Highly Robust Prediction of Lung Nodule Malignancy by Deep Learning Model: A Multiracial, Multinational Study

Hao Wu[*#], Wen Tang [*], Chu Wu, Yufeng Deng, and Rongguo Zhang

Infervision Medical Technology Co., Ltd.
[*] These two authors have equal contribution to this work.
[#] Corresponding author: whao@infervision.com

Abstract

Purpose
Although statistical models have been employed to detect and classify lung nodules using deep learning-extracted and clinical features, there is a lack of model validation in independent, multinational datasets from computed tomography (CT) scans and patient clinical information. To this end, we developed a deep learning-based algorithm to predict the malignancy of pulmonary nodules and validated its performance in three independent datasets containing multiracial and multinational populations.

Methods
In this study, a convolutional neural network-based algorithm to predict lung nodule malignancy was built based on CT scans and patient-wise clinical features (i.e. sex, spiculation, and nodule location). The model consists of three steps: (1) a deep learning algorithm to automatically extract features from CT scans, (2) clinical features were concatenated with the nodule features after dimension reduction by the principal component analysis (PCA), and (3) a multivariate logistic regression model was employed to classify the malignancy of the lung nodules. The model was trained by a dataset containing 1,556 nodules from 813 patients from the National Lung Screening Trial (NLST). The performance of the model was evaluated on three independent, multi-institutional datasets LIDC and Infervision Multi-Center (IMC) dataset, which contains 562 nodules from 293 patients, and 2044 nodules from 589 patients, respectively. The model accuracy was measured by the area under curve (AUC) of receiver operating characteristic (ROC) analysis.

Results
The study shows that the AUCs of ROCs on the NLST dataset, LIDC dataset, and IMC dataset are 0.91, 0.86, and 0.95, respectively. The inclusion of clinical features does not significantly improve the model performance. Quantitatively, the summed-up weight on the prediction accuracy of the 10 nodule features extracted by the deep learning algorithm equals to 0.091, while the weight of patient sex, nodule spiculation, and location is 0.031, 0.052, and 0.008, respectively.

Conclusion
The convolutional neural network-based model for lung nodule classification could be generalized to multiple datasets containing diverse populations. The addition of three patient clinical features to the nodule features extracted by deep learning does not boost the performance of the model.

Keywords: lung cancer diagnosis, pulmonary nodule, deep learning, computed tomography

Introduction

Over 12 million US adults experience a diagnostic error every year in outpatient settings [1]. Potentially 80,000 preventable deaths in US hospitals associated with failures in diagnosis every year. In addition, diagnostic errors lead to other serious harms or permanent disability [2]. Therefore, it is important to minimize diagnostic errors.

Lung cancer is the most common cancer in both men and women in the U.S. Lung cancer screening with low-dose computed tomography (LDCT) has been shown to reduce mortality by 20–43% and is included in US screening guidelines [5-7]. The Lung-RADS guidelines were published by the American College of Radiology (ACR), where Lung-RADS provide the reporting guidelines for LDCT lung cancer screening, in order to standardize image interpretation by radiologists and management recommendations. The Lung-RADS guideline is based on a variety of image findings, but primarily nodule size and growth [7].

Having a high nodule malignancy prediction is very important because of the high clinical and financial costs of missed diagnosis, late diagnosis and unnecessary biopsy procedures resulting from false negatives and false positives [8-9]. The diagnostic error of lung cancer leads to "high severity harms" and "malpractice claims". The total payments summed to more than a hundred billion dollars from 2006-2015 for lung cancer cases [2]. Persistent inter-reader variability and incomplete characterization of comprehensive imaging findings remain as limitations of the lung cancer screening program [10-11]. Clinical judgement failures, or the failure of recognition, is the second leading cause of diagnostic errors [3]. Lung cancers can be missed when sizes of lesions are small and hard to see [4]. Increased workload, physician fatigue typically exacerbates the observer error. The limitations of the lung cancer screening program and the detection errors suggest opportunities for more computer systems to aid the radiologists to improve the detection performance and improve the inter-reader consistency.

Prior studies have reported that deep learning-based algorithms have shown excellent performance on the nodule malignancy classification. For example, Causey et al. [12] used the LIDC/IDRI cohort to train a sophisticated CNN classification model (i.e. NoduleX) and achieved high accuracy for nodule malignancy classification, with an AUC of 0.99. Although with excellent prediction, NoduleX was trained and tested on the same database that has a relatively small number of CT images (thousands) and nodules. Furthermore, as they pointed out, their model may not be generalizable to other image datasets with different CT scan image quality or ground truthing methods. These drawbacks will limit its adoption and application in real scenarios.

Google AI's end-to-end approach has presented promising lung cancer detection results using deep learning [13]. They conducted clinical trials with a level similar to, or better than, that of radiologists. Although the results are promising, their model lacks validation from external databases and code availability. Moreover, radiologists indicated that Google AI's claims were two strong, and the recommendations from black-box nature of a proprietary AI system was likely to be accepted by medical professionals.

Previous statistical predictive tools based on patient and nodule characteristics have achieved high predictive accuracy of lung nodule malignancy. McWilliams et al. [14] developed multivariable logistic regression models with predictors including age, sex, family history of lung cancer, emphysema, nodule size, nodule position, and nodule type, using subjects from the Pan-Canadian Early Detection of Lung Cancer Study (PanCan) and the British Columbia Cancer Agency (BCCA). They found that out of all the clinical characteristics, patient age, nodule position, and nodule type play significant weights in the model's predictability. Use of this model has been recommended for pulmonary nodule risk estimation in lung cancer screening, although a later study reported that experienced and trainee radiologists had superior ability for lung cancer prediction compared with the multivariate model.

In this study, we aim to evaluate the prediction accuracy of lung nodule malignancy by deep learning model and validate its robustness on multiracial, multinational datasets. We also compared the model performance with *deep learning features* extracted from CT images only and *combined deep learning feature and clinical features*.

Materials and Methods

NLST dataset

The training cohort is consisted of the data from participants of the NLST trial. All positive cases are confirmed by biopsy in the trial. Negative cases are randomly selected from the NLST participants who are not diagnosed lung cancer in the time period of the trial. We screened out 624 sets of data that were diagnosed with lung cancer based on the report, including 633 malignant nodules, and 189 sets of data that were not diagnosed with lung cancer, including 923 non-malignant nodules.

LIDC dataset

The Lung Image Database Consortium (LIDC) database consists of diagnostic and lung cancer screening thoracic CT scans with annotated lesions. It contains 1018 subjects collected from 7 academic and 8 medical imaging companies. Each CT scan was independently reviewed by four experienced thoracic radiologists and was marked with lesions in three categories ("nodule >= 3 mm", "nodule < 3mm", and "no-nodule >= 3mm"). Each nodule in the "nodule >= 3 mm" class was then given a malignancy score and a detailed segmentation. The malignancy scores were defined as follows: 1 "Highly Unlikely for Cancer", 2 "Moderately Unlikely for Cancer", 3 "Indeterminate Likelihood", 4 "Moderately Suspicious for Cancer", 5 "Highly Suspicious for Cancer". Nodules containing the label of 1 "High Unlikely for Cancer" are considered benign (negative cases) and 5 "Highly Suspicious for Cancer" are considered malignant (positive cases).

We select nodules with label 1 "Highly Unlikely for Cancer" and label 5 "Highly Suspicious for Cancer" as Benign and Suspicious cases as the gold standard, respectively. Based on the report, we screened 106 sets of data that were diagnosed as lung cancer, including 239 malignant nodules, and 87 sets of data that were not diagnosed with lung cancer, including 323 non-malignant nodules.

Infervision Multi-Center Database

Infervision multi-center dataset consists of CT scans from 6 hospitals at different geographic areas in China, including Dalian Zhongshan Hospital, Affiliated Hospital of Shaanxi University of Traditional Chinese Medicine, Jiangsu University Affiliated Hospital, Fujian Medical University Affiliated Union Hospital, Wuhan Tongji Hospital, and Shengjing Hospital. Patients with lung cancer were confirmed by surgical biopsies. All the data and reports have been desensitized. In this dataset, we screened 502 sets of data diagnosed as lung cancer based on the report, including 300 malignant nodules, and 87 sets of data from undiagnosed lung cancer, including 1744 non-malignant nodules. And all of those nodules have been marked by radiologists according to the report, in order to unify the form with the other two dataset annotations.

Data Preprocessing

Preprocessing on CT scans: (a) Resample. The CT slice spacings were different. In order to obtain isotropic 3D volume data of nodules, the 3D nearest neighbor algorithm was used to resample the entire set of CTs of the patients, so that the spacing in each direction of x, y, and z was fixed to 0.625. We did the same for the 2D marker boxes to ensure the markers remain unchanged. (b) Extract the nodule area. The 2D markers frame were merged to obtain the 3D markers, and the target areas were extracted from the 3D markers. According to the merging rule: if IOU>0.3 in the 2D marker boxes of the adjacent layers, the merging could be performed, which means that the two 2D marker boxes belong to the same 3D marker. The final 3D marks were the smallest rectangle that can contain all the merged 2D marks. To extract the cube areas, we took the maximum rectangle side length as the side length and the rectangle center point as the center to extract the cube nodule areas.

Processing clinical information: The 3 clinical information were all enumerated information and the 3 characteristic values were assigned to the 3 clinical information. The assignment methods are: gender (male: 1, female: 0), glitch (with: 1, without: 0), upper lung (In: 1, not in: 0)

Data Augmentation

Data augmentation was performed by randomly selecting several of the following data enhancements methods, including random rotation, folding, center blocking, and brightness change. to perform data enhancement. The resulting datasets were uniformly resized into the dimension of 64x64x64.

Deep Learning Feature extraction

The feature extraction model is based on the backbone of the 3-D ResNet34 architecture. A 3-D (X, Y, Z) image volume is extracted and processed through successive 3-D (X, Y, Z) convolution and max pooling layers to produce spatial features that are gathered in a fully-connected layer into a 1-D "feature vector" and then to a final classification layer where a sigmoid function provides an output prediction.

Combined Deep learning Feature and Clinical Features

We selected three clinical features—sex, nodule upper lung, and speculation—that have the highest weights according to the Brock University cancer prediction equation. The Brock equation has been adopted by LungRADs.

These three clinical features of each patient were annotated by experienced thoracic radiologists. We combine these three features with the nodule features extracted from the corresponding CT scans by our deep learning model after principal component analysis (PCA) to avoid overfitting. We then performed a multivariate logistic regression to classify benign versus malignant nodules.

Training, Validation and Testing

The feature extraction convolutional neural network was trained on the NLST dataset using four GTX 1080 with a batch size of 32, a dropout rate of 0.5, Adam optimizer, and a learning rate of 0.0001. We performed five-fold cross-validation on the NLST dataset to get 5 sub-models, and 5 sub-models were tested on the test dataset. The final test result was the average of the results of the 5 sub-models.

Results

Study cohorts

A total of 1,556 nodules from 813 patients were collected from the NLST cohort to train the deep learning model. The detailed demographics are shown in Table 1. The flowchart of patient selection is shown in Figure 1.

Variables		No Lung Cancer	Lung Cancer	Total
NLST Dataset				
Patients				
	Women	100 (52.91%)	255 (40.861%)	355 (43.66%)
	Men	89 (47.09%)	369 (59.14%)	458 (56.34%)
Nodules				1556
	Women	464 (50.27%)	259 (28.06%)	723 (46.47%)
	Men	459 (49.73%)	374 (71.94%)	833 (53.53%)
Spiculation				
	No	922 (99.89%)	393 (62.08%)	1315 (84.51%)
	Yes	1 (0.11%)	240 (37.92%)	241 (15.49%)
Upper Lung				1456
	No	409 (44.31%)	244 (45.78%)	653
	Yes	514 (55.69%)	289 (54.22%)	803
LIDC Dataset				
Patients				
	Women	43 (49.42%)	53 (50.00%)	96 (49.74%)
	Men	44 (50.58%)	53 (50.00%)	97 (50.26%)
Nodules				1556
	Women	140 (43.34%)	128 (53.55%)	268 (47.60%)
	Men	183 (56.66%)	111 (45.45%)	295 (52.40%)
Spiculation				
	No	68 (28.45%)	14 (4.32%)	82 (14.56%)
	Yes	171 (71.55%)	310 (95.68%)	481 (85.44%)

Upper Lung				1456
	No	184 (56.79%)	111 (46.44%)	295 (53.40%)
	Yes	140 (43.21%)	128 (53.56%)	268 (46.60%)
IMC Dataset				
Patients				
	Women	183 (61.00%)	300 (59.96%)	484 (60.35%)
	Men	117 (39.00%)	202 (40.04%)	318 (39.65%)
Nodules				1556
	Women	1156 (66.28%)	300 (28.06%)	1456 (46.47%)
	Men	588 (33.72%)	202 (71.94%)	790 (53.53%)
Spiculation				
	No	0	105 (20.91%)	105 (84.51%)
	Yes	0	397 (79.09%)	397 (15.49%)
Upper Lung				
	No	0	232 (46.21%)	232 (46.21%)
	Yes	0	270 (53.79%)	270 (5.79%)

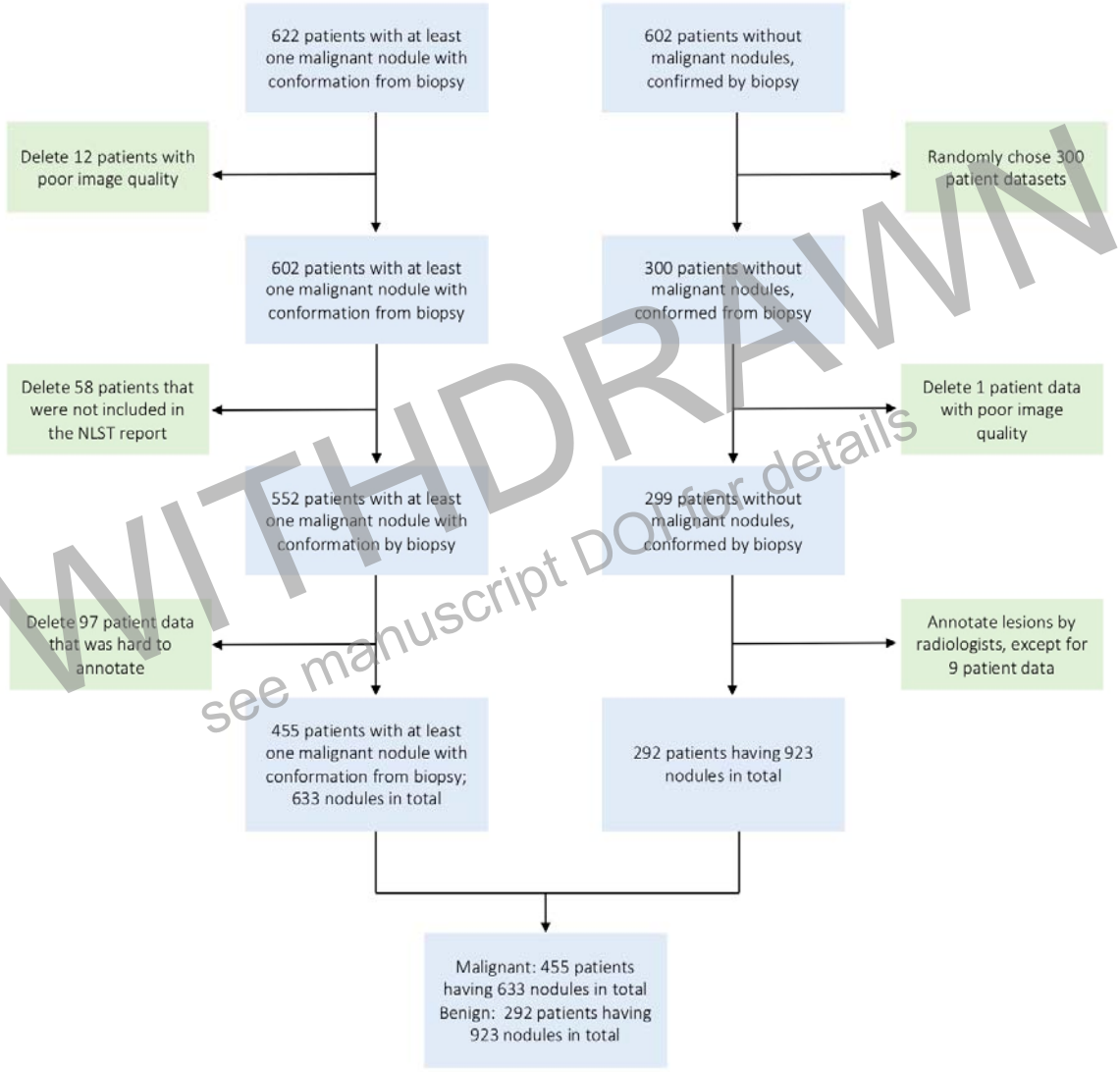

Fig 1. The protocol for patient selection from the NLST cohort. The same protocol is applied to the LIDC and IMC cohorts.

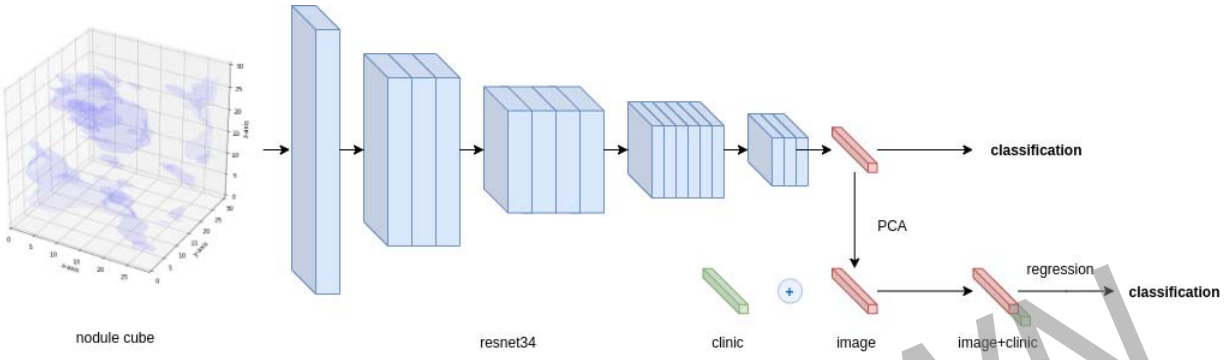

Fig 2. Framework of the proposed model.

Fig 2 illustrates our residual neural network model for the prediction of lung nodule malignancy. Our model consists of two major modules. The first module employs a convolutional neural network based on the architecture of ResNet-34 to extract high dimensional deep learning features from CT scans. Unlike previous pure deep learning-based approaches, which generate classification purely by the high dimensional features, we then reduce the dimension of the output of the first network as to avoid possible overfitting. After this, the dimensionally reduced output is combined with curated patient-wise clinical features by outer concatenation (an operation similar to outer product) and then fed into a multivariate logistic regression to predict the malignancy of the pulmonary nodules.

Discussion

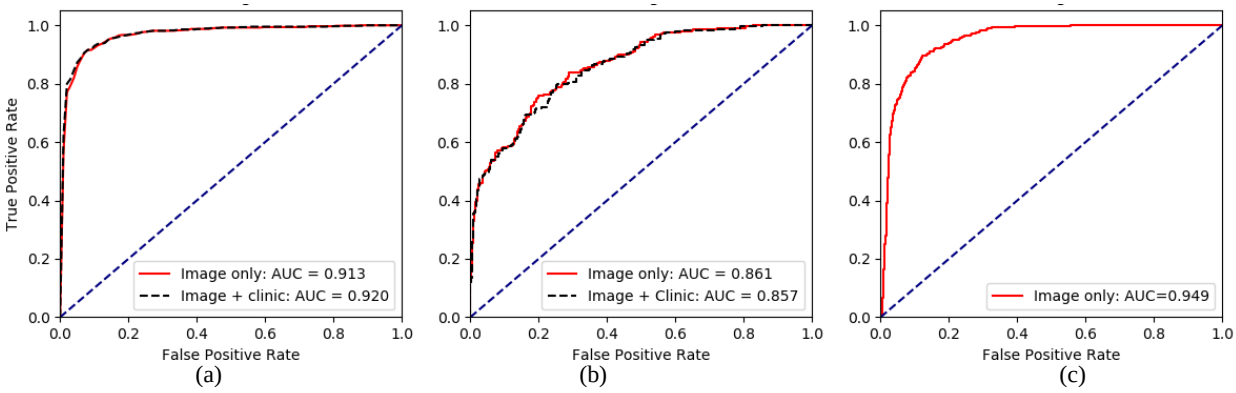

Figure 3. AUCs of the models using deep learning features extracted from CT images only (image only, red solid line) and combined deep learning feature and clinical feature (image + clinical, blue dash line) on the NLST dataset (a), LIDC dataset (b), and IMC dataset (c).

For training and validation, we performed analysis on 1,556 nodules with annotations (633 malignant nodules, 923 benign nodules) from the NLST cohort using 5-fold cross-validation. We found that, both the deep learning features and machine learning approache can achieve AUCs > 0.90 on the NLST cohort, and the integration of three clinical features does not remarkably increase the model performance. Unlike other deep learning-based algorithms for lung nodule malignancy prediction, which lack validation from external datasets, we also achieved high accuracy of prediction in other two datasets to evaluate the model's robustness in datasets containing many different metrics such as patients' race, country of residency, and the gold standard, which has been rarely reported. We found that both the pure *deep learning feature* and *DL feature plus clinical features* have good performance on the two validation datasets. The AUCs for LIDC, and LCR are 0.861 (pure DL) / 0.857 (DL + clinical), and 0.949 (pure DL),

respectively. Noticeably, LIDC has lower AUC value than the other two datasets. It might be because the LIDC's gold standard for malignancy is different than others. That is, the malignant nodules in the LIDC dataset are those labeled as "Highly Suspicious" by radiologists, meanwhile in the NLST and LIDC datasets, malignant nodules are confirmed by surgery biopsy.

Further, we found that the concatenation of the clinical features with the deep learning extracted features has a marginal impact on the recall and precision of the model. As aforementioned, we reduced the dimension of the deep learning features by selecting the 10 most weighted features that account for a total weight of 0.978. These 10 deep learning features are combined with the three clinical features as the input of the multivariate logistic regression model. We found that the prediction accuracy (AUCs, recall, and precision) was almost the same compared to the pure deep learning features without dimensionality reduction. Quantitatively, the summed-up weight of the 10 nodule features extracted by the deep learning algorithm equals to 0.091, while the weight of patient sex, nodule spiculation, and location is 0.031, 0.052, and 0.008, respectively. We speculate that, compared to the patient clinical features, image-derived features may provide a more direct correlation with the malignancy of the lung nodules.

Table 2. Cancer Risk Stratification on NLST dataset.

		Recall	Precision	TP	FP	GT
DL	Benign	0.841	0.644	159	88	189
	Malignancy	0.859	0.947	536	30	624
DL + clinical		Recall	Precision	TP	FP	GT
	Benign	0.873	0.620	165	101	189
	Malignancy	0.838	0.956	523	24	626

Table 3. Cancer Risk Stratification on LIDC dataset.

		Recall	Precision	TP	FP	GT
DL	Benign	0.802	0.817	256	58	323
	Malignancy	0.757	0.739	181	64	239
DL + clinical		Recall	Precision	TP	FP	GT
	Benign	0.746	0.834	241	48	323
	Malignancy	0.799	0.700	191	82	239

Table 4. Cancer Risk Stratification on IMC dataset.

		Recall	Precision	TP	FP	GT
DL	Benign	0.878	0.967	1531	53	1744
	Malignancy	0.894	0.678	449	213	502
DL + clinical		Recall	Precision	TP	FP	GT
	Benign	/	/	/	/	/
	Malignancy	/	/	/	/	/

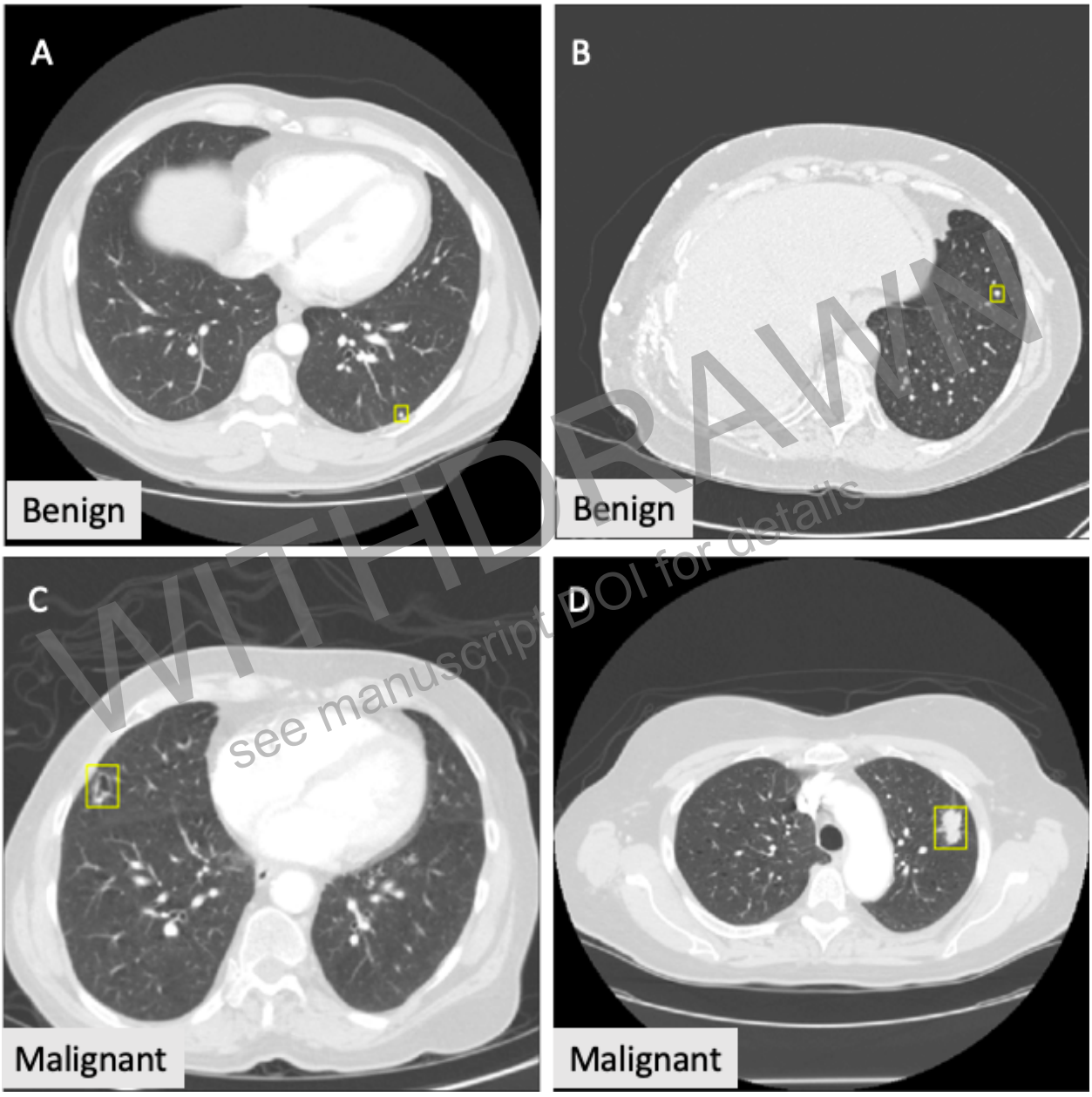

Fig. 4. A visual comparison between the nodules with low (A, B) or high (C, D) likelihood of being malignant. The model prediction rating for A, B, C, D are 0.03, 0.44, 0.73, and 0.99, respectively. The yellow square represents the nodule lesions.

Fig. 4 shows four CT images from the LIDC cohort. Fig. 4(A) represents a male with a nodule having no stipulation and not on the left upper lung, with a model predicted malignancy score of 0.03 and a label of benign. Fig. 4(B) denotes a male with a nodule having no stipulation and not on the left upper lung, with a model predicted malignancy score of 0.44 and a label of benign. Fig. 4(C) represents a female with a nodule having no stipulation and not on the left upper lung, with a model predicted malignancy score of 0.73 and a label of benign of being malignant. Fig. 4(D) denotes a male with a nodule having no stipulation and not on the left upper lung, with a model predicted malignancy score of 0.99 and a label of malignant.

We have many limitations in this study. First, the three datasets are composed of CT scans from multiple institutions from the US and China. The image quality may vary due to different ways of practices. Second, the ground truths

for nodule malignancy of the three datasets are different. As aforementioned, in the NLST dataset and Infervison multi-center dataset, the malignant nodules are confirmed by surgery biopsy. In the IMC dataset, however, the malignant nodules were identified and annotated by radiologists. Third, although the three datasets we chose in this study include patients with multiracial and multinational backgrounds, the dataset volume is still relatively small. We have been collecting more real-world data, and further study will be performed in the futher.

In conclusion, we evaluate the generalizability of a deep learning-based approach in the prediction of lung nodule malignancy from CT scans and clinical features. We found that the convolutional neural network algorithm trained by the NLST dataset achieved good prediction accuracy on the LIDC dataset and IMC dataset. The model shows good robustness on the three independent cohorts containing patients with diverse backgrounds such as race, ethnicity, socioeconomic status, and geography. Future, we compared models with the deep learning features only and combined deep learning and clinical features, and found no remarkable difference in the prediction accuracy. We believe this work showed potential to promote the recognition and adoption of deep learning-based approaches for lung cancer diagnosis in the clinical settings.

Authors' contributions

H.W. conceived the study concepts and design, result interpretation, and manuscript preparation. W.T. conducted the model development and training. C.W. performance data collection, annotation, and analysis. Y.D. and R.Z. contributed to experimental design, data analysis, and manuscript editing.

Compliance with ethical standards

Conflict of interest
The authors declare no conflicts of interest.

Reference

1. Rationale of DXQI Seed Grant Program, https://www.improvediagnosis.org/dxqi/
2. Serious misdiagnosis related harms in malpractice claims, https://www.degruyter.com/view/j/dx.2019.6.issue-3/dx-2019-0019/dx-2019-0019.xml?format=INT
3. Whang et al. The Causes of medical malpractice suits
4. Ciello et al. Missed lung cancer: when, where, why?
5. National Lung Screening Trial Research Team et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N. Engl. J. Med. 365, 395–409 (2011).
6. Black, W. C. et al. Cost-effectiveness of CT screening in the National Lung Screening Trial. N. Engl. J. Med. 371, 1793–1802 (2014).
7. Lung CT screening reporting & data system. American College of Radiology https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Lung-Rads
8. Field, J. K. et al. UK Lung Cancer RCT Pilot Screening Trial: baseline findings from the screening arm provide evidence for the potential implementation of lung cancer screening. Thorax 71, 161–170 (2016).

9. McMahon, P. M. et al. Cost-effectiveness of computed tomography screening for lung cancer in the United States. J. Thorac. Oncol. 2011;1841.
10. van Riel, S. J. et al. Observer variability for Lung-RADS categorisation of lung cancer screening CTs: impact on patient management. Eur. Radiol. 29, 924–931 (2019).
11. Singh, S. et al. Evaluation of reader variability in the interpretation of follow-up CT scans at lung cancer screening. Radiology 259, 263 (2011).
12. Causey J, Zhang J, Ma S, Jiang B, Qualls J, Politte D, Prior F, Zhang S et al. Highly accurate model for prediction of lung nodule malignancy with CT scans. Scientific Reports. 2018;9286.
13. Ardila D, Kiraly A et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat. Med. 2019;954.
14. McWilliams A, M. B., Tammemagi M, et al. Probability of cancer in pulmonary nodules detected on first screening CT. N. Engl. J. Med. 2013;910.