Automatic head computed tomography image noise quantification with deep learning

Inkinen, Satu I.

2022-07

Inkinen , S I , Mäkelä , T , Kaasalainen , T , Peltonen , J , Kangasniemi , M & Kortesniemi , M 2022 , ' Automatic head computed tomography image noise quantification with deep learning ' , Physica Medica , vol. 99 , pp. 102-112 . https://doi.org/10.1016/j.ejmp.2022.05.011

http://hdl.handle.net/10138/351307
https://doi.org/10.1016/j.ejmp.2022.05.011

Physica Medica 99 (2022) 102–112

Contents lists available at ScienceDirect

Physica Medica

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

Automatic head computed tomography image noise quantification with deep learning

Satu I. Inkinen [a,*], Teemu Mäkelä [a,b], Touko Kaasalainen [a], Juha Peltonen [a], Marko Kangasniemi [a], Mika Kortesniemi [a]

[a] HUS Diagnostic Center, Radiology, Helsinki University and Helsinki University Hospital, Haartmaninkatu 4, 00290 Helsinki, Finland
[b] Department of Physics, University of Helsinki, P.O. Box 64, FI-00014 Helsinki, Finland

ARTICLE INFO

Keywords:
Anthropomorphic phantom
Brain
Computed tomography
Deep learning
Image quality
Noise

ABSTRACT

Purpose: Computed tomography (CT) image noise is usually determined by standard deviation (SD) of pixel values from uniform image regions. This study investigates how deep learning (DL) could be applied in head CT image noise estimation.

Methods: Two approaches were investigated for noise image estimation of a single acquisition image: direct noise image estimation using supervised DnCNN convolutional neural network (CNN) architecture, and subtraction of a denoised image estimated with denoising UNet-CNN experimented with supervised and unsupervised noise2noise training approaches. Noise was assessed with local SD maps using 3D- and 2D-CNN architectures. Anthropomorphic phantom CT image dataset (N = 9 scans, 3 repetitions) was used for DL-model comparisons. Mean square error (MSE) and mean absolute percentage errors (MAPE) of SD values were determined using the SD values of subtraction images as ground truth. Open-source clinical head CT low-dose dataset (N_{train} = 37, N_{test} = 10 subjects) were used to demonstrate DL applicability in noise estimation from manually labeled uniform regions and in automated noise and contrast assessment.

Results: The direct SD estimation using 3D-CNN was the most accurate assessment method when comparing in phantom dataset (MAPE = 15.5%, MSE = 6.3HU). Unsupervised noise2noise approach provided only slightly inferior results (MAPE = 20.2%, MSE = 13.7HU). 2DCNN and unsupervised UNet models provided the smallest MSE on clinical labeled uniform regions.

Conclusions: DL-based clinical image assessment is feasible and provides acceptable accuracy as compared to true image noise. Noise2noise approach may be feasible in clinical use where no ground truth data is available. Noise estimation combined with tissue segmentation may enable more comprehensive image quality characterization.

Introduction

Currently, digital radiology is producing a continuously increasing amount of image data with a steadily increasing part of 3D imaging studies in which the contribution from computed tomography (CT) has been substantial [1–3]. CT also provides a dominating part of the total radiation exposure in radiology, even 70% [4], which makes this single modality especially relevant for optimisation. Assessment of clinical image quality plays an important role in the optimisation process which seeks to combine and balance the level of image quality in relation to radiation dose. Head scans are among the most common CT studies with a general diagnostic task to distinguish between relevant brain tissue including white and gray matter, ventricles, vascular structures, bone and subcutaneous soft tissue. In the context of head CT clinical image quality assessment, these tissue types are therefore the most relevant targets for image quality quantification.

The purpose of medical imaging is to provide reliable information for accurate diagnosis and subsequent clinical decisions for effective patient care. Image quality refers to how well the acquired images can serve the purpose of diagnostics while taking into account the existing diagnostic process and recommendations of acceptable images in various diagnostic tasks [5]. Image quality may be characterised by parameters ranging from physical parameters (noise, contrast, spatial resolution and derivatives) to clinical parameters (sensitivity, specificity, accuracy and derivatives) [6]. In x-ray imaging methods, improved image quality has traditionally been achieved by using higher radiation exposure as it

* Corresponding author.
E-mail address: satu.inkinen@hus.fi (S.I. Inkinen).

https://doi.org/10.1016/j.ejmp.2022.05.011
Received 20 January 2022; Received in revised form 2 April 2022; Accepted 25 May 2022
Available online 4 June 2022

Table 1
Imaging parameters for phantom dataset and its division for training, validation and test sets.

Dataset	Tube peak kilovoltage (kV)	Table height	Tube current (mAs)	CTDIvol (mGy)
Train	120	Centered	50	9.2
Train	120	Centered	100	18.5
Train	120	Centered	250	46.5
Train	120	Centered	300	55.8
Validation	120	Centered	150	27.7
[1]Test	120	Centered	200	37.0
Test	100	Centered	200	23.9
Test	120	3 cm down	200	37.0
Test	120	3 cm up	200	37.0

[1] Corresponding clinical head CT scan protocol.

reduces image noise while increasing dose. Optimisation, however, involves an inferred and a careful balance between radiation dose and image quality with a fundamental aim to ensure adequate diagnostic image quality with minimal radiation exposure to patients. Thus, the absolute condition of successful optimisation is the adequacy of diagnostic image quality. Traditional optimisation principle has been described by the ALARA acronym [7] – as low as reasonably achievable – which has focused more on radiation safety and radiation dose minimisation. The main reason for the dose focus in medical imaging optimisation has been the ease of measuring and monitoring of the physical dose output of imaging equipment – in comparison to image quality - as a part of regular quality assurance in radiology and in compliance with international technical standards. On the other hand, diagnostic or clinical image quality has been much harder to determine in a reliable, repeatable and unambiguous way. [5,8].

Physical image quality parameters are usually measured from technical test objects (phantoms), and they have been utilized as a part of quality assessment (QA) for a long time. However, the ability of physical image parameters determined *in phantoms* to describe more comprehensive diagnostic image quality *in patients* is limited. More specifically, phantoms do not provide a sufficiently versatile surrogate to patients as regards to individual anatomical variability, gender and age representations, tissue compositions, physiological motion and numerous pathological stages deviating from healthy individuals [5,8].

Another approach to assess diagnostic image quality is to use human or model observers [6,9]. However, human observer studies are laborious and present limitations by intra- and inter-observer variability limiting the reliability of such assessment results especially with a small number of observers/expert reviewers. These factors inevitably reduce the applicability of such subjective assessment in routine clinical level image quality monitoring. On the other hand, objective clinical level image quality assessment with model observers or equivalent methods is a demanding and non-trivial task with varying clinical representations of patients. However, there are some initial developments in this direction [10–13]. Regardless of these challenges, image quality assessment remains a pivotal target for both radiological optimisation and QA.

New developments in artificial intelligence (AI) especially in the regime of deep learning (DL) has brought new methods to various applications of healthcare [14]. Diagnostic radiology involving large amounts of standardised image data has been an attractive target of DL methods. Altogether, medical imaging benefits from an abundance of training data, transferability of DL models from previously trained image-based DL networks and increasing access to datasets complemented with ground truth labeling [15]. Due to this potential versatility, DL methods may be also considered for QA purposes and specifically for image quality assessment. For example, Kretz et al. 2020 developed a DL method for automated mammography image quality assurance from a technical image quality phantom [16], and in another study, a DL was utilized for motion corruption assessment of MRI brain scans in retake evaluation [17].

Patient-specific image quality assessment of CT images has been focusing on noise and HU value measurement using traditional image quality metrics and image processing methods [6,10,18]. Noise magnitude estimation from clinical CT images is usually performed by assessing local standard deviation (SD in Hounsfield units) in pixels from uniform regions. However, this approach limits the investigation only for uniform image regions and as for example the clinical CT head scans the different anatomical structures pose a challenge in comprehensive noise assessment biasing the SD assessment in different head regions. One approach to overcome this issue is to use subtracted adjacent Z-slice images [18] but this approach also has inherent limitations as the slice thickness selection affects the resulting outcome and tissue boundaries are visible in the subtracted image. To obtain a real noise realization image, a dual acquisition is needed with image subtraction. However, this is not clinically feasible as the patient dose would double. A study by Abadi et al 2017 focused on automated assessment of organ-based distribution of Hounsfield units from clinical chest CT images using segmentation with thresholding and image processing [10]. Even though this method is applicable to chest CT it is difficult to generalize to other imaging protocols and anatomical regions. DL-based methods could offer more flexible and generalizable solutions.

In this study, DL is applied to clinical head CT image quality assessment of image noise to overcome the aforementioned challenges in noise estimation. The primary goal is to develop a fast and accurate DL method which directly estimates image noise and noise magnitude (described as SD) from single CT scan corresponding to normal clinical CT image acquisition. We assess both supervised and unsupervised training schemes while evaluating several DL architectural options using anthropomorphic phantom data from several varied and repeated CT scans. In addition, we also demonstrate with openly available clinical CT head dataset, how the developed DL noise and SD assessment pipeline can be incorporated into an automated image quality assessment framework. This framework allows image quality monitoring which is pivotal for a comprehensive optimisation process.

Material and methods

Phantom dataset

An anthropomorphic dosimetry phantom (CIRS ATOM 702-D, Norfolk, USA) was scanned with a Revolution EVO (GE Healthcare, Boston, MA, USA) CT system using nine different scan settings (Table 1). The CIRS ATOM 702-D is female phantom model, and it contains bone and soft tissue structures, but excludes specific soft tissue such as white and gray matter as separate materials. The slice thickness (0.625 mm), pixel spacing (0.488 mm) and collimation (20 mm) were kept constant during different scans. Rotation time of 1 s was kept constant over all scans. The scan acquisition was repeated three times. Scan data was reconstructed with a standard kernel and ASiR-V30% statistical iterative reconstruction. The image stacks were divided into training, validation, and testing sets based on scan settings (Table 1). For the test set, two scans were performed with varying vertical off-centering. This was performed to gain variability in the scan positions which also occurs in clinical practice [19].

Noise image and noise magnitude estimation using deep learning

To improve the sensitivity of the noise image and noise magnitude estimation by assessing the local SD map, different DL noise estimation approaches were investigated using the phantom dataset with three repeated CT acquisitions ($\boldsymbol{X}_{1,2,3}$). These acquisitions enabled ground truth (GT) labeling as the ground truth noise image and subsequent local SD maps can be assessed by subtracting the slice images of two repeated CT scan acquisitions. Several convolutional neural network models were experimented with three distinguishable approaches: 1. Direct local SD map estimation (supervised learning), 2. Direct noise image estimation

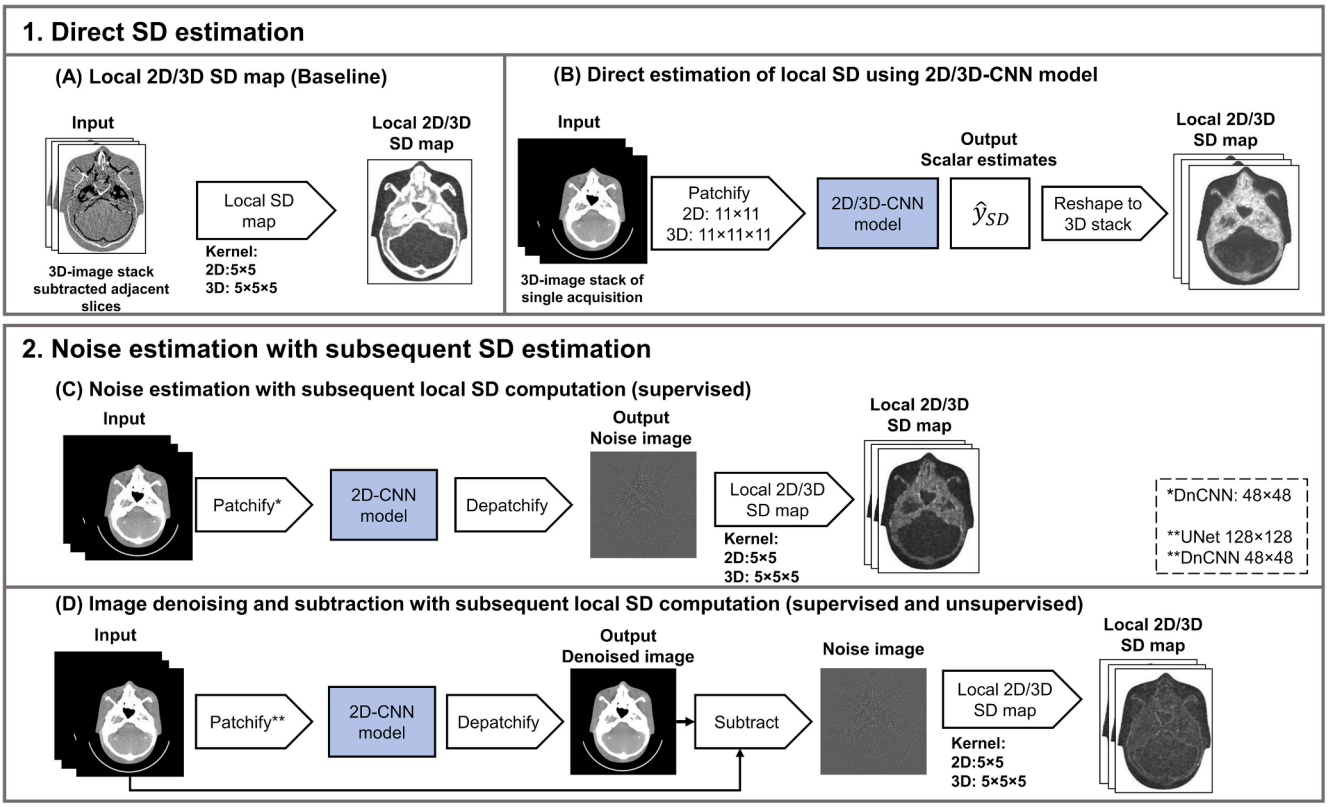

Fig. 1. Workflow diagrams of different noise estimation schemes: A) depicts reference local SD estimation scheme without any noise estimation (baseline). The SD values are very high at the tissue boundaries. B) depicts direct SD estimation workflow where the 2D or 3D CNN network estimates local SD values directly from the input image. C) illustrates the 2D-CNN approaches where first a noise image is estimated with DL with subsequent local SD map computation and D) illustrates the denoising DL networks in which denoised image is subtracted from the input image to obtain the noise image for the local SD estimation.

with subsequent local SD map computation (supervised learning) and 3. Differentiation of noise using denoising CNN and subsequent local SD map computation with supervised and unsupervised learning approaches (Fig. 1).

Local standard deviation derived from subtracted adjacent slice images (Baseline)

We computed the local SD value maps in 2D and 3D using the 5×5 and $5 \times 5 \times 5$ window kernels, respectively, over single acquisition. In this reference approach, the noise image is first estimated from the difference image of adjacent slices and divided by a factor of $\sqrt{2}$ to account for error propagation [18,20] (Fig. 1A). This is used as a baseline result for comparison, as the underlying challenge using the local SD values directly from single acquisition is that the anatomical structures distort the estimated volumetric SD maps (Fig. 1A).

Direct estimation of local standard deviation using convolutional neural network model

This direct approach aims to estimate SD maps ($\boldsymbol{Y}_{SD}$) directly from single acquisition images using CNN model (Fig. 1B). First, noise realization map was computed using the second ($\boldsymbol{X}_2$) and third ($\boldsymbol{X}_3$) CT scan acquisitions phantom image stacks difference $\boldsymbol{X}_{2\text{-}3}$ down-scaled with $\sqrt{2}$ division in order to normalise noise level according to Poisson statistics. The subscript denotes the acquisition number. Subsequently, GT labels (i.e. SD maps) were computed from noise realization image stack with rolling standard deviation filtering with 5×5 and $5 \times 5 \times 5$ window kernels for 2D and 3D cases, respectively. The input tensor ($\boldsymbol{x}_1$) sizes for the 2D and 3D neural networks were $1 \times 11 \times 11$ and $1 \times 11 \times 11 \times 11$, respectively. Both 2D and 3D patches were extracted from the head region of the image stack ($\boldsymbol{X}_1$) and no overlap was allowed between patches in training and validation set. The CNN-networks consisted of series convolutional filters followed by batch normalization and rectified linear unit activations (ReLU) (Fig. 2A), and it learns to estimate a mapping from the input patch ($\boldsymbol{x}_1$) to a scalar local SD value ($\hat{y}_{SD}$) using supervised learning with $\{\boldsymbol{x}_1, y_{SD}\}$ training pairs where y_{SD} is the SD at the center of the patch (Fig. 2A). The input image range was shifted with −200 HU and scaled with 1372 HU which was the maximum HU value in the phantom dataset, before feeding in the network. This was done to keep the network activation values small, i.e., to avoid the back-propagation gradients to explode. The output was again scaled back to SD value range before computing mean squared (MSE) loss function in training phase.

Direct noise estimation with subsequent SD computation

For noise estimation we experimented with DnCNN and UNet convolutional neural net models in 2D (Fig. 1C and Fig. 2B and 2C) [21,22]. This supervised learning process uses 2D training pair patch axial images $\{\boldsymbol{x}_1, \boldsymbol{x}_{1\text{-}2}\}$ of input image and subtracted image (Table 2). The selected window combinations were 48×48 and 128×128 for DnCNN and UNet, respectively. The smaller 48×48 patch was extracted from the head region and the larger 128×128 patch was taken from the whole image area. For both patches, no overlap was allowed. The input image range was scaled from −1024 to 1812 HUs (minimum and maximum range in phantom dataset) before feeding in the network and rescaled back to HU value range before computing mean squared (MSE) loss function between $\{\boldsymbol{x}_1, \boldsymbol{x}_{1\text{-}2}\}$ training pair in training phase. The estimated noise image was scaled with $\sqrt{2}$ division prior to SD map computation also in this method.

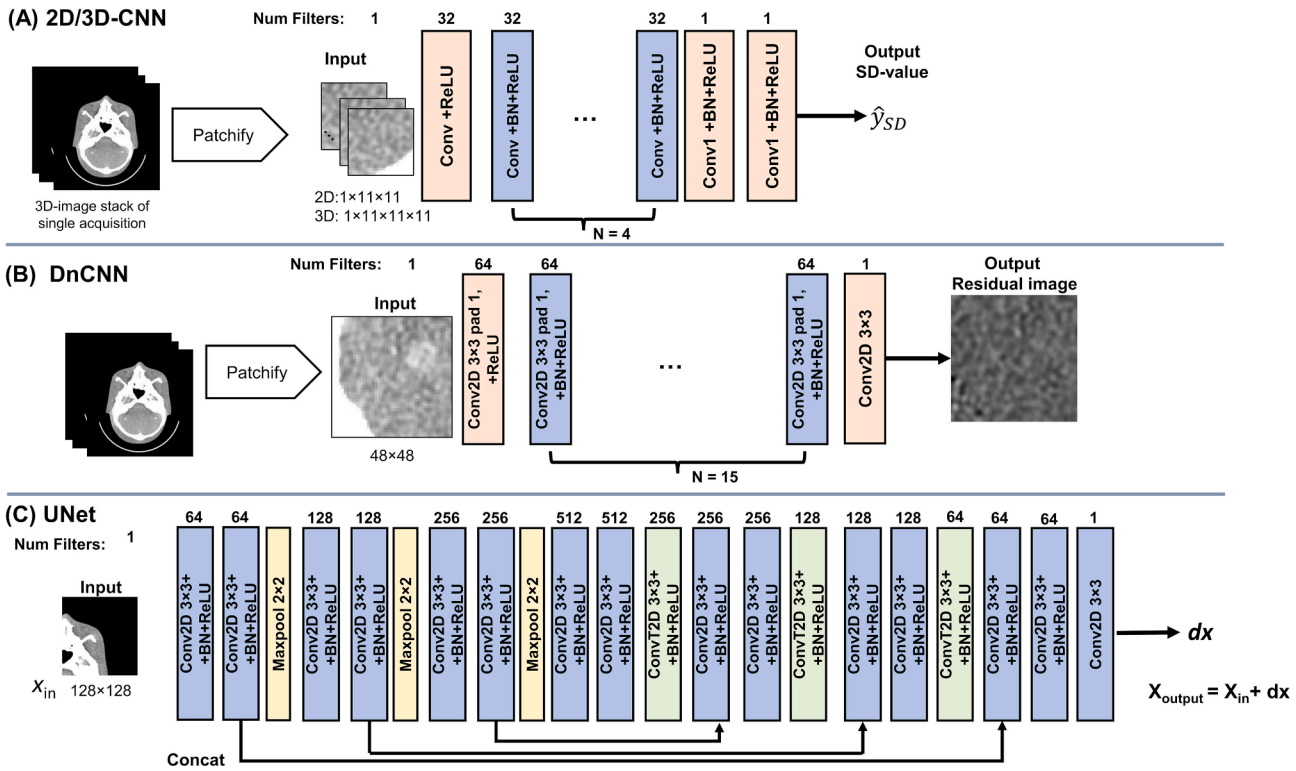

Fig. 2. Illustration of the different convolutional neural network architechtures. In (A) 2D and 3D CNNs are trained to estimate local 2D/3D SD values from $1 \times 11 \times 11$ and, $1 \times 11 \times 11 \times 11$ inputs, respectively. The kernel size for convolution filters (conv, conv1) for 2D and 3D networks were (3×3, 1×1) and ($3 \times 3 \times 3$, $1 \times 1 \times 1$), respectively. (B) DnCNN structure is used to estimate the noise image as well as in the unsupervised denoising and (C) Residual UNet is used for denoising assessment of noise in both supervised and unsupervised setting. *N* denotes the number of Convolution, batch normalization and ReLU operations and *dx* denotes the residual output of convolutional network which added to the input image.

Table 2
Dataset splits for different patch sizes for the phantom dataset and model hyperparameters. The dataset size varies due to different patch window sizes. Also, no overlap was allowed between patches.

Dim.	Model	Training dataset	Batch size	Patch size	Initial learning rate	Epochs	Num. of training samples	Training time (HH:MM)
2D	DnCNN	Phantom	128	48 × 48	1e-4	50	42 600	01:13
2D	UNet	Phantom	64	128 × 128	1e-4	25	23 104	01:23
2D	2DCNN	Phantom	128	11 × 11	1e-3	50	828 288	00:45
3D	3DCNN	Phantom	128	11 × 11 × 11	1e-3	50	75 288	00:07
2D	DnCNN	Clinical	64	48 × 48	1e-4	50	132 700	03:27
2D	UNet	Clinical	64	128 × 128	1e-4	25	21 232	01:25

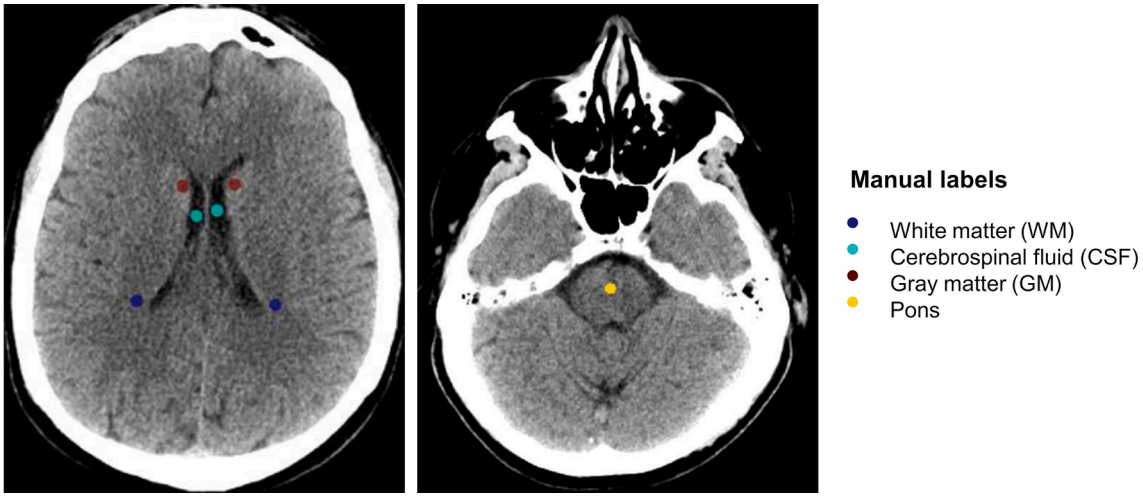

Fig. 3. Example of manual annotations made for the test set (subject 01380). The diameter of the annotated region is 10 pixels.

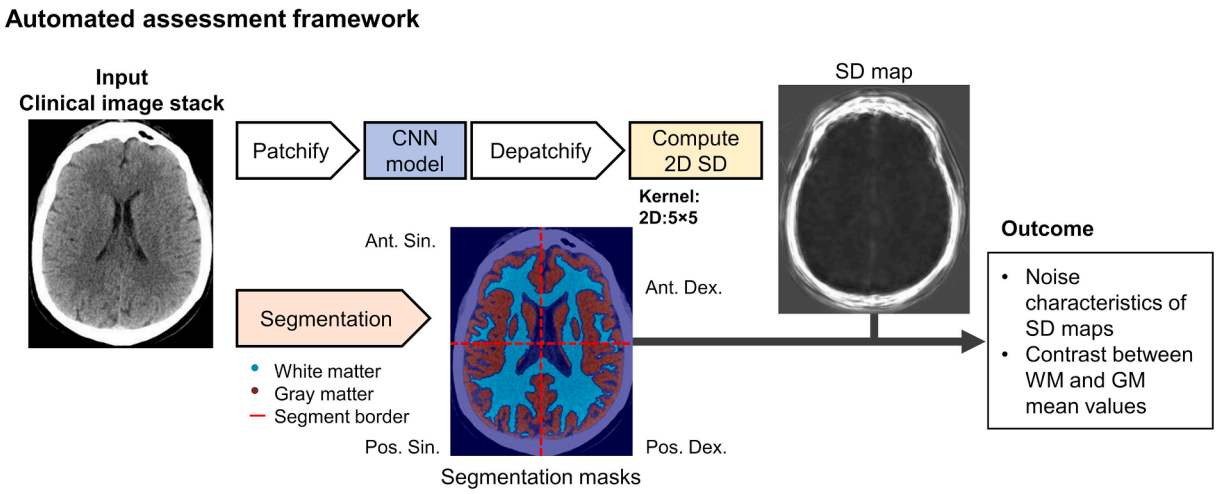

Fig. 4. Illustration of the automated assessment framework for the clinical dataset. The segmentation of white and gray matter is further divided into four segments (red lines) in which the SD characteristics and contrast is assessed. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Local 2D SD map errors from different model outcomes evaluated on phantom test data (mean ± SD). The unsupervised methods were also trained on clinical dataset as no GT data is required. For MSE and MAPE, the SD values are standard deviation of squared errors and standard deviation of absolute percentage error, respectively.

Model	Model output	Training dataset	Patch size	MSE (HU)	MAPE (%)
2DCNN (super)	SD estimate	Phantom	11 × 11	10.6 ± 35.5	25.6 ± 22.6
UNet (unsuper)	Denoised img.	Phantom	128 × 128	19.0 ± 61.6	33.9 ± 31.8
UNet (super)	Denoised img.	Phantom	128 × 128	26.9 ± 69.7	38.0 ± 20.3
DnCNN (super)	Noise image	Phantom	48 × 48	25.3 ± 66.9	36.6 ± 20.9
Local SD[1]	None	None	None	565.9 ± 2319.4	97.7 ± 201.2
UNet (unsuper)	Denoised img.	Clinical	128 × 128	18.8 ± 60.4	32.9 ± 29.5
DnCNN (usurper)	Denoised img.	Clinical	48 × 48	89.5 ± 3906.5	39.2 ± 142.8

Local SD is computed from the adjacent slice images from the single acquisition

Table 4
Local 3D SD map errors from different model outcomes evaluated on phantom test data (mean ± SD). The unsupervised methods were also trained on clinical dataset as no ground truth data is required. For MSE and MAPE, the SD values are standard deviation of squared errors and standard deviation of absolute percentage error, respectively.

Model	Model output	Training dataset	Patch size	MSE (HU)	MAPE (%)
3DCNN	SD estimate	Phantom	11 × 11 × 11	6.3 ± 22.0	15.5 ± 13.0
UNet (unsuper)	Denoised img.	Phantom	128 × 128	13.7 ± 47.0	20.2 ± 17.7
UNet (super)	Denoised img.	Phantom	128 × 128	25.2 ± 53.2	37.3 ± 14.4
DnCNN (super)	Noise image	Phantom	48 × 48	23.2 ± 50.1	35.1 ± 15.3
Local SD[1]	None	None	None	1245.0 ± 4036.0	133.7 ± 255.3
UNet (unsuper)	Denoised img.	Clinical	128 × 128	14.8 ± 47.780	21.1 ± 17.3
DnCNN (usurper)	Denoised img.	Clinical	48 × 48	84.0 ± 3887.9	27.2 ± 116.7

Local SD is computed from the adjacent slice images from the single acquisition

Learned denoising and image subtraction with subsequent SD computation

As a final approach, we experimented with DL denoising schemes in both supervised and unsupervised noise2noise learning schemes (Fig. 1D) [23]. In this approach, the CNN learns to denoise the input image. In supervised setting, we fed training patch pairs $\{\boldsymbol{x}_1, \bar{\boldsymbol{x}}\}$, where $\bar{\boldsymbol{x}}$ denotes average image over three repeated scan acquisitions. For the unsupervised noise2noise learning, the input ($\boldsymbol{x}_1$) is corrupted with additive Gaussian noise. We performed the Gaussian noise corruption after tensor normalization using Gaussian noise with unit normal distribution scaled with 0.01. For denoising, we applied the UNet architecture using 128x128 window size (Table 2, Fig. 2C) and MSE loss. The large 128 × 128 patching was taken from the whole image area i.e. it was not restricted inside the phantom. The SD values were estimated from subtraction image stacks between original input image (without noise corruption) and denoised image.

Model training

For model training, we used Python with PyTorch (v.1.8.1) GPU version with the Nvidia GTX1080 Ti 11 Gbps. The training data was read in the RAM (64 GB) during the initial data loading phase to optimize performance. The computation times for training varied between different networks as the number epochs and batch size were tuned for each model using a validation set (Table 2). MSE loss was used as a loss function for all models with ADAM optimizer with default parameters and training parameters are presented in Table 2. For all models, the learning rate was reduced using cosine annealing scheduler with the number of iterations set to equal number of epochs [24]. After the DL models were trained, the computation time to obtain the final SD maps varied between different DL approaches such that the fastest computation time was with the UNet noise estimation with subsequent SD map computation within 27 and 52 s for the 3D and 2D SD maps for one phantom image volume of 512 × 512 × 361, respectively. The slowest method was the direct estimation of local SD using convolutional neural network model. As in 3DCNN and 2DCNN, the corresponding computational times were 01:02:15 and 00:42:48, respectively. However, this

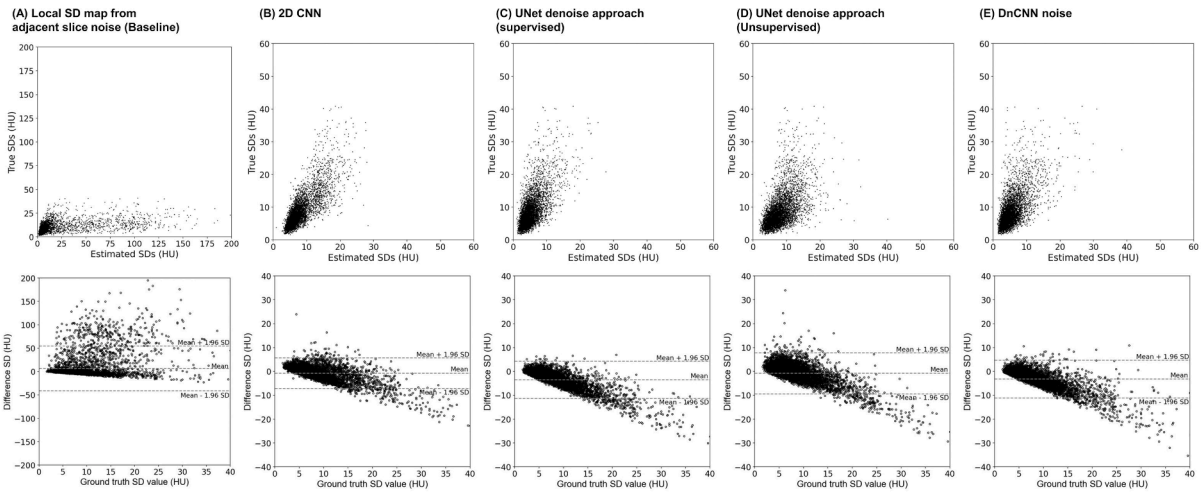

Fig. 5. Top row scatter plots between the ground truth 2D SD values and bottom row corresponding Bland-Altman plots. Plots represent N = 5000 points randomly sampled from test set data within the phantom region.

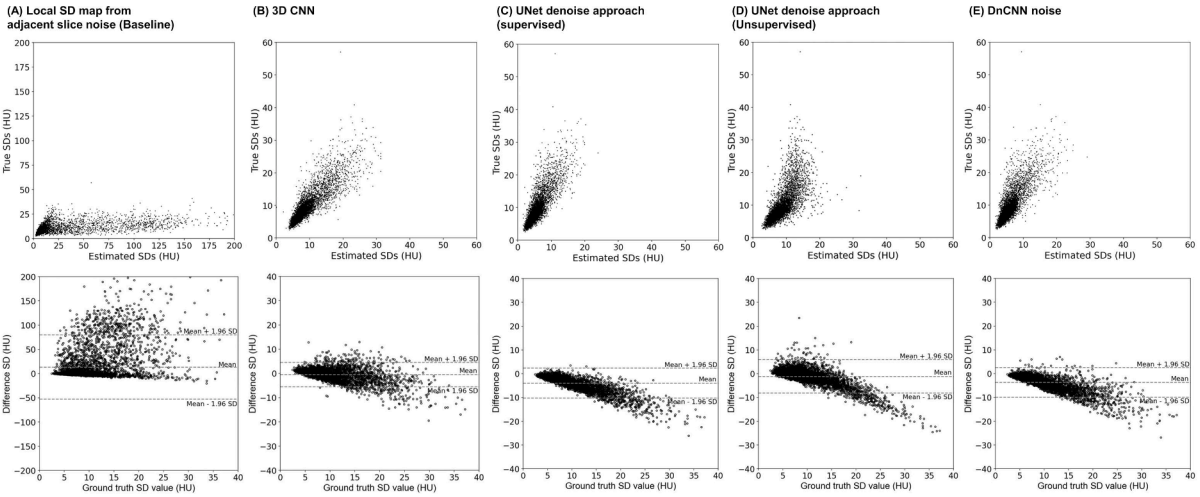

Fig. 6. Top row Scatter plots between the ground truth 3D SD values and corresponding Bland-Altman plots. Plots represent N = 5000 points randomly sampled from test set data within the phantom region.

computation time could be reduced to one-third if this voxel SD computation would be limited to the phantom region.

Performance evaluation metrics

Mean square error (MSE) and mean absolute percentage error (MAPE) were computed for a phantom test dataset to assess and compare the SD map accuracy of the different DL approaches. In addition, Bland-Altman plots of 5000 datapoints randomly sampled from the phantom test dataset SD maps were computed for each model to evaluate trend behavior in SD estimation accuracy.

Clinical dataset

Openly available Low Dose CT Grand Challenge dataset were used for an assessment of the developed noise estimation framework [25,26]. A subset of this dataset containing only brain scans were collected from The Cancer Imaging Archive [27]. The subjects were scanned with a Somatom Definition Flash (Siemens, Erlangen, Germany) CT scanner in axial scanning mode. The tube peak kilovoltage was set to 120 kVp. The slice thickness was 5 mm and pixel spacing 0.488 mm. The reconstruction kernel was H40s and the dataset was divided into train, validation and test sets with 37, 2 and 10 subjects with corresponding average tube exposures of 287, 350 and 277 mAs, respectively.

Automated noise estimation framework: Assessment on clinical data

After phantom test data comparison, additional experiments were performed with the openly available clinical head dataset data with the best performing supervised models trained on phantom data and unsupervised models trained with clinical data. Supervised models could not be trained with clinical data as no GT SD values were available for this dataset.

First, the performance of the models was assessed using manually annotated circular (diameter = 10 pixels) region-of-interest (ROI) from uniform anatomical regions: gray matter (GM) on caudate nucleus (Caput), white matter (WM) on centrum semiovale, cerebrospinal fluid and pons (Fig. 3). The uniformity region covered three adjacent slice images (Fig. 3 shows center slice image). The small ROI were chosen such that no anatomical borders were located within the regions. Therefore, the SD values estimated from single acquisition was used as the reference. Comparative local 5 × 5 2D SD maps were computed for different DL methods. Subsequently, mean SD value from the same annotated region of each SD map was computed, and MSE between the

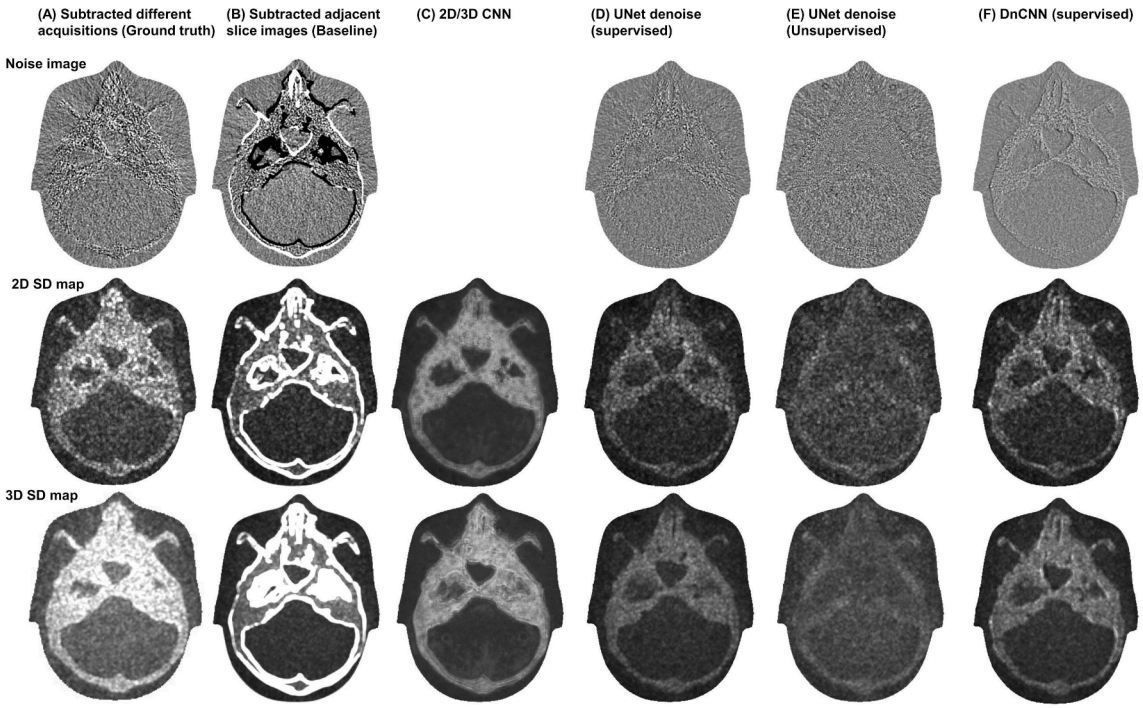

Fig. 7. Top Row estimated noise images with different methods, middle corresponding local 2D SD maps, and bottom corresponding local 3D SD maps on phantom data. The 2D- and 3D-CNN models directly estimates the SD values from single acquisition. Windowing: [-50, 50] HU (top), [0,40] HU (middle, bottom).

Table 5
Image quality results (MSE) from different model outcomes on clinical data (mean ± SD). Here the reference data is SD values determined from subtracted adjacent slice images. For MSE, the SD values are standard deviation of squared errors.

Model	Patch size	GM left (HU)	GM right (HU)	WM left (HU)	WM right (HU)	Pons (HU)	CSF left (HU)	CSF right (HU)
2DCNN	11 × 11	0.50 ± 0.40	0.37 ± 0.34	0.32 ± 0.25	0.43 ± 0.38	0.25 ± 0.29	0.44 ± 0.25	0.26 ± 0.16
UNet (super)	128 × 128	1.88 ± 1.29	1.84 ± 0.74	1.81 ± 1.40	1.37 ± 0.82	1.86 ± 1.08	1.64 ± 0.74	2.23± 0.74
UNet (unsuper)	128 × 128	0.22 ± 0.49	0.22 ± 0.28	0.25 ± 0.64	0.10 ± 0.18	0.18 ± 0.20	0.09 ± 0.08	0.17± 0.16
DnCNN (unsuper)	48 × 48	0.33 ± 0.72	0.32 ± 0.36	0.37 ± 0.66	0.15 ± 0.21	0.15 ± 0.19	0.17 ± 0.21	0.45± 0.36

mean SD values of reference and different DL methods was determined. The 2D SD assessment was chosen over 3D due to the large slice thickness of the clinical dataset.

Then, we developed an automated assessment framework of SD maps for clinical head CT image quality assessment which automatically assess SD value distributions, and contrast estimate between gray and white matter, and visualizes the results from the segmented regions (Fig. 4). We included this framework to test the feasibility of SD value estimation combined with contrast estimation on clinical images.

The clinical dataset was automatically segmented to GM and WM for the framework. The segmentation was automatically performed in MATLAB using CTseg [28]. The segmentation provides tissue probability maps which were thresholded above 0.3 to produce GM and WM masks. The overlapping regions of GM and WM masks were excluded and finally the segmented masks were eroded using morphological disk operator (diam = 3 pix) to avoid tissue borders in SD maps.

After segmentation, the SD and contrast assessment was performed as follows. The masks were divided into posterior left, posterior right, anterior left and anterior right regions (Fig. 4) and SD distribution characteristics and contrast between WM and GM mean HU values were computed for these four regions as well as from the total segmented volumes of GM and WM. Boxplots showing mean SD and contrast values for different segmentation regions in the test set are presented for different DL methods.

Results

Model performance comparison on phantom data

The most accurate SD estimation was provided with the 3DCNN architecture which directly estimates the SD values (Tables 3 and 4). Both, supervised and unsupervised denoising UNet models performed well on the phantom dataset. The Bland Altman analysis shows that especially high SD values are systematically underestimated by the DL models, but this effect is least present with the supervised 3DCNN model (Figs. 5 and 6). The visualizations of noise images and SD maps reveal that even though DL models perform better than the baseline, anatomical borders are still visible when compared to GT noise images (Fig. 7).

Image quality assessment for clinical data

When comparing the subtraction-based SD map values from the uniform manually labeled regions (Fig. 3), the 2DCNN and unsupervised UNet (trained with separate clinical dataset) models provided the smallest MSE errors (Table 5). All noise images provided by the different noise estimation methods show remaining traces of anatomical structures (Fig. 8). However, with DL methods, the anatomical borders are more suppressed than when compared with the baseline method using adjacent slices. Specifically, the SD maps derived from the direct

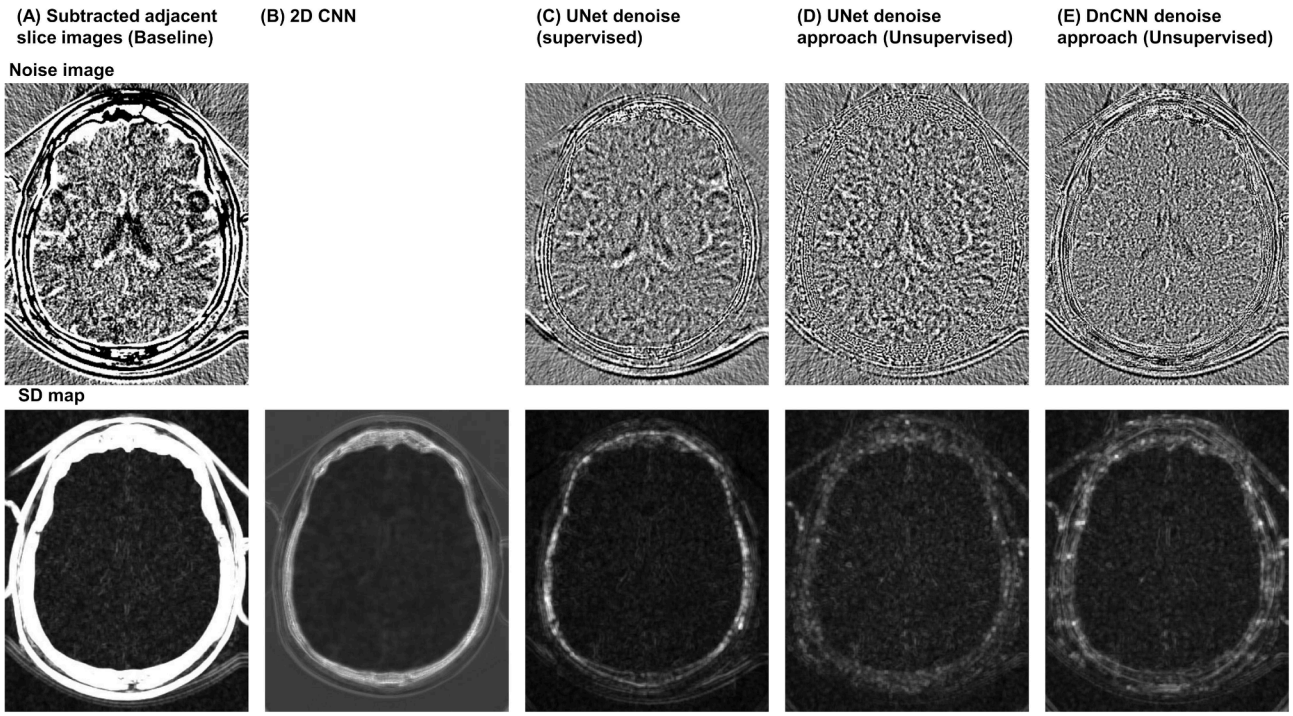

Fig. 8. Top Row estimated noise images with different methods and bottom row corresponding local SD maps (subject 01380). The 2DCNN model directly estimates the SD map. Windowing: [-50, 50] HU (top), [0,30] HU (bottom).

estimation method (2DCNN) showed least anatomical soft tissue structures in the SD maps (Fig. 8). Overall, the results demonstrate varying performance and feasibility of the automated image quality assessment framework while focusing on different locations in the head CT scan regions (Fig. 9). The WM regions show lower SD values compared to GM regions (Fig. 9, Table 5).

Discussion

In this study, we investigated how DL could be applied in head CT scan noise estimation using both phantom and clinical datasets, and by using supervised and unsupervised learning. To summarize, based on phantom assessment with ground truth data available, the direct mapping from input patch to SD values gave the most accurate noise estimation. However, in clinical situations where ground truth data from repeated acquisitions is not available, the unsupervised noise2noise denoising UNet could be used as an alternative approach. In this two-step approach, the first step provides an estimate of the noise image which can then be used in the second step in noise magnitude estimation via local SD values. The noise measurements can be complemented by automatic GM and WM segmentation to provide clinically relevant contrast information.

The traditional measurements of image quality have been focusing on technical image quality mainly by test object (phantom) acquisitions which can describe the performance of an imaging equipment in a repeatable and reproducible way. New imaging technology, especially in CT, has introduced challenges to this traditional approach due to more complex image post-processing and reconstruction methods, which makes it more difficult to infer clinically adequate image quality from these conventional technical measurements [6,29,30]. In addition, modern CT scanners used sophisticated patient size and shape dependent tube current and voltage modulation techniques for dose optimization, which in turn, further introduces patient-specific variations in image quality. Also, due to vendors' different technical solutions, clinical images have different appearances (noise, texture, artefact prominence). This creates additional challenges to e.g. cross-manufacturer protocol harmonization. Therefore, it is important to develop more sophisticated, robust, and patient-specific image quality assessment methods for CT.

Previous clinical image quality assessment studies have focused on assessing CT image noise properties, for example, by evaluating noise in air regions outside the patient [31] and with ranges of kernels sampling the homogeneous regions of the image data [18] providing promising results. However, it should be noted that noise assessment from in air regions outside patient are not reliable for iterative or DL-based reconstructions, as noise texture can differ significantly outside patient in those reconstruction techniques. Also image contrast has been measured in later studies [10], and noise measurements from clinical CT data have been utilised in further image quality metrics based on radiologist grading [12].

In contrast to previous studies, our work utilized data-driven DL approaches in the image noise estimation. The underlying challenge is the presence of anatomical structures in the noise images where the image quality assessment is performed. The common approach is to limit the image quality assessment on the uniform body regions where the local SD values can be measured directly or by computing noise images from subtracting adjacent slices [20,31]. These methods are inherently limited, as the first approach noise estimation is applicable to only part of the image area, whereas the second approach induces challenges of imprinted anatomical structures and low-frequency correlated noise. Our work utilized DL that can provide feasible tools for CT image noise estimation extending from phantom images to clinical images with variable characteristics introduced by actual patients and scan settings. Although this study focused on head CT scans, a robust, general-purpose estimator for clinical images is desired. Building a sufficiently versatile training set, if supervised learning is used, remains challenging.

The most relevant previous study was published by Christianson et al. involving automated technique to measure noise in clinical CT images. Specifically, the noise parameter was defined as a global - noise based on the mode of the kernel-based SD map histogram peak which corresponded to homogeneous tissue areas. Their results were validated by comparing the global noise with the reference

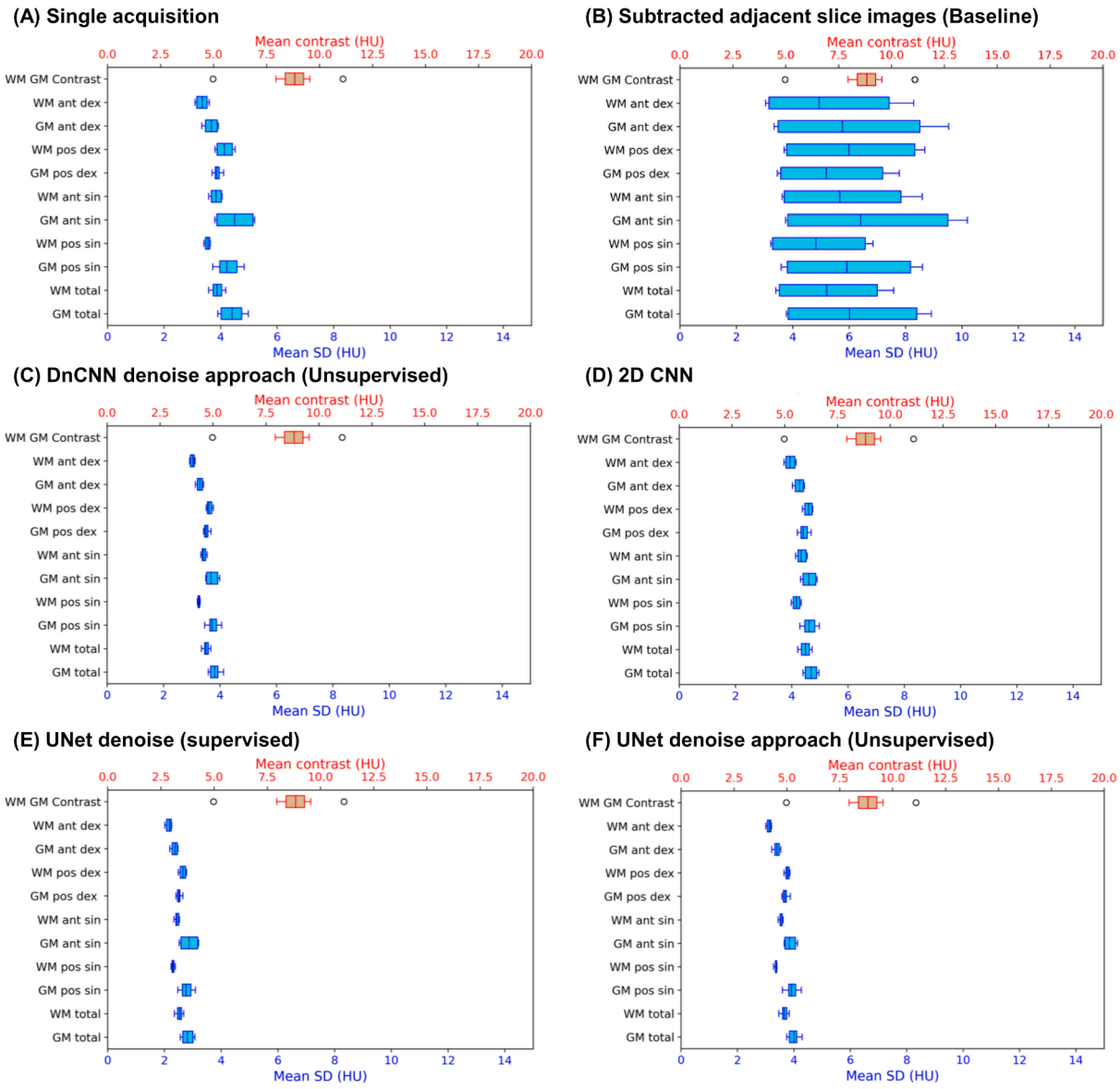

Fig. 9. Box plots of mean HU values over clinical test set (N = 10) for different DL assessment methods, and the contrast values computed directly from clinical images thus does not vary. ant = anterior, pos = posterior, sin = left, dex = right. Total = total segmented volume.

image noise, which was determined by the subtracted images from tissue phantom and additionally by comparing global noise with observer study results from six patient CT images (three thoracic and three abdominal images) [18]. Our method was taking a few steps forward by providing local noise maps from clinical CT images instead of providing a single global noise value. Therefore, our methods provide comprehensive noise evaluation from the CT images covering non-uniform body regions and offer localized noise information which can be targeted to certain diagnostically relevant tissue regions, in our case, WM and GM structures. Along with the noise estimation results, the contrast assessment of GM and WM was also enabled by the preceding step of automatic segmentation by applying CTseg algorithm.

Despite the promising applicability of DL in CT clinical image quality assessment, there are several limitations in study that need to be addressed. We focused only on the head CT scans and the models were developed using only head CT data. This region was chosen because it is among the most common CT examination regions [32]. The idea of DL-based noise measurement is to learn to ignore the background anatomical variation, and we acknowledge that the trained models are not directly applicable to other body parts. The DL models should be trained with data comprising the corresponding anatomy and morphological characteristics. Therefore, in future, these DL algorithms should be made more generalizable to images from any part of the body. Further, the organ segmentation algorithm must be applicable to different body regions for organ specific image quality assessment. The phantom data was acquired using only one CT scanner and a single reconstruction algorithm which likely poses a limitation for the model generalizability. Also, the models had high uncertainties i.e. large SD values when assessed using MSE and MAPE values, and when compared to GT images this variability could be located in the bony regions. More versatile model should be trained and validated with a broader range of data from several scanners, scan parameters, and reconstruction kernels enabling model adaptation to different noise textures and additional mean HU-value/contrast measurements. In addition, more versatile phantom models should be applied in the data collection with more realistic brain tissue structures with WM and GM matter regions and different anatomical variabilities.

The normalization strategy adopted in CNN training was to use global normalization values. In future studies, also local image-based approach should be studied as it might take better into account the

Table A1
Dataset division head CT subjects from the Low dose CT dataset. Full dose reconstructions were used in this study.

Set	Subject
Test	1380
Test	2547
Test	13097
Test	15218
Test	17764
Test	23583
Test	28274
Test	29387
Test	32550
Test	32955
Validation	40362
Validation	40362
Training	4471
Training	15303
Training	28961
Training	29550
Training	30594
Training	33088
Training	37785
Training	39290
Training	40845
Training	41056
Training	43589
Training	45803
Training	47448
Training	48089
Training	51553
Training	54239
Training	57154
Training	59029
Training	61183
Training	66237
Training	67026
Training	68375
Training	68739
Training	70856
Training	75759
Training	79504
Training	80904
Training	87201
Training	88419
Training	88686
Training	88964
Training	90289
Training	90926
Training	91238
Training	94317
Training	95978
Training	99877

scanner-wise and kVp based variability in HU values ranges.

It should be noted that each DL method had a linearly decreasing trend in the Bland-Altman plots showing that high noise magnitude regions i.e., the SD values from tissue borders and bony regions were underestimated. Discrepancy can partly be due to the spatial extent of the SD calculation window. However, in head imaging, the diagnostic interest is mainly in the soft tissue regions of GM and WM except for potential fractures and other specific applications. Therefore, this underestimation does not directly affect the most relevant results. Even though the direct SD estimation using CNN had the smallest error in our phantom dataset it may not be as flexible as the other methods involving a preceding step of determining the noise image as the kernel size has to be fixed prior to the training for the ground truth label. Furthermore, as compared to other processing methods, the computation of the SD map using CNN is a slow process because the whole image stack must be processed with sliding windowing though the network structure.

As the clinical open dataset had a large slice thickness of 5 mm and thus was a highly anisotropic, we applied 2D SD assessment in our framework. However, we did not have this limitation with the isotropic anthropomorphic phantom data (slice thickness was 0.625 mm) but on the other hand it lacked diagnostically important anatomical WM and GM soft tissue structures. Therefore, we applied the openly available dataset in our image quality assessment framework demonstration involving noise and contrast quantification. As a beneficial reference aspect, the latter dataset is available to other researchers as well, to investigate and to benchmark their own frameworks. The discrepancy between phantom and clinical data was unfortunate, as it would have been very interesting to compare model performances between phantom and clinical data having exactly same acquisition protocols. This should be accounted in the future studies for instance using local patient databases and selecting the same imaging protocol for the phantoms used.

In manual annotations, the ROI size was kept similar for each anatomical location (GM, WM, Pons, CSF) to make comparison between anatomical regions similar. The size of the ROIs was limited to cover only ten pixels in diameter. This was chosen as the CSF diameter cerebral ventricle was narrow for test set cases limiting the size. However, the limitations of the manual assessment using labeling could be overcome by using an automated methods as demonstrated using the segCT method. However, the downside of automated segmentation is that there is a risk of anatomical borders to be included in different soft tissue masks, as brain tissue segmentation from CT images is not a trivial segmentation task especially if there are tissue pathologies present in the tissue. Therefore, future studies are warranted to investigate in more detail how automated assessment performs compared with manual labeling in the image quality assessment task.

Our results presented the mean local SD values of each region in the clinical image quality assessment framework summarising the SD distribution characteristics. The SD distribution may carry additional information which could be utilized further in image quality characterisation. These image quality characteristics could also be combined with other measures of imaging performance such as with radiation exposure monitoring data [18] serving the comprehensiveness of quality assurance, consistency of imaging quality, malfunction detection, and optimisation of CT scan protocols over various vendors and scanner models. Future studies should focus on expanding the clinical image assessment to other image quality measures, while automating the analysis process. Finally, as the image quality parameters are sometimes challenging to interpret and translate to diagnostically acceptable image quality, future studies should include data-driven models aiming to match the machine predictions with expert diagnostic quality estimates in Likert or binary scale.

Conclusions

Deep learning-based clinical image assessment from head CT images is feasible and provides acceptable accuracy as compared to true image noise. The unsupervised noise2noise approach may be feasible in clinical use where no ground truth data is available. DL-based noise estimation combined with automated tissue segmentation for contrast measurement enables more comprehensive image quality characterization. The developed method provides a promising QA and optimisation tool for head CT examinations, which represent the most common CT examination worldwide.

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.

Acknowledgements

This study was supported by HUS Diagnostic Center research funding [grant number Y780021015]. Presentations and publications shall

acknowledge grants [grant numbers EB017095, EB017185] (Cynthia McCollough, PI) from the National Institute of Biomedical Imaging and Bioengineering.

Appendix A

References

[1] Levin DC, Parker L, Rao VM. Recent Trends in Imaging Use in Hospital Settings: Implications for Future Planning. J Am Coll Radiol 2017;14:331–6. https://doi.org/10.1016/j.jacr.2016.08.025.

[2] Maxwell S, Ha NT, Bulsara MK, Doust J, Mcrobbie D, O'Leary P, et al. Increasing use of CT requested by emergency department physicians in tertiary hospitals in Western Australia 2003–2015: an analysis of linked administrative data. BMJ Open 2021;11(3):e043315.

[3] Rubin GD. Computed Tomography: Revolutionizing the Practice of Medicine for 40 Years. Radiology 2014;273:S45–74. https://doi.org/10.1148/radiol.14141356.

[4] Bly R, Järvinen H, Kaijaluoto S, Ruonala V. CONTEMPORARY COLLECTIVE EFFECTIVE DOSE TO THE POPULATION FROM X-RAY AND NUCLEAR MEDICINE EXAMINATIONS—CHANGES OVER LAST 10 YEARS IN FINLAND. Radiat Prot Dosimetry 2020;189:318–22. https://doi.org/10.1093/rpd/ncaa045.

[5] Martin CJ, Sharp PF, Sutton DG. Measurement of image quality in diagnostic radiology. Appl Radiat Isot 1999;50:21–38. https://doi.org/10.1016/S0969-8043(98)00022-0.

[6] Verdun FR, Racine D, Ott JG, Tapiovaara MJ, Toroi P, Bochud FO, et al. Image quality in CT: From physical measurements to model observers. Phys Medica 2015;31(8):823–43.

[7] Kress LV. Preface, executive summary and glossary. Ann ICRP 2007;37:9–34. https://doi.org/10.1016/j.icrp.2007.10.003.

[8] Samei E, Järvinen H, Kortesniemi M, Simantirakis G, Goh C, Wallace A, et al. Medical imaging dose optimisation from ground up: expert opinion of an international summit. J Radiol Prot 2018;38(3):967–89.

[9] Hernandez-Giron I, Calzado A, Geleijns J, Joemai RMS, Veldkamp WJH. Comparison between human and model observer performance in low-contrast detection tasks in CT images: application to images reconstructed with filtered back projection and iterative algorithms. Br J Radiol 2014;87:20140014. https://doi.org/10.1259/bjr.20140014.

[10] Abadi E, Sanders J, Samei E. Patient-specific quantification of image quality: An automated technique for measuring the distribution of organ Hounsfield units in clinical chest CT images. Med Phys 2017;44:4736–46. https://doi.org/10.1002/mp.12438.

[11] Peltonen JI, Mäkelä T, Salli E. MRI quality assurance based on 3D FLAIR brain images. Magn Reson Mater Physics, Biol Med 2018;31:689–99. https://doi.org/10.1007/s10334-018-0699-3.

[12] Cheng Y, Abadi E, Smith TB, Ria F, Meyer M, Marin D, et al. Validation of algorithmic CT image quality metrics with preferences of radiologists. Med Phys 2019;46(11):4837–46.

[13] Cheng Y, Smith TB, Jensen CT, Liu X, Samei E. Correlation of Algorithmic and Visual Assessment of Lesion Detection in Clinical Images. Acad Radiol 2020;27:847–55. https://doi.org/10.1016/j.acra.2019.07.015.

[14] Manco L, Maffei N, Strolin S, Vichi S, Bottazzi L, Strigari L. Basic of machine learning and deep learning in imaging for medical physicists. Phys Medica 2021;83:194–205. https://doi.org/10.1016/j.ejmp.2021.03.026.

[15] Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer 2018;18:500–10. https://doi.org/10.1038/s41568-018-0016-5.

[16] Kretz T, Mueller K-R, Schaeffter T, Elster C. Mammography image quality assurance using deep learning. IEEE Trans Biomed Eng 2020;67:3317–26. https://doi.org/10.1109/TBME.2020.2983539.

[17] Sreekumari A, Shanbhag D, Yeo D, Foo T, Pilitsis J, Polzin J, et al. A deep learning-based approach to reduce rescan and recall rates in clinical MRI examinations. Am J Neuroradiol 2019;40(2):217–23.

[18] Christianson O, Winslow J, Frush DP, Samei E. Automated technique to measure noise in clinical CT examinations. Am J Roentgenol 2015;205:W93–9. https://doi.org/10.2214/AJR.14.13613.

[19] Kaasalainen T, Palmu K, Lampinen A, Kortesniemi M. Effect of vertical positioning on organ dose, image noise and contrast in pediatric chest CT—phantom study. Pediatr Radiol 2013;43:673–84. https://doi.org/10.1007/s00247-012-2611-z.

[20] Tian X, Samei E. Accurate assessment and prediction of noise in clinical CT images. Med Phys 2015;43:475–82. https://doi.org/10.1118/1.4938588.

[21] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation, 2015, p. 234–41. 10.1007/978-3-319-24574-4_28.

[22] Zhang K, Zuo W, Chen Y, Meng D, Zhang L. Beyond a gaussian denoiser: residual learning of deep CNN for image denoising. IEEE Trans Image Process 2017;26:3142–55. https://doi.org/10.1109/TIP.2017.2662206.

[23] Lehtinen J, Munkberg J, Hasselgren J, Laine S, Karras T, Aittala M, et al. Noise2Noise: Learning Image Restoration without Clean Data. 35th Int Conf Mach Learn ICML 2018 2018;7:4620–31.

[24] Loshchilov I, Hutter F. SGDR: Stochastic gradient descent with warm restarts. 5th Int Conf Learn Represent ICLR 2017 - Conf Track Proc 2017:1–16.

[25] Moen TR, Chen B, Holmes DR, Duan X, Yu Z, Yu L, et al. Low-dose CT image and projection dataset. Med Phys 2021;48(2):902–11.

[26] McCollough CH, Chen B, Holmes D, Duan X, Yu Z, Yu L, et al. Data from Low Dose CT Image and Projection Data, 2020. 10.7937/9npb-2637.

[27] Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The cancer imaging archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 2013;26(6):1045–57.

[28] Brudfors M, Balbastre Y, Flandin G, Nachev P, Ashburner J. Flexible Bayesian Modelling for Nonlinear Image Registration 2020:1–13. 10.1007/978-3-030-59716-0_25.

[29] Gong H, Yu L, Leng S, Dilger SK, Ren L, Zhou W, et al. A deep learning- and partial least square regression-based model observer for a low-contrast lesion detection task in CT. Med Phys 2019;46(5):2052–63.

[30] Singh R, Wu W, Wang G, Kalra MK. Artificial intelligence in image reconstruction: the change is here. Phys Medica 2020;79:113–25. https://doi.org/10.1016/j.ejmp.2020.11.012.

[31] Malkus A, Szczykutowicz TP. A method to extract image noise level from patient images in CT. Med Phys 2017;44:2173–84. https://doi.org/10.1002/mp.12240.