Diagnostic and Interventional Imaging 102 (2021) 691–695

Original article

Deep learning for lung disease segmentation on CT: Which reconstruction kernel should be used?

Trieu-Nghi Hoang-Thi[a,b], Maria Vakalopoulou[c], Stergios Christodoulidis[c], Nikos Paragios[c,d], Marie-Pierre Revel[a,b], Guillaume Chassagnon[a,b,*]

[a] *Université de Paris, Faculté de Médecine, 75006 Paris, France*
[b] *Department of Radiology, Hôpital Cochin, AP-HP.centre, 75014 Paris, France*
[c] *Université Paris-Saclay, CentraleSupélec, Mathématiques et Informatique pour la Complexité et les Systèmes, 3 91190 Gif-sur-Yvette, France*
[d] *TheraPanacea, 75014 Paris, France*

ARTICLE INFO

Keywords:
Deep learning
Multidector computed tomography
Lung

ABSTRACT

Purpose: The purpose of this study was to determine whether a single reconstruction kernel or both high and low frequency kernels should be used for training deep learning models for the segmentation of diffuse lung disease on chest computed tomography (CT).

Materials and methods: Two annotated datasets of COVID-19 pneumonia (323,960 slices) and interstitial lung disease (ILD) (4,284 slices) were used. Annotated CT images were used to train a U-Net architecture to segment disease. All CT slices were reconstructed using both a lung kernel (LK) and a mediastinal kernel (MK). Three different trainings, resulting in three different models were compared for each disease: training on LK only, MK only or LK+MK images. Dice similarity scores (DSC) were compared using the Wilcoxon signed-rank test.

Results: Models only trained on LK images performed better on LK images than on MK images (median DSC = 0.62 [interquartile range (IQR): 0.54, 0.69] *vs.* 0.60 [IQR: 0.50, 0.70], $P < 0.001$ for COVID-19 and median DSC = 0.62 [IQR: 0.56, 0.69] *vs.* 0.50 [IQR 0.43, 0.57], $P < 0.001$ for ILD). Similarly, models only trained on MK images performed better on MK images (median DSC = 0.62 [IQR: 0.53, 0.68] *vs.* 0.54 [IQR: 0.47, 0.63], $P < 0.001$ for COVID-19 and 0.69 [IQR: 0.61, 0.73] *vs.* 0.63 [IQR: 0.53, 0.70], $P < 0.001$ for ILD). Models trained on both kernels performed better or similarly than those trained on only one kernel. For COVID-19, median DSC was 0.67 (IQR: =0.59, 0.73) when applied on LK images and 0.67 (IQR: 0.60, 0.74) when applied on MK images ($P < 0.001$ for both). For ILD, median DSC was 0.69 (IQR: 0.63, 0.73) when applied on LK images ($P = 0.006$) and 0.68 (IQR: 0.62, 0.72) when applied on MK images ($P > 0.99$).

Conclusion: Reconstruction kernels impact the performance of deep learning-based models for lung disease segmentation. Training on both LK and MK images improves the performance.

1. Introduction

Over the last decade, deep learning (DL) has become the method of choice for many applications in medical image analysis, especially for segmentation tasks [1–3]. The performance gain in image segmentation achieved with DL methods over traditional machine learning methods makes it now possible to accurately segment complex entities within the lung, such as interstitial lung disease (ILD) or coronavirus disease 2019 (COVID-19) pneumonia [4–7].

A particularity of chest computed tomography (CT) is that images are usually reconstructed using two different reconstruction kernels, a high frequency kernel, lung kernel (LK) for the evaluation of the lung parenchyma and a standard reconstruction kernel, mediastinal kernel (MK) for the evaluation of the mediastinum. Recent studies have shown that the performance of DL-based tools for lung nodule detection and characterization is influenced by the acquisition and reconstruction settings, such as the choice of the reconstruction kernel [8,9]. However, there are currently no data regarding the influence of the reconstruction kernels on lung disease segmentation using data-driven methods. We hypothesized that combining the reconstruction kernels could improve model performances by allowing the model to get rid of the variability related to the reconstruction parameters.

The purpose of this study was to determine whether a single reconstruction kernel or two reconstruction kernels should be used

Abbreviations: COVID-19, Coronavirus disease 2019; CT, Computed tomography; DL, Deep learning; DSC, Dice similarity coefficient; ILD, Interstitial lung disease; IQR, interquartile range; LK, Lung kernel; MK, Mediastinal kernel

[*] Corresponding author.
E-mail address: guillaume.chassagnon@aphp.fr (G. Chassagnon).

https://doi.org/10.1016/j.diii.2021.10.001

when training DL for segmentation of diffuse lung disease on chest CT. For this, a DL architecture [10] was evaluated on two different CT datasets of diffuse lung disease

2. Material and methods

2.1. Image datasets and annotation transfer

This retrospective multi-center study was performed in accordance with The Code of Ethics of the World Medical Association (Declaration of Helsinki).

Data from two independent and previously published datasets were used to assess the impact of reconstruction kernels on DL-based segmentation. One dataset has been created to develop segmentation models for COVID-19 pneumonia (COVID dataset) [7] while the other was created to develop segmentation models ILD in systemic sclerosis (ILD dataset) [5]. Both datasets originally only included images reconstructed with a LK (LK images). For the purpose of this study, the corresponding images reconstructed with a MK (MK images) were retrieved. Annotations produced by the annotation of original LK images were projected on the MK images.

When the LK and MK images had the same slice thickness and reconstruction field of view, the mask of annotation was directly transferred to MK images. When the slice thickness and/or reconstruction field of view were different, the metadata were used to perform the required rescaling and cropping, so that the two acquired lung volumes could correspond.

2.2. COVID-19 dataset

This dataset comprised 180 unenhanced chest CT examinations from COVID-19 patients acquired at six university hospitals using four different CT models from three manufacturers. All images had been reconstructed using iterative reconstruction and a slice thickness ranging from 0.625- to 1 mm. Images had been annotated by 15 radiologists with one to seven years of experience in chest imaging. On each CT slice, all the COVID-19 related CT abnormalities (ground glass opacities, band consolidations, and reticulations) [11–13] had been segmented as a single

Fifty patients from three centers composed the training and validation dataset. All images of these 50 CT examinations (321,360 slices) had been annotated slice-by-slice. The 130 patients from the remaining three centers composed the test dataset. In this test dataset, only 20 slices per CT examination (2600 slices), equally spaced from the superior border of aortic arch to the lowest diaphragmatic dome had been annotated and each slide had been annotated by two different radiologists.

2.3. ILD dataset

This dataset comprised 31 unenhanced chest CT examinations from systemic sclerosis patients acquired at one university hospital using four different CT scanners from two manufacturers. Six patients out of the 37 patients from the original cohort were excluded because of the unavailability of MK images. Lung CT images were reconstructed with a slice thickness of 0.625- to 1.5 mm, using filter back projection or iterative reconstruction algorithms. ILD segmentation had been performed by four radiologists with one to four years of experience in chest imaging. All the ILD-related anomalies (ground glass opacities, reticulations, traction bronchiectasis and honeycombing) were segmented as a single class (ILD).

Chest CT examinations from 11 patients composed the training and validation dataset. All images of this dataset had been annotated by one radiologist (3884 slices). Chest CT examinations from the remaining 20 patients composed the test dataset. For each CT examination, 20 CT slices equally spaced from the lung apices to the right diaphragmatic dome (20 slices) had been annotated by three different radiologists (400 images in total).

2.5. DL-based segmentation methods

For benchmarking the performance of the DL architecture, a standard U-Net architecture [10] consisting of five blocks with a downsampling operation applied every two consequent Conv2D-BN-ReLU layers was used. Additionally, five decoding blocks were used for the decoding path, were at each block a transpose convolution was performed to up-sample the input. Skip connections were also employed between the encoding and decoding paths. Each two-dimensional slice extracted from the CT examination was given as an input to the network together with its annotation of the abnormal regions. Weighted cross entropy has been used as loss function for the training of the model with the weight of the pathological regions being inverse proportional to the appearance of each class, corresponding to 60 for the COVID dataset and to 65 for the ILD dataset. Adam optimizer was used with a learning rate of 0.0001 for the optimization. For fair comparison, the same number of iterations was kept for the training of all models.

Three different trainings, resulting in three different models were compared for each disease: training on LK images only, MK images only or LK+MK images. Thus, the same CT slices were used for each of the trainings but these CT slices had been reconstructed with only one kernel (LK or MK) or the two kernels (LK and MK).

2.6. Statistical analysis

Statistical analysis was performed using ‘R’ software (version 3.6.3, R Foundation, Vienna, Austria). The quality of the segmentations was assessed using the Dice similarity coefficient (DSC) [14]. As manual segmentations from several radiologists were available for the two test datasets, for each CT examination the average DSC between the model and the observers was calculated. Comparison between models was performed using the Wilcoxon signed-rank test. To compensate for multiple comparisons, a Bonferroni correction was applied. A P value < 0.05 was considered to indicate significant difference.

3. Results

3.1. Impact of the reconstruction kernel to build models

To evaluate the impact of the reconstruction kernel on models' generalizability, we first compared the performance of the segmentation models in test datasets combining LK and MK images from the same CT slices (Table 1, Fig. 1). For COVID-19 pneumonia segmentation, the model trained on LK images performed better than the one trained on MK images (median DSC = 0.61 [interquartile range (IQR): 0.52, 0.70] *vs.* 0.57 [IQR: 0.50, 0.67], respectively; $P < 0.001$). Conversely, for ILD segmentation, the model trained on MK images performed better than the one trained on LK images (median DSC = 0.65 [IQR: 0.58, 0.70] *vs.* 0.53 [IQR: 0.49, 0.63], respectively; $P = 0.007$). For both COVID-19 and ILD segmentation, training on a dataset combining the two reconstruction kernels (median DSC = 0.67 [IQR: 0.59, 0.72] for COVID-19 and 0.69 [IQR: 0.63, 0.73] for ILD) significantly improved the performance compared to models trained on LK images alone ($P < 0.001$ for COVID-19 and ILD segmentations) or MK images alone ($P < 0.001$ for COVID-19 and ILD segmentations).

When applying the model to a test dataset consisting solely of LK or MK images, models trained on both kernels performed as well as or better than models trained only on one of the two kernels. For COVID-19 pneumonia segmentation, the model trained on both kernels performed better than the one trained on LK images when applied on LK images only (median DSC = 0.67 [IQR: 0.59, 0.73] *vs.*

Table 1
Dice similarity scores for automated and manual segmentations depending on the reconstruction kernel used in the training and test datasets. The presented Dice similarity scores are the ones measured in the test cohorts.

	Applied on lung kernel		Applied on mediastinal kernel		Applied on both kernels	
	COVID-19	ILD	COVID-19	ILD	COVID-19	ILD
Trained on lung kernel	0.62 (0.54, 0.69) [0.27–0.90]	0.62 (0.56, 0.69) [0.38–0.87]	0.60 (0.50, 0.70) [0.24–0.90]	0.50 (0.43, 0.57) [0.31–0.83]	0.61 (0.52, 0.70) [0.26–0.90]	0.53 (0.49, 0.63) [0.37–0.85]
Trained on mediastinal kernel	0.54 (0.47, 0.63) [0.24–0.89]	0.63 (0.53, 0.70) [0.27–0.80]	0.62 (0.53, 0.68) [0.24–0.89]	0.69 (0.61, 0.73) [0.34–0.88]	0.57 (0.50, 0.67) [0.24–0.89]	0.65 (0.58, 0.70) [0.30–0.84]
Trained on both kernels	0.67 (0.59, 0.73) [0.30–0.92]	0.69 (0.63, 0.73) [0.32–0.88]	0.67 (0.60, 0.74) [0.29–0.92]	0.68 (0.62, 0.72) [0.29–0.88]	0.67 (0.59, 0.72) [0.30–0.92]	0.69 (0.63, 0.73) [0.30–0.88]

Note: Dice similarity scores are presented as median with interquartile ranges in parentheses and ranges in brackets; ILD = Interstitial lung disease.

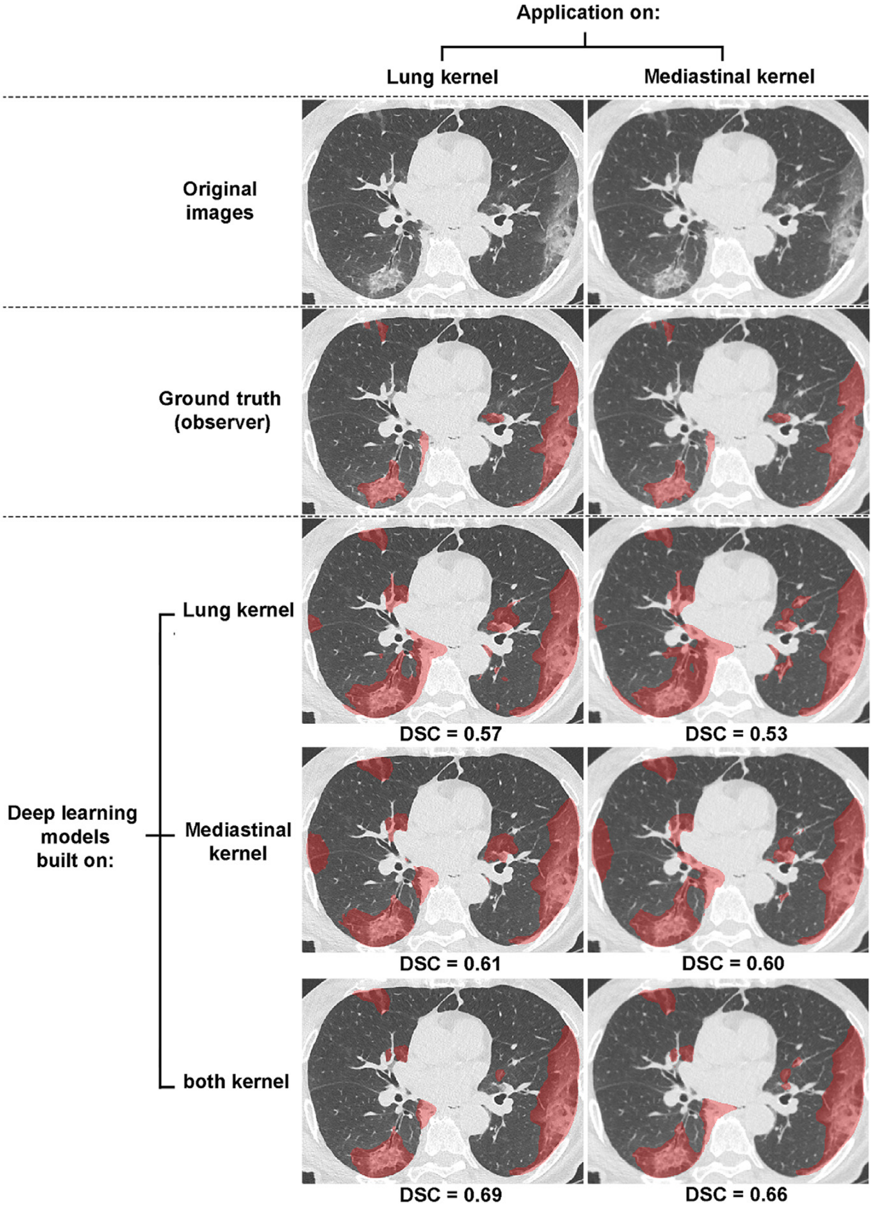

Fig. 1. Results of the COVID-19 segmentation models in a 42-year-old man with COVID-19 pneumonia. The deep learning model trained on the images reconstructed with the two kernels performs better than the models trained on the images reconstructed with only one of the kernels, whatever the kernel on which it is applied. The better performances are illustrated by the greater dice similarity score (DSC). Models trained on only one kernel are responsible for greater oversegmentation.

0.62 [IQR: 0.54, 0.69]; $P < 0.001$) and better than the model trained on MK images when applied on MK images only (median DSC = 0.67 [IQR: 0.60, 0.74] *vs.* 0.62 [IQR: 0.53, 0.68]; $P < 0.001$). For the ILD segmentation, the model trained on both kernels performed better than the one trained on LK images when applied on LK images (median DSC = 0.69 [IQR: 0.63, 0.73] *vs.* 0.62 [IQR: 0.56, 0.69]; $P = 0.006$) and as well as the model trained on MK images when applied on MK images (median DSC = 0.68 [IQR: 0.62, 0.72] *vs.* 0.69 [IQR: 0.61, 0.73]; $P > 0.99$).

3.2. Selection of the reconstruction kernel for application of trained models

To assess the impact of reconstruction kernel selection for model application, the performances of the models depending on the reconstruction kernel used in the test datasets were compared (Table 1, Fig. 1). Models trained only on LK images performed better on LK images than on MK images (median DSC = 0.62 [IQR: 0.54, 0.69] *vs.* 0.60 [IQR: 0.50, 0.70]; $P < 0.001$ for COVID-19 segmentation and 0.62 [IQR: 0.56, 0.69] *vs.* 0.50 [IQR: 0.43, 0.57]; $P < 0.001$ for ILD segmentation). Similarly, models trained only on MK images performed better on MK images (median DSC = 0.62 [IQR: 0.53, 0.68] *vs.* 0.54 [IQR: 0.47, 0.63]; $P < 0.001$ for COVID-19 segmentation and 0.69 [IQR: 0.61, 0.73] *vs.* 0.63 [IQR: 0.53, 0.70]; $P < 0.001$ for ILD segmentation).

When the segmentation models had been trained on a combination of the two kernels, the differences in performance according to the reconstruction kernel of the test dataset were smaller with a slight advantage to the LK. For COVID-19 segmentation, DSC was not significantly different whether applied to LK images (median DSC = 0.67 [IQR: 0.59, 0.73]) or MK images (median DSC = 0.67 [IQR: 0.60, 0.74]) ($P > 0.99$). On the opposite, the DSC was slightly greater when the ILD segmentation model was applied to LK images (median DSC =0.69 [IQR: 0.63, 0.73] *vs.* 0.68 [IQR: 0.62, 0.72]; $P < 0.001$).

4. Discussion

To our knowledge, no studies have assessed the influence of the reconstruction kernel on image segmentation using DL. Using two different datasets and a popular DL architecture, we demonstrated that the choice of the reconstruction kernel significantly impacts the performance of the models and that training algorithms on a combination of lung and mediastinal kernels significantly improves the performance.

Regardless of the lung disease, our results show that the models performed better when reconstruction kernels were the same in the training and the test datasets. This demonstrates that while being more efficient than traditional machine learning methods, DL is still sensitive to image characteristics such as the reconstruction kernel. Decreasing the sharpness of the reconstruction kernel reduces the amount of noise in the image, which is already known to have a significant impact on emphysema quantification [15,16]. For systemic sclerosis-related ILD, Kim et al. demonstrated that by using image denoising methods, they increased the performance of their ILD segmentation model based on texture analysis with classical machine learning methods [17]. Although images reconstructed with a MK are less noisy, our study does not demonstrate the superiority of this kernel over LK in the setting of lung disease segmentation using DL. LK is usually the preferred kernel for visual analysis of lung parenchyma.

For pulmonary nodules, a few studies have also reported an impact of the reconstruction kernel on the performance of DL-based computer-aided detection tools and showed conflicting results regarding the best kernel to be used [8,9]. The selection of the reconstruction kernel is also known to impact the results of tools based on radiomics and texture analysis approaches, such as for pulmonary nodule characterization [18–22]. Recently, Choe et al. have suggested that the use of chest CT image conversion using CNN can reduce the effect of the reconstruction kernels on radiomics parameters for lung nodule assessment [23].

An important result of our study is that we demonstrated that combining images reconstructed from the same CT slice using LK and MK improves learning performances. This combination is easy to perform as all chest CT acquisitions are usually reconstructed with both kernels and both sets of images are routinely stored. Moreover, the annotation time is not increased because manual segmentations performed on images reconstructed with one kernel can be easily applied on those reconstructed with the other kernel and has more influence when the dataset is smaller. Thus, the combination of the two kernels allows data augmentation without having to increase the number of patients or the annotation time, which are two factors limiting the size of medical imaging datasets. Combining the two kernels may allow the algorithm to free itself from the variability due to image noise. Interestingly, when the algorithm was trained on a dataset combining images reconstructed with both kernels, the choice of the kernel for test had almost no impact on the results and depended on the lung disease.

This study has several limitations. First, all the manual segmentations we used were performed on LK images. However, the fact that models trained on MK images performed better on the test dataset containing images also reconstructed with a MK suggests this had no significant impact. Second, only one DL architecture was used. Yet, the U-Net architecture is the most popular for segmentation tasks in medical imaging, while the small size of the ILD dataset did not allow the training of 3D deep learning architectures.

In conclusion, our work highlights the importance of the reconstruction kernel when working on a DL-based method to segment diffuse lung diseases. We have shown that learning on a combination of LK and MK images from the same CT slice should be done as it improves the performance and generalizability of the models. Importantly, this can be done without increasing the annotation time. The combination of two different kernels could also improve segmentation tasks for extra-thoracic applications, such as musculoskeletal imaging where images are commonly reconstructed with standard and bone reconstruction algorithms.

Credit author statement

T.-N. H.-T.: Study design; data collection; analysis and interpretation of the results; draft manuscript edition; manuscript final version approval

M.K.: Study design; data analysis; interpretation of the results manuscript final version approval

S.C.: Study design; data analysis; interpretation of the results; manuscript final version approval

N.P.: Study design; supervision; manuscript final version approval

M.-P. R.: Study design; supervision; manuscript final version approval

G.C.: Study design; data collection; analysis and interpretation of the results; supervision; manuscript final version approval

Declaration of Competing Interest

The authors have no conflict of interest in relation to this study and no relationships with any industry related to this work.

Human rights

This work has been performed in accordance with The Code of Ethics of the World Medical Association (Declaration of Helsinki).

Funding

This work did not receive any grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author contributions

All authors attest that they meet the current International Committee of Medical Journal Editors (ICMJE) criteria for Authorship.

References

[1] Chassagnon G, Vakalopoulou M, Paragios N, Revel M-P. Artificial intelligence applications for thoracic imaging. Eur J Radiol 2020;123:108774.
[2] Nakaura T, Higaki T, Awai K, Ikeda O, Yamashita Y. A primer for understanding radiology articles about machine learning and deep learning. Diagn Interv Imaging 2020;101:765–70.
[3] Blanc D, Racine V, Khalil A, Deloche M, Broyelle J-A, Hammouamri I, et al. Artificial intelligence solution to classify pulmonary nodules on CT. Diagn Interv Imaging 2020;101:803–10.
[4] Anthimopoulos M, Christodoulidis S, Ebner L, Christe A, Mougiakakou S. Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. IEEE Trans Med Imaging 2016;35:1207–16.
[5] Chassagnon G, Vakalopoulou M, Régent A, Zacharaki EI, Aviram G, Martin C, et al. Deep learning–based approach for automated assessment of interstitial lung disease in systemic sclerosis on CT images. Radiol Artif Intell 2020;2:e190006.
[6] Li K, Fang Y, Li W, Pan C, Qin P, Zhong Y, et al. CT image visual quantitative evaluation and clinical classification of coronavirus disease (COVID-19). Eur Radiol 2020;30:4407–16.
[7] Chassagnon G, Vakalopoulou M, Battistella E, Christodoulidis S, Hoang-Thi T-N, Dangeard S, et al. AI-driven quantification, staging and outcome prediction of COVID-19 pneumonia. Med Image Anal 2021;67:101860.
[8] Fu B, Wang G, Wu M, Li W, Zheng Y, Chu Z, et al. Influence of CT effective dose and convolution kernel on the detection of pulmonary nodules in different artificial intelligence software systems: a phantom study. Eur J Radiol 2020;126:108928.
[9] Blazis SP, Dickerscheid DBM, Linsen PVM, Martins Jarnalo CO. Effect of CT reconstruction settings on the performance of a deep learning based lung nodule CAD system. Eur J Radiol 2021;136:109526.
[10] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation editors. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Med image comput comput assist interv. MICCAI; 2015. p. 234–41.
[11] Chassagnon G, Regard L, Soyer P, Revel M-P. COVID-19 after 18 months: where do we stand? Diagn Interv Imaging 2021;102:491–2.
[12] Li J, Long X, Wang X, Fang F, Lv X, Zhang D, et al. Radiology indispensable for tracking COVID-19. Diagn Interv Imaging 2021;102:69–75.
[13] Jalaber C, Lapotre T, Morcet-Delattre T, Ribet F, Jouneau S, Lederlin M. Chest CT in COVID-19 pneumonia: a review of current knowledge. Diagn Interv Imaging 2020;101:431–7.
[14] Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26:297–302.
[15] Boedeker KL, McNitt-Gray MF, Rogers SR, Truong DA, Brown MS, Gjertson DW, et al. Emphysema: effect of reconstruction algorithm on CT imaging measures. Radiology 2004;232:295–301.
[16] Gierada DS, Bierhals AJ, Choong CK, Bartel ST, Ritter JH, Das NA, et al. Effects of CT section thickness and reconstruction kernel on emphysema quantification relationship to the magnitude of the CT emphysema index. Acad Radiol 2010;17:146–56.
[17] Kim HJ, Li G, Gjertson D, Elashoff R, Shah SK, Ochs R, et al. Classification of parenchymal abnormality in scleroderma lung using a novel approach to denoise images collected via a multicenter study. Acad Radiol 2008;15:1004–16.
[18] Zhao W, Zhang W, Sun Y, Ye Y, Yang J, Chen W, et al. Convolution kernel and iterative reconstruction affect the diagnostic performance of radiomics and deep learning in lung adenocarcinoma pathological subtypes. Thorac Cancer 2019;10:1893–903.
[19] He L, Huang Y, Ma Z, Liang C, Liang C, Liu Z. Effects of contrast-enhancement, reconstruction slice thickness and convolution kernel on the diagnostic performance of radiomics signature in solitary pulmonary nodule. Sci Rep 2016;6:34921.
[20] Shafiq-ul-Hassan M, Zhang GG, Hunt DC, Latifi K, Ullah G, Gillies RJ, et al. Accounting for reconstruction kernel-induced variability in CT radiomic features using noise power spectra. J Med Imaging 2017;5:1.
[21] Chen A, Karwoski RA, Gierada DS, Bartholmai BJ, Koo CW. Quantitative CT analysis of diffuse lung disease. Radiographics 2020;40:28–43.
[22] Ligero M, Jordi-Ollero O, Bernatowicz K, Garcia-Ruiz A, Delgado-Muñoz E, Leiva D, et al. Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis. Eur Radiol 2021;31:1460–70.
[23] Choe J, Lee SM, Do K-H, Lee G, Lee J-G, Lee SM, et al. Deep learning–based image conversion of CT reconstruction kernels improves radiomics reproducibility for pulmonary nodules or masses. Radiology 2019;292:365–73.