IOPscience Journals Books Publishing Support Login

PDF

Help

PAPER

Deep learning versus iterative reconstruction on image quality and dose reduction in abdominal CT: a live animal study

Jason Z Zhang[1], Halemane Ganesh[2], Flavius D Raslau[2], Rashmi Nair[2], Edward Escott[2], Chi Wang[3], Ge Wang[4] and Jie Zhang[5,2]

Published 8 July 2022 • © 2022 Institute of Physics and Engineering in Medicine

Physics in Medicine & Biology, Volume 67, Number 14

Citation Jason Z Zhang *et al* 2022 *Phys. Med. Biol.* **67** 145009

DOI 10.1088/1361-6560/ac7999

jnzh222@uky.edu

[1] Math, Science, and Technology Center, Lexington, KY 40513, United States of America

[2] Department of Radiology, University of Kentucky College of Medicine, Lexington, KY 40536 United States of America

[3] Department of Statistics, University of Kentucky, Lexington, KY 40536, United States of America

[4] Department of Biomedical Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180, United States of America

[5] Author to whom any correspondence should be addressed.

Ge Wang https://orcid.org/0000-0002-2656-7705

Jie Zhang https://orcid.org/0000-0003-2006-9964

1. Received 3 February 2022
2. Revised 24 May 2022
3. Accepted 16 June 2022
4. Published 8 July 2022

Buy this article in print

Journal RSS

Sign up for new issue notifications

Abstract

Objective. While simulated low-dose CT images and phantom studies cannot fully approximate subjective and objective effects of deep learning (DL) denoising on image quality, live animal models may afford this assessment. This study is to investigate the potential of DL in CT dose reduction on image quality compared to iterative reconstruction (IR). *Approach.* The upper abdomen of a live 4 year old sheep was scanned on a CT scanner at different exposure levels. Images were reconstructed using FBP and ADMIRE with 5 strengths. A modularized DL network with 5 modules was used for image reconstruction via progressive denoising. Radiomic features were extracted from a region over the liver. Concordance correlation coefficient (CCC) was applied to quantify agreement between any two sets of radiomic features. Coefficient of variation was calculated to measure variation in a radiomic feature series. Structural similarity index (SSIM) was used to measure the similarity between any two images. Diagnostic quality, low-contrast detectability, and image texture were qualitatively evaluated by two radiologists. Pearson correlation coefficient was computed across all dose-reconstruction/denoising combinations. *Results.* A total of 66 image sets, with 405 radiomic features extracted from each, are analyzed. IR and DL can improve diagnostic quality and low-contrast detectability and similarly modulate image texture features. In terms of SSIM, DL has higher potential in preserving image structure. There is strong correlation between SSIM and radiologists' evaluations for diagnostic quality (0.559) and low-contrast detectability (0.635) but moderate correlation for texture (0.313). There is moderate correlation between CCC of radiomic features and radiologists' evaluation for diagnostic quality (0.397), low-contrast detectability (0.417), and texture (0.326), implying that improvement of image features may not relate to improvement of diagnostic quality. *Conclusion.* DL shows potential to further reduce radiation dose while preserving structural similarity, while IR is favored by radiologists and more predictably alters radiomic features.

PDF

Help

Export citation and abstract BibTeX RIS

← **Previous** article in issue

Next article in issue →

Introduction

Iterative reconstruction (IR), as an alternative reconstruction algorithm to traditional filtered back projection (FBP), remains the most accessible computed tomography (CT) dose reduction technology. IR can remove image noise from low-dose CT (LDCT) using a variety of mathematical models (Padole *et al* 2015, Macri *et al* 2016, Morimoto *et al* 2017). However, the use of the higher IR strengths will produce overly 'smooth' or 'plastic' image texture that is deemed undesirable by radiologists. Similar to FBP, IR algorithms are also limited by low-contrast detectability at large radiation exposure reductions (Mileto *et al* 2019).

PDF
Help

With recent advances in artificial intelligence and improvements in hardware performance, deep learning (DL) has been applied in LDCT denoising and shows promising results (Jin *et al* 2017, Kaur *et al* 2018, Shan *et al* 2018, Yang *et al* 2018, Park *et al* 2019, Shan *et al* 2019, Gholizadeh-Ansari *et al* 2020, Greffier *et al* 2020, Hata *et al* 2020, Arndt *et al* 2021, Mohammadinejad *et al* 2021, Nam *et al* 2021, Noda *et al* 2021, Park *et al* 2021). Compared to IR, the existing studies show that DL approach demonstrates improved or comparable noise suppression and structural fidelity (Shan *et al* 2019, Lenfant *et al* 2020, Rozema *et al* 2020, Shin *et al* 2020). In some cases, i.e. CT pulmonary angiography, DL shows more potential in radiation dose reduction or image quality improvement (Lenfant *et al* 2020, Singh *et al* 2020). However, DL may also degrade spatial resolution (Shin *et al* 2020).

While IR has been widely adopted in clinical applications, DL is still in its early stage in CT image reconstruction and denoising (Arndt *et al* 2021). The performance of DL algorithms is highly dependent on the selection of neural network models and the datasets for training, validating and testing (Shin *et al* 2020). Quantitative analysis based on signal-to-noise ratio and contrast-to-noise ratio and subjective evaluation by radiologists may not fully elucidate the impact of DL algorithms on image quality, especially image texture variations (Rozema *et al* 2020).

To date, most studies have used routine-dose CT and simulated LDCT images (Shan *et al* 2019), limited pairs of LDCT and standard-dose CT (SDCT) data sets (Hata *et al* 2020, Park *et al* 2021), or phantoms/cadavers (Rozema *et al* 2020, Shin *et al* 2020, Racine *et al* 2021) for DL training and performance evaluation. This study is to investigate the potential of DL in CT dose reduction while preserving image texture when compared to IR. We performed a live animal study which allows repeatable testing under controlled conditions and provides a more realistic setting to evaluate the potential of DL in CT denoising. Similar to other studies (Hata *et al* 2020, Nam *et al* 2021), we adopted radiologists' subjective assessment of overall image quality and low-contrast detectability as metrics for evaluation. Furthermore, we directly investigated radiomic feature variations to better understand the impact of the DL algorithm on underlying attributes of image quality, along with subjective texture changes visually perceived by radiologists. A unique feature of our comparative study is that well-controlled animal scans have been used to compare a state-of-the-art Siemens IR and a high-impact end-to-process adaptive deep denoising network (MAP-NN).

Material and methods

Animal subject

The animal used for the study was a 4 year old, 72 kg, hornless, white Dorper ewe procured from the University of Kentucky Research Sheep Center, Department of Food and Animal Sciences. The Division of Laboratory Animal Resources Experimental Surgery staff and a veterinarian oversaw anesthesia induction, monitoring, and animal transport. The animal fasted overnight and was induced with a mixture of midazolam (0.4 mg kg^{-1}, IV) þ ketamine (5.5 mg kg^{-1}, IV) and orotracheal intubated and maintained on isoflurane (1.75%–2.0%) in 100% O2 with breathing self-regulated. Isotonic crystalloid (0.9% NaCl, 5–10 ml kg^{-1} h^{-1}) was administered for the duration of the study. At the conclusion of imaging, the animal was euthanized by sodium pentobarbital overdose while remaining under anesthesia.

This study was reviewed and approved by the University of Kentucky Institutional Animal Care and Use Committee and conducted in accordance with principles in the National Research Council, 2011, Guide for the Care and Use of Laboratory Animals.

Scanning protocol

The upper abdomen of the sheep was scanned under anesthesia on a Siemens Force CT scanner (Siemens, Erlangen, Germany). The sheep was in a supine position in the gantry for the acquisition of CT images. A routine abdominal CT protocol was used to acquire CT images to establish the starting reference dose. The refence protocol was tuned by adjusting the tube current until two radiologists were satisfied with the diagnostic image quality. Tube voltage was fixed at 120 kV, pitch at 0.6, and collimation at 192 × 0.6 mm. Automated exposure control (CAREDose 4D and CARE kV) was turned off to control tube current in each scan. Tube current was varied from 360 mAs (reference) to 30 mAs in 10 levels. The CTDIvol were 18.11 mGy, 16.74 mGy, 15.40 mGy, 13.39 mGy, 11.69 mGy, 10.21 mGy, 8.35 mGy, 5.02 mGy, 3.35 mGy, and 1.67 mGy, with corresponding size-specific dose estimate (SSDE) of 23.54 mGy, 21.76 mGy, 20.02 mGy, 17.41 mGy, 15.19 mGy, 13.27 mGy, 10.86 mGy, 6.53 mGy, 4.36 mGy, and 2.17 mGy. CT Images were reconstructed using both Advanced Modeled Iterative Reconstruction (ADMIRE) with strengths 1–5 and FBP at each exposure level. The Br40d reconstruction kernel was used for image reconstruction, with a FOV of 404 × 404 mm and matrix of 512 × 512, yielding 0.79 × 0.79 in-plane spatial resolution with slice thickness of 5 mm. A total of 60 combinations were generated by 10 dose levels × 6 reconstruction levels (IR with 5 reconstruction strengths and FBP).

ADMIRE was introduced in 2014 by Siemens Healthineers. It is a statistical IR method that, uses a regularization mechanism working in a 3D voxel neighborhood, separates noise from actual anatomical structures, and preserves anatomical texture (Ramirez-Giraldo and G K A R R 2018, Kataria *et al* 2021). Currently, ADMIRE is widely implemented in clinical practice, showing comparable reconstruction times to FBP but better dose reduction potential.

DL-based CT denoising

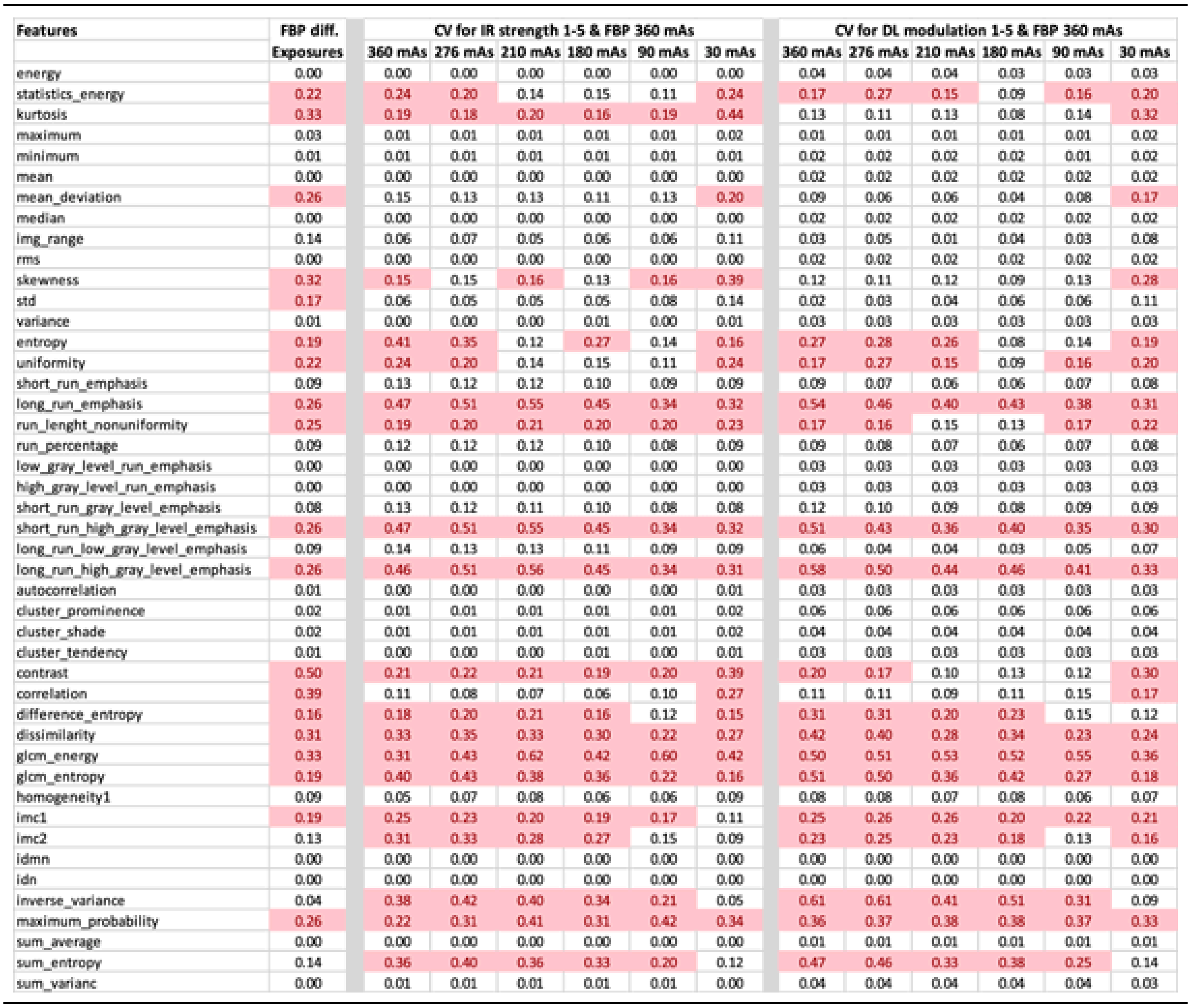
A recently-developed, open-source modularized deep neural network (MAP-NN) was used for CT image reconstruction at each exposure level via progressive denoising (Shan *et al* 2019). In general, conventional networks are end-to-end denoising networks that produce denoised LDCT images directly. MAP-NN decomposed the overall network into 5 identical network modules that allows maximization of the diagnostic performance. The learning workflow allows radiologists-in the-loop to optimize the denoising depth in a task-specific fashion (Shan *et al* 2019). Each module denoises an inputted image and passes the denoised image to the next module, which denoises it further. All 5 modules combined with 10 exposure levels produced 50 sets of DL denoised images.

MAP-NN can also reduce the noise level of normal dose CT (NDCT) images, since MAP-NN does not only learn to denoise but also encode a noise reduction direction in the module, which is not possible for conventional end-to-end denoising networks. MAP-NN has been trained with the Mayo LDCT Dataset, and tested on separate chest and abdominal CT exams from Massachusetts General Hospital (Shan *et al* 2019). The assessment on different training sets has been addressed, thus, no additional training was performed in this study.

Radiomics feature extraction

A region of interest (ROI) was manually selected over the liver. Radiomic features were extracted from the ROI using a Python-based package on GitHub developed based on Aerts *et al* (2014). There was a total of 45 first and second order features, and 360 multiscale texture features filtered by 3D Wavelet transform (45 × 8 oct-bands). The same ROI was used for all the image sets. The list of radiomic features is included in table 1.

Table 1. Coefficient of variation (CV) for FBP, IR strength 1–5 with FBP 360 mAs, and DL modulation 1–5 at different exposures with FBP 360 mAs at different exposures. 45 first- and second-order image features are included. CV of >15%, indicating large variation, are highlighted.

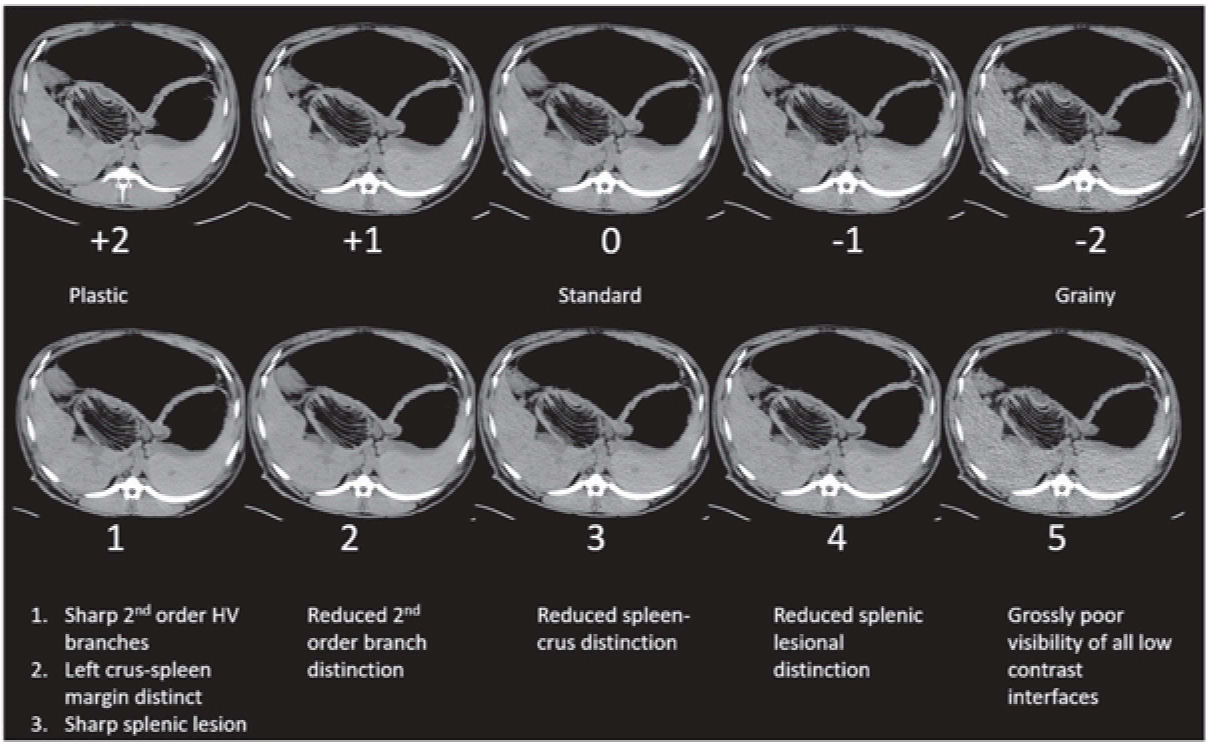


Features	FBP diff.	CV for IR strength 1-5 & FBP 360 mAs						CV for DL modulation 1-5 & FBP 360 mAs					
	Exposures	360 mAs	276 mAs	210 mAs	180 mAs	90 mAs	30 mAs	360 mAs	276 mAs	210 mAs	180 mAs	90 mAs	30 mAs
energy	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.04	0.04	0.03	0.03	0.03
statistics_energy	0.22	0.24	0.20	0.14	0.15	0.11	0.24	0.17	0.27	0.15	0.09	0.16	0.20
kurtosis	0.33	0.19	0.18	0.20	0.16	0.19	0.44	0.13	0.11	0.13	0.08	0.14	0.32
maximum	0.03	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.02
minimum	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.01	0.02
mean	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.02	0.02	0.02	0.02
mean_deviation	0.26	0.15	0.13	0.13	0.11	0.13	0.20	0.09	0.06	0.06	0.04	0.08	0.17
median	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.02	0.02	0.02	0.02
img_range	0.14	0.06	0.07	0.05	0.06	0.06	0.11	0.03	0.05	0.01	0.04	0.03	0.08
rms	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.02	0.02	0.02	0.02
skewness	0.32	0.15	0.15	0.16	0.13	0.16	0.39	0.12	0.11	0.12	0.09	0.13	0.28
std	0.17	0.06	0.05	0.05	0.05	0.08	0.14	0.02	0.03	0.04	0.06	0.06	0.11
variance	0.01	0.00	0.00	0.00	0.01	0.00	0.01	0.03	0.03	0.03	0.03	0.03	0.03
entropy	0.19	0.41	0.35	0.12	0.27	0.14	0.16	0.27	0.28	0.26	0.08	0.14	0.19
uniformity	0.22	0.24	0.20	0.14	0.15	0.11	0.24	0.17	0.27	0.15	0.09	0.16	0.20
short_run_emphasis	0.09	0.13	0.12	0.12	0.10	0.09	0.09	0.09	0.07	0.06	0.06	0.07	0.08
long_run_emphasis	0.26	0.47	0.51	0.55	0.45	0.34	0.32	0.54	0.46	0.40	0.43	0.38	0.31
run_lenght_nonuniformity	0.25	0.19	0.20	0.21	0.20	0.20	0.23	0.17	0.16	0.15	0.13	0.17	0.22
run_percentage	0.09	0.12	0.12	0.12	0.10	0.08	0.09	0.09	0.08	0.07	0.06	0.07	0.08
low_gray_level_run_emphasis	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.03	0.03	0.03	0.03	0.03
high_gray_level_run_emphasis	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.03	0.03	0.03	0.03	0.03
short_run_gray_level_emphasis	0.08	0.13	0.12	0.11	0.10	0.08	0.08	0.12	0.10	0.09	0.08	0.09	0.09
short_run_high_gray_level_emphasis	0.26	0.47	0.51	0.55	0.45	0.34	0.32	0.51	0.43	0.36	0.40	0.35	0.30
long_run_low_gray_level_emphasis	0.09	0.14	0.13	0.13	0.11	0.09	0.09	0.06	0.04	0.04	0.03	0.05	0.07
long_run_high_gray_level_emphasis	0.26	0.46	0.51	0.56	0.45	0.34	0.31	0.58	0.50	0.44	0.46	0.41	0.33
autocorrelation	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.03	0.03	0.03	0.03	0.03
cluster_prominence	0.02	0.01	0.01	0.01	0.01	0.01	0.02	0.06	0.06	0.06	0.06	0.06	0.06
cluster_shade	0.02	0.01	0.01	0.01	0.01	0.01	0.02	0.04	0.04	0.04	0.04	0.04	0.04
cluster_tendency	0.01	0.00	0.00	0.00	0.01	0.00	0.01	0.03	0.03	0.03	0.03	0.03	0.03
contrast	0.50	0.21	0.22	0.21	0.19	0.20	0.39	0.20	0.17	0.10	0.13	0.12	0.30
correlation	0.39	0.11	0.08	0.07	0.06	0.10	0.27	0.11	0.11	0.09	0.11	0.15	0.17
difference_entropy	0.16	0.18	0.20	0.21	0.16	0.12	0.15	0.31	0.31	0.20	0.23	0.15	0.12
dissimilarity	0.31	0.33	0.35	0.33	0.30	0.22	0.27	0.42	0.40	0.28	0.34	0.23	0.24
glcm_energy	0.33	0.31	0.43	0.62	0.42	0.60	0.42	0.50	0.51	0.53	0.52	0.55	0.36
glcm_entropy	0.19	0.40	0.43	0.38	0.36	0.22	0.16	0.51	0.50	0.36	0.42	0.27	0.18
homogeneity1	0.09	0.05	0.07	0.08	0.06	0.06	0.09	0.08	0.08	0.07	0.08	0.06	0.07
imc1	0.19	0.25	0.23	0.20	0.19	0.17	0.11	0.25	0.26	0.26	0.20	0.22	0.21
imc2	0.13	0.31	0.33	0.28	0.27	0.15	0.09	0.23	0.25	0.23	0.18	0.13	0.16
idmn	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
idn	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
inverse_variance	0.04	0.38	0.42	0.40	0.34	0.21	0.05	0.61	0.61	0.41	0.51	0.31	0.09
maximum_probability	0.26	0.22	0.31	0.41	0.31	0.42	0.34	0.36	0.37	0.38	0.38	0.37	0.33
sum_average	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.01
sum_entropy	0.14	0.36	0.40	0.36	0.33	0.20	0.12	0.47	0.46	0.33	0.38	0.25	0.14
sum_varianc	0.00	0.01	0.01	0.01	0.01	0.01	0.00	0.04	0.04	0.04	0.04	0.04	0.03

Radiologists' evaluation

A total of 110 combinations of ten dose levels with FBP, five ADMIRE IR strengths, and five DL denoising levels were generated. For overall image quality and texture fidelity, all images in each combination of the 110 combinations were evaluated, similar to existing studies (Hata *et al* 2020, Nam *et al* 2021, Park *et al* 2021). For low-contrast detectability, one axial image at the level of the hepatic hilum from each combination, which showed anatomical structures of interest most visibly, was focused on. The method used to establish the reference standard was the ability to identify normal anatomic structures directly correlating with the diagnostic quality of the scan. The animal model did not have any lesions, but given that the detection of a lesion or identification of normal anatomic structures are interrelated during subjective radiologist interpretation and directly correlate with the diagnostic quality of the scan, the ability to identify normal anatomic structures was used as the reference standard. CT images were anonymized and randomized before being evaluated blindly by two abdominal subspecialty trained radiologists (both with greater than 10 years of experience) on a standard independent PACS station. The images were evaluated in two sessions. Overall image quality and texture were evaluated in one session, while low-contrast detectability was evaluated in the other session. There were more than two weeks between the sessions.

The overall diagnostic image quality, image texture, and low-contrast detectability were qualitatively assessed on multiple point scales (figure 1). A 5-point metric was used for overall diagnostic image quality: 5 = diagnostic, 4 = adequate, 3 = acceptable, 2 = marginal, 1 = not diagnostic. In the case of image texture, 0 was set as standard texture, positive numbers represented increased plasticity, and negative numbers represented increased grain (figure 1). The evaluation of texture was subjective and was determined by radiologists based on FBP. Low-contrast detectability evaluation focused on the second order branch of the hepatic veins, the left spleen-crus distinction, and a splenic lesion: 1 = sharp distinction for all low contrast interfaces; 2 = reduced 2nd order branch distinction; 3 = reduced spleen-crus distinction; 4 = reduced splenic lesion distinction; 5 = grossly poor visibility of all low contrast interfaces. These structures were chosen for the evaluation of low-contrast detectability since the lowest contrast is between the second order branches, followed by the spleen-crus, then the splenic lesion. Therefore, there would not be a time where the splenic lesion was less distinct, but the 2nd order branches were preserved. Figure 1 shows examples of images and criteria for assessment.


+2
+1
0
-1
-2
Plastic
Standard
Grainy
1
2
3
4
5
1. Sharp 2nd order HV branches
2. Left crus-spleen margin distinct
3. Sharp splenic lesion
Reduced 2nd order branch distinction
Reduced spleen-crus distinction
Reduced splenic lesional distinction
Grossly poor visibility of all low contrast interfaces

PDF

Help

Figure 1. Qualitative reference standards for radiologists' evaluation of image texture (upper image row) and contrast (lower image row). The standard texture was determined subjectively by radiologists based on FBP at routine dose, while the evaluation of low-contrast detectability focused on the second order branch of the hepatic veins, the left spleen-crus distinction, and a splenic lesion.

Data analysis

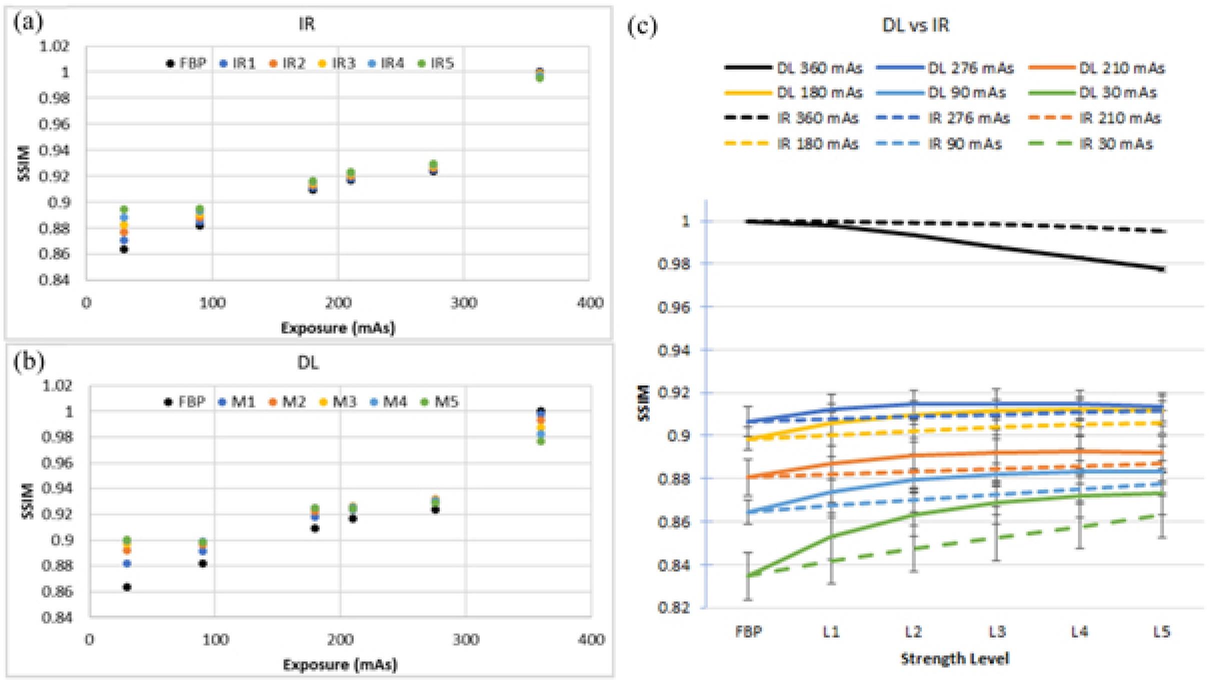
The concordance correlation coefficient (CCC) was applied to quantify the agreement between any two sets of Wavelet radiomic features. CV was used to test feature variation for a single image feature with different exposures or reconstruction/denoising strengths. The structural similarity index measure (SSIM) was calculated to evaluate the image quality of denoised images and compare its performance with that of FBP and IR. Pearson correlation coefficient (PCC) between SSIM and qualitative radiologists' evaluation was calculated for overall diagnostic image quality, low-contrast detectability, and image texture. The CCCs of the radiomics features were also correlated with radiologists' evaluation of diagnostic quality, low-contrast detectability, and image texture, respectively. Note that radiologists' evaluation used the lowest value for the best low-contrast detectability while CCC and SSIM use the highest value for the best similarity. To ensure a consistent direction for all evaluation systems, we considered the opposite (negative) of the evaluation score for low-contrast detectability in the PCC calculations. A similar method was applied to radiologists' evaluation for image texture, where 0 indicated standard texture and a negative or positive number indicated image texture being either grainy or plastic. To align with the trends of CCC and SSIM, we used the negative absolute value of the evaluation score for image texture in the PCC calculations. The correlation strength was assessed by general guidelines: 0.1–0.3 = weak correlation; 0.3–0.5 = moderate correlation; >0.5 = strong correlation (Cohen 1988). The 95% confidence intervals (CIs) were calculated for SSIM, CCC, and PCC. To calculate CIs for the more complex CCC, jackknifing with Fisher's Z was used to provide the coverage probabilities closest to nominal (Feng *et al* 2014).

PDF
Help

Results

Based on visual inspection and SSIM, four exposure levels (300 mAs, 240 mAs, 150 mAs, and 60 mAs) have large effects from respiratory motion, which results in positional changes. Considering the potential effects of motion on radiomic features, they are excluded from analysis. This exclusion has minimal impact on the results since the datasets for analysis still cover the range of exposure levels from 360 to 30 mAs at reasonable intervals.

Figure 2 shows the SSIM for both ADMIRE (IR) and DL-modulated images at 5 strength levels and a pair comparison as a function of radiation dose (mAs). IR and DL show comparable trends in SSIM with changes in exposure, but DL shows a larger SSIM range at each exposure level. DL shows a non-linear relationship between the increase in modulation strength and SSIM, while IR shows a linear relationship. This indicates that choosing appropriate modulation strength for DL is more challenging. Note that DL modulation strength levels do not necessarily correspond with IR strength levels.



PDF
Help

Figure 2. Structural similarity index measure (SSIM) as a function of radiation dose (mAs) for difference reconstruction algorithms, ADMIRE with 5 strengths (a), deep-learning with 5 modulations (b), and a pair comparison between ADMIRE and DL as a function of reconstruction strengths for each radiation dose (mAs) level (c). For each data set obtained by identical radiation dose, an image reconstructed by FBP with a tube current of 360 mAs was selected as reference image. Note: The DL modulation strength levels do not necessarily correspond with IR strength levels.

Table 1 illustrates CVs for FBP among different exposures, as well as IR strength 1–5 and DL modulation 1–5 at different exposures, with FBP 360 mAs as reference. 45 first- and second-order image features are included. CV of >15% are highlighted. The variations in image features show consistency between different exposures, IR, and DL, indicating that LDCT images can be denoised by either IR or DL to improve image quality while preserving image quality. For example, the variation of contrast and kurtosis can be reduced to within 10%. However, appropriate dose level, algorithm, and reconstruction strength level must be selected to avoid negative results. The appropriate use of IR and DL depends on the exposure (radiation dose level), i.e. for uniformity, IR at 90 mAs and DL at 180 mAs produce lowest CVs.

Figure 3 shows a heatmap of CCC of the Wavelet features versus FBP 360 mAs as a function of reconstruction/denoising strengths for each dose level. A larger number (yellowish) indicates a better agreement between two sets of radiomic features. The heatmap demonstrates how radiation exposure level and denoising level modulate radiomic features, with FBP 360 mAs as reference. Overall, IR shows a better texture feature conservation compared to DL. IR also follows a predictable trend in altering radiomic features with the change of exposure level and the change of the denoising strengths than DL, which has more variable performance.

PDF
Help

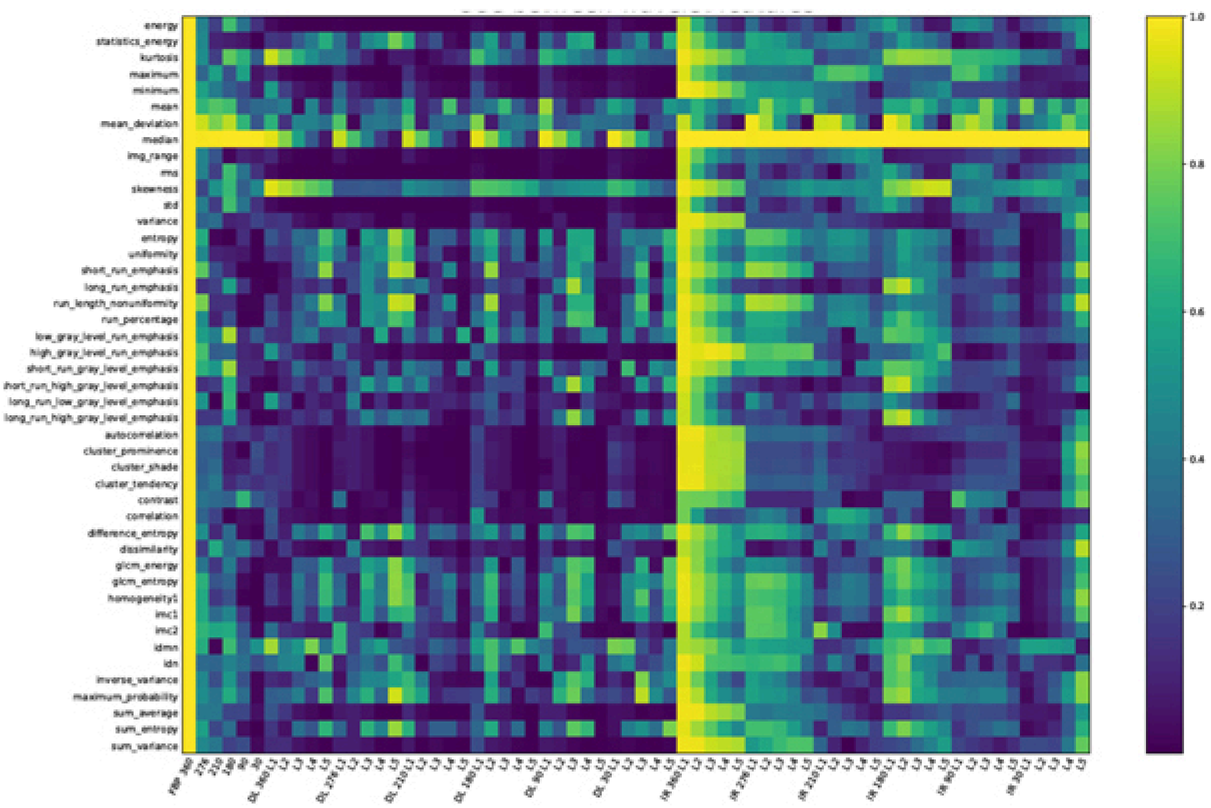
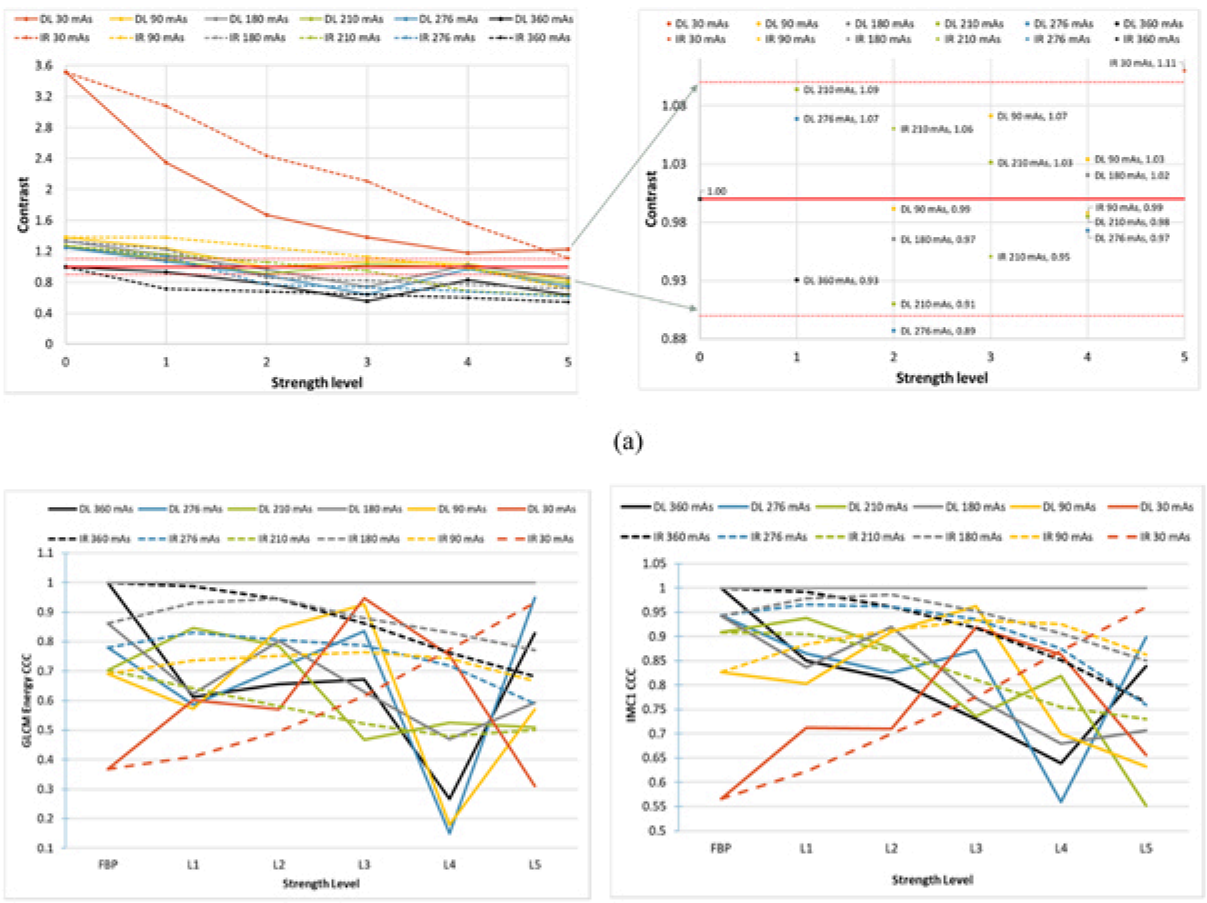


Figure 3. A heatmap of concordance correlation coefficient (CCC) of the Wavelet features verse FBP 360 mAs as a function of reconstruction/denoising strengths for each dose level. A larger number (more yellow) indicates a better agreement between two sets of radiomic features. The *Y* axis lists 45 Wavelet radiomic features. The *X* axis consists of image reconstruction/denoising at different exposure levels, with DL on the left side and IR on the right side. The exposure levels are 360 mAs, 276 mAs, 210 mAs, 180 mAs, 90 mAs, and 30 mAs, respectively.

For a closer look at figure 3, figure 4 shows a pair comparison of contrast between IR and DL and two examples of CCC of 3D Wavelet filtered features (GLCM-Energy, Informational Measure of Correlation 1) as a function of reconstruction/denoising strengths for each dose level. Contrast values are normalized by the value of the FBP 360 mAs image (reference). Note that here, contrast is a radiomic feature, different from the subjective low-contrast detectability evaluated by radiologists as shown in figure 1. The values within 10% of the baseline are spotlighted on the right. DL has more points within 10% of baseline, while IR shows a more predictable trend. Based on figures (b) and (c), IR shows a more predicable trend in altering high-order Wavelet features, depending on exposure levels. Meanwhile, DL has potential to improve image features at lower dose levels, but is more erratic.



(a)

(b)

(c)

Figure 4. Pair comparison of contrast, a radiomic feature, between ADMIRE (IR) and deep-learning (DL) (a) and two examples of concordance correlation coefficient of 3D Wavelet filtered features, GLCM-energy (b) and informational measure of correlation 1 (c), as a function of reconstruction/denoising strengths for each radiation dose (mAs) level.

For each dataset, an image reconstructed by FBP with a tube current of 360 mAs was selected as reference image, and contrast values are normalized by the value of the image reconstructed by FBP with a tube current of 360 mAs. The right graph is zoomed in to show the values at the baseline (FBP 360 mAs) +/– 10%.

Table 2 shows qualitative ratings of low-contrast detectability, texture, and overall diagnostic scores by two radiologists. Note that as explained in the materials and methods, a lower score indicates better low-contrast detectability, while for diagnostic quality, a higher score is better. For image texture, 0 is the standard and negative indicates grainy while positive indicates plastic. Overall, IR is favored by radiologists at most exposure levels for diagnostic quality, low-contrast detectability, and texture. A conspicuous and important exception is that DL appears to have greater potential to improve image texture at very low exposures, i.e. 90 and 30 mAs.

PDF
Help

Table 2. Qualitative ratings of contrast, texture, and overall diagnostic scores by two radiologists. Note that as explained in the materials and methods, a lower score indicates better low-contrast detectability, while for diagnostic quality, a higher score is better. For image texture, 0 is the standard and negative indicates grainy while positive indicates plastic.

Low-contrast detectability	FBP	DL1	DL2	DL3	DL4	DL5	IR1	IR2	IR3	IR4	IR5
360 mAs	1.5	1.5	2.0	2.0	2.0	3.0	1.5	1.5	1.5	1.5	2.0
276 mAs	1.5	1.5	1.5	2.5	3.5	2.5	1.0	1.0	1.5	1.5	1.5
210 mAs	3.0	2.5	2.0	2.5	3.0	4.0	2.0	2.5	2.0	1.5	2.5
180 mAs	3.5	3.5	2.5	3.5	2.5	3.0	2.5	2.5	2.5	2.0	3.0
90 mAs	4.0	4.0	4.0	4.5	5.0	4.5	3.0	3.5	4.0	3.0	2.5
30 mAs	3.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
Texture	FBP	DL1	DL2	DL3	DL4	DL5	IR1	IR2	IR3	IR4	IR5
360 mAs	–0.5	0.0	0.0	1.0	1.5	1.5	0.0	0.0	0.5	0.5	1.0
276 mAs	0.0	0.0	0.0	1.0	2.0	1.5	–0.5	0.0	0.5	1.5	1.5
210 mAs	–0.5	–1.0	1.0	1.0	1.5	2.0	0.5	–0.5	–0.5	1.0	1.0
180 mAs	–0.5	–0.5	0.5	1.0	0.5	2.0	–0.5	0.0	0.0	0.5	1.5

Low-contrast detectability	FBP	DL1	DL2	DL3	DL4	DL5	IR1	IR2	IR3	IR4	IR5
90 mAs	–1.5	–1.5	–1.5	0.0	–0.5	0.5	–1.5	–1.0	–1.5	–0.5	0.5
30 mAs	–1.5	–2.0	–2.0	0.0	0.0	0.0	–2.0	–2.0	–2.0	–2.0	0.0

Diagnostic	FBP	DL1	DL2	DL3	DL4	DL5	IR1	IR2	IR3	IR4	IR5
360 mAs	5.0	4.5	3.5	3.0	4.0	2.5	4.5	4.5	4.5	4.5	4.5
276 mAs	5.0	4.5	5.0	4.0	4.5	3.5	5.0	5.0	5.0	4.5	4.5
210 mAs	3.5	3.5	4.5	4.0	3.0	2.0	4.0	4.0	4.5	4.5	4.5
180 mAs	3.5	4.0	4.0	3.5	3.5	3.5	4.5	3.5	4.5	4.5	3.5
90 mAs	2.0	3.0	2.0	1.5	2.0	1.5	2.5	3.0	2.0	3.5	3.0
30 mAs	3.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

PDF

Help

Figure 5 the PCC between radiologists' evaluation and CCC of radiomic features as well as radiologists' evaluation and SSIM for overall diagnostic quality (a), image contrast (b), and image texture (c), respectively. Both CCC and SSIM are calculated with FBP 360 mAs as reference. There is a moderate correlation between radiologists' evaluation and CCC for diagnostic quality (0.397, 95% CI [0.172, 0.583]), low-contrast detectability (0.417, 95% CI [0.194, 0.598]), and texture (0.326, 95% CI [0.09, 0.526]). There is a strong correlation between radiologists' evaluation and SSIM for diagnostic quality (0.559, 95% CI [0.367, 0.706]) and contrast (0.635, 95% CI [0.465, 0.761]) but moderate correlation for texture (0.313, 95% CI [0.077, 0.516]). The calculated PCC between SSIM and CCC is 0.137 (95% CI [–0.108, 0.367]), indicating there is weak correlation between radiomic feature variation and image structure similarity.

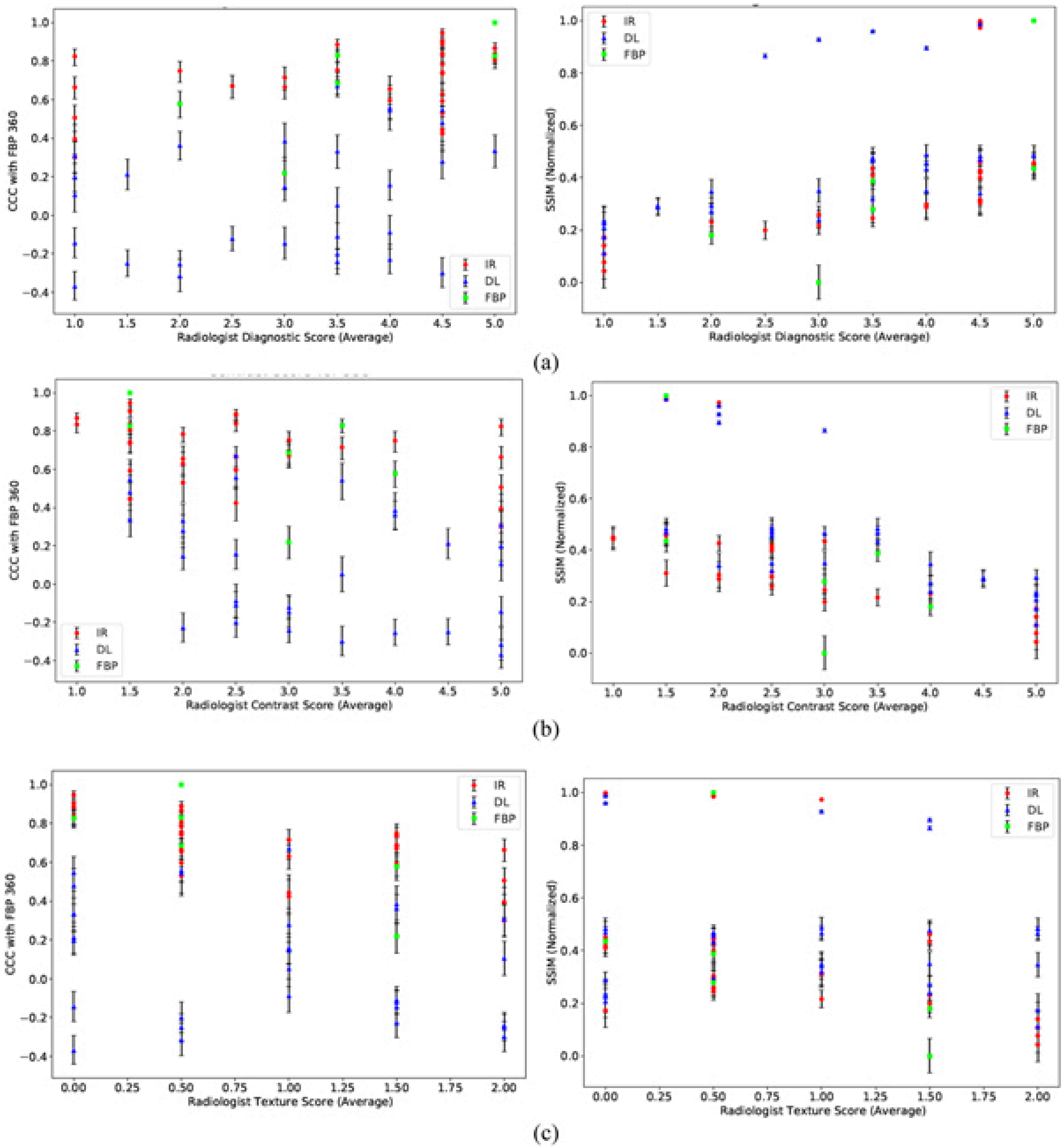


Figure 5. The pearson correlation coefficient (PCC) between radiologists' evaluation and CCC of Wavelet features (left column) and radiologists' evaluation and SSIM (right column) for diagnostic quality (a), low-contrast detectability (b), and image texture (c), respectively. Both CCC and SSIM are calculated with FBP 360 mAs as reference. There is a moderate correlation between radiologists' evaluation and CCC for diagnostic quality (0.397, 95% CI [0.172, 0.583]), low-contrast detectability (0.417, 95% CI [0.194, 0.598]), and texture (0.326, 95% CI [0.09, 0.526]). There is a strong correlation between

radiologists' evaluation and SSIM for diagnostic quality (0.559, 95% CI [0.367, 0.706]) and contrast (0.635, 95% CI [0.465, 0.761]) but moderate correlation for texture (0.313, 95% CI [0.077, 0.516]).

Discussion

Typically, deep convolutional neural network-based image denoising methods are evaluated by radiologists or in terms of metrics such as SSIM (Ti *et al* 2021, Usui *et al* 2021, Yoon *et al* 2021), with a focus on image quality to meet diagnostic standards. However, SSIM does not reveal if the DL results preserve the visibility of subtle lesions or if they alter the CT image properties such as the noise texture (KC *et al* 2021). With the rapid growth of radiomics study, image features are of another major concern. It becomes necessary to look into image feature variations while evaluating image quality. In this study, we calculate the SSIM. Most importantly, we incorporate the qualitative assessment of image texture as determined by radiologists. Along with radiologists' subjective evaluation, we perform a more nuanced study of DL low-contrast features and their noise textures to better understand the impact of the DL algorithm on these underlying attributes of image quality.

PDF
Help

Based on SSIM, DL-based image denoising outperforms IR methods (figure 2). Nevertheless, radiologists' evaluation shows that IR is favored for reducing radiation dose while preserving image quality. Our observation agrees with another study where IR showed better overall image quality and less subjective noise (Nam *et al* 2021). This may be partially due to their familiarity with IR in their routine practice. When higher denoising strengths are appropriately paired with radiation doses at the low to very low spectrum, the undesirable texture 'smoothing' effects are mitigated. This conclusion is consistent with previous results (Raslau *et al* 2019, Shan *et al* 2019, Noda *et al* 2021, Usui *et al* 2021). However, our data suggest that the DL denoising method may have more potential to improve the texture specifically at low doses (table 2). With regard to image feature variation, an interesting finding is that IR and DL show many similarities in altering the same types of radiomic features (table 1). Both can improve image features such as skewness, mean_deviation, contrast, entropy and more, while both worsen some image features, such as GLCM-energy, IMC1, IMC2, and sum_entropy. DL may have more potential at low radiation dose levels to reach comparable image quality with high radiation dose images compared to IR; however, IR follows a more predictable trend in modulating radiomic features, as shown in figure 4 with the focused analysis of two Wavelet features (GLCM-energy and IMC1). DL has more variations in dose and strength combination for improving image quality, especially texture feature, making it more difficult to predict its performance.

We use CCC to measure the agreement between two sets of radiomic features and CV to measure the relative dispersion of data points in a radiomic feature series (e.g. IR strengths of 1–5) around the mean. CCC and CV have been widely used to evaluate reproducibility of radiomic features (Balagurunathan *et al* 2014, van Timmeren *et al* 2016a, 2016b, Larue *et al* 2017, Weller *et al* 2017, Moradmand *et al* 2020). In general, radiomic features are considered reproducible if CCC > 0.85 or CV <10%. This study mainly focuses on the changes of radiomic features as a function of reconstruction/denoising strengths and exposure levels. Therefore, we do not set a specific CCC threshold. A CV threshold of >15% is set to indicate large variation.

PDF
Help

The PCC analysis shows there is a strong correlation between radiologists' evaluation and SSIM, indicating that SSIM may be an appropriate metric to evaluate image quality. This observation agrees with some existing studies that show SSIM demonstrates good agreement with human observers in tasks using reference images (Renieblas *et al* 2017). However, this may be limited to diagnostic quality and low-contrast detectability in routine practice. Neither SSIM nor CCC of radiomic features has a strong correlation with image texture as perceived by radiologists, which agrees with Mason *et al* (2020) where SSIM shows lower Spearman rank order correlation coefficient with radiologists' scoring than other image quality metrics. This observation indicates that SSIM may be appropriate only for some specific tasks. Denoising based on the modulation of radiomic features may not be useful since there is a moderate correlation at the lower end with radiologists' evaluation. One may expect radiomic features to adequately represent image texture, but our results suggest that changes in radiomic features may not necessarily be appreciated visually or by the structure similarity measurement.

We employ a live animal model because *in vivo* evaluation of low-contrast detectability (e.g. splenic lesion distinction) was deemed paramount. The use of an anthropomorphic phantom cannot reach this goal. Another reason is that attention was brought to the importance of image texture, which is not adequately captured by current quantitative metrics. Therefore, a direct comparison is created for the qualitative assessment of image texture.

There are some limitations worthy of mention. One limitation is that our findings are based on a single IR algorithm (ADMIRE) from one vendor and a single DL denoising method, thus, they cannot be generalized. Given that different IR algorithms and DL denoising methods work differently and possess different balances of advantages and disadvantages, how these results translate to other techniques has not been evaluated at this stage. MAP-NN is adopted because it is an end-to-process denoising mapping with radiologists in the loop so that a denoising process can be effectively and efficiently guided for a specific task. Hopefully, this study will help raise awareness regarding the differences between DL and IR, especially in the context of radiomics. Another potential criticism may emerge from DL denoising training. The DL method was trained using clinical data, while we apply the method directly to denoise CT images of the animal.

PDF
Help

Transfer learning may improve the performance of the DL denoising method, but it was not performed in our study due to limited animal data. Oftentimes, in clinical practice, there is no additional training involved for a specific task or a subject; hence, testing the generalizability is important in this field. Another concern is that only one subject was included in this study, and that the low-contrast detectability evaluation was based on the hepatic veins, the left spleen-crus distinction, and a splenic lesion, which may not fully represent task-driven diagnostic image quality assessment. A further potential limitation is that evaluation by a larger number of radiologists may help reduce variability of the qualitative evaluation.

In conclusion, our live animal study advances our knowledge regarding the application of DL denoised methods in radiation dose reduction while preserving image quality with attention to the relationships between SSIM, radiologists' evaluation, and radiomic features. The strong correlation between SSIM and radiologists' evaluation suggests SSIM may be used to measure image quality for diagnostic quality and low-contrast detectability. However, the visually perceived image texture cannot be adequately presented by SSIM, and differ from radiomic features extracted from the images. This conclusion should be taken into consideration in future radiomic studies. Recently, FDA-approved DL techniques (e.g. TrueFidelity) have been gradually introduced to clinical practice (Racine *et al* 2021). Further studies should be performed with these commercialized reconstruction algorithms.

You may also like

JOURNAL ARTICLES

Radiomics-guided radiation therapy: opportunities and challenges

Prediction of ovarian cancer prognosis using statistical radiomic features of ultrasound images

Applications and limitations of radiomics

PST-Radiomics: a PET/CT lymphoma classification method based on pseudo spatial-temporal radiomic features and structured atrous recurrent convolutional neural network

The stability of oncologic MRI radiomic features and the potential role of deep learning: a review

Predicting programmed death-ligand 1 expression level in non-small cell lung cancer using a combination of peritumoral and intratumoral radiomic features on computed tomography

PDF

Help

PDF

Help

PDF

Help

PDF

Help

PDF

Help

PDF

Help

PDF

Help

PDF

Help

PDF

Help

PDF

Help

PDF

Help

PDF

Help

PDF

Help

PDF

Help

PDF

Help

PDF

Help

PDF

Help

IOPSCIENCE

Journals

Books

IOP Conference Series

About IOPscience

Contact Us

Developing countries access

IOP Publishing open access policy

Accessibility

IOP PUBLISHING

Copyright 2024 IOP Publishing

Terms and Conditions

Disclaimer

Privacy and Cookie Policy

PUBLISHING SUPPORT

Authors

Reviewers

Conference Organisers