Re-thinking and Re-labeling LIDC-IDRI for Robust Pulmonary Cancer Prediction

Hanxiao Zhang[1], Xiao Gu[2], Minghui Zhang[1], Weihao Yu[1], Liang Chen[3],
Zhexin Wang[3(✉)], Feng Yao[3], Yun Gu[1,4(✉)], and Guang-Zhong Yang[1(✉)]

[1] Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China
{geron762,gzyang}@sjtu.edu.cn
[2] Imperial College London, London, UK
[3] Department of Thoracic Surgery, Shanghai Chest Hospital,
Shanghai Jiao Tong University, Shanghai, China
wangzhexin001@hotmail.com
[4] Shanghai Center for Brain Science and Brain-Inspired Technology,
Shanghai, China

Abstract. The LIDC-IDRI database is the most popular benchmark for lung cancer prediction. However, with subjective assessment from radiologists, nodules in LIDC may have entirely different malignancy annotations from the pathological ground truth, introducing label assignment errors and subsequent supervision bias during training. The LIDC database thus requires more objective labels for learning-based cancer prediction. Based on an extra small dataset containing 180 nodules diagnosed by pathological examination, we propose to re-label LIDC data to mitigate the effect of original annotation bias verified on this robust benchmark. We demonstrate in this paper that providing new labels by similar nodule retrieval based on metric learning would be an effective re-labeling strategy. Training on these re-labeled LIDC nodules leads to improved model performance, which is enhanced when new labels of uncertain nodules are added. We further infer that re-labeling LIDC is current an expedient way for robust lung cancer prediction while building a large pathological-proven nodule database provides the long-term solution.

Keywords: Pulmonary nodule · Cancer prediction · Metric learning · Re-labeling

1 Introduction

The LIDC-IDRI (Lung Image Database Consortium and Image Database Resource Initiative) [1] is a leading source of public datasets. Since the introduction of LIDC, it is used extensively for lung nodule detection and cancer prediction using learning-based methods [4,6,11,12,15–17,21,23].

When searching papers in PubMed[1] with the following filter: *("deep learning" OR convolutional) AND (CT OR "computed tomography") AND (lung OR pulmonary) AND (nodule OR cancer OR "nodule malignancy") AND (prediction OR classification)*, among 53 papers assessed for eligibility of nodule malignancy classification, 40 papers used LIDC database, 5 papers used NSLT (National Lung Screening Trial) database[2] [10,18,19] (no exact nodule location provided), and 8 papers used other individual datasets. LIDC is therefore the most popular benchmark in cancer prediction research.

A careful examination of the LIDC database, however, reveals several potential issues for cancer prediction. During the annotation of LIDC, characteristics of nodules were assessed by multiple radiologists, where the rating of malignancy scores (1 to 5) was based on the assumption of a **60-year-old male smoker**. Due to the lack of clinical information, these malignancy scores were subjective. Although a subset of LIDC cases possesses patient-based pathological diagnosis [13], its nodule-level binary labels can not be confirmed.

Since it is hard to recapture the pathological ground truth for each LIDC nodule, we apply the extra SCH-LND dataset [24] with pathological-proven labels, which is used not only for establishing a truthful and fair evaluation benchmark but also for transferring pathological knowledge for different clinical indications.

In this paper, we first assess the nodule prediction performances of LIDC driven model in six scenarios and their fine-tuning effects using SCH-LND with detailed experiments. Having identified the problems of the undecided binary label assignment scheme on the original LIDC database and unstable transfer learning outcomes, we seek to re-label LIDC nodule classes by interacting with the SCH-LND. The first re-labeling strategy adopts the state-of-the-art nodule classifier as an end-to-end annotator, but it has no contribution to LIDC re-labeling. The second strategy uses metric learning to learn similarity and discrimination between the nodule pairs, which is then used to elect new LIDC labels based on the similarity ranking in a pairwise manner between the under-labeled LIDC nodule and each nodule of SCH-LND. Experiments show that the models trained with re-labeled LIDC data created by metric learning model not only resolve the bias problem of the original data but also transcend the performance of our model, especially when the new labels of the uncertain subset are added. Further statistical results demonstrate that the re-labeled LIDC data suffers class imbalance problem, which indicates us to build a larger nodule database with pathological-proven labels.

2   Materials

LIDC-IDRI Database: According to the practice in [14], we excluded CT scans with slice thickness larger than 3 mm and sampled nodules identified by at least three radiologists. We only involve solid nodules in SCH-LND and LIDC databases because giving accurate labels for solid nodules is of great challenge.

---

[1] https://pubmed.ncbi.nlm.nih.gov/.
[2] https://cdas.cancer.gov/datasets/nlst/.

Extra Dataset: The extra dataset called SCH-LND [24] consists of 180 solid nodules (90 benign/90 malignant) with exact spatial coordinates and radii. Each sample is very rare because all the nodules are confirmed and diagnosed by immediate pathological examination via biopsy with ethical approval.

To regulate variant CT formats, CT slice thickness is resampled to 1mm/pixel if it is larger than $1\,\mathrm{mm/pixel}$, while the X and Y axes are fixed to 512×512 pixels. Each pixel value is unified to the HU (Hounsfield Unit) value before nodule volume cropping.
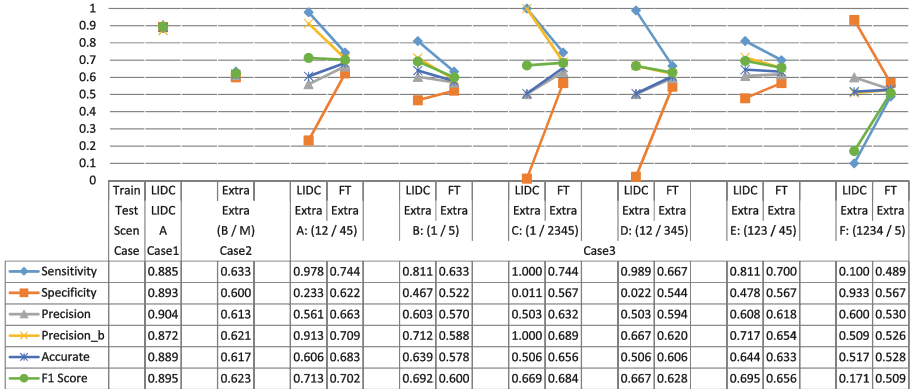
3   Study Design
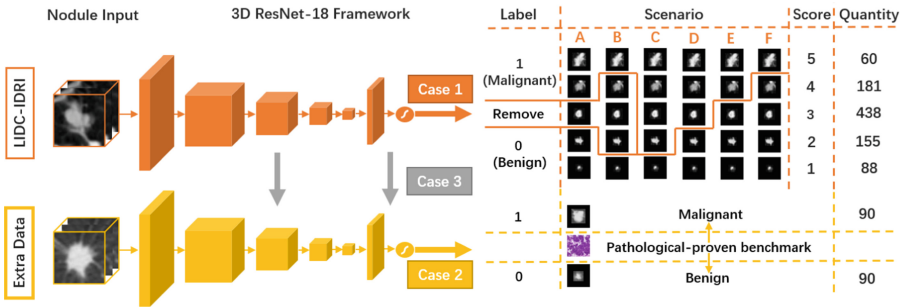
Fig. 1. Illustration of the study design for nodule cancer prediction. **Case 1:** training from scratch over the LIDC database after assigning nodule labels according to the average malignancy scores in 6 scenarios. **Case 2:** training over extra data based on accurate pathological-proven labels by 5-fold cross-validation. **Case 3:** testing or fine-tuning LIDC models of Case 1 using extra data.

The preliminary study follows the instructions of Fig. 1 where two types of cases (Case 1 and Case 2) conduct training and testing in each single data domain and one type of case (Case 3) involves domain interaction (cross-domain testing and transfer learning) between LIDC and SCH-LND. In Case 1 and Case 3, we identify 6 different scenarios by removing uncertain average scores (Scenarios A and B) or setting division threshold (Scenarios C, D, E, and F) to assign binary labels for LIDC data training. Training details are described in Sect. 5.1.

To evaluate the model performance comprehensively, we additionally introduce Specificity (also called Recall_b, when treating benign as positive sample) and $\mathrm{Precision}_b$ (Precision in benign class) [20], besides regular evaluation metrics including Sensitivity (Recall), Precision, Accuracy, and F1 score.

Based on the visual assessment of radiologists, human-defined nodule features can be easily extracted and classified by a commonly used model (3D ResNet-18 [5]), whose performance can emulate the experts' one (Fig. 2, Case 1). Many studies still put investigational efforts for better results across the LIDC board, overlooking inaccurate radiologists' estimations and bad model capability in the real world. However, once the same model is revalidated under

	Case1	Case2	A: (12/45)		B: (1/5)		C: (1/2345)		D: (12/345)		E: (123/45)		F: (1234/5)	
Train	LIDC	Extra	LIDC	FT	LIDC	FT	LIDC	FT	LIDC	FT	LIDC	FT	LIDC	FT
Test	LIDC	Extra	Extra	Extra	Extra	Extra	Extra	Extra	Extra	Extra	Extra	Extra	Extra	Extra
Scen	A	(B / M)						Case3						
Sensitivity	0.885	0.633	0.978	0.744	0.811	0.633	1.000	0.744	0.989	0.667	0.811	0.700	0.100	0.489
Specificity	0.893	0.600	0.233	0.622	0.467	0.522	0.011	0.567	0.022	0.544	0.478	0.567	0.933	0.567
Precision	0.904	0.613	0.561	0.663	0.603	0.570	0.503	0.632	0.503	0.594	0.608	0.618	0.600	0.530
Precision_b	0.872	0.621	0.913	0.709	0.712	0.588	1.000	0.689	0.667	0.620	0.717	0.654	0.509	0.526
Accurate	0.889	0.617	0.606	0.683	0.639	0.578	0.506	0.656	0.506	0.606	0.644	0.633	0.517	0.528
F1 Score	0.895	0.623	0.713	0.702	0.692	0.600	0.669	0.684	0.667	0.628	0.695	0.656	0.171	0.509

Fig. 2. Performance comparisons between different Cases or Scenarios (Scen) in Fig. 1. For instance, 'A:(12/45)' represents 'Scenario A' that treats LIDC scores 1 & 2 as benign labels and scores 4 & 5 as malignant labels. FT denotes fine-tuning using extra data by 5-fold cross-validation based on the pre-trained model in each scenario.

the pathological-proven benchmark (Fig. 2, Case 3, Scenario A), its drawback is objectively revealed that LIDC model decisions take up too many false-positive predictions. These two experimental outcomes raise a suspicion that whether the visual assessment of radiologists might have a bias toward malignant class.

To resolve this suspicion, we compare the performances of 6 scenarios in Case 3. Evidence reveals that, under the testing data from SCH-LND, the number of false-positive predictions has a declining trend when the division threshold moves from the benign side to the malignant side, but the bias problem is still serious when reaching Scenario E, much less of Scenario A and B. Besides, as training on the SCH-LND dataset from scratch can hardly obtain a high capacity model (Fig. 2, Case 2), we use transfer learning in Case 3 to get the model fine-tuned on the basis of weights of different pre-trained LIDC models.

Observing the inter-comparison within each scenario in Case 3, transfer learning can push scattered metric values close. However, compared with Case 2, the fine-tuning technique would bring both positive and negative transfer, depending upon the property of the pre-trained model.

Thus, either for training from scratch or transfer learning process, the radiologists' assessment of LIDC nodule malignancy can be hard to properly use. In addition to its inevitable assessment errors, there is a thorny problem to assign LIDC labels (how to set division threshold) and removing uncertain subset (waste of data). We thus expect to re-label the LIDC malignancy classes with the interaction of SCH-LND, to correct the assessment bias as well as utilize the uncertain nodules (average score = 3). Two independent approaches are described in the following section.

4   Methods

We put forward two re-labeling strategies to obtain new ground truth labels on the LIDC database. The first strategy generates the malignancy label from a machine annotator: the state-of-the-art nodule classifier that has been pre-trained on LIDC data and fine-tuned on SCH-LND to predict nodule class. The second strategy ranks the top nodules' labels using a machine comparator: a metric-based Network that measures the correlation between nodule pairs.

Considering that the knowledge from radiologists' assessments could be a useful resource, in each strategy, two modes of LIDC re-labeling are proposed. **For Mode 1 (Substitute):** LIDC completely accepts the re-label outcomes from other label machines. **For Mode 2 (Consensus):** The final LIDC re-label results would be decided by the consensus of label machine outcomes and its original label (Scenario A). In other words, this mode will leave behind the nodules with the same label and discard controversial ones, which may cause data reduction. We evaluate the LIDC re-labeling effect by using SCH-LND to test the model which is trained with re-labeled data from scratch.

4.1   Label Induction Using Machine Annotator

The optimized model with fine-tuning technique can correct the learning bias initiated by LIDC data. Some fine-tuned models even surpass the LIDC model performance in large scales of evaluation metrics. We wonder whether the current best performance model can help classify and annotate new LIDC labels. Experiments will be conducted using two annotation models from Case 2 and Case 3 (Scenario A) in Sect. 3.

4.2   Similar Nodule Retrieval Using Metric Learning
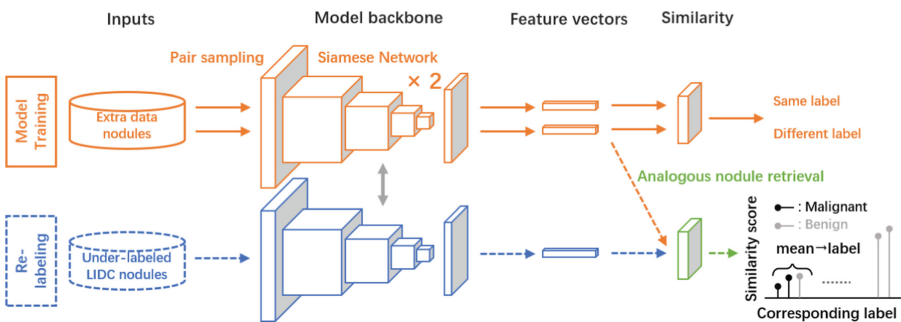
Fig. 3. The second strategy of LIDC re-labeling that using a metric learning model to search for the most similar nodules and give new labels.

Metric learning [2,7] provides a few-shot learning approach that aims to learn useful representations through distance comparisons. We use Siamese Network

[3,9] in this study which consists of two networks whose parameters are tied to each other. Parameter tying guarantees that two similar nodules will be mapped by their respective networks to adjacent locations in feature space.

For training a Siamese Network in Fig. 3, we pass the inputs in the set of pairs. Each pair is randomly chosen from SCH-LND and given the label whether two nodules of this pair are in the same class. Then these two nodule volumes are passed through the 3D ResNet-18 to generate a fixed-length feature vector individually. A reasonable hypothesis is given that: if the two nodules belong to the same class, their feature vectors should have a small distance metric; otherwise, their feature vectors will have a large distance metric. In order to distinguish between the same and different pairs of nodules when training, we apply contrastive loss over the Euclidean distance metric (similarity score) induced by the malignancy representation.

During re-labeling, we first pair each nodule from SCH-LND used in training up with an under-labeled LIDC nodule and sort each under-labeled nodule partner by their similarity scores. Then the new LIDC label is awarded by averaging the labels of the top 20% partner nodules in the ranking list of similarity scores.

5   Experiments and Results

5.1   Implementation

We apply 3D ResNet-18 [5] in this paper with adaptive average pooling (output size of $1 \times 1 \times 1$) following the final convolution layer. For the general cancer prediction model, we use a fully connected layer and a Sigmoid function to output the prediction score (binary cross-entropy loss). While for Siamese Network, we use a fully connected layer to generate the feature vector (8 neurons). Due to various nodule sizes, the batch size is set to 1, and group normalization [22] is adopted after each convolution layer.

All the experiments are implemented in PyTorch with a single NVIDIA GeForce GTX 1080 Ti GPU and learned using the Adam optimizer [8] with the learning rate of 1e–3 (100 epochs) and that of 1e–4 for fine-tuning in transfer learning (50 epochs). The validation set occupies 20% of the training set in each experiment. All the experiments and results involving or having involved the training of SCH-LND are strictly conducted by 5-fold cross-validation.

5.2   Quantitative Evaluation

To evaluate the first strategy using machine annotator, we first use Case 2 model to re-label LIDC nodules (a form of 5-fold cross-validation) other than the uncertain subset (original average score = 3). The re-labeled nodules are then fed into the 3D ResNet-18 model, which will be trained from scratch and tested on the corresponding subset of SCH-LND for evaluation. The result (4th row) shows

Table 1. Performances of different re-labeling methods based on each mode of re-labeling strategies. Under-labeled LIDC data are chosen by their original average score.

Row	Baselines	Method	Training	Testing	Sensitivity	Specificity	Precision	Precision$_b$	Accuracy	F1
1		Case 3-A	LIDC	Extra	0.9778	0.2333	0.5605	0.9130	0.6056	0.7126
2		Case 2	Extra	Extra	0.6333	0.6000	0.6129	0.6207	0.6167	0.6230
3		Siamese	Extra	Extra	0.6667	0.6000	0.6250	0.6429	0.6333	0.6452
	LIDC re-labeling									
	Strategy	Mode	Method	Under-label	Sensitivity	Specificity	Precision	Precision$_b$	Accuracy	F1
4	Annotator	Substitute	Case 2	1;2;4;5	0.5778	0.5667	0.5714	0.5730	0.5722	0.5746
5			Case 3-A		0.4630	0.6667	0.5814	0.5538	0.5648	0.5155
6		Consensus	Case 2		0.8778	0.3667	0.5809	0.7500	0.6222	0.6991
7			Case 3-A		0.8556	0.3778	0.5789	0.7234	0.6167	0.6906
8	Comparator	Substitute	Siamese	**1;2;4;5**	0.6111	0.6556	0.6395	0.6277	0.6333	0.6250
9				**1;2;3;4;5**	0.6778	**0.6667**	**0.6703**	0.6742	**0.6722**	**0.6740**
10		Consensus		1;2;4;5	0.8000	0.3778	0.5625	0.6538	0.5889	0.6606
11				1;2;3;4;5	0.7333	0.5889	0.6408	0.6883	0.6611	0.6839

that although this action greatly fixes label bias to a balanced state, this group of new labels can hardly build a model tested well on SCH-LND. Contrary to common sense, the state-of-the-art nodule classifier makes re-label performance worse (5th row), which is much lower than that of learning from scratch using SCH-LND (2nd row), indicating that the best model optimized with fine-tuning technique is not suitable for LIDC re-labeling. The initial two experiments adopting Mode 2 (Consensus) achieved better comprehensive outcomes than Mode 1 (Substitute) but with low Specificity (Table 1).

Metric learning takes a different re-label strategy that retrieves similar nodules according to the distance metric. Metric learning on a small dataset can obtain a better performance (3rd row) compared with general learning from scratch (2nd row). The re-label outcomes (8th and 9th row) also show great comprehensive improvement over baselines by Mode 1, where the re-labeling of uncertain nodules (average score = 3) is an important contributing factor.

Overall, there is a trade-off between Mode 1 and Mode 2. But Mode 2 seems to remain the LIDC bias property because testing results often have low Specificity and introduce data reduction. Re-labeling by consensus (Mode 2) may integrate the defects of both original labels and models, especially for malignant labels, while re-labeling uncertain nodules can help mitigate the defect of Mode 2.

We finally re-labeled the LIDC database with the Siamese Network trained using all of SCH-LND. As shown in Fig. 4, our re-labeled results are in broad agreement with the low malignancy score ones. In score 3 (uncertain data), the majority of the nodules are re-labeled to benign class, which explains the better performance when the nodules of score 3 are assigned to benign label in Scenario E (Fig. 2, Case 3). The new labels correct more than half of the original nodule labels with score 4 which could be the main reason leading to the data bias.

5.3   Discussion

Re-labeling through metric learning is distinct from the general supervised model in two notable ways. First, the input pairs generated by random sampling for
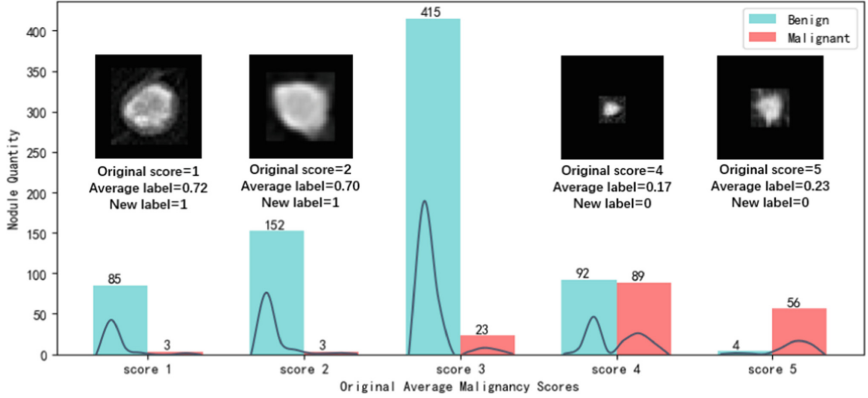
Fig. 4. Statistical result of LIDC re-labeling nodules (benign or malignant) in terms of original average malignancy scores, where the smooth curve describes the simplified frequency distribution histogram of average label outputs. For each average score of 1, 2, 4, and 5, one nodule re-labeling example with the opposite class (treat score 1 and 2 as benign; treat 4 and 5 as malignant) is provided.

metric learning provide a data augmentation effect to overcome overfitting with limited data. Second, under-labeled LIDC data take the average labels of top-ranked similarity nodules to increase the confidence of label propagation. These two points may explain why general supervised models (including fine-tuning models) perform worse than metric learning in re-labeling task. Unfortunately, after re-labeling, the class imbalance problem emerged (748 versus 174), while bringing up new limits in model training performance in the aforementioned experiments.

Moreover, due to the lack of pathological ground truth, the relabel outcomes of this study should always remain suspect until the LIDC clinical information is available. Considering a number of subsequent issues that LIDC may arise, sufficient evidence in this paper explores the motive for us to promote the ongoing collection work of a large pathological-proven nodule database, which is expected to become a powerful open-source database for the international medical imaging and clinical research community.

6   Conclusion and Future Work

The LIDC-IDRI database is currently the most popular public database of lung nodules with specific spatial coordinates and experts' annotations. However, because of the absence of clinical information, deep learning models trained based on this database have poor generalization capability in lung cancer prediction and downstream tasks. To challenge the low confidence labels of LIDC, an extra nodule dataset with pathological-proven labels was used to identify the annotation bias problems of LIDC and its label assignment difficulties. With the

robust supervision of SCH-LND, we used a metric learning-based approach to re-label LIDC data according to the similar nodule retrieval. The empirical results show that with re-labeled LIDC data, improved performance is achieved along with the maximization of LIDC data utilization and the subsequent class imbalance problem. These conclusions provide a guideline for further collection of a large pathological-proven nodule database, which is beneficial to the community.

References

1. Armato, S.G., III., et al.: The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Med. Phys. **38**(2), 915–931 (2011)
2. Bellet, A., Habrard, A., Sebban, M.: Metric Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, vol. 9, no. 1, pp. 1–151 (2015)
3. Guo, Q., Feng, W., Zhou, C., Huang, R., Wan, L., Wang, S.: Learning dynamic siamese network for visual object tracking. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1763–1771 (2017)
4. Han, F., et al.: Texture feature analysis for computer-aided diagnosis on pulmonary nodules. J. Digit. Imaging **28**(1), 99–115 (2015)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
6. Hussein, S., Cao, K., Song, Q., Bagci, U.: Risk stratification of lung nodules using 3D CNN-based multi-task learning. In: Niethammer, M., et al. (eds.) IPMI 2017. LNCS, vol. 10265, pp. 249–260. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59050-9_20
7. Kaya, M., Bilge, H.Ş: Deep metric learning: a survey. Symmetry **11**(9), 1066 (2019)
8. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
9. Koch, G., Zemel, R., Salakhutdinov, R.: Siamese neural networks for one-shot image recognition. In: ICML Deep Learning Workshop, vol. 2. Lille (2015)
10. Kramer, B.S., Berg, C.D., Aberle, D.R., Prorok, P.C.: Lung cancer screening with low-dose helical CT: results from the national lung screening trial (NLST) (2011)
11. Liao, Z., Xie, Y., Hu, S., Xia, Y.: Learning from ambiguous labels for lung nodule malignancy prediction. arXiv preprint arXiv:2104.11436 (2021)
12. Liu, L., Dou, Q., Chen, H., Qin, J., Heng, P.A.: Multi-task deep model with margin ranking loss for lung nodule analysis. IEEE Trans. Med. Imaging **39**(3), 718–728 (2019)
13. McNitt-Gray, M.F., et al.: The lung image database consortium (LIDC) data collection process for nodule detection and annotation. Acad. Radiol. **14**(12), 1464–1474 (2007)

14. Setio, A.A.A., et al.: 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 (2017)
15. Shen, W., et al.: Learning from experts: developing transferable deep features for patient-level lung cancer prediction. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 124–131. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_15
16. Shen, W., Zhou, M., Yang, F., Yang, C., Tian, J.: Multi-scale convolutional neural networks for lung nodule classification. In: Ourselin, S., Alexander, D.C., Westin, C.-F., Cardoso, M.J. (eds.) IPMI 2015. LNCS, vol. 9123, pp. 588–599. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19992-4_46
17. Shen, W., et al.: Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. Pattern Recogn. **61**, 663–673 (2017)
18. National Lung Screening Trial Research Team: The national lung screening trial: overview and study design. Radiology **258**(1), 243–253 (2011)
19. National Lung Screening Trial Research Team: Reduced lung-cancer mortality with low-dose computed tomographic screening. N. Engl. J. Med. **365**(5), 395–409 (2011)
20. Wu, B., Sun, X., Hu, L., Wang, Y.: Learning with unsure data for medical image diagnosis. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 10590–10599 (2019)
21. Wu, B., Zhou, Z., Wang, J., Wang, Y.: Joint learning for pulmonary nodule segmentation, attributes and malignancy prediction. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 1109–1113. IEEE (2018)
22. Wu, Y., He, K.: Group normalization. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
23. Xie, Y., et al.: Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest CT. IEEE Trans. Med. Imaging **38**(4), 991–1004 (2018)
24. Zhang, H., Gu, Y., Qin, Y., Yao, F., Yang, G.-Z.: Learning with sure data for nodule-level lung cancer prediction. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12266, pp. 570–578. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59725-2_55