Medical Image Analysis 83 (2023) 102627

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Trustworthy learning with (un)sure annotation for lung nodule diagnosis with CT

Hanxiao Zhang [a,1], Liang Chen [b,1], Xiao Gu [c], Minghui Zhang [a], Yulei Qin [d], Feng Yao [b], Zhexin Wang [b,*], Yun Gu [a,e,*], Guang-Zhong Yang [a,*]

[a] Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China
[b] Department of Thoracic Surgery, Shanghai Chest Hospital, Shanghai Jiao Tong University, Shanghai, China
[c] Imperial College London, London, UK
[d] Youtu Lab, Tencent, Shanghai, China
[e] Shanghai Center for Brain Science and Brain-Inspired Technology, Shanghai, China

ARTICLE INFO

Keywords:
Lung nodule
Computer-aided diagnosis
Deep learning
Explainable artificial intelligence

ABSTRACT

Recent evolution in deep learning has proven its value for CT-based lung nodule classification. Most current techniques are intrinsically black-box systems, suffering from two generalizability issues in clinical practice. First, benign-malignant discrimination is often assessed by human observers without pathologic diagnoses at the nodule level. We termed these data as "unsure-annotation data". Second, a classifier does not necessarily acquire reliable nodule features for stable learning and robust prediction with patch-level labels during learning. In this study, we construct a sure-annotation dataset with pathologically-confirmed labels and propose a collaborative learning framework to facilitate sure nodule classification by integrating unsure-annotation data knowledge through nodule segmentation and malignancy score regression. A loss function is designed to learn reliable features by introducing interpretability constraints regulated with nodule segmentation maps. Furthermore, based on model inference results that reflect the understanding from both machine and experts, we explore a new nodule analysis method for similar historical nodule retrieval and interpretable diagnosis. Detailed experimental results demonstrate that our approach is beneficial for achieving improved performance coupled with trustworthy model reasoning for lung cancer prediction with limited data. Extensive cross-evaluation results further illustrate the effect of unsure-annotation data for deep-learning based methods in lung nodule classification.

1. Introduction

Lung cancer is one of the major causes of cancer-related death worldwide in the last 10 years (Siegel et al., 2021; Sung et al., 2021). Screening for lung cancer with low-dose helical computed tomography (CT) has been shown in the National Lung Screening Trial (NLST) to reduce mortality from lung cancer by 20% in high-risk individuals relative to screening with chest radiography (National Lung Screening Trial Research Team, 2011b). Driven by CT data, deep learning is advantageous for fast computer-aided diagnosis (CAD) of lung cancer which involves lung nodule detection and benign-malignant classification. In this paper, rather than nodule detection, we mainly focus on nodule malignancy classification, which presents a challenge due to the diversified shapes, textures and contextual environments of lung nodules (Qin et al., 2021; McWilliams et al., 2013).

In recent years, many works have directed toward nodule classification based on Convolutional Neural Networks (CNNs). To improve the performance of nodule heterogeneity discrimination of CNN models, a diversity of approaches have been proposed, such as model ensemble with multi-level inputs (Shen et al., 2015, 2017; Xu et al., 2020; Xie et al., 2017, 2018a), multi-task learning (MTL) with auxiliary tasks (e.g., nodule segmentation (Wu et al., 2018; Yang et al., 2019), reconstruction (Xie et al., 2019), attribute regression (Liu et al., 2019)), and relational learning from multiple nodules within a patient (Liao et al., 2019; Yang et al., 2020; Liu et al., 2021) (see details in Section 2.1). However, these studies suffer from labeling accuracy and trustworthiness in model reasoning.

* Corresponding authors.
E-mail addresses: wzx1953@shchest.org (Z. Wang), geron762@sjtu.edu.cn (Y. Gu), gzyang@sjtu.edu.cn (G.-Z. Yang).
[1] Contributed equally to preparation of the manuscript.

https://doi.org/10.1016/j.media.2022.102627
Received 14 December 2021; Received in revised form 22 July 2022; Accepted 10 September 2022
Available online 27 September 2022

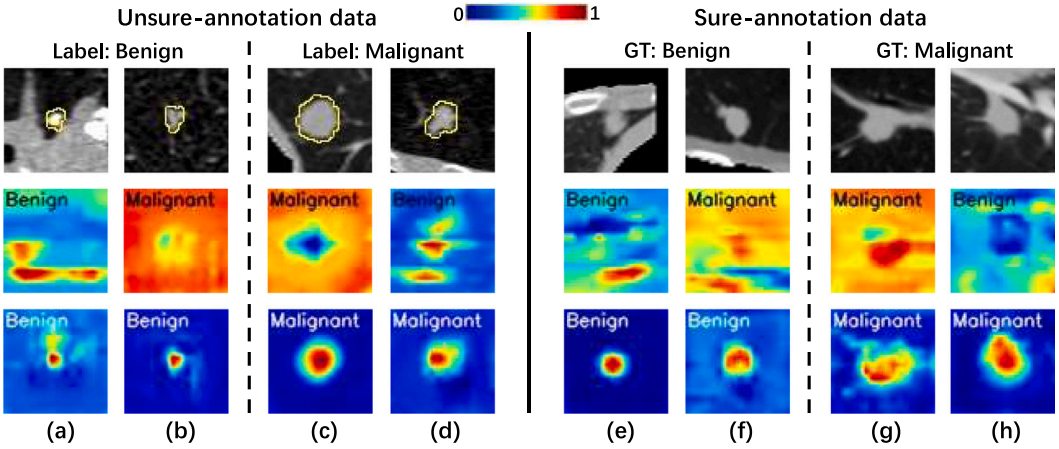

Fig. 1. Some examples of model interpretation. The first row shows different nodule inputs from unsure-annotation data and sure-annotation data, where the yellow contours on unsure-annotation data inputs are nodule segmentation by radiologists. The second row uses the CAM (Colormap Jet) to interpret which parts of the input contribute to the final prediction during ResNet reasoning. The third row shows the CAMs of our proposed model. Both the input images and CAMs are taken from the middle cross-sections of their 3D patches. The "Label" of unsure-annotation data is manually assigned by a malignancy score threshold 3 here. "GT" represents the ground truth of sure-annotation data that is confirmed by pathologic examination. For cases (a) and (e), although ResNet predicts "benign" on benign nodules correctly, this correct prediction comes from the misinterpreted evidence apart from the nodule regions. For cases (b) and (f), ResNet misclassifies benign nodules to "malignant", where the CAM highlights both nodule and background regions. Our model can activate reliable features that are faithful to the nodule regions.

First, most of the existing methods focuses on improving nodule malignancy classification accuracy within typical publicly available dataset such as LIDC-IDRI (Lung Image Database Consortium and Image Database Resource Initiative) (Armato III et al., 2011). During the annotation of LIDC-IDRI, the characteristics of nodules were assessed by multiple radiologists, for which the score of "likelihood of malignancy" was subjectively rated on a five-point scale under the assumption that the CT scan was originated from a 60-year-old male smoker (Armato III et al., 2011; McNitt-Gray et al., 2007), introducing inherent ambiguity problems (e.g., inter-rater variability, nodule malignancy uncertainty). Although many attempts have been made to overcome these problems (Section 2.2), models learned from LIDC-IDRI dataset could only mimic the radiologists' reasoning by statistical fitting rather than conduct benign-malignant nodule classification in real clinical practice, since LIDC-IDRI dataset lacks definite pathologically-proven ground truth. Other work in Section 2.2 applied public data from TIANCHI challenges[2], NLST trial[3] (National Lung Screening Trial Research Team, 2011a,b), and Kaggle's 2017 Data Science Bowl (DSB) competition (NLST subset)[4] to predict lung cancer by semi-supervised learning (Xie et al., 2019) or patient-level multiple instance learning (MIL) (Ardila et al., 2019; Liao et al., 2019; Ozdemir et al., 2019; Liu et al., 2021). Unfortunately, end-to-end training for nodule-level prediction was hindered by these datasets due to lack of complete annotations such as position coordinates and pathologic diagnosis of each nodule.

In this paper, we term these datasets as "unsure-annotation data" by its nature of uncertainty in classification label and specific target, which has been largely unaddressed by the medical image analysis community. The "unsure data" was first raised by Wu et al. (2019) referring to some cases of image-based disease prediction that lack deterministic "disease/normal" labels, such as LIDC-IDRI dataset. We further expand the scope of "unsure-annotation data" for data with uncertain lesion locations (e.g., NLST trial data (National Lung Screening Trial Research Team, 2011a,b)). This definition may have some overlaps with "noisy-label" and "weak-label", but "unsure annotation" puts more emphasis on the unconfident labels of lesion types or unspecific location of lesion detections in medical images. To circumvent these issues, we constructed a new, high-quality annotated dataset with definite pathologically-confirmed labels and specific locations for nodules, as described in Section 4.1.

Second, in addition to the problem of unsure-annotation data, further validation and explanation are desirable to ensure the systems is trustworthy (Jacobs and van Ginneken, 2019). The purpose of an explainable AI (XAI) system is to make its behavior more interpretable and credible to humans by providing explanations and evidence (Gunning et al., 2019). Class Activation Mapping (CAM) (Zhou et al., 2016) could help to retrospectively interpret the CNN's reasoning process by performing a weighted sum of the final feature maps, which enables a model to disclose salient information and lend insights into failure cases.

To illustrate failure cases visually, we use CAM in Fig. 1 to analyze some validation nodules misclassified or misinterpreted by a ResNet (He et al., 2016) trained with unsure LIDC-IDRI dataset or our sure-annotation dataset using binary cross-entropy loss. For instance, in the misinterpreted cases (a) and (e), the class "benign" has a bias toward background regions with high correlation as a distractor for malignancy prediction, indicating that the model may not acquire the incentive to learn benign evidence on nodule regions using training data that is only annotated with patch-level classes. However, generalization performance may degrade if the testing data does not have the same correlation (Zhang et al., 2021).

Evidently yet easily overlooked, failure cases in Fig. 1 deviate from the requirement of the nodule malignancy classification task that the reliance of the model on reliable and faithful features must be guaranteed (Samek et al., 2019; Gu et al., 2020), especially when the training data is small or in the cases of out-of-distribution (OoD).

Motivated by the above challenges, this paper proposes a trustworthy learning framework for CT-based lung cancer diagnosis based on the incorporation of both sure and unsure-annotation data.

First of all, we design a three-branch synergic model that not only learns to classify the sure-annotation data nodules in the primary task, but also learns to diagnose like a radiologist in two auxiliary tasks for conducting nodule segmentation and malignancy score regression using unsure-annotation data. This integration approach can alleviate the negative impact of unsure-annotation data risk while elaborately adapting the unsure-annotation data knowledge for sure-annotation data learning.

Then, to endow the model with the ability to learn reliable features that are focused on the nodule regions, we leverage the CAM not only an afterthought but also a first-class citizen during training.

[2] https://tianchi.aliyun.com/competition/entrance/231601/introduction/
[3] https://cdas.cancer.gov/datasets/nlst/
[4] https://www.kaggle.com/c/data-science-bowl-2017/

To be specific, we propose a novel loss function for model online regularization called adaptive CAM-SEM-Loss (ad-CSL) by introducing "interpretability constraints" (Gunning et al., 2019), which drives the model to express the malignancy features from the nodule regions and suppress the features in background regions under the supervision of nodule segmentation maps (SEM).

In addition, based on the intra-class variance of different nodules' benign-malignant predictions (machine reasoning) and malignancy score regression outputs (mimic expert reasoning), we explore a new nodule diagnosis strategy that automatically retrieves the most similar nodules identified in a historical database, relative to a testing nodule. This can provide more clues and evidence for radiologists and clinicians by referring to the prior knowledge of similar historical nodule cases.

Our main contributions of this work can be summarized as follows:

1. A new issue for jointly learning with unsure-annotation data and our newly constructed sure-annotation data is highlighted and our work represents the first attempt for addressing this issue systematically.
2. A synergic model is proposed to integrate the unsure-annotation data knowledge with two auxiliary tasks and ultimately promote the performance of sure-annotation data classification.
3. A novel regularization scheme is proposed, which feeds back the online generated CAM to modify the classification process in such a way that the model could be more robust by learning the faithful nodule features.
4. An effective nodule diagnosis strategy is developed, which is practical for clinical usage and extensive experiments are performed to investigate the effect of unsure-annotation data and the associated problems.

2. Related work

2.1. CNN-based lung cancer prediction

Lung cancer prediction typically refers to the classification of benign-malignant nodules, which is indispensable to the powerful Convolutional Neural Networks (CNNs) in recent years. Many attempts have been done to improve the classification performance of CNN-based methods.

First, ensemble model learning is commonly used to extract the multi-level input features. Shen et al. (2015) proposed a weight-shared network (MCNN) to learn discriminative features from 3D nodule patches with different scales, which was then simplified by applying a multi-crop pooling strategy in MC-CNN (Shen et al., 2017). Using multi-scale input, MSCS-DeepLN (Xu et al., 2020) combined three independent sub-networks to generate the final ensemble prediction. For better model nodule heterogeneity, Xie et al. (2017) developed a transferable ensemble model using three input patches characterizing overall appearance, nodule shapes, and voxel values, respectively. As suggested by Setio et al. (2016), MV-KBC (Xie et al., 2018a) extended the former work (Xie et al., 2017) by decomposing a 3D nodule volume onto nine fixed view planes and fed 2D patches into 27 sub-networks.

Second, multi-task learning (MTL) helps exploit shareable knowledge in nodule-involved tasks. Based on the LIDC-IDRI dataset, (Hussein et al., 2017) implicitly explored the potential of nodule attributes to improve the malignancy prediction by using MTL. Moreover, Chen et al. (2016) modeled the internal relationship between the nodule attributes and malignancy. Furthermore, Liu et al. (2019) designed an MTL model that renders a mutual influence between the nodule classification and attribute score regression tasks. Meanwhile, Wu et al. (2018) and Yang et al. (2019) conducted joint learning for nodule segmentation and malignancy prediction within an U-Net (Ronneberger et al., 2015) structured model.

Third, relational learning may explore the incremental value of nodule data. In the Kaggle's 2017 Data Science Bowl (DSB) competition, the first-place model (Liao et al., 2019) formulated the cancer prediction as a multiple instance learning (MIL) problem that evaluated the cancer probability of a patient with multiple detected nodules. Yang et al. (2020) learned the relations between multiple solitary nodules from a single patient within the LIDC-IDRI dataset. Considering that some contextual features could be malignancy-related in the sense of statistics, Liu et al. (2021) fused the features from the nodule and its surrounding structures to learn the malignancy patterns via an attention mechanism, which was also evaluated on the 2017 DSB dataset using MIL.

2.2. Learning from unsure-annotation nodule data

In lung cancer prediction, most of the current work are based on one static dataset LIDC-IDRI (Armato III et al., 2011). Learning from this unsure-annotation dataset has several obstacles.

First, assessed by multiple radiologists, the malignancy rating scores for each nodule may encounter a stochastic inter-observer variability. (Carrazza et al., 2016) has discovered the latent negative effect of aggregating raters' disagreements. Liao et al. (2021) proposed a 'divide-and-rule' model (MV-DAR) to learn from ambiguous labels by alleviating the value of inconsistent and unreliable nodule annotation. Nevertheless, label consensus matters in the classification task.

Second, current work commonly formulated the LIDC-IDRI nodule malignancy prediction as a binary classification task. The binary labels are crudely assigned using a simple method (Han et al., 2013) that based on the average malignancy score by enforcing a hard predefined threshold of score. A large number of nodules with an average score 3 were often discarded as uncertain class. To make use of these uncertain nodules, Wu et al. (2019), Lei et al. (2020b, 2021) applied ordinal regression to learn the relationship among the three classes. Furthermore, Liu et al. (2019) employed a Siamese network with a margin ranking loss to model the malignancy score difference.

Third, due to the lack of nodule-level pathologic diagnoses, models can only learn from experts' knowledge which is subjective and maybe inaccurate relative to ground truth labels. LIDC-IDRI contains a small set of cases (157 patients) with diagnosis data at the patient-level (McNitt-Gray et al., 2007) where four ratings (0: unknown, 1: non-malignant disease, 2: primary lung cancer, 3: metastatic lesion) were recorded along with five diagnosis methods. Based on this dataset, Shen et al. (2016) developed a MIL framework for patient-level lung cancer prediction, indicating the potential of using definite pathologically-proven CT data for lung cancer diagnosis.

In addition to LIDC-IDRI, datasets from the ANODE09 (Van Ginneken et al., 2010), LUNA16 (Setio et al., 2017) (a subset of LIDC-IDRI) and TIANCHI challenges are less directly used for lung cancer prediction but for nodule detection as they only provided the nodule locations and diameters. However, researchers can leverage these malignancy-unlabeled data to support the supervised classification model by semi-supervised learning (Xie et al., 2019). Besides, the dataset of the National Lung Cancer Screening Trial (NLST) (National Lung Screening Trial Research Team, 2011a,b) preserved numerous CT scans in a randomized controlled trial of screening tests for lung cancer. The ground truth for cancer on the NLST dataset was biopsy or surgically confirmed, while the cancer-negative cases were only defined by a minimum of one-year follow-up. The nodule locations were not collected by the NLST trial, unfortunately. Thus, lung cancer identification using the NLST dataset (Ardila et al., 2019) or Kaggle's 2017 DSB dataset (Liao et al., 2019; Ozdemir et al., 2019; Liu et al., 2021) (NLST subset) rely on accurate nodule detection and segmentation in advance, which is more appropriate to be treated as a long-term patient-level prediction task.

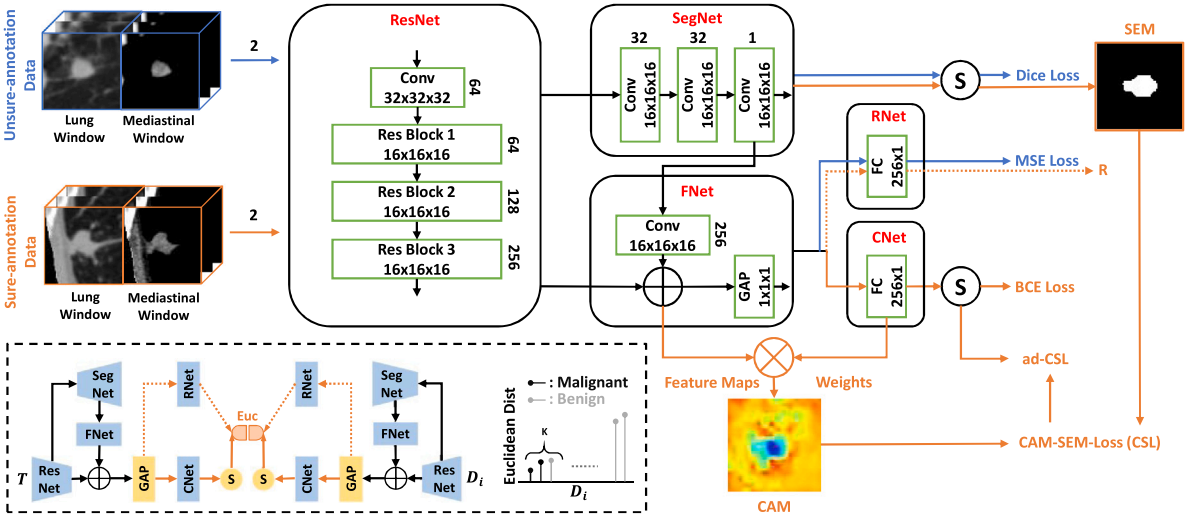

Fig. 2. Overview of our proposed framework for lung nodule diagnosis. (a) The synergic model includes five modules for joint learning from unsure and sure-annotation data. Except for the last convolution layer of ResNet, SegNet and FNet that is followed by a group normalization, other convolution layers are all followed by a group normalization and a ReLU activation. ReLU activation is also performed before SegNet and after the feature addition in FNet. The resolution and channel number of each convolution or block output are denoted. (b) Adaptive CAM-SEM-Loss (ad-CSL) leverages the class activation map (CAM), nodule segmentation map (SEM) and malignant probability to regulate the synergic model for stable learning from reliable features of sure-annotation data. (c) The new nodule diagnosis method is illustrated at the bottom left corner. Two weight-shared pre-trained synergic models are employed to calculate the Euclidean Distance between the outputs of a nodule pair include the testing nodule T and an identified nodule D_i in the historical database. The top K historical nodules with the closest Euclidean Distance relative to the testing one are selected for new malignancy assessment and diagnosis interpretation.

2.3. Interpretable nodule diagnosis

AI systems for medical imaging diagnosis are usually applied in a black-box manner, which is highly desirable to gain the trust of users (Gunning et al., 2019; Samek et al., 2019). CAM (Zhou et al., 2016) is often used to interpret what a model has learned. The CAM visualization for nodule classifiers is not guaranteed to activate nodule regions, implying a less reliable feature learning. These classifiers are prone to overfitting as they have the strong ability of data fitting to obtain the mapping from the patch-level input to its label, especially with limited supervision in a small dataset. The performance of OoD generalization may deteriorate when the model captures the spurious and irrelevant features within the training data domain whereas the testing data does not have the same correlation. Many regularization solutions have been designed to solve this issue, such as Dropout (Srivastava et al., 2014), Stochastic Depth (Huang et al., 2016), Label smoothing (Szegedy et al., 2016), Cutout (DeVries and Taylor, 2017), Mixup (Zhang et al., 2017) and Cutmix (Yun et al., 2019) . An alternative solution is to design different attention modules to obtain more distinguishing feature representations (Zhang et al., 2019; Gu et al., 2020). However, these methods rely on relatively large datasets and may still locate the most discriminative region of the model's interest in the training domain. Wang et al. (2021) instead proposed a CAM-loss to constrain the embedded feature maps by minimizing the difference between class activation map and class-agnostic activation map in natural images. Inspired by this attempt, we further model the CAM map explicitly as part of the training by guiding the CAM attention to the region of human interest.

3. Methods

As shown in Fig. 2, the principal part of our proposed framework is a synergic model (Section 3.1) that consists of five modules for joint learning with sure and unsure-annotation data in three tasks. In addition to the common supervisions for these three tasks, we introduce a novel regularization loss (Section 3.2) for faithful nodule feature learning. Moreover, a new nodule diagnosis method is described in Section 3.3. In Fig. 2, the blue lines indicate operations independently performed for the stream of unsure-annotation data. The orange lines represent the data stream of sure-annotation data, where the dashed lines denote the data stream only for sure-annotation data inference in our new nodule diagnosis process. The black lines are joint canals for both data streams.

3.1. Synergic model

To better integrate the knowledge from LIDC-IDRI (unsure-annotation data) and sure-annotation data while diminishing the negative effect of domain shift in data and label space, we design a synergic model consisting of five modules (Fig. 2): a backbone for feature extraction (ResNet), a branch for segmentation task (SegNet), a link bridge for feature fusion (FNet), and another two symmetric branches for nodule malignancy score regression task (RNet) and benign-malignant classification task (CNet). The primary sure-annotation data task of classification and the auxiliary unsure-annotation data tasks of segmentation and regression are online trained simultaneously in an end-to-end manner. The input of the synergic model is a couple of 3D patches, which are randomly selected from unsure and sure-annotation data, respectively. The five modules of the synergic model are delved as follows:

ResNet module: The ResNet module for nodule feature extraction is composed of a convolutional head (size $7 \times 7 \times 7$, stride 2, padding 3) and three stacked Residual blocks (two convolutional layers in each block) with the projection shortcut done by $1 \times 1 \times 1$ convolution to match dimensions as used in He et al. (2016) . As noticed in Zhou et al. (2016), the final convolutional layer before global average pooling (GAP) (Lin et al., 2013) should have a higher spatial resolution for better CAM expression. To this end, we make the following modifications as compared to the original ResNet model: (1) we remove the max-pooling layer before the first Residual block; (2) in Residual blocks 2 and 3, we set stride 1 for each convolutional operation to avoid feature down-sampling, and use dilated convolution (Yu and Koltun, 2015) with spacing 2 in the second Residual block and space 4 in the third Residual block to obtain the large receptive field; (3) group normalization (Wu and He, 2018) is used after each convolutional operation due to small batch size setting in the whole synergic model. Thus, the resolution of ResNet output is $16 \times 16 \times 16$, which provides a

reasonable representation in terms of semantic information and spatial contexture.

SegNet module: Our SegNet module takes inspirations from U-Net (Ronneberger et al., 2015), a common structure utilized in most biomedical image segmentation works. Differently, instead of performing segmentation as the main purpose, the auxiliary SegNet module should ultimately serve the primary task of sure-annotation data nodule classification. Therefore, we give the top priority to the high demand for fast and simple module integration. Without loss of generality, we adopt the encoder–decoder structure as the nodule segmentation implementation, where the ResNet module is regarded as a ready-made encoder to extract high-dimensional features, and the SegNet module is a lightweight decoder that utilizes the encoded features to recover the nodule segmentation distribution. This lightweight decoder is composed of two consecutive convolutional layers with a kernel size of (3, 3, 3) for feature channel reduction (see details in Fig. 2) and one convolutional layer with a (1, 1, 1) kernel to generate the distribution map of 1 channel, which is then used to output the segmentation results through a Sigmoid function. We do not use skip connections between ResNet and SegNet due to its light-constructed design. Since there is no up-sampling operation in the SegNet, the spatial resolution of feature maps in this module remains unchanged after ResNet. During SegNet optimization, we down-sample the original LIDC-IDRI segmentation ground truth to the output image size by a fast zero-order spline interpolation. Dice coefficient loss is chosen for segmentation supervision of unsure-annotation data nodules, which is defined as

$$L_{seg}^{unsure} = 1 - \frac{2\sum_i^N y_{s,i}\, g_{s,i} + \varepsilon}{\sum_i^N y_{s,i} + \sum_i^N g_{s,i} + \varepsilon} \tag{1}$$

where $y_{s,i}$ and $g_{s,i}$ denote the predicted probability and ground truth class of the ith voxel belonging to the nodule region, respectively. N is the number of voxels. ε is a smooth factor to avoid dividing zero. Although the malignancy scores of LIDC-IDRI is unsure labels, its segmentation ground truth is relatively accurate.

FNet module: The FNet module integrates double-way knowledge from two sources. One stems from the output of the left ResNet module in Fig. 2, which stands for high-level semantic information. The other one is generated from the upper SegNet module, introducing the distribution feature of nodule segmentation. Since the output layer of SegNet is the feature map of one channel, we apply one convolutional layer with 256 kernels (kernel size: $3 \times 3 \times 3$, stride: 1, padding: 1) after the SegNet output to match the number of the ResNet output channels. With the same feature resolution and the number of feature channels, we fuse the feature maps from ResNet and segmentation-specific representations from the SegNet together by addition function, followed by a ReLU operation. Different from other work (Wu et al., 2018; Liu et al., 2019) that concatenates the feature neutrons in the fully connected structure, we directly fuse and activate the feature representation maps in the spatial dimension, which provides a straightforward interpretation mode when inferring CAM. After performing the global average pooling (GAP), the produced features are fed into RNet and CNet for final predictions. Such joint learning design has two benefits. On one hand, by encoding the extracted nodule segmentation information into the subsequent layers, the model localization ability and prediction performance could be enhanced. On the other hand, the gradients from RNet and CNet can also propagate to the SegNet, which can adaptively adjust the coordination between the segmentation task and prediction tasks while ultimately better serving the classification task.

RNet module: In RNet module, we formulate the nodule malignancy learning task using unsure-annotation data as a regression problem. Compared with the commonly-used methods that simply cast this task as a binary or multiple classification problem which assumes independence between classes, the regression approach has several benefits, such as (1) evading the unreasonable step of nodule label assignment in most of the related work in Section 2; (2) avoiding the huge waste of indeterminate data, especially in a data-hungry situation; (3) and flexibly leveraging the ordinal relationship of certain features from different malignancy scores of unsure-annotation data. Even though the malignancy scores may not match the true benign-malignant label, the increased suspicion level to nodule malignancy reflects the experts' recognition in terms of the severity of nodule disease. Thus, it is of great value if the model exploits experts' knowledge in RNet. Also given the evidence that there is a high correspondence between the nodule malignancy score and other characteristics (Hussein et al., 2017; Chen et al., 2016; Liu et al., 2019), encoding the intrinsic malignancy relationship into model could implicitly help obtain the nodule's heterogeneity information in size, shape and texture. The main reasons for performing regression task in RNet are as follows: (1) to make the most of unsure-annotation data knowledge (e.g., following the ordinal relationship, avoiding label assignment and data waste); (2) to ensure that the RNet is more informative to represent the radiologists' inference (see Section 3.3 and Section 4.8). Additional discussions on different modes for RNet are discussed in Section 5.1. In RNet, we use a fully connected (FC) layer that outputs one neuron to generate the unsure-annotation data malignancy score prediction. Meanwhile, RNet can also produce the experts' assessment for sure-annotation data nodules, which is vital evidence for our new nodule diagnosis strategy in Section 3.3. We use mean square error (MSE) loss to minimize the distance error between the output value and ground truth of unsure-annotation data, which is defined as

$$L_{reg}^{unsure} = \left\| y_r - g_r \right\|_2^2 , \tag{2}$$

where y_r is the output of RNet and g_r is the malignancy score of unsure-annotation data nodule.

CNet module: In the primary task module of CNet, there is a FC layer that outputs one neuron followed by a Sigmoid function to generate the probability scores for nodule benign-malignant classification. The rich knowledge gained before CNet includes (1) high-level semantic information of nodule input extracted by ResNet; (2) spatial structured-features obtained from SegNet and FNet such as nodule shape, size and location; (3) and implicitly encoded features from RNet that reflect experts' recognition to the nodule malignancy and its correlated attributes. In this task, we use the binary cross-entropy (BCE) loss to optimize the sure-annotation data error, which is defined as

$$L_{cls}^{sure} = g_c \log x_c + (1 - g_c) \log (1 - x_c) , \tag{3}$$

where x_c is the output of Sigmoid function and g_c is the benign-malignant label of sure-annotation data.

To resolve the large domain shift not only in data space but also in label space between unsure and sure-annotation data, we leverage a "divide-and-conquer" approach to bypass the domain conflicts after the GAP layer, so that working as an auxiliary task, the regression process will not have negative interference on the major classification task. Instead, in such a synergic way, CNet could learn more informative knowledge that single sure-annotation data supervision cannot provide.

3.2. Learning with CAM-SEM-Loss

Although using contextual knowledge of nodules may improve the prediction performance by data fitting algorithms, subtle statistical correlations among nodule input variables have long been an annoying problem, making models potentially more prone to overfitting on the training data. We hypothesize that in a small medical image dataset, the object background could be regarded as a potential confounder that makes the model run against the notion of stable learning (Kuang et al., 2018). Thus, rather than purely fitting the observed training data, we expect to develop a predictive model that is not only robust to the wild environment changes but also faithful to the nodule regions.

In classification tasks that are supervised by patch-level annotation, CAM is usually used as a visualization tool when a model finishes optimization, but few studies fed it back to the training progress (Wang et al., 2021). To alleviate the underperformance of model trustworthiness in visual interpretability, in Fig. 2, we propose to construct a new loss function for model regularization, called CAM-SEM-Loss (CSL), by leveraging the online generated CAM with the supervision of nodule segmentation maps (SEM). The detailed description is shown below.

For a given nodule input in each training batch, the FNet finally uses ReLU to activate the fused feature maps. Let $f_k(x, y, z)$ represents the activation of unit k at 3D spatial location (x, y, z). Then, for unit k with length L, width W and height H, the result of performing GAP, F_k, is $\frac{1}{L \times W \times H} \sum_{x,y,z} f_k(x, y, z)$. Thus, the input to the Sigmoid function after CNet, S, is $\sum_k \omega_k F_k$, where ω_k is the weight of CNet's fully connected layer corresponding to the unit k before backpropagation and optimization of each batch. Essentially, ω_k indicates the current positive relevance of F_k for the predicted malignant class or the negative relevance for the benign class. Finally, the output of the Sigmoid function, P, is given by $\frac{1}{1+e^{(-S)}}$ to represent the nodule malignancy score. By plugging F_k into the malignancy score S, we obtain

$$\begin{aligned} S &= \frac{1}{L \times W \times H} \sum_k \omega_k \sum_{x,y,z} f_k(x, y, z) \\ &= \frac{1}{L \times W \times H} \sum_{x,y,z} \sum_k \omega_k f_k(x, y, z) \end{aligned} \tag{4}$$

We define CAM as the class activation map by forward propagation, where each spatial element is given by

$$CAM(x, y, z) = \sum_k \omega_k f_k(x, y, z) \tag{5}$$

Thus, $S = \frac{1}{L \times W \times H} \sum_{x,y,z} CAM(x, y, z)$, where $CAM(x, y, z)$ directly indicates the malignancy attention at spatial location (x, y, z). Since we use Sigmoid as our final activation function, in the attention area of class activation map, $CAM(x, y, z)$ will have a lower value if the model outputs a low malignancy score and a higher value for a high malignancy score.

To standardize the CAM visualization, we introduce the malignancy score P into the original CAM. The new CAM_c for predicted class c is defined by

$$CAM_c(x, y, z) = (P - Threshold) \sum_k \omega_k f_k(x, y, z), \tag{6}$$

where $Threshold$ is the division point to classify the benign-malignant nodule, which is normally set to 0.5 and under the following relation

$$c = \begin{cases} 0\,(benign) & P < Threshold \\ 1\,(malignant) & P \geqslant Threshold \end{cases} \tag{7}$$

The min–max normalization is finally used to rescale each element of CAM_c with $CAM_c(x, y, z) \in [0, 1]$.

Besides, in the nodule segmentation map, we define $SEM(x, y, z) \in (0, 1)$ as the foreground-background prediction of the pixel at the location (x, y, z). By performing Sigmoid on SegNet outputs, most $SEM(x, y, z)$ would quickly distribute near either 0 or 1 after the training starts, providing the high confident guidance to the updating $CAM_c(x, y, z)$ with the information of whether this position belongs to nodule regions or not. This inspires us to leverage the SEM knowledge and give a interpretability constraint to CAM_c by regulating more attention on nodule regions during training. For this purpose, we first obtain the approximate average CAM values of nodule regions and background regions, which are defined by

$$AvgCAM_{ndl} = \frac{\sum_{x,y,z} CAM_c(x, y, z)\, SEM(x, y, z)}{\sum_{x,y,z} SEM(x, y, z)} \tag{8}$$

$$AvgCAM_{bkg} = \frac{\sum_{x,y,z} CAM_c(x, y, z)\,(1 - SEM(x, y, z))}{\sum_{x,y,z}(1 - SEM(x, y, z))} \tag{9}$$

where $AvgCAM_{ndl} \in [0, 1]$, $AvgCAM_{bkg} \in [0, 1]$.

Then, to drive the model to learn more discriminative feature representations from a nodule perspective, we formulate the CAM-SEM-Loss as follows

$$L_{CSL}^{sure} = max\left\{0,\ AvgCAM_{bkg} - AvgCAM_{ndl} + \delta\right\}, \tag{10}$$

which enforces $AvgCAM_{ndl} \geqslant AvgCAM_{bkg} + \delta$ in the sure-annotation data training progress, where δ is the margin parameter that adjusts the attention bias between nodule and background regions. Moreover, L_{CSL}^{sure} does not neglect the contextual information of a nodule because feature learning from the background regions is not inhibited.

Considering that the CAM-SEM-Loss would treat equally to those predictions with different malignant probability scores, we further extend CAM-SEM-Loss with the knowledge of uncertainty that the prediction with a lower confidence score (P around $Threshold$) should gain less supervision from CAM-SEM-Loss. To this end, the adaptive CAM-SEM-Loss (ad-CSL) is designed by multiplying the l_1 distance between $Threshold$ and P over CAM-SEM-Loss, which is given by:

$$L_{ad-CSL}^{sure} = 2\,\|P - Threshold\|_{l_1}\ L_{CSL}^{sure}, \tag{11}$$

where the coefficient is 2 to match the value range of CAM-SEM-Loss.

In summary, the total loss for training includes four losses in our framework:

$$L_{total} = L_{cls}^{sure} + \alpha\, L_{ad-CSL}^{sure} + \beta\, L_{seg}^{unsure} + \gamma\, L_{reg}^{unsure}, \tag{12}$$

where α, β and γ are three hyper-parameters to balance these terms which are all set to 1 in our study.

3.3. Nodule diagnosis with synergic model

We establish a new nodule diagnosis method by a retrieval algorithm based on the prior nodule information. Specifically, using the trained model K, we first obtain the outputs for a testing nodule and each prior identified nodule i in the historical database, defined as T and D_i, respectively. Generally, the historical sure-annotation data are also used for training the synergic model if nodule locations are available.

From the perspective of synergic model, the outputs of each nodule have three forms including (1) the nodule classification score from CNet, represented as the machine inference; (2) the nodule regression score from RNet, originated from the expert knowledge; and (3) the concatenation of (1) and (2).

Then, we rank the Euclidean Distance of outputs between T and each D_i, and acquire the ranking list of top K nodules with the closest Euclidean Distance to the testing nodule, which is defined by

$$List_i^K = Rank^K\left\{\left|T,\ D_i\right|_{Euc}, i = 1, 2, \ldots, num_D\right\}, \tag{13}$$

where num_D is the number of reference nodules in the historical database. It should be noted that if we use the regression score from RNet for nodule retrieval, the RNet output should be normalized by the transformation $\frac{y_r - 1}{(5-1)}$ before computing the Euclidean Distance in order for them to be aligned with the range of classification score $x_c \in (0, 1)$.

Afterwards, a new diagnosis score of the testing nodule is awarded by averaging the labels of top K reference cases in $List_i^K$, which is given by

$$Diag = \frac{1}{K} \sum_{j=i} Label_j, i \in List_i^K, \tag{14}$$

where $Diag \in [0, 1]$ and K is empirically set to 20 in this study.

By reading and matching these closely related historical nodule cases which lead to similar outcomes, clinicians can acquire more evidence and clues for the testing nodule diagnosis.

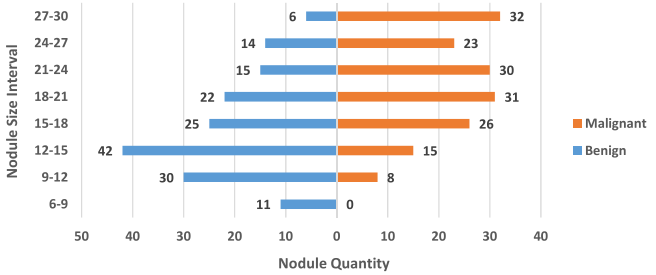

Fig. 3. Size (diameter, mm) distribution for malignant and benign nodules of sure-annotation data.

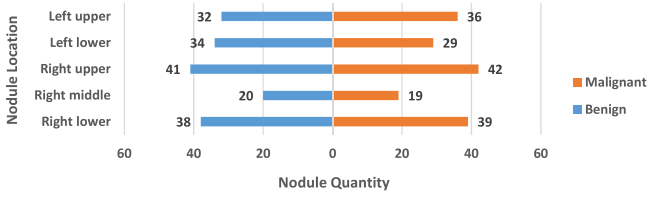

Fig. 4. Location (lung lobe) distribution for malignant and benign nodules of sure-annotation data.

Table 1
Nodule types of the sure-annotation dataset, which is according to the 2021 World Health Organization lung tumor classification guide (Nicholson et al., 2022) and a 2019 lung nodule management guide (Loverdos et al., 2019).

Type	Total
Cancer (malignant)	
Adenocarcinomas	
Minimally invasive adenocarcinoma, nonmucinous	1
Invasive pulmonary adenocarcinoma (IA[a])	121
Squamous cell carcinomas (SCC[b])	21
Large cell carcinoma	3
Adenosquamous carcinoma	2
Pulmonary blastoma	3
Small cell carcinoma	6
Metastatic carcinoma	5
Others[c]	3
Non-cancer (benign)	
Hamartoma	34
Mycobacteria	3
Fungi	12
Round pneumonia	7
Organizing pneumonia	13
Lymphaden	10
Granuloma	47
Proliferation of fibrous tissue	35
Sclerosing pneumocytoma	4
	330

[a]IA is a major type of the cancerous nodules which contains many subtypes in our sure-annotation dataset (e.g., Acinar adenocarcinoma, Papillary adenocarcinoma, Micropapillary adenocarcinoma, Solid adenocarcinoma).

[b]SCC includes subtypes in our dataset (e.g., keratinizing, nonkeratinizing).

[c]Others includes some long tail diseases such as Lymphoepithelial carcinoma, Atypical carcinoid, and Mixed carcinoma.

Table 2
TNM stages of malignant nodules (except metastatic carcinomas) of sure-annotation data, which is based on the eighth edition of the TNM classification for lung cancer staging proposed by the International Association for the Study of Lung Cancer (IASLC) (Goldstraw et al., 2016).

Overall TNM stage	Total
IA	94
IB	35
IIA	1
IIB	10
IIIA	14
IVA	6
	160[a]

[a]Note: 5 metastatic carcinomas are not available with TNM stage.

4. Experiments

4.1. Datasets

Unsure-annotation data: The unsure-annotation dataset used in this study comes from the LIDC-IDRI database (Armato III et al., 2011), which consists of 1018 CT scans with annotations of the nodules given by multiple radiologists. According to the practice in LUNA16 (Setio et al., 2017), CT scans with a slice thickness smaller than 3 mm were included (888 CTs), where nodules $\geq$ 3 mm and $<$ 30 mm accepted by at least 3 radiologists were considered as positive samples (1186 nodules). On top of that, we only involve the 919 majority solid nodules (average texture score = 5). To alleviate the inter-observer variability as regards to malignancy voting, we first discard the nodules with the mean absolute difference (MAD) (Yitzhaki et al., 2003) among malignancy scores larger than 0.6 and calculate the average malignancy scores for the remaining 686 nodules. In total, the number of nodules with an average score of 1, 2, 3, 4 and 5 is 88, 101, 338, 100 and 59, respectively. Different from any other work, further label assignment is not required in our method.

To obtain the nodule segmentation label, we first fill the internal area of the nodule boundary delineated by each radiologist. Then a 50% consensus criterion (Kubota et al., 2011) is adopted to generate a single ground-truth boundary. That is, if the current voxel is annotated by two or more radiologists, the voxel point will be regarded as a nodule label. Otherwise, it will be labeled as background.

Sure-annotation data: The sure-annotation dataset consists of 330 solid nodules (165 benign/165 malignant) obtained from 317 patients' chest CT scans (179 male/138 female). To form the sure-annotation data cohort, 350 random patients with lung nodules were collected retrospectively in Shanghai Chest Hospital from February 2017 to January 2021. Pathological examination for each target nodule of these patients was available via surgical resection. The collection and analysis of image data and medical records were approved by the ethical committee of the Shanghai Chest Hospital and adhered to the tenets of the Declaration of Helsinki. In this study, we only included the nodules with a diameter less than 30 mm (consistent with the accepted upper limit of nodule size (Hansell et al., 2008)) and larger than or equal to 3 mm (lower limit for practical consideration (Armato III et al., 2004)). Eligible 165 benign nodules were selected with required nodule size, nodule texture (solid), and CT quality. Eligible malignant nodules were sampled and matched to the same quantity of benign nodules. The last CT scans before surgery were chosen to enable a small time gap between nodule images and the pathological ground truth. Together, 330 nodules from 317 patients' CT scans finally formed our full sure-annotation data cohort. Each nodule had a single pathological-proven class (benign or malignant). Three CT scans contained both benign and malignant nodules. The age of patients ranged from 25 to 81 years, with an average of 57.58 ($\pm$ 10.99) years. CT scans in this dataset were acquired by multiple manufacturers of GE Medical Systems, Philips and United Imaging Health (UIH), where the slice thickness ranges from 0.5 to 3.0 mm with an average of 1.14 ($\pm$ 0.32) mm and the pixel spacing varied from 0.34 to 0.98 mm with an average of 0.58 ($\pm$ 0.22) mm. The distributions of nodule size, nodule location, nodule type, and cancer stage are reported in Fig. 3, Fig. 4, Table 1, and Table 2, respectively.

4.2. Data preprocessing

Considering that the valid nodules are restricted inside the two lungs, we think it reasonable to first perform robust lung segmentation

Table 3
Quantitative performance of the synergic model for the ultimate sure-annotation data nodule benign-malignant classification task by 5-fold cross-validation (under the threshold of 0.5), including combination with different modules, comparison with attention mechanism and different loss functions.

	Tasks		Modules	Params ($\times 10^6$)	Results (%) (mean ± standard deviation)							vis[a]
	Sure	Unsure			Sensitivity	Specificity	Precision	Precision$_b$	Accuracy	AUC	F1-score	
A	Cls	–	C	3.6282	64.24 ± 4.85	58.79 ± 8.70	61.28 ± 4.36	62.05 ± 3.92	61.52 ± 4.02	69.53 ± 3.39	62.54 ± 3.18	✓
B	Cls	Seg	C,S	3.8770	67.27 ± 9.27	64.85 ± 14.16	66.93 ± 9.89	66.49 ± 7.32	66.06 ± 7.46	74.67 ± 8.39	66.47 ± 6.74	–
C	Cls	Reg	C,R	3.6284	66.67 ± 6.64	67.27 ± 10.03	67.72 ± 6.02	67.04 ± 3.55	66.97 ± 4.00	76.80 ± 2.90	66.84 ± 3.68	–
D	Cls	Seg+Reg	C,S,R	3.8773	68.48 ± 8.48	**69.70** ± 3.83	**69.29** ± 1.98	69.35 ± 4.91	69.09 ± 3.26	76.51 ± 2.96	68.64 ± 4.94	✓
E	Cls	Seg+Reg	C,S,R,ARL[b]	3.8773	69.70 ± 6.06	67.27 ± 8.44	68.57 ± 4.55	69.15 ± 2.80	68.48 ± 2.61	76.95 ± 6.43	68.80 ± 2.44	✓
F	Cls	Seg+Reg	C,S,R,F	3.8842	75.15 ± 4.02	63.03 ± 11.08	67.69 ± 5.04	71.63 ± 1.15	69.09 ± 3.66	77.23 ± 6.23	70.94 ± 1.43	✓
G	Cls	Seg+Reg	C,S,R,F,ARL[b]	3.8842	71.52 ± 3.09	66.06 ± 9.07	68.28 ± 6.76	69.62 ± 4.45	68.79 ± 5.47	76.38 ± 5.03	69.75 ± 4.46	–
H	Cls (CL[c])	Seg+Reg	C,S,R,F	3.8842	72.73 ± 4.29	62.42 ± 6.24	66.09 ± 4.06	69.59 ± 3.97	67.58 ± 3.90	77.34 ± 5.14	69.17 ± 3.51	✓
I	Cls (CSL)	Seg+Reg	C,S,R,F	3.8842	71.52 ± 7.81	64.85 ± 12.36	68.16 ± 6.83	69.80 ± 2.97	68.18 ± 3.46	77.37 ± 5.91	69.15 ± 2.67	✓
J	Cls (as-CSL)	Seg	C,S,F	3.8839	74.55 ± 7.81	61.82 ± 14.42	67.23 ± 7.02	70.93 ± 4.10	68.18 ± 5.34	76.33 ± 5.51	70.11 ± 4.00	–
K	Cls (ad-CSL)	Seg+Reg	C,S,R,F	3.8842	**76.97** ± 4.11	64.24 ± 9.27	68.67 ± 5.67	**73.42** ± 4.87	**70.61** ± 5.30	**77.65** ± 5.64	**72.46** ± 4.20	✓

[a]CAMs shown in Fig. 6.
[b]ARL–Attention Residual Learning (Zhang et al., 2019).
[c]CL–CAM-Loss (Wang et al., 2021).

of chest CT scans as a prerequisite for automated nodule analysis, which is conducted by: (1) binarization using OTSU (Otsu, 1979) as threshold selection method; (2) extraction of the largest connected components A; (3) hole filling on A to obtain B; (4) coarse lung segmentation C by subtracting A from B; (5) denoising on C to obtain D; (6) closing operation on D using non-flat morphological structuring element to obtain E; (7) binarization on E using manually set threshold (0.5) to obtain the robust segmented lung mask F; (8) multiplication on the raw CTs and the mask F. Steps (6) and (7) are key guarantees for robust lung segmentation where abnormalities are often present. Otherwise, many juxta-pleural nodules will be missed due to erroneous lung segmentation (Armato III and Sensakovic, 2004).

We clipped the Hounsfield unit (HU) values into the lung window interval [-1000, 400 HU] and mediastinal window interval [-160, 240 HU], respectively. Then each window is normalized to the linear range of [0, 1]. To resolve the anisotropic nature of CTs, the 3D images were resampled to a fixed 0.5 mm/voxel along all three axes using spline interpolation. To extract 3D patches, we first consider a cube of $64 \times 64 \times 64$ voxel, which could completely enclose the nodule with a diameter less than 30 mm. Then we concatenate the cubes from the lung and mediastinal window to obtain the 2-channel nodule input.

The nodule volume augmentation method includes random flipping over the three axes, random rotation around the three axes with angles chosen from 90°, 180°, or 270°, and random transposing by reversing or permuting between every two axes of the 3D image.

4.3. Experiment setting

All the experiments are implemented in PyTorch (Paszke et al., 2019) with a single NVIDIA GeForce GTX 1080 Ti GPU and learned using Adam optimizer (Kingma and Ba, 2014) with the initialized learning rate of 1e-3 and the maximum epoch number of 100 (batchsize = 1). The iteration times in each epoch is based on the number of sure training data. The validation set occupies 20% of the training set in each experiment to monitor the performance of training model. All the experiments and results involving or having involved the training of sure-annotation data are strictly conducted by 5-fold cross-validation, in which four subsets were employed to train the model, one subset was used as testing data.

4.4. Evaluation metrics

To evaluate the model performance comprehensively, our evaluation metrics include Sensitivity (Recall), Specificity (also called Recall$_b$, if treating benign as positive sample), Precision, Precision$_b$ (Precision in benign class), Accuracy with the cut-off value of 0.5, AUC (area under the receiver operating characteristic curve) and F1-score.

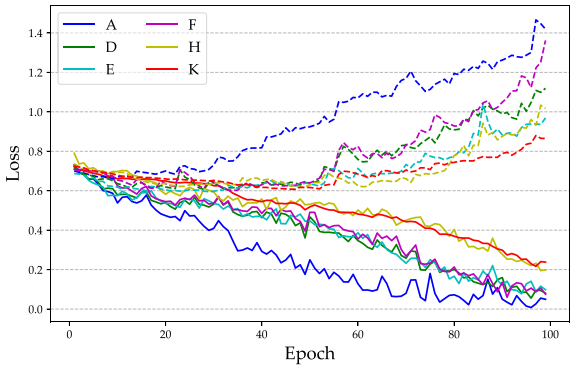

Fig. 5. Binary cross-entropy (BCE) loss error vs. epoch for different models in Table 3 in training (80%) and validation (20%) performances. Solid lines represent training losses and dashed lines denote validation losses.

4.5. Quantitative evaluation of multi-task learning in synergic model

In this subsection, we conduct ablation study and comparison experiments for quantitative evaluation of sure-annotation data performance with synergic model. To observe the contributions of different modules in our synergic model, we first combine different modules based on model A. As the numerical results illustrated in Table 3, model A, consisting of a ResNet (backbone) and a CNet, can hardly obtain the malignancy discrimination ability with only the sure-annotation data knowledge. This is probably restricted by the limited amount of sure-annotation data that cannot drive the model to extract the malignancy-related features supervised by definite nodule labels. Model B, C and D first attempt to integrate the sure and unsure-annotation data using an online co-training network in nodule malignancy classification task.

Although there is a large domain shift between two datasets (in both data and label space), additional learning with unsure-annotation data nodules simultaneously in auxiliary tasks, model B, C and D outperform model A in all the evaluation metrics by an average 4.7%, 5.6% and 7.3%, respectively. Such significant promotion indicates the potential of leveraging unsure-annotation data knowledge for sure-annotation data discrimination. Among these three structures, model D obtains the superimposing positive effects of adding SegNet and RNet modules. Based on model D, model F bridges the nodule segmentation features to the end of backbone for both unsure and sure-annotation data, achieving better performance in Sensitivity, Precision$_b$, AUC and F1-score.

We choose two typical existing techniques that are possible to facilitate nodule heterogeneity discrimination. The first technique is the attention mechanism. We replace the Residual blocks of model D

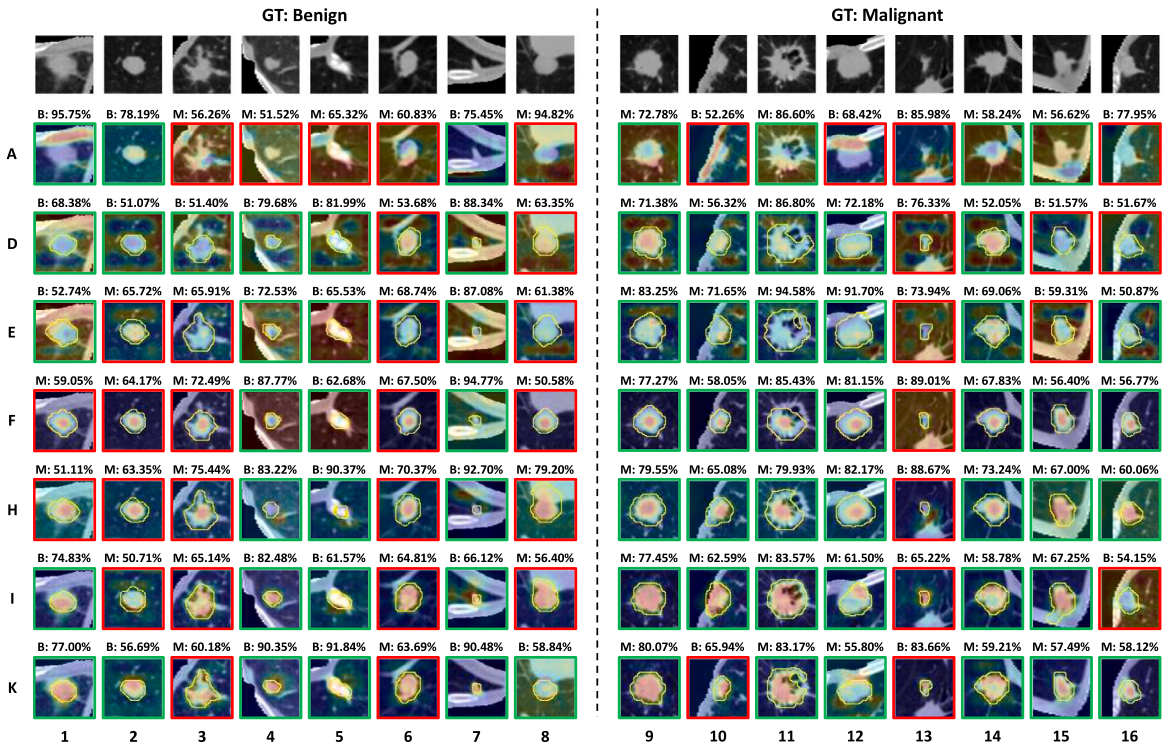

Fig. 6. Visualization of CAMs obtained by models chosen from Table 3 which are A, D (A + SegNet + RNet), E (D + ARL), F (D + FNet), H (F + CL), I (F + CSL) and J (F + ad-CSL). The first row shows the nodule inputs from sure-annotation data. "B" and "M" are the benign and malignant prediction results under the threshold of 0.5. The scores are the probabilities for each predicted class. The yellow contours on each image depict the automatic segmentation of sure-annotation data nodules generated from SegNet followed by Sigmoid function (threshold: 0.5). Green or red border to an image denotes the true or mistaken prediction, respectively. Both the input images and CAMs are taken from the central slices of their 3D patches.

Table 4
Performance of the models in Fig. 5 tested in a completely independent LUNGx dataset (%; under the threshold of 0.5).

Model	Sen	Spe	Pre	Pre_b	Acc	AUC	F1
A	65.85	50.00	56.25	60.00	57.83	63.41	60.67
D	70.73	64.29	65.91	69.23	67.47	66.84	68.24
E	73.17	61.9	65.22	70.27	67.47	**67.48**	68.97
F	**75.61**	66.67	68.89	73.68	71.08	66.84	72.09
H	70.73	59.52	63.04	67.57	65.06	65.45	66.67
K	**75.61**	**69.05**	**70.45**	**74.36**	**72.29**	67.31	**72.94**

and F by ARL (attention residual learning) blocks which work well for 2D skin lesion classification (Zhang et al., 2019). As suggested by Zhang et al. (2019), we apply spatial attention which is more suitable than channel attention and mixed attention to help focus on semantic regions of a target object. As the results of model E and G shown in Table 3, spatial ARL block has limited improved performance or local degradation compared with model D and F. One possible reason is that attention mechanisms like (Chen et al., 2017; Hu et al., 2018) would further enhance the correlated feature learning inside the model which may not fit for small-sample tasks. The second technique is CAM-loss (CL) (Wang et al., 2021) which regulates the feature maps by minimizing the difference between CAM and class-agnostic activation map (CAAM) in natural image classification tasks. We test the effect of CL based on model F. The results of model H show that except AUC, other metrics drop comprehensively when applied this extra loss in our task. We explain this phenomenon in the next Section 4.6.

To evaluate our loss function, we first use CAM-SEM-Loss (CSL) in model I. Compared to model F, model I appears a performance degradation which is similar to model H. Under such circumstances, an upturn happened to model K with adaptive CAM-SEM-Loss (ad-CSL) by encoding the uncertainty of nodule prediction to CSL, which outperforms any other model broadly, especially for Sensitivity, $Precision_b$, Accuracy and F1-score. In addition, Fig. 5 shows that, model K has a higher training error but lower validation error in BCE loss for sure-annotation data classification compared with other models, which indicates that our method has a positive effect on reducing overfitting.

Moreover, we have conducted a generalizable study to test the models in a completely independent public dataset: LUNGx Challenge dataset (Kirby et al., 2016) (approximating sure-annotation data), which consists of 70 thoracic CT scans with a total of 42 benign nodules (including 6 confirmed cases based on pathologic assessment, 15 confirmed cases based on nodule stability for at least 2 years, and 21 confirmed cases based on nodule resolution) and 41 malignant nodules (pathology-confirmed). The pre-trained models are taken from those in Fig. 5 which are trained and validated with all our sure-annotation data. As shown in Table 4, the improvements of adding different modules are in agreement with the corresponding performances in Table 3. The constraint of ad-CSL benefits model K significantly in cases of testing data with potential domain shifts, proving that our method could be beneficial to learn domain-invariant nodule features for better generalizability.

Up to the present, these quantitative results in Table 3, Fig. 5 and Table 4 affirm the superiority of aggregating multi-task modules and ad-CSL loss for discriminating high-correlated features. Most published papers ended the experiments here, which constrained the analytical insight instead of advancing it. As a matter of fact, quantitative evaluation might be inadvertently involved in a hidden fraud that the nodule malignancy prediction is not derived from the reasonable evidence for model reasoning, but rather from the outcomes of data fitting.

4.6. Qualitative evaluation of synergic model

For further visualization analysis, we adopt the class activation mapping (CAM) for sure-annotation data classification to reveal the attention regions identified by different models in Table 3. Because all

the models are constructed in a common structure ending with a GAP and a fully connected layer , we apply Eq. (6) followed by a min–max normalization to express the interpretation results in Fig. 6.

As illustrated in model A, there is no guarantee for a common black-box CNN model to learn nodule-relevant features with limited knowledge and supervision. As a result, the small amount of sure training data causes overfitting quickly since an early stage, as illustrated in Fig. 5. Based on the cases of A-1, A-10 and A-12 in Fig. 6, we can interpret that model A may remember the juxta-pleural feature of benign nodules during training but it suffers from domain bias when the model is applied on testing data.

With the incorporation of unsure-annotation data knowledge, model D appears some regular patterns of CAMs that are guided by the nodule segmentation map. As marked by the yellow contour lines in Fig. 6, sure-annotation data nodules can be well automatically segmented by the light-weight SegNet branch which is trained by unsure-annotation data. Although there is no direct shortcut that introduces the segmentation knowledge to the end of backbone in model D, an effective constraint can be exhibited in D-1, D-10 and D-12. However, in some CAMs, model D not only highlights the features in nodule regions but also activates massive background regions because nodule segmentation learns both foreground and background information. Nevertheless, jointly learned with sure and unsure-annotation data, model D obtained a significant improvement in Table 3 but implicitly failed in interpretation performance.

Compared to model D, we observe a worse CAM result in model E that the spatial ARL block (Zhang et al., 2019) aggravates the incorrect concentration on background for benign predictions (e.g., E-1, E-4, E-5, E-13 and E-15) and reduces the nodule feature learning for malignant prediction (e.g., E-6, E-8 and E-9). This indicates that attention techniques may compound the preconception error if the correct features are not guaranteed in advance.

For model F that bridges the SegNet features to the final convolution layer of model D using an FNet, its visual saliency maps preserve an evident involvement of nodule segmentation knowledge, whereas most benign predictions preserve similar unexpected attention as the model D and E. Added with CAM-loss (Wang et al., 2021) which was applied in natural image classification tasks, model H could erase part of the background attention of benign predictions, but it failed to highlight their nodule regions. This may be caused by the inherent differences between classification tasks using natural images and nodule images. In contrast to natural images that possess large training samples with obvious class-discriminative features, (1) nodule malignancy classification is a one-object and two-category task; (2) the benign-malignant information of sure nodule data is visually interchangeable even for experimented radiologists; (3) as a carrier for radiomics, the representation of a CT scan is restricted by its image acquisition mode that invasive pathological knowledge may not be captured; (4) sure nodule data-hungriness issue makes CNNs unreliable; (5) the attention maps cannot serve as reliable priors for CAM-Loss in nodule task (CAM-loss is under the assumption that their attention maps can still serve as reliable priors for tasks).

By incorporating CSL loss, a powerful interpretability constraint designed for nodule images, the model I can put more emphasis on the feature variables in nodule area either for malignant predictions or benign predictions (e.g., I-4, I-7 and I-13). Benefiting from the encoding of prediction uncertainty, the model K optimized with additional ad-CSL presents a stronger attention ability that its produced CAM maps are highly calibrated with nodules, and therefore make more reliable predictions in Table 3 and faithful feature representations in Fig. 6. This is attributed to the key role of adaptive strategy in ad-CSL that enables the model to first strengthen discrimination ability and segmentation performance if a nodule prediction is of low confidence while focusing more on semantic information of nodules. Note that, it does not matter if the prediction is incorrect because the wrong predictions should also have the correct CAMs.

Table 5
Performance comparison of different synergic model structures and other state-of-the-art results for lung nodule classification using LIDC-IDRI dataset (%) .

	Methods	Sen	Spe	Pre	Pre_b	Acc	AUC	F1
Others	Shen et al. (2017)	77.00	93.00	–	–	87.14	93.00	–
	Hussein et al. (2017)	–	–	–	–	91.26	–	–
	Xie et al. (2018b)	84.40	90.88	82.09	–	88.73	94.02	83.23
	Xie et al. (2017)	83.83	94.56	88.40	–	91.01	95.35	86.07
	Xie et al. (2018a)	86.52	94.00	87.75	–	91.60	95.70	87.13
	Xie et al. (2019)	84.94	96.28	–	–	92.53	95.81	–
	Xu et al. (2020)	85.58	95.87	90.39	–	**92.64**	94.00	87.91
Ours	C	76.15	87.23	84.19	81.44	82.17	90.53	79.65
	C,S	79.90	90.97	88.34	84.38	85.91	92.99	83.83
	C,S,F	**88.10**	90.94	89.53	**90.40**	89.65	95.23	88.57
	C,S,F,ad-CSL	85.54	**96.81**	**95.85**	88.85	91.67	**96.10**	**90.38**

4.7. The classification performance of synergic model and comparison with state-of-the-art using unsure-annotation dataset

Table 5 displays the results of our models and other state-of-the-art methods using only LIDC-IDRI dataset. Following the common label assignment approach in other methods, we choose average score 3 as a division point and assign benign (malignant) labels to these nodules with average score 1 or 2 (4 or 5). Nodules with average score 3 were not considered in this study. The 5-fold cross-validation results show that our synergic model structure (without the application of RNet for regression task) can also be applied to unsure-annotation data for its nodule benign-malignant determination (binary classification). Compared within our different models, the module SegNet and module FNet can make a significant contribution in terms of each evaluation metric. When the model with SegNet and FNet is additionally optimized by ad-CSL loss function ($\delta = 0.5$), it can achieve much higher Specificity and Precision but lower Sensitivity and Precision_b, indicating that ad-CSL ($\delta = 0.5$) works more effectively for the prediction of benign nodules which are defined by unsure-annotation data.

Compared with other state-of-the-art methods, our model obtain the best results in Specificity, Precision, AUC and F1-score on minority samples with small batch size and little parameter tuning. Nevertheless, comparison between our model and state-of-the-art methods is beyond the scope of this study that mainly contributes to the integration of two datasets' knowledge and regularization method for faithful feature learning. More importantly, unsure-annotation data should deliver the dominant position to sure-annotation data under the consideration of scientific strictness and clinical validity.

In addition, there probably be a trade-off between good quantitative results and faithful nodule feature learning for unsure-annotation data, where the malignancy scores could be only derived from visible features by multiple radiologists whose decisions are made under the assumption that all the CT scans belong to a 60-year-old male smoker. We hence argue that without the pathologically-proven labels for training and verification, many proposed methods, which have achieved prominent cancer prediction results on unsure-annotation data, should be further reexamined in clinical practice. Meanwhile, such disenchantment should be set and reinforced in our community.

Moreover, there also exists another trade-off between learning from nodule regions and background. As demonstrated in Section 5.2 and Section 5.3, sure-annotation data could achieve better performance if we regularize the attention bias to nodule regions. We do not deny the value of contextual information for nodule discrimination (Liu et al., 2021). Instead, we believe that for small training samples, concentrating more attention on nodule regions could help to learn domain-invariant features while reducing over-fitting, because background variables are more likely to be a confounding factor in such situation. In fact, due to the difficulty in data acquisition, this task is hard to have massive sure-annotation data.

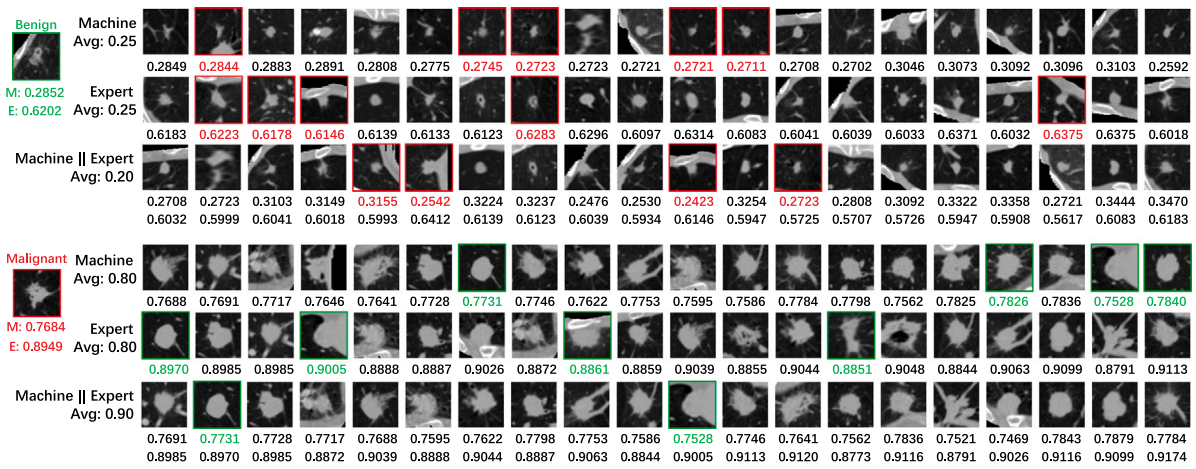

Fig. 7. The illustration of similar nodule retrieval (K = 20) using the new computer-aided diagnosis (CAD) method. The upper half shows similar nodules relative to a testing nodule whose ground truth is benign while the lower half tests on a malignant nodule. For machine reasoning, the scores below each image denote the malignant probability generated by CNet followed by a Sigmoid. For expert reasoning, the scores represent the normalized malignancy score regression output from RNet. Red or green border to an image denotes the wrong retrieved class for benign or malignant nodule, respectively.

Table 6
Performance of new nodule diagnosis strategy and five comparison retrieval methods by treating the new diagnosis score as malignant probability and conducting evaluation under the same metrics on sure-annotation data (%).

Method	Sen	Spe	Pre	Pre_b	Acc	AUC	F1
(1) Contrastive (Koch et al., 2015)	71.52	59.39	63.78	67.59	65.45	68.70	67.43
(2) Triplet (Hoffer and Ailon, 2015)	71.52	60.61	64.48	68.03	66.06	71.69	67.82
(3) Matching (Vinyals et al., 2016)	70.91	62.42	65.36	68.21	66.67	71.70	68.02
(4) Relation (Sung et al., 2018)	73.94	58.79	64.21	69.29	66.36	70.96	68.73
(5) Margin ranking (Liu et al., 2019)	70.30	64.85	66.67	68.59	67.58	73.80	68.44
(6) Machine (CNet)	76.97	62.42	68.22	72.48	69.70	76.22	72.01
(7) Expert (RNet)	76.97	64.85	69.17	73.38	70.91	77.07	72.73
(8) Machine ‖ Expert	**78.18**	**65.45**	**70.65**	**74.15**	**71.82**	**77.67**	**73.88**

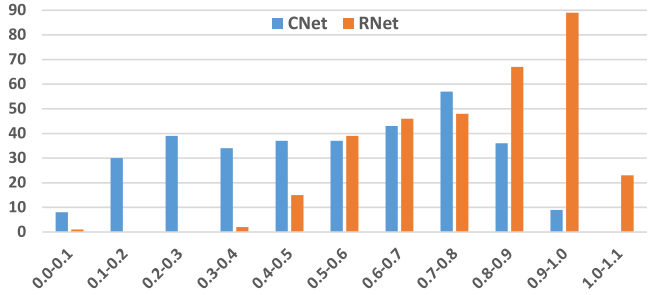

Fig. 8. The distributions of malignant probabilities generated by CNet (+ Sigmoid) and malignancy scores generated d by RNet. To produce this, model K was trained based on the whole sure-annotation data. Afterwards, the outputs from CNet (+ Sigmoid) and RNet are collected. Note: the overstepping abscissa value (1.0–1.1) results from the application of regression task in RNet.

4.8. Evaluation of new nodule diagnosis strategy

We hypothesize that the variance information of intra-class prediction can bring about the similarity of a nodule pair in different feature spaces and these predictions are able to represent the overall assessment of a nodule characteristic such as malignant probability through CNet and malignancy score through RNet. Based on this hypothesis, Section 3.3 proposed a new computer-aided diagnosis (CAD) strategy that retrieves the most similar nodules in a historical database and generates another diagnosis score for testing nodules.

Under such a diagnosis strategy, we treat the new diagnosis score as malignant probability and evaluate using the same metrics for testing nodules. As shown in Table 6, based on model K, the performance of machine reasoning (model 6) is slightly worse than that of expert reasoning (model 7), which is broadly in line with the performance of model K in Table 3 using the traditional diagnosis method. This indicates the potential to leverage the expert knowledge from unsure domain in testing phase although we do not align the distribution between sure-annotation data and unsure-annotation data in RNet. After adding the conditions for the judgment of similarity by concatenating the cognition from machine and experts (model 8), one can achieve better performance of nodule classification, even surpassing its original model K.

It is worth noting that although RNet cannot directly contribute to cancer prediction and learning with unsure-annotation data, it can nevertheless improve nodule classification performance (model A vs. model C, model B vs. model D, model J vs. model K in Table 3) interactively in the multi-task learning framework. Moreover, in our nodule retrieval study, RNet can provide a malignancy regression score for the nodule to be diagnosed and help differentiate nodule variance for better nodule retrieval performance.

To evaluate the performance gain compared to existing retrieval-based methods, we show in Table 6 detailed quantitative results obtained. For a fair comparison with our method without modifying the main architecture of the original model, we added another branch (parallel to RNet and CNet) for the first four methods in Table 6 to generate feature embedding (8 neurons) by an FC layer. Information for the five comparative methods includes: (1) Siamese network (Koch et al., 2015) with contrastive loss (margin = 2) (Hadsell et al., 2006); (2) Triplet network with triplet loss (margin = 2) (Hoffer and Ailon, 2015); (3) Matching network (Vinyals et al., 2016) that replacing the original Cosine Distance with Euclidean Distance before Softmax to be consistent with other methods' similarity function; (4) Relation network (Sung et al., 2018) with a relation module consisting with two FC layers (16×8 and 8×1 dimensional, respectively) after feature embedding; (5) Margin ranking loss (Liu et al., 2019) that is additionally added after the output of CNet and trained in a pair-wise manner. Results show that the retrieval performances of other methods are inferior to our approach mainly due to: (1) Metric learning methods are better suited to multi-class problems than binary-class tasks; (2) Additional loss and

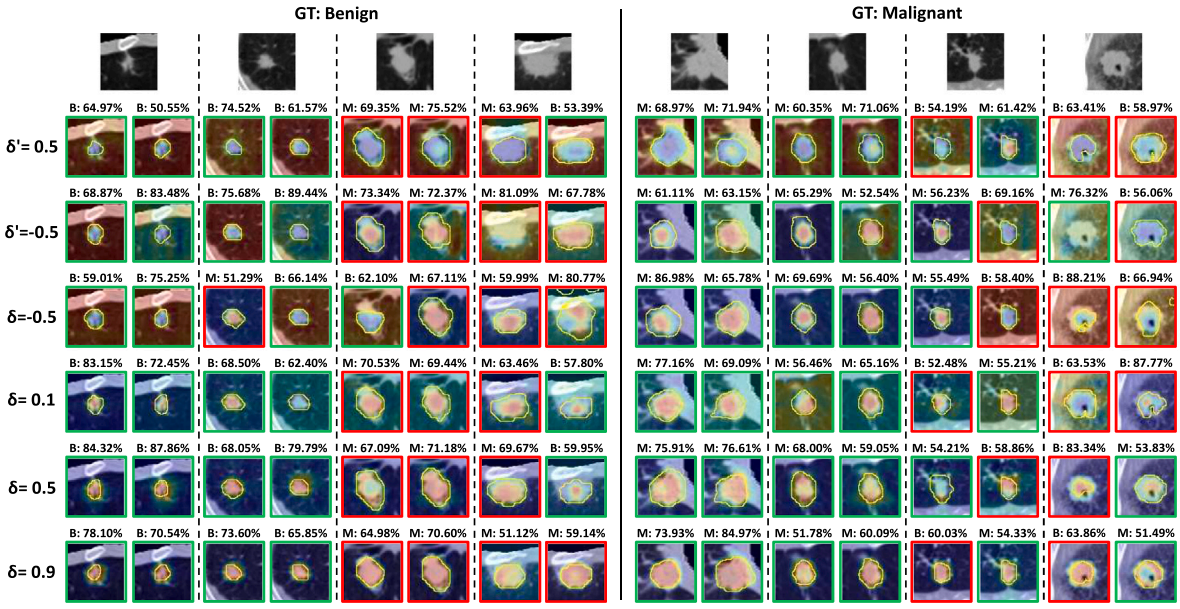

Fig. 9. Comparison of CAMs generated by models optimized with CSL or ad-CSL using different margin parameters. δ denotes the margin parameter with a bias toward nodule regions while δ' represents that with a bias toward background regions. CAMs in the left and right columns of each nodule image come from the model using CSL and ad-CSL, respectively.

similarity constraint are possible to make the pair-wise nodule relation "over-confident" with strong supervision. In contrast, our method first conducts benign/malignant nodule classification and subsequently uses the score variance of intra and inter-class for similar nodule retrieval. This two-step approach not only saves the computational resources (single network vs. Siamese/triplet network) during training but also takes the advantage of unsure/sure-annotation data knowledge during inference for improved nodule retrieval and classification. In Fig. 7, we present an intuitive nodule diagnosis result to demonstrate the advantages mentioned above. Other than the binary prediction that most studies considered, we further provide the top 20 similar nodules as diagnosis references that make our system more user-friendly and practical. Moreover, prior knowledge of historical nodule cases can also be reused to facilitate nodule diagnosis in clinical practice.

The aforementioned nodule retrieval performance and illustration have experimentally verified the effectiveness of our retrieval approach that is based on the hypothesis about the intra-class variance of model outputs. To further test our hypothesis, we present in Fig. 8 the distributions of the classification scores (through CNet) and regression scores (through RNet) for sure-annotation data. As shown in the figure, the distribution of classification scores for sure-annotation data is dispersed from 0 to 1, which creates favorable conditions for nodule retrieval. While trained by unsure-annotation data, RNet tends to generate high regression scores for most of the sure-annotation data, and the variance of regression scores remains apparent. Accordingly, the concatenation of classification and regression scores is more suitable for representing the difference for similar nodule clustering.

5. Discussion

5.1. Comparing different modes for RNet

We compare the influence of different modes for RNet in Table 7 based on model C and model D that both apply the RNet to learn malignancy scores of unsure-annotation data. First, cross-entropy (CE) loss is commonly used for multi-class discrimination tasks. However, this loss function has never been applied to the five-category classification task of LIDC-IDRI malignancy scores, which is often formulated as binary classification. The second mode (Ord) applies ordinal regression for RNet to model the ranking of malignancy scores. For this mode, we use the effective method from (Diaz and Marathe, 2019) that converts unsure-annotation data labels into soft probability distributions pairing with cross-entropy loss. For these two modes, we modify the output neuron number of the FC layer to 5 and add a Softmax function to generate the prediction in RNet.

Table 7
Comparison between models with different modes for RNet (%).

Model	Reg/Cls	Sen	Spe	Pre	Pre_b	Acc	AUC	F1
C	CE	68.48	63.03	65.37	67.65	65.76	72.95	66.33
	Ord	68.48	62.42	64.51	66.63	65.45	72.36	66.39
	MSE	66.67	67.27	67.72	67.04	66.97	76.80	66.84
D	CE	67.88	67.27	67.94	67.94	67.58	77.02	67.52
	Ord	67.88	66.67	67.15	67.48	67.27	75.02	67.48
	MSE	68.48	69.70	69.29	69.35	69.09	76.51	68.64

The result in Table 7 shows that the regression mode (MSE) obtained the best performances in both model C and model D. The traditional cross-entropy loss, which fails to model the ordinal relationship among five malignancy classes, has worse overall performances compared to the regression mode (MSE), especially in model C. We also observe that ordinal regression mode (Ord) is less likely to have good cooperation with the major task during optimization, leading to the obstruction of sure-annotation data learning. Thus, regression mode is a simple and effective way to leverage the ordinal variables that resided in unsure-annotation data and finally contribute to the sure-annotation data classification. The results in Table 7 also demonstrate that the segmentation task can improve the model discrimination ability comprehensively under the comparison between model C and model D.

5.2. Margin parameter δ

In this experiment, we explore the influence of margin parameter δ in CAM-SEM-Loss (CSL) and adaptive CAM-SEM-Loss (ad-CSL). As can be seen in Table 8 and Fig. 9, to better illustrate the working mechanism of these two loss functions, we add mode 1 ($bkg \geqslant ndl + \delta'$) that enable the model to learn more discriminative features in the background, whose CSL is formulated as

$$L_{CSL}^{sure} = max\left\{0,\ AvgCAM_{ndl} - AvgCAM_{bkg} + \delta'\right\}, \tag{15}$$

Table 8
Performances of models using CAM-SEM-Loss (CSL) or ad-CSL with different margin parameters δ/δ' in two modes (%).

Mode	δ/δ'	Loss	Sen	Spe	Pre	Pre_b	Acc	AUC	F1
bkg $\geqslant$ ndl+δ'	0.5	CSL	67.27	68.48	68.38	67.68	67.88	77.36	67.68
		ad-CSL	51.52	72.73	71.5	61.85	62.12	76.8	53.29
	−0.5	CSL	70.91	61.21	65.48	68.25	66.06	76.11	67.26
		ad-CSL	73.94	68.48	70.07	72.87	71.21	75.74	71.80
ndl $\geqslant$ bkg +δ	−0.5	CSL	73.33	64.24	67.51	70.80	68.79	77.04	70.14
		ad-CSL	73.33	63.64	67.05	70.37	68.48	75.45	69.98
	0.1	CSL	71.52	63.64	66.57	68.94	67.58	76.58	68.87
		ad-CSL	76.97	64.85	68.83	73.89	70.91	77.76	72.54
	0.5	CSL	71.52	64.85	68.16	69.80	68.18	77.37	69.15
		ad-CSL	76.97	64.24	68.67	73.42	70.61	77.65	72.46
	0.9	CSL	64.85	73.94	72.05	67.94	69.39	77.12	67.79
		ad-CSL	81.82	58.79	67.42	77.35	70.30	77.59	73.30

Driven by Eq. (15), we successfully control the CAM to focus on the background in most cases of $bkg \geqslant ndl + 0.5$ in Fig. 9. Although the degradation of quantitative performance appears for ad-CSL in Table 8, there is an interesting phenomenon that CSL can still maintain acceptable evaluation results. This indicates that learning from the background can still realize data fitting to some extent when no reliable features are provided.

In normal circumstances, we regulate δ with positive values in mode 2 ($ndl \geqslant bkg + \delta$), where ad-CSL is prone to have greater results than CSL in overall quantitative evaluation Table 8. When δ rises to 0.9, an aggressive parameter that forces the model to pay nearly all the concentration on nodule regions, its salient visualization map will present a "cleaner" attention in the background regions compared to $\delta = 0.1$ and $\delta = 0.5$ for both CSL and ad-CSL in Fig. 9. This action guarantees the high faithful feature learning from nodule regions, but it will cause performance fluctuation, especially for Sensitivity and Specificity.

While in cases of $ndl \geqslant bkg + 0.1$ and $ndl \geqslant bkg + 0.5$ that obtain the top overall quantitative performance in Table 8, we can see in Fig. 9 that CSL and ad-CSL still consider the feature learning from nodule background as we formulate (ad-) CSL based on the relativity between foreground and background instead of simply penalizing the CAM values of background to zero.

Besides, we evaluate the performance when setting δ/δ' with a negative value −0.5 for both modes. Such a conservative parameter setting may have few impacts on CAM results which show the a similar attention pattern compared with model F in Fig. 6.

Overall, we give the conclusion for this experiment: (1) δ can adjust the attention weight between nodule regions and background regions; (2) ad-CSL performs better than CSL in quantitative evaluation when guiding attention to the nodule regions; (3) there is an attention balance between nodule and its background for good lung cancer prediction; (4) the quantitative performance cannot reflect the reliability of model learning, misleading the observers accidentally.

5.3. Influence of using different input patches

We argue that nodule inputs matter for the whole deep learning process. As shown in Fig. 10, we consider four input patches that could be used for nodule benign-malignant classification task: (1) the original 64 × 64 × 64 voxel size cube (64); (2) the nodule size cube (x) that is cropped according to the nodule coordinate and radius; (3) a new 64 × 64 × 64 voxel size cube (x-resizing-64) that is generated by resizing the nodule size cube using cubic interpolation; and (4) another 64 × 64 × 64 voxel size cube (x-padding-64) that is produced by performing zero padding on the nodule size cube. For unsure-annotation data, the ground truth of the segmentation map follows the transformation of its nodule data. The patches of x and x-resizing-64 can be approximately regarded as nodule-level inputs.

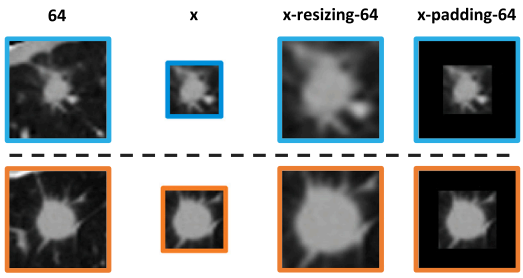

Fig. 10. Sample images of sure-annotation data with different input patches. Up: benign nodule; down: malignant nodule.

We choose Model A and Model D in this experiment because FNet seems redundant for the other three new input patches that they have already possessed the knowledge of nodule size and location. We additionally conduct ad-CSL ($\delta = 0.5$) based on model D to evaluate the effect of interpretability constrain on these different inputs.

As shown in Table 9, it can be observed that training with model D leads to a broad growth relative to model A, indicating the significant value of integrating unsure-annotation data knowledge for each input. If adding ad-CSL during model D training, the performance of using original input 64 could also get promoted even without FNet. However, ad-CSL can hardly empower model D with improved overall quantitative performance for input x-resizing-64 and input x-padding-64 in Accuracy and F1-score. Meanwhile, the model fed with input x produces more false-positive predictions. It is not difficult to give the reason for such a situation. If a kind of input has already embedded the function of nodule region expression and background suppression into itself, ad-CSL would lose its value. Therefore, learning with more nodule semantic information, model A fed with input x and input x-padding-64 can outperform the same model fed with input 64 by only using sure-annotation data.

In this experiment, several phenomena were revealed: (1) The background of a nodule could also contribute to the extraction of discriminative nodule features corresponding to the final prediction (64 vs. x-padding-64); (2) Only learning from the CT information in the nodule regions may not achieve prominent performance in malignancy classification (e.g., x, x-resizing-64) without further consideration of contextual information from regions outside the nodule patch and the ability to model tissue heterogeneity, especially with small-quantity data.

5.4. Cross-evaluation and other integration methods of two datasets

In this subsection, we first investigate the effect of adopting a malignancy classifier on unsure-annotation data through cross-evaluation of two datasets and then study on other data integration methods. As can be seen in Table 10, this study mainly uses two model structures that apply RNet module and CNet module on ResNet backbone, respectively.

For the first structure with RNet, we conduct regression task based on two datasets and optimize the model by only using MSE loss. Given the fact that the ground truth of sure-annotation data has a high confidence level, we align the benign-malignant labels of sure-annotation data to the lowest-highest malignancy scores of unsure-annotation data. The model was separately trained on the unsure-annotation data only and on a mixed data consisting of both datasets.

The second structure is the same as model A that is designed for classification task. Thus, in this task, we first identified 5 scenarios by setting different division thresholds to assign binary labels for unsure-annotation data. The samples of unsure-annotation data nodules with different average malignancy scores are shown in Fig. 11.

Table 9
Performances of model A, model D and model D+ad-CSL trained with inputs with different patches (%).

Model	Patch	Sensitivity	Specificity	Precision	$Precision_b$	Accuracy	AUC	F1-score
Model A	64	64.24 ± 4.85	58.79 ± 8.70	61.28 ± 4.36	62.05 ± 3.92	61.52 ± 4.02	69.53 ± 3.39	62.54 ± 3.18
	x	66.06 ± 14.01	61.21 ± 14.65	63.69 ± 7.57	65.45 ± 7.16	63.64 ± 6.57	72.58 ± 2.64	64.00 ± 8.04
	x-resize-64	63.64 ± 11.18	56.97 ± 5.88	59.49 ± 2.23	61.88 ± 4.99	60.30 ± 3.24	65.25 ± 4.24	61.10 ± 6.21
	x-padding-64	69.70 ± 10.14	64.24 ± 7.02	66.19 ± 3.23	68.74 ± 5.87	66.97 ± 3.76	72.40 ± 3.11	67.52 ± 5.40
Model D	64	68.48 ± 8.48	69.70 ± 3.83	69.29 ± 1.98	69.35 ± 4.91	69.09 ± 3.26	76.51 ± 2.96	68.64 ± 4.94
	x	69.70 ± 9.58	61.21 ± 6.18	64.11 ± 5.68	67.36 ± 7.02	65.45 ± 6.17	73.13 ± 3.93	66.64 ± 7.16
	x-resize-64	69.09 ± 2.27	64.24 ± 7.52	66.23 ± 4.77	67.37 ± 2.67	66.67 ± 3.71	70.69 ± 5.21	67.53 ± 2.68
	x-padding-64	72.12 ± 9.47	63.03 ± 9.47	66.31 ± 5.73	69.73 ± 7.21	67.58 ± 6.25	72.97 ± 5.47	68.83 ± 6.47
Model D + ad-CSL	64	73.33 ± 4.85	66.67 ± 9.58	69.23 ± 5.16	71.39 ± 3.45	70.00 ± 4.33	76.01 ± 5.35	71.01 ± 3.39
	x	76.36 ± 6.47	57.58 ± 8.57	64.50 ± 5.05	71.01 ± 5.64	66.97 ± 5.20	72.21 ± 4.02	69.79 ± 4.78
	x-resize-64	69.70 ± 8.57	63.03 ± 11.08	66.05 ± 4.57	68.13 ± 4.24	66.36 ± 2.61	72.51 ± 3.88	67.30 ± 2.64
	x-padding-64	67.27 ± 9.85	66.06 ± 8.44	66.75 ± 4.15	67.51 ± 5.47	66.67 ± 3.95	72.47 ± 2.35	66.58 ± 5.23

Table 10
Performances of cross-evaluation between sure & unsure-annotation data and different integration methods of two datasets (%). "ft" denotes the mode of transfer learning that pre-training model using unsure-annotation data and fine-tuning model using sure-annotation data. The column "Scenario" represents the different label assignment methods (benign/malignant) for unsure-annotation data based on their average malignant scores (five-point scale).

Module	Scenario	Training data & mode	Testing data	Sensitivity	Specificity	Precision	$Precision_b$	Accuracy	AUC	F1-score
RNet		Unsure	Sure	95.15 ± 3.64	24.85 ± 4.45	55.90 ± 1.32	85.27 ± 10.21	60.00 ± 2.06	74.77 ± 6.06	70.39 ± 1.53
		Unsure+sure	Sure	73.94 ± 9.88	61.82 ± 12.21	66.58 ± 6.86	70.77 ± 7.51	67.88 ± 6.46	75.37 ± 5.75	69.59 ± 6.12
CNet	1/2345	Unsure	Unsure	96.49 ± 0.64	80.85 ± 7.37	97.16 ± 1.10	77.34 ± 2.46	94.47 ± 0.70	97.49 ± 1.22	96.82 ± 0.39
		Unsure	Sure	100.00 ± 0.00	1.21 ± 2.42	50.31 ± 0.63	100.00 ± 0.00	50.61 ± 1.21	60.75 ± 8.80	66.94 ± 0.55
		Unsure+sure	Sure	90.30 ± 8.44	13.94 ± 6.80	51.16 ± 1.95	64.79 ± 20.94	52.12 ± 3.40	62.26 ± 6.66	65.23 ± 3.44
		Unsure+sure(ft)	Sure	64.85 ± 8.04	67.27 ± 9.66	66.95 ± 4.71	65.88 ± 4.37	66.06 ± 4.13	72.87 ± 5.56	65.49 ± 4.45
	12/345	Unsure	Unsure	90.56 ± 5.80	52.92 ± 3.25	83.48 ± 0.80	71.18 ± 14.41	80.18 ± 3.65	80.97 ± 2.82	86.80 ± 2.78
		Unsure	Sure	100.00 ± 0.00	0.61 ± 1.21	50.15 ± 0.31	100.00 ± 0.00	50.30 ± 0.61	67.09 ± 6.14	66.80 ± 0.27
		Unsure+sure	Sure	84.24 ± 6.47	33.94 ± 13.61	56.47 ± 3.20	68.20 ± 1.97	59.09 ± 3.71	69.99 ± 2.95	67.33 ± 0.82
		Unsure+sure(ft)	Sure	67.88 ± 6.53	66.67 ± 12.71	67.92 ± 6.85	67.27 ± 4.40	67.27 ± 5.72	73.02 ± 6.63	67.52 ± 4.67
	12/45	Unsure	Unsure	76.15 ± 4.92	87.23 ± 6.88	84.19 ± 7.28	81.44 ± 2.78	82.17 ± 3.14	90.53 ± 1.73	79.65 ± 3.22
		Unsure	Sure	89.09 ± 5.28	31.52 ± 12.80	56.91 ± 3.75	75.38 ± 6.27	60.30 ± 4.43	70.74 ± 5.40	69.25 ± 2.01
		Unsure+sure	Sure	76.97 ± 1.48	57.58 ± 8.57	64.77 ± 4.63	71.05 ± 4.20	67.27 ± 4.66	73.90 ± 4.10	70.27 ± 3.13
		Unsure+sure(ft)	Sure	73.33 ± 5.55	60.00 ± 7.02	64.82 ± 5.24	69.22 ± 6.06	66.67 ± 5.50	74.75 ± 6.37	68.77 ± 5.03
	123/45	Unsure	Unsure	58.47 ± 4.78	95.42 ± 2.41	80.09 ± 8.71	88.40 ± 1.23	86.86 ± 2.41	87.48 ± 1.86	67.41 ± 5.43
		Unsure	Sure	80.00 ± 5.94	52.73 ± 9.51	63.14 ± 4.78	72.44 ± 6.50	66.36 ± 5.37	71.31 ± 6.94	70.42 ± 4.30
		Unsure+sure	Sure	69.09 ± 13.19	69.70 ± 13.28	70.51 ± 7.24	70.17 ± 6.35	69.39 ± 5.86	76.64 ± 5.27	68.80 ± 7.73
		Unsure+sure(ft)	Sure	67.88 ± 5.94	66.06 ± 8.44	67.02 ± 5.11	67.35 ± 3.73	66.97 ± 4.22	73.72 ± 6.61	67.24 ± 4.11
	1234/5	Unsure	Unsure	45.45 ± 11.76	97.13 ± 1.09	60.33 ± 9.36	95.01 ± 0.92	92.70 ± 1.26	93.54 ± 3.33	51.11 ± 10.23
		Unsure	Sure	32.73 ± 14.77	93.94 ± 3.32	85.94 ± 9.12	58.74 ± 4.97	63.33 ± 6.74	74.97 ± 6.77	45.13 ± 17.09
		Unsure+sure	Sure	59.39 ± 12.94	70.91 ± 10.43	67.51 ± 6.39	64.27 ± 6.61	65.15 ± 6.06	71.55 ± 7.25	62.42 ± 8.38
		Unsure+sure(ft)	Sure	69.70 ± 8.99	61.21 ± 14.26	65.02 ± 8.15	66.70 ± 7.42	65.45 ± 7.81	71.72 ± 7.99	66.88 ± 6.91

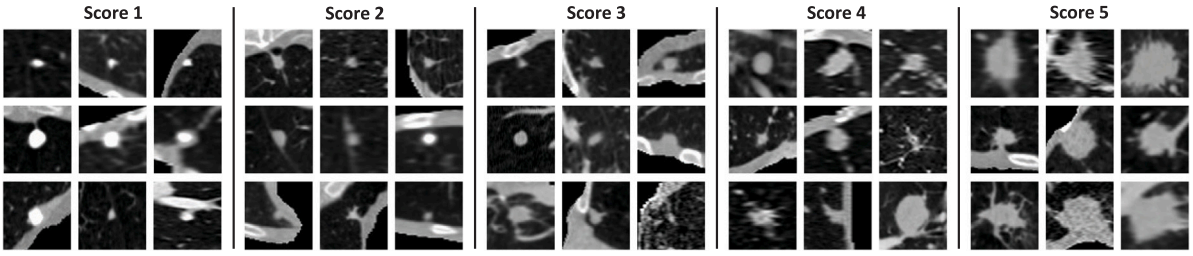

Fig. 11. Sample images of LIDC-IDRI (unsure-annotation data) nodules. Nodules with different average malignant scores are randomly selected after pre-processing and illustrated here.

In each scenario, we conducted four experiments: (1) model A was independently trained and tested on only unsure-annotation data with 5-fold cross-validation; (2) model A was trained on the whole unsure-annotation data and tested on sure-annotation data; (3) model A was trained on a merged two datasets. (4) model A was pre-trained using unsure-annotation data (50 epochs; learning rate: 1e-3) and fine-turned using sure-annotation data on the whole pre-trained model (50 epochs; learning rate: 1e-4).

From Table 10, observing the results of experiments that conducted both training and testing using unsure-annotation data only, we can find that good evaluation performance can be achieved within the unsure-annotation data domain, especially for scenario 1/2345 and scenario 12/45 which removed the uncertain score. According to the nodules samples shown in Fig. 11, most samples with average score 1 are calcified nodules, which have significant visible differences from other nodules with higher scores. The differences between nodules with scores 1&2 and those with scores 4&5 are distinct as well. As far as we know, scenario 12/45 was broadly used in many work for binary classification of LIDC-IDRI nodules. Compared between scenario 12/345 and scenario 123/45, better overall classification performance could be achieved by grouping the uncertain score into benign class, which is consistent with the experimental results in Han et al. (2015)

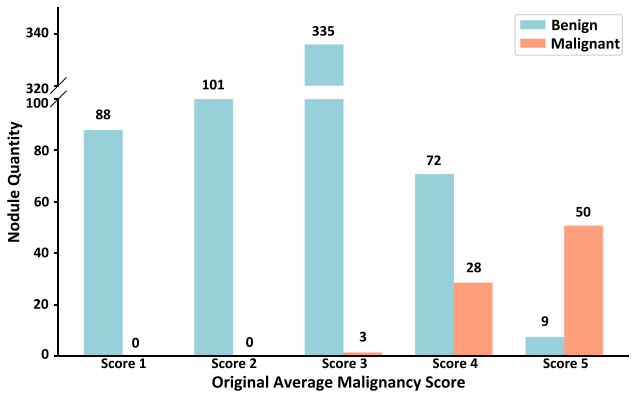

Fig. 12. Statistical result of LIDC-IDRI re-labeling nodules (benign or malignant) in terms of original average malignancy scores.

. This paper concluded that these uncertain nodules are more similar to benign ones. However, this inner-dataset experimental conclusion is still under suspicion as the ground truth for LIDC-IDRI is not available.

Once the scenario 12/45 is evaluated using sure-annotation data, model trained with unsure-annotation data tends to generate many false-positive predictions while most evaluation results fall sharply, especially for Specificity. Evidence reveals that monitored by sure testing data, false-positive problem could be alleviated successively when the division threshold moves from low-score side to high-score side, indicating the ordinal relationship and knowledge bias of unsure-annotation data malignancy score in the view of sure-annotation data. It can be noted from Fig. 8 that RNet tends to produce high regression scores when testing on sure-annotation data, demonstrating that our sure-annotation data are generally regarded as high likelihood of malignancy according to the radiologists' knowledge, which is another indication for the domain shifts between two datasets.

Mixed data learning and transfer learning (Zhang et al., 2020) are two effective methods for multi-data integration. As shown the results with the training mode of "Unsure + sure" that two datasets are fed into one task model simultaneously, improved performance could be achieved either for regression task or classification task. Moreover, transfer learning that gets the model fine-tuned could further correct the domain bias. However, these two methods will inevitably encounter the thorny problem of unsure-annotation data label assignment as well as knowledge waste of a large uncertain subset.

5.5. Re-labeling LIDC-IDRI

We use the new nodule diagnosis method in Section 3.3 for LIDC-IDRI nodule relabel. As shown in Fig. 12, our re-labeled results are in broad agreement with the low original malignancy score ones. In score 3, the majority of the nodules are re-labeled to benign class, which could explain the better performance when the nodules of score 3 are assigned to the benign label in Scenario 123/45 in Section 5.4. The new labels correct more than half of the original nodule labels with score 4 which could be the chief criminal leading to the data bias trouble because of the small suspicion on score 5 nodules. Moreover, due to the lack of pathological ground truth, the relabel outcomes of this study should always remain suspect until the LIDC-IDRI clinical information is available.

5.6. The reflections about the performances for unsure/sure-annotation data

We finally discuss an interesting phenomenon in this subsection: with the similar number of training samples (scenario 12/45 in Section 5.4), model trained and tested using the only unsure-annotation data show much better classification performance than using only sure-annotation data. The considerations are analyzed as follows:

Relative performance for unsure-annotation data:

Due to a lack of suitable reference to assess the likelihood of malignancy, low-level visible features (e.g. size, shape, brightness) are likely to be regarded as scoring criteria by radiologists' observation. Built on consensus agreement within multiple radiologists, apparent features of these nodules can be easily extracted and classified by a commonly used model, whose power can successfully emulate the radiologist's one.

Fig. 11 also indicates that unsure-annotation data nodules with the same average scores often share similar features, whereas sure-annotation data presents more heterogeneous characteristics for intra-class nodules. Thus, Lei et al. (2020a) is an effective scheme for LIDC-IDRI nodule classification by focusing on fine-grained features such as nodule shape and margins as its attention mechanism mimics radiologists' reasoning. In our method, we encode the fine-grained features into our model through SegNet and FNet. However, without the sure-annotation data, this model will always take the human ability as a golden standard rather than real malignancy labels. This motivated us to build the sure-annotation dataset.

Relative performance for sure-annotation data:

According to the good performance on unsure-annotation data when using the same classifier, we can rule out the hypothesis that the model may have limited discrimination capability for lung nodule heterogeneity modeling. However, no matter what input patch is used in Section 5.3, the model performance using sure-annotation data cannot come close to that using unsure-annotation data. These observations suggest a number of concerns about the current methods: (1) The use of CT imaging information alone may not be sufficient clinically for definite diagnosis. Although Ardila et al. (2019) and Venkadesh et al. (2021) have shown the possibility of deep learning systems to outperform radiologists in predicting malignancy using biopsy-proven labels, the potential of learning intrinsic correlation between CT (data from domain A) and pathological-proven standard (label from domain B) remains challenging. In addition, the prediction performance may be related to the data cohort (e.g., nodule quantity, nodule size, nodule type, cancer stage) used in each study. (2) In terms of benign-malignant nodule labeling, it may not be clinically sufficient if we subsume them under just two broad categories because there are some unspecified, borderline, or uncertain behavior as mentioned in the 2021 WHO lung tumor classification guide (Nicholson et al., 2022). It would therefore be more appropriate to first classify the major types of nodules according to their visibility in CT volumes and then further determine their tissue characteristics.

6. Conclusion and future work

In summary, we raised the vital issues that are commonly ignored in lung cancer prediction task from the aspect of unsure-annotation data and unreliable model reasoning. For better verifiability of nodule diagnosis algorithm and authenticity that simulates the real clinical world, we constructed a sure-annotation dataset with pathologically-confirmed labels. A synergic model was first proposed to integrate unsure-annotation data based on its properties and ultimately boost the classification performance of sure-annotation data. Then, our ad-CSL loss treats the CAM not only a post-hoc interpretation to analyze a nodule classification process, but also as a participant to modify the classification process in such a way that the model could pay more attention to the faithful nodule features and gain improved generalizability. Moreover, similar nodule retrieval empowers a CAD system more practical for clinical application during collaboration with clinicians. It is obvious that discriminating nodules of sure-annotation data is more difficult because they contain more complicated heterogeneity which is hard to model and there remains a critical dispute on CT-based manifestation of pathological diagnosis. Besides, data scarcity

still makes current nodule research a great challenge. In the future, it is imperative to enable the deep learning system to possess the capacity of causal inference for lung cancer prediction. We will explore the approaches for explainable decisions that lead to a more creditable and trustworthy diagnosis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Some data will be available for research purposes upon reasonable request.

Acknowledgments

This work was partly supported by Medicine-Engineering Interdisciplinary Research Foundation of Shanghai Jiao Tong University, China (YG2021QN128), Shanghai Sailing Program, China (20YF1420800), National Nature Science Foundation of China (No. 62003208), Shanghai Municipal of Science and Technology Project, China (Grant No. 20JC1419500), and Science and Technology Commission of Shanghai Municipality, China (Grant 20DZ2220400).

References

Ardila, D., Kiraly, A.P., Bharadwaj, S., Choi, B., Reicher, J.J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., et al., 2019. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat. Med. 25 (6), 954–961.

Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al., 2011. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Med. Phys. 38 (2), 915–931.

Armato III, S.G., McLennan, G., McNitt-Gray, M.F., Meyer, C.R., Yankelevitz, D., Aberle, D.R., Henschke, C.I., Hoffman, E.A., Kazerooni, E.A., MacMahon, H., et al., 2004. Lung image database consortium: developing a resource for the medical imaging research community. Radiology 232 (3), 739–748.

Armato III, S.G., Sensakovic, W.F., 2004. Automated lung segmentation for thoracic CT: impact on computer-aided diagnosis1. Academic Radiol. 11 (9), 1011–1021.

Carrazza, M., Kennedy, B., Rasin, A., Furst, J., Raicu, D., 2016. Investigating the effects of majority voting on CAD systems: a LIDC case study. In: Medical Imaging 2016: Computer-Aided Diagnosis. Vol. 9785, International Society for Optics and Photonics, 978533.

Chen, S., Qin, J., Ji, X., Lei, B., Wang, T., Ni, D., Cheng, J.-Z., 2016. Automatic scoring of multiple semantic attributes with multi-task feature leverage: A study on pulmonary nodules in CT images. IEEE Trans. Med. Imaging 36 (3), 802–814.

Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.-S., 2017. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5659–5667.

DeVries, T., Taylor, G.W., 2017. Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552.

Diaz, R., Marathe, A., 2019. Soft labels for ordinal regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4738–4747.

Goldstraw, P., Chansky, K., Crowley, J., Rami-Porta, R., Asamura, H., Eberhardt, W.E., Nicholson, A.G., Groome, P., Mitchell, A., Bolejack, V., et al., 2016. The IASLC lung cancer staging project: proposals for revision of the TNM stage groupings in the forthcoming (eighth) edition of the TNM classification for lung cancer. J. Thoracic Oncol. 11 (1), 39–51.

Gu, R., Wang, G., Song, T., Huang, R., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., Zhang, S., 2020. CA-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. IEEE Trans. Med. Imaging 40 (2), 699–711.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.-Z., 2019. XAI—Explainable artificial intelligence. Science Robotics 4 (37).

Hadsell, R., Chopra, S., LeCun, Y., 2006. Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). Vol. 2, IEEE, pp. 1735–1742.

Han, F., Wang, H., Zhang, G., Han, H., Song, B., Li, L., Moore, W., Lu, H., Zhao, H., Liang, Z., 2015. Texture feature analysis for computer-aided diagnosis on pulmonary nodules. J. Digit. Imaging 28 (1), 99–115.

Han, F., Zhang, G., Wang, H., Song, B., Lu, H., Zhao, D., Zhao, H., Liang, Z., 2013. A texture feature analysis for diagnosis of pulmonary nodules using LIDC-IDRI database. In: 2013 IEEE International Conference on Medical Imaging Physics and Engineering. IEEE, pp. 14–18.

Hansell, D.M., Bankier, A.A., MacMahon, H., McLoud, T.C., Muller, N.L., Remy, J., 2008. Fleischner society: glossary of terms for thoracic imaging. Radiology 246 (3), 697–722.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Hoffer, E., Ailon, N., 2015. Deep metric learning using triplet network. In: International Workshop on Similarity-Based Pattern Recognition. Springer, pp. 84–92.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132–7141.

Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K.Q., 2016. Deep networks with stochastic depth. In: European Conference on Computer Vision. Springer, pp. 646–661.

Hussein, S., Cao, K., Song, Q., Bagci, U., 2017. Risk stratification of lung nodules using 3D CNN-based multi-task learning. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 249–260.

Jacobs, C., van Ginneken, B., 2019. Google's lung cancer AI: a promising tool that needs further validation. Nat. Rev. Clin. Oncol. 16 (9), 532–533.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kirby, J.S., Armato, S.G., Drukker, K., Li, F., Hadjiiski, L., Tourassi, G.D., Clarke, L.P., Engelmann, R.M., Giger, M.L., Redmond, G., et al., 2016. LUNGx challenge for computerized lung nodule classification. J. Med. Imaging 3 (4), 044506.

Koch, G., Zemel, R., Salakhutdinov, R., 2015. Siamese neural networks for one-shot image recognition. In: ICML Deep Learning Workshop. Vol. 2, Lille.

Kuang, K., Cui, P., Athey, S., Xiong, R., Li, B., 2018. Stable prediction across unknown environments. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 1617–1626.

Kubota, T., Jerebko, A.K., Dewan, M., Salganicoff, M., Krishnan, A., 2011. Segmentation of pulmonary nodules of various densities with morphological approaches and convexity models. Med. Image Anal. 15 (1), 133–154.

Lei, Y., Shan, H., Zhang, J., 2021. Meta ordinal weighting net for improving lung nodule classification. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 1210–1214.

Lei, Y., Tian, Y., Shan, H., Zhang, J., Wang, G., Kalra, M.K., 2020a. Shape and margin-aware lung nodule classification in low-dose CT images via soft activation mapping. Med. Image Anal. 60, 101628.

Lei, Y., Zhu, H., Zhang, J., Shan, H., 2020b. Meta ordinal regression forest for learning with unsure lung nodules. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine. BIBM, IEEE, pp. 442–445.

Liao, F., Liang, M., Li, Z., Hu, X., Song, S., 2019. Evaluate the malignancy of pulmonary nodules using the 3-d deep leaky noisy-or network. IEEE Trans. Neural Netw. Learn. Syst. 30 (11), 3484–3495.

Liao, Z., Xie, Y., Hu, S., Xia, Y., 2021. Learning from ambiguous labels for lung nodule malignancy prediction. arXiv preprint arXiv:2104.11436.

Lin, M., Chen, Q., Yan, S., 2013. Network in network. arXiv preprint arXiv:1312.4400.

Liu, L., Dou, Q., Chen, H., Qin, J., Heng, P.-A., 2019. Multi-task deep model with margin ranking loss for lung nodule analysis. IEEE Trans. Med. Imaging 39 (3), 718–728.

Liu, M., Zhang, F., Sun, X., Yu, Y., Wang, Y., 2021. CA-net: Leveraging contextual features for lung cancer prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 23–32.

Loverdos, K., Fotiadis, A., Kontogianni, C., Iliopoulou, M., Gaga, M., 2019. Lung nodules: a comprehensive review on current approach and management. Ann. Thorac. Med. 14 (4), 226.

McNitt-Gray, M.F., Armato III, S.G., Meyer, C.R., Reeves, A.P., McLennan, G., Pais, R.C., Freymann, J., Brown, M.S., Engelmann, R.M., Bland, P.H., et al., 2007. The lung image database consortium (LIDC) data collection process for nodule detection and annotation. Academic Radiol. 14 (12), 1464–1474.

McWilliams, A., Tammemagi, M.C., Mayo, J.R., Roberts, H., Liu, G., Soghrati, K., Yasufuku, K., Martel, S., Laberge, F., Gingras, M., et al., 2013. Probability of cancer in pulmonary nodules detected on first screening CT. N. Engl. J. Med. 369 (10), 910–919.

National Lung Screening Trial Research Team, 2011a. The national lung screening trial: overview and study design. Radiology 258 (1), 243–253.

National Lung Screening Trial Research Team, 2011b. Reduced lung-cancer mortality with low-dose computed tomographic screening. N. Engl. J. Med. 365 (5), 395–409.

Nicholson, A.G., Tsao, M.S., Beasley, M.B., Borczuk, A.C., Brambilla, E., Cooper, W.A., Dacic, S., Jain, D., Kerr, K.M., Lantuejoul, S., et al., 2022. The 2021 WHO classification of lung tumors: impact of advances since 2015. J. Thorac. Oncol. 17 (3), 362–387.

Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. 9 (1), 62–66.

Ozdemir, O., Russell, R.L., Berlin, A.A., 2019. A 3D probabilistic deep learning system for detection and diagnosis of lung cancer using low-dose CT scans. IEEE Trans. Med. Imaging 39 (5), 1419–1429.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. 32, 8026–8037.

Qin, Y., Gu, Y., Zhang, H., Yang, J., Wang, L., Yao, F., Zhu, Y.-M., 2021. Relationship between pulmonary nodule malignancy and surrounding pleurae, airways and vessels: a quantitative study using the public LIDC-IDRI dataset. arXiv preprint arXiv:2106.12991.

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

Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R., 2019. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Vol. 11700, Springer Nature.

Setio, A.A.A., Ciompi, F., Litjens, G., Gerke, P., Jacobs, C., Van Riel, S.J., Wille, M.M.W., Naqibullah, M., Sánchez, C.I., Van Ginneken, B., 2016. Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks. IEEE Trans. Med. Imaging 35 (5), 1160–1169.

Setio, A.A.A., Traverso, A., De Bel, T., Berens, M.S., Van Den Bogaard, C., Cerello, P., Chen, H., Dou, Q., Fantacci, M.E., Geurts, B., et al., 2017. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. Med. Image Anal. 42, 1–13.

Shen, W., Zhou, M., Yang, F., Dong, D., Yang, C., Zang, Y., Tian, J., 2016. Learning from experts: Developing transferable deep features for patient-level lung cancer prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 124–131.

Shen, W., Zhou, M., Yang, F., Yang, C., Tian, J., 2015. Multi-scale convolutional neural networks for lung nodule classification. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 588–599.

Shen, W., Zhou, M., Yang, F., Yu, D., Dong, D., Yang, C., Zang, Y., Tian, J., 2017. Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. Pattern Recognit. 61, 663–673.

Siegel, R.L., Miller, K.D., Fuchs, H.E., Jemal, A., 2021. Cancer statistics, 2021. CA: Cancer J. Clin. 71 (1), 7–33.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.

Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F., 2021. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: Cancer J. Clin. 71 (3), 209–249.

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M., 2018. Learning to compare: Relation network for few-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1199–1208.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2818–2826.

Van Ginneken, B., Armato III, S.G., de Hoop, B., van Amelsvoort-van de Vorst, S., Duindam, T., Niemeijer, M., Murphy, K., Schilham, A., Retico, A., Fantacci, M.E., et al., 2010. Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the ANODE09 study. Med. Image Anal. 14 (6), 707–722.

Venkadesh, K.V., Setio, A.A., Schreuder, A., Scholten, E.T., Chung, K., Wille, M.M.W., Ginneken, B.v., Prokop, M., Jacobs, C., 2021. Deep learning for malignancy risk estimation of pulmonary nodules detected at low-dose screening CT.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al., 2016. Matching networks for one shot learning. Adv. Neural Inf. Process. Syst. 29.

Wang, C., Xiao, J., Han, Y., Yang, Q., Song, S., Huang, G., 2021. Towards learning spatially discriminative feature representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1326–1335.

Wu, Y., He, K., 2018. Group normalization. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 3–19.

Wu, B., Sun, X., Hu, L., Wang, Y., 2019. Learning with unsure data for medical image diagnosis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10590–10599.

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

Xie, Y., Xia, Y., Zhang, J., Feng, D.D., Fulham, M., Cai, W., 2017. Transferable multi-model ensemble for benign-malignant lung nodule classification on chest CT. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 656–664.

Xie, Y., Xia, Y., Zhang, J., Song, Y., Feng, D., Fulham, M., Cai, W., 2018a. Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest CT. IEEE Trans. Med. Imaging 38 (4), 991–1004.

Xie, Y., Zhang, J., Xia, Y., 2019. Semi-supervised adversarial model for benign–malignant lung nodule classification on chest CT. Med. Image Anal. 57, 237–248.

Xie, Y., Zhang, J., Xia, Y., Fulham, M., Zhang, Y., 2018b. Fusing texture, shape and deep model-learned information at decision level for automated classification of lung nodules on chest CT. Inf. Fusion 42, 102–110.

Xu, X., Wang, C., Guo, J., Gan, Y., Wang, J., Bai, H., Zhang, L., Li, W., Yi, Z., 2020. MSCS-deepln: Evaluating lung nodule malignancy using multi-scale cost-sensitive neural networks. Med. Image Anal. 65, 101772.

Yang, J., Deng, H., Huang, X., Ni, B., Xu, Y., 2020. Relational learning between multiple pulmonary nodules via deep set attention transformers. In: 2020 IEEE 17th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 1875–1878.

Yang, J., Fang, R., Ni, B., Li, Y., Xu, Y., Li, L., 2019. Probabilistic radiomics: ambiguous diagnosis with controllable shape analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 658–666.

Yitzhaki, S., et al., 2003. Gini's mean difference: A superior measure of variability for non-normal distributions. Metron 61 (2), 285–316.

Yu, F., Koltun, V., 2015. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122.

Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y., 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6023–6032.

Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O., 2021. Understanding deep learning (still) requires rethinking generalization. Commun. ACM 64 (3), 107–115.

Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D., 2017. Mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412.

Zhang, H., Gu, Y., Qin, Y., Yao, F., Yang, G.-Z., 2020. Learning with sure data for nodule-level lung cancer prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 570–578.

Zhang, J., Xie, Y., Xia, Y., Shen, C., 2019. Attention residual learning for skin lesion classification. IEEE Trans. Med. Imaging 38 (9), 2092–2103.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2921–2929.