

Predicting malignancy of pulmonary ground-glass nodules and their invasiveness by random forest

Xueyan Mei^{1*}, Rui Wang^{2*}, Wenjia Yang³, Fangfei Qian³, Xiaodan Ye⁴, Li Zhu⁴, Qunhui Chen⁴, Baohui Han³, Timothy Deyer^{5,6}, Jingyi Zeng⁷, Xiaomeng Dong⁸, Wen Gao², Wentao Fang²

¹Department of Applied Biomedical Engineering, Whiting School of Engineering, Johns Hopkins University, Baltimore, MD, USA; ²Department of Thoracic Surgery, ³Department of Pulmonary Medicine, ⁴Department of Radiology, Shanghai Chest Hospital, Shanghai Jiao Tong University, Shanghai 200030, China; ⁵East River Medical Imaging, New York, NY, USA; ⁶Department of Radiology, Weill Cornell Medicine, New York, NY, USA; ⁷Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL, USA; ⁸Department of Data Science Analytics, University of Oklahoma, Norman, OK, USA

Contributions: (I) Conception and design: X Mei, R Wang, W Fang; (II) Administrative support: Q Chen, B Han, W Gao; (III) Provision of study materials or patients: X Ye, L Zhu, Q Chen; (IV) Collection and assembly of data: W Yang, F Chen, R Wang, W Fang; (V) Data analysis and interpretation: X Mei, T Deyer, J Zeng, X Dong; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

*These authors contributed equally to this work.

Correspondence to: Wentao Fang, MD. Department of Thoracic Surgery, Shanghai Chest Hospital, Shanghai Jiao Tong University, Huaihai West Road No. 241, Shanghai 200030, China. Email: vwtfang@hotmail.com.

Background: The purpose of this study was to develop a predictive model that could accurately predict the malignancy of the pulmonary ground-glass nodules (GGNs) and the invasiveness of the malignant GGNs.

Methods: The authors built two binary classification models that could predict the malignancy of the pulmonary GGNs and the invasiveness of the malignant GGNs

Results: Results of our developed model showed random forest could achieve 95.1% accuracy to predict the malignancy of GGNs and 83.0% accuracy to predict the invasiveness of the malignant GGNs.

Conclusions: The malignancy and invasiveness of pulmonary GGNs could be predicted by random forest.

Keywords: Ground-glass nodule (GGN); random forest

Submitted Jan 14, 2018. Accepted for publication Jan 16, 2018.

doi: 10.21037/jtd.2018.01.88

View this article at: <http://dx.doi.org/10.21037/jtd.2018.01.88>

Introduction

Lung cancer is the leading cause of cancer-related deaths worldwide (1). The U.S. national lung screening trial (NLST) demonstrated that a 20% reduction in lung cancer mortality for low-dose computed tomography (LDCT) compared to chest radiography (2). However in the NLST trial, 96.4% of participants with pulmonary nodules were benign. The risk of major complications for invasive diagnostic procedures was 4.5 per 10,000 persons screened while 25% of surgically resected nodules were benign (2). An accurate and practical non-invasive tool that can predict the malignancy of lung nodules will thus greatly reduce costs and surgical complications and avoid unnecessary

invasive diagnostic procedures or surgery.

With the wide application of CT screening, the detection of ground-glass nodules (GGNs) has become increasingly common, especially in pulmonary nodules less than 2 cm (2-5). Nodules can present as a pure GGN having only a ground-glass component and mixed GGN having both ground-glass and solid components on CT scan (6). Manual interpretation of nodule characteristics such as contour, shape, and margin is a widely accepted method for predicting malignancy of lung nodules. However, automated interpretation of the radiological characteristics of GGNs by deep learning techniques may potentially increase efficiency, improve reproducibility, and improve diagnostic accuracy.

In 2011, a new lung adenocarcinoma classification system was proposed by the International Association for the Study of Lung Cancer (IASLC)/American Thoracic Society (ATS)/European Respiratory Society (ERS) (7). According to this new classification, adenocarcinomas were classified as atypical adenomatous hyperplasia (AAH), adenocarcinoma in situ (AIS), minimally invasive adenocarcinoma (MIA) and invasive adenocarcinoma with lepidic, acinar, papillary, micropapillary or solid growth patterns (7). The presence of invasive components, especially with a solid or micropapillary growth pattern, correlated with poor prognosis in lung cancers (8,9). On CT scans, less aggressive subtypes such as AIS, MIA and lepidic adenocarcinomas frequently present as pure GGN (9). While the presence of aggressive subtypes, such as solid and micropapillary growth patterns, greatly impacts the disease recurrence and overall survival for patients with GGN (10,11). At present, sublobar resection is advocated for AIS/MIA while lobectomy and lymph node dissection is the standard treatment for invasive adenocarcinoma. Consequently, pre-operative planning in patients with pulmonary GGNs could be greatly improved if deep learning technique could differentiate the subtypes of adenocarcinoma based on their radiologic characteristics.

Herein, we focus on creating a machine learning model to quantify radiological imaging features of GGNs and thereby predict their malignancy and invasiveness. We first built a model to predict the benignity or malignancy of pulmonary GGNs. For those nodules identified as malignant, another binary classification model was built to predict the invasiveness of the nodule.

Methods

Data

The authors retrospectively collected CT images of pulmonary GGN from 1,177 patients who underwent either a sublobar resection or lobectomy from 2015 to 2017 at Shanghai Chest Hospital. CT scans were conducted using a 64-detector CT row scanner (Brilliance 64; Philips, Eindhoven, The Netherlands). Thin-section helical CT scans (1.0 mm collimation, 0.4-second gantry rotation time, 120 kVp, 349 mA) were obtained from the lung apices to the level of the mid portion of both kidneys. The image data were reconstructed with a slice thickness of 1.0 mm using soft tissue and lung algorithms. All CT data were interfaced with our picture archiving and communication

system (Kingstar Winning, Shanghai, China).

The data description and collection were conducted by three radiologists at Shanghai Chest Hospital. The images were evaluated in the following settings: lung window center—520 HU/lung window width—1,450 HU; mediastinal window center—40 HU/mediastinal window width—350 HU. For each case, the demographic data was documented. The following nodule image features were evaluated by consensus of the three radiologists:

- (I) Lobulation (yes/no);
- (II) Pleural retraction (yes/no);
- (III) Peritumoral vascularity (yes/no);
- (IV) Air bronchogram (yes/no);
- (V) Calcification (yes/no);
- (VI) Spiculation (yes/no);
- (VII) Bubble lucency (yes/no);
- (VIII) Nodule density (mixed/pure);
- (IX) Lung window largest diameter;
- (X) Lung window vertical diameter;
- (XI) Mediastinal window largest diameter;
- (XII) Mediastinal window vertical diameter;
- (XIII) Average CT value.

Malignancy and invasiveness predictions

We used random forest along with logistic regression, decision tree, support vector machine and AdaBoosting to predict the malignancy of the pulmonary GGNs and the invasiveness of malignant GGNs. Random forest is a classification method that consists of multiple nodes of decision trees. Random forest can overcome the overfitting issues to the training set generated by decision trees. The forests were comprised by using randomly selected input variables or combinations of variables at each node to grow each tree.

Results

Patient characteristics

The demographics and clinicopathologic characteristics of the cohort are shown in *Table 1*. Upon histological examination, 115 (9.8%) cases were benign, 474 (40.2%) AAH/AIS/MIA, 146 (12.4%) lepidic predominant lung adenocarcinoma, 345 (29.3%) acinar/papillary adenocarcinoma, 73 (6.2%) micropapillary/solid adenocarcinoma and 24 (2%) invasive mucinous adenocarcinoma (IMA).

Table 1 Descriptive characteristics of variable

Feature	Benign (n=115)	AAH/AIS/MIA (n=474)	Lepidic (n=146)	Acinar/papillary (n=345)	Micropapillary/solid (n=73)	IMA (n=24)
Age (y)						
<60	82 (71%)	337 (71%)	64 (44%)	181 (53%)	36 (49%)	16 (67%)
≥60	33 (29%)	137 (29%)	82 (56%)	164 (47%)	37 (51%)	8 (33%)
Gender						
Male	56 (49%)	119 (25%)	52 (36%)	130 (38%)	34 (47%)	9 (38%)
Female	59 (51%)	355 (75%)	94 (64%)	215 (62%)	39 (53%)	15 (62%)
Size in lung window (mm)	12.6±5.28	10.3±3.75	19.78±8.3	18.8±7.37	24.4±7.28	15.55±5.54
Solid proportion (CTR %)	1.98±3.49	0.48±1.5	4.0±5.7	7.09±7.29	15.83±8.2	6.65±4.99
CT value (HU)	-476.9±187.48	-528.55±142.05	-422.6±138.71	-289.96±157.72	-135.04±119.54	-272.46±143.56
Ground-glass component	-554.56±168.7	-569.71±107.62	-550±146.64	-466.23±124.7	-421.18±117.76	-491.37±127.5
Solid component	-138.85±152.65	-219.15±123.36	-142.98±114.73	-84.71±109.83	-1.14±47.39	-74.08±103.23
Location						
RUL	45 (39%)	177 (37%)	56 (38%)	118 (34%)	29 (40%)	6 (25%)
RML	9 (8%)	29 (6%)	11 (7.5%)	31 (9%)	3 (4%)	1 (4%)
RLL	20 (17.5%)	66 (14%)	27 (18.5%)	53 (15%)	13 (18%)	9 (38%)
LUL	22 (19%)	138 (29%)	41 (28%)	100 (29%)	22 (30%)	2 (8%)
LLL	19 (16.5%)	64 (14%)	11 (7.5%)	43 (12%)	6 (8%)	6 (25%)
GGN						
pGGN	40 (35%)	318 (67%)	16 (11%)	21 (6%)	0 (0%)	0 (0%)
mGGN	75 (65%)	156 (33%)	130 (89%)	324 (94%)	73 (100%)	24 (100%)
Pathology stage						
Benign/AAH	/	/	/	/	/	/
0	/	157 (33%)	/	/	/	/
IA1	/	304 (64%)	25 (17%)	85 (25%)	4 (5%)	11 (46%)
IA2	/	1 (0.2%)	96 (66%)	189 (55%)	25 (34%)	11 (46%)
IA3	/	/	18 (12%)	57 (16%)	16 (22%)	0 (0%)
IB	/	/	7 (5%)	14 (4%)	28 (38%)	2 (8%)

AAH, atypical adenomatous hyperplasia; AIS, adenocarcinoma in situ; MIA, minimally invasive adenocarcinoma; IMA, invasive mucinous adenocarcinoma; CTR, ratio of the diameter of the solid component relative to the maximal tumor diameter; RUL, right upper lobe; RML, right middle lobe; RLL, right lower lobe; LUL, left upper lobe; LLL, left lower lobe; GGN, ground-glass nodule.

Malignancy predictions

Random forest obtained 95.1% accuracy and 99.1% sensitivity. AdaBoosting and knearest neighbors (KNN) both achieved 99.4% percent sensitivity. Logistic regression had 80.0% specificity. *Table 2* shows the detailed results of predictions

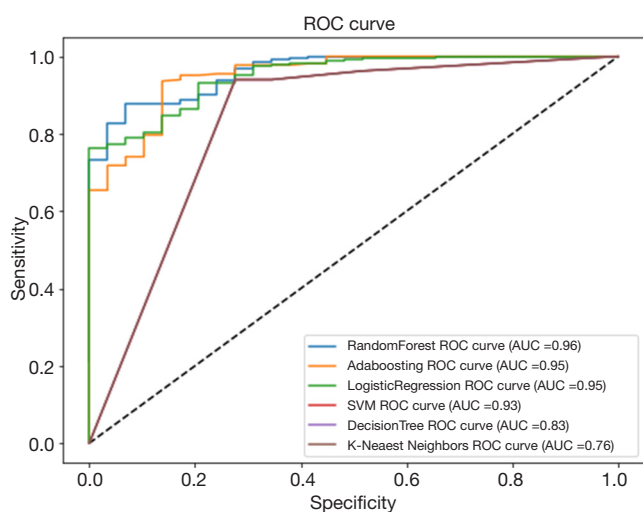
using different classification algorithms.

Figure 1 demonstrates the receiver operating characteristic (ROC) curves of the malignancy prediction models. As can be seen in the figure random forest outperformed [area under the curve (AUC) =96%] the other models [AdaBoosting (AUC

Table 2 Results of malignancy predictions (benign vs. malignant)

Algorithms	Accuracy (%)	Sensitivity (%)	Specificity (%)
Random forest	95.1	99.1	58.6
Logistic regression	86.9	87.7	80.0
Decision tree	91.7	95.6	55.9
AdaBoosting	93.4	99.4	37.9
SVM	91.3	94.8	58.6
KNN	93.4	99.4	37.9

SVM, support vector machine; KNN, k-nearest neighbors.

**Figure 1** ROC curve of malignancy predictive models. ROC, receiver operating characteristic; AUC, area under the curve.

=95%), logistic regression (AUC =95%), SVM (AUC =93%), decision tree (83%) and K-nearest neighbors (76%).

Invasiveness predictions

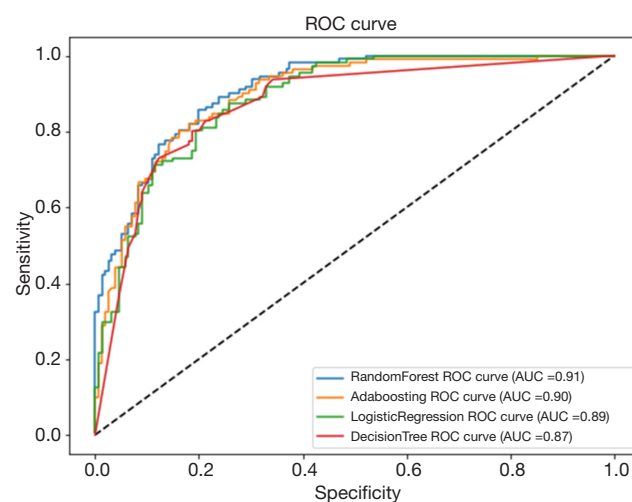
Random forest achieved the best overall accuracy, whereas logistic regression had the highest sensitivity and AdaBoosting had the highest specificity. *Table 3* shows a detailed comparison of the performance of each algorithm in predicting the invasiveness of the malignant GGNs.

Algorithmic performance for invasiveness is shown in the ROC curves in *Figure 2*. Random forest achieved the highest AUC at 91%. AdaBoosting, logistic regression and decision tree achieved 90%, 89% and 87% respectively (*Figure 2*).

Table 3 Results of invasiveness predictions [pre-invasion lesions (AAH, AIS, and MIA) vs. invasive adenocarcinoma]

Algorithms	Accuracy (%)	Sensitivity (%)	Specificity (%)
Logistic regression	81.5	81.1	81.8
Random forest	83.0	80.7	84.6
Decision tree	76.7	72.8	79.5
AdaBoosting	82.1	74.4	87.6

AAH, atypical adenomatous hyperplasia; AIS, adenocarcinoma in situ; MIA, minimally invasive adenocarcinoma.

**Figure 2** ROC curve of invasiveness predictive models. ROC, receiver operating characteristic; AUC, area under the curve.

Discussions

With the wide application of computed tomography screening for lung cancer worldwide, more GGNs are being found. The positive relationship of lesion size and CT features such as contour, shape, and margin to the likelihood of malignancy in solid pulmonary nodules has been clearly demonstrated (12). Historically, these features were thought to be less helpful in differentiating GGNs. However our study demonstrates that application of machine learning to evaluation of the CT characteristics of GGNs can help predict malignancy and invasiveness with high sensitivity and specificity.

Our model for the classification of pulmonary GGNs achieved 99.1% and 95.1% sensitivity and accuracy respectively, which would facilitate identification of malignant nodules in a screening setting. Automated interpretation

of radiological image traits for lung nodules has potential benefits such as increasing efficiency, reproducibility and improving prognosis by providing early detection and treatment. For malignancy predictions, the ensemble model random forest achieved the best performance among other classifiers. This is because the bagging strategy of random forest is more generalized and is able to incorporate different types of features. The majority voting method of random forest can effectively reduce misclassifications. In comparison with random forest bagging approach, the boosting strategy of AdaBoost tends to more likely overfit resulting in its lower accuracy and specificity score.

AAH, AIS and MIA refer to pre-invasive lesions in the lung. The disease-free survival rate is 100% in patients with AAH, AIS or MIA lesions following sublobar resection (10). However, lobectomy with systemic lymph node dissection or sampling is a standard surgical procedure for invasive adenocarcinoma (10). Considering that invasiveness of adenocarcinomas is hard to decide accurately by intraoperative frozen section, pre-treatment differentiation of the invasiveness of the nodule is helpful for pre-operative planning. As with malignancy prediction, the robustness of the bagging strategy of random forest worked well with invasiveness prediction resulting in higher accuracy, sensitivity and AUC.

To take most advantage of machine learning algorithms, it is useful to combine the strengths of multiple classifiers and minimize the misclassification errors. Although random forest already benefits from bagging among a lot of decision trees, it is not enough to only consider one silo classifier. In the future, we will try some more advanced ensemble approaches, such as ensemble on various basic models or stacking, to optimize the accuracy in prognosis.

One limitation of this study is the use of radiologic features, instead utilizing the imaging information directly from the CT images. Future research will use the imaging dataset directly to develop deep learning algorithms for the same purpose. This has been shown to be quite successful in thyroid nodules but not yet tried in lung cancers (13,14).

In conclusion, our study is the first attempt to differentiate benign and malignant pulmonary GGNs and to predict invasiveness of malignant nodules using machine-learning models. Further study is warranted so that artificial intelligence could be incorporated into clinical practice to improve management outcomes.

Acknowledgements

Funding: This work is supported by Shanghai Science

and Technology Commission Foundation project (No. 14411950800) (W Gao).

Footnote

Conflicts of Interest: The authors have no conflicts of interest to declare.

Ethical Statement: All the data of this study was retrospectively collected, thereby the ethical approval patients' informed consent was waived by the institutional review board of Shanghai Chest Hospital.

References

1. Torre LA, Bray F, Siegel RL, et al. Global cancer statistics, 2012. *CA Cancer J Clin* 2015;65:87-108.
2. National Lung Screening Trial Research Team, Aberle DR, Adams AM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med* 2011;365:395-409.
3. Lee HY, Lee KS. Ground-glass opacity nodules: histopathology, imaging evaluation, and clinical implications. *J Thorac Imaging* 2011;26:106-18.
4. Takahashi M, Shigematsu Y, Ohta M, et al. Tumor invasiveness as defined by the newly proposed IASLC/ATS/ERS classification has prognostic significance for pathologic stage IA lung adenocarcinoma and can be predicted by radiologic parameters. *J Thorac Cardiovasc Surg* 2014;147:54-9.
5. Kovalchik SA, Tammemagi M, Berg CD, et al. Targeting of low-dose CT screening according to the risk of lung-cancer death. *N Engl J Med* 2013;369:245-54.
6. Yankelevitz DF, Yip R, Smith JP, et al. CT Screening for Lung Cancer: Nonsolid Nodules in Baseline and Annual Repeat Rounds. *Radiology* 2015;277:555-64.
7. Travis WD, Brambilla E, Noguchi M, et al. International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. *J Thorac Oncol* 2011;6:244-85.
8. Hung JJ, Yeh YC, Jeng WJ, et al. Predictive Value of the International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society Classification of Lung Adenocarcinoma in Tumor Recurrence and Patient Survival. *J Clin Oncol* 2014;32:2357-64.
9. Miao Y, Zhang J, Zou J, et al. Correlation in histological

- subtypes with high resolution computed tomography signatures of early stage lung adenocarcinoma. *Transl Lung Cancer Res* 2017;6:14-22.
10. Dembitzer FR, Flores RM, Parides MK, et al. Impact of histologic subtyping on outcome in lobar vs sublobar resections for lung cancer: a pilot study. *Chest* 2014;146:175-81.
 11. Hung JJ, Yeh YC, Jeng WJ, et al. Predictive value of the international association for the study of lung cancer/ American Thoracic Society/European Respiratory Society classification of lung adenocarcinoma in tumor recurrence and patient survival. *J Clin Oncol* 2014;32:2357-64.
 12. Asamura H, Hishida T, Suzuki K, et al. Radiographically determined noninvasive adenocarcinoma of the lung: survival outcomes of Japan Clinical Oncology Group 0201. *J Thorac Cardiovasc Surg* 2013;146:24-30.
 13. Mei XY, Dong XM, Trafalis T, et al. Convolutional Auto-Encoders Based Feature Extraction for Thyroid Nodule Ultrasound Image. 17th International Symposium on Mathematical and Computational Biology, November 2017.
 14. Mei XY, Dong XM, Deyer T, et al. Thyroid Nodule Benignity Predict by Deep Feature Extraction. 2017 IEEE 17th International Conference Bioinformatics and Bioengineering (BIBE), 2017:24145.

Cite this article as: Mei X, Wang R, Yang W, Qian F, Ye X, Zhu L, Chen Q, Han B, Deyer T, Zeng J, Dong X, Gao W, Fang W. Predicting malignancy of pulmonary ground-glass nodules and their invasiveness by random forest. *J Thorac Dis* 2018;10(1):458-463. doi: 10.21037/jtd.2018.01.88