Böttcher *et al. European Radiology Experimental* (2025)9:4
https://doi.org/10.1186/s41747-024-00539-w

European Radiology EXPERIMENTAL

ORIGINAL ARTICLE Open Access

Evaluation of a content-based image retrieval system for radiologists in high-resolution CT of interstitial lung diseases

Benjamin Böttcher[1,2], Marly van Assen[1], Roberto Fari[1,3], Philipp L. von Knebel Doeberitz[1,4], Eun Young Kim[1,5], Eugene A. Berkowitz[1], Felix G. Meinel[2] and Carlo N. De Cecco[1*]

Abstract

Background This retrospective study aims to evaluate the impact of a content-based image retrieval (CBIR) application on diagnostic accuracy and confidence in interstitial lung disease (ILD) assessment using high-resolution computed tomography CT (HRCT).

Methods Twenty-eight patients with verified pattern-based ILD diagnoses were split into two equal datasets (1 and 2). The images were assessed by two radiology residents (3rd and 5th year) and one expert radiologist in four sessions. Dataset 1 was used for sessions A and C, assessing diagnostic accuracy and confidence with mandatory and without CBIR software. Dataset 2 was used for sessions B and D with optional CBIR use, assessing time spending and frequency of CBIR usage. Accuracy was assessed on the CT pattern level, comparing readers' diagnoses with reference diagnoses and CBIR results with region-of-interest (ROI) patterns.

Results Diagnostic accuracy and confidence of readers showed an increasing trend with CBIR use compared to no CBIR use (53.6% *versus* 35.7% and 50.0% *versus* 32.2%, respectively). Time for reading significantly decreased in both datasets (A *versus* C: 104 s *versus* 54 s, $p < 0.001$; B *versus* D: 88.5 s *versus* 70 s, $p = 0.009$), whereas time for research increased with CBIR software use (A *versus* C: 31 s *versus* 81 s, $p = 0.040$). CBIR results showed a high pattern-based accuracy of overall 73.4%. Comparison between readers indicates a slightly higher accuracy of CBIR results when more than one ROI was used as input (77.7% *versus* 70.1%).

Conclusion CBIR software improves in-training radiologist diagnostic accuracy and confidence while reducing interpretation time in ILD assessment.

Relevance statement Content-based image retrieval software improves the assessment of interstitial lung diseases (ILD) in high-resolution CT, especially for radiology residents, by increasing diagnostic accuracy and confidence while reducing interpretation time. This can provide educational benefits and more time-efficient management of complex cases.

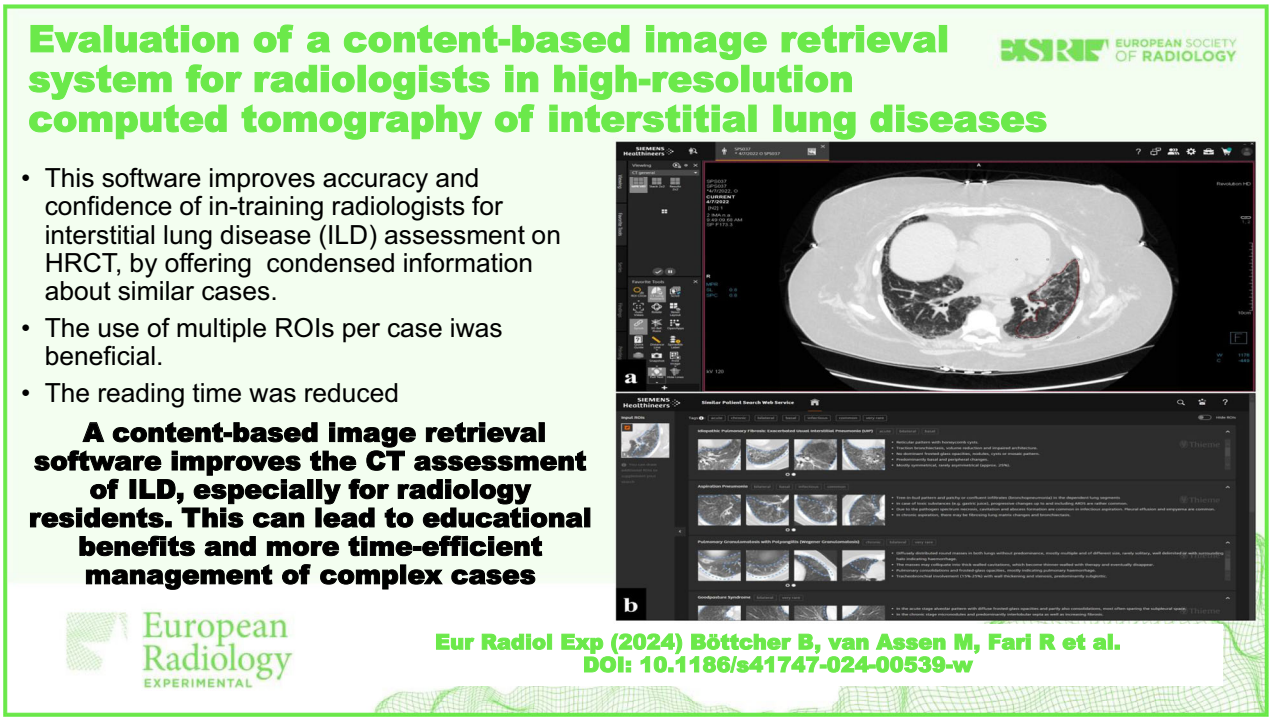
Key Points

- A content-based image retrieval (CBIR) software improves diagnostic accuracy and confidence for in-training radiologists for interstitial lung disease (ILD) assessment on computed tomography (CT).
- A CBIR application provides condensed information about similar HRCT cases reducing time for ILD assessment.
- CBIR algorithms benefit from the input of multiple regions of interest per ILD case.

*Correspondence:
Carlo N. De Cecco
carlo.dececco@emory.edu
Full list of author information is available at the end of the article

Keywords Artificial intelligence, Diagnosis (computer-assisted), Lung diseases (interstitial), Tomography (x-ray computed)

Graphical Abstract



Background

Interstitial lung diseases (ILD) are very heterogeneous pathologies of the respiratory system in which inflammation and fibrosis cause destruction of lung structure. Classification of these entities is complex and various multidisciplinary guidelines are established to stratify clinical decisions in this field [1–5]. Although improvements were made, defining a precise diagnosis is still challenging, and agreement between multidisciplinary teams is only moderate [6].

High-resolution computed tomography (CT) provides detailed images and is a cornerstone in the diagnostic pathway of ILD [1–5]. Despite new developments in both acquisition technology and post-processing increasing image quality [7–9], discrimination between various ILD patterns is a major challenge with only moderate inter-observer agreement, even in experienced radiologists [10, 11].

Content-based image retrieval (CBIR) systems for CT datasets are designed to perform an image-based query to receive similar CT scans from a database. These systems use image information, including but not limited to pattern or attenuation values from a selected input, for example, a region-of-interest (ROI), to identify similar CT images. Additionally, retrieved CT datasets can be supplemented with assigned pathologies and more detailed case-related information to assist radiologists in diagnostic decision-making.

Early investigations showed promising results for CBIR systems used for chest CT scans, however, major limitations, such as low computational power, hampered further development [12]. The rise of artificial intelligence and growing computing resources led to significant progress in CBIR application development focusing on different aspects of chest CT scans such as lung nodule differentiation [13], distinction of lung cancer from atypical tuberculosis [14], or assessment of obstructive lung diseases [15]. In particular, a small number of CBIR systems for ILD assessment on high-resolution CT datasets have been proposed in recent years [16–18]. Validation of such CBIR applications in a clinical setting focuses on the accuracy of retrieved images and on the diagnostic confidence of readers as well as the time needed per case [19–22].

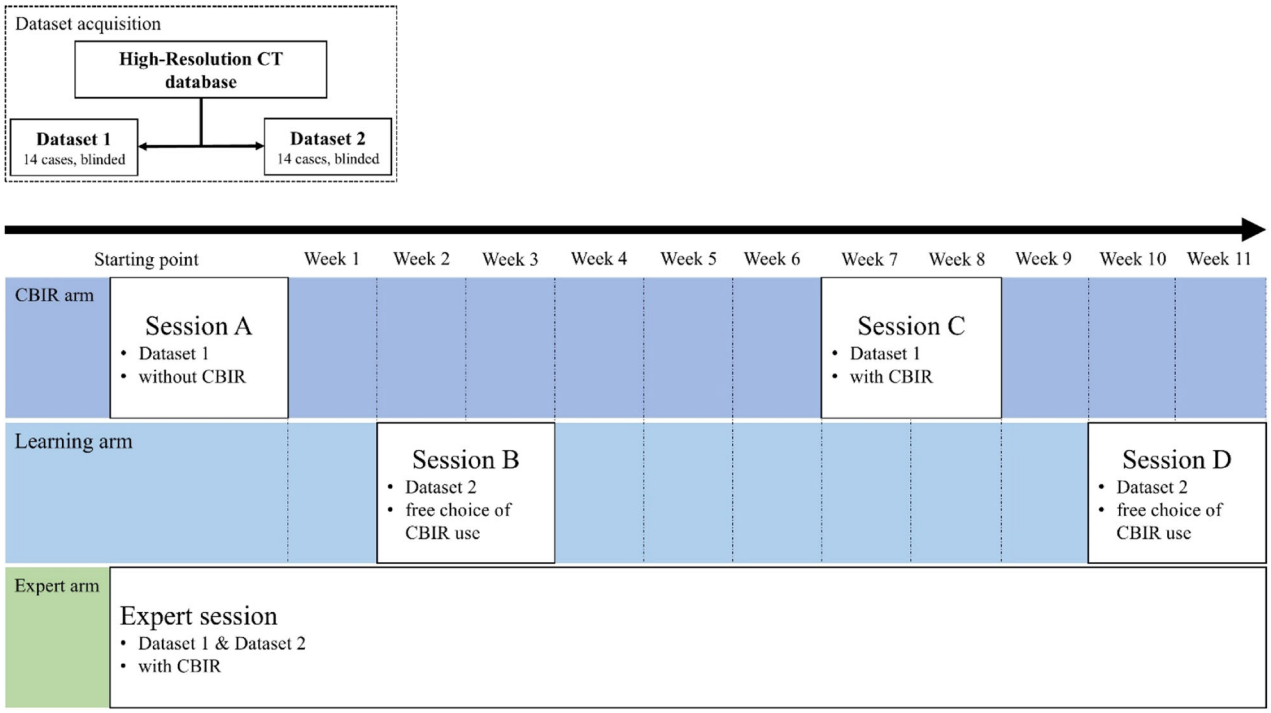


Fig. 1 Study design of CBIR software evaluation. First, the retrospectively selected high-resolution computed tomography datasets ($n = 28$) were assigned to one of two datasets (each $n = 14$). Subsequently, sessions from A to D were performed by resident readers at several time points, as visualized above. The expert performed a single session with both datasets. CBIR, Content-based image retrieval

Despite this growing data on the performance of CBIR systems in clinical scenarios, there is no data available on interobserver variability of input ROI placement. This is somewhat remarkable since the ROIs placed by the user are critical for the accuracy of the retrieved images and are not yet standardized. Therefore, this study aims to evaluate an AI-based CBIR software with a special focus on the ROI used as input and on how the variability of ROI placements might impact the accuracy of CBIR software.

Methods

Ethical approval and patient selection

This single-center cohort study was approved by the responsible Institutional Review Board (IRB number: STUDY00002503; November 8, 2021), and the need for written informed consent was waived.

High-resolution axial thin slice CT image datasets (slice thickness 0.625–1.5 mm) were selected retrospectively according to the following criteria: (i) patients of ≥ 18 years with (ii) a chest CT scan between 2010 and 2021 with verified pattern-based ILD diagnosis. The following common ILD patterns were included: usual interstitial pneumonia, non-specific interstitial pneumonia, and fibrotic as well as non-fibrotic hypersensitivity pneumonitis. The reference diagnosis was defined and verified patterns based on the radiology report and electronic health records, including multidisciplinary board decisions. Four patients without pathological findings on CT were added. Exams with non-diagnostic image quality were excluded.

Study design

The study design is visualized in Fig. 1. Overall, four reading sessions were conducted with a fully blinded case presentation using dataset 1 for sessions A and C and dataset 2 for sessions B and D. In total, 28 CT cases were included and split into two datasets of 14 cases each. Each dataset contained different patients but a similar distribution of cases. Two radiology residents (readers 1 and 2) in their 3rd and 5th year of training (less than one year of chest CT reading experience and two years of experience, respectively) separately performed all sessions and were allowed to change image windowing settings to their preferences and to use digital or non-digital sources of information to assist diagnostic decision-making in all sessions. Since the case presentation was fully blinded, there was no access to medical health records during all sessions. Readers' diagnostic decisions were given as open answers not limited to the included ILD entities. Prior to the first session, both

readers were introduced to the CBIR software package using the user manual and a small number of test cases which were not included in the test datasets.

Sessions A and C aimed to assess diagnostic confidence, diagnostic accuracy, and time needed for image analysis and diagnostic decision-making without (session A, baseline) and with CBIR software use (session C), in the following referred as CBIR arm. Session C was scheduled six weeks after session A to avoid recall bias. Usage of software tools was expected to change with gaining experience and knowledge of software functions and benefits. Therefore, sessions B and D aimed to assess the learning effects of CBIR system use, including time spent on software use and frequency of CBIR software use (learning arm). Each session was scheduled at least one week after sessions A and C. In the learning arm (sessions B and D) readers had free choice to use the CBIR application or not. Additionally, an experienced expert radiologist (+10 years in reading chest CT) performed a separate session (session E) with mandatory CBIR use in each case on both datasets. An independent observer recorded reading times, diagnoses, the confidence of diagnosis, the relevance of software use, results of the CBIR software, and the number of ROIs to ensure a continuous workflow. Feedback on correct or incorrect diagnostic decisions after the sessions was not provided.

CBIR software

All reading sessions were conducted on a dedicated workstation using commercially available software (syngo.via, version VB50, Siemens Healthineers, Erlangen, Germany) with the fully integrated CBIR application (Similar Patient Search Web Service, version VA41B, Siemens Healthineers, Erlangen, Germany). The CBIR software utilized in this study is founded upon an image embedding function that maps a ROI to a fixed-length feature representation. The embedding function takes the form of a deep residual convolutional neural network, more specifically, a ResNeXt-50 (32 × 4d) [23]. The network backbone was initialized by model weights pretrained on an ImageNet dataset [24]. Input images of the training set were normalized using a mean of 0 and a standard deviation of 1. Model training was performed on a multi-center CT database using metric learning techniques, including triplet loss [25], N-pair loss [26], and an Adam optimizer with a batch size of 256 and 5-fold cross-validation, ensuring that pathological patterns are effectively segregated in the feature space.

A region-of-interest (ROI) based query is used to retrieve cases from a multi-site database of more than 800 CT cases covering 78 ILD diagnoses, see Fig. 2a. The CBIR software allows the use of multiple ROIs simultaneously, and the user can include or exclude individual ROIs from image query manually. Retrieved CT cases are displayed in a list with associated pathologies, starting with the entity with the highest image-based similarity at the top. An example is visualized in Fig. 2b. Each pathology can be selected, and the software provides multiple similar CT scans from its database and additional information, such as definitions, imaging signs, clinical aspects, differential diagnoses, and tips and pitfalls, see Fig. 3.

Analysis of time, accuracy, and ROIs

Overall time per case was recorded. For detailed analysis, time was measured as time for reading the CT images and research time spent using the CBIR software and/or other sources of information. The accuracy of readers was analyzed by direct comparison between readers' and verified reference diagnoses (diagnosis-based approach). To assess CBIR software accuracy, verified reference diagnoses and CBIR software results were assigned to characteristic CT imaging patterns (CT-patterns-based approach) by a board-certified radiologist specialized in cardiothoracic imaging (*e.g.*, usual interstitial pneumonia was assigned with honeycombing, bronchiectasis, and reticular abnormalities). In addition, each ROI was assigned to an anatomical location and CT pattern. CBIR software accuracy was calculated based on matches between assigned CT patterns of CBIR results and CT patterns within the input ROIs. This pattern-based approach was chosen because the algorithm only had access to image information within ROIs, whereas readers had access to the overall image dataset in each case. The software results were considered accurate if at least one ROI CT pattern was matched.

The objective analysis was performed on overall cases and on per-reader and per-session levels. For per-reader analysis, results from all sessions were evaluated separately for readers 1 and 2 to detect inter-reader variabilities. Per-session analysis was performed to identify changes between sessions A-D by evaluating results from both readers together separated for each session.

Analysis of diagnostic confidence and CBIR relevance

Readers were asked to scale their confidence in diagnosis on a 5-point Likert scale: 1 = very low confidence in diagnosis; 2 = low confidence in diagnosis; 3 = intermediate confidence in diagnosis; 4 = high confidence in diagnosis; and 5 = very high confidence in diagnosis. Additionally, readers were asked to state if CBIR software was relevant to the diagnostic decision-making process. The software was considered relevant if retrieved images and information helped the reader in the diagnostic

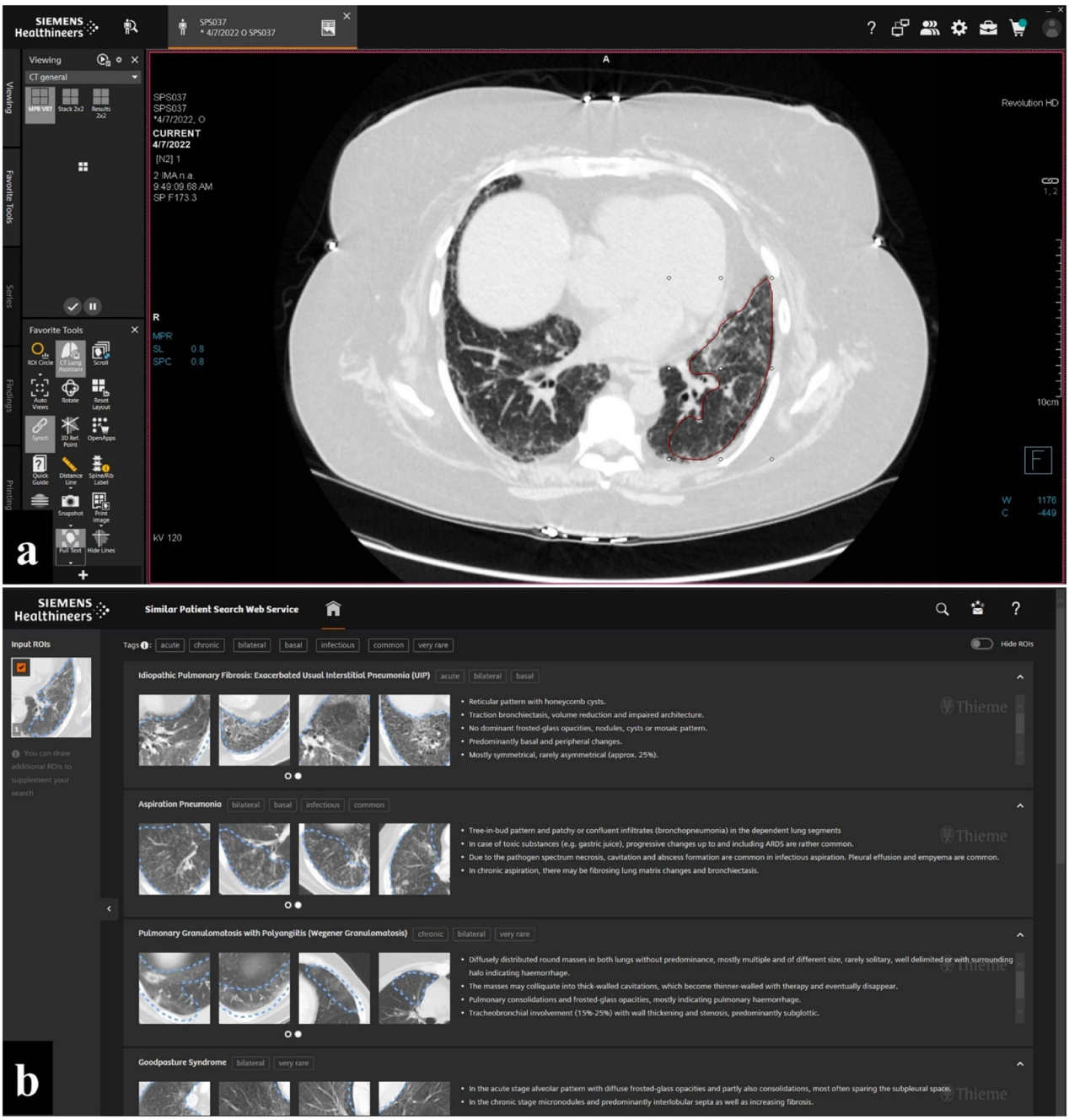


Fig. 2 User interface of the content-based image retrieval application (Similar Patient Search Web Service (Siemens Healthineers). **a** Example of an input ROI on an axial high-resolution chest computed tomography in lung window settings. **b** List of retrieved cases with assigned pathologies starting with the most similar at the top. ROI, Region-of-interest

decision-making or if confidence in diagnosis was improved.

Statistical analysis

Wilcoxon signed-rank test was used to evaluate differences in time overall, time used for reading, and time used for research. Differences in the number of ROIs used by each reader were assessed by using the χ^2 test. The McNemar test was employed to investigate discrepancies in diagnostic accuracy and CBIR utilization. Statistical analysis was performed using SPSS statistics, version 28.0.0 (IBM Corporation). Statistical significance was considered for a p-value < 0.050.

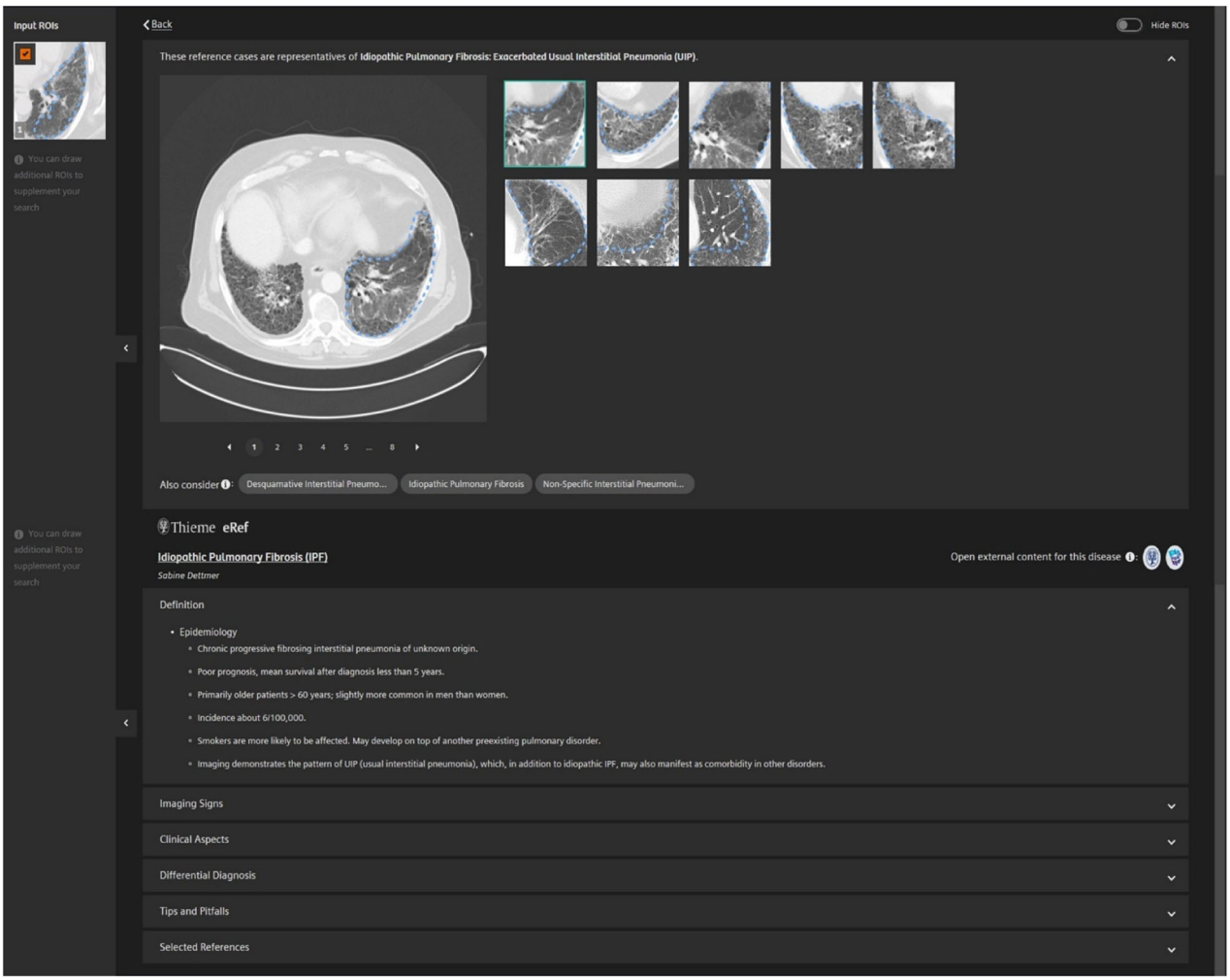


Fig. 3 User interface of the content-based image retrieval application after selecting a pathology from the results list. At the top, multiple scans with high similarity to the input ROI can be assessed. Below detailed information of the assigned pathology is provided to assist diagnostic decision-making. ROI, Region-of-interest

Results

Patient characteristics

The CT database used in this study consisted of 28 high-resolution CT scans from 12 women (42.9%) and 16 men (57.1%) with a mean age of 65 years (range 29–87) and a mean body mass index of 29.9 kg/m^2. CT imaging was performed at 100 or 120 kVp with automated tube current modulation according to clinical protocols. The slice thickness ranged from 0.625–1.5 mm. CT scans were acquired on 10 different CT scanner models from two different vendors (GE Healthcare and Siemens Healthineers). The database consisted of eight CT scans of each included ILD pattern and four scans without pathological changes.

Diagnostic accuracy of readers (diagnosis-based)

Over all sessions, readers' primary diagnosis matched the reference diagnosis in 53/112, 47.3% of cases. In per-session analysis, diagnostic accuracy showed an increasing trend over all sessions from session A (10/28, 35.7%) to session D (15/28, 53.6%, $p = 0.302$) (Table 1).

On the per-reader level, the accuracy of reader 1 showed an increasing trend with the use of CBIR software from session A (5/14, 35.7%) to session C (8/14, 57.1%, $p = 0.375$). This trend was not observed for reader 2 (5/14, 35.7% to 5/14, 35.7%). The second and third differential diagnoses given by readers showed lower accuracy (second diagnosis 18/112, 16.1%; third diagnosis 24/112, 21.4%) throughout all sessions.

Table 1 Accuracy of primary diagnosis by readers and primary retrieved results by content-based image retrieval software

	Session A *Without CBIR use*	Session B *Optional CBIR use*	Session C *Mandatory CBIR use*	Session D *Optional CBIR use*
Reader				
Overall	53/112, 47.3%			
	10/28, 35.7%	15/28, 53.6%	13/28, 46.4%	15/28, 53.6%
Reader 1	28/56, 50.0%			
	5/14, 35.7%	7/14, 50.0%	8/14, 57.1%	8/14, 57.1%
Reader 2	25/56, 44.6%			
	5/14, 35.7%	8/14, 57.1%	5/14, 35.7%	7/14, 50.0%
CBIR software				
Overall	–	45/61, 73.4%		
		14/18, 77.8%	17/28, 60.7%	14/15, 93.3%
Reader 1		21/27, 77.7%		
Reader 2		24/34, 70.1%		
Expert		16/24, 66.7%		

Reader accuracy was tested against reference diagnosis. CBIR software accuracy was tested against computed tomography patterns within input regions of interest.
CBIR Content-based image retrieval

Diagnostic accuracy of CBIR software (CT-pattern-based)

Accuracy analyses are summarized in Table 1. Primary retrieved cases of the CBIR software matched at least one CT pattern of input ROIs in 45/61 (73.4%) of all cases where CBIR software was used ($n = 61$). The second and third cases retrieved by CBIR showed lower accuracy (second result 21/61, 34.4%; third result 28/61, 46%). Per-session analyses found a significant increase between session B (14/18, 77.8%) to session D (14/15, 93.3%, $p = 0.031$), most likely because of changes in ROI placements. To calculate CBIR accuracy in session C (mandatory CBIR use), normal cases without pathological findings were excluded from analysis because normal cases are not considered by the software.

At per-reader analysis, CBIR software accuracy revealed slight differences between readers 1 and 2 (21/27, 77.7% *versus* 24/34, 70.1%). Primary results of the CBIR software when used by the expert radiologist matched at least one ROI CT pattern in 16/24 (66.7%) of cases.

Diagnostic confidence

Results of confidence rating on the 5-point Likert scale were grouped into low confidence (score 1 or 2), moderate confidence (score 3), and high confidence (score 4 or 5). A comparison of sessions A and C showed an increasing trend of confidence levels from low to moderate. Confidence in sessions B and D was similar but higher compared to sessions A and C. This might be caused by the fact that CBIR results and reader´s diagnosis mismatch and, therefore, decrease confidence if CBIR use is mandatory. The expert radiologist had a higher confidence consistently throughout all sessions. Confidence ratings are displayed in Table 2.

Relevance of CBIR software

Table 2 shows the results of CBIR relevance analysis. CBIR software was rated relevant for diagnostic decision-making in 32/61, 52.5% of all cases with CBIR use (sessions B, C, and D). Per-session analysis showed that CBIR software was considered relevant for the diagnostic decision-making in more cases in sessions with optional CBIR use (session B 11/18, 61.1%; session D 10/15, 66.7%) compared to session C with mandatory CBIR use (11/28, 39.3%).

In per-reader analysis, reader 1 rated CBIR software relevant for diagnostic decisions in 15/27 (55.6%) of cases over all sessions. Reader 2 rated CBIR system relevance slightly lower with 50% of all cases. In contrast, the expert radiologist rated CBIR software results relevant for diagnostic decision-making in 27/28 (96.4%) of all cases, which was unexpected given that the greatest benefit of CBIR use was anticipated for resident readers. This discrepancy may be attributed to disparate interpretations of the retrieved CT cases and associated pathologies, as it can be assumed that in-training radiologists with less experience may encounter greater challenges in interpreting CBIR results compared to expert readers.

CBIR software use

The use of CBIR software was not allowed in session A and was mandatory in session C. In session B and D readers had free choice to use CBIR software during assessment. The CBIR application was used in 18/28 (64.3%) of cases in session B and in 15/28 (53.6%) of cases in session D. Readers used the CBIR software once per case, multiple uses in the same CT dataset were technically allowed but was not observed during all sessions.

Table 2 Subjective analysis of confidence in diagnosis and relevance of content-based image retrieval software results

	Session A *Without CBIR use*	Session B *Optional CBIR use*	Session C *Mandatory CBIR use*	Session D *Optional CBIR use*
Confidence in diagnosis				
Low (score 1 or 2)	13/28, 46.4%	3/28, 10.7%	8/28, 28.6%	4/28, 14.3%
Moderate (score 3)	6/28, 21.4%	11/28, 39.3%	11/28, 39.3%	10/28, 35.7%
High (score 4 or 5)	9/28, 32.2%	14/28, 50.0%	9/28, 32.1%	14/28, 50.0%
Relevance of CBIR result for diagnostic decision				
Overall	–	32/61, 52.5%		
		11/18, 61.1%	11/28, 39.3%	10/15, 66.7%
Reader 1		15/27, 55.6%		
		3/6, 50.0%	6/14, 42.9%	6/7, 85.7%
Reader 2		17/34, 50.0%		
		8/12, 66.7%	5/14, 35.7%	4/8, 50.0%
Expert		27/28, 96.4%		

CBIR software was considered relevant if retrieved images and information helped the reader in the diagnostic decision-making or if confidence in diagnosis was improved. *CBIR* Content-based image retrieval

Table 3 Interpretation time analysis on the per-session level

	Time overall	Time for reading	Time for research
Session A Without CBIR use	166 s (112–231.5 s)	104 s (81.5–148.5 s)	31 s (0.0–92.3 s)
Session B Optional CBIR use	159 s (74.3–257.0 s)	88.5 s (64.3–154.3 s)	55.5 s (0.0–114.0 s)
Session C Mandatory CBIR use	145 s (89.3–201.0 s)	54 s (37.5–91.0 s)	81 s (41.5–121.8 s)
Session D Optional CBIR use	102.5 s (44.0–190.0 s)	70 s (33.5–111.3 s)	31.5 s (0.0–81.3 s)

Times are displayed in seconds (s) as median (interquartile interval). Time overall = time used from opening the case to set final differential diagnoses, Time for reading = time used for image assessment. Time for research = time used for searching in digital or non-digital sources of information and CBIR software use. *CBIR* Content-based image retrieval

Per-reader analysis showed that CBIR software use of reader 1 slightly increased from session B (6/14, 42.9%) to D (7/14, 50%), whereas software use of reader 2 showed a decreasing trend (12/14, 85.7% *versus* 8/14, 57.1%, $p = 0.125$). This deviation in CBIR use may be caused by different ratings of the relevance of CBIR results for diagnostic decision-making, as reader 2 perceived CBIR relevance slightly lower compared to reader 1. It can be assumed that a decreased relevance of CBIR application results for diagnostic decision-making results in a reduction of CBIR software use.

Time analysis

Table 3 presents time analysis on a per-session level. Time for reading decreased between subsequent sessions in both study arms (session A, 104 s *versus* session C, 54 s, $p < 0.001$; session B, 88.5 s *versus* session D, 70 s, $p = 0.009$). In the CBIR arm, time for research increased with CBIR software use (session A, 31 s *versus* session C, 81 s, $p = 0.040$). The overall time needed per case showed no significant difference between sessions A (166 s) *versus* session C (145 s, $p = 0.356$), whereas it decreased between session B (159 s) *versus* session D (102.5 s, $p = 0.006$) when CBIR use was optional.

ROI subanalysis

Overall, 124 ROIs were placed during all sessions, a mean of 1.4 ROIs per patient. Figure 4 visualizes the ROIs used by readers and the expert in an identical case. Per-reader analysis showed different ROI approaches between readers 1 and 2. Reader 1 used a multiple-ROI approach with a mean of 2.26 ROIs per patient, whereas reader 2 used a single-ROI approach with a mean of 1.06 ROIs per patient ($p < 0.001$). The expert used a single-ROI approach with 1 ROI per case throughout all cases. Reader 2 placed a majority of ROIs in central and peripheral regions (32/35, 91.4%). Reader 1 and the expert used ROIs with peripheral distribution in 35/61, (57.4%) and 13/28 (46.4%) of cases, respectively. They used ROIs with central and peripheral distribution in only 26/61 (42.6%) of cases for reader 1 and 13/28 (46.4%) of cases for the expert reader.

Discussion

This study evaluated the impact of CBIR software on diagnostic accuracy, confidence, and interpretation time

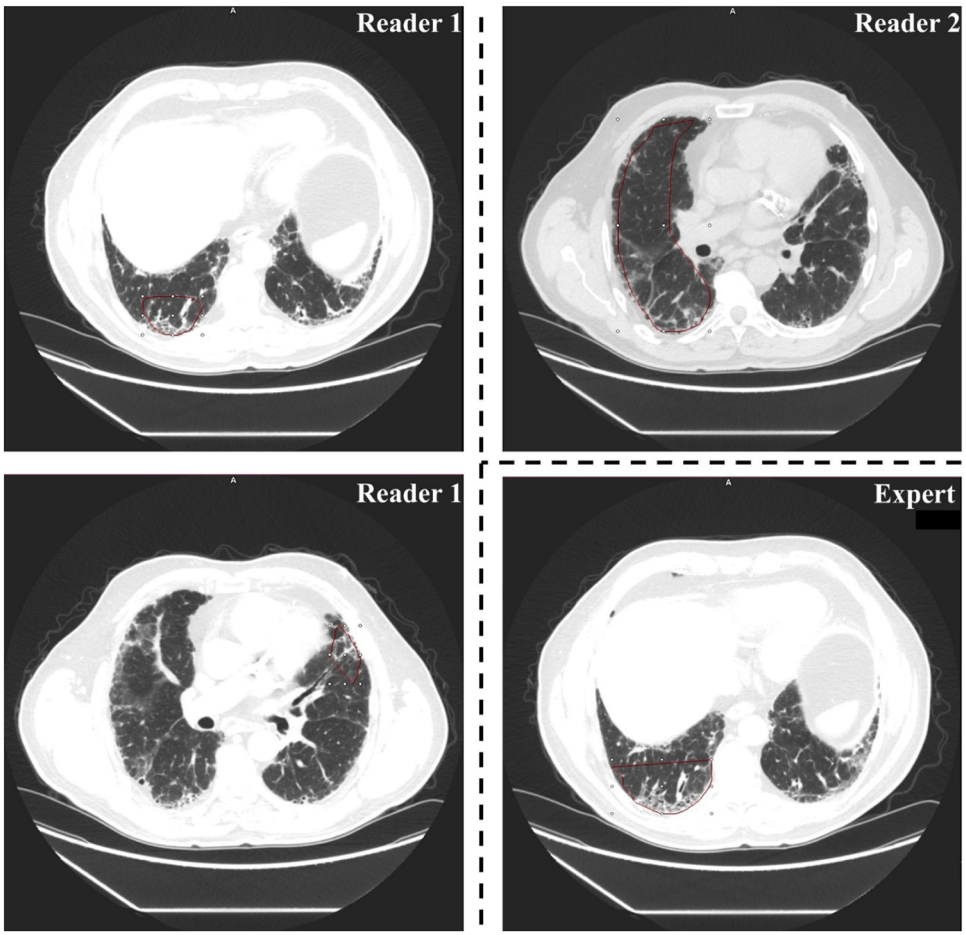


Fig. 4 Examples of ROIs used by different users in an identical scan. These images show axial high-resolution computed tomography slices of a patient with an usual interstitial pneumonia pattern. Left: ROIs placed by reader 1. Upper right: ROI placed by reader 2. Bottom right: ROI placed by expert radiologist. ROI, Region-of-interest

for in-training radiologists with a special focus on inter-observer input variability. Results demonstrated an increasing trend of diagnostic accuracy and confidence with the use of the CBIR software. Time spent on CBIR application use decreased with a growing number of use cases. Finally, analysis of input ROIs indicated the beneficial impact of multiple ROIs on the accuracy of the CBIR system.

The accuracy of readers showed an increasing trend with CBIR software use (35.7% *versus* 53.6%, $p = 0.302$), indicating its utility, particularly in low-experienced users. Pogarell et al [20] investigated a similar CBIR software showing improved diagnostic accuracy with CBIR use for students (30% *versus* 85.3%) and residents (60.7% *versus* 93.3%). A recent study by Haubold et al [22] reported for the same CBIR application a significant increase in diagnostic accuracy of residents with CBIR use (18.4% *versus* 33.6%). In contrast to our study design, readers in their study were not allowed to use any additional source of information for ILD assessment during the first session without CBIR application use. This may explain the relatively low increase of accuracy in our study compared to Haubold et al [22]. Additionally, Pogarell et al [20] and Haubold et al [22] included a great variety of ILD entities which can be expected to lead to a greater impact of CBIR use. Röhrich et al [21] evaluated a CBIR application including residents and attending radiologists. Their study also showed an increasing diagnostic accuracy with CBIR software use (34.7% *versus* 42.2%). The CBIR software retrieval output evaluated by Röhrich et al [21] was CT pattern-based, whereas the application investigated in our study provided a disease-based retrieval output. These discrepancies in CBIR results output format may cause differing interpretations by radiologists impacting diagnostic accuracy. Choe et al [19] reported a positive

trend of diagnostic accuracy using an in-house developed CBIR (46.1% *versus* 60.9%), including radiology and non-radiology residents, whereas our study focused on radiology residents. This and the lower number of ILD entities in our study might explain differing absolute accuracy values. Overall, the findings of our study and of previous published work suggest that CBIR applications lead to improved diagnostic accuracy of radiologists assessing ILD on chest CT. The fact that diagnostic accuracy was improved even when attending radiologists where included as in the study of Röhrich et al [21] indicates that CBIR applications are also valuable tools for more experienced users. Interestingly, the expert reader in our study perceived a high relevance of CBIR results in 96.4% of all cases, which supports this hypothesis.

In our study, CBIR accuracy showed that in 73.4% of cases, software results matched CT patterns of input ROIs. The database of the CBIR application evaluated in our study does not include normal cases and, therefore, requires sufficient discrimination between normal and pathologic CT cases by users. Radiology residents in our study showed 100% accuracy in detecting normal cases. This indicates that false positive results of CBIR software, when used in non-pathological CT datasets, are a minor risk since these cases can be expected to be detected by the user correctly. Choe et al [19] reported a software accuracy of 80% for matching the same disease class as the query CT, while Hwang et al [16] investigating the same algorithm showed an accuracy of 93.3% for the first retrieved case. Both Choe et al [19] and Hwang et al [16] used CT cases from the CBIR database for image query, whereas our study uses a fully independent dataset. This may cause discrepancies in CBIR accuracy values compared to our study since machine learning algorithms perform better in already-seen cases. Further, the CBIR algorithm used by Choe et al [19] and Hwang et al [16] assessed the whole lung volume. It remains unclear if whole lung assessment leads to improved accuracy because external testing of the algorithm from Choe et al [19] and Hwang et al [16] is missing. The software used in our study uses ROI-based input volumes potentially increasing user variability and error. Further research is needed to explore whether there are relevant differences in retrieval accuracy of CBIR applications using the whole lung volume compared to the ROI-based approach.

ROI analysis in our study showed that readers 1, 2, and the expert used 2.3, 1.1, and 1.0 ROI per patient, respectively. CBIR pattern accuracy was 77.7% (reader 1), 70.1% (reader 2), and 66.7% (expert), indicating an increasing trend of accurate pattern-based retrievals with an increasing number of ROIs per case. ROI use variability in terms of size and location could explain the differences between the CBIR accuracy of reader 2 and the expert. In most cases, reader 2 included peripheral and central lung volume together in relatively large ROIs, whereas the expert tends to do the same in only about half of the cases (91.4% *versus* 46.4%). These results indicate that CBIR software accuracy tends to increase with a growing number and increasing size of ROIs, maximizing the input information.

Time analysis in our study showed an increase in time for research per case when the CBIR software was introduced. However, in subsequent sessions, this time tended to decrease. Time overall per case decreased significantly in subsequent sessions with optional CBIR use. Haubold et al [22] reported an initial increase in time per case of +34% observed during the first half of CBIR use cases which significantly decreased to +7% in the second half. These findings are in line with the decreasing time trends observed in our study. Röhrich et al [21] reported similar results, showing a decrease in overall time used per case of 31% in readings with CBIR. These results support our hypothesis of a significant learning process for users with growing experience in CBIR software use. Like ROI placements, it can be assumed that users are improving the handling of CBIR applications in a reasonable number of use cases, leading to an overall decrease in ILD case interpretation times.

Considering the beneficial impact of CBIR applications on diagnostic accuracy and confidence of radiologists in ILD assessment on chest CT, coupled with an expected favorable learning curve in software use, clinical implementation of CBIR systems for daily use seems feasible. In addition, CBIR systems may be suitable for radiology training programs targeting attending radiologists in private practice or medical students to facilitate the acquisition of knowledge and skills in this field.

This study has several limitations that need to be mentioned. First, our case collection included only a low number of ILD entities with a potential impact on accuracy analysis and power for getting statistical significance. Further, we defined an independent reference standard for the pattern-based accuracy analysis. It is undisputable that ILDs include numerous pathologies, however, a not negligible number of ILD entities are very rare. This is why we focused on the most common pathologies, representing most differential diagnoses in clinical routine. Second, our study provides only a very limited number of in-training readers and a single expert reader who performed the reading sessions in rather small CT datasets. This might have led to a limitation of CT patterns included in the evaluation in terms of severity and variability of occurrence. Further studies with a larger

number of readers may be necessary to validate our findings and to explore the full extent of ROI variability's impact on CBIR accuracy.

In conclusion, the use of CBIR software has been demonstrated to improve diagnostic accuracy and confidence in in-training radiologists while reducing interpretation time in ILD assessment.

Abbreviations

CBIR	Content-based image retrieval
CT	Computed tomography
ILD	Interstitial lung disease
ROI	Region-of-interest

Acknowledgements

All authors declare that there were no Large Language Models (LLMs) used for writing or editing the submitted work.

Author contributions

All authors of the manuscript have made a significant contribution to the manuscript with regards to its conception, writing, and final approval. All authors reviewed the final version of the manuscript and agreed to submit it to the *European Radiology Experimental* for publication.

Funding

This study was conducted with research funding from Siemens Medical Solutions USA Inc. (Malvern, PA, United States of America).

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request. The source code of the evaluated content-based image retrieval system is not accessible to the public.

Declarations

Ethics approval and consent to participate

Ethical approval was authorized by the responsible Institutional Review Board of Emory University (IRB number: STUDY00002503; November 8, 2021) and need for written informed consent was waived.

Consent for publication

Not applicable.

Competing interests

CNDC receives research funding from Siemens Healthineers and Cleerly. MvA receives research funding from Siemens Healthineers and is a member of the scientific editorial board (section: Artificial intelligence, augmented reality, computer science, and radiomics) for *European Radiology Experimental;* as such, they have not participated in the selection nor review process for this article. FGM has received research funding from GE Healthcare and speaker honoraria from GE Healthcare, Circle Cardiovascular Imaging, and Bayer Vital. Others have no disclosures to report.

Author details

[1]Division of Cardiothoracic Imaging, Department of Radiology and Imaging Sciences, Emory University Hospital, Atlanta, GA, USA. [2]Institute of Diagnostic and Interventional Radiology, Pediatric Radiology and Neuroradiology, University Medical Centre Rostock, Rostock, Germany. [3]Clinical and Experimental Medicine PhD Program, University of Modena and Reggio Emilia, Modena, Italy. [4]Institute of Clinical Radiology and Nuclear Medicine, University Medical Center Mannheim, Medical Faculty Mannheim of Heidelberg University, Mannheim, Germany. [5]Department of Radiology, Gil Medical Center, Gachon University College of Medicine, Namdong-Daero 774 Beon-Gil, Namdong-gu, Incheon, South Korea.

Received: 21 July 2024 Accepted: 15 November 2024
Published online: 13 January 2025

References

1. Travis WD, Costabel U, Hansell DM et al (2013) An official American Thoracic Society/European Respiratory Society statement: update of the international multidisciplinary classification of the idiopathic interstitial pneumonias. Am J Respir Crit Care Med 188:733–748. https://doi.org/10.1164/rccm.201308-1483ST
2. Lynch DA, Sverzellati N, Travis WD et al (2018) Diagnostic criteria for idiopathic pulmonary fibrosis: a Fleischner Society White Paper. Lancet Respir Med 6:138–153. https://doi.org/10.1016/s2213-2600(17)30433-2
3. Raghu G, Remy-Jardin M, Ryerson CJ et al (2020) Diagnosis of hypersensitivity pneumonitis in adults. An official ATS/JRS/ALAT Clinical Practice Guideline. Am J Respir Crit Care Med 202:e36–e69. https://doi.org/10.1164/rccm.202005-2032ST
4. Adegunsoye A, Ryerson CJ (2021) Diagnostic classification of interstitial lung disease in clinical practice. Clin Chest Med 42:251–261. https://doi.org/10.1016/j.ccm.2021.03.002
5. Raghu G, Remy-Jardin M, Richeldi L et al (2022) Idiopathic pulmonary fibrosis (an update) and progressive pulmonary fibrosis in adults: an official ATS/ERS/JRS/ALAT Clinical Practice Guideline. Am J Respir Crit Care Med 205:e18–e47. https://doi.org/10.1164/rccm.202202-0399ST
6. Walsh SLF, Wells AU, Desai SR et al (2016) Multicentre evaluation of multidisciplinary team meeting agreement on diagnosis in diffuse parenchymal lung disease: a case-cohort study. Lancet Respir Med 4:557–565. https://doi.org/10.1016/s2213-2600(16)30033-9
7. Xu X, Sui X, Song L et al (2019) Feasibility of low-dose CT with spectral shaping and third-generation iterative reconstruction in evaluating interstitial lung diseases associated with connective tissue disease: an intra-individual comparison study. Eur Radiol 29:4529–4537. https://doi.org/10.1007/s00330-018-5969-y
8. Zhao R, Sui X, Qin R et al (2022) Can deep learning improve image quality of low-dose CT: a prospective study in interstitial lung disease. Eur Radiol 32:8140–8151. https://doi.org/10.1007/s00330-022-08870-9
9. Kim CH, Chung MJ, Cha YK et al (2023) The impact of deep learning reconstruction in low dose computed tomography on the evaluation of interstitial lung disease. PLoS One 18:e0291745. https://doi.org/10.1371/journal.pone.0291745
10. Flaherty KR, Andrei A-C, King TE et al (2007) Idiopathic interstitial pneumonia: do community and academic physicians agree on diagnosis? Am J Respir Crit Care Med 175:1054–1060. https://doi.org/10.1164/rccm.200606-833OC
11. Walsh SLF, Calandriello L, Sverzellati N et al (2016) Interobserver agreement for the ATS/ERS/JRS/ALAT criteria for a UIP pattern on CT. Thorax 71:45–51. https://doi.org/10.1136/thoraxjnl-2015-207252
12. Depeursinge A, Fischer B, Müller H et al (2011) Prototypes for content-based image retrieval in clinical practice. Open Med Inform J 5:58–72. https://doi.org/10.2174/1874431101105010058
13. Wei G, Qiu M, Zhang K et al (2020) A multi-feature image retrieval scheme for pulmonary nodule diagnosis. Medicine (Baltimore) 99:e18724. https://doi.org/10.1097/MD.0000000000018724
14. Zhang K, Qi S, Cai J et al (2022) Content-based image retrieval with a convolutional siamese neural network: distinguishing lung cancer and tuberculosis in CT images. Comput Biol Med 140:105096. https://doi.org/10.1016/j.compbiomed.2021.105096
15. Choe J, Choi HY, Lee SM et al (2024) Evaluation of retrieval accuracy and visual similarity in content-based image retrieval of chest CT for obstructive lung disease. Sci Rep 14:4587. https://doi.org/10.1038/s41598-024-54954-5
16. Hwang HJ, Seo JB, Lee SM 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. https://doi.org/10.3348/kjr.2020.0603
17. Oosawa A, Kurosaki A, Kanada 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 57:490–498. https://doi.org/10.1016/j.resinv.2019.03.015

18. Walsh SLF, Calandriello L, Silva M et al (2018) Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort study. Lancet Respir Med 6:837–845. https://doi.org/10.1016/s2213-2600(18)30286-8
19. Choe J, Hwang HJ, Seo JB et al (2022) Content-based image retrieval by using deep learning for interstitial lung disease diagnosis with chest CT. Radiology 302:187–197. https://doi.org/10.1148/radiol.2021204164
20. Pogarell T, Bayerl N, Wetzl M et al (2021) Evaluation of a novel content-based image retrieval system for the differentiation of interstitial lung diseases in CT examinations. Diagnostics 11. https://doi.org/10.3390/diagnostics11112114
21. Röhrich S, Heidinger BH, Prayer F et al (2023) Impact of a content-based image retrieval system on the interpretation of chest CTs of patients with diffuse parenchymal lung disease. Eur Radiol 33:360–367. https://doi.org/10.1007/s00330-022-08973-3
22. Haubold J, Zeng K, Farhand S et al (2023) AI co-pilot: content-based image retrieval for the reading of rare diseases in chest CT. Sci Rep 13:4336. https://doi.org/10.1038/s41598-023-29949-3
23. Xie S, Girshick R, Dollár P et al (2016) Aggregated residual transformations for deep neural networks. https://arxiv.org/pdf/1611.05431
24. Russakovsky O, Deng J, Su H et al (2014) ImageNet large scale visual recognition challenge. https://doi.org/10.48550/arXiv.1409.0575
25. Schroff F, Kalenichenko D, Philbin J (2015) FaceNet: A unified embedding for face recognition and clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015): Boston, Massachusetts, USA, 7–12 June 2015. IEEE, Piscataway, NJ, pp 815–823
26. Sohn K (2016) Improved deep metric learning with multi-class N-pair loss objective. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. Curran Associates Inc, Red Hook, NY, USA, pp 1857–1865

Publisher's Note