Hiroshima J. Med. Sci.
Vol. 71, No. 1·2, 9~22, June, 2022
HIMJ 71–2

Content-based CT image retrieval system using deep learning: Preliminary assessment of its accuracy for classifying lesion patterns and retrieving similar cases among patients with diffuse lung diseases

Hiroaki TERADA[1)], Toru HIGAKI[1)], Hiroaki TAKEBE[2)], Takayuki BABA[2)], Nobuhiro MIYAZAKI[2)], Yasutaka MORIWAKI[2)], Hiroaki SAKANE[1)], Wataru FUKUMOTO[1)], and Kazuo AWAI[1,*)]

1) Diagnostic Radiology, Hiroshima University, 1-2-3 Kasumi, Minami-ku, Hiroshima 734-8551, Japan
2) FUJITSU LABORATORIES LTD., 4-1-1 Kamikodanaka, Nakahara-ku, Kawasaki 211-8588, Japan

ABSTRACT

Practical image retrieval systems must fully use image databases. We investigated the accuracy of our content-based computer tomography (CT) image retrieval system (CB-CTIRS) for classifying lesion patterns and retrieving similar cases in patients with diffuse lung diseases. The study included 503 individuals, with 328 having diffuse lung disease and 175 having normal chest CT scans. Among the former, we randomly selected ten scans that revealed one of five specific patterns [consolidation, ground-glass opacity (GGO), emphysema, honeycombing, or micronodules: two cases each]. Two radiologists separated the squares into six categories (five abnormal patterns and one normal pattern) to create a reference standard. Subsequently, each square was entered into the CB-CTIRS, and the F-score used to classify squares was determined. Next, we selected 15 cases (three per pattern) among the 503 cases, which served as the query cases. Three other radiologists graded the similarity between the retrieved and query cases using a 5-point grading system, where grade 5 = similar in both the opacity pattern and distribution and 1 = different therein. The F-score was 0.71 for consolidation, 0.63 for GGO, 0.74 for emphysema, 0.61 for honeycombing, 0.15 for micronodules, and 0.67 for normal lung. All three radiologists assigned grade 4 or 5 to 67.7% of retrieved cases with consolidation, emphysema, or honeycombing, and grade 2 or 3 to 67.7% of the retrieved cases with GGO or micronodules. The retrieval accuracy of CB-CTIRS is satisfactory for consolidation, emphysema, and honeycombing but not for GGO or micronodules.

Key words: *content-based image retrieval, diffuse lung disease, chest CT*

INTRODUCTION

Annually, an estimated 62 million patients have undergone diagnostic computer tomography (CT) scans in the United States[1)] and massive imaging data are stored in the picture archiving and communication systems (PACS) of hospitals. A national-level database that integrates the PACS of many hospitals throughout the country was proposed in Japan[2)]. While the stored image data of individual patients are used for follow-up and treatment planning, data of individuals other than the patient are used only for educational or research purposes rather than for diagnosing a wider spectrum of patients. This may be because there is no practical application that can perform content-based image retrieval systems (CBIRS) on the existing PACS.

When diagnosing a case in which a specific or differential diagnosis is elusive, diagnostic radiologists consult their personal collections or search the literature for similar cases with definitive diagnoses. The retrieved similar images in linked clinical and pathological databases may help obtain a potential diagnosis or clinically useful information[3)].

CBIRS has been studied in various disease groups such as diffuse lung disease (DLD)[3–13)]. DLD consists of a wide spectrum of diseases, each of which manifests itself as a combination of various lung opacities (lung pattern)[14)]. Therefore, we believe that CBIRS for DLDs must achieve accurate classification of each lesion pattern and retrieval of similar cases as a prerequisite for its diagnostic usefulness. However, few studies have investigated the classification and retrieval accuracy of the CBIRS for each lung pattern[8,13)]. We developed a CBIRS for CT scans (CB-CTIRS) using a deep-learning technique targeted at DLDs and evaluated its accuracy in classifying lesion patterns and its ability to retrieve similar cases.

* Corresponding author: Kazuo Awai
Diagnostic Radiology, Hiroshima University, 1-2-3 Kasumi, Minami-ku, Hiroshima 734-8551, Japan
Tel: +81-82-257-5257, Fax: +81-82-257-5259, E-mail: awai@hiroshima-u.ac.jp

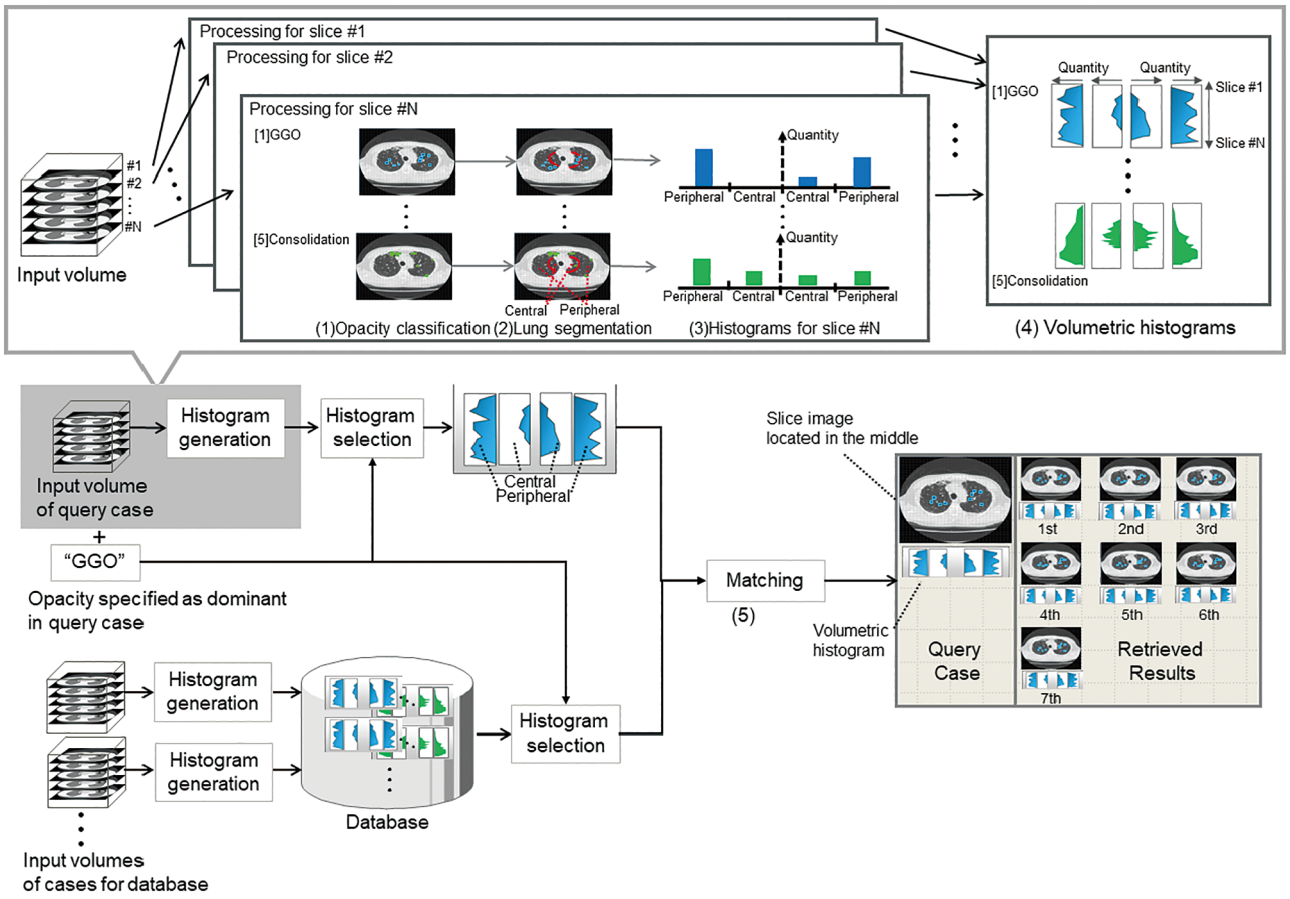

Figure 1 Our content-based CT image retrieval system: (1) The lung fields on CT images are extracted and divided into 1-cm squares on a grid, and each square is classified as exhibiting one of five abnormal patterns or as a normal lung with the aid of a trained deep convolutional neural network (DCNN). (2) Identifies each square as residing in central- or peripheral zones in the lung field. (3) Generate a histogram showing the number of squares classified as a specific pattern type and their location (central or peripheral zone, left or right lung) in each image. (4) Generates histograms of 3D volume data including the whole lung. (5) Calculate the similarity between the query case histograms and histograms of all cases in the database using a histogram intersection algorithm. The retrieved database cases are presented in the descending order of similarity.

MATERIALS AND METHODS

Outline of our content-based CT image retrieval system (CB-CTIRS)

Here, we present an outline of our CB-CTIRS, and additional technical details can be found in the Supplementary Materials.

With the CB-CTIRS, lung fields on the CT image were extracted, grid divided into 1-cm squares, and each square was classified into five abnormal pattern categories (consolidation, ground-glass opacity, micronodules, emphysema, and honeycombing) or as a normal lung using a trained deep convolutional neural network (DCNN) [(1) in Figure 1]. The five types of patterns are defined in the glossary of terms for thoracic imaging[15]. Details of the extraction and segmentation of the lung fields and the training process for the DCNN are described in the Supplementary Materials.

CB-CTIRS generates a histogram showing the location of a particular segment (central or peripheral zone, left or right lungs [(3) in Figure 1] after segmenting the lung fields into central and peripheral zones [(2) in Figure 1]. It also creates a histogram showing the 3D volumetric histogram of the whole lung [(4) in Figure 1].

In the retrieval process, the CB-CTIRS matches one abnormal pattern (consolidation, ground-glass opacity, emphysema, honeycombing, or micronodules) that was manually specified by a radiologist in the query case with the histogram of the corresponding abnormal pattern contained in the database.

The time required for the retrieval of target cases from 1000 CT scans is 0.2 s (CPU: Xeon(R) CPU E3-1275 v6 [3.80 GHz], Memory:16 GByte, GPU:Quadro P4000).

Study population

This retrospective study was approved by our institutional review board, and informed patient consent was waived because we used existing CT images in this study.

Between November 2017 and June 2018, we obtained chest CT scans from 10,563 patients with suspected lung disease or who were being followed up for existing lung lesions. When patients underwent multiple chest CT studies during the study period, we included only the first CT scan. Consequently, 2,412 scans were available for this study. From these, we excluded 1,909 scans from patients with intrathoracic malignant tumours (n = 1,885), those with poor image quality owing to inadequate breath-holding (n = 11), and those from patients

Table 1 Clinical diagnoses of 328 patients with abnormal chest CT findings

Clinical Diagnosis	Patient number
Idiopathic pulmonary fibrosis (IPF)	57
Non-specific interstitial pneumonia (NSIP)	54
Cryptogenic organic pneumonia (COP)	15
Non-tuberculous mycobacterial infection	44
Pulmonary emphysema	39
Sarcoidosis	31
Pneumonia	16
Post inflammatory change	10
Bronchiectatis	10
Lymphangioleiomyomatosis	5
Pneumoconiosis	4
Granulomatosis with polyangiitis	4
Combined pulmonary fibrosis and emphysema	4
Old tuberculosis	3
Chronic hypersensitivity pneumonia	3
Chronic bronchitis	3
Pulmonary edema	2
Pulmonary cryptococcosis	2
Secondary organizing pneumonia	2
Hypersensitivity pneumonia	2
Chronic eosinophilic pneumonia	2
ANCA associated vasculitis	2
Other diseases*	14
Total	328

*Other diseases indicate disease which only one patient had shown.
Abbreviation
ANCA: antineutrophilic cytoplasmic antibody

whose unilateral or bilateral lung volume was markedly decreased because of massive amounts of pleural fluid (n = 13). Therefore, 503 patients (257 men and 246 women) were finally used in the study. The median age of the 503 patients was 66 years (range: 23–88 years). CT yielded normal findings in 175 patients and abnormal results in 328 patients (Table 1).

Chest CT scans

All scans were performed on one of five scanners (Aquilion One, Canon Medical Systems; Aquilion One Genesis, Canon Medical Systems; Aquilion Precision, Canon Medical Systems; LightSpeed 64 VCT, GE Healthcare; and Revolution CT, GE Healthcare). The scanning protocol for the Aquilion instruments had the following parameters: 0.5 × 80 mm detector configuration, tube rotation time of 0.50 s, pitch factor of 1.388, scanning field of view (FOV) ranging between 30 cmand 45 cm, a voltage of 120 kV, image noise of 12 with automatic tube current modulation (ATCM), reconstruction "FC52" with AIDR 3D weak, and slice thickness and interval of 1.00 mm. For the LightSpeed 64 VCT, the detector configuration was 0.625 × 64 mm, tube rotation time was 0.50 s, pitch factor was 1.375, scanning FOV was ranging between 30 cm and 45 cm, a voltage of 120 kV, image noise of ten with ATCM, reconstruction kernel "lung" with filtered back projection, and slice thickness and interval of 1.25 mm. For the Revolution CT scanner, it was single-energy scan mode, detector with 0.625 × 128 mm configuration, tube rotation time of 0.50 s, 0.992 pitch factor, scanning FOV between 30 cm and 45 cm, voltage of 120 kV, image noise of ten with ATCM, reconstruction kernel "lung" with ASiR-V of 30%, and slice thickness and interval of 1.25 mm. Contrast enhancement was required for each patient.

Reference standard of dominant lung patterns for the 503 test cases

Two board-certified radiologists (#1 and #2) with 32 and 14 years of experience reading chest CT scans, respectively, subjectively and consensually determined the dominant pattern [consolidation, ground-glass opacity (GGO), emphysema, honeycombing, or micronodules] and recorded the 1st- and 2nd-largest pattern volume as the 1st and 2nd dominant patterns in each patient. We served the 1st dominant pattern determined by the two radiologists as the reference standard for the lung pattern for each case. We included five lung pattern categories, and their definitions were established by others[15]. Another representative pattern other than the above five patterns seen in the DLDs is reticular opacity. Regarding reticular opacity on CT, some refer to thickening of the interlobular septal wall[15,16], while others refer to the lobular inner line seen in idiopathic pulmonary fibrosis;[17] however, its appearance on CT scans is different. Because it lacks a clear definition and no scans showed reticular opacity dominance, we excluded it in our study.

Among the 328 CT scans with abnormal findings, radiologists 1 and 2 identified the 1st dominant pattern consolidation (n = 7), GGO (n = 103), emphysema (n = 93), honeycombing (n = 77), and micronodules (n = 48). The 2nd dominant pattern was consolidation (n = 14), GGO (n = 44), emphysema (n = 47), honeycombing (n = 136), and micronodules (n = 87).

Accuracy of our CB-CTIRS in the classification of lung patterns

First, we evaluated the accuracy of our CB-CTIRS for classifying lung patterns on horizontal slices using a 1-cm sized square by square reference standard determined by radiologists (as described below).

Using a random table, we selected two cases each that revealed consolidation, GGO, emphysema, honeycombing, or micronodules as the 1st dominant pattern from 328 cases with abnormal patterns. Subsequently, for each case, we selected approximately 20 horizontal slices with equal slice intervals, ranging from the level of the upper end of the aortic arch to the level just above the diaphragm. The lung field on each slice was divided into 1-cm squares (Figure 2), and the two radiologists (#1 and #2) independently and subjectively classified all squares as consolidation, GGO, emphysema, honeycombing, or micronodules or recorded them as normal or undetermined. Disagreements were resolved through consensus. The CB-CTIRS then classified each square into one of the five patterns or as normal.

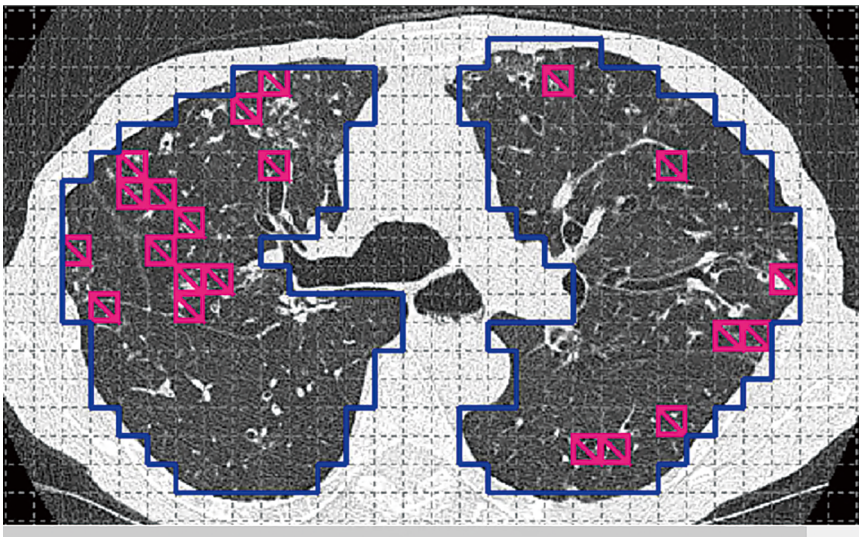

Figure 2 The extracted lung field is covered with a grid of 1-cm squares. The area inside the blue line is the lung field area extracted by the content-based CT image retrieval system (CB-CTIRS). Red squares were identified by the system as micronodules.

Squares recorded as undetermined by the radiologists were excluded from classification by the CB-CTIRS.

Finally, the precision (positive predictive value), recall (sensitivity), and F-score (harmonic mean of recall and precision) of the CB-CTIRS for classifying lung patterns were evaluated. The precision and recall are defined as follows[18]:

Precision = (*True positive*)/(*True Positive* + *False Positive*)

Recall = (*True positive*)/(*True Positive* + *False negative*)

The F-score was defined as follows:

F-score = (2 × *Recall* × *Precision*)/(*Recall* + *Precision*)

To analyse the relationship between the pattern classification recorded by radiologists 1 and 2 (the reference standard) and the classification yielded by our CB-CTIRS, we created a confusion matrix in which each row lists the number of squares with a specific pattern in the reference standard, and each column lists the number of squares classified as a specific pattern by our CB-CTIRS.

Visual analysis of the retrieval accuracy of our CB-CTIRS

Second, we evaluated the accuracy of our CB-CTIRS in retrieving morphologically similar cases to a query case. Three additional board-certified radiologists (#3, # 4, and # 5) with 18, 12, and 9 years of experience interpreting CT scans assessed the retrieval accuracy of our CB-CTIRS, respectively. They were not involved in the development of a503 case reference standard . From the 328 cases with abnormal patterns, we randomly selected three query cases that revealed consolidation, GGO, emphysema, honeycombing, or micronodules (five patterns × three query cases = 15 query cases). Then, using the CB-CTIRS, for each of the 15 query cases, we retrieved three morphologically similar cases from the remaining 502 (503-1) cases. Thus, 45 retrieved cases [(five patterns × three query cases × three retrieved cases = 45 cases] were submitted to the three radiologists for visual evaluation . We selected three query cases for each pattern based on the hypothesis that their use would reveal the performance trend of our CB-CTIRS for that pattern . In the retrieval process by the CB-CTIRS, one radiologist (radiologist #1) who was not involved in the visual analysis of the retrieval performance of the CB-CTIRS, specified the target abnormal pattern to match in each query case. When a query case revealed more than one abnormal pattern, we focused on a single (specific) pattern in the query case to evaluate the pattern similarity in the retrieved cases.

The three radiologists (#3, #4, and #5) assigned Grade 5 when the retrieved and query cases were morphologically similar with respect to the pattern and distribution, Grade 4 when they were fairly similar, Grade 3 when the dominant pattern was very similar to the two images but its distribution was dissimilar, Grade 2 when the pattern was fairly similar but the lesion distribution was different, and Grade 1 when both the pattern and distribution differed.

Cohen's kappa coefficient of concordance was used to determine interobserver agreement among the three radiologists. Kappa values between 0 and 0.20 indicated poor-, those between 0.21 and 0.40 indicated fair-, those between 0.41 and 0.60 indicated moderate-, those between 0.61 and 0.80 indicated good-, and those greater or equal to 0.81 indicated excellent agreement[19].

RESULTS

Classifying lung pattern with the CB-CTIRS

The division of lung slices into 1-cm squares resulted in a total of 14,925 squares. After excluding 235 squares from ten cases that had been recorded as undetermined , radiologists #1 and #2 consensually classified the remaining 14,690 squares as consolidation (n = 922), GGO (n = 1,670), emphysema (n = 3,672), honeycombing (n = 1,255), or micronodules (n = 1,368); 5,803 squares were classified as normal lung.

The recall, precision, and F-score obtained using the

CB-CTIRS are listed in Table 2. The F-score for consolidation and emphysema was the highest (> 0.70); it was the lowest for micronodules (0.15). The F-score for GGO, honeycombing, and normal lungs ranged from 0.61 to 0.67.

Analysis of the confusion matrix (Table 3) showed that 1,624 of 5,803 squares (28.0%) designated normal lungs in the reference standard were misclassified as abnormal patterns by the CB-CTIRS. These were misclassified as consolidation (n = 162, 2.8%), GGO (n = 401, 6.9%), emphysema (n = 400, 6.9%), honeycombing (n = 525, 9.0%), and micronodules (n = 136, 2.3%). However, 2,570 squares were misclassified as normal lungs by the CB-CTIRS, 72 (1.1%) as consolidation, 365 (5.4%) as GGO, 911 (13.5%) as emphysema, and 1,169 (17.3%) as micronodules.

Retrieval performance of the CB-CTIRS

The results of the three radiologists (#3, #4, and #5) are shown in Table 4(A)–4(C).

Of the nine candidate cases for consolidation retrieved by the CB-CTIRS, radiologist #3 evaluated eight of nine cases to be Grade 4 or 5. Likewise, radiologists #4 and #5 evaluated seven and nine cases, respectively, to be Grade 4 or 5 in nine candidate cases for consolidation. Of the nine candidate cases for emphysema retrieved by the CB-CTIRS, the three radiologists evaluated all nine cases to be Grade 4 or 5. Of the nine candidate cases for honeycomb, the three radiologists also evaluated all nine cases to be Grade 4 or 5 in visual assessment. Of the nine candidate cases for GGO, radiologists #3, #4, and #5 evaluated three, two, and zero cases, respectively, as Grade 4 or 5. Only one of the candidates for micronodules was assigned Grade 4 or 5 by each of the three radiologists. The kappa coefficients for interobserver agreement were 0.92 [95% confidence interval (CI], 0.86–0.97) for observers 1 and 2, 0.78 (95% CI, 0.64–0.91) for observers 1 and 3; and 0.77 (95% CI, 0.62–0.93) for observers 2 and 3. Thus, the interobserver agreement was good or excellent.

We presented three representative cases (Figures 3, 4. and 5).

Table 2 Recall-, precision-, and F-scores of our content-based CT image retrieval system for classifying lung patterns

Opacity	Recall	Precision	F-score
Consolidation	0.75	0.67	0.71
Ground-glass opacity	0.63	0.62	0.63
Emphysema	0.66	0.83	0.74
Honeycombing	0.78	0.50	0.61
Micronodule	0.09	0.39	0.15
Normal lung	0.72	0.62	0.67
Average	0.61	0.61	0.59

DISCUSSION

In the evaluation of the accuracy for classifying patterns of our CB-CTIRS, the recall values tended to be low even when the CB-CTIRS classification and the reference standard were in relatively good agreement. Except for micronodules, which were above 0.63, that is, with respect to most squares, the CB-CTIRS pattern classification and the readers' reference standard was in agreement. However, the pattern classification ability of CB-CTIRS for micronodules was unsatisfactory. The poor ability to classify micronodules may be attributed to the small number of cases with micronodules (n = 10, Table 2 in the Supplementary Materials) in the training process for our CB-CIRS. Micronodules may be of high or low density, their size ranges from less than 1 mm to 2 mm, and their distribution in the lung depends on the disease; some are centrilobularly, while others are randomly distributed[15]. To improve the system's classification ability for micronodules, the number of training cases must be increased, and a wider spectrum of characteristics must be included.

Analysis of the CB-CTIRS confusion matrix revealed that many normal lung squares in the reference standard were misidentified as GGO or emphysema by the CB-CTIRS, while many GGO and emphysema squares in the reference standard were miscategorised as normal lung by the CB-CTIRS. We suspect that this is attributable to the low contrast between these lesions and the normal lung tissue. We believe that classification performance can also be improved by increasing the number of training cases for cases with GGO or emphysema.

Visual assessment by the three radiologists showed

Table 3 Confusion matrix of the classification of lung patterns yielded by the content-based CT image retrieval system

		Classification by the CT image retrieval system					
		Consolidation	Ground-glass opacity	Emphysema	Honeycombing	Micronodule	Normal lung
Reference Standard	Consolidation	692 (75%)	83 (9%)	0 (0%)	75 (8%)	0 (0%)	72 (8%)
	Ground-glass opacity	116 (7%)	1057 (63%)	4 (0%)	110 (7%)	18 (1%)	365 (22%)
	Emphysema	6 (0%)	397 (1%)	2434 (66%)	244 (7%)	38 (1%)	911 (25%)
	Honeycombing	50 (4%)	917 (7%)	75 (6%)	978 (78%)	8 (1%)	53 (4%)
	Micronodule	3 (0%)	327 (2%)	17 (1%)	21 (2%)	126 (9%)	1169 (85%)
	Normal lung	162 (3%)	4017 (7%)	400 (7%)	525 (9%)	136 (2%)	4179 (72%)

Number indicates number of squares.
Number in parentheses indicates [number of squares determined to be a certain pattern by the CB-CTIRS]/[total number of squares of that pattern in the reference standard] × 100.

Tables 4 A–C Performance of the content-based CT image retrieval system evaluated by 3 radiologists

Table 4A Radiologist #3

Query Case No.	Query case pattern	Visual score of retrieved case		
		1st-most similar	2nd-most similar	3rd-most similar
1	Consolidation	4	5	4
2	Consolidation	4	3	4
3	Consolidation	5	4	4
4	Ground glass opacity	2	3	4
5	Ground glass opacity	3	5	2
6	Ground glass opacity	5	2	2
7	Emphysema	4	5	4
8	Emphysema	4	4	4
9	Emphysema	4	5	4
10	Honeycombing	5	5	5
11	Honeycombing	4	5	5
12	Honeycombing	5	4	5
13	Micronodule	2	2	5
14	Micronodule	2	3	2
15	Micronodule	3	2	3

Table 4B Radiologist #4

Query Case No.	Query case pattern	Visual score of retrieved case		
		1st- most similar	2nd-most similar	3rd-most similar
1	Consolidation	4	5	4
2	Consolidation	4	3	4
3	Consolidation	4	4	3
4	Ground glass opacity	2	3	3
5	Ground glass opacity	2	4	2
6	Ground glass opacity	4	2	2
7	Emphysema	4	5	4
8	Emphysema	4	4	4
9	Emphysema	4	5	4
10	Honeycombing	5	4	5
11	Honeycombing	4	4	5
12	Honeycombing	5	4	5
13	Micronodule	2	2	5
14	Micronodule	2	2	2
15	Micronodule	3	2	3

Table 4C Radiologist #5

Query Case No.	Query case pattern	Visual score of retrieved case		
		1st-most similar	2nd-most similar	3rd-most similar
1	Consolidation	4	5	4
2	Consolidation	4	4	4
3	Consolidation	4	4	5
4	Ground glass opacity	2	3	3
5	Ground glass opacity	2	3	2
6	Ground glass opacity	3	2	3
7	Emphysema	4	5	4
8	Emphysema	4	4	4
9	Emphysema	4	4	4
10	Honeycombing	5	4	5
11	Honeycombing	4	4	5
12	Honeycombing	4	4	5
13	Micronodule	3	3	4
14	Micronodule	2	3	2
15	Micronodule	3	2	3

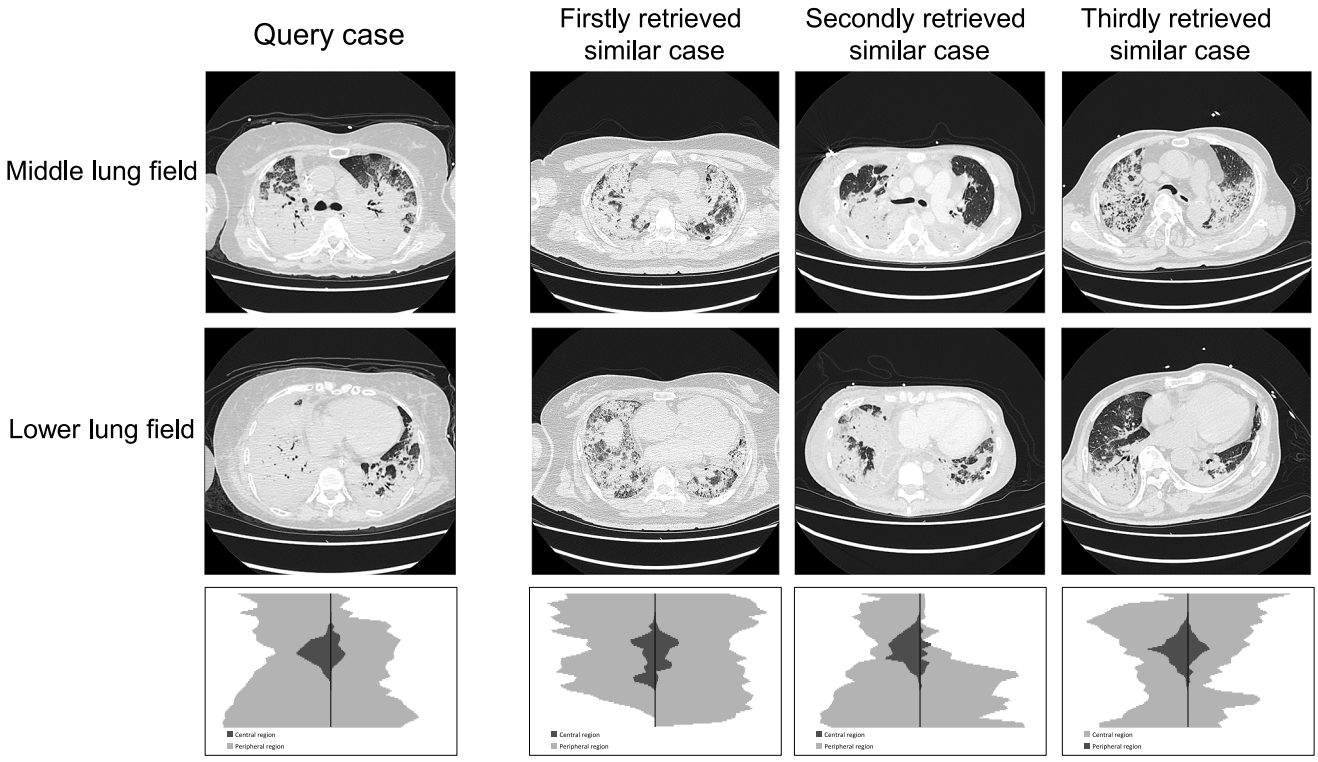

Figure 3 Representative Case 1: The query case is a patient with severe pneumonia showing extensive consolidation throughout the lungs. The leftmost column shows the middle and lower lung fields of the query case. The 2nd-, 3rd-, and 4th columns are scans with, in order, the highest to lower degrees of similarity.

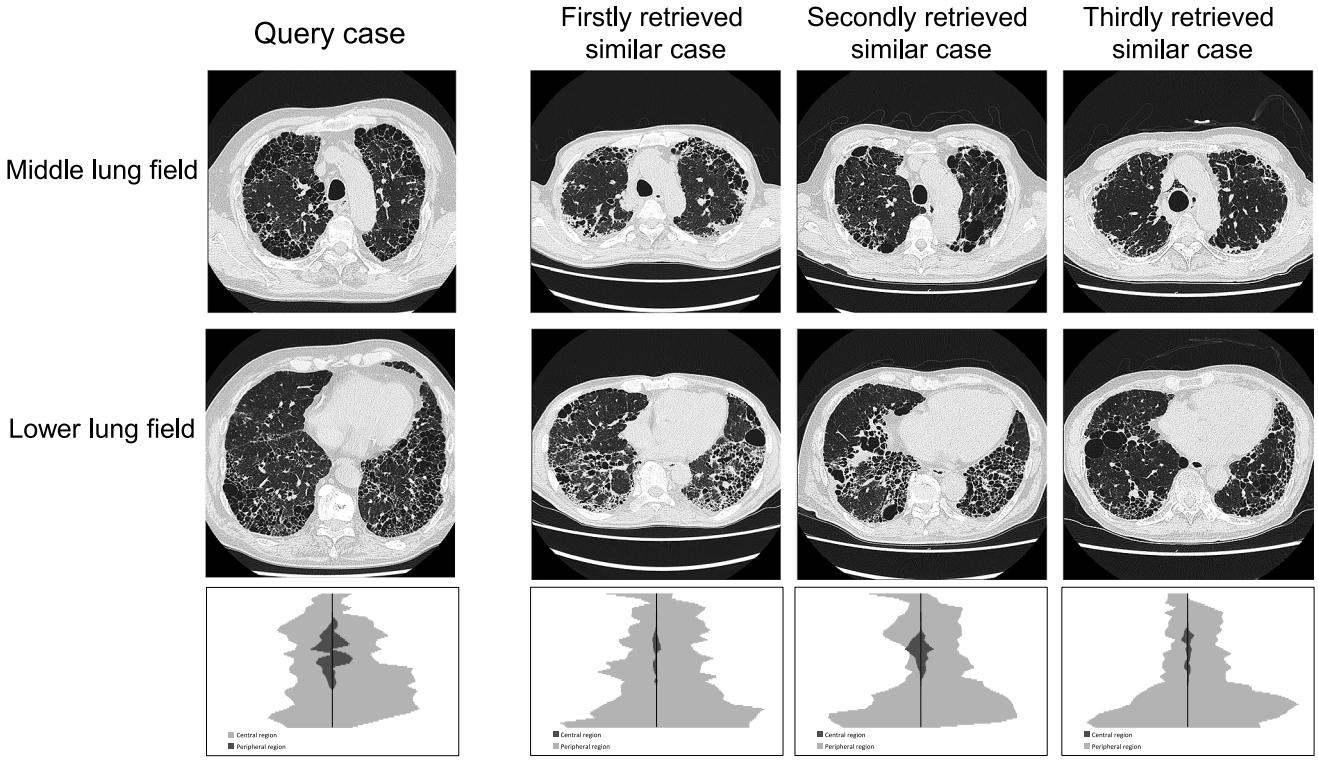

Figure 4 Representative Case 2: The query case is a patient with interstitial pneumonia with honeycombing of typical distribution. The leftmost column shows the middle and lower lung fields of the query case. The 2nd-, 3rd-, and 4th columns are scans with, in order, the highest to lower degrees of similarity.

that the retrieval performance of the CB-CTIRS was relatively good for consolidation, emphysema, and honeycombing. In contrast, the retrieval performance was unsatisfactory for micronodules and GGOs. We believe that poor retrieval performance for these patterns may result from poor classification ability for micronodules and GGO, and improvement in classification ability may improve retrieval ability.

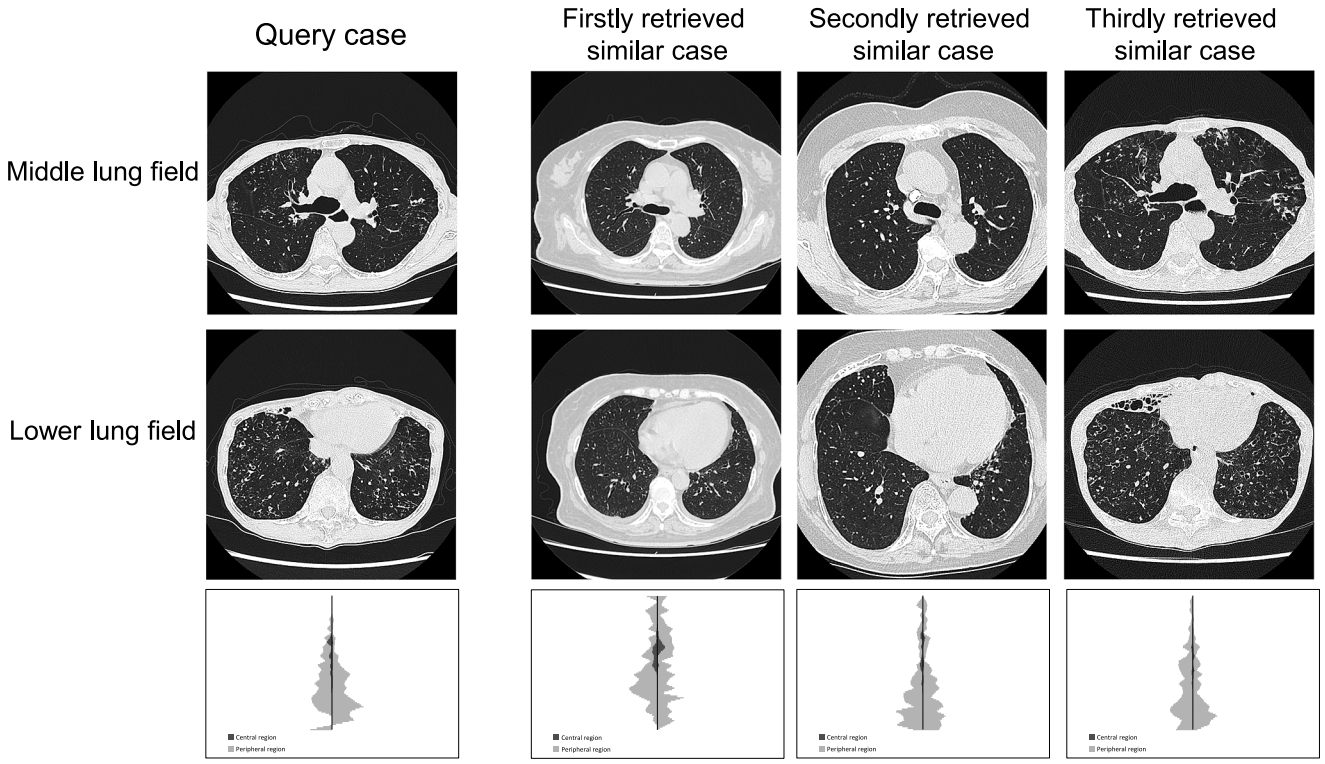

Figure 5 Representative Case 3: The query case is a patient with micronodules. The leftmost column shows the middle and lower lung fields of the query case. The 2nd-, 3rd-, and 4th columns are scans with, in order, the highest to lower degrees of similarity. Although the histograms of retrieved cases were all similar to the histograms of query cases, the three radiologists judged the similarity of retrieved cases to be score 2 or 3 for first or secondly retrieved similar cases and score 4 or 5 for thirdly retrieved similar cases.

In our CB-CTIRS, the radiologist must specify an abnormal lung pattern of interest in each query case. Radiologists should identify lung patterns that may be key to diagnosis in each case. When a radiologist retrieves similar cases for a query case with an unknown diagnosis, he/she wants to retrieve cases, which have patterns similar to the key lung patterns. Therefore, we incorporated the process of manually specifying lung patterns into the CBIRS. Because the inclusion of multiple lung patterns may be useful for retrieving similar cases, we are developing a more complex retrieval system that combines the presentation of multiple lung patterns.

Hwang et al. reported the retrieval performance of CBIRS in patients with three major classes of diffuse interstitial lung disease (DILD)[8]. In their visual similarity assessment graded from 5 = almost identical – 1 = different disease, three radiologists assigned a similarity score of 4 or 5 to 71.3–73.0% of the retrieved chest CT scans. They concluded that their CBIRS performed well in retrieving similar images. However, as their database included only three classes of DILD rather than various lung diseases, the retrieval performance may have been overestimated. However, we used a database that contained various lung diseases in efforts to develop a CBIRS with clinical robustness.

To the best of our knowledge, our CB-CTIRS is the first system to retrieve images based not only on lesion similarity but also on lesion distribution similarity. The latter is important for diagnosing DLDs, because some manifest characteristic distributions in the lungs[14]. Because it remains unclear whether our CB-CTIRS can assist in the diagnosis of undiagnosed cases, we plan to conduct an additional clinical study to investigate its clinical utility.

Thus, our system may reduce the time required for diagnosis. First, we included a relatively small number of patients (n = 503). If our CB-CTIRS was applied to a database consisting of 1,000 cases that included ten cases similar to the query case, and 3 min were required for reading each case, the reading time would be 3,000 min (50 hr). However, more time is spent on sorting cases based on similarity. Under the same conditions, our CB-CTIRS, which grants access to a large image database, can retrieve and sort similar cases in 0.2 s. Thus, the diagnostician will spend only 30 min (10 cases × 3 min), an acceptable length of reading time in daily clinical practise, to reach a diagnosis.

Our study had some limitations. First, this was a single-centre investigation, and the number of cases was limited. To develop a robust CB-CTIRS, more cases need to be trained. Furthermore, the performance of our CB-CTIRS must be verified against a large image database. In this regard, our study was a preliminary study. Second, in our visual analysis of the CB-CTIRS, we focused on one type of pattern to evaluate retrieval performance, although patients with abnormal lung patterns may harbour more than one type of pattern. We are currently developing a CB-CTIRS that can comprehensively determine similarities in cases with multiple patterns. Third, five different CT systems were used in this study. CT

image quality depends on the CT system, reconstruction kernel, and scan parameters of each CT system. This may have affected the classification of lesions, such as GGO. Fourth, our CB-CTIRS two-dimensionally analyses axial CT images. If CT images can be three-dimensionally analysed using the CB-CTIRS, the accuracy of pattern classification and similarity image retrieval can be improved. In the future, it will be necessary to analyse the CT images in three dimensions. Finally, the reference standard for the five types of abnormal patterns in individual patients was subjectively determined by the consensus reading of two radiologists. According to Watadani et al.[20], radiologists often disagree with the CT interpretation of honeycombing, primarily because it is mimicked in the presence of other conditions. Although our readers were experienced in the interpretation of CT scans and made their judgments carefully, lesions recorded as honeycombing may have been admixed with other lesions.

The retrieval performance of our CB-CTIRS was acceptable for consolidation, emphysema, and honeycombing; however, it was unsatisfactory for GGO and micronodules. To be clinically useful, the pattern classification ability and retrieval performance must be improved by increasing the number of training cases. We are now planning a clinical study to investigate whether our CB-CTIRS can assist in cases where disease diagnosis is difficult.

Abbreviations

CT: computed tomography
CB-CTIRS: content-based CT image retrieval system
GGO: ground-glass opacity
PACS: picture archiving and communication systems
DCNN: deep convolutional neural network
CBIRs: content-based image retrieval systems
ATCM: automatic tube current modulation

ACKNOWLEDGEMENTS

This work was supported by Fujitsu Laboratories Ltd.

(Received November 8, 2021)
(Accepted December 9, 2021)

REFERENCES

1. Aisen, A.M., Broderick, L.S., Winer-Muram, H., Brodley, C.E., Kak, A.C., Pavlopoulou, C., et al. 2003. Automated storage and retrieval of thin-section CT images to assist diagnosis: system description and preliminary assessment. Radiology 228: 265–270.
2. Depeursinge, A., Vargas, A., Gaillard, F., Platn, A., Geissbuhler, A. Poletti, P.A., et al. 2012. Case-based lung image categorization and retrieval for interstitial lung diseases: clinical workflows. Int. J. Comput. Assist. Radiol. Surg. 7: 97–110.
3. Deepak, S. and Ameer, P.M. 2020. Retrieval of brain MRI with tumor using contrastive loss based similarity on GoogLeNet encodings. Comput. Biol. Med. 125: 103993.
4. Dhara, A.K., Mukhopadhyay, S., Dutta, A., Garg, M. and Khandelwal, N. 2017. Content-Based Image Retrieval System for Pulmonary Nodules: Assisting Radiologists in Self-Learning and Diagnosis of Lung Cancer. J. Digit. Imaging 30: 63–77.
5. Davis, J. and Goadrich, M. 2006. The Relationship Between Precision-Recall and ROC Curves. Paper presented at: 23rd International Conference on Machine Learning Pittsburgh.
6. Endo, M., Aramaki, T., Asakura, K., Moriguchi, M., Akimaru, M., Osawa, A., et al. 2012. Content-based image-retrieval system in chest computed tomography for a solitary pulmonary nodule: method and preliminary experiments. Int. J. Comput. Assist. Radiol. Surg. 7: 331–338.
7. Fujimoto, K., Taniguchi, H., Johkoh, T., Kondoh, Y., Ichikado, K., Sumikawa, H., et al. 2012. Acute exacerbation of idiopathic pulmonary fibrosis: high-resolution CT scores predict mortality. Eur. Radiol. 22: 83–92.
8. Hwang, H.J., Seo, J.B., Lee, S.M., Kim, E.Y., Park, B., Bae, H.J., et al. 2021. Content-Based Image Retrieval of Chest CT with Convolutional Neural Network for Diffuse Interstitial Lung Disease: Performance Assessment in Three Major Idiopathic Interstitial Pneumonias. Korean J. Radiol. 22: 281–290.
9. Hansell, D.M., Prmstrong, P., Lynch, D.A. and McAdams, H.P. 2005. Basic HRCT patterns of lung disease. In: Imaging of the deseases of the chest, 4th edition. 4th ed.: Elsevier Mosby; 161–163.
10. Hansell, D.M., Bankier, A.A., MacMahon, H., McLoud, T.C., Muller, N.L. and Remy, J. 2008. Fleischner Society: glossary of terms for thoracic imaging. Radiology 246: 697–722.
11. Jiang, M., Zhang, S., Li, H. and Metaxas, D.N. 2015. Computer-aided diagnosis of mammographic masses using scalable image retrieval. IEEE Trans. Biomed. Eng. 62: 783–792.
12. Li, Q., Li, F., Shiraishi, J., Katsuragawa, S., Sone, S. and Doi, K. 2003. Investigation of new psychophysical measures for evaluation of similar images on thoracic computed tomography for distinction between benign and malignant nodules. Med. Phys. 30: 2584–2593.
13. Mehre, S.A., Dhara, A.K., Garg, M., Kalra, N., Khandelwal, N. and Mukhopadhyay, S. 2019. Content-Based Image Retrieval System for Pulmonary Nodules Using Optimal Feature Sets and Class Membership-Based Retrieval. J. Digit. Imaging 32: 362–385.
14. Nishie, A., Kakihara, D., Machitori, A., Aoki, S., Jinzaki, N., Tomiyama, N., et al. 2020. Japan Safe Radiology 2020. Paper presented at: EuroSafe Imaging 2020.
15. Napel, S.A., Beaulieu, C.F., Rodriguez, C., et al. 2010. Automated retrieval of CT images of liver lesions on the basis of image similarity: method and preliminary results. Radiology 256: 243–252.
16. Oosawa, A., Kurosaki, A., Kanada, S., Takahashi, Yui., Ogawa, K., Hanada, S., et al. 2019. Development of a CT image case database and content-based image retrieval system for non-cancerous respiratory diseases: Method and preliminary assessment. Respir Investig.
17. Power, S.P., Moloney, F., Twomey, M., James, K., O'Connor, O.J. and Maher, M.M. 2016. Computed tomography and patient risk: Facts, perceptions and uncertainties. World J. Radiol. 8: 902–915.
18. Rossi, S.E., Erasmus, J.J., Volpacchio, M., Franquet, T., Castiglioni, T. and McAdams, H.P. 2003. "Crazy-paving" pattern at thin-section CT of the lungs: radiologic-

pathologic overview. Radiographics 23: 1509–1519.
19. Svanholm, H., Starklint, H., Gundersen, H.J., Fabricius, J., Barlebo, H. and Olsen, S. 1989. Reproducibility of histomorphologic diagnoses with special reference to the kappa statistic. APMIS 97: 689–698.
20. Watadani, T., Sakai, F., Johkoh, T., Noma, S., Akira, M., Fujimoto, K., et al. 2013. Interobserver variability in the CT assessment of honeycombing in the lungs. Radiology 266: 936–944.

SUPPLEMENTARY MATERIALS

In this supplemental material, we describe the technical aspects of our content-based CT image retrieval system.

(1) CT data to train the deep learning software

To train the deep convolutional neural networks, we accessed existing chest CT images from 50 patients with an established lung diagnosis and 12 subjects whose lung scans were normal, although diffuse lung disease was suspected. All 62 scans were performed at our institute between March 2015 and April 2018. The clinical diagnoses of the 50 patients are presented in Table S1. The 62 study subjects comprised 46 men and 16 women with a median age of 68 years (range, 19–88 years).

The CT scans were performed on one of five scanners [Aquilion One, Aquilion One Genesis, Aquilion Precision (Canon Medical Systems), LightSpeed 64 VCT, and Revolution CT (GE Healthcare)]. The protocol for the Aquilion scanners used a detector configuration of 0.5 × 80 mm, tube rotation time of 0.50 sec, pitch factor of 1.388, scanning field of view (FOV) 30–45 cm, voltage 120 kV, image noise of 12 with automatic tube current modulation (ATCM), reconstruction "FC52" with AIDR 3D weak, and a slice thickness and interval of 1.00 mm. The LightSpeed 64 VCT used a detector configuration of 0.625 × 64 mm, tube rotation time of 0.50 sec, pitch factor of 1.375, scanning FOV of 30–45 cm, a voltage of 120 kV, image noise of 10 with ATCM, reconstruction kernel "lung" with filtered back projection (FBP), and a slice thickness and interval of 1.25 mm. For the Revolution CT scanner, single-energy scan mode was used with a detector configuration of 0.625 × 128 mm, tube rotation time of 0.50 sec, pitch factor of 0.992, scanning FOV of 30–45 cm, a voltage of 120 kV, image noise of 10 with ATCM, reconstruction kernel "lung" with ASiR-V 30%, and a slice thickness and interval of 1.25 mm. Contrast enhancement was required for the individual patients.

Table S2 lists the dominant pattern identified from the 62 lung scans.

Table S1 Clinical diagnosis of 50 patients whose CT data were used for training the deep learning software.

Clinical Diagnosis	Number of patients
Interstitial pneumonia	20
Pneumonia	10
Pneumoconiosis	2
Pulmonary emphysema	7
Pneumocystis carinii pneumonia	3
Non-tuberculous mycobacterial infection	2
Pulmonary oedema	3
Alveolar protenosis	1
ANCA-associated vasculitis	1
Sarcoidosis	1
Total	50

ANCA: antineutrophil cytoplasmic antibody

(2) Reference standard for the lung patterns

We divided the lung field of each of the 62 CT images into 1-cm squares. Two board-certified radiologists with 32 and 14 years of experience in reading chest CT scans consensually recorded each square as showing consolidation, ground-glass opacity (GGO), emphysema, honeycombing, micronodules, or normal lungs. Each pattern was defined based on the Glossary of Terms for Thoracic Imaging. The pattern with the largest volume in the lungs of individuals was defined as the dominant pattern.

(3) Automatic lung-pattern classification

We generated 512 × 512-pixel images; CT values ranging from −1350 to +150 Hounsfield units (HU) were normalised to a range from 0 to 255 based on the window width and level, slope, and intercept of the Digital Imaging and Communications in Medicine (DICOM) tags. C# was the programming language used.

(3-1) Identification of abnormal patterns

We divided each 512 × 512-pixel image into 16 × 16-pixel for classifying approximately 1-cm squares to generate training and evaluation datasets, and then overlaid each square with an image of the corresponding pattern that had been recorded by the two radiologists. We created paired data using a) 224 × 224-pixel images that were resized from the 16 × 16-pixel square images to perform classification with deep learning, using C++ and OpenCV 3.4.6, and b) their corresponding pattern label. We then divided the number of cases to obtain a

Table S2 Number of CT scans and their pattern type, and the number of pre- and post-augmentation squares used for training the deep learning software.

Pattern	Consolidation	Ground-glass opacity	Emphysema	Honeycombing	Micronodule	Normal lung
Number of cases	11	9	7	16	10	12
Number of patches before augmentation	3433	3901	2836	2947	3042	11137
Number of patches after augmentation	31223	35471	25796	26803	27668	101297

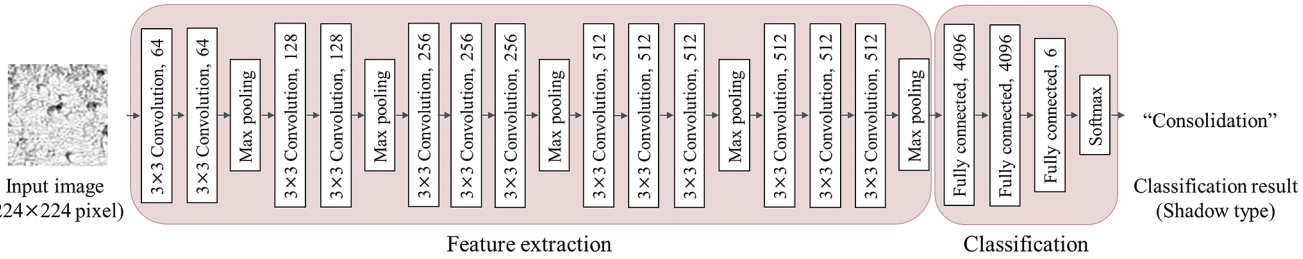

Figure S1 The structure of the convolutional neural network was based on VGG16, a representative model used in image recognition.

training-to-evaluation ratio of 9:1.

(3-2) Extraction of lung areas

Images normalised to 0–255 CT values were generated in JPEG format and used as training and evaluation datasets. The right lung (pixel value, 127), left lung (pixel value, 255), and miscellaneous regions (pixel value, 0) were labelled on 512 × 512-pixel images. Data showing the values of the right and left lung, and of the miscellaneous CT regions were recorded in the PGM format. Paired data were generated using the JPEG images and the corresponding PGM images and used as training and evaluation datasets.

(4) Data augmentation using the point spread function

To avoid a decrease in the accuracy of pattern classification resulting from image-quality differences attributable to the use of different scanners and protocols, we augmented the data with the aid of the point spread function (PSF). We replaced the PSF characteristics of CT images with different PSF characteristics during PSF-driven data augmentation. This resulted in CT images that were artificially scanned with different scanners and protocols; the PSF was obtained from the scanned images by scanning CT phantoms. We used 14 PSF types for data augmentation: FC08, FC08-H, FC52, and FC86 (Aquilion ONE), FC52, FC86, and LUNG (Aquilion Precision), BONE, BONEPLUS, LUNG, and SOFT (Revolution CT), and BONE, BONEPLUS, and SOFT (Revolution CT DECT). The number of squares before and after augmentation used for training the deep learning software is shown in Table S2.

(5) Deep learning using a convolutional neural network

We applied supervised convolutional neural network (CNN) training. The computer specifications used in the training procedure were the Ubuntu 16.04 operating system (OS), Xeon E5-2680 v4 2.4 GHz CPU, 128 GB random access memory (RAM), and NVIDIA Tesla P100 16 GB GPU. The CNN was created using Python 2.7, and the CNN framework used was the Caffe 1.1 framework. The CNN structure (Figure S1) was based on VGG16, a representative model used for image recognition.

The five types of opacity identified by VGG16 were consolidation, GGO, honeycombing, emphysema, and micronodules; normal lungs were also identified. The training dataset comprised 248,258 images subjected to PSF-based data augmentation; 31,223 images showed consolidation, 35,471 GGO, 26,803 honeycombing, 25,796 emphysemas, and 27,668 micronodules; 101,297 were normal lung images.

Training was performed with 130 epochs and a batch size of 32, with 224 × 224-pixel square images that were extracted from 62 patients and resized from 16 × 16-pixel square images. Images characterised with 13 convolution layers (activation function: rectified linear unit [ReLU]) and five pooling layers (system: max) were input to the CNN. The probability of each input image being assigned to a shadow was determined using three fully connected layers and one softmax layer. The most probable opacity was considered to be the opacity of the input image. During the training phase, the results yielded by the CNN and labelled opacity type were used in loss-function calculations (loss function: cross-entropy). Each parameter of the CNN was updated to minimise the loss function, which is the difference between the CNN results and the labelled opacity type, using backpropagation. The stochastic gradient descent method was used for optimisation with a learning rate of 10^{-3}, momentum of 0.9, and weight decay of 0.0005.

(6) Extraction and lung segmentation into central and peripheral regions on horizontal CT images

Figure S2 presents an overview of the method used to extract the distribution of lung opacities from the CT images. After the lung regions were extracted from each input image, the image was divided into a grid comprised of 1-cm squares, and the opacity type within a grid was identified. Then, based on the central and peripheral region models, the extracted lung region was segmented into the central and peripheral regions. The number of squares containing a specific type of opacity in the central and peripheral regions was calculated and presented as a histogram. The data obtained from the histograms of each image slice were modelled across the long axis of the human body.

(6-1) Extraction of lung regions from CT images

In this process, we included lung CT images from 158 patients; 4,740 images (30 per patient) were used as the training data set for extraction. We used 1,185 batches consisting of four images each for the training process. The training set consisted of 30 image slices, which were extracted at equal intervals from the 158 lung CT images.

We extracted the lung regions from the CT images

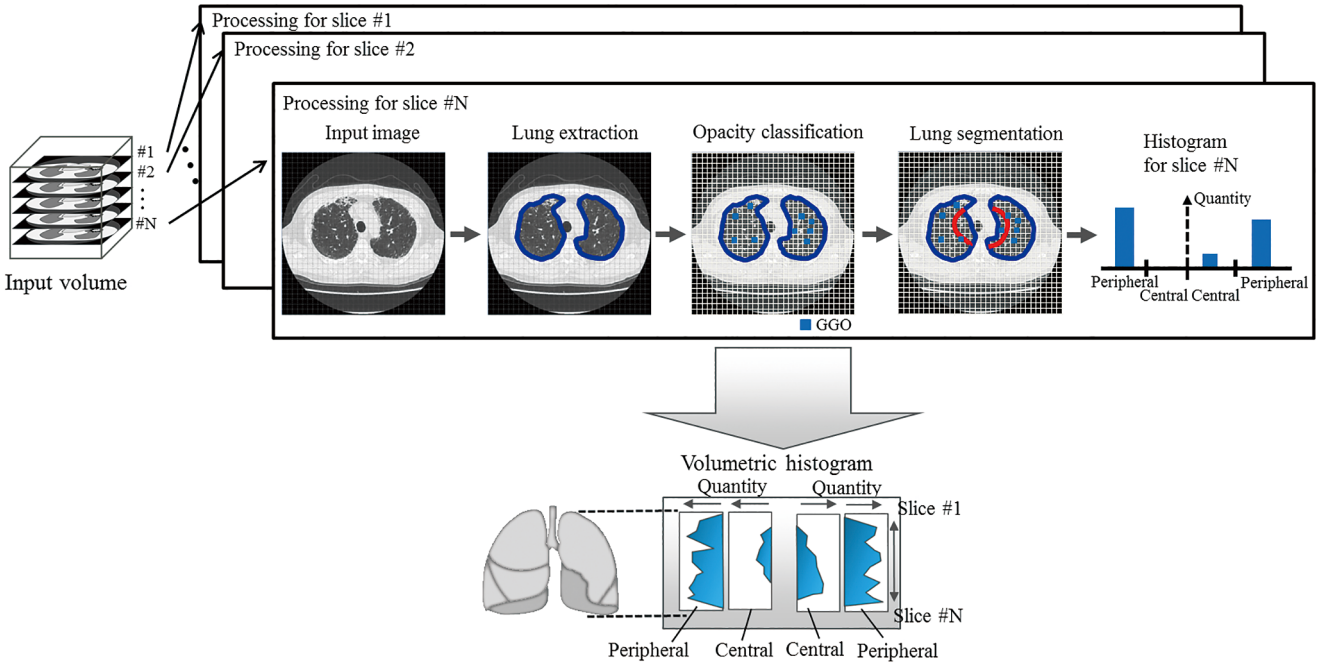

Figure S2 Overview of the method used to extract the distribution of lung patterns from CT images.

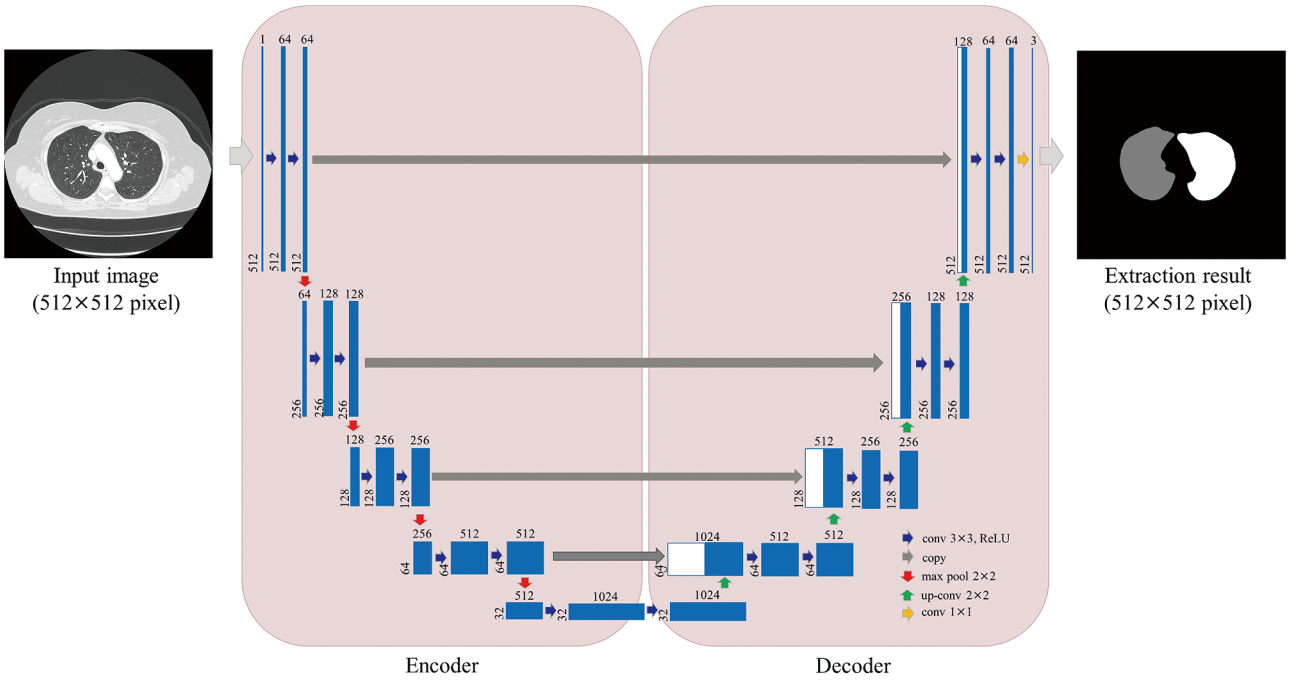

Figure S3 Extraction of the lung fields on CT images using deep learning with U-Net.

using deep learning with U-Net (Figure S3). U-Net was created in Python 3.6 with the Tensorflow 1.4.0 framework. The computer specifications were Ubuntu 16.04 OS, Xeon E5-2680 v4 2.4 GHz CPU, 128 GB RAM, and NVIDIA Tesla P100 16 GB GPU.

The three image classes identified by U-Net were the left lung, right lung, and miscellaneous classes. We calculated the 3-class probability of each pixel on the input image and considered the image class with the highest probability as the outcome result. Before performing convolution layer calculations, we applied zero padding to the U-Net input images to normalise the size of the input and output images. Training was performed for 1000 epochs. During the training phase, the U-Net results and labelled 3-class were used in the loss-function calculations (loss function: cross-entropy). Each parameter of U-Net was updated to minimise the loss function, which is the difference between the U-Net results and the labelled 3-class, using backpropagation. Optimisation was performed using ADAM with a learning rate of 10-4, β_1 0.9, β_2 0.999.

(6-2) Central and peripheral lung region modelling

We divided the lung regions on CT images into central and peripheral regions (Figure S4a) and modelled the region structures, as shown in Figure S4b. The mod-

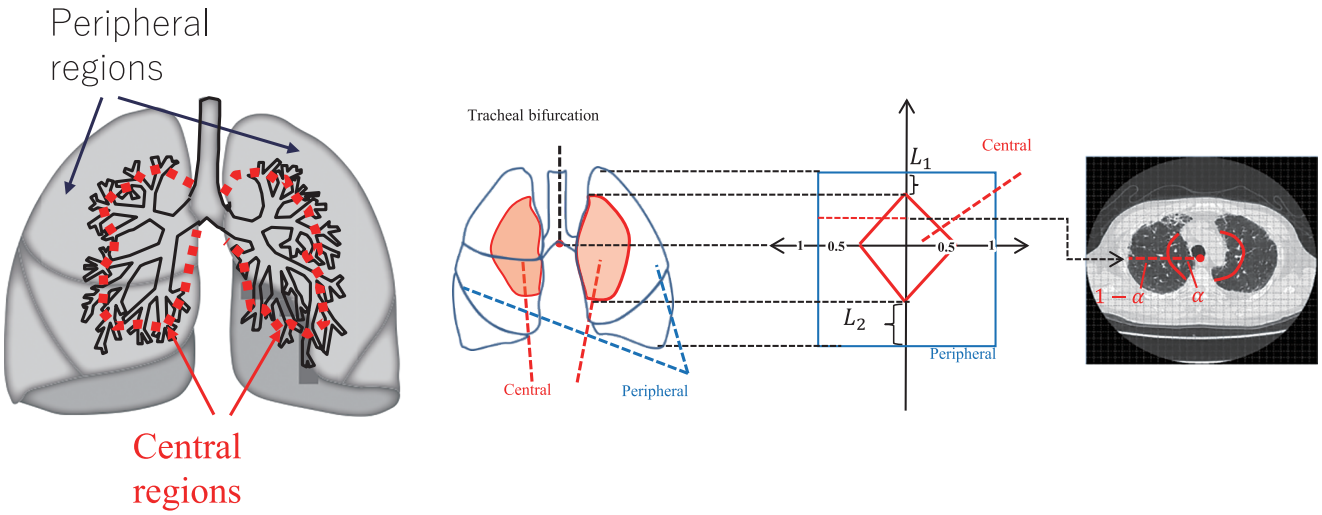

Figure S4 Method for dividing the lung field from CT images into central and peripheral regions. a. Conceptualization of the central and peripheral regions (the left part). b. Schematic for dividing the lung from CT images into central and peripheral regions (the right part).

elled central region was from L_1 (top of the lung) to L_2 (bottom of the lung); in the horizontal plane, its area was largest at the tracheal bifurcation. Therefore, this was considered the internal region, and the lung was divided roughly into two equal parts (Figure S4b, centre). Here, the tracheal bifurcation was extracted as the point where the tracheal region was extracted from each image slice from the upper to the lower portion of the CT image branches into two parts. Parameter α expresses the cross-sectional area of the central region (Figure S4b, right). It was estimated as 0.5. For the tracheal bifurcation region, it was estimated as 0 for the upper and lower ends of the lung, and as a linearly interpolated value for all other sections. Our method segments each image slice into central and peripheral regions by calculating α from this model and generating a curve that passes through the internal division point that divides the lung region into a (1- α):α ratio.

(7) Determination of the morphological similarity between query and database images

The morphological similarity of the query and database images of the patients was determined by calculating the histogram similarity of the opacity type specified as dominant in the query image. The similarity between histograms was calculated using the histogram intersection method, which calculates the morphological similarity of two histograms, $H_1[i]$ and $H_2[i]$, using the equation $similarity = \sum_i \min(H_1[i], H_2[i])$.