



Application of deep learning (3-dimensional convolutional neural network) for the prediction of pathological invasiveness in lung adenocarcinoma

著者 (英)	Masahiro Yanagawa, Hirohiko Niioka, Akinori Hata, Noriko Kikuchi, Osamu Honda, Hiroyuki Kurakami, Eiichi Morii, Masayuki NOGUCHI, Yoshiyuki Watanabe, Jun Miyake, Noriyuki Tomiyama
journal or publication title	Medicine
volume	98
number	25
page range	e16119
year	2019-06
権利	(C) 2019 the Author(s). Published by Wolters Kluwer Health, Inc. This is an open access article distributed under the terms of the Creative Commons Attribution-Non Commercial-No Derivatives License 4.0 (CCBY-NCND), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.
URL	http://hdl.handle.net/2241/00159155

doi: 10.1097/MD.00000000000016119



Application of deep learning (3-dimensional convolutional neural network) for the prediction of pathological invasiveness in lung adenocarcinoma

A preliminary study

Masahiro Yanagawa, MD, PhD^{a,*}, Hirohiko Niioka, PhD^b, Akinori Hata, MD, PhD^a, Noriko Kikuchi, MD^a, Osamu Honda, MD, PhD^a, Hiroyuki Kurakami, PhD^c, Eiichi Morii, MD, PhD^d, Masayuki Noguchi, MD, PhD^e, Yoshiyuki Watanabe, MD, PhD^f, Jun Miyake, PhD^g, Noriyuki Tomiyama, MD, PhD^a

Abstract

To compare results for radiological prediction of pathological invasiveness in lung adenocarcinoma between radiologists and a deep learning (DL) system.

Ninety patients (50 men, 40 women; mean age, 66 years; range, 40–88 years) who underwent pre-operative chest computed tomography (CT) with 0.625-mm slice thickness were included in this retrospective study. Twenty-four cases of adenocarcinoma in situ (AIS), 20 cases of minimally invasive adenocarcinoma (MIA), and 46 cases of invasive adenocarcinoma (IVA) were pathologically diagnosed. Three radiologists of different levels of experience diagnosed each nodule by using previously documented CT findings to predict pathological invasiveness. DL was structured using a 3-dimensional (3D) convolutional neural network (3D-CNN) constructed with 2 successive pairs of convolution and max-pooling layers, and 2 fully connected layers. The output layer comprises 3 nodes to recognize the 3 conditions of adenocarcinoma (AIS, MIA, and IVA) or 2 nodes for 2 conditions (AIS and MIA/IVA). Results from DL and the 3 radiologists were statistically compared.

No significant differences in pathological diagnostic accuracy rates were seen between DL and the 3 radiologists ($P > .11$). Receiver operating characteristic analysis demonstrated that area under the curve for DL (0.712) was almost the same as that for the radiologist with extensive experience (0.714; $P = .98$). Compared with the consensus results from radiologists, DL offered significantly inferior sensitivity ($P = .0005$), but significantly superior specificity ($P = .02$).

Despite the small training data set, diagnostic performance of DL was almost the same as the radiologist with extensive experience. In particular, DL provided higher specificity than radiologists.

Abbreviations: 3D-CNN = 3-dimensional convolutional neural network, AIS = adenocarcinoma in situ, AUC = area under the curve, C1 = convolution 1, C2 = convolution 2, CAD = computer-aided diagnosis, CI = confidence interval, DL = deep learning, GGN = ground-glass nodule, IVA = invasive adenocarcinoma, M1 = max pooling 1, M2 = max pooling 2, MIA = minimally invasive adenocarcinoma, ROC = receiver operating characteristic.

Keywords: artificial intelligence, convolutional neural network, deep learning, lung adenocarcinoma, pathological invasiveness

Editor: Hisashi Oshiro.

The authors have no conflicts of interest to disclose.

^a Department of Radiology, Osaka University Graduate School of Medicine,

^b Institute for Dataability Science, ^c Department of Medical Innovation, Osaka University, ^d Department of Pathology, Osaka University Graduate School of Medicine, Suita-city, Osaka, ^e Department of Diagnostic Pathology, University of Tsukuba, Tsukuba-city, Ibaraki, ^f Department of Future Diagnostic Radiology, Osaka University Graduate School of Medicine, ^g Global Center for Medical Engineering and Informatics, Osaka University, Suita-city, Osaka, Japan.

* Correspondence: Masahiro Yanagawa, The Department of Radiology, Osaka University Graduate School of Medicine, 2-2 Yamadaoka, Suita-city, Osaka, 565-0871, Japan (e-mail: m-yanagawa@radiol.med.osaka-u.ac.jp).

Copyright © 2019 the Author(s). Published by Wolters Kluwer Health, Inc. This is an open access article distributed under the terms of the Creative Commons Attribution-Non Commercial-No Derivatives License 4.0 (CCBY-NC-ND), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.

Medicine (2019) 98:25(e16119)

Received: 12 March 2019 / Received in final form: 21 May 2019 / Accepted: 25 May 2019

<http://dx.doi.org/10.1097/MD.00000000000016119>

1. Introduction

CT has been widely used for qualitative and quantitative imaging of lung cancer. Recent advances in CT techniques have enabled the provision of thin-slice image data with high image quality with a short period. As symbolized by the appearance of the term “big data” in the mid-2000s, large amounts of data are now easily obtained. In the medical field, clinical data such as CT images are constantly accumulating. Quick and efficient extraction of valuable information from among large amounts of clinical data is thus important. Computer-aided diagnosis (CAD) systems have the potential to improve the clinical diagnostic process by offering correct classification decisions and volumetric measurements, representing promising avenues for providing accurate, reproducible quantification of lung cancer.^[1–3] Semi-automatic and/or automatic differentiation between benign and malignant tumors is one of the major tasks of CAD.^[4–8] However, many problems have to be overcome for CAD systems in terms of the integration and selection of image features because of the dependence on multiple image-processing tasks.^[3,9]

Recently, improvements in computing capacity, the evolution of deep learning algorithms, and the appearance of big data have brought about a boom in tertiary artificial intelligence. Deep learning (DL), structured by building a model to imitate the human brain, is one of the artificial intelligence systems based on neural networks. DL techniques are currently considered state of the art for the classification of images^[10] and applied for some fields of medical images.^[11–14] The neural network begins by simulating neural cells and trying to simulate the human brain using a simulation model called a perceptron. A multilayer perceptron is constructed by making and arranging layers with perceptrons in which all nodes in the model are fully connected, allowing the solution of more complicated problems. Artificial intelligence technology has been applied in the field of thoracic imaging, and has developed in the following areas^[15–30]: detection of pulmonary nodules; differentiation between benign and malignant lesions; diagnosis of diffuse lung diseases (i.e., retrieval system for resemble cases); and improvement of 3D-analysis and image quality (e.g., Pixelshine, a noise-reduction algorithm using machine learning; AlgoMedica, Sunnyvale, CA).

Lung adenocarcinoma is the most common histopathological subtype of lung cancer. Early diagnosis of pathological invasiveness using CT may alter the course of treatment of adenocarcinomas and subsequently improve the prognosis.^[21,22] However, radiological prediction of pathological invasiveness is very difficult. Moreover, to the best of our knowledge, no CAD systems have yet been shown to predict pathological invasiveness.

Under the hypothesis that DL might be able to predict pathological invasiveness, DL intended to be capable of such prediction was developed in cooperation with the department of technology in our institution. The purpose of our study was to compare results for radiological prediction of pathological invasiveness in lung adenocarcinoma between radiologists and DL.

2. Materials and methods

2.1. Study population

This study was approved by the internal ethics review board at our institute. The need to obtain informed consent for this retrospective review of patient records and images and the use of patient biomaterials was waived in this study. Consecutive 90 patients (50 men, 40 women; mean age, 66 years; range, 40–88 years) with 90 nodules who had undergone surgery at our institution between 2009 and 2011 were included (Fig. 1). All patients had undergone preoperative thin-section CT of the chest. Patients who had received previous treatments of the lungs or other organs were excluded from the study. Patients with histological subtypes other than adenocarcinoma were also excluded.

Pathological diagnoses were made by 2 independent pathologists according to the 2015 World Health Organization (WHO) Classification of lung tumors^[23] as adenocarcinoma in situ (AIS), minimally invasive adenocarcinoma (MIA), or

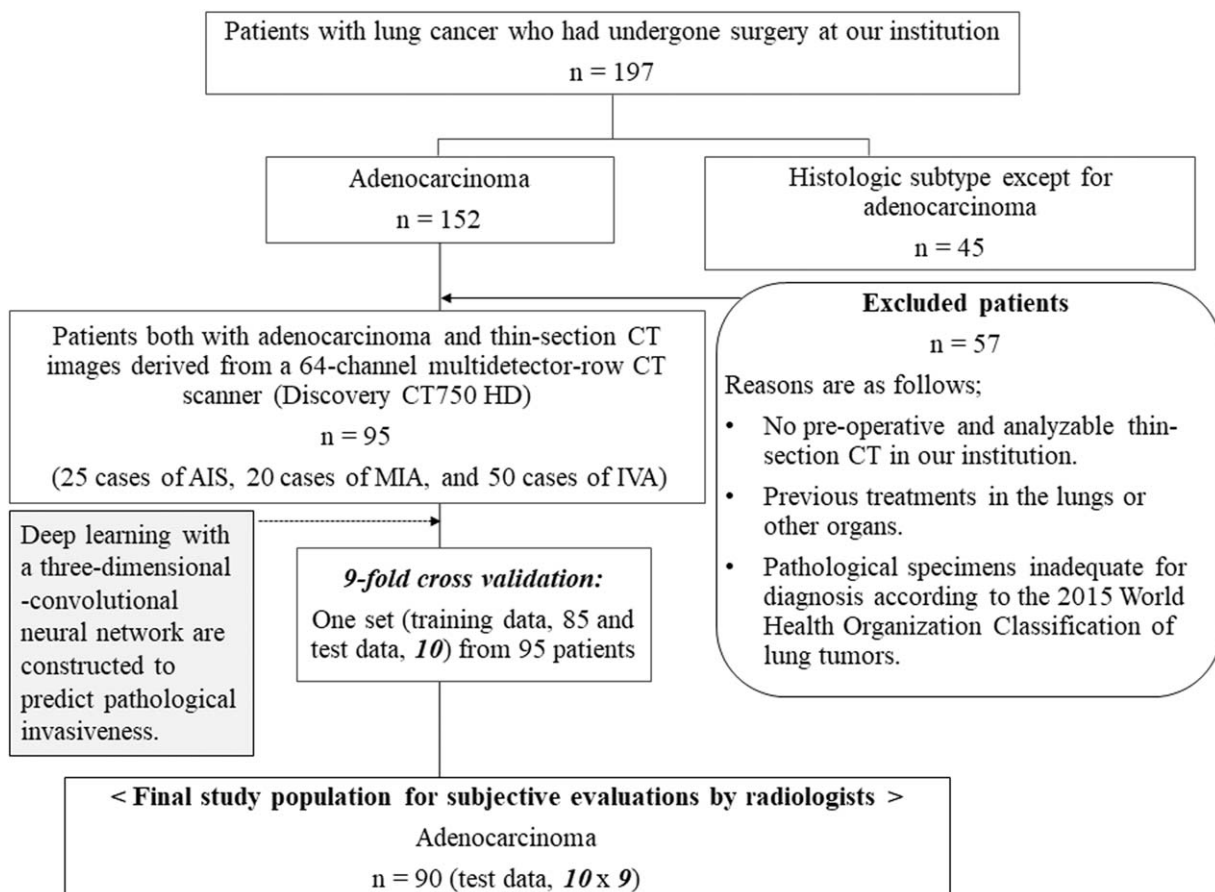


Figure 1. Flowchart of patient population in the study.

invasive adenocarcinoma (IVA). Histological diagnoses of AIS, MIA, and IVA were confirmed by consensus decisions.

2.2. CT protocols

Ninety patients underwent scanning using a 64-channel multi-detector-row CT scanner (Discovery CT750 HD; GE Healthcare, Milwaukee, WI) with the following protocols: detector collimation, 0.625 mm; detector pitch, 0.984; gantry rotation period, 0.4 seconds; matrix size, 512 × 512 pixels; X-ray voltage, 120 kVp; tube current, auto exposure control (mA); field of view, 34.5 cm for full lung; high-resolution mode with 2496 views per rotation. All targeted lung CT images were reconstructed using a 20-cm field of view from thin-section CT images reconstructed with a high spatial frequency algorithm at 0.625-mm thicknesses using a 30% adaptive statistical iterative reconstruction.

2.3. Subjective evaluation by radiologists

Three chest radiologists (AH [R1], NK [R2], and OH [R3], with 8, 9, and 26 years of experience, respectively) were instructed only to evaluate each nodule. They independently diagnosed each nodule by predicting pathological invasiveness from CT findings on a 21-inch monochrome liquid crystal display monitor without prior knowledge of histopathological diagnoses. CT findings were classified according to previously reported criteria^[24–31]: irregular margin; air bronchogram with disruption and/or irregular dilatation; pleural indentation; and solid component in a part-solid nodule (size ≤ 5 mm or > 5 mm). In cases of ground-glass nodule (GGN), nodule density (dense or inhomogeneous) was evaluated

according to the previous report^[27]: dense GGN, with CT value > -400 Hounsfield unit; and inhomogeneous GGN, complicated distribution of solid-like portions. Final diagnosis (i.e., AIS, or MIA, or IVA) was decided comprehensively according to these CT findings. Nodule distribution (GGN, part-solid GGN, and solid) of 90 nodules was decided by consensus. Total nodule size and solid component size were measured by the principal investigator (M.Y., with 17 years of experience).

2.4. Objective evaluation by the deep learning

DL with a 3-dimensional (3D)-convolutional neural network (CNN) was used in the present study. Tensorflow™ (version 0.12.1, Google Inc., Mountain View, CA) was used as the DL framework. The 3D CT images including a nodule and a surrounding normal lung parenchyma were classified by the 3D-CNN. Input data size was 30 × 30 × 30 voxels (11.7-mm cube in real space). The 3D-CNN structure was constructed with 2 successive pairs of convolution (C1 and C2) and max-pooling layers (M1 and M2), and 2 fully connected layers (Fig. 2). Kernel sizes were 3 × 3 × 3 voxels (C1), 2 × 2 × 2 voxels (M1), 2 × 2 × 2 voxels (C2), and 3 × 3 × 3 voxels (M2). Numbers of convolution filters were 32 and 64 for each convolution layer (C1 and C2), respectively. The number of nodes in the first fully connected layer was 8000. The output layer comprised 3 nodes to recognize the 3 conditions of adenocarcinoma (AIS, or MIA, or IVA), or 2 nodes for 2 conditions (AIS or MIA/IVA). A rectified linear unit (ReLU) was used for the activation function in layers other than the output layer. A softmax function was used for converting output values to probabilities in the output layer. The CT images

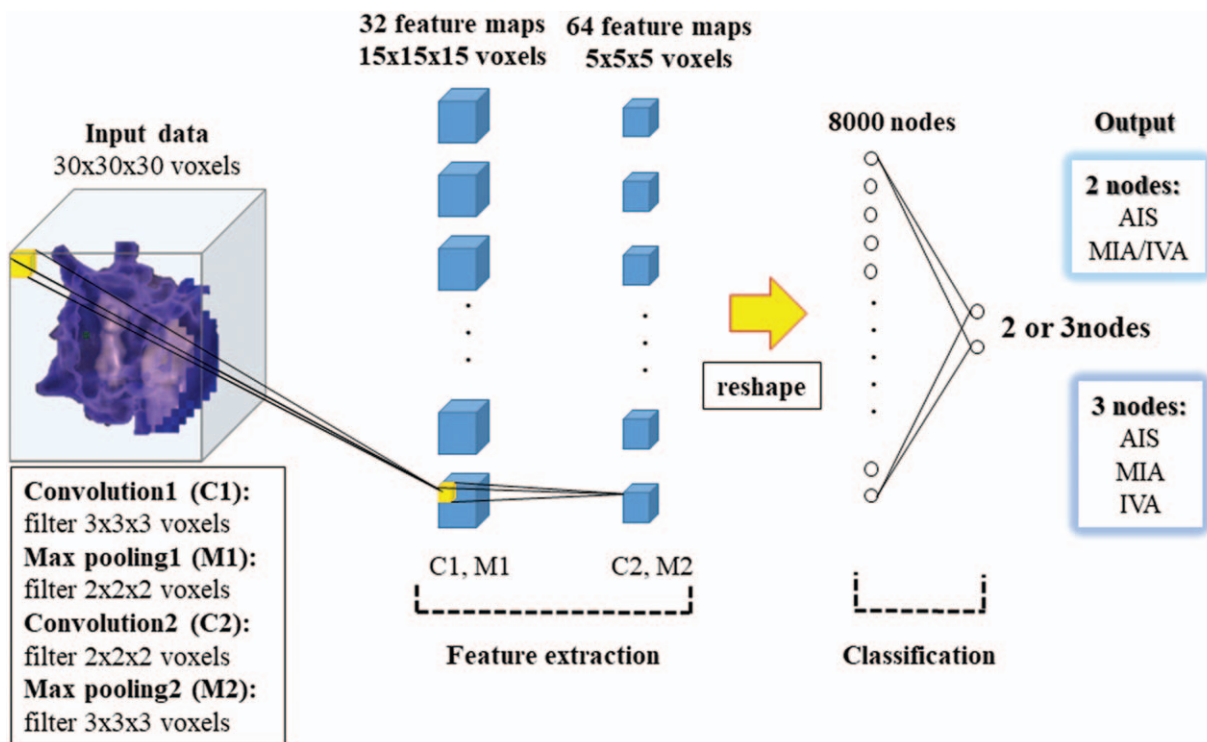


Figure 2. Three-dimensional convolutional neural network (3D-CNN) structure. The 3D-CNN structure consisted of feature extraction and classification. Feature extraction was constructed using 2 successive pairs of convolution (C1 and C2) and max-pooling layers (M1 and M2). The classification was of 2 fully connected layers. The first 8000 fully connected layers were classified into 3 nodes (AIS, or MIA, or IVA) or 2 nodes (AIS or MIA/IVA). AIS=adenocarcinoma in situ, IVA=invasive adenocarcinoma, MIA=minimally invasive adenocarcinoma.

Table 1
Total nodule size and solid component size for each nodule type.

Nodule type	Histology	N	Total nodule size	Solid component size
			(mean ± SD, range) [mm]	(mean ± SD, range) [mm]
GGN	AIS	16	16.3 ± 6.3, 7–29	0
	MIA	3	15.7 ± 11.2, 6–28	0
	IVA	5	17.6 ± 2.4, 15–21	0
Part-solid	AIS	8	18.0 ± 6.3, 9–26	7.1 ± 3.1, 3–12
	MIA	12	19.3 ± 4.8, 13–27	9.3 ± 3.9, 4–16
	IVA	20	24.7 ± 4.8, 14–30	17.6 ± 6.5, 8–27
Solid	AIS	0	N/A	N/A
	MIA	5	14.6 ± 5.5, 6–20	14.6 ± 5.5, 6–20
	IVA	21	20.1 ± 6.4, 9–30	20.1 ± 6.4, 9–30

AIS=adenocarcinoma in situ, GGN=ground-glass nodule, IVA=invasive adenocarcinoma, MIA=minimally invasive adenocarcinoma, SD=standard deviation.

of 95 nodules (25 cases of AIS, 20 cases of MIA, and 50 cases of IVA) scanned under the same conditions were divided into cases of training data, 85 and test data, 10. To predict pathological invasiveness, 9-fold cross-validation was performed on 90 cases of CT image data (Fig. 1). In each training and prediction process, data was augmented by adding some noise to 85 cases of training data.

2.5. Statistical analyses

The following statistical analyses were performed using commercially available software (MedCalc version 13.1.2.0, 64 bit; Frank Schoonjans, Mariakerke, Belgium). Accuracy rates among DL and the 3 radiologists were statistically analyzed using repeated-measures analysis of variance, conducted with Bonferroni correction applied for multiple comparisons. A receiver operating characteristic (ROC) analysis was used to determine area under the curves (AUC) of DL and the 3 radiologists, respectively: each CT diagnostic performance (i.e., 0=AIS, 1=MIA/IVA) in DL and the 3 radiologists was used as a variable, and each pathological diagnosis (i.e., 0=pathological AIS, 1=pathological MIA/IVA) was used as a classification variable. Multivariate ROC analysis was performed to determine the statistical significance of the difference among 4 AUCs (DL and the 3 radiologists). The additional statistical analysis was performed using commercially available software (R version 3.4.1, 64 bit; R Core Team, Vienna, Austria [https://www.R-project.org/]). Sensitivity and specificity among DL and the consensus result of radiologists were compared using the McNemar test with Bonferroni correction. Corrected values of $P < .05$ were considered significant.

3. Results

3.1. Demographics of our study population

Final study population for subjective evaluations by radiologists included 90 patients with 90 nodules (Fig. 1). Nine data set groups were created from 95 patients: each data set group consisted of training data, 85 and test data, 10. As a result of 9-fold cross-validation, the diagnostic performance of DL for pathological invasiveness in 90 nodules was obtained. The same 90 nodules were evaluated by each radiologist. Nodule distribution of 90 nodules was as follows: 24 cases of GGN, 40 cases of part-solid GGN, and 26 cases of solid nodule. The relationship between nodule type and histopathological results is summarized in Table 1. GGN and part-solid GGN included all

histological subtypes. Solid nodules included only MIA and IVA. Total nodule size and solid component size for each nodule type are summarized in Table 1.

3.2. Pathological diagnostic accuracy rates among DL and the 3 radiologists

In differentiating among AIS, MIA, and IVA, no significant differences in pathological diagnostic accuracy rates were seen among all observers (DL, R1, R2, and R3; $P > .105$). Results are summarized in Table 2. In 11 cases, only DL could accurately differentiate among AIS (n=4), MIA (n=6), and IVA (n=1) (Fig. 3). The 4 cases of AIS comprised 1 GGN and 3 part-solid GGNs, and the 6 cases of MIA comprised 5 part-solid GGNs and 1 solid nodule. The case of IVA represented GGN.

Similarly, in differentiation between AIS (without pathological invasiveness) and MIA/IVA (with pathological invasiveness), no significant differences in pathological diagnostic accuracy rates were evident among all observers (DL, R1, R2, and R3) ($P > .120$). Results are summarized in Table 2.

3.3. Diagnostic performance: differentiation between AIS and MIA/IVA

Results of multivariate ROC analyses were as follows: the AUC for DL, 0.712 (95% confidence interval [CI], 0.607–0.803); for

Table 2
Pathological diagnostic accuracy rates among observers.

Differentiation among AIS, MIA, and IVA		
Observer	Diagnostic accuracy rates	95% confidence interval
DL	46/90 (51)	40.58–61.64
R1	55/90 (61)	42.83–63.84
R2	48/90 (53)	50.84–71.38
R3	58/90 (64)	54.36–74.53
Differentiation between AIS and MIA/IVA		
Observer	Diagnostic accuracy rates	95% confidence interval
DL	66/90 (73)	64.02–82.65
R1	72/90 (80)	71.58–88.42
R2	67/90 (74)	65.26–83.63
R3	75/90 (83)	75.48–91.18

AIS=adenocarcinoma in situ, DL=deep learning, IVA=invasive adenocarcinoma, MIA=minimally invasive adenocarcinoma, R1=radiologist with 8 years of experience, R2=radiologist with 9 years of experience, R3=radiologist with 26 years of experience. Figures in parentheses show percent.

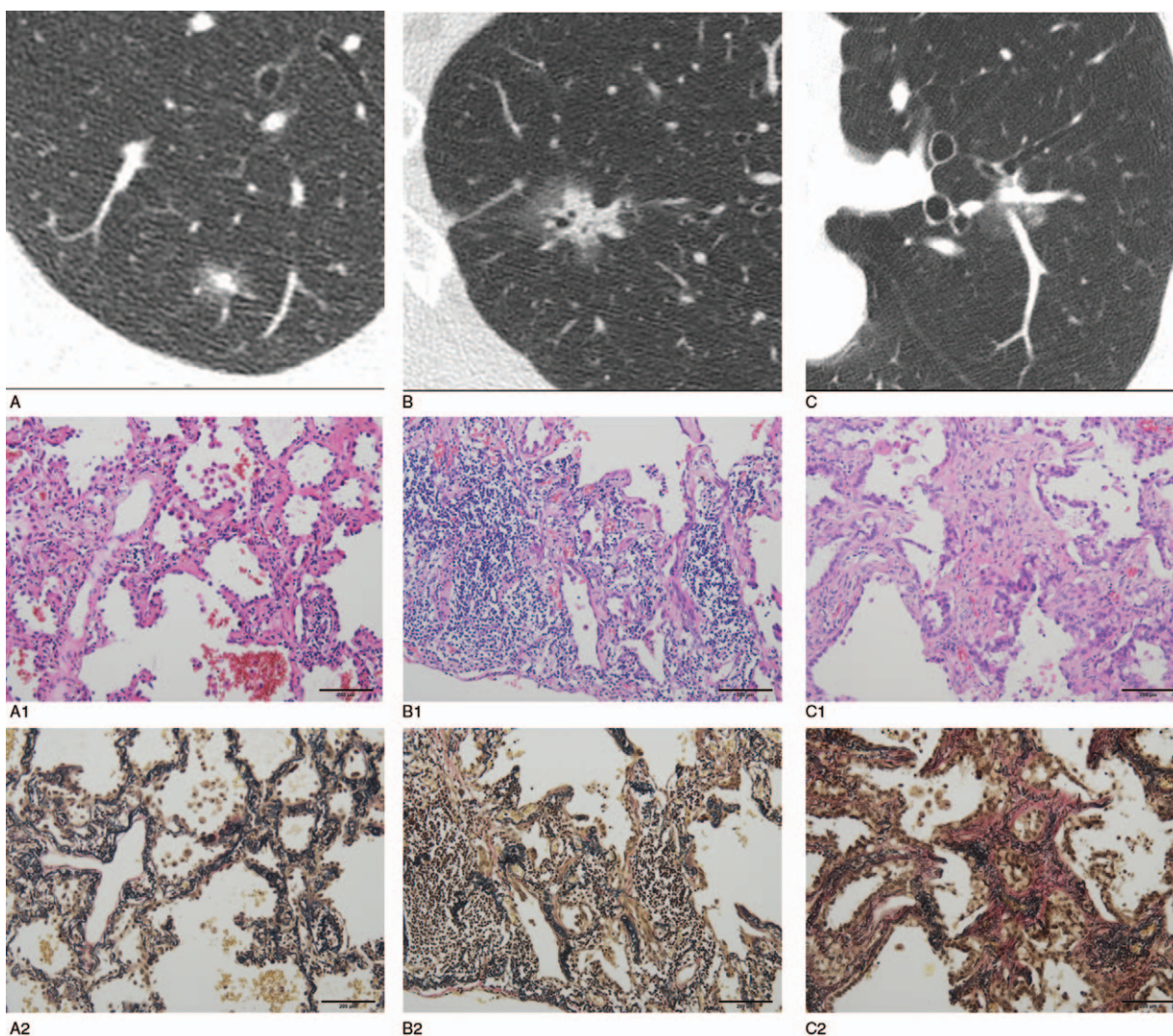


Figure 3. Cases in which only the DL system could achieve accurate diagnosis. Pathologic diagnosis of AIS was made using hematoxylin and eosin stain (A-1) and elastica-van Gieson stain (A-2). Only the DL system could diagnose this part-solid GGN (total nodule size, 9 mm; solid component size, 3 mm) as AIS (A-3). All radiologists misdiagnosed the lesion as MIA. Pathologic diagnosis of MIA was made using hematoxylin and eosin stain (B-1) and elastica-van Gieson stain (B-2). Only the DL system could diagnose this part-solid GGN (total nodule size, 20 mm; solid component size, 12 mm) as MIA (B-3). All radiologists misdiagnosed the lesion as IVA. Pathologic diagnosis of IVA was made using hematoxylin and eosin stain (C-1) and elastica-van Gieson stain (C-2). Only the DL system could diagnose this GGN (total nodule size, 17 mm) as IVA (C-3). All radiologists misdiagnosed the lesion as AIS or MIA. AIS=adenocarcinoma in situ, GGN=ground-glass nodule, IVA=invasive adenocarcinoma, MIA=minimally invasive adenocarcinoma.

R1, 0.665 (95% CI, 0.557–0.761); for R2, 0.574 (95% CI, 0.465–0.678); and for R3, 0.714 (95% CI, 0.609–0.804). The AUC (0.712) for DL was almost the same as that (0.714) for the most experienced radiologist (R3; $P=.983$), who had a significantly higher AUC than the radiologist with the least amount of experience (R2; $P=.026$). Compared with the consensus result for radiologists, DL offered significantly inferior sensitivity ($P=.0005$), but significantly superior specificity for diagnosing the invasiveness of adenocarcinoma ($P=.02$) (Table 3).

4. Discussion

This study has shown that regardless of the small training data set, diagnostic performance of DL was almost the same as that of the most experienced radiologist, who showed a significantly higher AUC than the radiologist with the least amount of

experience. In particular, although the sensitivity of DL was inferior to that of radiologists, DL provided higher specificity for diagnosing the invasiveness of adenocarcinoma than radiologists. In the present study, DL more accurately suggested the percentage of adenocarcinomas correctly identified as not having pathological invasiveness. If many more learning cases could be used as training data, the likelihood that the performance of DL will exceed that of human beings seems extremely high.

CAD systems can allow radiologists to quantitatively analyze imaging data but work on the basis of differing algorithms for feature extraction, selection, and integration, affecting the differentiating abilities of each CAD system. Moreover, complicated and laborious processes are needed to improve the performance of CAD systems. On the other hand, particularly in terms of image recognition, DL systems based on the CNN structure with convolution and max-pooling layers can provide

Table 3
Diagnostic performance among observers: sensitivity and specificity.

Observer		Point estimate	95% confidence interval
R1	Sensitivity	0.955	0.873, 0.991
	Specificity	0.375	0.188, 0.594
R2	Sensitivity	0.939	0.852, 0.983
	Specificity	0.208	0.071, 0.422
R3	Sensitivity	0.970	0.895, 0.996
	Specificity	0.458	0.256, 0.672
DL	Sensitivity	*0.758	0.636, 0.855
	Specificity	#0.667	0.447, 0.844
The consensus result of radiologists	Sensitivity	*0.970	0.895, 0.996
	Specificity	#0.292	0.126, 0.511

DL=deep learning, R1=radiologist with 8 years of experience, R2=radiologist with 9 years of experience, R3=radiologist with 26 years of experience.

* The sensitivity of DL was significantly inferior to the consensus result of radiologists ($P=.0005$).

The specificity of DL was significantly superior to the consensus result of radiologists ($P=.02$).

simple and alternative results (i.e., invasiveness or non-invasiveness) without using specific algorithms. In the present study, our DL system easily achieved an accuracy rate almost identical to that of radiologists in differentiating not only between AIS, MIA, and IVA but also between AIS and MIA/IVA. Although no significant differences in pathological diagnostic accuracy rates were seen among all observers (DL, R1, R2, and R3), high dimensional image data from pixel data to whole-image data could be processed within a few seconds simply by inputting image data into the DL system. Unlike radiologists, the extreme and growing processing power represents a key advantage of DL systems.

In diagnostic performance to differentiate between AIS and MIA/IVA, the AUC for the DL system was almost the same as that for the most experienced radiologist, who in turn displayed a significantly higher AUC than the radiologist with the least amount of experience. In previous studies into diagnoses for lung cancer,^[15,16] CAD systems using DL techniques were demonstrated to offer superior diagnostic accuracy of lung cancer and superior feature extraction for diagnosing pulmonary nodule compared to conventional CAD. In the future, malignancy diagnosis and prognostic prediction using DL systems may be incorporated into our clinical setting. In our study, however, compared with radiologists, DL showed significantly inferior sensitivity, but significantly superior specificity. DL has the possibility of correctly identifying adenocarcinoma without pathological invasiveness. This result may be due to the fact that only DL could accurately predict pathological diagnoses that were sometimes difficult for radiologists to differentiate. Higher specificity of DL might suppress overdiagnoses of radiologists, resulting in a positive effect on the management or treatment strategies. Identifying subgroups of patients without pathological invasiveness might be helpful in selecting patients suitable for watchful waiting. Generally, DL systems are functionally black boxes, meaning that the process by which a DL reaches a conclusion is unknown. One possible reason was that given the small number of cases, the cancer characteristics used as training data happened to be consistent with those only DL can capture. In other words, DL may not have always captured the essential features of cancer in this study. Further analysis using a larger cohort is needed to validate our results.

DL systems will provide useful and informative results, but may not provide a useful alternative for clinicians. DL systems do not have the ability to manage and decide treatment strategies, but it is very important for radiologists to be able to beneficially utilize the information from DL systems. The combination of radiologists and DL should preferably be superior to DL alone. In fact, DL systems can provide useful results for nodule detection,^[32] and a deep 3D CNN could achieve high nodule detection sensitivity even at 0.25 false-positive results per scan. However, differentiation between benign and malignant lesions may not necessarily be satisfactory results. Basically, radiologists often diagnose pulmonary nodules by morphologically evaluating the margins and internal characteristics according to previous data.^[24–31] Naturally, limitations to diagnostic performance exist. For example, localized GGN included all pathological subtypes of adenocarcinoma.^[33] However, some radiologists might not be able to accurately differentiate IVA from GGN on CT images. Therefore, unlike nodule detection, in differentiating between benignancy and malignancy, some cases might be encountered in which a DL system identifies malignant lesions that the radiologist cannot believe or verify. Much higher accuracy, sensitivity, and specificity are thus needed for DL to contribute to the diagnosis of malignancy.

Our study shows several limitations. First, the biggest limitation is the small dataset. Overfitting is thus one of the disadvantages to DL, which may be more prominent in cases with a small number of training samples. Our DL system fit the present study data well but might not generalize well to unseen cases. Further studies using a larger training data set are needed to validate our results. Second, our study was retrospective, and selection bias was thus inevitable. Third, only primary sites were evaluated. If data for lymph node metastases and lung tissues surrounding primary sites had been used for the construction of DL systems, diagnostic performance might have been influenced. Analysis of additional training cases would have been interesting. Finally, 3D data from CT images were analyzed using the DL system. However, not all cross-sections from each pathological specimen could be evaluated in the present study. Therefore, invasive components in pathological specimens may not necessarily have represented the true invasive component itself.

In conclusion, despite the small training data set, DL showed an accuracy rate almost equal to that of radiologists. The AUC of DL was almost the same as that of the radiologist with the most experience, which in turn was significantly higher than that of the radiologist with the least experience. DL systems can predict pathological invasiveness in lung adenocarcinoma from CT images, particularly with high specificity. DL can provide useful and informative results but does not replace the radiologist. Likewise, DL does not have the ability to manage or decide treatment strategies, so radiologists need to intelligently use the information derived from DL. We expect that radiologists will play important roles in developing and using artificial intelligence technologies.

Author contributions

Conceptualization: Masahiro Yanagawa, Hirohiko Niioka.

Data curation: Masahiro Yanagawa, Akinori Hata, Noriko Kikuchi, Osamu Honda.

Formal analysis: Masahiro Yanagawa, Hirohiko Niioka, Hiroyuki Kurakami, Eiichi Morii, Masayuki Noguchi.

Investigation: Masahiro Yanagawa, Hirohiko Niioka, Akinori Hata, Noriko Kikuchi, Osamu Honda, Hiroyuki Kurakami, Eiichi Morii, Masayuki Noguchi, Yoshiyuki Watanabe.

Methodology: Masahiro Yanagawa, Hirohiko Niioka, Hiroyuki Kurakami, Yoshiyuki Watanabe, Jun Miyake, Noriyuki Tomiyama.

Project administration: Masahiro Yanagawa.

Software: Masahiro Yanagawa, Hirohiko Niioka.

Supervision: Masahiro Yanagawa, Eiichi Morii, Masayuki Noguchi, Yoshiyuki Watanabe, Jun Miyake, Noriyuki Tomiyama.

Validation: Masahiro Yanagawa, Hirohiko Niioka.

Visualization: Masahiro Yanagawa, Hirohiko Niioka, Eiichi Morii, Masayuki Noguchi.

Writing – original draft: Masahiro Yanagawa.

Writing – review & editing: Jun Miyake, Noriyuki Tomiyama.

References

- Yankelevitz DF, Reeves AP, Kostis WJ, et al. Small pulmonary nodules: volumetrically determined growth rates based on CT evaluation. *Radiology* 2000;217:251–6.
- Wormanns D, Kohl G, Klotz E, et al. Volumetric measurements of pulmonary nodules at multi-row detector CT: in vivo reproducibility. *Eur Radiol* 2004;14:86–92.
- Goo JM. A computer-aided diagnosis for evaluating lung nodules on chest CT: the current status and perspective. *Korean J Radiol* 2011;12:145–55.
- Yanagawa M, Tanaka Y, Kusumoto M, et al. Automated assessment of malignant degree of small peripheral adenocarcinomas using volumetric CT data: correlation with pathologic prognostic factors. *Lung Cancer* 2010;70:286–94.
- de Hoop B, Gietema H, van de Vorst S, et al. Pulmonary ground-glass nodules: increase in mass as an early indicator of growth. *Radiology* 2010;255:199–206.
- Naidich DP, Bankier AA, MacMahon H, et al. Recommendations for the management of subsolid pulmonary nodules detected at CT: a statement from the Fleischner Society. *Radiology* 2013;266:304–17.
- Yanagawa M, Tanaka Y, Leung AN, et al. Prognostic importance of volumetric measurements in stage I lung adenocarcinoma. *Radiology* 2014;272:557–67.
- Colombi D, Manna C, Montermini I, et al. Semiautomatic analysis on computed tomography in locally advanced or metastatic non-small cell lung cancer: reproducibility and prognostic significance of unidimensional and 3-dimensional measurements. *J Thorac Imaging* 2015;30:290–9.
- Way TW, Sahiner B, Chan HP, et al. Computer-aided diagnosis of pulmonary nodules on CT scans: improvement of classification performance with nodule surface features. *Med Phys* 2009;36:3086–98.
- Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge. *Int J Comput Vis* 2015;115:211–52.
- Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542:115–8.
- Keremany DS, Goldbaum M, Cai W, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* 2018;172:1122–31.
- Bejnordi BE, Beta M, Diest PJ, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 2017;318:2199–210.
- Niioka H, Asatani S, Yoshimura A, et al. Classification of C2C12 cells at differentiation by convolutional neural network of deep learning using phase contrast images. *Hum Cell* 2018;31:87–93.
- Sun W, Zheng B, Qian W. Automatic feature learning using multichannel ROI based on deep structured algorithms for computerized lung cancer diagnosis. *Comput Biol Med* 2017;289:530–9.
- Hua KL, Hsu CH, Hidayati SC, et al. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. *Oncotargets Ther* 2015;8:2015–22.
- Labaki WW, Han MK. Artificial intelligence and chest imaging. Will deep learning make us smarter. *Am J Respir Crit Care Med* 2018;197:148–50.
- Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* 2017;284:574–82.
- Bar Y, Diamant I, Wolf L, Greenspan H. Deep learning with non-medical training used for chest pathology identification. In: Hadjiiski LM, Tourassi GD, eds. Proceedings of SPIE: medical imaging 2015–computer-aided diagnosis. Vol 9414. Bellingham, Wash: International Society for Optics and Photonics, 2015; 94140 V.
- Shin HC, Roth HR, Gao M, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans Med Imaging* 2016;35:1285–98.
- Maeshima AM, Tochigi N, Yoshida A, et al. Histological scoring for small lung adenocarcinomas 2 cm or less in diameter: a reliable prognostic indicator. *J Thorac Oncol* 2010;5:333–9.
- Tsutani Y, Miyata Y, Mimae T, et al. The prognostic role of pathologic invasive component size, excluding lepidic growth, in stage I lung adenocarcinoma. *J Thorac Cardiovasc Surg* 2013;146:580–5.
- Travis WD, Brambilla E, Nicholson AG, et al. The 2015 World Health Organization classification of lung tumors: impact of genetic, clinical and radiologic advances since the 2004 classification. *J Thorac Oncol* 2015;10:1243–60.
- Stroom J, Blaauwgeers H, van Baardwijk A, et al. Feasibility of pathology-correlated lung imaging for accurate target definition of lung tumors. *Int J Radiat Oncol Biol Phys* 2007;69:267–75.
- Macpherson RE, Higgins GS, Murchison JT, et al. Non-small-cell lung cancer dimensions: CT-pathological correlation and interobserver variation. *Br J Radiol* 2009;82:421–5.
- Lee SM, Goo JM, Lee KH, et al. CT findings of minimally invasive adenocarcinoma (MIA) of the lung and comparison of solid portion measurement methods at CT in 52 patients. *Eur Radiol* 2015;25:2318–25.
- Yanagawa M, Johkoh T, Noguchi M, et al. Radiological prediction of tumor invasiveness of lung adenocarcinoma on thin-section CT. *Medicine (Baltimore)* 2017;96:e6331.
- Lim HJ, Ahn S, Lee KS, et al. Persistent pure ground-glass opacity lung nodules ≥ 10 mm in diameter at CT scan: histopathologic comparisons and prognostic implications. *Chest* 2013;144:1291–9.
- Lampen-Sachar K, Zhao B, Zheng J, et al. Correlation between tumor measurement on computed tomography and resected specimen size in lung adenocarcinomas. *Lung Cancer* 2012;75:332–5.
- Kakinuma R, Noguchi M, Ashizawa K, et al. Natural history of pulmonary subsolid nodules: a prospective multicenter study. *J Thorac Oncol* 2016;11:1012–28.
- Nakamura H, Saji H, Ogata A, et al. Lung cancer patients showing pure ground-glass opacity on computed tomography are good candidates for wedge resection. *Lung Cancer* 2004;44:61–8.
- Jin H, Li Z, Tong R, et al. A deep 3D residual CNN for false-positive reduction in pulmonary nodule detection. *Med Phys* 2018;45:2097–107.
- She Y, Zhao L, Dai C, et al. Preoperative nomogram for identifying invasive pulmonary adenocarcinoma in patients with pure ground-glass nodule: a multi-institutional study. *Oncotarget* 2017;8:17229–38.