



Optimization Strategies for Interactive Classification of Interstitial Lung Disease Textures

Thessa T. J. P. Kockelkorn^{1*}, Rui Ramos², José Ramos², Pim A. de Jong³, Cornelia M. Schaefer-Prokop^{4,5}, Rianne Wittenberg^{3†}, Audrey M. Tiehuis³, Jan C. Grutters^{6,7}, Max A. Viergever¹ and Bram van Ginneken^{1,5}

¹Image Sciences Institute, University Medical Center Utrecht, Utrecht, Netherlands, ²Instituto de Engenharia Biomédica, Faculdade de Engenharia da Universidade do Porto, Porto, Portugal, ³Department of Radiology, University Medical Center Utrecht, Utrecht, Netherlands, ⁴Department of Radiology, Meander Medical Center, Amersfoort, Netherlands, ⁵Diagnostic Image Analysis Group, Radboudumc, Nijmegen, Netherlands, ⁶Department of Pulmonology, Center of Interstitial Lung Diseases, St Antonius Hospital, Nieuwegein, Netherlands, ⁷Heart and Lungs Division, University Medical Center Utrecht, Utrecht, Netherlands

OPEN ACCESS

Edited by:

Antoine Vacavant,
University of Auvergne, France

Reviewed by:

Alex Pappachen James,
Nazarbayev University, Kazakhstan
Diana Rodrigues Pina,
São Paulo State University, Brazil

*Correspondence:

Thessa T. J. P. Kockelkorn
thessa@gmail.com

†Present address:

Rianne Wittenberg,
Department of Radiology, VU Medical
Center, Amsterdam, Netherlands

Specialty section:

This article was submitted to
Computer Image Analysis,
a section of the journal
Frontiers in ICT

Received: 31 August 2016

Accepted: 13 December 2016

Published: 27 December 2016

Citation:

Kockelkorn TTJP, Ramos R, Ramos J, de Jong PA, Schaefer-Prokop CM, Wittenberg R, Tiehuis AM, Grutters JC, Viergever MA and van Ginneken B (2016) Optimization Strategies for Interactive Classification of Interstitial Lung Disease Textures. *Front. ICT* 3:33. doi: 10.3389/fict.2016.00033

For computerized analysis of textures in interstitial lung disease, manual annotations of lung tissue are necessary. Since making these annotations is labor intensive, we previously proposed an interactive annotation framework. In this framework, observers iteratively trained a classifier to distinguish the different texture types by correcting its classification errors. In this work, we investigated three ways to extend this approach, in order to decrease the amount of user interaction required to annotate all lung tissue in a computed tomography scan. First, we conducted automatic classification experiments to test how data from previously annotated scans can be used for classification of the scan under consideration. We compared the performance of a classifier trained on data from one observer, a classifier trained on data from multiple observers, a classifier trained on consensus training data, and an ensemble of classifiers, each trained on data from different sources. Experiments were conducted without and with texture selection (ts). In the former case, training data from all eight textures was used. In the latter, only training data from the texture types present in the scan were used, and the observer would have to indicate textures contained in the scan to be analyzed. Second, we simulated interactive annotation to test the effects of (1) asking observers to perform ts before the start of annotation, (2) the use of a classifier trained on data from previously annotated scans at the start of annotation, when the interactive classifier is untrained, and (3) allowing observers to choose which interactive or automatic classification results they wanted to correct. Finally, various strategies for selecting the classification results that were presented to the observer were considered. Classification accuracies for all possible interactive annotation scenarios were compared. Using the best-performing protocol, in which observers select the textures that should be distinguished in the scan and in which they can choose which classification results to use for correction, a median accuracy of 88% was reached. The results obtained using this protocol were significantly better than results obtained with other interactive or automatic classification protocols.

Keywords: interactive annotation, interstitial lung disease, texture, classification, computer-aided diagnosis

INTRODUCTION

In medical image analysis, obtaining reliable ground truth annotations is of pivotal importance for the training, testing, and comparison of algorithms. Several factors may hamper the construction of such a ground truth dataset. To start, making annotations is expensive, as it requires a substantial amount of human observer effort. But before observers can start annotating, a selection of scans needs to be made. This selection process has a large influence on the quality and usability of the resulting dataset. When choosing images from a single institution, with similar acquisition and reconstruction parameters, training and testing are done on a homogeneous dataset. Good performance of an algorithm on such a dataset is by no means a guarantee of comparable results on other datasets. However, when choosing to collect a set of images with varying acquisition parameters, training becomes more complicated (van Opbroek et al., 2015). In this case, probably more training data are needed to obtain results similar to the ones that can be obtained using a homogeneous dataset. Finally, another complication arises when the ground truth is difficult to establish. The presence or absence of a bone fracture will in most cases not give rise to much debate, inasmuch as most bone fractures can be established objectively. For more subtle lesions, matters readily become more complicated, since different observers may have different opinions on the interpretation of these lesions [e.g., Watadani et al. (2013) and van Riel et al. (2015)].

In this work, we consider interstitial lung disease (ILD), a group of around 200 inflammatory and fibrotic lung diseases that mainly affect the tissue and space around the air sacs of the lungs. These diseases have distinct and also considerably overlapping imaging features. Since the individual diseases have substantially different treatment options and prognosis, it is important to make the correct diagnosis. Computed tomography (CT) scans play a pivotal role in the interdisciplinary process of making a diagnosis of ILD (Aziz et al., 2006). Automatic classification of normal and abnormal lung tissue in CT scans of ILD patients has been studied extensively, focusing on finding optimal feature sets and classifiers (Uppaluri et al., 1999; Xu et al., 2006; Zavaletta et al., 2007; Park et al., 2009; Depeursinge et al., 2010, 2012; Huber et al., 2011, 2012; Vasconcelos et al., 2015). Recently, deep-learning techniques have been applied to the problem (Anthimopoulos et al., 2016; Gao et al., 2016).

This study focuses on the process of obtaining annotations that can be used in ILD texture analysis. For scans from patients with ILD, all complicating aspects mentioned above may occur when compiling an annotated dataset. First, manual delineation of all lung textures present in a volumetric CT scan is a labor-intensive task, especially when the disease affected a large part of the lungs. Second, thoracic CT scans of ILD patients may be made using various CT protocols (Prosch et al., 2013). Finally, analysis of imaging features and therefore also annotations vary substantially even among experienced radiologists. Therefore, obtaining ground truth annotations is not trivial.

Since automatic ILD annotation systems may not be able to adapt to various CT acquisition protocols, let alone to different annotation preferences, we have developed a system for

interactive annotation of 3D volumes of interest (VOIs), which we applied to scans of ILD patients (Kockelkorn et al., 2010, 2016). This method allows observers to quickly annotate ILD textures in CT scans. A human observer trains the system continuously by correcting classification results in a slice-by-slice manner. In this way, the algorithm becomes increasingly better in annotating textures. The smaller the amount of user input required to obtain completely annotated lungs, the easier it becomes to obtain a large number of annotated datasets, which can then be used to study effects of varying acquisition parameters on the one hand and different annotation preferences on the other hand.

In this work, we aim to optimize the interactive annotation procedure, in order to decrease the numbers of VOIs for which the observer has to correct the computer-generated label. We investigate how annotations of VOIs in other previously annotated ILD scans can be used for classification of unseen VOIs in the scan under consideration. In addition, we evaluate various ways in which users transfer their knowledge to the interactive annotation environment. Finally, we compared different strategies for selection of the VOIs that are shown to the observer for correction.

This paper is structured as follows: Section “Materials” describes the CT scans used for the experiments. Section “Methods” details the processes of automatic and interactive classification, followed by an outline of the experiments that were performed in Section “Manual Annotation of VOIs.” Section “Results” contains the results of the experiments. In Section “Discussion,” the main insights resulting from this work are summarized and their relevance is discussed.

MATERIALS AND METHODS

Materials

For this project, 23 clinically indicated, standard-dose thoracic CT scans of ILD patients were collected retrospectively. Scans were acquired between 2004 and 2010 at the St. Antonius Ziekenhuis Nieuwegein, the Netherlands, on a Philips Mx8000 IDT or a Philips Brilliance iCT scanner (Philips Medical Systems, Best, the Netherlands). Scans were taken at full inspiration with patients in supine position, without contrast material. Data were acquired in spiral mode and reconstructed to 512×512 or 768×768 matrices. Patient and scanning protocol parameters are summarized in **Table 1**.

Methods

Preprocessing

Interactive annotation of VOIs in the lungs has been described previously (Kockelkorn et al., 2016). To summarize, lungs in the CT scans were segmented (van Rikxoort et al., 2009) and subdivided into roughly spherical VOIs, containing only one type of texture, using the algorithm described in Kockelkorn et al. (2016). On average, lungs contained 2,114 VOIs (range: 1,148–3,313).

Features and Classifier

For all VOIs, 72 rotationally invariant features were calculated. Scans were filtered using Gaussian, Laplacian, gradient

TABLE 1 | Patient and scan characteristics.

Scan number	Patient age	Patient sex	Number of volumes of interest	In-plane resolution (mm)	Slice spacing (mm)	Peak voltage (kV)	Tube current (mA)
1	67	Female	2,021	0.605	0.8	120	120
2	20	Female	2,786	0.574	0.8	120	217
3	33	Male	3,084	0.688	0.8	120	144
4	45	Male	1,647	0.873	0.8	120	144
5	42	Female	2,040	0.658	0.8	120	144
6	61	Male	1,234	0.781	1.0	120	192
7	24	Male	2,818	0.645	0.7	120	206
8	32	Male	1,459	0.758	0.8	120	144
9	62	Female	1,368	0.688	0.5	120	188
10	38	Female	2,386	0.586	1.0	120	270
11	57	Female	1,148	0.660	0.8	120	90
12	71	Male	1,761	0.707	0.8	120	144
13	51	Male	2,017	0.652	0.8	120	144
14	57	Male	1,692	0.750	0.8	120	90
15	33	Male	1,701	0.666	0.8	120	90
16	52	Male	1,810	0.676	0.5	120	188
17	30	Male	3,177	0.658	0.8	120	90
18	39	Female	3,313	0.627	0.5	120	188
19	58	Male	1,687	0.688	0.8	120	90
20	49	Female	2,233	0.580	0.8	120	150
21	54	Female	1,681	0.672	0.8	120	90
22	53	Female	2,540	0.411	4.0	120	125
23	32	Female	2,579	0.652	0.8	120	90

magnitude, and three Hessian Eigenvalue-based filters. Each of these six filters was applied at three scales ($\sigma = 1, 2, \text{ and } 4$ voxels). This resulted in 18 filtered images. In each filtered image, mean, SD, kurtosis, and skew of the VOI histogram were calculated and used as features. Because training of the classifier is done while the observer is annotating a scan, we used an Extra-Trees classifier [99 trees, 10 features per node, a minimum number of samples after each split of 1; parameters based on the recommendations of Geurts et al. (2006)]. This classifier can be trained relatively fast, since the training process can be divided over multiple cores.

Manual Annotation of VOIs

Three radiologists independently performed manual labeling of all VOIs in a subset of the set of 23 scans. Two radiologists labeled 6 scans and one radiologist labeled 21 scans. In total, 17 scans were annotated by one observer, 2 were annotated by 2 observers, and 4 were annotated by 3 observers.

Observers were instructed to annotate the following textures (see **Figure 1**. Examples of the eight texture classes.):

Normal tissue: lung tissue without any abnormalities;

Ground glass: increased lung density, in which underlying structures are still visible;

Consolidation: increased lung density, in which underlying structures are no longer visible;

Honeycombing: cystic destruction of subpleural lung parenchyma: there are cysts of varying diameter (0.3–1.0 cm) in several layers and cysts share relatively thick walls;

Decreased density: decreased density compared with normal lung parenchyma, with or without surrounding walls;

Crazy paving: regular pattern of ground glass with a reticular pattern;

NSIP pattern: ground glass with architectural distortion, traction bronchiectasis, or irregular lines; and

Nodular pattern: sharply defined nodular densities (1–4 mm) in a random or paralympathic (paraseptal) distribution. Nodules can also have branching structures (tree-in-bud).

Thus, in this work an 8-class classification problem is studied. **Table 2** shows the percentages of VOIs assigned to each of the eight classes in the resulting dataset.

Simulated Interactive Annotation

We used simulation software to investigate the effect of design choices in interactive annotation protocols on the percentage of VOIs that were correctly classified. Interactive annotation is schematically depicted in **Figure 2**. Initially, an axial slice was chosen at random, and VOIs intersecting with this slice were labeled automatically. Automatically generated labels were compared to the manual labels that were provided by the human observers and incorrect labels were changed. All labeled VOIs were used to train an Extra-Trees classifier. The second axial slice was chosen, and all VOIs intersecting with the slice were classified by the classifier. Also in this slice, automatic classifications were compared to manual annotations, and incorrect labels were changed. The VOIs in this second slice were added to the training dataset, on which the Extra-Trees classifier was retrained. The cycle of correction, retraining, and classification was repeated until at least 50% of all VOIs were annotated. The remainder of the scan was then classified automatically and subsequently checked against the manual annotations. The

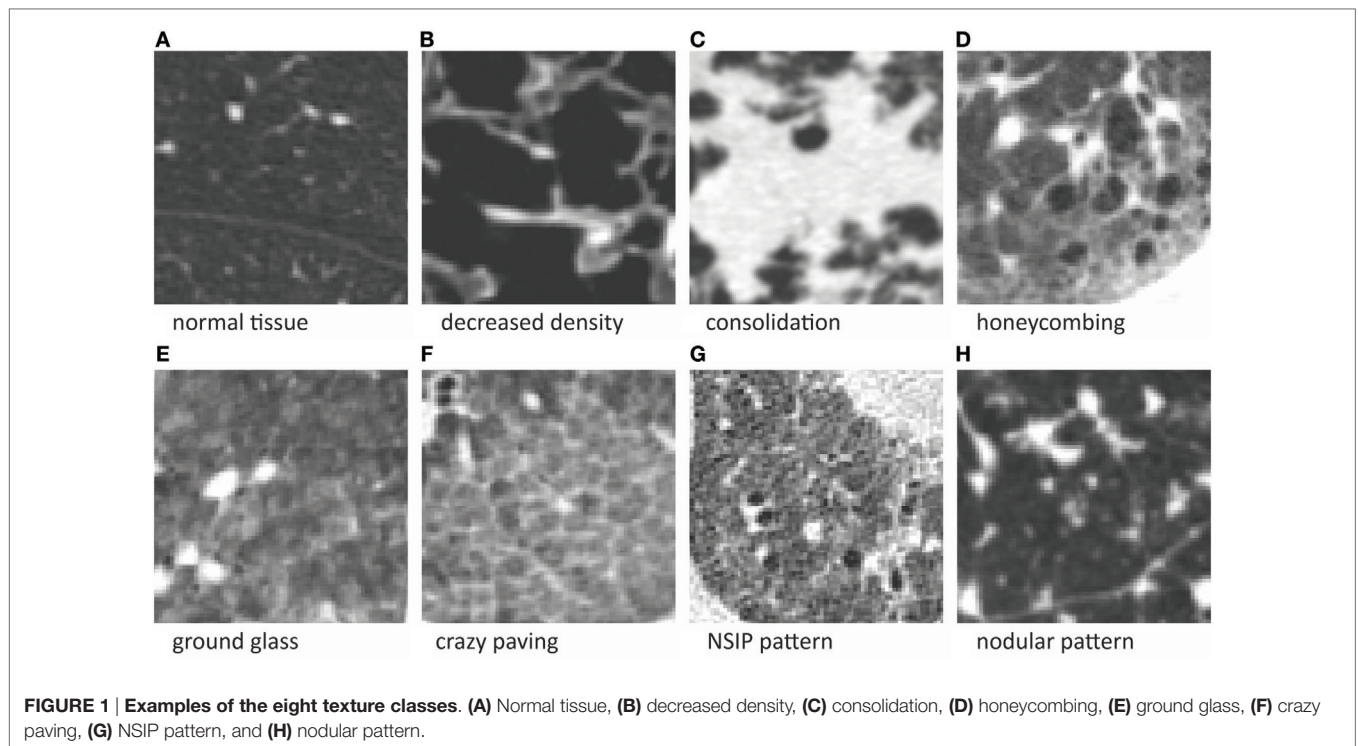


TABLE 2 | Percentages of volumes of interest assigned to the 8 texture classes.

Normal tissue (%)	Decreased density (%)	Consolidation (%)	Honeycombing (%)	Ground glass (%)	Crazy paving (%)	NSIP pattern (%)	Nodular pattern (%)
55	15	1	2	8	6	1	11

number of correctly classified VOIs was used as a performance indicator.

Experiments

To determine an optimal protocol for interactive annotation, three types of experiments were conducted. First, automatic classification experiments were done to evaluate which type of previous training data yielded the best classification results in unseen scans. Second, we evaluated different interactive annotation scenarios. We investigated several ways to exploit prior knowledge about the annotation task. Finally, experiments were conducted to compare four strategies for slice selection. As all experiments included random selection of training samples, they were repeated five times, after which results were averaged.

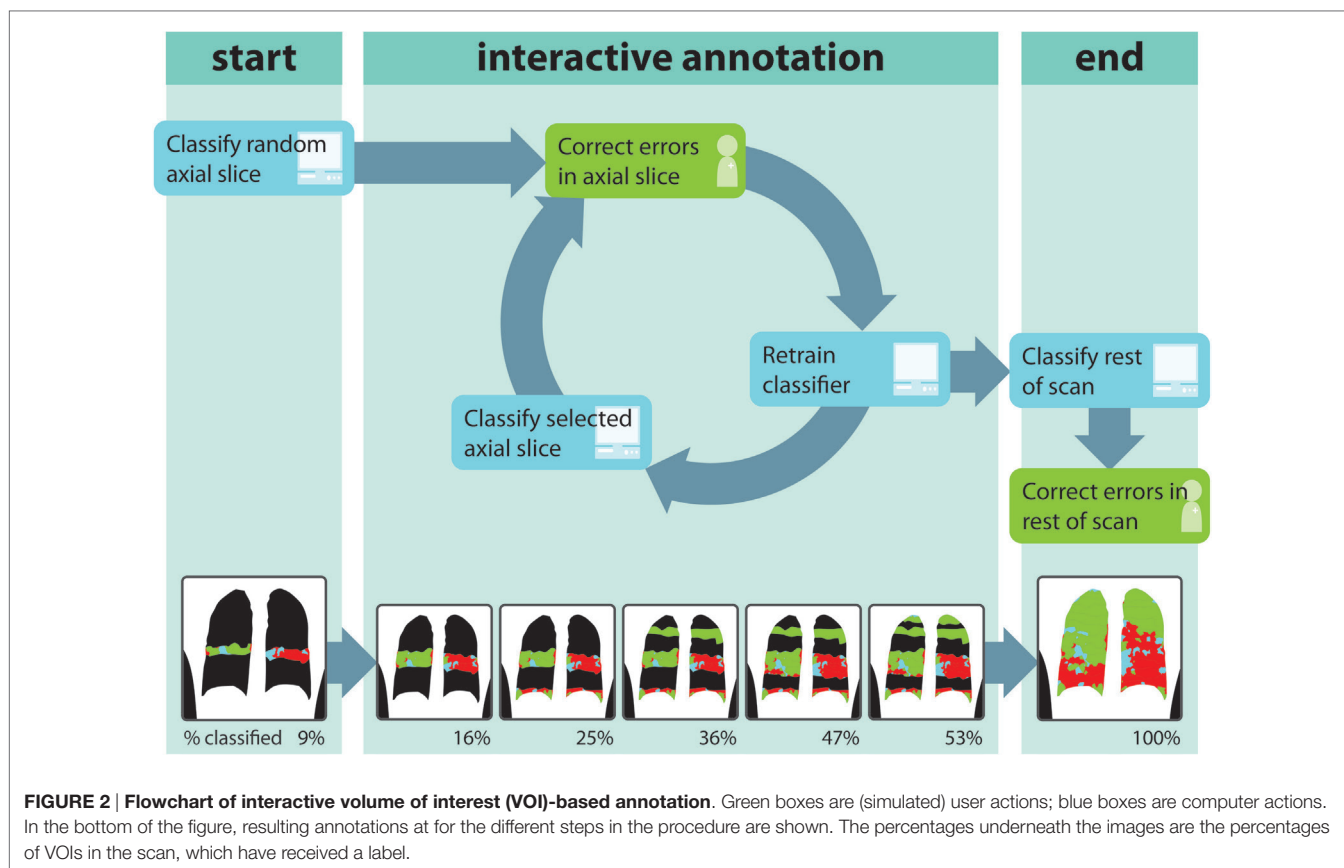
Automatic Classification Experiments

Four automatic classification protocols were tested in a leave-one-scan-out setup (see **Figure 3**). In all automatic protocols, 100, 250, or 500 training VOIs were selected for each texture pattern. If a texture was represented by less than the intended number of samples, all samples were used. In the first approach (a1), we trained an Extra-Trees classifier on training data from other scans annotated by the same observer who annotated the scan

under consideration. In the second approach (a2), the classifier was trained using training data from other scans annotated by all three observers. In the third approach, we trained a classifier on consensus training data (a3c). This training data set was obtained by selecting VOIs that were labeled two or three times. In the first case, the two observers had to agree on the label. In the second case, all three observers had to agree. If less than 100 consensus samples were available for a texture, regular training samples from all users were used. In the fourth approach, we used an ensemble classification strategy. For each dataset annotated by observer *x*, five Extra-Trees classifiers were trained on different types of training data:

- I. training data from other scans annotated by observer *x*;
- II. training data from other scans annotated by all observers;
- III. consensus training data from other scans annotated by all observers;
- IV. training data from other scans annotated by observer *y*; and
- V. training data from other scans annotated by observer *z*.

The final class label was determined by voting. In case of a tie, the class with the highest posterior probability was chosen. Results were obtained for fully automatic classification of all annotated datasets.



Experiments with approaches a1, a3, a3c, and ae were performed with and without texture selection (ts). Without ts, training data from all eight texture types were used to train the classifier. With ts, only training data from the texture types present in the scan were used (see Texture Selection).

Interactive Classification Experiments

Interactive classification was simulated in a leave-one-scan-out setup, as described in Kockelkorn et al. (2016). We investigated three manners to decrease annotation effort.

Texture Selection

First, we studied the effect on classification accuracy of indicating five VOIs of each texture present in the scan before the start of annotation in the different interactive protocols. This approach is illustrated in **Figure 4A**. The major difference in these protocols and the ones in **Figure 4B** was the way in which the VOIs in the first axial slice were classified. In the left protocol (i-ts), classification was performed interactively by a classifier trained on the VOIs selected by the simulated observer before training of the interactive classifier. In the protocols using automatic classification results for the classification of the first slice (i-a1, i-a3, i-a3c, and i-ae), only training samples from the texture categories indicated by the user before annotation were used. Obviously, the classification problem is simplified considerably if the system knows from the start which of the eight texture classes are present in the scan.

Training Data from Previously Annotated Scans

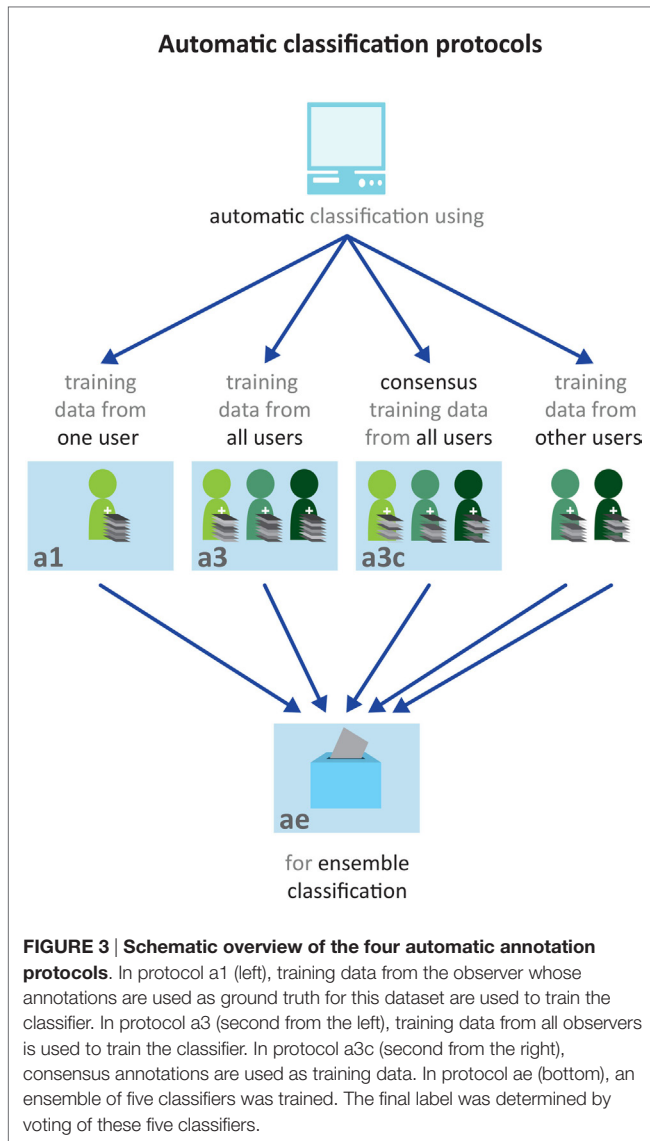
Second, we investigated the effect of using classifiers trained on training data obtained from previously annotated scans. These automatic classification results could be used at the beginning of interactive annotation, when little or no training data from the scan under consideration is available.

We tested the following annotation protocols, schematically depicted in **Figure 4A** (with ts) and **Figure 4B** (without ts):

- i: completely interactive annotation, without the use of previous training data;
- i-a1, i-a3, i-a3c, and i-ae: classification of VOIs in the first slice using one of the four automatic classification methods described above, followed by interactive annotation; and
- i-cc: annotation in which observers could determine per slice which classification results they wanted to use as a starting point for corrections—this could be i, a1, a3, a3c, or ae.

Classifier Choice

Third, we investigated the scenario in which observers were given the option to choose from different classification results when correcting the labels of the individual VOIs. The protocols based on this approach are i-cc-ts in **Figure 4A** and i-cc in **Figure 4B**. Users could choose between interactive classification results and results of a classifier trained on data from other scans annotated by the observers themselves (a1), a classifier trained on data from other scans annotated by all observers (a3), a classifier trained on



consensus training data from other scans (a3c), and the ensemble classification method (ae). For correction of the first axial slice in the absence of previous training data, without *ts*, the user could choose for heuristic labeling instead of interactive labeling. From these different classification results, the one requiring the lowest number of corrections was selected in the simulations.

Slice Selection

Finally, the way in which slices are presented to the observer influences the efficiency of the training of the classifier. We compared random slice selection with selection based on cumulative uncertainty U of n unassigned VOIs per slice:

$$U = \sum_{v=1}^n 1 - c_v$$

In this formula, the uncertainty of the classifier for VOI v was calculated by subtracting the confidence c_v from one. c_v was the highest posterior probability among the classes. By adding the

uncertainties for all unassigned VOIs in a slice, its cumulative uncertainty was determined. In each classification and correction cycle, the slice with the highest cumulative uncertainty was chosen. We chose to use cumulative instead of average uncertainty, since the first method favored slices with large numbers of VOIs. Larger numbers of VOIs per slice meant that the cycle of retraining, classification, and correction had to be repeated a lower number of times.

In our previous study, we subdivided the lungs into five levels in axial direction, numbered 1–5 from the apex to the base of the lungs (Kockelkorn et al., 2016). Slices are chosen from alternating levels in the following order: level 3, level 5, level 2, level 4, and level 1. This sequence is repeated if necessary. In this work, we compared random and uncertainty-based slice selection in a scenario in which this subdivision was used. In addition, we tested the effects of random and uncertainty-based slice selection when the lungs were not subdivided and slices could be chosen from the apex to the base of the lungs in each classification and correction cycle. This resulted in the following slice selection methods:

- I. random selection from the entire lungs;
- II. uncertainty-based selection from the entire lungs;
- III. random selection from different levels in the lungs; and
- IV. uncertainty-based selection from different levels in the lungs.

Evaluation

In all automatic experiments, classification accuracy was calculated per annotated dataset. For the interactive classification experiments, classification accuracy was calculated for each classified slice, for classification of the remainder of the scan after training on at least 50% of all VOIs, and for the complete dataset. In all cases, results for the five repetitions per experiments were averaged per annotated dataset. If a scan was annotated by more than one observer, results were calculated for all two or three annotations separately.

Repeated measures ANOVA was performed to test the difference in overall accuracy between the interactive protocols without *ts* (i, i-a1, i-a3, i-a3, i-ae, and i-cc), between the best-performing interactive protocol with and without *ts*, and between the best-performing interactive protocol with *ts* and the four automatic protocols (a1, a3, a3c, and ae) with *ts*.

RESULTS

Interobserver Agreement

Four scans in our dataset were annotated by three observers and two scans were annotated by two observers. **Table 3** shows the results of the comparison of the labels of the VOIs that were annotated more than once. Four thousand seven hundred twenty-four VOIs were annotated twice. Interobserver agreement was 69%. Eight thousand four hundred ninety-eight VOIs were annotated three times. A total of 35% of the VOIs were given the same label by all three observers. For another 35% of the VOIs, two observers agreed on the label and one observer had a different opinion. The final 30% of the VOIs received a different label from each of the three observers.

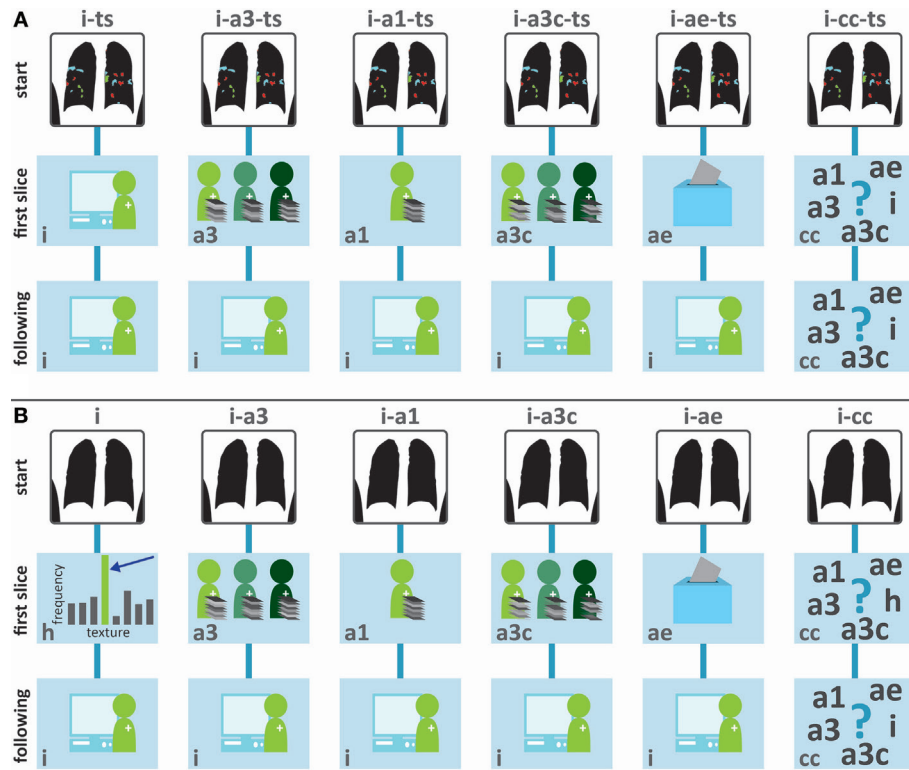


FIGURE 4 | (A) Schematic overview of the tested interactive annotation protocols with texture selection (ts). The top row displays how classification is initiated: the user selected five example VOIs of each texture class present in the scan. The images in the second row indicate the classification procedure for the VOIs in the first slice. This can be done interactively, by training the classifier on the VOIs that the user indicated before the start of annotation (i, interactive), by using a classifier trained on the automatic classification protocols in **Figure 3** (a1, a3, a3c, and ae), or by letting observers choose the classification results that they want to correct (cc, classifier choice). The third row indicates how the following slices are classified: either interactively (i) or by letting observers choose the classification results that they want to correct (cc). **(B)** Schematic overview of the tested interactive annotation protocols without ts. The top row displays how classification is initiated: in this case, no user action is required at this stage. The second and third rows are similar to panel **(A)**.

TABLE 3 | Interobserver agreement in % for VOIs that were annotated two or three times.

	Nr of VOIs	% of VOIs that received		
		3 labels	2 labels	1 label
Two times annotated	4,724	–	31	69
Three times annotated	8,498	30	35	35

Figure 5 shows an example of an axial slice (a), with in the second row manual VOI annotations as made by observer 1 (d), observer 2 (e), and observer 3 (f). All observers agree that this slice contains normal tissue and crazy paving, but the distribution of the textures varies between the observers. In addition, observer 2 has indicated areas of ground glass, which are absent in the annotations of observers 1 and 3.

Automatic Classification Results

Use of Previous Training Data

In **Figure 6A**, the results of the use of different protocols for automatic classification—namely using training data from the

observer who provided annotations for the scan under consideration (a1), using training data from all observers (a3), using consensus training data (a3c), and building an ensemble classifier (ae)—are displayed. The lower border of each box is the first quartile (Q1); the upper border is the third quartile (Q3). The median value is indicated by the horizontal bar inside the box. Upper and lower whiskers extend to the maximum and minimum data point, respectively. Results are shown for three different intended sizes of the training dataset: 100, 250, and 500 training samples per texture class. For all four classification protocols, adding more training samples yielded higher median classification accuracy. This also held for the maximum accuracy and for Q3. The minimum accuracy was 0 or close to 0 for all different protocols and for the three different intended training dataset sizes. The interquartile range (IQR), calculated as Q3–Q1, increased with an increasing intended number of training samples per texture class for the protocols a1 and a3c.

Table 4 shows the median, minimum, and maximum accuracies for all four automatic classification protocols, for each of the three observers, and for all observers taken together.

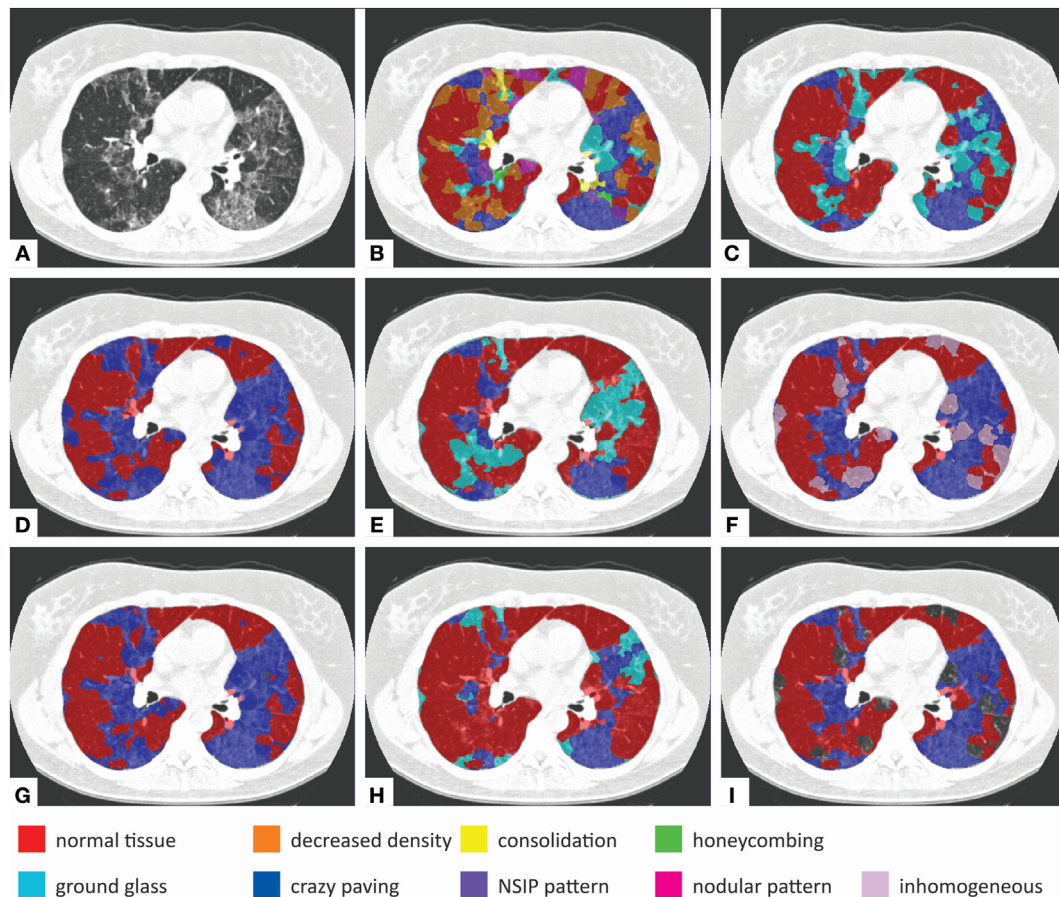


FIGURE 5 | Examples of automatic and interactive classification results. (A) Axial slice of a computed tomography scan of an interstitial lung disease patient. **(B)** Results of automatic classification using training data from all observers without texture selection (ts). **(C)** Results of automatic classification with ts. **(D)** Manual VOI labeling by observer 1. **(E)** Manual VOI labeling by observer 2. **(F)** Manual VOI labeling by observer 3. **(G)** Interactive classification results after training on at least 50% of all VOIs for observer 1. **(H)** Interactive classification results after training on at least 50% of all VOIs for observer 2. **(I)** Interactive classification results after training on at least 50% of all VOIs for observer 3.

The intended number of training samples per texture class was 500. The protocol in which consensus training data were used yielded the highest median percentage of correctly classified VOIs for observer 1, observer 3, and for all observers together. The protocol yielding the highest median accuracy for observer 2 was a1, in which only training data from this observer was used. The minimum percentage of correctly classified VOIs varied between the different observers: for observer 1, this percentage varied from 22 to 35%. For observer 3, values ranged between 8% for protocol a1 and 18% for protocol a3c. Minimum values were lowest for observer 2.

Texture Selection

Figure 6B shows the effect of ts on classification accuracy for the four automatic protocols. When using ts, median classification accuracy substantially increased for all protocols as compared to the results without ts. Minimum and maximum accuracy also showed this increase. Similar to the situation in which no ts was

applied, median classification accuracy increased with increasing size of the training dataset.

In **Table 5**, median, minimum, and maximum classification accuracy after ts are shown for the individual observers and for all observers together. The spread in median accuracies between the different protocols was smaller than when no ts was applied. For all observers, median accuracy ranged from 73% for protocols a3c and a3 to 77% for protocol a1. This latter was an increase of 27% points as compared to the scenario in which no ts was performed. The same could be seen for all individual observers: the spread in median accuracy between the classification protocols was also smaller when ts was applied.

Figures 5B,C show the effect of ts. In both panels, VOIs were automatically classified using training data from all observers. In **Figure 5B**, this was done without ts, in **Figure 5C** with ts. Without ts, the classifier predicted the occurrence of all texture classes in this slice; with ts, the areas containing normal tissue as indicated by all observers are more accurately classified.

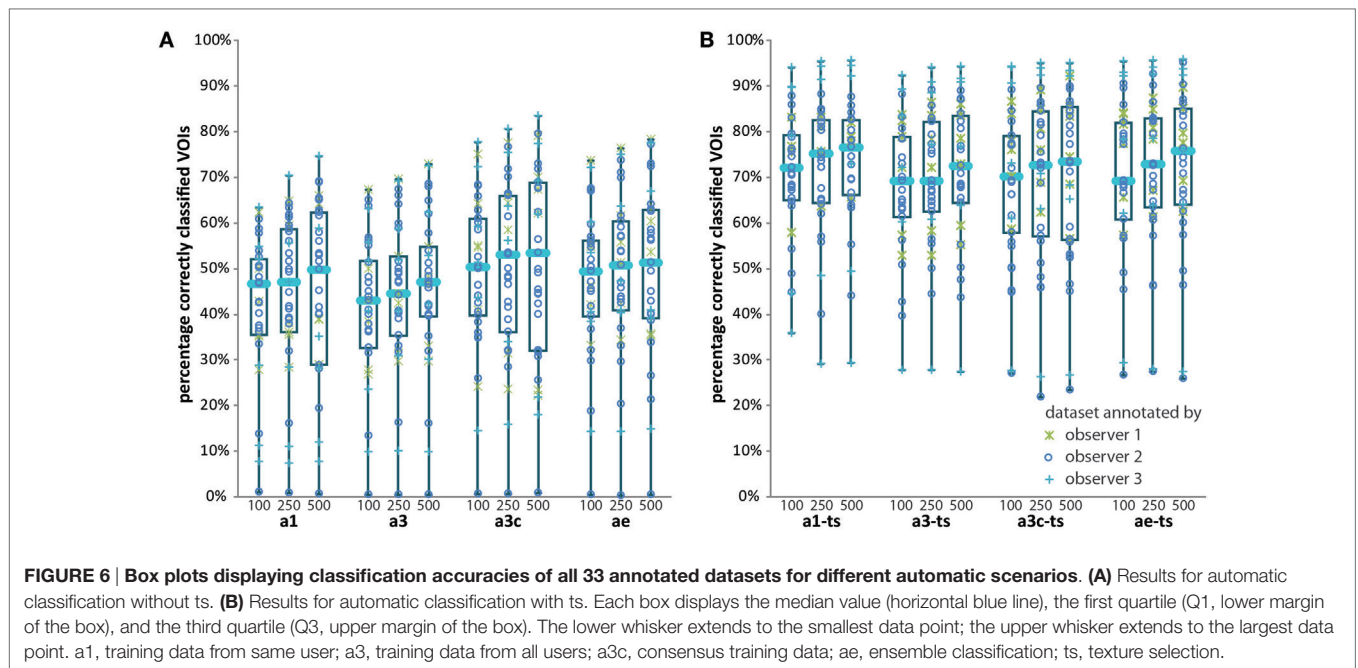


FIGURE 6 | Box plots displaying classification accuracies of all 33 annotated datasets for different automatic scenarios. **(A)** Results for automatic classification without ts. **(B)** Results for automatic classification with ts. Each box displays the median value (horizontal blue line), the first quartile (Q1, lower margin of the box), and the third quartile (Q3, upper margin of the box). The lower whisker extends to the smallest data point; the upper whisker extends to the largest data point. a1, training data from same user; a3, training data from all users; a3c, consensus training data; ae, ensemble classification; ts, texture selection.

TABLE 4 | Median (and range) of percentages of correctly classified VOIs for observer 1, observer 2, observer3, and for all 33 annotated datasets in the four automatic annotation protocols without texture selection.

Protocol	Median (min–max)% of correctly classified VOIs per protocol			
	Obs 1	Obs 2	Obs 3	All
One user (a1)	39 (29–66)	53 (1–69)	32 (8–75)	50 (1–75)
Three users (a3)	47 (30–73)	47 (0–68)	48 (10–72)	47 (0–73)
Consensus (a3c)	65 (22–79)	49 (1–79)	66 (18–83)	53 (1–83)
Ensemble (ae)	57 (35–78)	51 (0–77)	41 (15–77)	51 (0–78)

For each texture class, the intended number of training samples was 500.

TABLE 5 | Median (and range) of percentages of correctly classified VOIs for observer 1, observer 2, observer3, and all 33 annotated datasets in the four automatic annotation protocols with texture selection.

Protocol	Median (min–max)% of correctly classified VOIs			
	Obs 1	Obs 2	Obs 3	All
One user (a1)	78 (66–82)	74 (44–88)	83 (29–96)	77 (29–96)
Three users (a3)	76 (55–86)	69 (44–89)	84 (27–94)	73 (27–94)
Consensus (a3c)	79 (56–92)	73 (23–90)	80 (27–95)	73 (23–95)
Ensemble (ae)	78 (62–90)	73 (26–95)	84 (27–96)	76 (26–96)

For each texture class, the intended number of training samples was 500. Results are taken over all annotated datasets.

Interactive Classification Results

Texture Selection

The following paragraphs contain the results of interactive annotation experiments. First, we studied the effect of ts in interactive classification. In 20 of the 33 annotated datasets, 2 types of textures were found. In nine scans, three textures were selected, and in the remaining four scans, four textures were selected. In **Figure 7**, classification accuracies per slice and for

the remainder of the scan after training on at least 50% of the VOIs in the scan are shown for interactive protocols *i* (without ts) and *i*-ts (with ts). In both protocols, no previously annotated training data were used. In the first axial slice, median classification accuracy was the same whether VOIs were all labeled as normal tissue (in *i*) or VOIs were classified using a training dataset of five samples per texture class present in the scan. However, the IQR and the total accuracy range were larger if no ts was performed. The main advantage of ts in the first slice is that the minimum percentage of VOIs that are correctly classified was 38% instead of 0%. A similar effect could be seen in the second, third, and fourth slice. For slice 5 and classification of the remainder of the VOIs in the scan, results were similar for both approaches. This plot indicates that using training data from previously annotated scan may only be beneficial in classification of the first axial slice. Therefore, previous training data were only used for classification of the first axial slice in protocols *i*-a1(-ts), *i*-a3(-ts), *i*-a3c(-ts), and *i*-ae(-ts).

In the bottom row of **Figure 5**, interactive classification results are shown for one axial slice, for observer 1 (g), observer 2 (h), and observer 3 (i). These manual VOI annotations are shown directly above. Results are shown for classification of the remainder of the VOIs, after training on at least 50% of all VOI. For individual observers, the classifier was trained differently, which is reflected in the results in the bottom row. This indicated that the interactive classifier was able to adapt to the annotation preferences of the observers.

Use of Previous Training Data and Classifier Selection

In **Figure 8**, classification accuracy for the first slice, classification accuracy for the remainder of the scan, and overall classification accuracy are displayed, without (a) and with (b) ts. Median, minimum, and maximum values are given in **Table 6**.

The red boxes in **Figure 8** represent results for classification of the remainder of the scan. Since this was done in the same way in each interactive protocol, the results do not differ between the different approaches. The results for classification of the first axial slice, displayed by the blue boxes, displayed a large spread of median values: from 47% for protocol i-a3, in which previous

training data were used for classification of the first axial slice, to 82% for protocol i-cc-ts, where observers initiated annotation by selecting five example VOIs for each texture present in the scan and where they could choose which classification results to start from when correcting VOI labels. Protocols i-ts and i-cc-ts were the only two protocols for which the minimum number of VOIs correctly classified is not close to 0, with 38 and 51% respectively. In the overall results, median accuracies ranged from 82% for protocol i-a3 to 88% for protocol i-cc-ts. The minimum overall percentage of correctly classified VOIs was also largest for protocol i-cc-ts: 72%. For all interactive protocols, ts increased overall classification accuracy by 1–5% points.

Repeated measures ANOVA was performed to test for the difference in accuracy between the six interactive protocols without ts (i, i-a1, i-a3, i-a3c, i-ae, and i-cc). Protocol i-cc performed significantly better than the other ones [$F(5,28) = 29.1$; $p < 0.01$]. In addition, repeated measures ANOVA was used to test the significance of texture selection on classification accuracy by comparing the accuracies of protocols i-cc and i-cc-ts. Texture selection had a significant effect on accuracy [$F(1,32) = 11.6$; $p = 0.002$]. Finally, we tested whether interactive classification performed better than automatic classification by performing repeated measures ANOVA on the four automatic protocols (a1, a3, a3c, and ae) and the best-performing interactive protocol. Interactive classification performed significantly better than the four automatic protocols [$F(4,27) = 21.6$; $p < 0.001$]. *Post hoc* analysis indicated no significant differences between the individual automatic protocols.

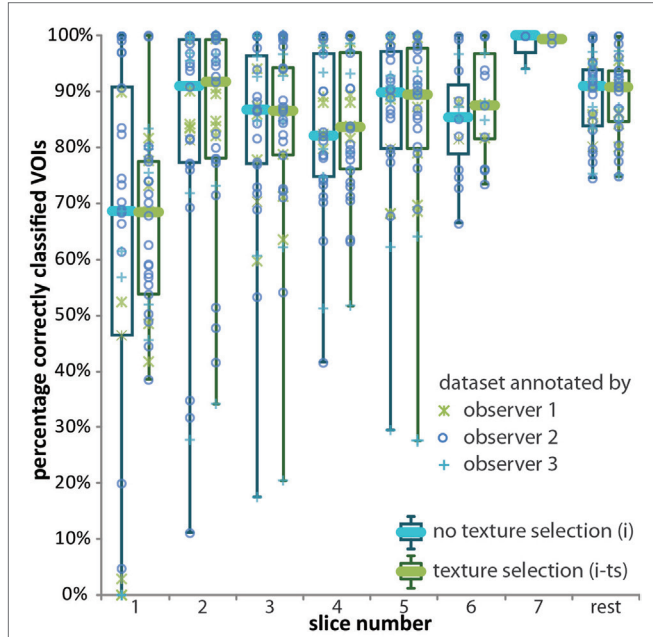


FIGURE 7 | Box plot displaying the classification accuracies of all 33 annotated datasets for interactive protocols i (blue boxes) and i-ts (green boxes). See the caption of **Figure 6** for an explanation of the boxes and whiskers.

Slice Selection Methods

Finally, we compared four different methods for slice selection in interactive classification. **Figure 9A** displays the distribution of the number of slices that had to be reviewed by an observer before

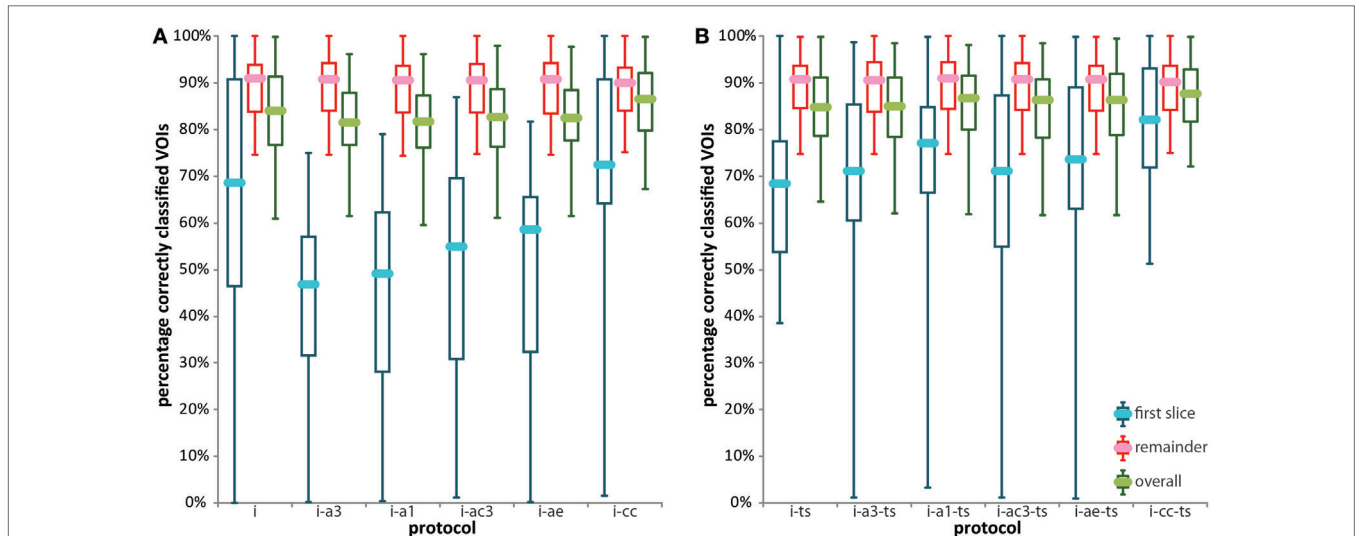


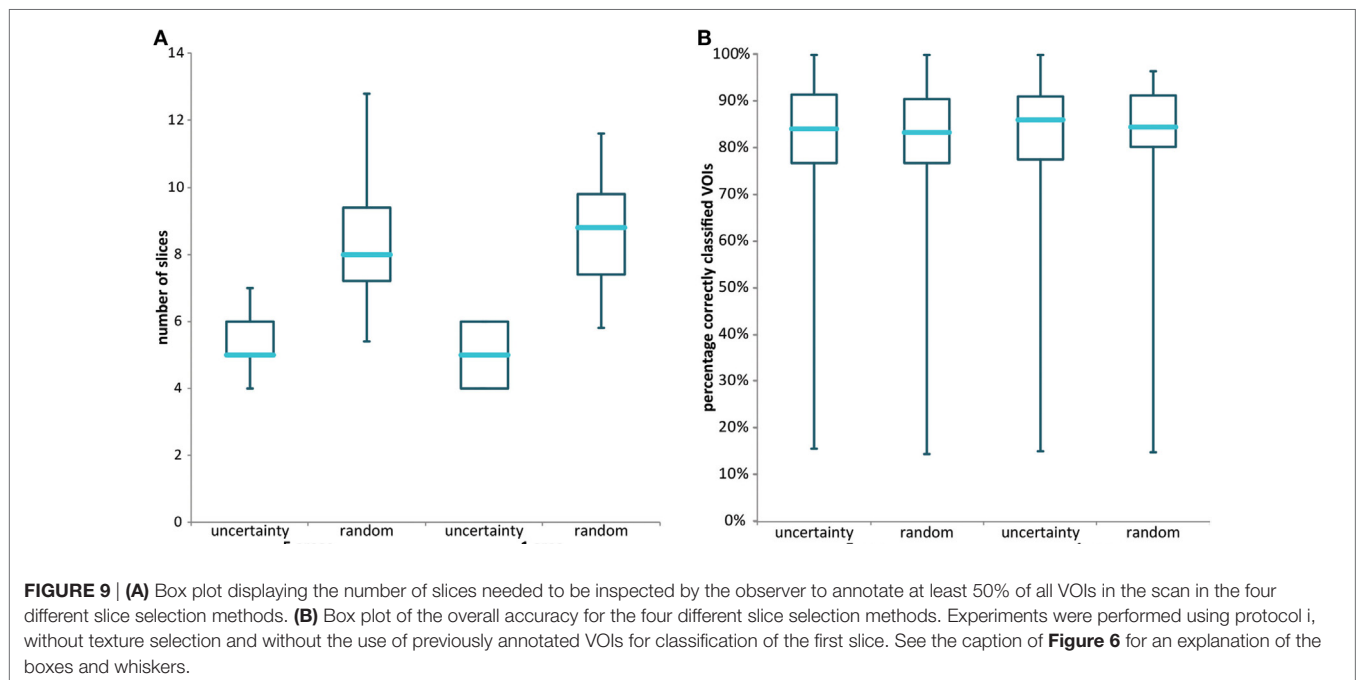
FIGURE 8 | Box plots displaying classification accuracies of all annotated datasets for the six interactive protocols without texture selection (ts) (A) and for the six protocols with ts (B). For each protocol, distribution of classification accuracy of the first slice (blue boxes), distribution of classification accuracy of the remainder of the scan after training on 50% of all VOIs (red boxes), and distribution of overall classification accuracy (green boxes) are shown. See the caption of **Figure 6** for an explanation of the boxes and whiskers.

TABLE 6 | Median (min–max)% of correctly classified VOIs for all annotated datasets in the 12 interactive annotation protocols.

		Median (min–max) accuracy for the first slice					
		i	i-a1	i-a3	i-a3c	i-ae	i-cc
No ts		69 (0–100)	49 (0–79)	47 (0–75)	55 (1–87)	59 (0–82)	72 (2–100)
ts		68 (38–100)	77 (3–100)	71 (1–99)	71 (1–100)	74 (1–100)	82 (51–100)
		Median (min–max) accuracy for the remainder of the scan					
		i	i-a1	i-a3	i-a3c	i-ae	i-cc
No ts		91 (75–100)	91 (74–100)	91 (75–100)	91 (75–100)	91 (75–100)	90 (75–100)
ts		91 (75–100)	91 (75–100)	91 (75–100)	91 (75–100)	91 (75–100)	90 (75–100)
		Median (min–max) overall accuracy					
		i	i-a1	i-a3	i-a3c	i-ae	i-cc
No ts		84 (61–100)	82 (59–96)	82 (61–96)	83 (61–98)	83 (62–98)	87 (67–100)
ts		85 (65–100)	87 (62–98)	85 (62–98)	86 (62–98)	86 (62–99)	88 (72–100)

Results are shown for classification of the first slice (top), for classification of the remainder of the scan after training on at least 50% of the VOIs in the scan (middle), and for all VOIs in the scan (bottom).

a1, training data from same user; a3, training data from all users; a3c, consensus training data; ae, ensemble classification; ts, texture selection.



the threshold of 50% of the VOIs in the scan was reached. In these experiments, classification protocol i was used, in which no ts was performed. Both random slice selection methods required more slices to be checked; median values were 8 when the lungs were divided into five areas and 9 when the lungs were not subdivided. This also means that more rounds of training the classifier and reclassification of the remaining VOIs had to be done. Dividing the scan into five areas, from which slices were selected alternately, led to a larger spread in the number of slices that had to be reviewed. The median number of slices, 5, was the same, whether the lungs are divided into five areas or not. **Figure 9B** shows that the distribution of interactive classification accuracy per

annotated dataset was comparable for all slice selection methods. Therefore, use of uncertainty-based slice selection is preferred over random slice selection. The number of areas in which the lungs were subdivided did not matter in these experiments.

DISCUSSION

Many studies have focused on automatic texture analysis in ILD. In general, these studies use user-defined ROIs/VOIs for which one set of annotations is used as the ground truth. In this work, we built on the interactive annotation approach that we have described earlier (Kockelkorn et al., 2016), in which all lung tissue

is annotated and ground truth is defined by the observer using the software. One of our aims was to optimize the interactive annotation process.

High interobserver variability is a known issue in ILD texture annotation. In two of our previous studies, both of which were performed on smaller datasets, we found interobserver agreements for two observers of 51% (Kockelkorn et al., 2010) and 63% (Kockelkorn et al., 2016). In this work, roughly 70% of all VOIs that were labeled at least twice received the same label from at least two observers. This also means for roughly one-third of the VOIs, all two or three observers assigned a different label to the same VOI. With observers having three different opinions on the texture label of a VOI, the problem of making a universally accepted annotated dataset becomes even more complicated. One of the issues we wanted to study was how we could use training data from previously annotated scans for classification at the start of interactive annotation, when no or little training data from the scan under consideration are available. Given the high interobserver variability, we hypothesized that using training data from the observer who is annotating the scan might have advantages over using training data obtained from all observers. Our results indicate that using training data from the same observer results in slightly higher classification accuracy, but only when combined with *ts*. Without *ts*, using consensus training data leads to the highest median accuracy in automatic classification. We also investigated the use of having a panel of classifiers deciding together on the label of VOIs, analogous to physicians making a diagnosis together. Using this ensemble approach did not lead to higher median classification accuracies when compared with the other classification protocols. In general, the differences between the median results of the various annotation methods are small and which methods yields the highest median accuracy varies for the individual observers.

Another aim of our study was to investigate how a classifier, trained on previously obtained training data, can decrease user effort necessary for interactive annotation of all lung VOIs. Since the automatic classification experiments did not indicate the superiority of one single automatic classification protocol, we tested the use of all automatic classification methods in our interactive classification framework. In the most basic protocol (i), no *ts* or automatic classification results were used. VOIs in the first axial slice were labeled as normal tissue, which is in this dataset with 55% the most common texture. The median labeling accuracy over all annotated datasets resulting from this approach was 69%, which is below median classification accuracy for all automatic protocols with class selection. Median interactive classification accuracy for all following slices is above median accuracy for all automatic protocols. Therefore, we conclude that automatic classification is only beneficial for classification of the first axial slice.

Other studies have reported higher automatic classification accuracies (Depeursinge et al., 2012; Song et al., 2013; Anthimopoulos et al., 2016); however, these were obtained on hand-drawn ROIs at specific user-selected locations instead of automatically generated VOIs spanning the entire lungs. In addition, not all classification approaches are suitable for the interactive annotation approach we propose: we opted for pre-calculated

features and a classifier who is trained relatively fast to reduce the time that the observer has to wait for classification results. Nevertheless, it would be possible to initiate interactive annotation by classification of the first slice using, for example, a deep-learning approach.

Besides correcting automatic classification results, observers can transfer knowledge of the annotation task in other ways. The first is by selecting examples of all tissue types present in the scan before the start of interactive classification. In automatic classification, *ts* resulted in a substantial increase of median classification accuracy, ranging from 20% points for the protocol using consensus training data (a3c) to 27% points for the protocol that used training data from the observer who provided the ground truth annotations for the scan under consideration (a1).

In interactive classification, using automatic classification results without *ts* leads to a decrease in classification accuracy, as compared with the protocols in which *ts* is applied. This decrease is not only seen in the first slice but also in the overall classification results: protocols i-a1, 1-a3, i-a3c, and i-ae display a decrease in median classification accuracy of 1–2% points as compared with protocol i. Therefore, automatic classification results should only be used for classification of the first axial slice if the observer performs *ts* before interactive annotation.

Finally, we noted that selecting slices to be corrected by the observer based on the cumulative uncertainty in the slice results in a smaller number of slices that should be reviewed before reaching at least 50% of the VOIs, as compared with random slice selection. A smaller number of slices to review means that the classifier has to be trained less often, which in turn decreases the time observers have to wait for new classification results to be generated. Dividing the lungs into five areas, from which slices are chosen in an alternate fashion, did not have advantages over considering slices from the entire lungs in each classification, retraining, and correction cycle. This was contrary to what we expected, since ILD textures may be localized.

This work presents several open issues. First, while the 23 thoracic CT scans included are an increase as compared with our previous studies, experiments should ideally be executed on a larger dataset, containing various ILD subtypes. Second, only part of the dataset was annotated by more than one observer. This did enable us to assess interobserver variability and to evaluate how the interactive annotation framework adapts to individual observers' annotation preferences. However, in order to get to a consensus dataset, three (or more) observers should individually annotate all scans and then discuss their results. Interactive annotation can facilitate this approach in future work. Third, this work does not compare the effects of using a different classifier or different features. In principle, any texture features could be inserted into the interactive framework. The same holds for the classifier, with the limitation that the chosen classifier should allow training while the observer awaits the results.

In future work, it would be interesting to compare the simulation results from this work to results obtained by human observers. Given the substantial interobserver variability, it is conceivable that observers are influenced by the classification results that are suggested by the algorithm. To investigate the

influence of suggested annotations, observers could be asked to annotate the same scan twice at different time points: once by completely manual labeling of all VOIs and once by interactive labeling. By comparing the interactive annotation results to the manual labels on the one hand and the automatically generated labels on the other hand, an estimate of the degree by which computer-generated labels influence annotation behavior of the observer could be made.

We have shown that automatic classification results can be beneficial in interactive annotation, but only when used in combination with ts. In addition, giving observers several different automatic classification results to choose from when correcting VOI labels decreased the median the number of corrections. Using the best-performing protocol, in which observers select the textures that should be distinguished in the scan and they are provided with alternative classification results in case interactive

classification accuracy is low, a median accuracy of 88% was reached. We therefore conclude that interactive annotation with ts and classifier choice could be a useful tool for annotating lung tissue in CT scans of ILD patients.

AUTHOR CONTRIBUTIONS

TK, JR, PJ, CS-P, JG, MV, and BG designed the study. TK, PJ, CS-P, and BG developed the methodology. JG collected the data. RR, AT, and RW analyzed the CT scans. TK analyzed the results. All the authors reviewed and approved the manuscript.

ACKNOWLEDGMENTS

The authors would like to thank Floor van Meer for performing the repeated measures ANOVA.

REFERENCES

- Anthimopoulos, M., Christodoulidis, S., Ebner, L., Christe, A., and Mougiakakou, S. (2016). Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. *IEEE Trans. Med. Imaging*, 35, 1207–1216. doi:10.1109/TMI.2016.2535865
- Aziz, Z. A., Wells, A. U., Bateman, E. D., Copley, S. J., Desai, S. R., Grutters, J. C., et al. (2006). Interstitial lung disease: effects of thin-section CT on clinical decision making. *Radiology* 238, 725–733. doi:10.1148/radiol.2381041823
- Depeursinge, A., Iavindrasana, J., Hidki, A., Cohen, G., Geissbuhler, A., Platon, A., et al. (2010). Comparative performance analysis of state-of-the-art classification algorithms applied to lung tissue categorization. *J. Digit. Imaging* 23, 18–30. doi:10.1007/s10278-008-9158-4
- Depeursinge, A., Van de Ville, D., Platon, A., Geissbuhler, A., Poletti, P. A., and Müller, H. (2012). Near-affine-invariant texture learning for lung tissue analysis using isotropic wavelet frames. *IEEE Trans. Inf. Technol. Biomed.* 16, 665–675. doi:10.1109/TITB.2012.2198829
- Gao, M., Bagci, U., Lu, L., Wu, A., Buty, M., Shin, H., et al. (2016). Holistic classification of CT attenuation patterns for interstitial lung diseases via deep convolutional neural networks. *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.* doi:10.1080/21681163.2015.1124249
- Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Mach. Learn.* 63, 3–42. doi:10.1007/s10994-006-6226-1
- Huber, M. B., Bunte, K., Nagarajan, M. B., Biehl, M., Ray, L. A., and Wismüller, A. (2012). Texture feature ranking with relevance learning to classify interstitial lung disease patterns. *Artif. Intell. Med.* 56, 91–97. doi:10.1016/j.artmed.2012.07.001
- Huber, M. B., Nagarajan, M. B., Leinsinger, G., Eibel, R., Ray, L. A., and Wismüller, A. (2011). Performance of topological texture features to classify fibrotic interstitial lung disease patterns. *Med. Phys.* 38, 2035–2044. doi:10.1118/1.3566070
- Kockelkorn, T. T. J. P., de Jong, P. A., Gietema, H. A., Grutters, J. C., Prokop, M., and van Ginneken, B. (2010). Interactive annotation of textures in thoracic CT scans. *Proc. SPIE* 7624, 76240X. doi:10.1117/12.844932
- Kockelkorn, T. T. J. P., de Jong, P. A., Schaefer-Prokop, C. M., Wittenberg, R., Tiehuis, A. M., Gietema, H. A., et al. (2016). Semi-automatic classification of textures in thoracic CT scans. *Phys. Med. Biol.* 61, 5906–5924. doi:10.1088/0031-9155/61/16/5906
- Park, S. O., Seo, J. B., Kim, N., Park, S. H., Lee, Y. K., Park, B. W., et al. (2009). Feasibility of automated quantification of regional disease patterns depicted on high-resolution computed tomography in patients with various diffuse lung diseases. *Korean J. Radiol.* 10, 455–463. doi:10.3348/kjr.2009.10.5.455
- Prosch, H., Schaefer-Prokop, C. M., Eisenhuber, E., Kienzl, D., and Herold, C. J. (2013). CT protocols in interstitial lung diseases – a survey among members of the European Society of Thoracic Imaging and a review of the literature. *Eur. Radiol.* 23, 1553–1563. doi:10.1007/s00330-012-2733-6
- Song, Y., Cai, W., Zhou, Y., and Feng, D. D. (2013). Feature-based image patch approximation for lung tissue classification. *IEEE Trans. Med. Imaging* 32, 797–808. doi:10.1109/TMI.2013.2241448
- Uppaluri, R., Hoffman, E. A., Sonka, M., Hunninghake, G. W., and McLennan, G. (1999). Interstitial lung disease: a quantitative study using the adaptive multiple feature method. *Am. J. Respir. Crit. Care Med.* 159, 519–525. doi:10.1164/ajrccm.159.2.9707145
- van Opbroek, A., Ikram, M. A., Vernooij, M. W., and de Bruijne, M. (2015). Transfer learning improves supervised image segmentation across imaging protocols. *IEEE Trans. Med. Imaging* 34, 1018–1030. doi:10.1109/TMI.2014.2366792
- van Riel, S. J., Sánchez, C. I., Bankier, A. A., Naidich, D. P., Verschakelen, J., Scholten, E. T., et al. (2015). Observer variability for classification of pulmonary nodules on low-dose CT images and its effect on nodule management. *Radiology* 277, 863–871. doi:10.1148/radiol.2015142700
- van Rikxoort, E. M., de Hoop, B., Viergever, M. A., Prokop, M., and van Ginneken, B. (2009). Automatic lung segmentation from thoracic computed tomography scans using a hybrid approach with error detection. *Med. Phys.* 36, 2934–2947. doi:10.1118/1.3147146
- Vasconcelos, V., Barroso, J., Marques, L., and Silva, J. S. (2015). Enhanced classification of interstitial lung disease patterns in HRCT images using differential lacunarity. *Biomed Res. Int.* 2015, 672520. doi:10.1155/2015/672520
- Watanabe, T., Sakai, E., 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. doi:10.1148/radiol.12112516
- Xu, Y., van Beek, E. J., Hwanjo, Y., Guo, J., McLennan, G., and Hoffman, E. A. (2006). Computer-aided classification of interstitial lung diseases via MDCT: 3D adaptive multiple feature method (3D AMFM). *Acad. Radiol.* 13, 969–978. doi:10.1016/j.acra.2006.04.017
- Zavaletta, V. A., Bartholmai, B. J., and Robb, R. A. (2007). High resolution multidetector CT-aided tissue analysis and quantification of lung fibrosis. *Acad. Radiol.* 14, 772–787. doi:10.1016/j.acra.2007.03.009

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Kockelkorn, Ramos, Ramos, de Jong, Schaefer-Prokop, Wittenberg, Tiehuis, Grutters, Viergever and van Ginneken. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.