Computers in Biology and Medicine 159 (2023) 106931

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Generative adversarial network with radiomic feature reproducibility analysis for computed tomography denoising

Jina Lee [a,b], Jaeik Jeon [a], Youngtaek Hong [a,d,*], Dawun Jeong [a,b], Yeonggul Jang [a,b], Byunghwan Jeon [c], Hye Jin Baek [e], Eun Cho [e], Hackjoon Shim [a], Hyuk-Jae Chang [a,f]

[a] CONNECT-AI Research Center, Yonsei University College of Medicine, Seoul, 03764, South Korea
[b] Brain Korea 21 PLUS Project for Medical Science, Yonsei University, Seoul, 03722, South Korea
[c] Division of Computer Engineering, Hankuk University of Foreign Studies, Yongin, 17035, South Korea
[d] Ontact Health, Seoul, 03764, South Korea
[e] Department of Radiology, Gyeongsang National University Changwon Hospital, Gyeongsang National University School of Medicine, Changwon, 51472, South Korea
[f] Division of Cardiology, Severance Cardiovascular Hospital, Yonsei University College of Medicine, Seoul, 03722, South Korea

ARTICLE INFO

Keywords:
Medical image denoising
Generative adversarial networks
Radiomics
Parameter tuning

ABSTRACT

Background: Most computed tomography (CT) denoising algorithms have been evaluated using image quality analysis (IQA) methods developed for natural image, which do not adequately capture the texture details in medical imaging. Radiomics is an emerging image analysis technique that extracts texture information to provide a more objective basis for medical imaging diagnostics, overcoming the subjective nature of traditional methods. By utilizing the difficulty of reproducing radiomics features under different imaging protocols, we can more accurately evaluate the performance of CT denoising algorithms.

Method: We introduced radiomic feature reproducibility analysis as an evaluation metric for a denoising algorithm. Also, we proposed a low-dose CT denoising method based on a generative adversarial network (GAN), which outperformed well-known CT denoising methods.

Results: Although the proposed model produced excellent results visually, the traditional image assessment metrics such as peak signal-to-noise ratio and structural similarity failed to show distinctive performance differences between the proposed method and the conventional ones. However, radiomic feature reproducibility analysis provided a distinctive assessment of the CT denoising performance. Furthermore, radiomic feature reproducibility analysis allowed fine-tuning of the hyper-parameters of the GAN.

Conclusion: We demonstrated that the well-tuned GAN architecture outperforms the well-known CT denoising methods. Our study is the first to introduce radiomics reproducibility analysis as an evaluation metric for CT denoising. We look forward that the study may bridge the gap between traditional objective and subjective evaluations in the clinical medical imaging field.

1. Introduction

Computed tomography (CT) is a reliable imaging test for diagnosing diseases. The quality of CT images is affected by various scanning parameters such as tube voltage and current. X-ray dose is directly related to image quality [1]; a high dose allows high-quality images to be acquired, but it can cause DNA damage and cell deformity [2]. Therefore, reconstruction algorithms try to improve the quality of CT images acquired at low-dose parameters. The iterative reconstruction algorithm [3] is a well-known technique for reducing radiation dose and improving image quality in CT. Recently, deep learning has shown promising performances in the field of computer vision and medical image processing [4,5]. Generative adversarial networks (GANs) [6], which are unsupervised learning networks, show excellent performance in CT image reconstruction with low noise [7,8]. Although GAN-based post-processing algorithms have been validated in objective and subjective analysis, their clinical adoption is still limited.

Peak signal-to-noise ratio (PSNR) [9], structural similarity index measure (SSIM) [10] and noise of the region of interest (ROI) are the most widely used objective image quality metrics. They are based on a pixel-wise difference, but the human visual system is not based on pixel-wise perception. Thus, the evaluation results do not correlate well with perceptual satisfaction. However, most CT denoising algorithms

* Corresponding author at: CONNECT-AI Research Center, Yonsei University College of Medicine, Seoul, 03764, South Korea.
E-mail address: hyt0205@gmail.com (Y. Hong).

https://doi.org/10.1016/j.compbiomed.2023.106931
Received 24 June 2022; Received in revised form 4 April 2023; Accepted 13 April 2023
Available online 20 April 2023

have been evaluated using PSNR, SSIM, and noise level [7,11–24]. Prior works [7,12,15,16,22,24] have shown superior performance with subjective analysis to that with objective analyses based on PSNR and SSIM. Since these metrics insufficiently represent texture details, subjective analysis is necessary. Qingsong et al. [7] evaluated the proposed CT denoising algorithm with subjective analysis regarding the noise suppression, artifact reduction, and overall quality. It is important to note that artifact reduction and the overall quality scores of their method were superior to the compared methods, but the PSNR and SSIM were not. Chenyu et al. [16] also demonstrated the performance superiority of the CycleGAN-based CT super-resolution algorithm with a subjective analysis based on the five aspects of image sharpness, image noise, contrast resolution, diagnostic acceptance, and overall quality. Their method scored higher points than the compared methods except for the image noise score, even though the PSNR and SSIM of the proposed method were not superior, which means a discrepancy between the subjective and objective evaluations.

Radiomics is an emerging image analysis technique used to extract sets of multiple features from radiographic images [25,26]. It encodes pixel relationships in the ROI (e.g., tumor) using a variety of matrices. Radiomics is increasingly used in diagnosis, treatment planning, and outcome prediction [25,26]. However, lack of feature reproducibility is its major limitation [27], mainly because radiomic features depend on imaging protocols. In prior works [28,29], reproducibility analysis was performed to explore robust features that do not rely on the imaging protocols.

In this study, we introduced radiomic feature reproducibility analysis as an assessment metric for post-processed images. The reproducibility of radiomics feature represents how similar the feature from the post-processed image is to that of the target image. Thus, we can evaluate the similarity of a texture feature of the denoised CT image to the that of target image. We designed a GAN-based CT image denoising algorithm and tuned the hyperparameter, considering radiomic feature reproducibility analysis results. The architecture of the denoising model was inspired by the Enhanced Super-Resolution Generative Adversarial Network (ESRGAN) [30]. We replaced the residual scaling parameters of ESRGAN with spatial and channel attention modules to improve the model performance. Furthermore, we added a dropout layer to the generator network and experimentally defined the dropout rate. Since the PSNR and SSIM are not sensitive to present detailed texture differences [31], they are not suitable for selecting the optimal dropout rate. We demonstrated that the radiomic feature reproducibility analysis enables efficient tuning of the dropout rate and a well-tuned dropout rate improves denoising performances. The overall workflow is shown in Fig. 1

In summary, objective texture evaluation is critical for developing CT denoising methods, but existing image quality analysis methods are inadequate for evaluating these algorithms. Our experiments showed that the proposed method is more effective for evaluating and configuring efficient CT denoising methods.

The remainder of this paper is organized as follows. Section 2 discusses related works on GAN for CT denoising and radiomic feature reproducibility analysis; Section 3 presents the experimental design; and Section 4 shows the experimental results. Finally, relevant issues and future research plans are discussed in Section 5.

The main contributions of this study are summarized as follows:

- The radiomic feature reproducibility analysis enables evaluation of texture details to overcome the shortcomings of traditional image evaluation metrics such as PSNR and SSIM for CT image denoising algorithms.
- The proposed model tuned with radiomic feature reproducibility analysis outperformed other well-known CT denoising methods

2. Related work

2.1. GAN based CT denoising studies

As research on medical image processing using GAN has recently increased, several GAN-based models have been proposed for the denoising problem of CT images. Pixel-wisely paired CT images are necessary for the denoising studies to translate texture patterns without structural changes. Thus, most CT denoising studies are performed using 2016 NIH-AAPM-Mayo Clinic Low Dose CT Grand Challenge (AAPM-Mayo) datasets comprising standard-dose CT images and pixel-wise corresponding quarter-dose CT images. Xin Yi et al. [14] conducted a CT denoising study using a conditional GAN with a sharpness detection network that had a constraint function to measure the sharpness of the generated CT image. They successfully demonstrated the effectiveness of the sharpness loss function, but the evaluation metrics were not enough to explicitly tell the differences. Wasserstein GAN with a gradient penalty and perceptual loss (WGAN-VGG) [7], which employed a pretrained convolutional network named as VGG, was used to measure the perceptual difference. VGG encodes images into low-dimensional representations using pretrained convolutional kernels, enabling comparison of perceptual features. The concept was excellent on improving denoising performance, but the traditional metrics could not evaluate perceptual quality properly. They also performed subjective analysis to support their denoising results. You et al. [12] proposed a structurally sensitive multi-scale GAN (SMGAN) to capture subtle structural features while maintaining visual sensitivity by employing SSIM as a perceptual constraint. The quantitative denoising performance of SMGAN was not superior to those of other methods, and a subjective analysis in terms of sharpness, noise suppression, diagnostic acceptability, contrast retention, and overall quality was performed by three radiologist readers. However, the noise suppression did not outperform the L2 objective-oriented convolutional neural network (CNN) method. This result shows that the traditional metrics are insufficient to evaluate denoising performance appropriately. Ma et al. [20] proposed a least squares GAN with SSIM and L1 objective functions. However, the translated image was still blurry, and the structure did not match well the input low-dose noisy CT. They calculated uniformity and entropy to evaluate the texture statistics. They employed the texture evaluation metrics; however, these were insufficient for proper comparison of denoising performances. Objective texture evaluation is important in developing CT denoising methods.

2.2. Radiomic feature reproducibility analysis

Machine learning has facilitated use of radiomics in diagnosing and predicting diseases using medical images [32–34]. Since radiomics is greatly dependent on imaging protocols, segmentation rule, and scan-rescan robustness, all the imaging parameters and analysis protocols need to be defined in advance of the radiomics study [27]. For this reason, radiomic features in CT may not be reproducible unless the imaging settings are pre-defined [35]. The lack of reproducibility can be mitigated using deep learning-based post-processing algorithms [36]. Recently, Lee et al. [37] proposed that deep learning-based image conversion methods improve the reproducibility of CT radiomic features. They converted CT images acquired with eight protocols to a reference protocol CT image and evaluated the reproducibility of radiomic features of the converted images. Choe et al. [38] investigated the effect of different reconstruction kernels on radiomic features, and the reproducibility of the radiomic feature was improved by using the CNN. In addition, some studies [39–41] reported improvements in radiomics feature reproducibility on well-processed CT images with postprocessing algorithms, especially generative adversarial networks. In this study, we propose to employ radiomic feature reproducibility analysis as an evaluation metric for CT image denoising performance. Furthermore, we employ radiomics feature reproducibility analysis for the network architecture configuration and hyperparameter tuning (dropout rate). To the best of our knowledge, this is the first attempt in this direction.

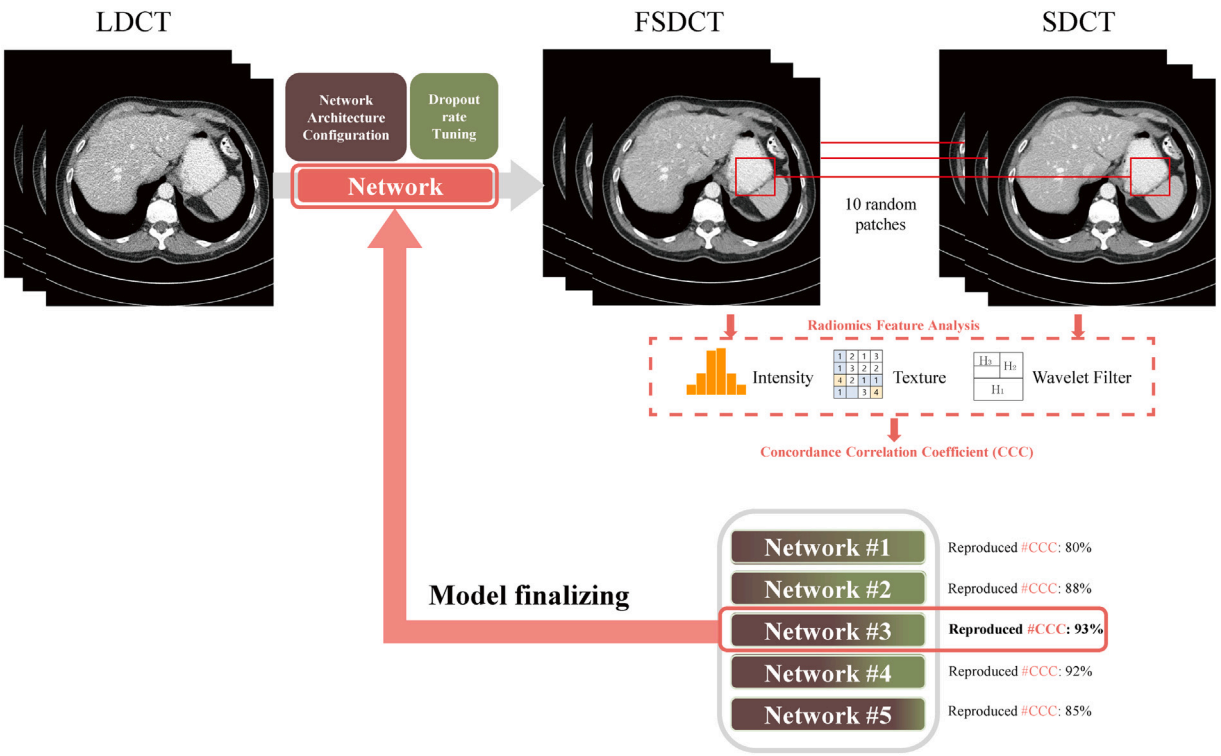

Fig. 1. Flow diagram of the overall workflow.

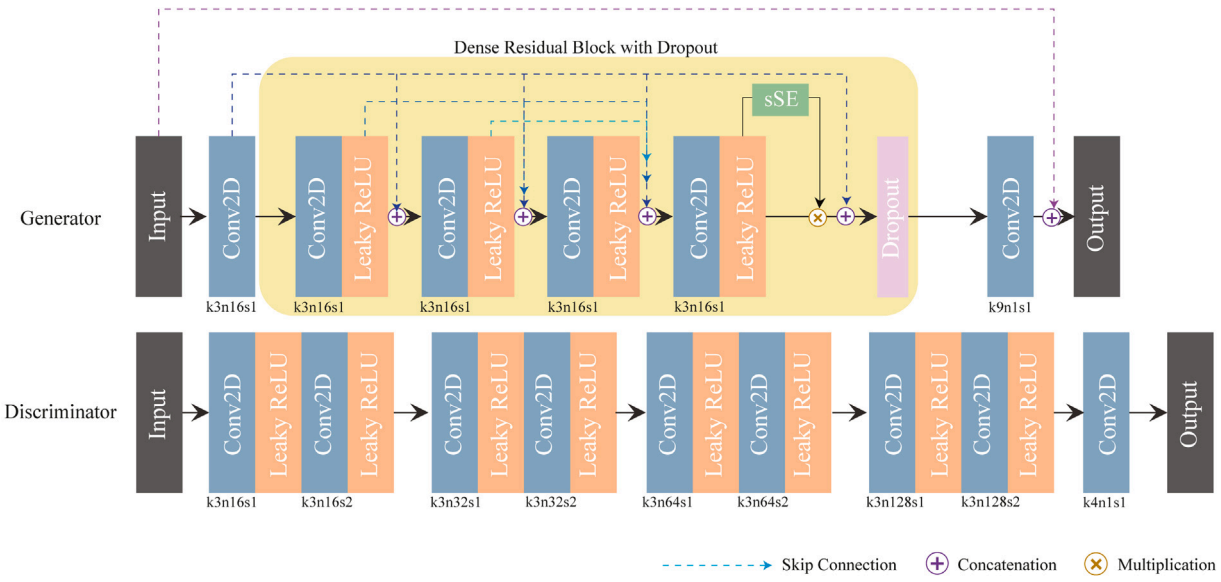

Fig. 2. Overall architecture of the proposed network. k stands for the size of the convolutional kernel, n stands for the number of the kernel, and s stands for the size of the convolutional stride. sSE stands for channel squeeze and spatial excitation module.

3. Proposed method

3.1. Problem statement for CT denoising

Assuming that $x \in \mathbb{R}^{H\times W\times D}$ denotes low-dose CT (LDCT) and $y \in \mathbb{R}^{H\times W\times D}$ denotes the corresponding standard-dose CT (SDCT), the relationship between the two can be defined as follows:

$$y = T(x), \tag{1}$$

where T is a translation function that converts x to y. H, W, D denotes height, width and depth respectively. In general, the noise distribution of CT images is modeled as a complex synthesis of Poisson quantum noise and Gaussian electronic noise [42]. Since the relationship T between LDCT and SDCT could not be accurately established, previous traditional denoising methods [43,44] had limited performances for CT denoising. Our network is trained to solve the inverse problem to produce the image $\tilde{x}$. It can be expressed as:

$$T^{-1}(y) = \tilde{x} \approx x, \tag{2}$$

while the network learns high-dimensional features using a non-linear function. Consequently, it is possible to generate a denoised image which is close to x.

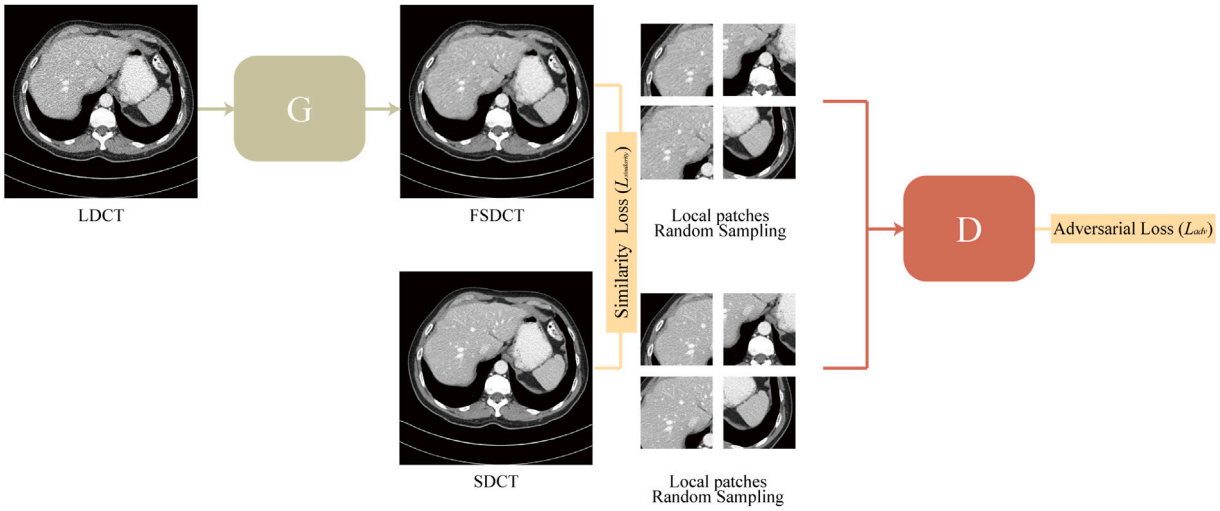

Fig. 3. Summary of the proposed network framework. G denotes the generator network, and D denotes the discriminator network. LDCT is low-dose computed tomography, SDCT is standard-dose computed tomography, and FSDCT is fake SDCT.

3.2. Network architecture

In previous studies [7,14,24], the generator network was a fully convolutional architecture; therefore, convolutional kernels could be trained with images of an arbitrary size. They used sampled patches to train the convolutional kernels in the generator network. However, we trained the convolutional kernels with images of 512 × 512 pixels originally to generate the output images of same size. The main reason for this approach is to measure perceptual differences between standard-dose CT (SDCT) and fake SDCT (FSDCT) over the original dimension, not over the locally sampled dimension. Meanwhile, the discriminator network was trained with randomly sampled local patches to distinguish the local textures of SDCT and FSDCT. The overall network architecture is depicted in Fig. 2 (see Fig. 3).

3.2.1. Discriminator network

The discriminator network D was trained to distinguish local patches of SDCT and those of FSDCT. We randomly sampled 64 × 64 local patches five times from SDCT and at the same location from FSDCT. D had four convolutional blocks and one convolutional layer. In the convolutional block, the first layer had 3 × 3 convolutional kernels with one stride and the second layer had a 3 × 3 convolutional kernel with two strides to reduce the dimension. These two convolutional layers had zero padding and were activated with a leaky ReLU [45] activation. This activation had a negative slope coefficient of 0.2. The number of convolutional feature maps was doubled through the next convolutional block. The last convolutional layer had a 4 × 4 convolutional kernel with one stride, and it had a single probability output.

3.2.2. Generator network

The generator network G took a 512 × 512 LDCT image as the input and produced the same size of FSDCT as the output. The input LDCT images were encoded to 32 feature maps through 3 × 3 convolutional kernels with a leaky ReLU [45] activation. The encoded feature map was fed to a dense residual block with dropout (DRBD). DRBD consisted of three consecutive convolution layers with a leaky ReLU activation and was connected by a residual operation. All leaky ReLU activations had 0.3 of the negative slope coefficients. A dropout operation [46] was applied after the block with a dropout rate of 0.5. We tested the impact of the dropout layer and explored for the optimal dropout rate. The experimental results could show the optimal dropout rate after the radiomic feature reproducibility analysis.

In ESRGAN, the residual scaling parameter was used to prevent unstable training caused by the large scale of residual features. We replaced this scaling factor with the attention module to apply a trainable discounting scale parameter. We employed Squeeze-and-Excitation (SE) module [47] and Channel Squeeze and Spatial Excitation (sSE) module [48] as an attention module. The SE module enhances the feature maps with a squeeze operation that summarizes the entire information about the feature map and an excitation operation that scales the importance of each feature map through this. In the sSE module, we first passed the input through a 1 × 1 convolutional layer with a sigmoid activation to obtain squeezed information. The weights and biases of this layer were initialized using Xavier uniform initialization [49]. We then multiplied the squeezed information with the original input to obtain a spatially re-calibrated feature map.

For all convolutional layers, except the last one, we used 3 × 3 convolutional kernels, and initialized their weights and biases using He normal initialization [50]. The last convolutional layer had a kernel size of 9 × 9, which facilitated faster learning with hyperbolic tangent activation [51]. We initialized the weights and biases of the last layer using Xavier uniform initialization. To preserve the dimension of the feature map, we applied zero-padding to all the convolutional layers.

3.3. Objective functions for noise reduction

3.3.1. Adversarial loss

We employed WGAN objective losses to translate the given LDCT to SDCT using the WGAN-GP framework [52]. The discriminator loss was defined as:

$$L_D = -\mathbb{E}[D(y)] + \mathbb{E}[D(G(x))] + \mathbb{E}[(\| \nabla_{\hat{x}} D(\hat{x}) \|_2 - 1)^2], \quad (3)$$

where $\mathbb{E}(\cdot)$ stands for the expected value operator, y denotes the given SDCT, x denotes the given LDCT, and $\hat{x}$ is a randomly sampled 2D Gaussian point on the straight line between y and $G(x)$. The first two terms are the discriminator loss using Wasserstein distance, and the last term is the gradient penalty. According to [52], the adversarial loss for the generator is defined as:

$$L_{adv} = -\mathbb{E}[D(G(x))] \quad (4)$$

3.3.2. Similarity loss

We adopted a multi-scale structural similarity index measure (MS-SSIM), for more flexibility in analysis [53] as a similarity loss to

Table 1
Objective evaluation results (mean ± standard deviation) on the test and cross-validation sets. **Red** and **Blue** indicate the best and second-best performance, respectively.

Model	Test set		Cross-Validation		Model parameters
	PSNR (dB)	SSIM	PSNR (dB)	SSIM	
SDCT	–	1.000±0.000	–	1.000±0.000	–
LDCT	27.19±1.604*	0.948±0.019*	27.595±1.041	0.946±0.012	–
BM3D	28.961±0.906*	0.949±0.016*	30.474±1.042	0.962±0.012	–
RED-CNN	30.771±1.279*	0.965±0.013*	31.156±0.741	0.966±0.008	1,848,865
AAPM Net	29.076±0.738*	0.960±0.011*	29.600±0.628	0.960±0.008	4,166,272
Framelet	30.937±1.479*	0.966±0.014	31.399±0.971	0.966±0.010	4,166,272
WGAN-VGG	27.479±0.911*	0.955±0.016*	28.530±0.345	0.960±0.005	7,756,002
Proposed	30.341±1.604	0.966±0.019	30.049±1.519	0.966±0.009	331,379

*Indicated statistical significance ($p < 0.05$).

obtain visually more perceptible images than a regular SSIM [10]. The MS-SSIM was formulated as:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{5}$$

$$MS\_SSIM(x,y) = \prod_{j=1}^{M} SSIM(x_j, y_j), \tag{6}$$

where C_1, C_2 are constant; μ_x, μ_y, σ_x, σ_y, σ_{xy} denote means, standard deviations, and cross-covariance of the given image pair (x, y); x_j, y_j are the given local image content at the jth level, and M is the number of scale levels. The MS-SSIM values are in the range of $[0, 1]$. The closer the source x is to the target y, the closer the MS-SSIM value is to 1. The optimizer acts to minimize the cost; therefore, the dissimilarity by MS-SSIM is the cost for the optimization, and it is expressed as follows:

$$L_{similarity} = 1 - MS\_SSIM(x,y). \tag{7}$$

3.3.3. Generator loss

The generator network was trained by combining the generator adversarial loss L_{adv} and similarity loss $L_{similarity}$. In summary, the overall generator loss was defined as:

$$L_G = L_{adv} + \alpha L_{similarity}, \tag{8}$$

where α is the coefficient of the similarity loss term. We experimentally set α to 1.0 on the generative loss function.

3.4. Radiomics feature reproducibility analysis

We utilized an open-source python package, Pyradiomics [54], to extract the radiomic features. The first step of radiomic feature extraction is the discretization of gray values. We used 25 as the bin-width value for features from the original intensity and 10 as the bin-width value for features from wavelet signals, for discretization from the given local patch. The size of the local patch was 64 × 64, and ten patches were randomly sampled from the given FSDCT and SDCT. We extracted 836 radiomic features, subdivided into first-order statistics, texture features, and wavelet features. The details of employed radiomic features are as follows: 18 first-order statistics features, 24 gray-level co-occurrence matrix (GLCM), 16 gray-level run-length matrix (GLRLM), 16 gray-level size zone matrix (GLSZM), 14 gray-level dependence matrix (GLDM), and 5 neighboring gray-tone difference matrix (NGTDM). The other 743 features were obtained from 8 of the wavelet decomposition derived images. For the reproducibility analysis, we calculated concordance correlation coefficients (CCCs), defined by Lin [55] as a measure of the similarity of a data set Y is to a "golden standard" X. We calculated the CCCs of the extracted 836 radiomic features from the patches of the denoised FSDCT image against target SDCT images. We considered a radiomic feature with CCC ≥ 0.85 as a significantly reproduced feature [56]. Following a previous study [38], we counted the significantly reproduced features (CCC ≥ 0.85) to present the denoising results effectively (see Fig. 8).

4. Experiments

4.1. Data sources

The dataset is "2016 NIH-AAPM-Mayo Clinic Low Dose CT Grand Challenge", containing fully anonymized patient data approved by Mayo Clinic. It consisted of contrast-enhanced abdominal CT images with a thickness of 3 mm and a 2-D size of 512 × 512. The quarter dose CT was simulated with a noise level corresponding to 25% of the full dose by adding Poisson noise before image reconstruction. In this study, we used 2254 CT slices as training data and 224 CT slices as test data.

4.2. Network training

In our experiments, according to the WGAN-GP, we trained the generator and discriminator in a ratio of 5:1. We employed the adam optimization algorithm [57] for our network training with a learning rate = 1×10^{-4}, $\beta_1 = 0$, and $\beta_2 = 0.9$. We trained the proposed network using a batch size of 3 for 2000 epochs. We implemented the network in TensorFlow 2.0.0 using an NVIDIA GeForce RTX 2080 Ti GPU.

4.3. Quantitative evaluation of denoising performance

For the denoising performance comparison, we employed one filtering-based method and four deep learning models. The block matching 3D (BM3D) algorithm [43] is a representative filtering-based method that has shown good denoising performance in medical imaging fields [18,44,58]. Residual encoder–decoder CNN (RED-CNN) [12] is a deep learning model that incorporates an auto encoder–decoder network architecture. Deep network in wavelet domain (AAPM Net) [11] suppresses CT-specific noise using the directional wavelet transform coefficient of an image. Deep Convolutional Framelet WavResNet (Framelet) [17] is an extension of AAPM Net. The model measures residual wavelet coefficients for all sub-band decompositions. WGAN-VGG [7] is a Wasserstein GAN-based model with a perceptual loss. We performed a qualitative comparison using representative abdominal CT images from the test data (Figs. 4, 6).

We conducted hyperparameter (dropout rate) tuning and ablation studies with a test set from a single patient to optimize the proposed model. To ensure the reliability of the results, we performed patient-level cross-validation with ten patients. We measured the PSNR and SSIM of the FSDCT. Furthermore, we employed multiscale SSIM to obtain a more accurate evaluation. Table 1 summarizes all the objective evaluation results. The PSNR of the proposed model was 30.341 dB, showing a difference of 0.596 dB against the Framelet, which had the best PSNR with 30.937 dB. The SSIM of the proposed model showed a mean value of 0.966, which is equivalent to that of the best-performing models. The results from the cross-validation were similar to those from the test set. In addition, the proposed model's inspected number of model parameters was 331,379, which is significantly lower than that of other models. For a more detailed insight into the results obtained

Table 2
Statistical properties of the ROI in Figs. 4 and 6.

Model	Fig. 4		Fig. 6	
	Mean	SD	Mean	SD
SDCT	102.847	164.666	125.667	139.928
LDCT	102.129	169.141	123.798	143.137
BM3D	101.994	161.767	123.790	136.075
RED-CNN	102.399	161.986	123.666	134.096
AAPM Net	101.876	158.104	123.986	131.759
Framelet	100.465	163.068	122.493	137.212
WGAN-VGG	98.276	165.154	120.156	137.645
Proposed	**102.327**	**165.029**	**124.977**	**138.339**

Table 3
Denoising performance evaluation with radiomic feature reproducibility analysis.

Model	Intensity	Wavelet	Total	Reproducibility
LDCT	22	135	157	18.7%
BM3D	27	177	204	24.4%
RED-CNN	27	178	205	24.5%
AAPM Net	18	140	158	18.9%
Framelet	32	170	202	24.1%
WGAN-VGG	52	230	282	33.7%
Proposed	**67**	**313**	**380**	**45.4%**
SDCT	92	744	836	100.0%

Table 4
Radiomic feature reproducibility analysis for the dropout rate tuning.

Dropout rate	Intensity	Wavelet	Total	Reproducibility
0.0	56	251	307	36.7%
0.25	52	296	348	42.4%
0.50 (Proposed)	**67**	**313**	**380**	**45.4%**
0.75	42	268	311	37.2%

by different methods, we inspected the statistical properties (mean and standard deviations of CT intensity) from the selected regions as shown in Figs. 4 and 6 (Table 2).

Although the proposed model achieved excellent overall performance, the performances were not distinct from the compared methods. The zoomed view of the ROIs in Figs. 4 and 6 are presented in Figs. 5 and 7. From the zoomed views, we recognize that the objective evaluation results did not match well with the visual perception of denoised images. This result indicates that the traditional objective metrics are not sensitive to detect detailed perceptual differences. In addition, we presented the absolute difference maps to visualize the denoising results in Fig. 9. The difference maps showed that the comparison methods reduce the noise heterogeneously or at the out of the field of organ.

4.4. Results of radiomic feature reproducibility analysis

BM3D showed the best PSNR (Table 1). The number of significantly reproduced radiomic features using BM3D was 204 (24.4%). However, the significantly reproduced radiomic features using the proposed method were 380 (45.4%). All the SSIM results were not distinct but the number of significantly reproduced radiomic features show that the proposed method outperformed the other methods. The number of significantly reproduced radiomic features is summarized in Table 3. The radiomic feature reproducibility analysis results are presented in Fig. 10. Fig. 10-(a) shows that the radiomic features of the LDCT are not well correlated with those of SDCT. However, Fig. 10-(h) is self-comparison and shows a complete correlation with SDCT. The feature reproducibility of LDCT Fig. 10-(a) can be regarded as a baseline. The proposed method showed the best feature reproducibility even though some radiomic features did not meet the criteria (CCC $\geq$ 0.85).

4.5. Dropout rate tuning

Hinton [59] suggested preventing overfitting in neural networks, which requires a grid-search over the dropout probabilities [60]. However, it is difficult to tune the dropout rate because denoising performance evaluation using SSIM and PSNR are often indistinguishable across different models. We implemented a detailed search space as a grid of dropout rates and found out that the best dropout rate was 0.50 in our experiments. We showed that the radiomic feature reproducibility analysis allows us to tune the dropout rate. In Table 4, setting the dropout rate to 0.25, 0.50, and 0.75 revealed performance differences clearly. We selected and used a dropout rate of 0.50, which showed the best performance in our experimental setting. We evaluated the effect of dropout with PSNR, SSIM, mean, SD, and radiomic feature reproducibility analysis. The evaluation results are summarized in Table 5. The effect of dropout is not distinct with PSNR, SSIM, mean, and SD, but the number of significantly reproduced radiomic features is distinct (without dropout vs. with dropout: 36.7% vs. 45.4%). We can confirm the improvement of denoising performance with the calibrated dropout regularization technique.

4.6. Ablation study on model architecture

We performed ablation studies to demonstrate the effectiveness of the network architecture design. We tested four model architectures: one DRDB with attention, one DRDB without attention, two DRDBs with attention, and three DRDBs with attention. The experiments were performed with fixed seeds to have deterministic results. We compared the performances by measuring PSNR and SSIM for all slices in the test patient set and radiomic feature reproducibility analysis for the same patches in Fig. 10. The one DRDB with attention module had 45.4% of significantly reproduced radiomic features while one DRDB without attention had 41.7% of significantly reproduced radiomic features. When the effect of the attention module was compared using PSNR, SSIM, mean, and SD, the effect was not distinct (with attention vs. without attention, PSNR: 30.082 vs. 29.932, SSIM: 0.961 vs. 0.960, mean: 124.977 vs. 125.703, SD: 138.339 vs. 137.465). This result indicates that the attention module contributes to better denoising and texture feature recovery. The two and three DRDBs with attention had 20.1% and 22.6% of significantly reproduced radiomic features, respectively. The two deeper models showed a decrease not only in quantitative and statistical measures but also in a number of significantly reproduced radiomic features. The ablation study results are summarized in Table 6.

4.7. Qualitative analysis

The visual assessment was performed in a blind manner by two invited radiologists: H.J.B. with 17 years of experience and E.C. with 12 years of experience. The radiologists evaluated the image sharpness, noise suppression, structure preservation, and overall quality using a five-point scale, 1: non-diagnostics, 2: bad, 3: acceptable, 4: good, and 5: excellent. We provided the test cases that were processed by different denoising methods. The scores are presented as the mean (average score of two radiologists) ±standard deviation (SD) and are summarized in Table 7. The results in Table 7 show that our proposed method and LDCT obtained the highest average sharpness score of 3.5 from the two radiologists. Our proposed method achieved the highest average score of 3.5 for noise suppression. LDCT obtained the highest score for structure preservation among the compared methods. However, our proposed method obtained good scores from both radiologists for structure preservation. The proposed method and LDCT also obtained the highest overall quality scores among the compared methods.

These results indicate that the proposed method performs well in terms of sharpness, noise suppression, and overall quality. Although LDCT was evaluated slightly better in the structure preservation category, the proposed method still scored well in this aspect.

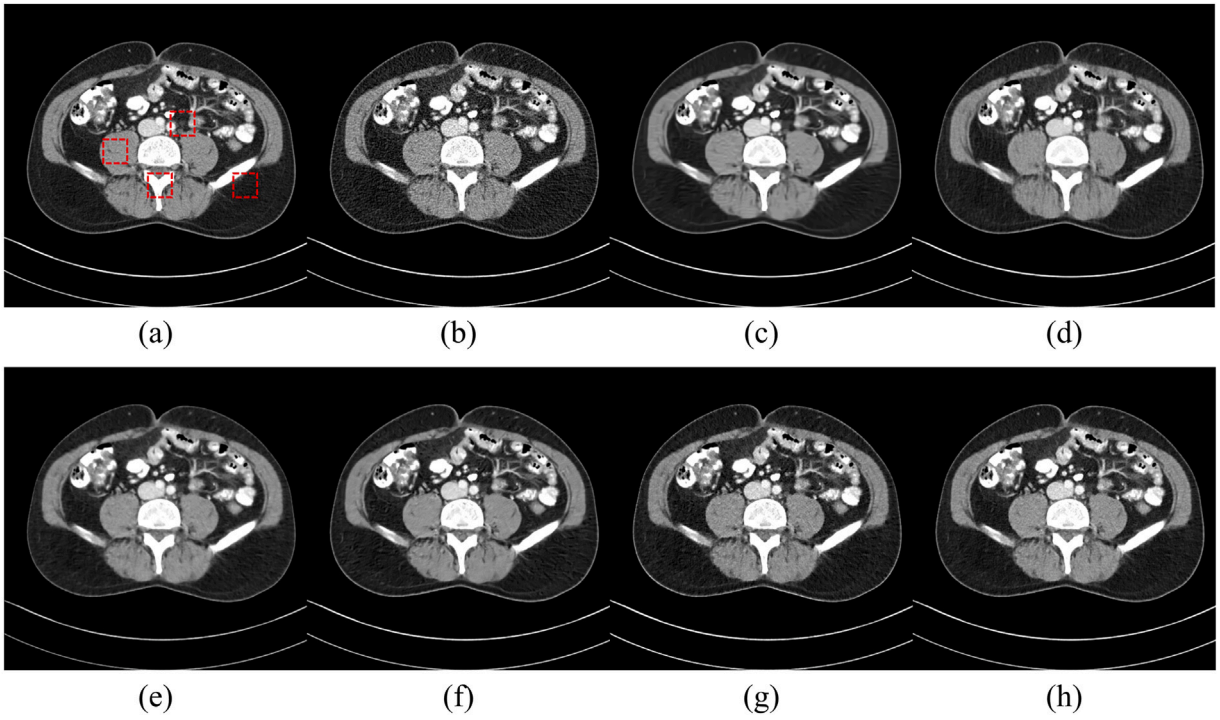

Fig. 4. Results from the test data sets: (a) SDCT; (b) LDCT; (c) BM3D; (d) RED-CNN; (e) AAPM Net; (f) Framelet; (g) WGAN-VGG; and (h) Proposed. The display window is [−160, 240] HU.

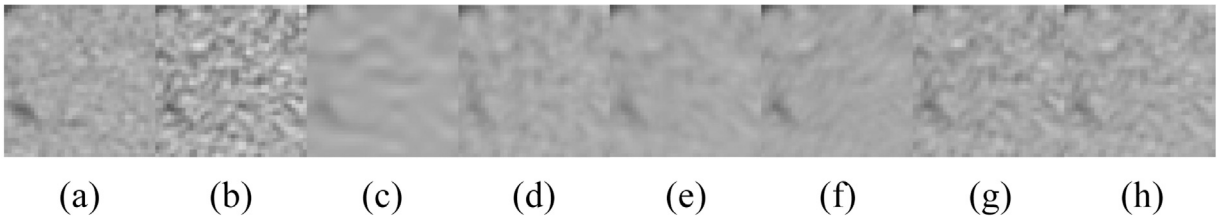

Fig. 5. The leftmost zoomed ROI in Fig. 4 for comparison: (a) SDCT; (b) LDCT; (c) BM3D; (d) RED-CNN; (e) AAPM Net; (f) Framelet; (g) WGAN-VGG; and (h) Proposed. The display window is [−160, 240] HU.

Table 5
Significance of dropout layer with quantitative, statistical and radiomic feature reproducibility analysis in Fig. 4.

Model	Quantitative analysis		Statistical analysis		Radiomic feature reproducibility analysis			
	PSNR	SSIM	Mean	SD	Intensity	Wavelet	Total	Reproducibility
LDCT	26.715	0.939	123.798	143.137	22	135	157	18.7%
Proposed w/o dropout	29.981	0.961	124.766	137.697	56	251	307	36.7%
Proposed w dropout	**30.082**	**0.961**	**124.977**	**138.339**	**67**	**313**	**380**	**45.4%**
SDCT (reference)	–	–	125.667	139.928	92	744	836	100.0%

Table 6
Ablation study for the model architecture configuration.

Module		Quantitative analysis		Statistical analysis		Radiomic feature reproducibility analysis			
DRBD	Attention	PSNR	SSIM	Mean	SD	Intensity	Wavelet	Total	Reproducibility
1	O	30.082	**0.961**	124.977	**138.339**	**67**	**313**	**380**	**45.4%**
1	X	29.932	0.960	**125.703**	137.465	64	285	349	41.7%
2	O	**30.374**	0.961	121.789	133.817	23	145	168	20.1%
3	O	27.168	0.927	128.163	133.728	22	167	189	22.6%
SDCT (reference)		–	–	125.667	139.928	92	744	836	100.0%

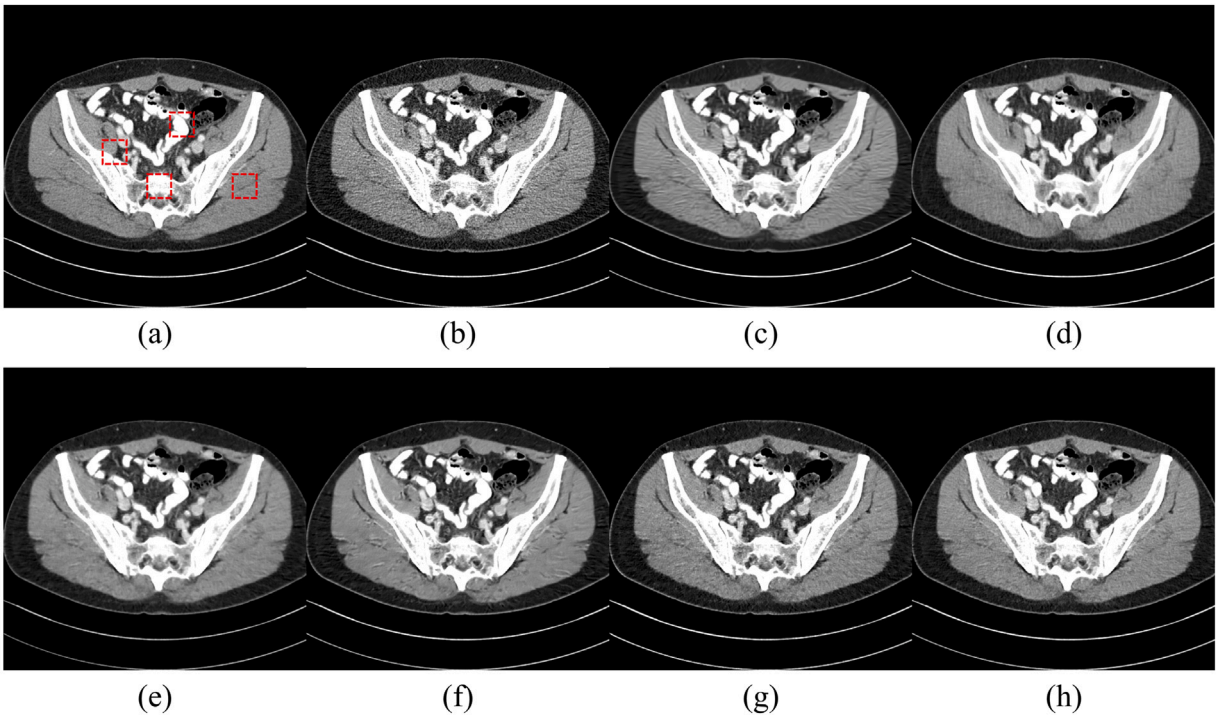

Fig. 6. Results from the test data sets: (a) SDCT; (b) LDCT; (c) BM3D; (d) RED-CNN; (e) AAPM Net; (f) Framelet; (g) WGAN-VGG; and (h) Proposed. The display window is [−160, 240] HU.

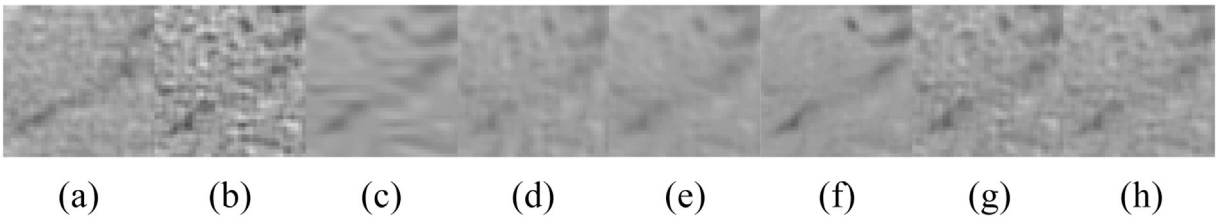

Fig. 7. The rightmost zoomed ROI in Fig. 6 for comparison: (a) SDCT; (b) LDCT; (c) BM3D; (d) RED-CNN; (e) AAPM Net; (f) Framelet; (g) WGAN-VGG; and (h) Proposed. The display window is [−160, 240] HU.

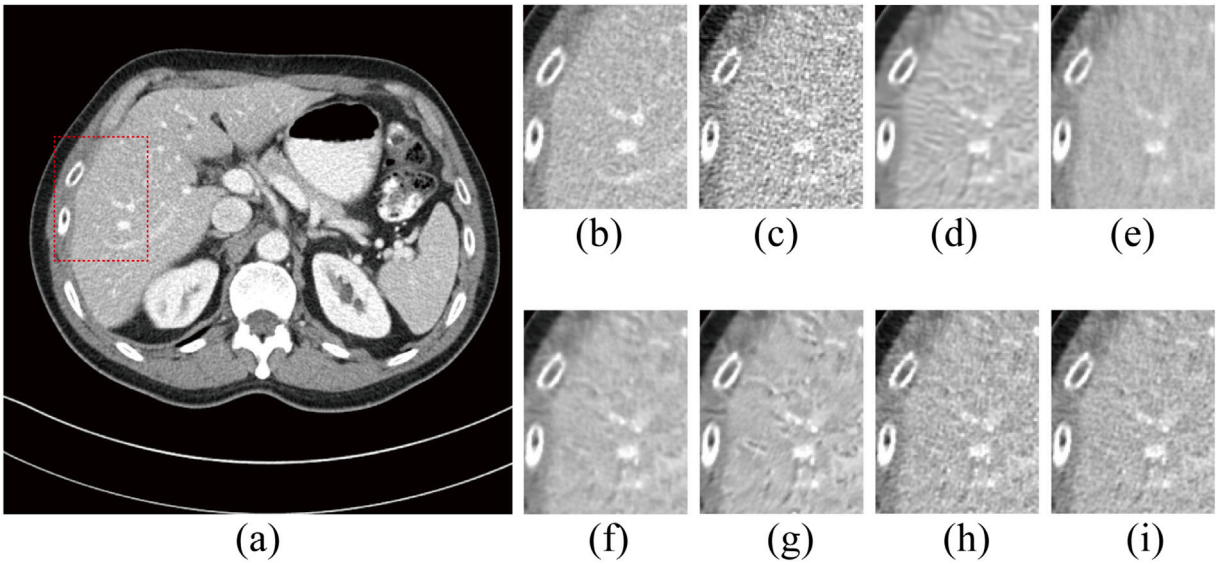

Fig. 8. Comparison of the denoising results of the vessel region from the test data slide (a) SDCT whole image and ROI; (b) SDCT; (c) LDCT; (d) BM3D; (e) RED-CNN; (f) AAPM Net; (g) Framelet; (h) WGAN-VGG; and (h) Proposed. The display window is [−160, 240] HU.

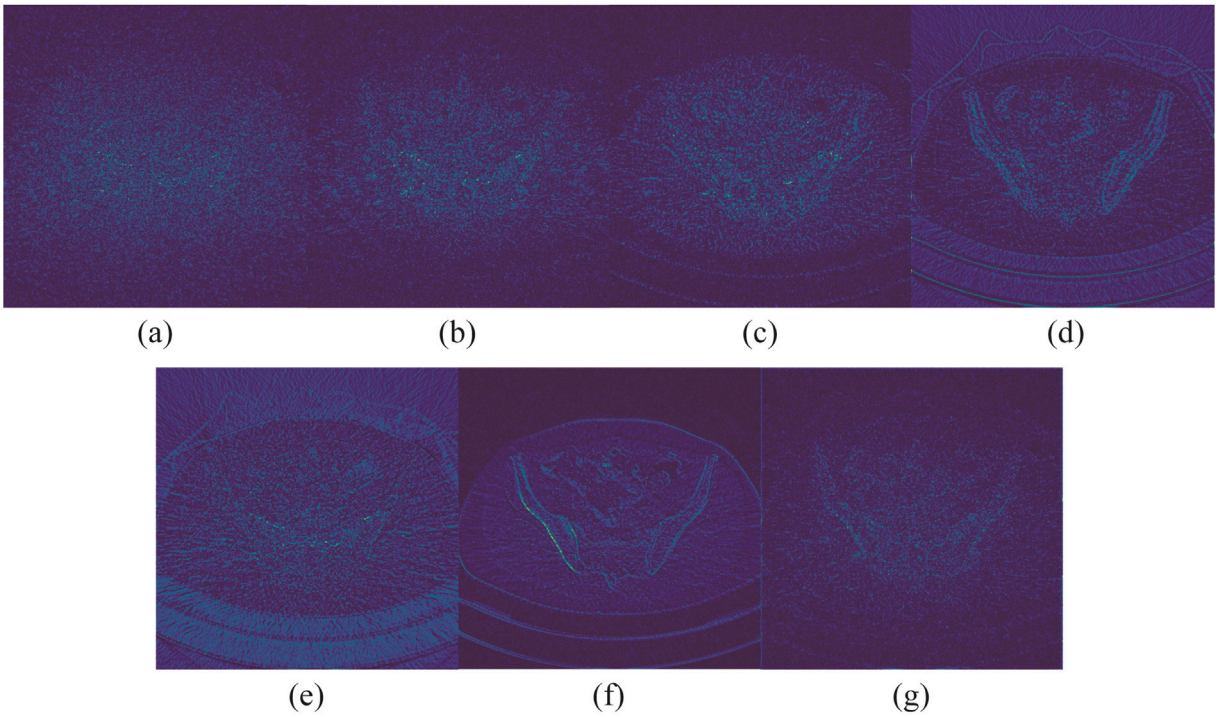

Fig. 9. Absolute difference images of Fig. 6: (a) LDCT; (b) BM3D; (c) RED-CNN; (d) AAPM Net; (e) Framelet; (f) WGAN-VGG; and (g) Proposed.

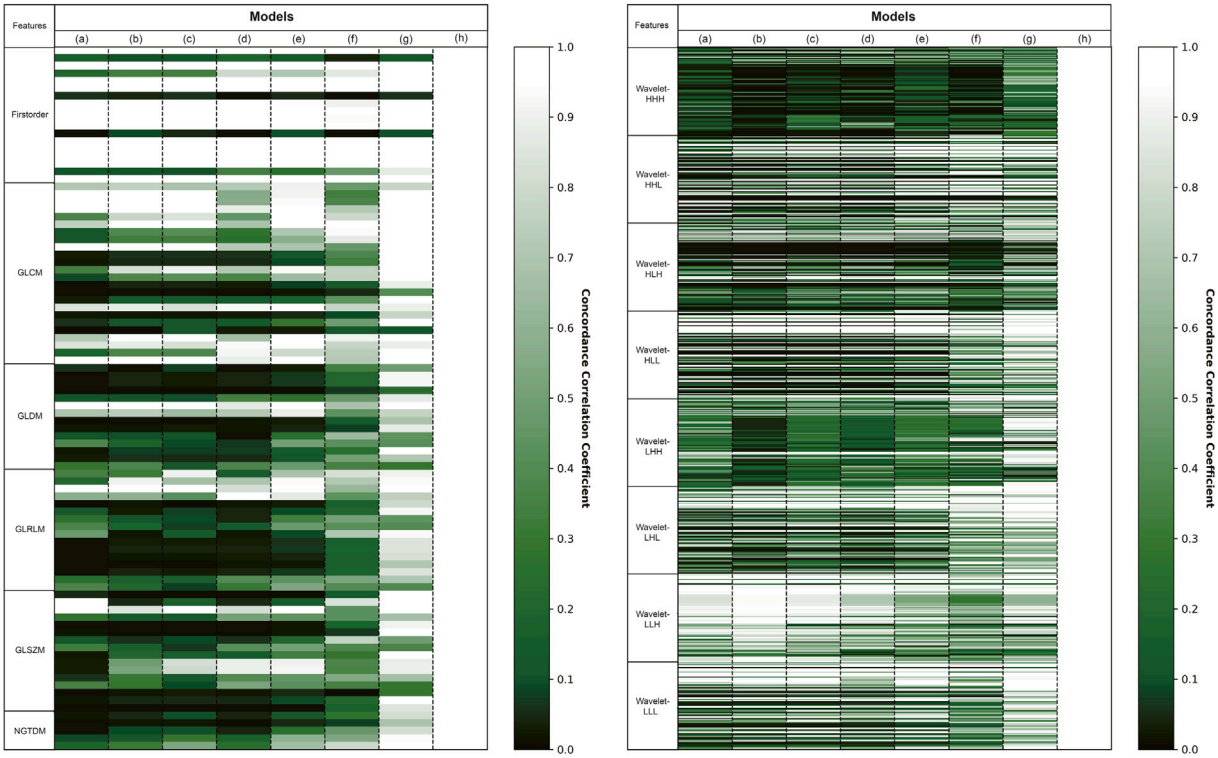

Fig. 10. Results of radiomic feature reproducibility analysis with concordance correlation coefficient on Fig. 4: (a) LDCT; (b) BM3D; (c) RED-CNN; (d) AAPM Net; (e) Framelet; (f) WGAN-VGG; (g) Proposed; and (h) SDCT.

Table 7
Visual assessment scores by two radiologist readers.

Model	Sharpness	Noise suppression	Structure preservation	Overall quality
SDCT	5.0 ± 0.00	5.0 ± 0.00	5.0 ± 0.00	5.0 ± 0.00
LDCT	**3.5 ± 0.71**	3.0 ± 1.41	**4.5 ± 0.71**	**3.5 ± 0.71**
BM3D	2.5 ± 0.71	1.0 ± 0.00	2.5 ± 0.71	1.5 ± 0.71
RED-CNN	2.5 ± 0.71	2.0 ± 0.00	3.0 ± 0.00	2.0 ± 0.00
AAPM Net	2.0 ± 0.00	2.0 ± 0.00	3.0 ± 0.00	2.0 ± 0.00
Framelet	2.5 ± 0.71	1.0 ± 0.00	2.0 ± 0.00	1.0 ± 0.00
WGAN-VGG	3.0 ± 0.00	2.5 ± 0.71	3.0 ± 0.00	3.0 ± 0.00
Proposed	**3.5 ± 0.71**	**3.5 ± 0.71**	4.0 ± 0.00	**3.5 ± 0.71**

5. Discussion

In this study, we proposed a GAN-based low-dose CT denoising method, which outperformed well-known CT denoising methods. It is difficult to determine the optimal hyperparameters of GAN with traditional image quality assessment metrics (Table 1); however, radiomic feature reproducibility analysis allows the sophisticated tuning of the hyper-parameters of GAN. This was possible since reproducibility analysis supports in-depth texture analysis of the CT image.

The traditional image quality assessment metrics such as PSNR and SSIM do not fully reflect perceptual difference since they are dependent on the per-pixel differences. For example, Fig. 4 shows a significant visual difference, but difference in evaluation metrics values is insignificant. However, radiomic feature reproducibility analysis distinctly shows the performance difference between the compared models (Table 3). The denoising performance evaluated using radiomic feature reproducibility analysis was distinct, and the proposed model explicitly showed the best denoising performance. The hypothesis that radiomics reproducibility analysis accurately reflects perceptual differences in image quality is supported by the qualitative analysis of radiologists in Table 7. Among the denoising methods, the proposed method rated the highest score in the overall quality category of visual assessment and showed the highest reproducibility in radiomics analysis. In contrast, RED-CNN and Framelet, which were highly rated in PSNR and SSIM, were evaluated as "unacceptable" by the radiologist. The results support that the traditional image quality evaluation methods may not fully capture perceptual differences. Furthermore, we explicitly showed that radiomic feature reproducibility analysis could allow accurate configuration of the network architecture and hyper-parameter setting. In addition, we performed patient-level cross-validation with ten patients. Table 1 presents the test set and cross-validation results. The difference between test set evaluation and cross-validation was 0.292 dB for PSNR, and there is no difference for SSIM. Thus, we experimentally confirmed that our model was optimized well.

We evaluated the effect of the dropout layer using radiomic feature reproducibility analysis. The proposed model showed similar denoising performance based on PSNR, SSIM, mean, and noise (SD) with and without dropout. It means that the traditional methods could not demonstrate the effect of the dropout. However, the results of reproducibility analysis were significantly different. The proposed model had reproducibility of 36.7% without the dropout layer and 45.4% with the dropout layer. In addition, we explored the optimal dropout rate for the proposed model. The results of the dropout rate tuning are shown in Table 4. The dropout of 0.50 was experimentally determined as optimal.

Recently, Gal and Ghahramani [61] showed that Monte-Carlo dropout (MC dropout) can be utilized as a Bayesian inference over a neural network's weights. This Bayesian inference can measure the uncertainty of the neural network [62]. Several works have used Bayesian uncertainty from MC dropout for denoising/image synthesis in CT. For example, [63] employed Bayesian inference to prevent overfitting in Deep Image Prior [64] and validate the estimated uncertainty using the uncertainty calibration error (UCE) metric. [65] used Bayesian epistemic uncertainty to integrate the physical model and deep learning synthetic images in the CT reconstruction method. Our result shows that radiomics reproducibility analysis enables the dropout rate tuning to improve denoising performance. In future, we plan to leverage the tuned dropout rate for MC dropout to obtain Bayesian uncertainty and to validate whether the estimated uncertainty can tackle the problem of hallucinations and artifacts in medical images [63].

The proposed network architecture was not based on the latest deep convolutional technique, but we demonstrated that the well-tuned GAN architecture outperformed well-known CT denoising methods. We explored the network architecture configuration with radiomic feature reproducibility analysis. We tested the effect of the attention module and the number of proposed DRBD modules. We adopted sSE attention modules to replace the residual scaling parameter of ESRGAN.

Table 7 shows that LDCT received generally favorable scores from the radiologists. This may be attributed to radiologists' prior familiarity with conventional CT reconstruction methods like filtered back-projection [66]. As a result, the radiologists may have been biased against artificial characteristics such as smoothing, which could have led to the undervaluation of the post-processed images, even if the objective analysis performance was high. To overcome this limitation and obtain a more reliable measure of denoising effectiveness, future research could use objective analysis methods that investigate the relationship between radiomics features and lesions in CT images. Such methods could provide an unbiased evaluation of denoising techniques, which may be particularly important given the potential for subjective bias in radiologist evaluations.

The proposed method may open the possibility of making an accurate diagnosis even with a low-dose CT image. Furthermore, radiomic feature reproducibility can be applied to any medical imaging modality such as magnetic resonance images, ultrasound images, and chest radiographs. Thus, our approach can be applied to any medical imaging processing algorithm development.

Although our proposed model demonstrated promising denoising performance, there are limitations to be addressed. First, all radiomic features were set to the same importance to evaluate the model performance. Radiomic features are prone to redundancy, so many radiomics studies try to screen redundant features and define a few important features. This study uses radiomic feature reproducibility to support in-depth image texture assessment, not to determine feature importance. Our future study will narrow down the number of radiomic features for the robust reproducibility analysis. Second, the radiomic feature reproducibility analysis cannot be employed as an objective function, because the texture matrix in the radiomic feature is not differentiable. For example, a gray-level zone matrix quantifies gray-level zones in an image. A gray-level zone is simply derived based on the number of connected voxels that share the same gray level intensity. We plan to develop differentiable texture matrices to employ radiomics analysis as an objective function in future work. Furthermore, we will validate the generalizability of radiomics analysis as an imaging performance evaluation method by performing the analysis on all other medical imaging tasks. Lastly, the denoising model should be trained on real clinical images to learn various noise patterns. Because it is challenging to create a simulated dataset with all the noise that actually occurs in real world, re-training or re-tuning for other noise properties is essential to using a model trained with the simulated paired datasets [7]. Recently, several denoising models have been released for unpaired datasets [67]. In future we aim to solve the denoising problem with an unpaired dataset obtained from actual clinical settings for various noise patterns.

6. Conclusion

In this study, we tuned hyperparameters of GAN with radiomic feature reproducibility analysis, and the tuned GAN-based denoising model outperformed well-known other CT denoising methods. Since the

texture assessment is an unmet need in medical image assessment, we believe that radiomic feature reproducibility analysis bridges the gap between the shortfalls of traditional objective evaluation and subjective evaluation. It may substitute subjective analysis on post-processed medical images.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, the Ministry of Food and Drug Safety) (Project Number: 202016B02)

References

[1] Hyun Woo Goo, CT radiation dose optimization and estimation: an update for radiologists, Korean J. Radiol. 13 (1) (2012) 1–11.
[2] Amy Berrington de Gonzalez, Sarah Darby, Risk of cancer from diagnostic X-rays: estimates for the UK and 14 other countries, Lancet 363 (9406) (2004) 345–351.
[3] Dominik Fleischmann, F. Edward Boas, Computed tomography—old ideas and new technology, Eur. Radiol. 21 (3) (2011) 510–517.
[4] Ge Wang, A perspective on deep imaging, Ieee Access 4 (2016) 8914–8924.
[5] Ge Wang, Jong Chu Ye, Klaus Mueller, Jeffrey A. Fessler, Image reconstruction is a new frontier of machine learning, IEEE Trans. Med. Imaging 37 (6) (2018) 1289–1296.
[6] Ian Goodfellow, et al., Generative adversarial nets, Adv. Neural Inf. Process. Syst. 27 (2014).
[7] Qingsong Yang, et al., Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss, IEEE Trans. Med. Imaging 37 (6) (2018) 1348–1357.
[8] Jelmer M. Wolterink, Tim Leiner, Max A. Viergever, Ivana Išgum, Generative adversarial networks for noise reduction in low-dose CT, IEEE Trans. Med. Imaging 36 (12) (2017) 2536–2545.
[9] Wenming Yang, Xuechen Zhang, Yapeng Tian, Wei Wang, Jing-Hao Xue, Qingmin Liao, Deep learning for single image super-resolution: A brief review, IEEE Trans. Multimed. 21 (12) (2019) 3106–3121.
[10] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, Eero P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[11] Eunhee Kang, Junhong Min, Jong Chul Ye, A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction, Med. Phys. 44 (10) (2017) e360–e375.
[12] Hu Chen, et al., Low-dose CT with a residual encoder-decoder convolutional neural network, IEEE Trans. Med. Imaging 36 (12) (2017) 2524–2535.
[13] Wei Yang, et al., Improving low-dose CT image using residual convolutional network, IEEE Access 5 (2017) 24698–24705.
[14] Xin Yi, Paul Babyn, Sharpness-aware low-dose CT denoising using conditional generative adversarial network, J. Digit. Imaging 31 (5) (2018) 655–669.
[15] Fenglei Fan, et al., Quadratic autoencoder (Q-AE) for low-dose CT denoising, IEEE Trans. Med. Imaging 39 (6) (2019) 2035–2050.
[16] Chenyu You, et al., CT super-resolution GAN constrained by the identical, residual, and cycle learning ensemble (GAN-CIRCLE), IEEE Trans. Med. Imaging 39 (1) (2019) 188–203.
[17] Eunhee Kang, Jong Chul Ye, et al., Wavelet domain residual network (WavResNet) for low-dose X-ray CT reconstruction, 2017, arXiv preprint arXiv:1703.01383.
[18] Dongwoo Kang, et al., Image denoising of low-radiation dose coronary CT angiography by an adaptive block-matching 3D algorithm, in: Medical Imaging 2013: Image Processing, Vol. 8669, International Society for Optics and Photonics, 2013, p. 86692G.
[19] Eunhee Kang, Won Chang, Jaejun Yoo, Jong Chul Ye, Deep convolutional framelet denosing for low-dose CT via wavelet residual network, IEEE Trans. Med. Imaging 37 (6) (2018) 1358–1369.
[20] Yinjin Ma, Biao Wei, Peng Feng, Peng He, Xiaodong Guo, Ge Wang, Low-dose CT image denoising using a generative adversarial network with a hybrid loss function for noise learning, IEEE Access 8 (2020) 67519–67529.
[21] Hu Chen, et al., Low-dose CT via convolutional neural network, Biomed. Opt. Express 8 (2) (2017) 679–694.
[22] Meng Li, William Hsu, Xiaodong Xie, Jason Cong, Wen Gao, SACNN: Self-attention convolutional neural network for low-dose CT denoising with self-supervised perceptual loss network, IEEE Trans. Med. Imaging 39 (7) (2020) 2289–2301.
[23] Zhixian Yin, Kewen Xia, Ziping He, Jiangnan Zhang, Sijie Wang, Baokai Zu, Unpaired image denoising via wasserstein GAN in low-dose CT image with multi-perceptual loss and fidelity loss, Symmetry 13 (1) (2021) 126.
[24] Chenyu You, et al., Structurally-sensitive multi-scale deep neural network for low-dose CT denoising, IEEE Access 6 (2018) 41839–41855.
[25] Stefania Rizzo, et al., Radiomics: the facts and the challenges of image analysis, Eur. Radiol. Exp. 2 (1) (2018) 1–8.
[26] Robert J. Gillies, Paul E. Kinahan, Hedvig Hricak, Radiomics: images are more than pictures, they are data, Radiology 278 (2) (2016) 563–577.
[27] Evangelos K. Oikonomou, Musib Siddique, Charalambos Antoniades, Artificial intelligence in medical imaging: a radiomic guide to precision phenotyping of cardiovascular disease, Cardiovasc. Res. 116 (13) (2020) 2040–2054.
[28] Rikiya Yamashita, et al., Radiomic feature reproducibility in contrast-enhanced CT of the pancreas is affected by variabilities in scan parameters and manual segmentation, Eur. Radiol. 30 (1) (2020) 195–205.
[29] Binsheng Zhao, et al., Reproducibility of radiomics for deciphering tumor phenotype with imaging, Sci. Rep. 6 (1) (2016) 1–7.
[30] Xintao Wang, et al., Esrgan: Enhanced super-resolution generative adversarial networks, in: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018.
[31] Christian Ledig, et al., Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4681–4690.
[32] Philippe Lambin, et al., Radiomics: the bridge between medical imaging and personalized medicine, Nat. Rev. Clin. Oncol. 14 (12) (2017) 749–762.
[33] Pengpeng Xu, et al., Radiomics: The next frontier of cardiac computed tomography, Circ.: Cardiovasc. Imaging 14 (3) (2021) e011747.
[34] Peng Lin, et al., A Delta-radiomics model for preoperative evaluation of Neoadjuvant chemotherapy response in high-grade osteosarcoma, Cancer Imaging 20 (1) (2020) 1–12.
[35] Roberto Berenguer, et al., Radiomics of CT features may be nonreproducible and redundant: influence of CT acquisition parameters, Radiology 288 (2) (2018) 407–415.
[36] Chang Min Park, Can artificial intelligence fix the reproducibility problem of radiomics? Radiology 292 (2) (2019) 374–375.
[37] Seul Bi Lee, et al., Deep learning-based image conversion improves the reproducibility of computed tomography radiomics features: A phantom study, Invest. Radiol. (2021).
[38] Jooae Choe, et al., Deep learning–based image conversion of CT reconstruction kernels improves radiomics reproducibility for pulmonary nodules or masses, Radiology 292 (2) (2019) 365–373.
[39] Hassan Bagher-Ebadian, Farzan Siddiqui, Chang Liu, Benjamin Movsas, Indrin J. Chetty, On the impact of smoothing and noise on robustness of CT and CBCT radiomics features for patients with head and neck cancers, Med. Phys. 44 (5) (2017) 1755–1770.
[40] Junhua Chen, Inigo Bermejo, Andre Dekker, Leonard Wee, Generative models improve radiomics performance in different tasks and different datasets: An experimental study, Phys. Medica 98 (2022) 11–17.
[41] Junhua Chen, Chong Zhang, Alberto Traverso, Ivan Zhovannik, Andre Dekker, Leonard Wee, Inigo Bermejo, Generative models improve radiomics reproducibility in low dose CTs: a simulation study, Phys. Med. Biol. 66 (16) (2021) 165002.
[42] Lin Fu, et al., Comparison between pre-log and post-log statistical models in ultra-low-dose CT reconstruction, IEEE Trans. Med. Imaging 36 (3) (2016) 707–720.
[43] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, Karen Egiazarian, Image denoising by sparse 3-D transform-domain collaborative filtering, IEEE Trans. Image Process. 16 (8) (2007) 2080–2095.
[44] P. Fumene Feruglio, Claudio Vinegoni, J. Gros, A. Sbarbati, R. Weissleder, Block matching 3D random noise filtering for absorption optical projection tomography, Phys. Med. Biol. 55 (18) (2010) 5401.
[45] Andrew L. Maas, Awni Y. Hannun, Andrew Y. Ng, Rectifier nonlinearities improve neural network acoustic models, in: Proc. Icml, Vol. 30, Citeseer, 2013, p. 3.
[46] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.
[47] Jie Hu, Li Shen, Gang Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[48] Abhijit Guha Roy, Nassir Navab, Christian Wachinger, Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2018, pp. 421–429.

[49] Xavier Glorot, Yoshua Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.
[50] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1026–1034.
[51] David E. Rumelhart, Geoffrey E. Hinton, Ronald J. Williams, Learning representations by back-propagating errors, Nature 323 (6088) (1986) 533–536.
[52] Martin Arjovsky, Soumith Chintala, Léon Bottou, Wasserstein generative adversarial networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 214–223.
[53] Zhou Wang, Eero P. Simoncelli, Alan C. Bovik, Multiscale structural similarity for image quality assessment, in: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, Vol. 2, Ieee, 2003, pp. 1398–1402.
[54] Joost J.M. Van Griethuysen, et al., Computational radiomics system to decode the radiographic phenotype, Cancer Res. 77 (21) (2017) e104–e107.
[55] I. Lawrence, Kuei Lin, A concordance correlation coefficient to evaluate reproducibility, Biometrics (1989) 255–268.
[56] Jurgen Peerlings, et al., Stability of radiomics features in apparent diffusion coefficient maps from a multi-centre test-retest trial, Sci. Rep. 9 (1) (2019) 1–10.
[57] Diederik P. Kingma, Jimmy Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.
[58] Ke Sheng, Shuiping Gou, Jiaolong Wu, Sharon X. Qi, Denoised and texture enhanced MVCT to improve soft tissue conspicuity, Med. Phys. 41 (10) (2014) 101916.
[59] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, Ruslan R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors, 2012, arXiv preprint arXiv:1207.0580.
[60] Yarin Gal, Jiri Hron, Alex Kendall, Concrete dropout, 2017, arXiv preprint arXiv:1705.07832.
[61] Yarin Gal, Zoubin Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: International Conference on Machine Learning, PMLR, 2016, pp. 1050–1059.
[62] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, Daan Wierstra, Weight uncertainty in neural network, in: International Conference on Machine Learning, PMLR, 2015, pp. 1613–1622.
[63] Max-Heinrich Laves, Malte Tölle, Tobias Ortmaier, Uncertainty estimation in medical image denoising with bayesian deep image prior, in: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis, Springer, 2020, pp. 81–96.
[64] Dmitry Ulyanov, Andrea Vedaldi, Victor Lempitsky, Deep image prior, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9446–9454.
[65] Pengwei Wu, Alejandro Sisniega, Ali Uneri, Runze Han, Craig Jones, Prasad Vagdargi, Xiaoxuan Zhang, Mark Luciano, William Anderson, Jeffrey Siewerdsen, Using uncertainty in deep learning reconstruction for cone-beam CT of the brain, 2021, arXiv preprint arXiv:2108.09229.
[66] Dan E. Dudgeon, Russell M. Mersereau, Multidimensional Digital Signal Processing, Prentice-Hall, 1984.
[67] Hyoung Suk Park, Jineon Baek, Sun Kyoung You, Jae Kyu Choi, Jin Keun Seo, Unpaired image denoising using a generative adversarial network in X-ray CT, IEEE Access 7 (2019) 110414–110425.