Interpretable Lung Cancer Diagnosis with Nodule Attribute Guidance and Online Model Debugging

Hanxiao Zhang[1], Liang Chen[2], Minghui Zhang[1], Xiao Gu[3], Yulei Qin[4], Weihao Yu[1], Feng Yao[2], Zhexin Wang[2(✉)], Yun Gu[1,5(✉)], and Guang-Zhong Yang[1(✉)]

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

Abstract. Accurate nodule labeling and interpretable machine learning are important for lung cancer diagnosis. To circumvent the label ambiguity issue of commonly-used unsure nodule data such as LIDC-IDRI, we constructed a sure nodule data with gold-standard clinical diagnosis. To make the traditional CNN networks interpretable, we propose herewith a novel collaborative model to improve the trustworthiness of lung cancer predictions by self-regulation, which endows the model with the ability to provide explanations in meaningful terms to a human-observer. The proposed collaborative model transfers domain knowledge from unsure data to sure data and encodes a cause-and-effect logic based on nodule segmentation and attributes. Further, we construct a regularization strategy that treats the visual saliency maps (Grad-CAM) not only as post-hoc model interpretation, but also as a rational measure for trustworthy learning in such a way that the CNN features are extracted mainly from intrinsic nodule features. Moreover, similar nodule retrieval makes a nodule diagnosis system more understandable and credible to humans-observers based on the nodule attributes. We demonstrate that the combination of the collaborative model and regularization strategy can provide the best performances on lung cancer prediction and interpretable diagnosis that can automatically: 1) classify the nodule patches; 2) analyse and explain a prediction by nodule segmentation and attributes; and 3) retrieve similar nodules for comparison and diagnosis.

Keywords: Lung cancer · Computer-aided diagnosis · Interpretable AI.

1 Introduction

Today's AI systems for CT-based lung cancer diagnosis are highly desirable to gain the trust of clinicians with high-quality data labels and dependable interpre-

H. Zhang and L. Chen–Joint first authors of this work.

tations [6, 10, 16]. However, based on standard Convolutional Neural Networks (CNNs), most recent approaches [20, 26, 27, 24, 14] focus on statistical performance of nodule heterogeneity discrimination within a given nodule dataset LIDC-IDRI [2], instead of model interpretation and generalizability.

Normally, saliency maps [31][17] can retrospectively provide insight and interpret the prediction by highlighting where the model is looking at. However, this cannot explain its predictions in the same way as a human, who can classify objects based on a taxonomy of attributes. This inspired us to design a model which explains its predictions using a set of human-understandable terms. During the annotation of LIDC-IDRI [2][15], nine nodule attributes were assessed by multiple radiologists, which are Subtlety (Sub), Internal Structure (IS), Calcification (Cal), Sphericity (Sph), Margin (Mar), Lobulation (Lob), Spiculation (Spi), Texture (Tex), and Malignancy (Mal). Except for Internal Structure (6 categories) and Calcification (4 categories), each of the attributes is rated on a five-point scale and holds a degree relation (see Fig 1). Among these attributes, the rating of Malignancy is especially subjective due to the lack of pathologically-proven labels [2]. We term this kind of data as 'unsure(-annotation) data' by its nature of uncertainty. In addition, the outline of each nodule is delineated by multiple radiologists, providing the knowledge of nodule segmentation which, together with nodule attributes, can be considered as understandable concepts for experts to interpret model decisions and make evidence-based diagnoses. This also calls for the need of fair evaluation with a 'sure dataset' that has definite benign-malignant nodule annotations confirmed by pathological examination.

Moreover, saliency maps typically rely on human-experts to examine the corresponding results. By disclosing the salient information of a 'black-box' AI system using interpretable tools, one can intuitively observe some failure cases that the diagnosis model fails to assimilate reliable features from nodule regions (Section 4.3 and Fig 2). These underlying problems are mainly owing to the limitations of deep learning that its model often learns through superficial correlations for data fitting, especially with limited supervision (e.g. patch-level labels) [5]. Due to data scarcity, such circumstance is common yet easily overlooked in medical image analysis [19]. However, saliency maps cannot directly adjust the model if improper regions of attention are highlighted, leading to false and confounding correlations[28]. This encourages us to endow the model with the ability of self-regulation that automatically justifies the feature attention monitored by Grad-CAM [17]. To this end, we use a regularization strategy where Grad-CAM is regarded not only as a post-hoc interpretation, but also as a participant to debug model paired with the reference of nodule segmentation maps.

The feasibility of leveraging Grad-CAM to debug a model has three considerations: 1) it passes the sanity checks to highlighting attentions while many other saliency methods are similar to 'edge detectors' [1][4]; 2) it applies to a wide variety of CNNs for class-discriminative localization [17]; and 3) it is sensitive to the properties of the model parameters, which helps to update model [1].

Further, attribute-based nodule retrieval has the potential to improve the interpretability for lung cancer diagnosis, since it searches for nodules in historically collected data that share similar human-understandable features relative to the one being diagnosed. This mimics the clinical procedure, where clinicians make diagnoses based on their prior knowledge and experience indicated by nodule attributes and segmentation.

The main contribution of this work includes: 1) establishment of a collaborative model for lung cancer prediction guided by the knowledge of nodule segmentation and attributes; 2) introduction of model debugging with Grad-CAM to ensure trustworthiness during training and testing; and 3) provision of interpretable diagnoses for clinicians by attribute-based nodule retrieval.

2 Materials

Unsure dataset: According to the practice in [18], we excluded CT scans in LIDC-IDRI [2] with slice thickness larger than 3 mm and selected nodules identified by at least three radiologists. On top of that, we only involve 919 solid nodules (average Texture score = 5). In our work, we do not consider the learning and generating of Internal Structure and Calcification because the inner-classes of these two attributes are extremely imbalanced in this dataset [23]. Accordingly, except for Texture, our work performs the regression of the other six attributes whose average ratings hold sequential degrees. Each nodule segmentation map is generated according to a 50% consensus criterion [13].
Sure dataset: The sure dataset consists of 617 solid nodules (316 benign/301 malignant) collected from 588 patients' CT scans retrospectively in Shanghai Chest Hospital with ethical approval. CT scans in this dataset were acquired by multiple manufacturers where the slice thickness ranges from 0.50 to 3.00 mm with an average of 1.14 ($\pm$ 0.26) mm and the pixel spacing varied from 0.34 to 0.98 mm with an average of 0.60 ($\pm$ 0.22) mm. Each nodule was labeled to a definite class (benign or malignant) confirmed by pathological-proven examination by surgical resection. The exact spatial coordinate and radius of each nodule were annotated by two board-certified radiologists and confirmed by one senior radiologist. In this study, we only include the nodules with a diameter between 3 and 30 mm [3][7]. Note that although there exist some other sure data from NLST trial [21][22], Kaggle's 2017 Data Science Bowl (DSB) competition[1] and LUNGx Challenge dataset [12], we do not include these datasets in our study due to the lack of complete annotations such as position coordinates and pathologic diagnosis.

3 Methodology

3.1 Collaborative Model Architecture with Attribute-guidance

In our study, we train a collaborative model (Fig 1) to jointly conduct nodule segmentation and attribute regression tasks based on the annotation knowledge

[1] https://www.kaggle.com/c/data-science-bowl-2017/

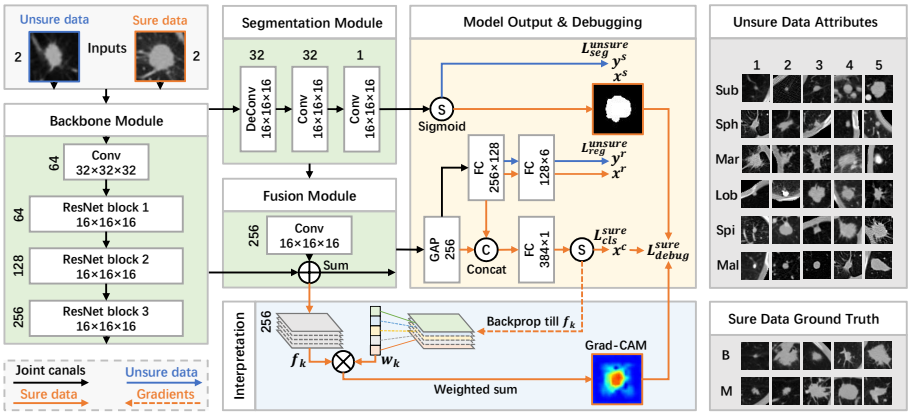

Fig. 1. The schematic illustration of the proposed collaborative model for joint learning with sure and unsure data. The basic modules (green bottom color) consist of three parts for feature extraction, nodule segmentation and feature fusion (follow the settings of [28]). In the next stage (yellow), model encodes interactive features for nodule attribute regression and classification, which are regulated with the rational measure of model interpretation (blue).

of unsure data and perform nodule benign-malignant classification learned from the nodule ground truth of sure data. The proposed collaborative model consists of a backbone for nodule feature extraction, a module for nodule segmentation, a fusion module that combines the features from backbone and segmentation head, and two interactive branches for nodule attribute regression and benign-malignancy classification.

The combined feature maps outputted by the fusion module are fed into the two branches for regression and classification tasks, which act in an interactive way to improve the discriminative ability for nodule prediction by exploring the correlation from attributes to benign-malignant classes. To this end, we first use a fully-connected (FC) layer to generate the intermediate embedding features, and apply another FC layer to output the six attribute scores, which are supervised by unsure data labels. For sure data classification, we first extract the attribute features from the first FC layer of the regression branch, and concatenate these features in the classification branch to make lung cancer prediction.

Different from other works [9][14], we treat the likelihood of Malignancy as a normal attribute rather than the outcome to determine whether a nodule is cancerous or not. This is mainly because: 1) the rating of Malignancy does not have a one-to-one connection with its binary benign-malignant label and retains an uncontrollable subjective bias [30][29]; 2) derived from the experts' knowledge, Malignancy reflects some observable nodule features such as size, shape and brightness; and 3) training six nodule attributes together can implicitly model the internal relationship between them. Such interactive architecture enables more guidance knowledge from nodule segmentation and attributes for sure data

to make a decision, although sure data do not have such detailed annotations. We formulate the loss function for the three aforementioned tasks as follows:

$$L_{tasks}^{(un)sure} = \underbrace{g^c log x^c + (1-g^c) log(1-x^c)}_{L_{cls}^{sure}} + \underbrace{1 - \frac{2\sum_i^N y_i^s g_i^s + \theta}{\sum_i^N y_i^s + \sum_i^N g_i^s + \theta}}_{L_{seg}^{unsure}} + \underbrace{\|y^r - g^r\|_2^2}_{L_{reg}^{unsure}} \quad (1)$$

in which, L_{cls}^{sure} is a binary cross-entropy (BCE) loss for the main classification task where x^c is the malignant probability after Sigmoid and g^c is the benign-malignant ground truth of sure data; L_{seg}^{unsure} is a Dice coefficient loss for the auxiliary segmentation task where y_i^s and g_i^s denote the predicted probability and class label of the i^{th} voxel, N is the number of voxels, and θ is a smoothing coefficient that prevents division by zero; L_{reg}^{unsure} is a mean square error (MSE) loss for the auxiliary attribute regression task where $y^r \in \mathbb{R}^{1\times n}$ is the regression output, $g^r \in \mathbb{R}^{1\times n}$ is the average attribute scores rated by radiologists, and n equals to 6 (sub, sph, mar, lob, spi and mal).

3.2 Debugging Model with Semantic Interpretation

To deal with the crisis of trustworthiness that happens in the reasoning process of a black-box model, we propose a controllable strategy to constrain the model to **diagnose** 'nodule' rather than arbitrary voxels in the sense of statistics. With the assistance of nodule segmentation map, Grad-CAM [17] is used to interpret and debug model online for trustworthy learning from nodule-relevant features.

Let $f_k(x, y, z)$ represents the unit k at 3D spatial location (x, y, z) of feature maps with length L, width W and height H outputted by the fusion module in Fig 1. To obtain the Grad-CAM, we first compute the gradients of the malignant probability x^c with respect to the feature map f_k, $\frac{\partial x^c}{\partial f_k}$. Then, the gradients are global-average-pooled to generate the neuron weights:

$$\omega_k = \frac{1}{L \times W \times H} \sum_x \sum_y \sum_z \frac{\partial x^c}{\partial f_k(x, y, z)} \quad (2)$$

Afterwards, due to using Sigmoid instead of Softmax, we perform a weighted sum of the feature maps f_k to obtain the Grad-CAM map with respect to x^c (benign: $x^c < 0.5$; malignant: $x^c \geqslant 0.5$):

$$Grad\text{-}CAM(x, y, z) = (x^c - 0.5) \sum_k \omega_k f_k(x, y, z) \quad (3)$$

which is then rescaled to [0,1] by min-max normalization.

To enable trustworthy learning, we regulate the Grad-CAM to concentrate attention on the nodule regions. Guided by the online generated nodule segmentation map, the average Grad-CAM values of nodule regions and background regions can be calculated, which are $Grad\text{-}CAM_{ndl}^{avg} \in [0, 1]$, and $Grad\text{-}CAM_{bkg}^{avg} \in$

$[0, 1]$. To drive the model to express the features of target object, we enforce $Grad\text{-}CAM_{ndl}^{avg}$ **larger than** $Grad\text{-}CAM_{bkg}^{avg}$, which is formulated as follows:

$$L_{debug}^{sure} = \|x^c - 0.5\|_{l_1} \, max\left\{0, Grad\text{-}CAM_{bkg}^{avg} - Grad\text{-}CAM_{ndl}^{avg} + \lambda\right\} \quad (4)$$

where λ is a margin parameter (empirically set to 0.5 in this work) and $\|x^c - 0.5\|_{l_1}$ is an adaptive coefficient that encodes the uncertainty of x^c so that model can strengthen the optimization for other tasks if a nodule prediction is of low confidence. In our practical application, we merged the item of $(x^c - 0.5)$ in Eq.(3) and Eq.(4), and made a simplification.

3.3 Explanation by Attribute-based Nodule Retrieval

To enable the interpretable lung cancer diagnosis, we can provide explainability through attribute-based nodule retrieval. Based on the nodule attribute scores $x^r \in \mathbb{R}^{1\times 6}$ generated by the collaborative model, we can retrieve K most similar nodules within the historically collected data for the one being diagnosed. The similarity metric used for retrieval is Euclidean Distance. By reading these closely related historical nodule cases, clinicians can acquire more understandable evidence and clues. Meanwhile, the auxiliary attribute scores work as assist-proofs for the diagnosis results and support the user's final decision.

4 Experiments and Results

4.1 Implementation

In data preprocessing, we first conduct lung segmentation to restrict the valid nodule regions inside the lungs. Then, inspired by the fact that radiologists change CT window widths and centers for nodule diagnosis, we mix lung window [-1000, 400 HU] and mediastinal window [-160, 240 HU] together to generate the nodule inputs. Each window is normalized to the range of [0, 1] and resampled to 0.5 mm/voxel along all three axes using spline interpolation. The final image volume extracted for each nodule is a cube of $64 \times 64 \times 64$ voxels with 2 channels. Data augmentation methods include random flipping, rotation and transposing.

All the experiments are implemented in PyTorch with a single NVIDIA GeForce GTX 1080 Ti GPU and learned using Adam optimizer [11] with the learning rate of 1e−3 (100 epochs). The batch size is set to 1 and group normalization [25] is used after each convolution operation. 5-fold cross-validation is performed, with 20% of the training set used for validation and early stopping.

4.2 Quantitative Evaluation

To provide the detailed evaluation of the model performance, we used evaluation metrics including Accuracy, AUC, F1-score, Sensitivity, Specificity, Precision,

Table 1. Quantitative classification performance of comparison methods and ablation study evaluated with sure data by 5-fold cross-validation (threshold=0.5).

		Method	Accuracy	AUC	F1-score	Sensitivity	Specificity	Precision	Precision_b
Baselines	1	3D ResNet[8]	64.03	73.26	63.09	64.05	64.00	63.63	65.97
	2	Transfer learning[30]	67.26	73.99	65.62	64.09	70.27	67.90	67.31
Ablation	3	-	67.29	76.28	65.85	64.40	70.03	69.00	67.24
	4	attr	67.28	77.32	66.03	65.40	69.06	68.56	67.94
	5	debug	69.39	76.16	67.92	66.44	**72.19**	69.77	69.30
	6	attr+debug	69.20	76.57	68.86	70.74	67.74	67.98	71.49
	7	mal+debug(CAM[31])	68.24	77.29	67.22	67.39	69.04	68.69	69.56
	8	attr+concat	69.70	76.89	69.16	69.72	69.67	69.63	71.08
	9	**attr+concat+debug**	**71.16**	**77.85**	**71.19**	**72.73**	69.67	**70.31**	**72.88**

and Precision_b (Precision in benign class). The results summarized in Table 1 illustrate the performance of nodule benign-malignancy classification tested on sure data in fair comparison with a normally-used 3D ResNet [8] and a state-of-the-art method [30] which also integrates the knowledge of unsure data. The results show that our best model (the last row) has the ability to predict lung cancer far better than the two baselines, especially for Accuracy, F1-score and Sensitivity. To analyze the impact of each component of our proposed method, we conducted ablation studies in the phase of 'Model Output & Debugging' in Fig 1 for: (3) only with basic modules; (4) only adding attribute regression (FC: 256×6); (5) only adding model debugging; (6) without attribute feature concatenation; (7) only adding one attribute ('malignancy', which is the most popular one) for regression and applying CAM [31] for debugging [28]; and (8) without model debugging with Grad-CAM. This shows retaining the single attribute regression or model debugging can barely exceed the performance of 3D ResNet, Transfer learning and the model only with basic modules. The integration of feature concatenation and model debugging plays an important role in improving the performance of nodule benign-malignant discrimination and have a positive effect on reducing overfitting.

4.3 Trustworthiness Check and Interpretable Diagnosis

Trustworthiness Check: Given the fact that there is no guarantee for a black-box model to learn nodule-relevant features with respect to model outputs, it is necessary for the human-experts to examine its trustworthiness before considering whether to adopt the model decisions. As illustrated in Fig 2, the saliency maps (Grad-CAM) of the 1^{st} and 2^{nd} rows present inexplicable patterns scattered in nodule patches. This implies that 3D ResNet and Transfer learning methods fail our trustworthiness check and can be misleading in real clinical practice. Compared with the 7^{th} method (the 3^{rd} row), our best method (the 4^{th} row) not only appears more effective constraint to extract reliable features in nodule regions, but also has a better quality of nodule segmentation (yellow outline) with a light-weight segmentation module. This benefits from the multi-attribute guidance for nodule discrimination and the superiority of Grad-CAM for model debugging, with their weights being updated by better achieving both

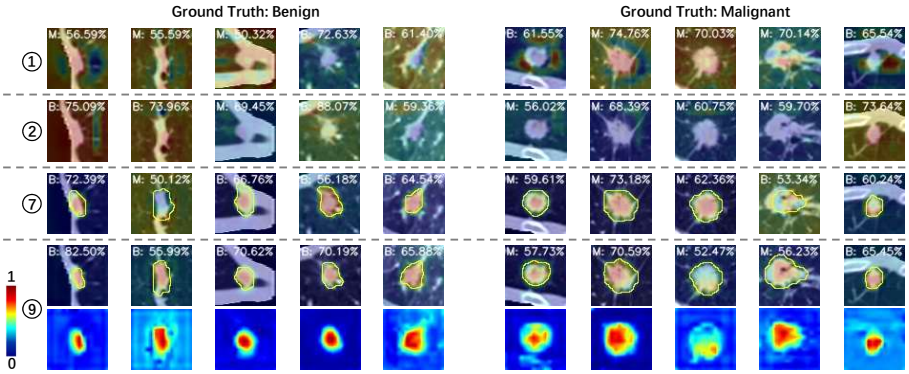

Fig. 2. Examples of saliency maps obtained by methods from Table 1. Examples are taken from the central slices of their 3D patches, where the scores are the predicted probabilities to each class and yellow contours denote the nodule segmentation outlines.

nodule segmentation and classification performance. Note that, our method does not completely inhibit feature learning from nodule background according to the last row.

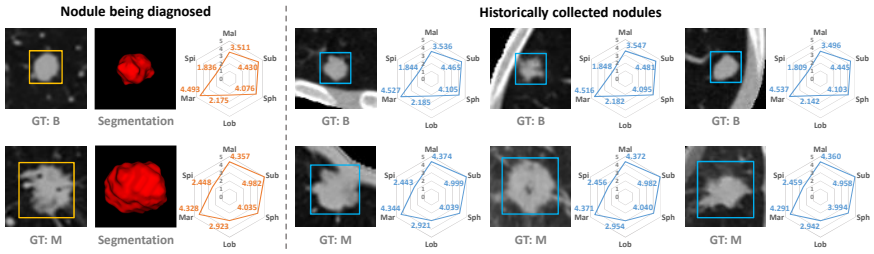

Fig. 3. Examples of the attribute-based retrieval for similar nodules (top 3, right part), with respect to the nodule being diagnosed (left part). Attribute scores and 3D segmentation maps are generated by the pre-trained model.

Interpretable Diagnosis: Fig 3 shows the examples of attribute-based nodule retrieval using our best model. For the nodule being diagnosed, our system, working as the role of an explainer, can generate its segmentation map and attribute scores, based on which, historically collected nodules with the most similar characteristics can be automatically recalled to support the clinicians to make confident diagnoses.

5 Conclusions

Under the fair evaluation of sure data, this paper introduced a new formulation to improve the performance of nodule classification, as well as enhance the trustworthiness of model reasoning and explainability for lung cancer diagnosis. Our superiority mainly comes from the effective cooperation of unsure and sure data knowledge and regulative application of model online debugging with semantic interpretation (Grad-CAM). These innovations empower a diagnosis system more credible and practical during collaboration with clinicians. We believe our formulation can be applied to other classification tasks, where the object segmentation (hand-crafted or automatic) and fine-grained attributes are available to provide regulation for interpretable learning and understandable diagnosis.

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

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. Advances in neural information processing systems **31** (2018)
2. 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.: The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical physics **38**(2), 915–931 (2011)
3. 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.: Lung image database consortium: developing a resource for the medical imaging research community. Radiology **232**(3), 739–748 (2004)
4. Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., Hoebel, K., Gupta, S., Patel, J., Gidwani, M., et al.: Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. Radiology: Artificial Intelligence **3**(6), e200267 (2021)
5. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)
6. Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.Z.: Xai—explainable artificial intelligence. Science Robotics **4**(37) (2019)
7. Hansell, D.M., Bankier, A.A., MacMahon, H., McLoud, T.C., Muller, N.L., Remy, J.: Fleischner society: glossary of terms for thoracic imaging. Radiology **246**(3), 697–722 (2008)

8. 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)
9. Hussein, S., Cao, K., Song, Q., Bagci, U.: Risk stratification of lung nodules using 3d cnn-based multi-task learning. In: International conference on information processing in medical imaging. pp. 249–260. Springer (2017)
10. Jacobs, C., van Ginneken, B.: Google's lung cancer ai: a promising tool that needs further validation. Nature Reviews Clinical Oncology **16**(9), 532–533 (2019)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
12. 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.: Lungx challenge for computerized lung nodule classification. Journal of Medical Imaging **3**(4), 044506 (2016)
13. Kubota, T., Jerebko, A.K., Dewan, M., Salganicoff, M., Krishnan, A.: Segmentation of pulmonary nodules of various densities with morphological approaches and convexity models. Medical Image Analysis **15**(1), 133–154 (2011)
14. Liu, L., Dou, Q., Chen, H., Qin, J., Heng, P.A.: Multi-task deep model with margin ranking loss for lung nodule analysis. IEEE transactions on medical imaging **39**(3), 718–728 (2019)
15. 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.: The lung image database consortium (lidc) data collection process for nodule detection and annotation. Academic radiology **14**(12), 1464–1474 (2007)
16. Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.R.: Explainable AI: interpreting, explaining and visualizing deep learning, vol. 11700. Springer Nature (2019)
17. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
18. 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.: Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. Medical image analysis **42**, 1–13 (2017)
19. Shad, R., Cunningham, J.P., Ashley, E.A., Langlotz, C.P., Hiesinger, W.: Designing clinically translatable artificial intelligence systems for high-dimensional medical imaging. Nature Machine Intelligence **3**(11), 929–935 (2021)
20. Shen, W., Zhou, M., Yang, F., Yu, D., Dong, D., Yang, C., Zang, Y., Tian, J.: Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. Pattern Recognition **61**, 663–673 (2017)
21. Team, N.L.S.T.R.: The national lung screening trial: overview and study design. Radiology **258**(1), 243–253 (2011)
22. Team, N.L.S.T.R.: Reduced lung-cancer mortality with low-dose computed tomographic screening. New England Journal of Medicine **365**(5), 395–409 (2011)
23. Wang, Q., Zhou, X., Wang, C., Liu, Z., Huang, J., Zhou, Y., Li, C., Zhuang, H., Cheng, J.Z.: Wgan-based synthetic minority over-sampling technique: Improving semantic fine-grained classification for lung nodules in ct images. IEEE Access **7**, 18450–18463 (2019)

24. 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)
25. Wu, Y., He, K.: Group normalization. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 3–19 (2018)
26. Xie, Y., Xia, Y., Zhang, J., Feng, D.D., Fulham, M., Cai, W.: Transferable multi-model ensemble for benign-malignant lung nodule classification on chest ct. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 656–664. Springer (2017)
27. Xie, Y., Xia, Y., Zhang, J., Song, Y., Feng, D., Fulham, M., Cai, W.: Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest ct. IEEE transactions on medical imaging **38**(4), 991–1004 (2018)
28. Zhang, H., Chen, L., Gu, X., Zhang, M., Qin, Y., Yao, F., Wang, Z., Gu, Y., Yang, G.Z.: Faithful learning with sure data for lung nodule diagnosis. arXiv preprint arXiv:2202.12515 (2022)
29. Zhang, H., Gu, X., Zhang, M., Weihao, Y., Chen, L., Wang, Z., Yao, F., Gu, Y., Yang, G.Z.: Re-thinking and re-labeling lidc-idri for robust pulmonary cancer prediction. arXiv preprint arXiv:2207.14238 (2022)
30. Zhang, H., Gu, Y., Qin, Y., Yao, F., Yang, G.Z.: Learning with sure data for nodule-level lung cancer prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 570–578. Springer (2020)
31. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)