed for each sign, the overall classification efficiency is bound to decline. Second, the multiclassification model is a model that outputs only 1 positive class from multiple classes. For example, a fused classification model was proposed to identify 9 kinds of lung CT signs (12). Image retrieval methods are also used to realize CT sign classification (13,14). He (15) proposed using images of signs generated by a generative adversarial network (GAN) to pretrain the convolutional neural network (CNN) model, and then use real data to fine-tune the model. These multiclassification models are trained on multiclassification sign datasets, which means that each image in the dataset contains only 1 sign. This leads to 2 major issues: (I) the process does not conform to the actual clinical application scenario, and (II) the data labeling process is extremely tedious. A nodule usually contains more than one sort of sign, so when the general area of a nodule is located, it is hoped that an efficient sign classification model can diagnose all signs contained in the entire area of the lesion at a time. The multiclassification model can output only 1 positive class at a time and therefore cannot meet our needs. Moreover, the acquisition of multiclassification sign annotation requires radiologists to additionally crop each sign, representing an inevitable and massive increase in the workload. As an MLC model can output multiple positive classes at a time, what is needed is an MLC model that can evaluate all the signs in the nodular lesion area simultaneously. Accordingly, radiologists would only need to annotate the nodules with global annotations of signs, which would greatly reduce the annotation workload.

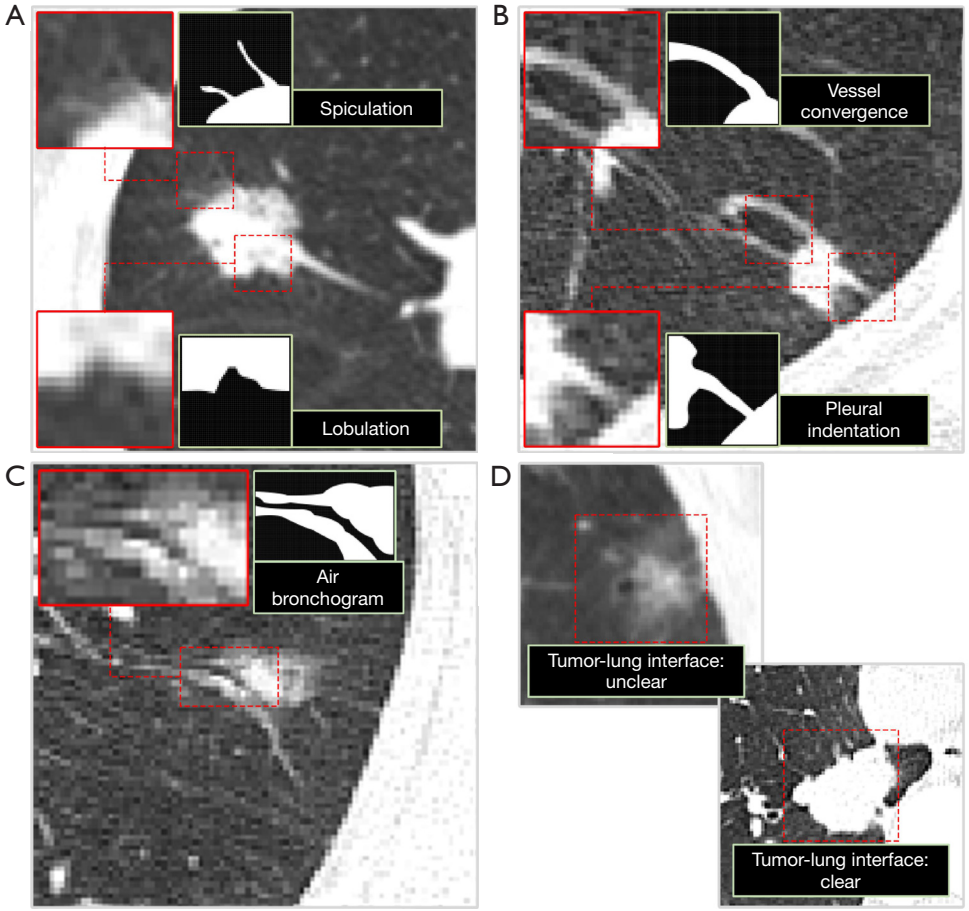

Figure 1 Examples of common lung nodule CT signs. (A) Spiculation and lobulation; (B) vessel convergence, pleural indentation; (C) air bronchogram; (D) tumor-lung interface. The nodule CT sign in the red dotted box is enlarged and displayed in the red solid line box in the upper left corner or lower left corner of the subfigure. To better display the characteristics of different nodule signs, the corresponding binary image is given in the light green box. CT, computed tomography.

Semi-supervised learning (SSL) can use a small amount of labeled data and a large amount of unlabeled data in the training process, in which unlabeled data can also be used to improve model performance. Considering the difficulty of acquiring sign annotation, this paper further explores the implementation of an SSL method in the MLC task of nodule signs. A variety of SSL methods have been proposed for image classification, among which consistency training (16,17) and pseudo-labeling (18-20) are the most commonly used. Consistency training is based on the consistency regularization method, which emphasizes that the realistic perturbations of the data points should not change the output of the model (21). Pseudo-labeling is based on entropy minimization (22), which combines unlabeled data and low-entropy predictions for model optimization in a supervised manner (23). Most of the advanced SSL methods proposed in recent years use both consistency regularization and entropy minimization (24-27). In recent years, the SSL methods mentioned above have also been widely used in the field of medical image classification. For example, virtual adversarial training and consistency regularization were combined for the ultrasound image screening of breast cancer and for the multiclassification tasks in ophthalmic diseases (28). It was also reported that a similarity metric function in the semantic representation space for pseudo-label selection was used to iteratively incorporate unlabeled samples so as to optimize the model for the diagnosis of benign and malignant lung nodules (29). Effective SSL methods have also been designed for the MLC of thorax disease, which have been tested on 2D chest X-ray

images. In contrast to the traditional consistency-based method which enforces the prediction consistency, a sample relation consistency method was proposed by Liu *et al.* (30), focusing on the consistency of the intrinsic relation among different unlabeled samples. To solve the imbalanced classification problem in SSL, Liu *et al.* (31) incorporated an anticurriculum learning method in pseudo-labeling. However, to the best of our knowledge, few studies have implemented SSL methods in MLC tasks of 3D CT. The advanced SSL methods that have been proposed (24-27) cannot be directly applied in the MLC of CT images. For one, these methods (using the softmax layer in the network) (24,26) are designed only for single-label classification. For another, the augmentation operation mixup (32) used in these SSL methods has been proven to be ineffective when extended to MLC (33). Additionally, strong augmentation is a crucial component in these SSL methods (25-27). However, some strong augmentation operations may severely alter the appearance of the images, making it difficult to identify lesion areas, which potentially makes them unsuitable for CT images.

Overall, our study aimed to overcome the following issues: (I) current classification models of CT signs are mostly designed for identifying individual signs, fail to judge all signs in the nodular lesion area simultaneously, and thus are impracticable for clinical application; (II) medical imaging labels are often difficult to obtain, and the process of annotating nodule CT signs is particularly time-consuming and labor intensive for radiologists. Therefore, we designed a 3D CNN which combines the dense block and convolutional block attention modules (CBAMs) for the MLC of nodule signs, realizing simultaneous recognition of multiple signs in nodular lesions for the first time and propose a new SSL method: the uncertainty-aware (UA) self-training framework with consistency regularization (USC). This method was applied in the MLC task of nodule signs and can also be extended to other similar scenarios. We present this article in accordance with the TRIPOD reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-23-40/rc).

Methods

Datasets

The lung nodules examined in this study were solitary pulmonary nodules (SPNs), which are defined as localized, circular or elliptical lesions in the lung with a diameter of ≤3 cm. Many studies on SPNs have used the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) public dataset. In addition to benign and malignant annotation, the LIDC-IDRI dataset also provides some sign annotations, such as lobulation and spiculation (34). However, to our knowledge, there are other signs that can also affect the diagnosis of nodules. The presence of spiculation, lobulation, and vessel convergence (VC), and particularly of pleural indentation (PI) and air bronchogram (AB), is often indicative of a malignant nodule (35). As the LIDC dataset could not meet our needs for the MLC of nodule signs, we collected CT images of lung nodules with the annotation of several common signs and constructed our own nodule sign dataset. This retrospective study was conducted in accordance with the Declaration of Helsinki (as revised in 2013) and approved by the ethics committee of the Shanghai Public Health Clinical Center. By virtue of the retrospective nature of study, the collection of patient CT images and relevant clinical information would not adversely affect the patient's rights or welfare, and thus need for individual consent was waived by the committee.

The datasets used in this study include 1 public dataset and 1 private dataset. The public dataset is the Lung Nodule Analysis 16 (LUNA16) dataset (36), which is a subset of the LIDC-IDRI dataset and was used for pretraining in this study. The LUNA16 dataset contains a total of 1,186 nodules, which were annotated by multiple radiologists for malignancy on a rating scale of 1–5, with 1 indicating the lowest malignancy and 5 indicating the highest malignancy. In this study, the average rating of multiple ratings for the same nodule was used as the final assessment of its malignancy, and nodules with an average rating ≤2.5 were considered to be benign, while those with an average rating ≥3.5 were considered to be malignant. As a result, a total of 811 nodules were selected as the pretraining dataset (*Table 1*).

The private dataset of nodule CT signs was collected and annotated by the Shanghai Public Health Clinical Center and contains 641 nodules from 3 hospitals: Zhongshan Hospital Affiliated with Fudan University, the Second People's Hospital of Ningbo Beilun District, and the Cancer Hospital Affiliated with Fudan University. Among the nodules, 454 samples are labeled. The data annotation in this study was performed by 2 radiologists (Y Zhuo, with 7 years of chest radiological experience and F Shan, with more than 18 years of chest radiological

Table 1 The label distribution of the public and private datasets

Label	LUNA16	Nodule CT signs					
	Malignance	Lobulation	Spiculation	TLI	AB	VC	PI
1	403	338	305	360	261	301	227
0	408	116	149	94	193	153	227

LUNA16, Lung Nodule Analysis 16; CT, computed tomography; TLI, tumor-lung interface; VC, vessel convergence; AB, air bronchogram; PI, pleural indentation.

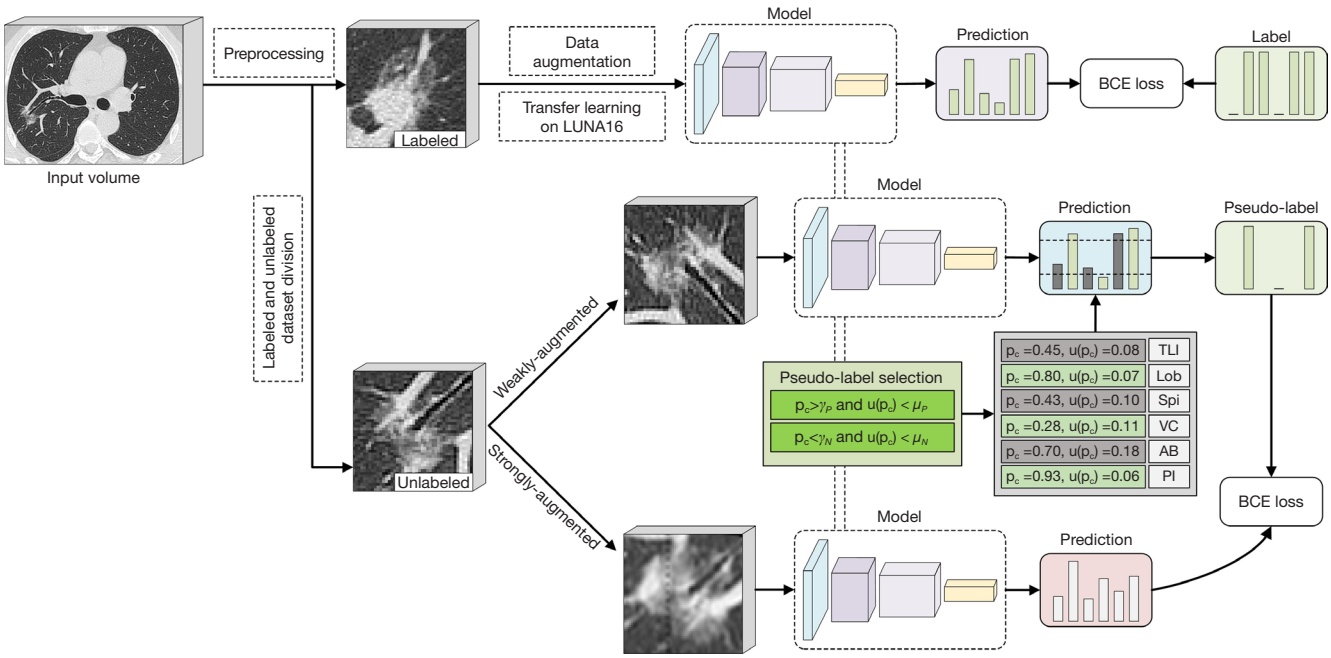

Figure 2 An overview of our proposed USC framework. After preprocessing, the image is divided into labeled data and unlabeled data. The supervised training (the first line) is performed on the labeled data, and the supervised BCE loss is calculated between the model predictions and labels; the consistency training (the second and third line) is performed on the unlabeled data. During the consistency training process, selected pseudo-labels of weakly augmented samples are used to supervise the prediction of strongly augmented samples, and the consistency BCE loss is calculated between the predictions and the pseudo-labels. Pseudo-label selection is based on confidence and uncertainty. The confidence thresholds of positive and negative pseudo-labels set in this paper are 0.75 and 0.35, respectively (γ_P =0.75 and γ_N =0.35), while the uncertainty thresholds are 0.08 and 0.12, respectively (μ_P =0.08 and μ_N =0.12). USC, uncertainty-aware self-training framework with consistency regularization; BCE, binary cross entropy; LUNA16, lung nodule analysis 16.

experience), and in cases of ambiguous labeling results, a consensus was reached through discussion. Six common signs related to lung nodules were annotated: lobulation, spiculation, tumor-lung interface (TLI), AB, VC, and PI. The distribution of label categories in the dataset is shown in *Table 1*. Generally, 1 indicates the presence of the sign in the nodule, and 0 indicates its absence. Specifically, for the TLI sign, 1 represents a clear interface, while 0 represents an unclear interface.

Workflow

The framework proposed in this paper consists of 3 main stages: (I) image preprocessing and data preparation stage, which involves preprocessing the CT images, performing the train-test split, and dividing the labeled and unlabeled data; (II) nodule sign classification model training stage for training the initial MLC model, in which our proposed 3D CNN (Model in *Figure 2*) is first pretrained with LUNA16

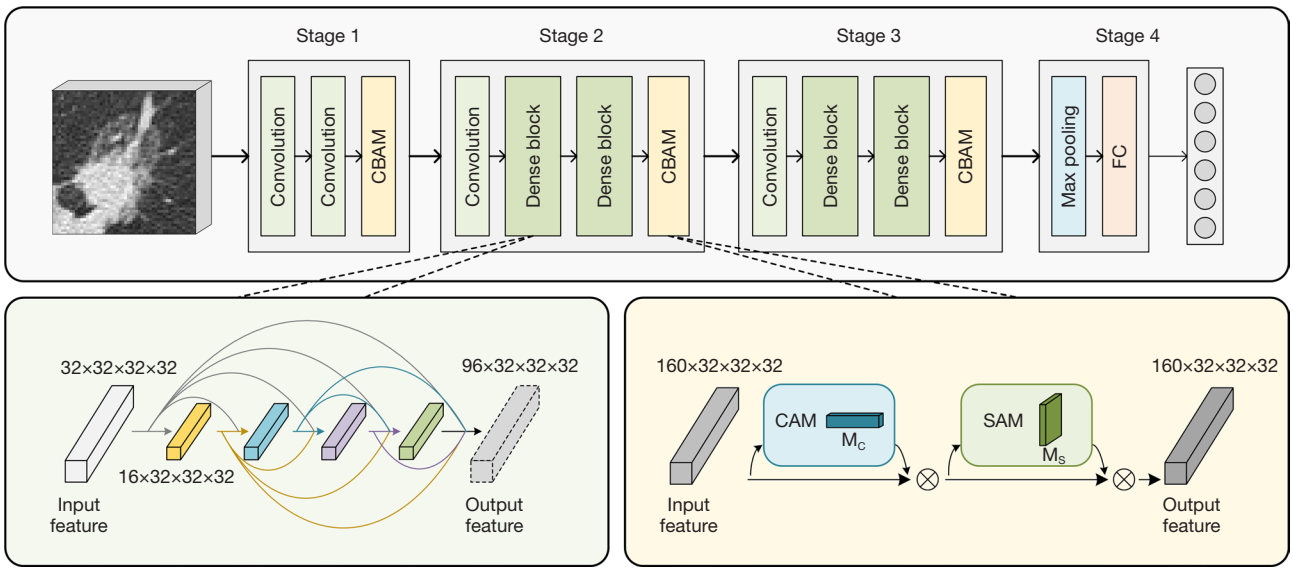

Figure 3 The overall architecture of our proposed 3D CNN. 3D CNN, 3-dimensional convolutional neural network; CBAM, convolutional block attention module; FC, fully connected layer; CAM, channel attention module; SAM, space attention module.

dataset and then fine-tuned with the labeled dataset of nodule signs; and (III) a USC semi-supervised learning stage for model optimization using both the labeled and unlabeled dataset. In the latter stage, the framework selectively uses the pseudo-labels generated by the fine-tuned MLC model and the strongly augmented samples in a supervised manner. Simultaneously, the labeled data are used to fine-tune the model continuously. These 2 processes are repeated in every epoch. The overall process of the framework is shown in *Figure 2*.

Data preprocessing

All nodules used in this study were between 5 and 30 mm in size, and the slice thickness was less than 1.5 mm. CT images that failed to meet the requirements were screened out from the dataset in advance. For all the CT images, we carried out the following preprocessing operations: (I) to have a clearer observation of the nodular signs, we adjusted the CT window width and level to the lung window [a window width of 1,500 Hounsfield unit (HU) and a window level of –500 HU]; (II) we resampled the CT image according to the spacing of 0.5 mm × 0.5 mm × 0.5 mm; (III) we used the mask center as the nodule center to extract a 3D cube of 64×64×64 pixels, which ensured that the cube contained all the nodule signs while reducing the inclusion of irrelevant lung tissue surrounding the nodule as much as possible.

MLC method

CNNs are commonly used in artificial intelligence-driven computer-aided diagnosis systems and have demonstrated promising prediction accuracy for lung nodule risk stratification (37). According to the input, CNN-based diagnosis models in CT image analysis can be mainly divided into the following 3 types: (I) 2D CNN using a single slice as input (38); (II) 2D CNN using multiple slices as input, generally consisting of images from different perspectives or scales (39,40); and (III) 3D CNN using the entire 3D volume as input (41,42). As various nodule signs may appear in different slices, finding a slice in the CT series that contains all the nodule signs is highly unrealistic. Another major problem related to the 2D CNN using multiple images as input is that it sees the CT series as many independent slices, ignoring useful continuous information between adjacent slices. Because the 3D convolutional kernel can leverage interslice context, we built a 3D CNN model for the MLC of nodule signs. Our proposed 3D CNN was constructed using dense blocks and CBAMs. The dense block enhances feature reuse through dense connections (43), while the CBAM builds both spatial and channel attention maps to incorporate global information into the network (44). Specifically, the 3D CNN contains 3 stages (*Figure 3*). Stage 1 contains two 3D convolutional layers with a kernel size of 3 and a CBAM (44). Both stages 2 and 3 are composed of a 3D

convolutional layer with a stride of 2, two dense blocks (43), a CBAM, and a dropout layer (dropout rate =0.3). Stage 4 consists of an adaptive max pooling layer, a fully connected (FC) layer, and a sigmoid layer. As shown in *Figure 3*, each dense block contains 4 bottlenecks with a growth rate of 16, meaning that after each dense block, the number of feature channels increases by 64. CBAM consists of 2 modules, the channel attention module (CAM) and the spatial attention module (SAM). CAM and SAM respectively compute the attention weights for feature channels and spatial locations to optimize the features jointly.

It is well known that the performance of deep learning models relies on representation learning based on massive data (45). Training deep learning models with a small dataset often leads to overfitting, so model pretraining has become a basic step. Many image classification tasks are first pretrained on ImageNet, but due to the large differences in features between natural images and medical images, especially grayscale images such as CT, pretraining on ImageNet has been proven to be ineffective in medical image-related tasks (46). Therefore, in this study, we used the LUNA16 dataset to pretrain the model.

Two classical models, 3D DenseNet (43) and 3D ResNet (47), were reimplemented for comparison. We also included 2 other models, Local_Global (38) and NASLung (42), for comparison; the former is a 2D model, while the latter is a 3D model, and both have achieved state-of-the-art performance in benign and malignant lung nodule classification. We examined their applicability to the classification of nodule signs.

SSL method

The training of the 3D model requires a large amount of labeled data, and as mentioned above, nodule sign annotation is labor intensive, while the acquisition of unlabeled nodule data is comparatively convenient. Consequently, we thought to use an SSL method to improve MLC performance. In this section, we describe, in detail, our proposed SSL method, USC. Our proposed method consists of 2 parts: self-training and consistency regularization.

For the MLC task, we use $X_l = \left\{\left(x_l^{(i)}, y_l^{(i)}\right)\right\}_{i=1}^{N_l}$ to denote a labeled dataset with N_l samples, where $x_l^{(i)}$ denotes the input sample, $y_l^{(i)} = \left[y_1^{(i)}, \ldots, y_C^{(i)}\right] \subseteq \{0,1\}^C$ denotes the corresponding ground truth label, and C denotes the number of labels for MLC. $y_c^{(i)} = 1$ indicates that the label exists in the i-th sample, while $y_c^{(i)} = 0$ indicates that the label does not exist in the i-th sample. Accordingly, in this paper, $y_c^{(i)}$ denotes whether a particular sign is present in the sample. Similarly, $X_u = \left\{\left(x_u^{(i)}\right)\right\}_{i=1}^{N_u}$ denotes an unlabeled dataset with N_u samples. Therefore, the entire dataset can be represented as $X=\{X_l, X_u\}$.

Consistency regularization

Recently, advanced methods in the semi-supervised field have typically used the idea of consistency regularization. Consistency regularization is based on the manifold assumption, which requires that perturbations applied to input do not cause a change in model output (23). The idea of consistency regularization is generally incorporated in semi-supervised methods in the form of the consistency loss term. More specifically, for an unlabeled sample x_u, the consistency loss is defined as follows:

$$\left\| P\left(y \mid \alpha\left(x_u\right), \theta\right) - P\left(y \mid \alpha'\left(x_u\right), \theta\right) \right\|_2^2 \quad [1]$$

where $P\left(y \mid x, \theta\right)$ is used to denote the output distribution of the model whose parameter is θ, and α and α' are used to denote the random perturbation applied to the sample. Due to the random perturbation, the inputs of the model are not the same, thus leading to a difference in the outputs. The L2 loss is used to measure the difference (i.e., the consistency loss), and the Kullback-Leibler (KL) divergence loss (48) and cross-entropy loss (25,26) are also used in some studies.

Self-training

The self-training algorithm regards high-confidence predictions as the pseudo-labels for unlabeled data (20). More specifically, it uses the base classification model trained on a small number of labeled samples to predict pseudo-labels for unlabeled samples and then selectively uses the pseudo-labeled samples in model optimization together with the labeled samples.

After the pseudo-label inference is completed by the base classification model, the common strategy of pseudo-label selection is to determine whether it reaches a certain confidence threshold. For MLC, the output of the sample through the sigmoid layer of the model represents the confidence that the label exists in the sample. We use $p^{(i)}$ to denote the model prediction on the i-th sample and $p_c^{(i)}$ to

denote the prediction of the *c*-th class in the *i*-th sample, with $c \in [1, C]$ and *C* being the total number of class labels for MLC. The pseudo-label selection conforms to the following formula:

$$g_c^{(i)} = I\left(p_c^{(i)} > \gamma_P\right) + I\left(p_c^{(i)} < \gamma_N\right) \quad [2]$$

where I denotes the indicator function, and γ_P and γ_N denote the confidence thresholds of positive and negative pseudo-labels, respectively. If $p_c^{(i)} > \gamma_P$, the positive pseudo-label corresponding to class *c* will be selected; if $p_c^{(i)} < \gamma_N$, the negative pseudo-label corresponding to class *c* will be selected.

UA self-training framework with consistency regularization

In the SSL method based on self-training, the quality of pseudo-labels will directly affect the classification performance. If the pseudo-labels we use in model training are reliable and stable, with the addition of plentiful pseudo-labeled samples, the robustness of the model will be continuously enhanced, while if incorrect pseudo-labels are overused, much noise will be introduced into the model training process, resulting in a decline in the performance of model classification. Using confidence and uncertainty simultaneously for pseudo-label selection can improve the accuracy of the pseudo-labels used for model training and greatly reduce the introduction of noise during training (20). To obtain more stable and reliable pseudo-labeled samples, we also incorporated the above strategy into our method. More specifically, as shown in *Figure 2*, we first use labeled samples to train the initial MLC model, and then the model is used to infer labels on weakly augmented unlabeled samples. During the inference process, Monte Carlo dropout (49) is used. This means the dropout layer is kept open in the network to add perturbations to the model. Inference is repeated for each unlabeled sample 10 times, and the mean and variance of multiple inference results are each calculated, with the mean being the confidence and the variance being the uncertainty. According to the confidence and uncertainty, we then decide whether to select the inference result of the sample as a pseudo-label. Only when both the confidence and uncertainty requirements are met, will we select the pseudo-label and use it for self-training. For unlabeled samples x_u, whether or not the prediction of the *c*-th class label in the *i*-th sample can be used as a pseudo-label is determined as follows:

$$g_c^{(i)} = I\left(u\left(p_c^{(i)}\right) < \mu_P\right) \times I\left(p_c^{(i)} > \gamma_P\right) + I\left(u\left(p_c^{(i)}\right) < \mu_N\right) \times I\left(p_c^{(i)} < \gamma_N\right) \quad [3]$$

where $p_c^{(i)}$ is the prediction of the *c*-th class in the sample $x_u^{(i)}$, and *u* is a function used for calculating the uncertainty of the prediction. γ_P and γ_N respectively represent the confidence thresholds, and similarly, μ_P and μ_N respectively represent the uncertainty thresholds. These 4 hyperparameters are used to judge whether the prediction is a qualified positive or negative pseudo-label. $g_c^{(i)} \in \{0,1\}$ acts as a flag, indicating if the prediction can be selected as a pseudo-label. Meanwhile, the same unlabeled sample will also be strongly augmented, the prediction of which can then be obtained through the model. Finally, a consistency loss is constructed between the prediction of the strongly augmented image and the pseudo-label based on $g_c^{(i)}$. The pseudocode of our proposed method can be found in *Figure 4*.

In the training of the overall network, the loss consists of 2 parts: the supervised loss of labeled data and the consistency loss of unlabeled data, which is simply $\ell_s + \lambda_u \ell_u$, where λ_u is a fixed coefficient used to represent the relative weight of the consistency loss. Both ℓ_s and ℓ_u are calculated in a custom binary cross-entropy (BCE) loss form, which can be uniformly expressed as follows:

$$\begin{aligned} L_{bce}\left(\hat{y}^{(i)}, \tilde{y}^{(i)}, g^{(i)}\right) &= \frac{1}{\sum g^{(i)}} \sum_{c=1}^{C} g_c^{(i)} \cdot D_{bce}\left(\hat{y}_c^{(i)}, \tilde{y}_c^{(i)}\right) \\ &= -\frac{1}{\sum g^{(i)}} \sum_{c=1}^{C} g_c^{(i)} \left[\tilde{y}_c^{(i)} \log \hat{y}_c^{(i)} + \left(1 - \tilde{y}_c^{(i)}\right) \log\left(1 - \hat{y}_c^{(i)}\right)\right] \end{aligned} \quad [4]$$

where $\hat{y}^{(i)}$, $\tilde{y}^{(i)}$, $g^{(i)}$ and are all *C*-dimensional vectors; $\hat{y}^{(i)}$ denotes the model prediction for the *i*-th sample; and $\tilde{y}^{(i)}$ denotes the label of the *i*-th sample. For the labeled sample, it is its original ground truth label, and for the incorporated pseudo-labeled sample, it is its pseudo-label. $g^{(i)}$ consists of $g_c^{(i)}$. For a labeled sample, its $g_c^{(i)}$ is all 1, while for a pseudo-labeled sample, the value depends on whether the inference outcome meets the requirements of confidence and uncertainty, as is defined in Equation [3]. If the inference outcome of the class *c* meets the requirements, $g_c^{(i)}$ equals 1; otherwise, $g_c^{(i)}$ equals 0. Thus, the indeterminate pseudo-label will be ignored in calculating the loss. $D_{bce}\left(\hat{y}_c^{(i)}, \tilde{y}_c^{(i)}\right)$ denotes the BCE between $\hat{y}_c^{(i)}$ and $\tilde{y}_c^{(i)}$.

Considering that few studies have been conducted on

Algorithm 1 Pseudocode of uncertainty-aware self-training with consistency regularization.

Input: A Batch of labeled CT volumes (batchsize=b) and corresponding labels $B_l = \{(x_l^{(i)}, \mathbf{y}_l^{(i)})\}_{i=1}^{N_l}$, a batch of unlabeled CT volumes $B_u = \{(x_u^{(i)})\}_{i=1}^{N_u}$, confidence threshold γ_p, γ_n, uncertainty threshold μ_p, μ_n, weighting coefficient λ_u.

Output: Network parameters θ .

for $i = 1$ to b **do**
  $\hat{x}_l^{(i)} \leftarrow \phi(x_l^{(i)})$ — Apply weak augmentation to labeled sample $x_l^{(i)}$
  $\hat{x}_u^{(i)} \leftarrow \phi(x_u^{(i)})$ — Apply weak augmentation to unlabeled sample $x_u^{(i)}$
  $\tilde{x}_u^{(i)} \leftarrow \Phi(x_u^{(i)})$ — Apply strong augmentation to unlabeled sample $x_u^{(i)}$
  $p_l^{(i)} \leftarrow P(\mathbf{y} \mid \hat{x}_l^{(i)}; \theta)$ — Compute prediction for weak-augmented $\hat{x}_l^{(i)}$
  $p_u^{(i)} \leftarrow P(\mathbf{y} \mid \hat{x}_u^{(i)}; \theta)$ — Compute prediction for weak-augmented $\hat{x}_u^{(i)}$
  $\tilde{p}_u^{(i)} \leftarrow P(\mathbf{y} \mid \tilde{x}_u^{(i)}; \theta)$ — Compute prediction for strong-augmented $\tilde{x}_l^{(i)}$
  for $c = 1$ to C **do** — Check the pseudo-label's qualification for each class in each $p_u^{(i)}$
    $g_c^{(i)} \leftarrow \mathbb{1}\left[\mu(p_c^{(i)}) < \mu_p\right] \mathbb{1}\left[p_c^{(i)} > \gamma_p\right] + \mathbb{1}\left[\mu(p_c^{(i)}) < \mu_n\right] \mathbb{1}\left[p_c^{(i)} < \gamma_n\right]$
  end for
end for
$\ell_s \leftarrow \sum_{i=1}^{b} \sum_{c=1}^{C} \mathcal{D}_{\text{bce}}\left[\hat{\text{y}}_c^{(i)}, \tilde{\text{y}}_c^{(i)}\right]$ — Compute the supervised loss
$\ell_u \leftarrow \sum_{i=1}^{b} \sum_{c=1}^{C} g_c^{(i)} \mathcal{D}_{\text{bce}}\left[\hat{\text{y}}_c^{(i)}, \tilde{\text{y}}_c^{(i)}\right]$ — Compute the consistency loss
$loss \leftarrow \ell_s + \lambda_u \ell_u$
update θ using Adam optimizer
return θ

Figure 4 Pseudocode of USC. CT, computed tomography; USC, uncertainty-aware self-training framework with consistency regularization.

SSL for an MLC task of 3D medical images, to further verify the effectiveness of our proposed semi-supervised method, we reimplemented 2 works: FixMatch, which is also based on self-training and consistency regularization methods and which is capable of achieving state-of-the-art performance across many standard SSL benchmarks (27); and UPS, which is one of the most recent of the proposed SSL methods that only uses the pseudo-labeling method and which performs on par with consistency regularization-based SSL methods (20). We adapted these 2 methods to be suitable for the MLC task and compared their performance with that of our proposed USC method.

Augmentation in our proposed USC

In our proposed USC method, 2 versions of 3D image augmentation methods are applied: "weak" and "strong". We adopt the flip-and-crop strategy in the weak augmentation. Specifically, it includes the random flip along the X, Y, and Z axes and a random crop operation after padding. As for strong augmentation, apart from the traditional augmentation transformations such as Flip, Rotate, ElasticTransform, and GaussianNoise, considering the characteristics of CT images, we also incorporate RandomCropBorder and RandomDropPlane operations (50). The specific augmentation transformations and corresponding probabilities of using them in the weak and strong augmentation methods are shown in *Table 2*.

Experiment

Dataset partition and experimental setup

In our experiments, the nodule sign dataset was split into training and testing sets at the nodule level for a realistic evaluation. For MLC method assessment, we used all 454 labeled samples and implemented a 5-fold cross-validation. For SSL method assessment, based on the same 5-fold cross-validation split, we selected a small portion of the training set as labeled data, removed the labels of the remaining training set, and mixed them with 187 unlabeled samples to form the unlabeled data. Eventually, the labeled data accounted for 15% of the entire training dataset, which conforms to the basic assumption of semi-supervised learning of labeled data being far less abundant than unlabeled data. The overall dataset partition is shown in *Figure 5*.

In the training phase of the MLC model, we used Adam optimizer with the learning rate ranging from 1e-3 to 3e-4 for different models with a batch size of 16 and trained each model for 100 epochs. During training, we performed the following augmentation operations on the input data: random flip along each of the axes and random crop (fist padding the nodule to 68×68×68 pixels and then randomly

Table 2 List of transformations used in USC

Variables	Transformation	Description
Weak augmentation	Flip	Random flip along each of the axes, P=0.5
	Crop	First pad the volume and then crop the volume at a random location, P=1.0
Strong augmentation	Rotate	Random rotation up to 20° along the vertical axis, P=0.5
	Flip	Random flip along each of the axes, P=0.5
	ElasticTransform	Applying elastic deformation to the volume, P=0.4
	GaussianNoise	Adding Gaussian noise to the volume, P=0.5
	RandomGamma	Randomly changing the contrast of the volume, P=0.8
	GridCutout	Dropping cuboid regions of a volume in grid fashion, P=0.4
	RandomCropBorder	Randomly removing some pixels from borders, P=0.4
	RandomDropPlane	Randomly removing a few intermediate 2D planes along the vertical axis, P=0.4

USC, uncertainty-aware self-training framework with consistency regularization; 2D, two-dimensional.

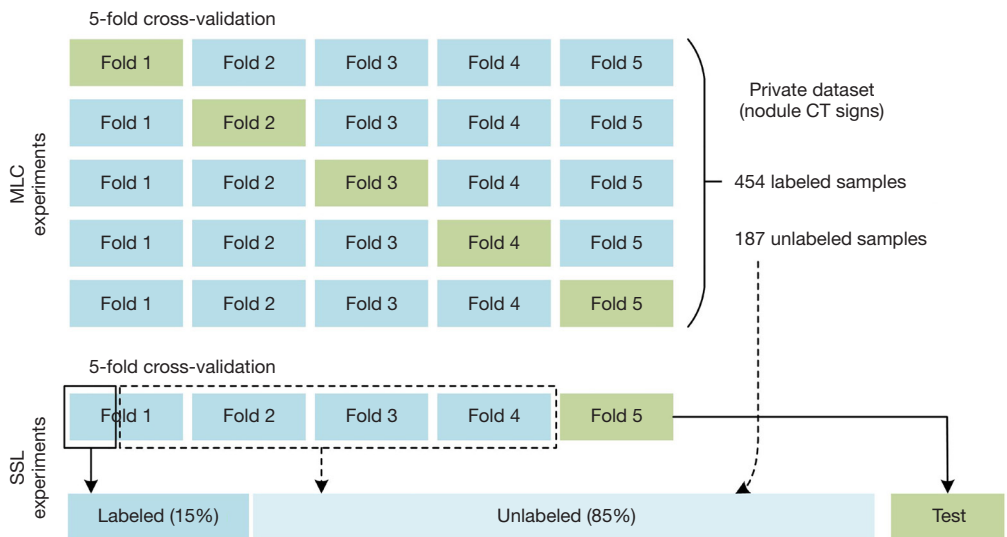

Figure 5 The dataset split into the MLC and SSL experiments. CT, computed tomography; MLC, multilabel classification; SSL, semi-supervised learning.

cropping a patch of 64×64×64 pixels).

In the experiments of SSL methods, we used our proposed 3D CNN with transfer learning as the MLC model. We first fine-tuned the pretrained MLC model with labeled data using the Adam optimizer with a learning rate of 1e-4 and a batch size of 16 for 100 epochs and then further optimized the model via our proposed USC method with a learning rate of 3e-5 for another 100 epochs. In the process of applying USC, we randomly resampled the labeled data to make it consistent in quantity with the unlabeled part to ensure that in each training iteration, the labeled data and unlabeled data were equal in number.

In all experiments, we used PyTorch (Linux Foundation) to build MLC models and the SSL frameworks and trained the models on 4 Titan Xp graphics cards (Nvidia). The source code is publicly available at https://github.com/oslo71/USC.

Metrics of performance evaluation

To comprehensively evaluate the MLC model and SLL

method we proposed for lung nodule sign classification, we calculated the accuracy, micro-F1, macro-F1, hamming loss, and mean average precision (mAP) as the performance metrics. The detailed definitions and formulae for these metrics are as follows.

The overall accuracy was calculated by taking the average of the accuracies of each individual class. The formula for calculating the overall accuracy is as follows:

$$Accuracy = \frac{1}{C}\sum_{c=1}^{C} Accuracy_c = \frac{1}{C}\sum_{c=1}^{C} \frac{TP_c + TN_c}{TP_c + TN_c + FP_c + FN_c} \quad [5]$$

We separately calculated 2 types of F1-score: micro-F1 and macro-F1. For macro-F1, we first calculated the precision and recall for each class, and then the macro-F1 was calculated by averaging on the F1-score of each individual class as follows:

$$Precision_c = \frac{TP_c}{TP_c + FP_c} \quad [6]$$

$$Recall_c = \frac{TP_c}{TP_c + FN_c} \quad [7]$$

$$macro\text{-}F1 = \frac{1}{C}\sum_{c=1}^{C} F1\text{-}score_c = \frac{1}{C}\sum_{c=1}^{C} \frac{2 \cdot Precision_c \cdot Recall_c}{Precision_c + Recall_c} \quad [8]$$

The micro-F1 was computed using the total true-positive (TP), false-positive (FP), and false-negative (FN) counts across all labels and samples as follows:

$$micro\text{-}Precision = \frac{\sum_{c=1}^{C} TP_c}{\sum_{c=1}^{C} TP_c + \sum_{c=1}^{C} FP_c} \quad [9]$$

$$micro\text{-}Recall = \frac{\sum_{c=1}^{C} TP_c}{\sum_{c=1}^{C} TP_c + \sum_{c=1}^{C} FN_c} \quad [10]$$

$$micro\text{-}F1 = 2 \cdot \frac{micro\text{-}Precision \times micro\text{-}Recall}{micro\text{-}Precision + micro\text{-}Recall} \quad [11]$$

The hamming loss was defined as the fraction of misclassified labels to the total number of labels of all samples as follows:

$$L_{Hamming} = \frac{1}{N_{labels}} \sum_{j=0}^{N_{labels}-1} I\left(\hat{y}_j^c \neq \tilde{y}_j^c\right) \quad [12]$$

The area under the precision–recall curve, referred to as average precision (AP), represents the AP of the classifier. By computing the AP for each class and taking the mean, we obtained the mAP of the MLC classifier.

We further calculated the area under the curve (AUC) for each sign class and the average overall AUC. We performed 5-fold cross-validation in both the MLC experiments and semi-supervised experiments, with all results being averaged on the test set.

Results

Performances of MLC models trained from scratch

In the experiment for the MLC of nodule signs, based on all the labeled data, we used the schemes of a 2D model and 3D models, respectively. The 2D model was the Local_Global model, and the 3D models were ResNet, DenseNet, NASLung, and the 3D CNN proposed in this paper. As is shown in *Table 3*, Our proposed CNN model achieved the best performance with an accuracy of 0.751, an mAP of 0.860, a hamming loss of 0.249, and an F1-micro and F1-macro of 0.820 and 0.812, respectively. The performances of NASLung and 3D DenseNet were slightly inferior to our proposed 3D CNN, but the number of parameters in these models are much higher than that of our proposed 3D CNN.

Medical-to-medical transfer learning

In our study, we adopted a medical-to-medical transfer learning strategy. The LUNA16 dataset was used to pretrain our proposed 3D CNN, and fine-tuning was then performed in the phase of MLC with the nodule sign dataset. During the model fine-tuning process, the first and second stages of the network were frozen, and only the remaining network parameters were updated. *Table 4* demonstrates that after applying the transfer learning strategy to our model, we further obtained performance improvements of 1.1% accuracy, 1.0% mAP, 0.6% F1-micro, and 1.1% hamming loss. For the results of 5-fold cross-validation, we compiled the confusion matrix of each fold to obtain the integrated confusion matrix, which is presented in *Figure 6*.

For each lung nodule sign in the MLC task, we used AUC to evaluate the classification performance of the model. *Table 5* shows that our proposed 3D CNN with transfer learning achieved an average AUC of 0.782 on all 6 signs and an average AUC of 0.805 on the 5 signs of lobulation, spiculation, VC, AB, and PI. However, for the TLI sign, our CNN performed relatively poorly.

Table 3 Performance comparison of the MLC models trained from scratch

Model	Parameter	Accuracy	mAP	F1-micro	F1-macro	Hamming loss
Local_Global (2D)	183.99 k	0.739	0.838	0.809	0.793	0.261
ResNet-34 (3D)	63,472.71 k	0.719	0.828	0.790	0.781	0.281
DenseNet-121 (3D)	11,248.77 k	0.748	0.845	0.817	0.808	0.252
NASLung (3D)	28,564.35 k	0.744	0.851	0.812	0.801	0.256
The proposed CNN	1,409.54 k	0.751*	0.860*	0.820*	0.812*	0.249*

*, the best performance in each column. MLC, multilabel classification; mAP, mean average precision; 2D, two-dimensional; 3D, three-dimensional; CNN, convolutional neural network; k, thousand.

Table 4 Performance of transfer learning

Model	Accuracy	mAP	F1-micro	F1-macro	Hamming loss
The proposed CNN	0.751	0.860	0.820	0.812	0.249
The proposed CNN + transfer learning	0.762*	0.870*	0.826*	0.816*	0.238*

*, the best performance in each column. mAP, mean average precision; CNN, convolutional neural network.

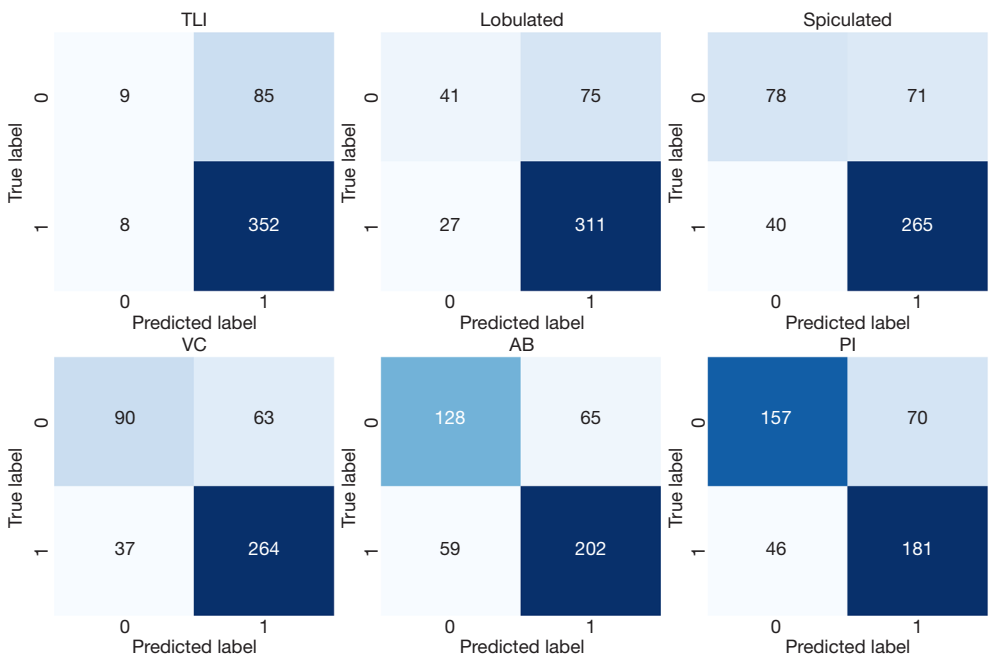

Figure 6 The confusion matrix of our proposed CNN with transfer learning. CNN, convolutional neural network; TLI, tumor-lung interface; VC, vessel convergence; AB, air bronchogram; PI, pleural indentation.

Comparison of MLC and separate binary classification

To prove the superiority of a single MLC model for all nodule signs compared to 6 separate classification models for each sign, we converted the multilabel dataset into a single-label dataset with multiple classes, used the same model architecture as our proposed 3D CNN to build 6 binary classification models, pretrained the models with LUNA16 dataset, and fine-tuned each model with the nodule sign dataset. The comparison is shown in *Table 6*. The AUC of our

Table 5 The AUC of 6 nodule signs

Model	TLI	Lobulation	Spiculation	VC	AB	PI	Average AUC
The proposed CNN + transfer learning	0.670	0.790	0.824	0.800	0.806	0.804	0.782

TLI, tumor-lung interface; VC, vessel convergence; AB, air bronchogram; PI, pleural indentation; AUC, area under receiver operating characteristic curve; CNN, convolutional neural network.

Table 6 Performance comparison of the MLC model and the binary classification model

Model	TLI	Lobulation	Spiculation	VC	AB	PI	Average AUC
Separate binary classification	0.693*	0.752	0.806	0.807*	0.805	0.792	0.776
Multilabel classification	0.670	0.790*	0.824*	0.800	0.806*	0.804*	0.782*

*, the best performance in each column. MLC, multilabel classification; AUC, area under receiver operating characteristic curve; TLI, tumor-lung interface; VC, vessel convergence; AB, air bronchogram; PI, pleural indentation.

Table 7 Ablation study

Method	Accuracy	mAP	F1-micro	F1-macro	Hamming loss	Average AUC
Supervised	0.662	0.730	0.762	0.742	0.338	0.631
USC, without UA	0.702	0.800	0.792	0.774	0.298	0.707
USC, with UA	0.714*	0.806*	0.800*	0.784*	0.286*	0.712*

Supervised: only using the labeled data; USC, without UA: selection of pseudo-labels using a confidence-based strategy only; USC, with UA: selection of pseudo-labels using a confidence-and-uncertainty-based strategy. *, the best performance in each column. USC, uncertainty-aware self-training framework with consistency regularization; UA, uncertainty aware; mAP, mean average precision; AUC, area under receiver operating characteristic curve.

MLC model was better than that of the binary classification model on the 4 signs of lobulation, spiculation, AB, and PI. Although in terms of TLI and VC, the performance of MLC was inferior to that of a binary classification model, the overall average AUC of the MLC model reached 0.782, which was 0.6% higher than the average AUC of the 6 binary classification models trained separately.

Performance of the SSL method USC

An ablation study was conducted to demonstrate the superiority of our proposed SSL method. To evaluate the effect of our proposed USC method and the confidence-and-uncertainty-based pseudo-label selection strategy, we also trained a supervised model and a model only using the confidence to select pseudo-labels. *Table* 7 shows that for the initial MLC model trained with a small labeled dataset, the performance was inadequate, with an accuracy of 0.662, an mAP of 0.730, and an average AUC of 0.631. With the introduction of pseudo-labeled samples, all classification metrics significantly improved. Under the confidence-based pseudo-label selection strategy (USC without UA selection), the accuracy reached 0.702, the mAP reached 0.800, and the average AUC reached 0.707. When our confidence-and-uncertainty-based pseudo-label selection strategy (USC with UA selection) was used, the metrics mentioned above were further improved (accuracy: 1.2%; mAP: 0.6%; average AUC: 0.5%). Obviously, the confidence-and-uncertainty-based pseudo-label selection strategy in USC could further enhance the performance of the MLC model as compared to the strategy that used confidence only.

In the process of pseudo-label selection, as shown in *Figure* 7, if only confidence was used (USC without UA), then all unlabeled samples are selected at the beginning of self-training. When the confidence and uncertainty are considered at the same time (USC with UA), the quantity of selected unlabeled samples grows rapidly from a relatively small number, and nearly all unlabeled samples will be selected after 60 epochs. Correspondingly, the total number of selected pseudo-labels under this strategy is

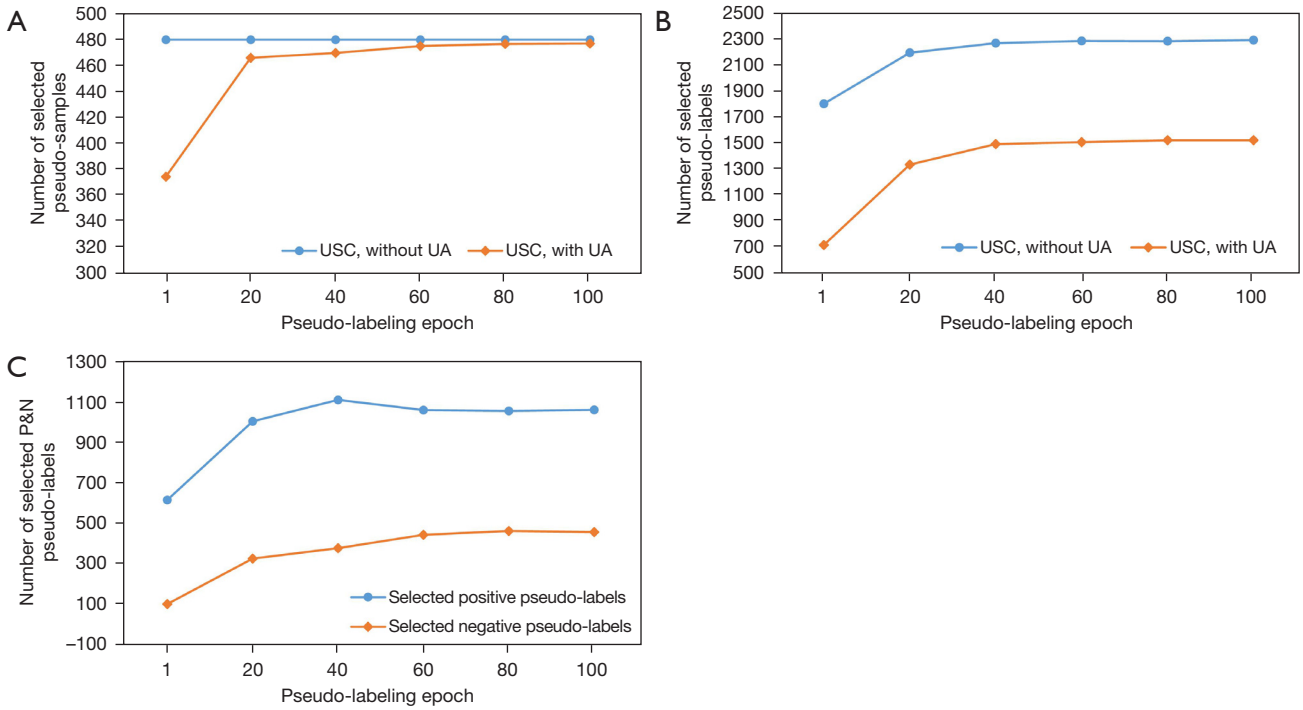

Figure 7 The total number of selected pseudo-samples and pseudo-labels. (A) Comparison of the total number of selected pseudo-samples between USC without UA and USC with UA. (B) Comparison of the total number of selected pseudo-labels between USC without UA and USC with UA. (C) Comparison of the number of selected positive and negative pseudo-labels in USC with UA. USC, uncertainty-aware self-training framework with consistency regularization; UA, uncertainty-aware.

only two-thirds that of the former strategy. Since different confidence and uncertainty thresholds are set for positive and negative pseudo-label selection, the number of positive and negative pseudo-labels used for model optimization is also quite different. Overall, the number of selected positive and negative pseudo-labels both tend to be stable after 60 epochs, and the number of positive pseudo-labels is about 3 times that of negative pseudo-labels.

In addition, as part of the unlabeled data were generated from the labeled data artificially, we also conducted a statistical analysis of the accuracy of the pseudo-labels for them. As depicted in *Figure 8A*, the USC with UA achieved significantly higher accuracy in selecting pseudo-labels than did the USC without UA, with an average accuracy reaching 78.3% after 60 epochs, which is 3.2% higher than that of the USC without UA. Furthermore, as shown in *Figure 8B*, the accuracy of positive pseudo-labels selected by the USC with UA method was much higher than that of the negative pseudo-labels, and the accuracy of the positive pseudo-labels gradually increased with the pseudo-labeling process, while the accuracy of the negative pseudo-labels decreased to some extent.

Comparison with other advanced SSL methods

We reimplemented the FixMatch and UPS method for comparison with our proposed USC. In the FixMatch experiment, the pseudo-labels were directly selected upon the confidence threshold for consistency training. In the UPS experiment, the pseudo-labeled samples were treated as labeled data to extend the dataset, and the model was then retrained using the extended dataset. *Table 8* shows the improvement in metrics that the above 2 methods and our USC method yielded for the MLC of nodule signs. We can see that on all metrics, our proposed USC achieved the most significant performance improvement.

Discussion

In this study, we realized the MLC of common CT imaging signs of lung nodules with our proposed 3D CNN, proposed a new SLL method for the MLC task of 3D

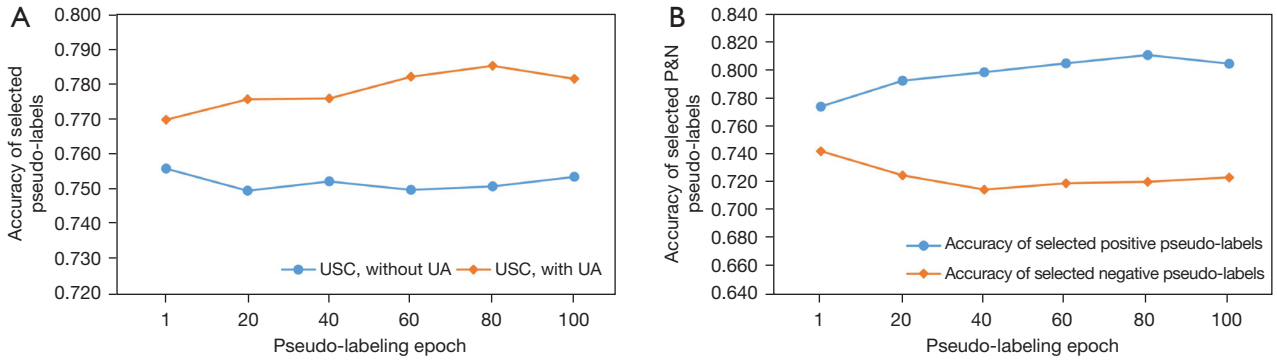

Figure 8 The accuracy of the selected pseudo-labels. (A) Comparison of the accuracy of selected pseudo-labels between USC without UA and USC with UA. (B) Comparison of the accuracy of selected positive and negative pseudo-labels in USC with UA. USC, uncertainty-aware self-training framework with consistency regularization; UA, uncertainty-aware.

Table 8 Performance improvement comparison between the USC and advanced SSL methods

Method	Accuracy	mAP	F1-micro	F1-macro	Hamming loss	Average AUC
FixMatch	4.0%	7.0%	3.0%	3.2%	4.0%	7.6%
UPS	3.7%	4.2%	2.9%	4.2%	3.7%	6.2%
USC	5.2%*	7.6%*	3.8%*	4.2%*	5.2%*	8.1%*

*, the best performance in each column. USC, uncertainty-aware self-training framework with consistency regularization; SSL, semi-supervised learning; UPS, uncertainty-aware pseudo-label selection framework (20); mAP, mean average precision; AUC, area under receiver operating characteristic curve.

medical images and applied it in a nodule sign classification task. Our proposed 3D CNN is constructed with dense block and CBAMs, achieving the best classification results in our experiments. Our proposed SSL method is called USC and compared with previously developed advanced SSL methods, exhibited better performance enhancement.

MLC is an important part of our work. Our proposed 3D CNN combines the efficient feature reuse of dense block with the spatial and channel attention provided by CBAMs. It can be seen from *Table 3* that the 3D CNN is better in terms of classification performance and parameter quantity. As shown in *Table 5*, the average AUC of the proposed model on the 6 nodule signs reached 0.782, and among 5 of these signs, reached 0.805. Moreover, the AUC of AB and PI, which are more clinically important and more indicative of malignant nodules, also exceeded 0.8, fully demonstrating our model's superiority and clinical value. However, for TLI, the AUC was far from satisfying. We believe that as it is difficult for radiologists to judge whether the TLI is clear, the model is also not able to learn salient features for this sign.

In addition, we used the same network structure to build a binary classification model for each sign for comparison. *Table 6* shows that although the binary classification models constructed separately showed better performance on some signs, their average AUC was not as good as that of the MLC model. We believe that it is because separate binary classification models fail to consider the implicit connections among features of different signs, while in the MLC model, the features can influence each other. Furthermore, considering that it is more costly to train a binary classification model for each sign, using an MLC model is more advantageous.

The SSL method is another focus of our work. In this paper, we propose a simple and efficient SSL method, USC, which combines the ideas of consistency training and confidence-and-uncertainty-based pseudo-labeling. We fully considered the distinctiveness of CT images compared with traditional images and MLC tasks compared with binary classification or multiclassification tasks in designing our semi-supervised method. For the former, we use a customized strong augmentation method adapted to the characteristics of CT images. The specific augmentation operations are shown in *Table 2*.

For the latter, we independently judge each class label of the sample during the pseudo-label selection process. Different selection criteria are set for positive pseudo-labels and negative pseudo-labels, respectively. Specifically, we use the initial classification model trained on labeled dataset to infer labels on the weakly augmented unlabeled images, select the predictions as pseudo-labels according to their confidence and uncertainty, and then calculate the consistency loss between the predictions of the corresponding strongly augmented images and the pseudo-labels. Compared with the traditional pseudo-label selection strategy, the biggest advantage of our proposed method is that we use uncertainty and confidence simultaneously as the criteria for pseudo-label selection, which helps us to select more stable and reliable pseudo-labels. As *Figure* 7 and *Figure 8* show, although fewer pseudo-labels are used for model optimization, these selected pseudo-labels are more stable and reliable, providing more useful information to the model and reducing the introduction of noise. As *Table* 7 shows, this practice further enhanced the accuracy, mAP, and average AUC by 1.0%, 0.6%, and 0.5% respectively, as compared with the traditional pseudo-label selection strategy, and in contrast to the supervised method, the overall enhancements in accuracy, mAP, and average AUC were 5.2%, 7.6%, and 8.1% respectively, exceeding those of other advanced methods (*Table 8*). In our experiments, probably due to the limited size of the dataset, the performance improvement derived from the confidence-and-uncertainty-based pseudo-label selection strategy was not significant, but it did improve the quality of pseudo-label selection. We believe that our method will demonstrate additionally significant advantages on larger datasets. In summary, our proposed semi-supervised method has great potential for application to other MLC tasks of 3D medical images.

Our study has some notable limitations which should be addressed. First, the CT signs related to lung nodules are not limited to the 6 types discussed in this paper. Therefore, we will collect additional CT images containing other types of signs to expand our model's classification capability. Second, the data we used in this study were gathered from 3 hospitals, but the domain differences caused by the introduction of multicenter data were not considered in our experiments. In the actual clinical application scenario, the labeled data and unlabeled data that we can use for model training are likely to originate from different centers or scanners with different parameters, which will inevitably lead to the problem of domain shift. In future work, we will continue to improve our MLC model and SSL method and explore how to integrate domain adaptation techniques into our proposed SSL method to further enhance its generalizability.

Conclusions

In this paper, we realized the MLC task of lung nodule signs with our proposed 3D CNN, which achieved satisfactory performance. We also proposed a simple and efficient SSL method: USC, which combines consistency regularization and confidence-and-uncertainty-based pseudo-label selection strategy, and applied it in our experiments. The experimental results show that the USC enables a more efficient selection of pseudo-labels and can also significantly improve model performance, achieving better improvement than other advanced SSL methods. We believe that our proposed USC method has considerable potential for other 3D medical image MLC tasks.

Acknowledgments

Funding: This work was supported by the National Key Research and Development Program of China (No. 2021YFC2500403), the S&T Program of Hebei (No. 21377734D), the National Natural Science Foundation of China (General Program; No. 82172030), the Clinical Research Plan of SHDC, China (No. SHDC2020CR3080B), and Peak Disciplines (Type IV) of Institutions of Higher Learning in Shanghai.

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-23-40/rc

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-23-40/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013) and was approved by the ethics committee

of the Shanghai Public Health Clinical Center. As this study was retrospective in nature, the collection of patient CT images and relevant clinical information would not adversely affect the patient's rights or welfare, and thus the need for individual consent was waived by the committee.

References

1. Chhikara BS, Parang K. Global Cancer Statistics 2022: the trends projection analysis. Chemical Biology Letters 2023;10:451.
2. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. CA Cancer J Clin 2022;72:7-33.
3. Chiarenza A, Esposto Ultimo L, Falsaperla D, Travali M, Foti PV, Torrisi SE, Schi-sano M, Mauro LA, Sambataro G, Basile A, Vancheri C, Palmucci S. Chest imaging using signs, symbols, and naturalistic images: a practical guide for radiologists and non-radiologists. Insights Imaging 2019;10:114.
4. Liu J, Yang X, Li Y, Xu H, He C, Qing H, Ren J, Zhou P. Development and valida-tion of qualitative and quantitative models to predict invasiveness of lung adenocar-cinomas manifesting as pure ground-glass nodules based on low-dose computed to-mography during lung cancer screening. Quant Imaging Med Surg 2022;12:2917-31.
5. Zhu W, Liu C, Fan W, Xie X. DeepLung: Deep 3D dual path nets for automated pulmonary nodule detection and classification. Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE; 2018:673-81.
6. Al-Shabi M, Shak K, Tan M. ProCAN: Progressive growing channel attentive non-local network for lung nodule classification. Pattern Recognition 2022;122:108309.
7. Wu B, Zhou Z, Wang J, Wang Y, editors. Joint learning for pulmonary nodule seg-mentation, attributes and malignancy prediction. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018); 04-07 April 2018; Washington, DC, USA. IEEE; 2018.
8. Li J, Zhou L, Zhan Y, Xu H, Zhang C, Shan F, Liu L. How does the artificial intelli-gence-based image-assisted technique help physicians in diagnosis of pulmonary adenocarcinoma? A randomized controlled experiment of multicenter physicians in China. J Am Med Inform Assoc 2022;29:2041-9.
9. Han G, Liu X, Zheng G, Wang M, Huang S. Automatic recognition of 3D GGO CT imaging signs through the fusion of hybrid resampling and layer-wise fine-tuning CNNs. Med Biol Eng Comput 2018;56:2201-12.
10. Han G, Liu X, Zhang H, Zheng G, Soomro NQ, Wang M, Liu W. Hybrid resampling and multi-feature fusion for automatic recognition of cavity imaging sign in lung CT. Future Generation Computer Systems 2019;99:558-70.
11. Qiu S, Sun J, Zhou T, Gao G, He Z, Liang T. Spiculation Sign Recognition in a Pul-monary Nodule Based on Spiking Neural P Systems. Biomed Res Int 2020;2020:6619076.
12. Ma L, Liu X, Song L, Zhou C, Zhao X, Zhao Y. A new classifier fusion method based on historical and on-line classification reliability for recognizing common CT imaging signs of lung diseases. Comput Med Imaging Graph 2015;40:39-48.
13. Ma L, Liu X, Gao Y, Zhao Y, Zhao X, Zhou C. A new method of content based med-ical image retrieval and its applications to CT imaging sign retrieval. J Biomed Inform 2017;66:148-58.
14. Ma L, Liu X, Fei B. A multi-level similarity measure for the retrieval of the common CT imaging signs of lung diseases. Med Biol Eng Comput 2020;58:1015-29.
15. He G. Lung CT imaging sign classification through deep learning on small data. arXiv:1903.00183 [Preprint]. 2019. Available online: https://doi.org/10.48550/arXiv.1903.00183
16. Laine S, Aila T. Temporal ensembling for semi-supervised learning. arXiv:1610.02242 [Preprint]. 2016. Available online: https://doi.org/10.48550/arXiv.1610.02242
17. Tarvainen A, Valpola H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017); Long Beach, CA, USA. 2017:1195-204.
18. Xie Q, Luong MT, Hovy E, Le QV. Self-training with noisy student improves imagenet classification. 2020 IEEE/CVF Conference on Computer Vision and Pat-tern Recognition (CVPR); 13-19 June 2020; Seattle, WA, USA.

IEEE; 2020.
19. Li Z, Ko B, Choi HJ. Naive semi-supervised deep learning using pseudo-label. Peer Peer Netw Appl 2019;12:1358-68.
20. Rizve MN, Duarte K, Rawat YS, Shah M. In defense of pseudo-labeling: An uncer-tainty-aware pseudo-label selection framework for semi-supervised learning. arXiv:2101.06329 [Preprint]. 2021. Available online: https://doi.org/10.48550/arXiv.2101.06329
21. Oliver A, Odena A, Raffel CA, Cubuk ED, Goodfellow I. Realistic evaluation of deep semi-supervised learning algorithms. Advances in Neural Information Processing Systems 31 (NeurIPS 2018); Montréal, Canada. 2018:3239-50.
22. Grandvalet Y, Bengio Y. Semi-supervised learning by entropy minimization. Pro-ceedings of the 17th International Conference on Neural Information Processing Systems (NIPS'04); Cambridge, MA, USA. 2004:529-36.
23. Yang X, Song Z, King I, Xu Z. A survey on deep semi-supervised learning. IEEE Transactions on Knowledge and Data Engineering. IEEE; 2022. doi: 10.1109/TKDE.2022.3220219.
24. Berthelot D, Carlini N, Goodfellow I, Papernot N, Oliver A, Raffel CA. Mixmatch: A holistic approach to semi-supervised learning. 33rd Conference on Neural Information Processing Systems (NeurIPS 2019); Vancouver, Canada. 2019:5049-59.
25. Xie Q, Dai Z, Hovy E, Luong T, Le Q. Unsupervised data augmentation for con-sistency training. 34th Conference on Neural Information Processing Systems (Neu-rIPS 2020); Vancouver, Canada. 2020:6256-68.
26. Berthelot D, Carlini N, Cubuk ED, Kurakin A, Sohn K, Zhang H, Raffel C. Re-MixMatch: Semi-Supervised Learning with Distribution Alignment and Augmenta-tion Anchoring. arXiv:1911.09785 [Preprint]. 2019. Available online: https://doi.org/10.48550/arXiv.1911.09785
27. Sohn K, Berthelot D, Carlini N, Zhang Z, Zhang H, Raffel CA, Cubuk ED, Kurakin A, Li C-L. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. 34th Conference on Neural Information Processing Systems (NeurIPS 2020); Vancouver, Canada. 2020:596-608.
28. Wang X, Chen H, Xiang H, Lin H, Lin X, Heng PA. Deep virtual adversarial self-training with consistency regularization for semi-supervised medical image classification. Med Image Anal 2021;70:102010.
29. Shi F, Chen B, Cao Q, Wei Y, Zhou Q, Zhang R, Zhou Y, Yang W, Wang X, Fan R, Yang F, Chen Y, Li W, Gao Y, Shen D. Semi-Supervised Deep Transfer Learning for Benign-Malignant Diagnosis of Pulmonary Nodules in Chest CT Images. IEEE Trans Med Imaging 2022;41:771-81.
30. Liu Q, Yu L, Luo L, Dou Q, Heng PA. Semi-Supervised Medical Image Classifica-tion With Relation-Driven Self-Ensembling Model. IEEE Trans Med Imaging 2020;39:3429-40.
31. Liu F, Tian Y, Chen Y, Liu Y, Belagiannis V, Carneiro G. ACPL: Anti-curriculum Pseudo-labelling for Semi-supervised Medical Image Classification. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 18-24 June 2022; New Orleans, LA, USA. IEEE; 2022.
32. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. mixup: Beyond empirical risk mini-mization. arXiv:1710.09412 [Preprint]. 2017. Available online: https://doi.org/10.48550/arXiv.1710.09412
33. Wang Q, Jia N, Breckon T. A baseline for multi-label image classification using an ensemble of deep convolutional neural networks. 2019 IEEE International Confer-ence on Image Processing (ICIP); 22-25 September 2019; Taipei, Taiwan. IEEE; 2019.
34. Armato SG 3rd, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys 2011;38:915-31.
35. Albert RH, Russell JJ. Evaluation of the solitary pulmonary nodule. Am Fam Physi-cian 2009;80:827-31.
36. Setio AAA, Traverso A, de Bel T, Berens MSN, Bogaard CVD, Cerello P, et al. Val-idation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. Med Image Anal 2017;42:1-13.
37. Li K, Liu K, Zhong Y, Liang M, Qin P, Li H, Zhang R, Li S, Liu X. Assessing the predictive accuracy of lung cancer, metastases, and benign lesions using an artificial intelligence-driven computer aided diagnosis system. Quant Imaging Med Surg 2021;11:3629-42.
38. Al-Shabi M, Lan BL, Chan WY, Ng KH, Tan M. Lung nodule classification using deep Local-Global networks. Int J Comput Assist Radiol Surg 2019;14:1815-9.
39. Abid MMN, Zia T, Ghafoor M, Windridge D. Multi-view convolutional recurrent neural networks for lung cancer nodule identification. Neurocomputing 2021;453:299-311.
40. Onishi Y, Teramoto A, Tsujimoto M, Tsukamoto T, Saito

K, Toyama H, Imaizumi K, Fujita H. Multiplanar analysis for pulmonary nodule classification in CT images using deep convolutional neural network and generative adversarial networks. Int J Comput Assist Radiol Surg 2020;15:173-8.
41. Mehta K, Jain A, Mangalagiri J, Menon S, Nguyen P, Chapman DR. Lung Nodule Classification Using Biomarkers, Volumetric Radiomics, and 3D CNNs. J Digit Im-aging 2021;34:647-66.
42. Jiang H, Shen F, Gao F, Han W. Learning Efficient, Explainable and Discriminative Representations for Pulmonary Nodules Classification. Pattern Recognition 2021;113:107825.
43. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolu-tional networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 21-26 July 2017; Honolulu, HI, USA. IEEE; 2017.
44. Woo S, Park J, Lee JY, Kweon IS, CBAM: Convolutional Block Attention Module. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, editors. Computer Vision – ECCV 2018. Lecture Notes in Computer Science. vol 11211. Springer, Cham; 2018:3-19.
45. Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, Santamaría J, Fadhel MA, Al-Amidie M, Farhan L. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 2021;8:53.
46. Alzubaidi L, Santamaría J, Manoufali M, Mohammed B, Fadhel MA, Zhang J, Al-Timemy AH, Al-Shamma O, Duan Y. MedNet: pre-trained convolutional neural network model for the medical imaging tasks. arXiv:2110.06512 [Preprint]. 2021. Available online: https://doi.org/10.48550/arXiv.2110.06512
47. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 27-30 June 2016; Las Vegas, NV, USA. IEEE; 2016.
48. Bachman P, Alsharif O, Precup D. Learning with pseudo-ensembles. Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS'14). MIT Press; 2014:3365-73.
49. Gal Y, Ghahramani Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML'16). 2016:1050-9.
50. Solovyev R, Kalinin AA, Gabruseva T. 3D convolutional neural networks for stalled brain capillary detection. Comput Biol Med 2022;141:105089.

Cite this article as: Zhan K, Wang Y, Zhuo Y, Zhan Y, Yan Q, Shan F, Zhou L, Chen X, Liu L. An uncertain-ty-aware self-training framework with consistency regularization for the multilabel classification of common computed tomography signs in lung nodules. Quant Imaging Med Surg 2023;13(9):5536-5554. doi: 10.21037/qims-23-40

