Physica Medica xxx (xxxx) 1–13

Contents lists available at ScienceDirect

Physica Medica

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

Addressing signal alterations induced in CT images by deep learning processing: A preliminary phantom study

Sandra Doria[a,b], Federico Valeri[c,d], Lorenzo Lasagni[c,d], Valentina Sanguineti[e,f], Ruggero Ragonesi[e,f], Muhammad Usman Akbar[e,f], Alessio Gnerucci[c,d], Alessio Del Bue[g], Alessandro Marconi[c], Guido Risaliti[c], Mauro Grigioni[h], Vittorio Miele[i], Diego Sona[e,j], Evaristo Cisbani[h,*], Cesare Gori[c,k], Adriana Taddeucci[l]

[a] Istituto di Chimica dei Composti Organo Metallici, Consiglio Nazionale delle Ricerche, Florence, Italy
[b] European Laboratory For Non Linear Spectroscopy, Università degli Studi di Firenze, Florence, Italy
[c] Dipartimento di Fisica e Astronomia, Università degli Studi di Firenze, Florence, Italy
[d] Scuola di Scienze della Salute Umana, Università degli Studi di Firenze, Florence, Italy
[e] Pattern Analysis & Computer Vision, Istituto Italiano di Tecnologia, Genoa, Italy
[f] Dipartimento di Ingegneria Navale, Elettrica, Elettronica e delle Telecomunicazioni, Università degli Studi di Genova, Genoa, Italy
[g] Visual Geometry and Modelling, Istituto Italiano di Tecnologia, Genoa, Italy
[h] Istituto Superiore di Sanità, Centro Nazionale Tecnologie Innvative in Sanità Pubblica, Rome, Italy
[i] Radiodiagnostica di Emergenza-Urgenza, Azienda Ospedaliero-Universitaria Careggi, Florence, Italy
[j] Fondazione Bruno Kessler, Trento, Italy
[k] Istituto Nazionale di Fisica Nucleare - Sezione di Firenze, Sesto Fiorentino, Florence, Italy
[l] Unità Operativa di Fisica Sanitaria, Azienda Ospedaliero-Universitaria Careggi, Florence, Italy

ARTICLE INFO

Keywords:
Artificial intelligence
Convolutional neural network
Computed tomography
Image quality
Radiomic features

ABSTRACT

Purpose: We investigate, by an extensive quality evaluation approach, performances and potential side effects introduced in CT images by Deep Learning (DL) processing.
Method: We selected two relevant processing steps, denoise and segmentation, implemented by two Convolutional Neural Networks (CNNs) models based on autoencoder architecture (encoder-decoder and UNet) and trained for the two tasks. In order to limit the number of uncontrolled variables, we designed a phantom containing cylindrical inserts of different sizes, filled with iodinated contrast media. A large CT image dataset was collected at different acquisition settings and two reconstruction algorithms. We characterized the CNNs behavior using metrics from the signal detection theory, radiological and conventional image quality parameters, and finally unconventional radiomic features analysis.
Results: The UNet, due to the deeper architecture complexity, outperformed the shallower encoder-decoder in terms of conventional quality parameters and preserved spatial resolution. We also studied how the CNNs modify the noise texture by using radiomic analysis, identifying sensitive and insensitive features to the denoise processing.
Conclusions: The proposed evaluation approach proved effective to accurately analyze and quantify the differences in CNNs behavior, in particular with regard to the alterations introduced in the processed images. Our results suggest that even a deeper and more complex network, which achieves good performances, is not necessarily a better network because it can modify texture features in an unwanted way.

1. Introduction

In the last decades Computed Tomography (CT) applications have gained a fundamental role in the field of diagnostic imaging [1,2], generating an increasing demand for visual examination by the radiologist. This need, together with the evolution of Artificial Intelligence (AI) technologies supported by the increasing availability of computational resources, has driven the scientific research towards the development of AI-based tools. In particular, various Deep Learning (DL) models have been designed by researchers for relevant tasks applied to CT

* Corresponding author.
E-mail address: evaristo.cisbani@iss.it (E. Cisbani).

https://doi.org/10.1016/j.ejmp.2021.02.022
Received 18 December 2020; Received in revised form 19 February 2021; Accepted 23 February 2021

images and have been largely documented in literature because of their potential support to the radiologist staff in the diagnosis from clinical images [3–9]. Indeed, Convolutional Neural Networks (CNNs) have shown remarkable effectiveness in several tasks, such as automatic localization and segmentation of low contrast objects and denoising of clinical images [10,5,11,6,7,12–14,3,15–17].

Many studies have evaluated the impact of DL algorithms on image quality by means of conventional image analysis, including estimation of signal to noise ratio (SNR), noise power spectrum (NPS) and modulation transfer function (MTF) ([18–29]). According to these image quality estimators, DL algorithms have shown the ability to decrease noise and remove related artifacts [30], even if the preservation of a good spatial resolution is still under debate ([19,21,24,26,27]). However, beyond the surprisingly good results in many domains, DL tools are black-boxes, which interpretation remains still obscure. As a consequence the images resulting from DL processing might present subtle alterations that might impact further processing steps. As a matter of facts, the nature and the extent of possible alterations of the information content in output images after the CNNs processing remains mostly unexplored.

The aim of this work is, therefore, to address the problem of possible alterations induced by DL processing methods on CT images. We performed such investigation by means of simplified images obtained from a specially designed homogeneous PMMA phantom with inserts of various diameters and contrasts. Indeed, one strength of this research consists in having collected a large amount of labeled CT images necessary to train and test the CNNs in highly controlled conditions.

We focused the analysis on two tasks: denoising and object segmentation. Indeed these two tasks are exemplar of the two main issues in the field of image evaluation: detectability preservation with increasing noise and areas identification in low contrast environment. All experiments were based on the adoption of two CNN architectures with different complexity: a standard encoder-decoder and its extension, the more powerful UNet model, largely applied in the field of medical imaging for segmentation and denoising tasks [31–47]. The choice of using these two models was guided by the intention to study how the increasing computational power affects the structure of information in the processed images. In addition, since there are some works reporting increased performance of CNNs addressing multi-tasks learning [48,49], we explored this solution modifying the models is such a way to solve both denoising and segmentation in one shot.

CNNs outputs were assessed through various metrics. For the segmentation results the Dice Similarity Coefficient (DSC) and the area under Receiving Operating Characteristic (ROC) curves [50–57] were adopted. Denoising results were assessed by means of conventional metrics (SNR, NPS, MTF). In addition, since radiomics is attracting interest because of its descriptive power [58–60], resulting images were also assessed by means of relevant radiomics features of the first and second order [61]. In this analysis we exploited the radiomics features with a novel perspective, considering them as "radiomics properties" characterizing the information structure. The comparison of images in terms of these "radiomics properties" allowed us to identify the "sensitive features", i.e. the features affected by the CNN denoising process, hence, more suitable for analyzing the alterations introduced by such process. This analysis allowed to better characterize the image alterations, adding further information to the one arising from conventional metrics. On the other hand, the same comparison allowed us to identify the "robust features", i.e. those insensitive to the CNN denoise process.

The investigation procedure we adopted resulted efficient enough to accurately detect and quantify the differences in CNNs behavior both for the attained result and for the alterations introduced in the processed images. Its sensitivity resulted also adequate to properly quantify the differences in noise shape associated to the reconstruction method (FBP or IR) [62–64].

2. Materials and methods

2.1. Phantom and CT acquisitions

We designed and manufactured a PolyMethyl MethAcrylate (PMMA) phantom (Fig. 1) of ellipsoidal shape consisting of four adjacent blocks (size $31 \times 21 \times 7$ cm each); one block is homogeneous in PMMA to produce background images; two blocks contain 5 cylindrical inserts each, with increasing diameters (3 mm, 4 mm, 5 mm, 6 mm, 7 mm) and 5 cm high; the inserts were filled with iodinated contrast media at two different concentrations to obtain signal differences of 80 HU (C_1 contrast) and 70 HU (C_2 contrast) with respect to background (125 HU) at 120 kV. In the fourth block four solid cylindrical inserts of different materials (acrylic, Teflon, polyzene, polyvinylchloride) and variable diameters were placed, in order to measure low and high contrast spatial resolution.

Acquisition for CNNs training and optimization was carried out by using a 128 slice CT scanner (Somatom Definition Flash, Siemens Healthcare) and selecting the standard oncological protocol for abdomen [1]. In Fig. 2 two pictures of the acquisition setup are shown.

CT scans of the entire phantom were performed at 8 + 1 different current settings (Table 1), in order to vary the image noise and to obtain clean (HD) reference images to generate ground truth for the neural network training. Table 1 reports the volumetric CT dose index (here named CTDI) for the current acquisition settings.

Reconstruction was performed by applying two different algorithms: FBP (convolution kernel B41s) and IR (SAFIRE, strength 3, convolution kernel IF41s). For each block of the phantom, 24 slices each 2 mm thick were taken into account. In order to have only one contrast object in each image, reconstructions were performed by setting Reconstructed FoV (RFoV) equal to 5×5 cm^2. Fig. 3 shows an image of the entire phantom and two 5×5 cm^2 images at different radiation levels.

2.2. Deep learning approach

We describe here our CNN approach for CT images processing. The relative background on machine learning (ML) and deep learning (DL) can be found in S.I. material (Section S2). All the computational codes were home-made developed using Pyhton programming language.

2.2.1. Data preprocessing

In order to set up the large number of images required both to train and validate the CNNs, we reconstructed multiple independent (5×5 cm^2 RFoV) images from the CT acquisitions of the phantom by positioning the RFoV, in HD images, in correspondence of each contrast insert and then repeating the reconstruction by randomly varying the center coordinates of the RFoVs (10 times for each insert) in order to generate more images containing the contrast objects in different locations (see also Fig. S1 in S.I. - Supplementary Information - material). Reconstruction of images from the acquisitions at the other 8 CTDI values was carried out by setting the same coordinates of the RFoV centers. Having selected a slice thickness of 2 mm, we were able to reconstruct 24 images, for each insert along the z axis of the phantom with the same x,y coordinates as mentioned above, therefore totaling $10 \times 10 \times 8 \times 24 = 19200$ images containing a contrast object. Same amount of images was obtained without any contrast object, from the homogeneous block of the phantom. At the end of this step each scans contains subsequent slices where the insert is always in the same position. In order to reduce overfitting (see Deep learning background Section S2 in S.I. and Ref. [65–71]), data augmentation techniques were then applied [68] to each slice, i.e. to the subset of 24 images recon-

[1] (Helical, 120 kVp, 200 mAs ref, acquisition= 128×0.6 mm, collimation= 64×0.6 mm, pitch= 1, scan Field of View (FoV)= 50 cm, rotation time 0.5 s, CareDose 4D = on, Care kV = off)

Fig. 1. Lateral view (a) and top view (b) of one of the two blocks containing 5 inserts filled with iodinated contrast media; (c) Picture of the fourth block, containing 4 solid inserts for MTF evaluation.

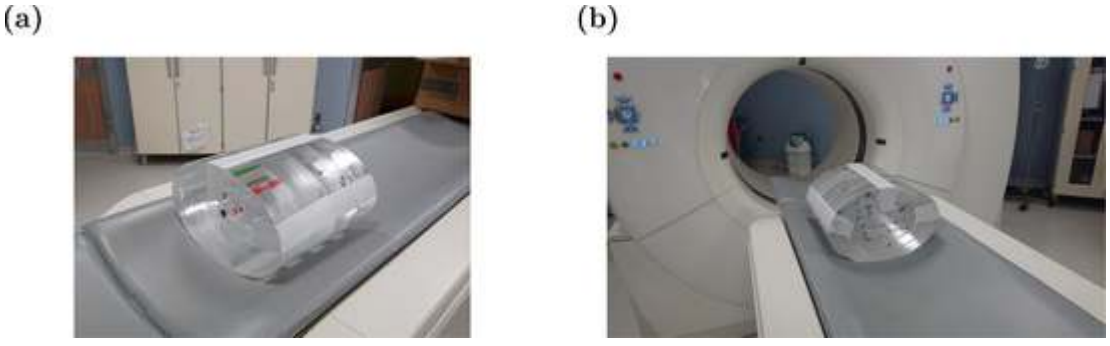

Fig. 2. Pictures of the acquisition setup in the CT room.

Table 1
Current settings for CT acquisitions, and corresponding expected image quality level (HD stands for High Dose index, the highest quality).

	Current x Rotation Time		
Quality	Reference	Average	CTDI
Level	[mAs]	[mAs]	[mGy]
1	100	64	4.4
2	120	76	5.1
3	140	89	6
4	160	102	6.9
5	180	115	7.8
6	200	128	8.6
7	220	142	9.6
8	240	154	10.2
HD	600	390	26.3

structed over the depth of the phantom (therefore with the same x, y coordinates) consisting of 90 degrees rotation and flipping. These augmentation operations add further variation on the inserts position from one slice to the other, in order to reduce the number of images containing the object in the same position. By repeating the procedure for both FBP and IR we obtained two separate dataset. We fed the neural network with the described dataset divided between train and validation data in $70:30$ ratio.

Ground truths for the denoise task were generated from the highest CTDI scan by averaging all the slices over the entire depth of the inserts (24 slices), in order to obtain very clean images with minimum noise. Binary masks for denoise metrics computation were then obtained from these ground truth images by means of a Gaussian adaptive thresholding algorithm (Functions *adaptiveThreshold()* of the *Imgproc class* of Open Computer Vision – OpenCV library [72]). Ground truths for segmentation task were obtained from such binary masks: firstly the coordinates of the center of the contrast object were determined and thereafter a binary image was generated locating in those coordinates the center of a disk with a diameter equal to the corresponding nominal one in the phantom.

2.2.2. CNNs architecture

We developed a neural network architecture to perform multiple tasks, namely segmentation and denoise of the reconstructed phantom CT images (input images). In order to get insight into the behavior of

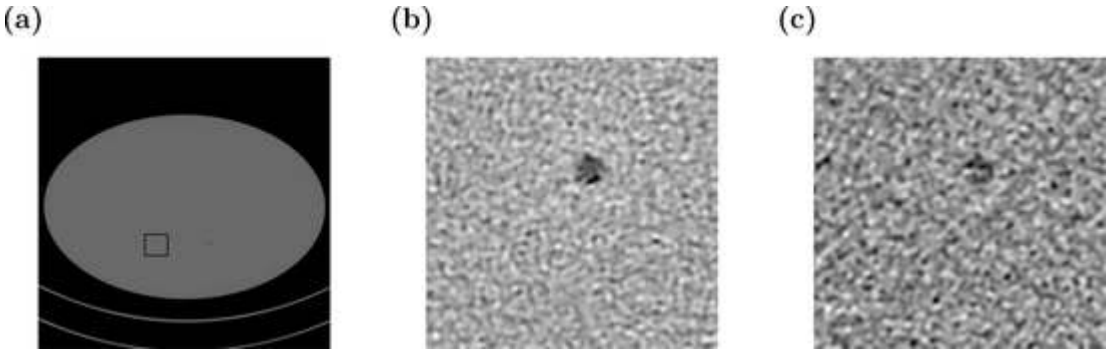

Fig. 3. a) Image (IR reconstruction) of the entire phantom acquisition (RFoV = 50×50 cm^2); b) and c) IR reconstructions around an insert (RFoV = 5×5 cm^2) of the phantom acquired at high dose index (CTDI = 26.3 mGy) and low dose index (CTDI = 4.4 mGy), respectively. The square in (a) identified the insert represented in (b) and (c).

the trained neural network in manipulating the input images, we implemented and tested different models, based on autoencoder architecture (Fig. 4), and then we evaluated their performances on the considered tasks. The achievement of the two tasks simultaneously was obtained by inserting two parallel branches (one for segmentation and one for denoise) at the end of the decoder step of the model. In this way we realized a double-task model, that recently have attracted interest in literature for the good performances reached in both the tasks trained [49]. Detailed schemes of CNN models used can be found in S.I. (Section S2).

Models implemented were:

- An encoder-decoder (Enc-Dec) consisting in 4 convolutional layers for encoding, followed by 2 fully connected layers interposed with dropout layers (dropout rate tuned to reach optimum at 0.1), after which the model splits into two branches made of three convolutional layers each, for optimization of the two tasks.
- A UNet model [31–33,47] adapted to the current tasks: a combination of max pooling, convolutional and fully connected layers for a total of 12 layers and 3 skip connections. Skip connections concatenate high resolution features produced by encoder to upsampled features of decoder to enable precise segmentation. The two branches for the separated tasks consist of three convolutional and two additional concatenation layer each.
- The UNet model was trained also separately in the two tasks of segmentation and denoise, by minimizing only the corresponding loss at one time. We will address these trained model as UNet-den and UNet-seg.

Following [73,74,33,75–82,49,83], a linear combination of two losses was implemented for the training: a *mean square error loss* and a *binary cross entropy loss* to optimize denoise and segmentation tasks, respectively. The relative weight of the two losses was tuned using validation set. Further details on loss tuning process and optimizer algorithm (*Adam*) can be found in S.I. (Section S2).

Training was carried out from scratch. Model's hyperparameters were fine-tuned using validation set. We employed Tensorflow framework [84] and GPU parallel programming to build and train the CNN models. All experiments were run on a Nvidia GeForce RTX 2080 Ti GPU with CUDA 10 support.

2.2.3. Performance metrics

We used the *Dice similarity coefficient* (DSC) to evaluate the spatial performances of the trained models in localizing the inserts and their shape. ROC (ROC) curves were computed to address the behavior of the algorithms in radiological terms. Finally, the area under the ROC curves (AUC) was computed as a function of CTDI for the two tasks to estimate the overall neural network performance by means of such a common metrics used in diagnostic imaging. Additional description of the metrics implemented can be found in S.I. (Section S2.).

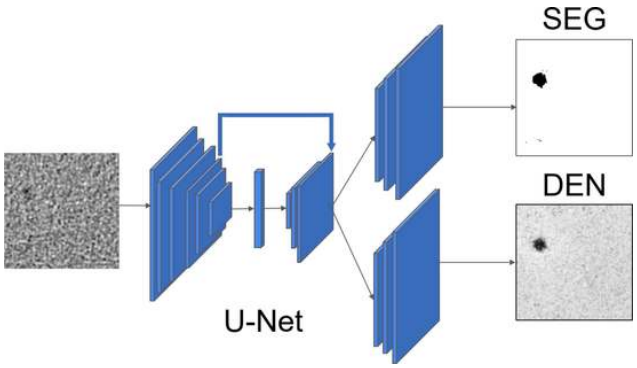

Fig. 4. A schematic version of our architecture inspired by UNet: a noisy image is given as input and model tries to reconstruct the clean image (right, down) and to highlight the position of the object (right, up).

2.3. Conventional images analysis

Conventional image quality parameters quantifying detectability (via SNR), noise frequency content (via NPS), spatial resolution (through MTF) were evaluated on original and denoised images and then compared. The ground truth images were used to automatically define the insert region as the region of interest (ROI) by means of pixel aggregation and adaptive threshold which maximize the SNR. Details on the above mentioned quality indices and their extraction procedures are reported in Section S3 of the S.I.

2.4. Radiomic features extraction

A Python based code was developed to extract features from every image using the open-source package Pyradiomics [61] and taking into account the standardization of the radiomics feature/biomarker initiative [85]. All features are computed by setting a fixed bin width equal to 25, after applying intensity normalization of the images in the preprocessing step.

For the study of the denoise process we concentrated on second order statistics features, which provide information on the texture of the examined region (see Table S1 in S.I. section for the complete list of these features). Both homogeneous images and those with insert were considered and a square ROI of fixed size was defined for each image to select a background region, not including insert; the size was equal to 200×200 pixels, the largest satisfying these conditions on all considered images.

At first, we identified the features most relevant and indicative for the characterization of noise and denoising process: these specific features were selected in terms of repeatability and sensitivity [86]. Repeatability refers to features that remain the same for the entire dataset of images acquired in the same CT acquisition conditions while the term sensitivity refers to the variability of a feature in relation to different image acquisition settings, or different reconstruction techniques.

Repeatability was quantified by means of the coefficient of variation (*CV*),*i.e.*, the ratio of the standard deviation to the absolute value of the mean, evaluated on each subset of original images at various *CTDI* values (from *4.4* to *20* mGy): the smaller the *CV* the better the repeatability. We arbitrarily assumed as repeable those features with $CV <= 15\%$ for all *CTDI* values, a threshold value halfway between 10 % and 20 % that are often chosen to categorize features into groups according to their coefficient of variation [41]. In order to investigate the sensitivity to noise of radiomic features, we plotted their average values, evaluated on the subsets of original images, as a function of *CTDI*; a linear fit was applied to the data (normalized by max) and the extrapolated slope parameter was taken as an estimate of the sensitivity to noise patterns.

We have also examined the sensitivity of each feature to different reconstruction methods, defined as the mean percentage difference between the values extracted from the original images reconstructed by the FBP and IR techniques: results show that, in general, those features sensitive to noise are also the most sensitive to the reconstruction method.

At last, all features were grouped in terms of their Spearman rank correlation [87] evaluated on the entire image dataset. Features with correlation larger than 0.9 [88] have been considered redundant, and therefore interchangeable.

The features representative of each redundant group, *i.e.* those with the highest sensitivity to noise, have been selected as candidates of sensitive characterization: *ShortRunEmphasis* and *LongRunEmphasis* are features that respectively measure the distribution of the lengths of short and long runs, *i.e.* those 1-dimensional (1D) structures of consecu-

tive pixels with the same gray level value; decrease of *ShortRunEmphasis* and/or increase of *LongRunEmphasis* are indices of coarser 1D noise texture. The *ZonePercentage* is a feature which quantifies 2-dimensional (2D) pattern and low values correspond to coarser 2D texture. *Busyness* belongs measures variations between adjacent pixels; as the *Busyness* decreases the spatial frequency of intensity changes decreases. For the mathematical definition of the above features see [61].

For each feature the percentage difference between the values extracted by denoised and corresponding original images was computed; feature varying on average by less than 10% were considered "robust" [89], *i.e.* noise insensitive.

Two unrelated shape features, *Sphericity* (that corresponds to roundness in 2D) and *MeshSurface*, have been extracted for the spatial and geometrical characterization of the denoising and segmentation processing: the former is the ratio of the perimeter of the object under analysis and the perimeter of a circle with the same surface area of the object; the latter estimates the surface area of the object [61].

3. Results

3.1. Deep learning approach

An example of the neural network output for the Enc-Dec and UNet models is shown in Fig. 5, in case of an original image, reconstructed via IR technique, containing an insert of 5 mm diameter with C_1 contrast. Different background textures are produced by the two models, and it is noticeable the better spatial accuracy of the UNet model respect to Enc-Dec in segmenting the insert, probably due to the presence of skip connections within the model architecture. Analogous examples for FBP reconstruction, in case of input images containing insert or background-only, are presented in Fig. S6 and S7 of S.I., respectively.

DSC was computed to estimate the performance of the trained models in the geometrical localization of the insert within the image. Figs. 6 a) and b) present plots of the computed DSCs as a function of CTDI, for inserts of different sizes and contrasts, in case of both denoise and segmentation tasks performed by the UNet model.

We combined signal detection theory (*i.e.* DSC metric) with ROC analysis, this latter being the preferred method to assess image quality in diagnostic radiology. AUC, computed from ROC curves, increases with CTDI (Fig. 7), with saturation at around 8 mGy and values above 0.95.

In order to better understand the CNNs behavior in addressing the two tasks and their mutual influence on the processed images, we carried out two separate training by minimizing a single loss, thus optimizing one single task at one time. Comparison between the UNet double task and single tasks (UNet-den and UNet-seg), and quantification of the superior performances of the UNet respect to the Enc-Dec, were carried out by evaluating AUC as a function of CTDI (Fig. 8).

3.2. Conventional image analysis

SNR was computed for all original images, reconstructed via FBP and IR techniques, and for the denoised images produced by the trained CNN models. The results are summarized in Fig. 9 where SNR mean values are reported as a function of CTDI.

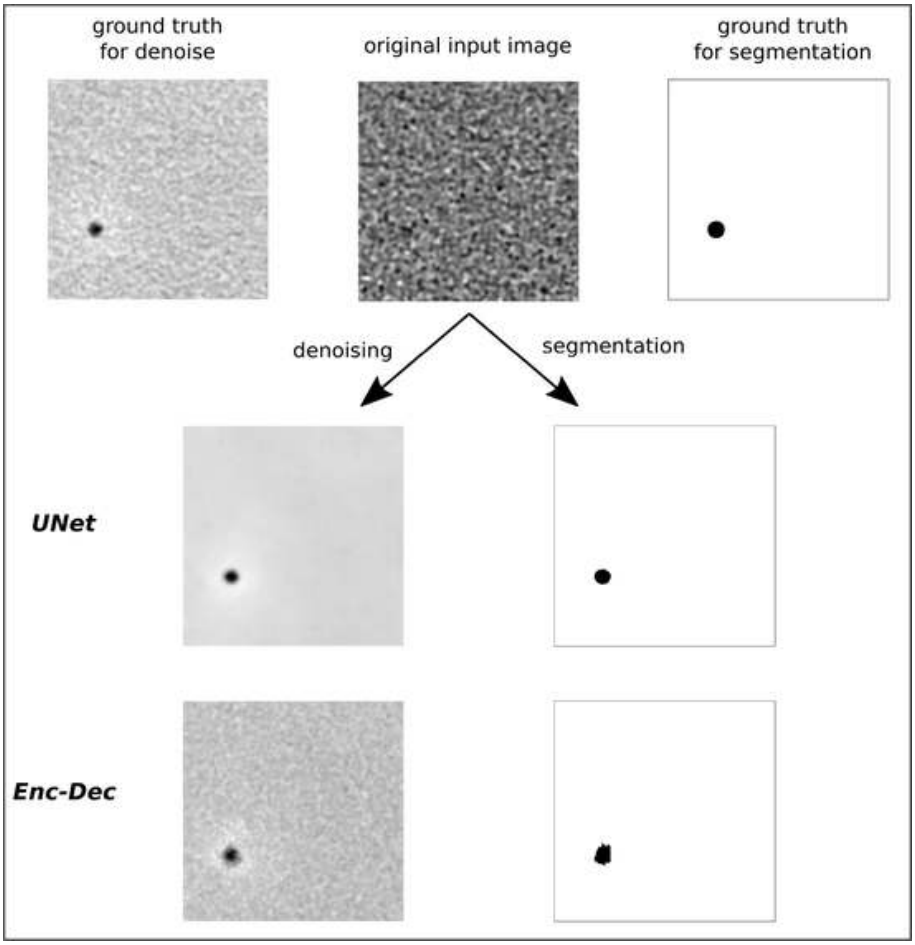

Fig. 5. Example of results from Enc-Dec and UNet trained models in denoising and segmenting the same original image (from validation dataset), reconstructed via IR technique and containing an insert of 5 mm diameter filled with C_1 contrast.

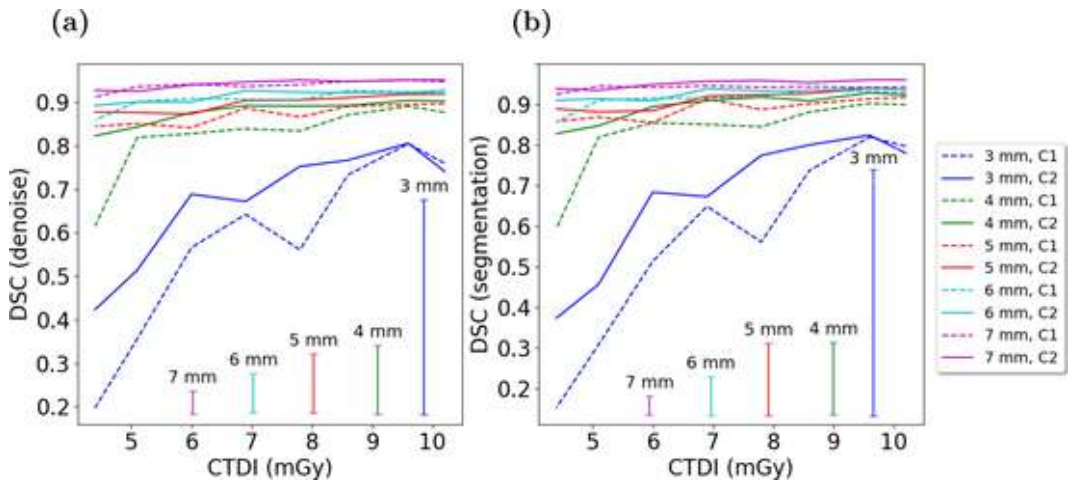

Fig. 6. DSCs as a function of CTDI, computed (a) on binary masks generated from UNet model denoised images and (b) from UNet model segmented images. Solid and dotted lines refer to inserts of C_2 and C_1 contrast respectively. Insert diameters range from 3 to 7 mm. Average error bars (equal to the standard deviation of the average) for each curve are shown in the bottom, indicating larger errors for small inserts diameters.

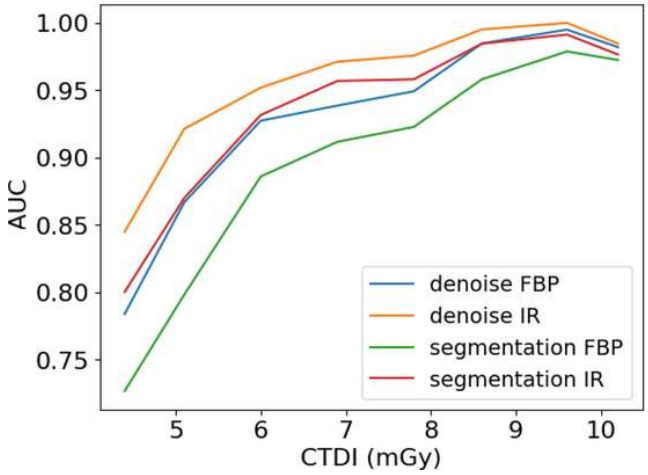

Fig. 7. Area under ROC curves for UNet segmentation and denoise, computed separately for FBP and IR reconstructions.

NPS was evaluated for original and ground truth images and for the corresponding denoised images produced by the trained CNNs, within a background region (256×256 pixels) where no insert is present. The results are shown in Fig. 10 for original images at CTDI = 4.4 mGy (no signicative difference has been found as a function of noise level).

MTFs were evaluated from the inserts in the ground truth images and from the images containing the solid inserts (no remarkable differences were found, as shown in Fig. S12 of S.I. section); we verified that the average operation to obtain ground truth from single slices didn't degrade spatial resolution (Fig. S.13 of S.I. section). These reference MTFs were then compared to those obtained from the same inserts in the denoised images produced by the UNet and UNet-den models. The MTF curves from the 7 mm diameter insert and C1 contrast are shown in Fig. 11; very similar results were obtained for all the other inserts of different size and contrast.

The circular-edge technique for the extraction of MTFs is not generally applicable to the Enc-Dec denoised images since therein inserts roundness is not well preserved, as discussed in the next section (see Fig. 15a).

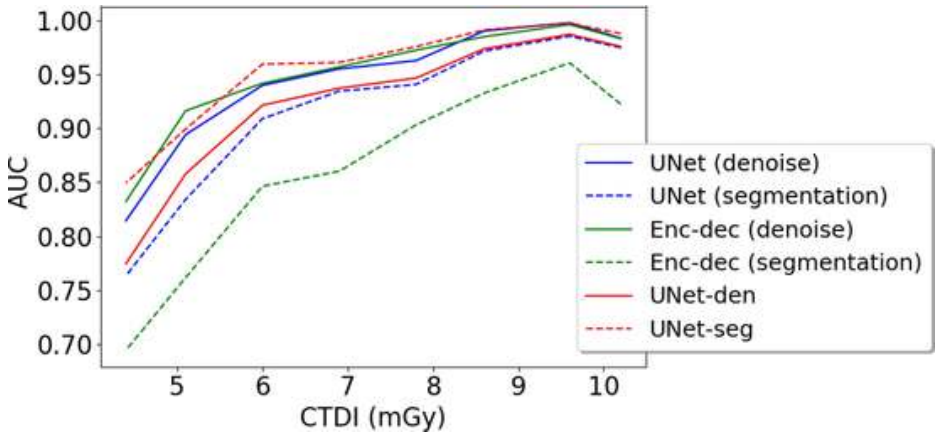

Fig. 8. AUC as a function of CTDI, computed on the images output of the trained models: UNet (blu curves) and Enc-Dec (green curves) are the double-task models, and in the parenthesis it is indicated the output images subset used to compute AUC (denoised -solid line- or segmented -dashed line- images); UNet-den and UNet-seg are the single task models, which produce, respectively, denoised (red solid line) and segmented (red dashed line) images used to compute AUC. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

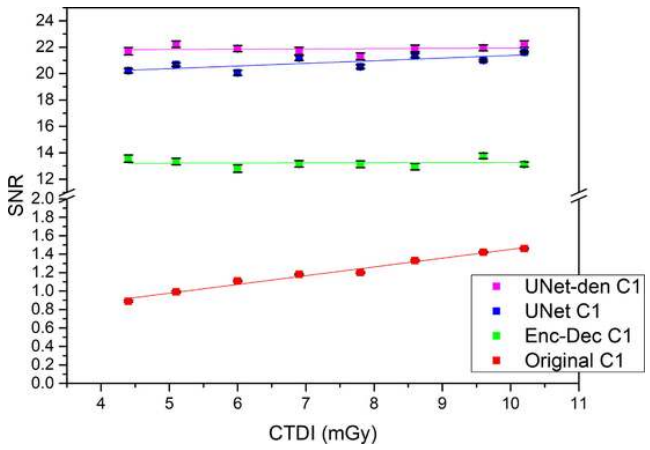

Fig. 9. SNR as a function of CTDI, calculated on original images (FBP reconstruction) containing the inserts with contrasts C1 and on denoised images produced by the different models. Lines are linear fits to emphasize the data trends. Note that a break is introduced on y axis, to better show the original's values. The errors bars, barely visible, are the Standard Deviation of the represented mean values. SNR in case of IR method is presented in Fig. S11 of S.I.

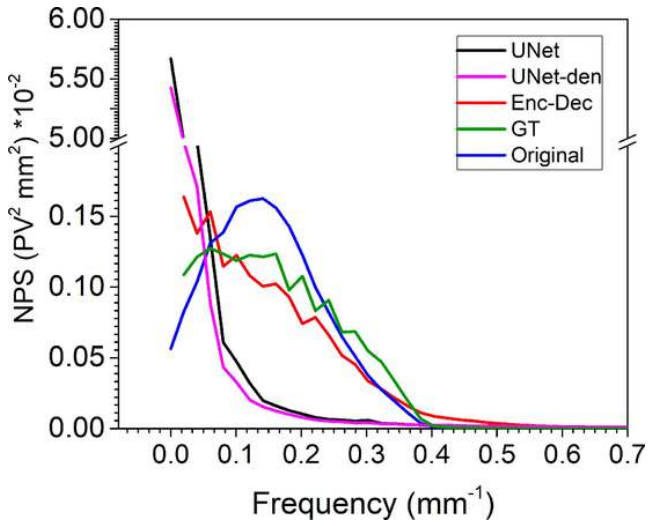

Fig. 10. NPS curves of the ground truth -GT- (green), original (blue) and denoised images produced by the trained CNNs: Enc-Dec (orange), UNet (black) and UNet-den (pink). A break on y axis is used to show the UNet and UNet-den remarkable peak near zero. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.3. Radiomic features analysis

The repeatable and sensitive radiomic features selected by mean of the procedure discussed in Section 2.4 are listed in Table 2, along with the corresponding repeatability and sensitivity indices. The distributions of the four selected features are shown in Fig. 12, in case of original images (FBP reconstruction at the lower dose index), ground truths and corresponding denoised images.

In Fig. 13 the values of *ShortRunEmphasis* feature, extracted from the denoised images and from the original images, are plotted as a function of the *CTDI*.

In Fig. 14 is reported the mean percentage difference between denoised and original images for all the radiomic features with respect to the denoising process. The values of all the standard features [85] were calculated from each subset (8 different dose indices) of input images and compared to those extracted from the corresponding CNNs

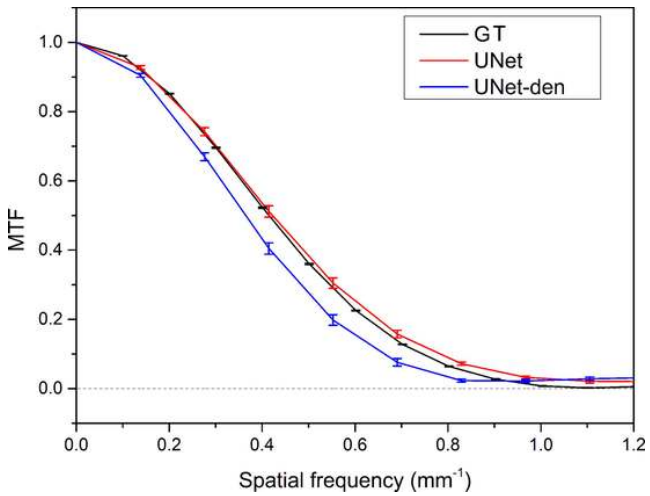

Fig. 11. MTF curves computed from the ground truth (GT) images (black) and the denoised output images of the CNN UNet (red) and UNet-den (pink). The error bars represent the Standard Deviation of the mean values. The test images used have insert with 7 mm diameter and C1 contrast. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
List of features sensitive to denoise process. The maximum correlation coefficient between these features is 0.58. (*a*) Coefficient of variation measured from the sample of ground truth images reconstructed by FBP method. (*b*) Slope parameter of the line fitted to the average values of the single feature as a function of *CTDI* for FBP method. (*c*) Mean percentage difference between FBP and IR reconstruction methods for images acquired with quality level 8.

	Repeatability	Sensitivity to noise	Sensitivity to reconstruction
	CV[(a)] (%)	Slope[(b)] ($_{10^{-2}}$)	Percentage difference[(c)] (%)
ShortRunEmphasis	5.7	1.01	11.7
LongRunEmphasis	6.4	−0.62	4.4
Busyness	6.7	0.30	−1.2
ZonePercentage	11.2	0.13	0.8

processed images. In this way the features less sensitive to noise and to the denoise process (*i.e.* more robust) were identified (Fig. 14). Different robust features were identified depending on the CNN models, based on the criterion described in 2.4 section. A summary table with the overall results is reported in S.I. (Table S2).

The distributions of the shape features, *Sphericity* and *MeshSurface*, for the different trained models are summarized in Fig. 15, while the detailed variation of the *Sphericity* respect to dose index, inserts size and contrasts for the UNet segmentation is shown in Fig. 16.

3.4. Test on dataset from a different CT scanner

The UNet model trained on Siemens dataset was finally tested in the denoise and segmentation tasks of images from a different dataset, acquired on the same phantom with a different CT scanner from Philips, to estimate the capacity of generalization of the trained model and to open new discussion about the complex interplay of factors involved in the CNNs optimization process and therefore in the quantitative evaluation of CNNs behavior. The description of the acquisition method and data preprocessing, as well as a full overview of the metrics evaluation computed on the UNet output images is reported in S.I. (Section S7).

The comparison between the performances of the trained CNN on validation (Siemens) and test (Philips) dataset, in terms of AUC, are shown in Fig. 17 as a function of CTDI.

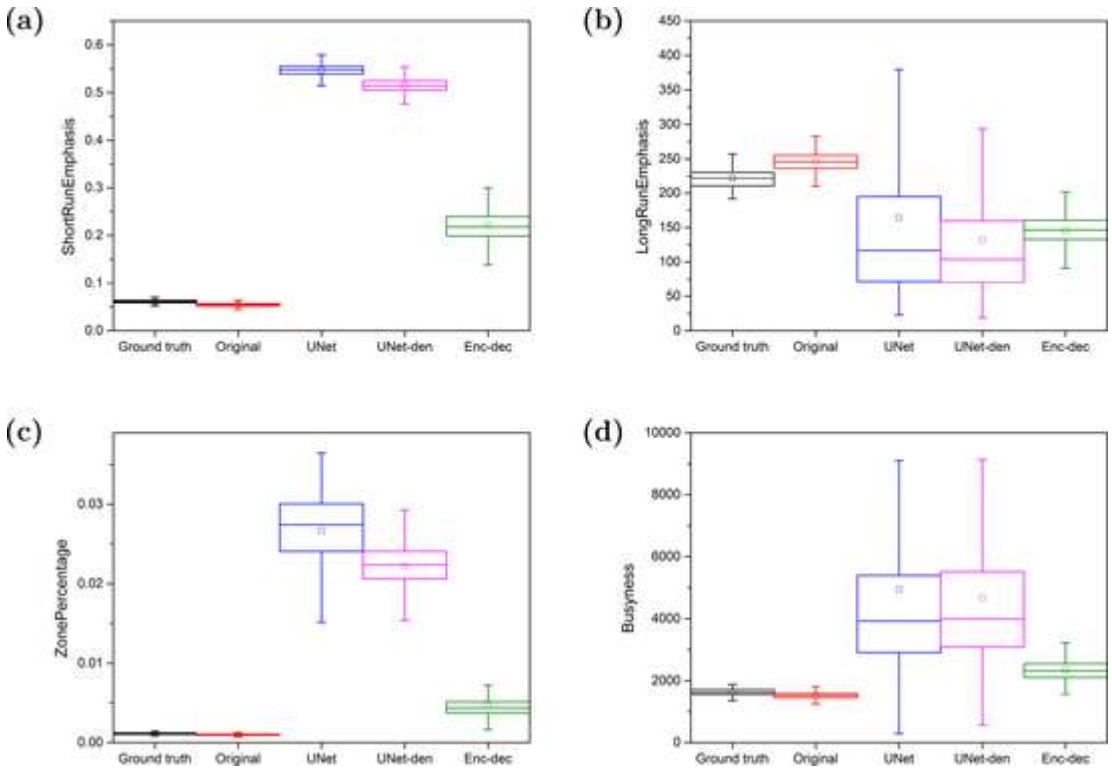

Fig. 12. Distributions of features listed in Table 2, corresponding to dose index $CTDI = 4.4$ mGy and reconstruction method FBP. The line spanning the full width of the box corresponds to the median of the distribution while the square box extends from 25 to 75 percentile values and the vertical bars represent the range of distribution not including outliers (matplotlib.pyplot.boxplot function) [90].

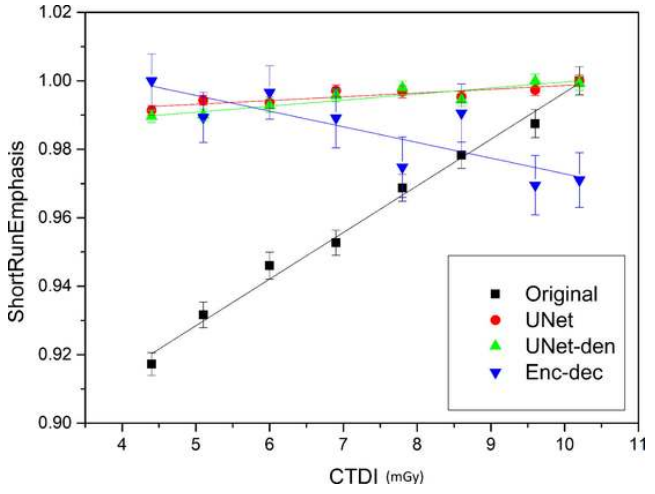

Fig. 13. ShortRunEmphasis feature mean values versus the *CTDI*, normalized to the maximum, from original and denoised images. Lines are fits of the corresponding data. The error bars represent the Standard Deviation of the mean values.

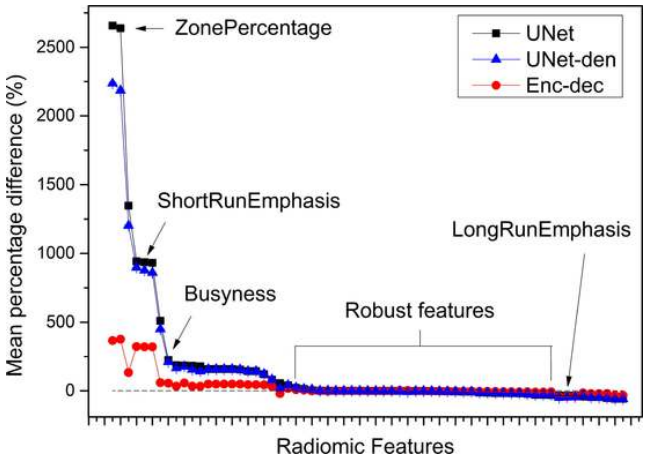

Fig. 14. Mean percentage difference between denoised and original images for all radiomic features and for three neural networks. The values are sorted in decreasing order on UNet. The four features displayed correspond to those listed in Table 2 and the robust features are listed in Table S2 in S.I.

4. Discussion

4.1. CNNs characterization

The results of the deep learning experiments were very promising, in terms of both the capacity of trained CNN models to correctly locate the contrast objects within the background pattern, and the quality of the denoised images produced. As it can be inferred by looking at Fig. 5, UNet model learned to discriminate the presence of the inserts even in condition of high noise level, where by eye it is difficult to identify the edge of the object, or even to recognize the presence of the object itself.

Different alteration of the background region are observed between the denoised imaged produced by Enc-Dec and UNet model: the Enc-Dec seems to partially preserve the noise texture of the ground truth images, while the UNet model strongly flatten the homogeneous region around the insert. These alterations have been attributed to the mean square error loss function [33,75,76,78,80,82]. However, we found substantial difference between the two models, and we believe that beyond spatial performances, also this kind of alterations must be taken under consideration when dealing with CNNs image processing, and should influence the choice of the model. Therefore we can state that even a complex network, which achieves good performances, is not necessary a better network because it can modify texture features in an

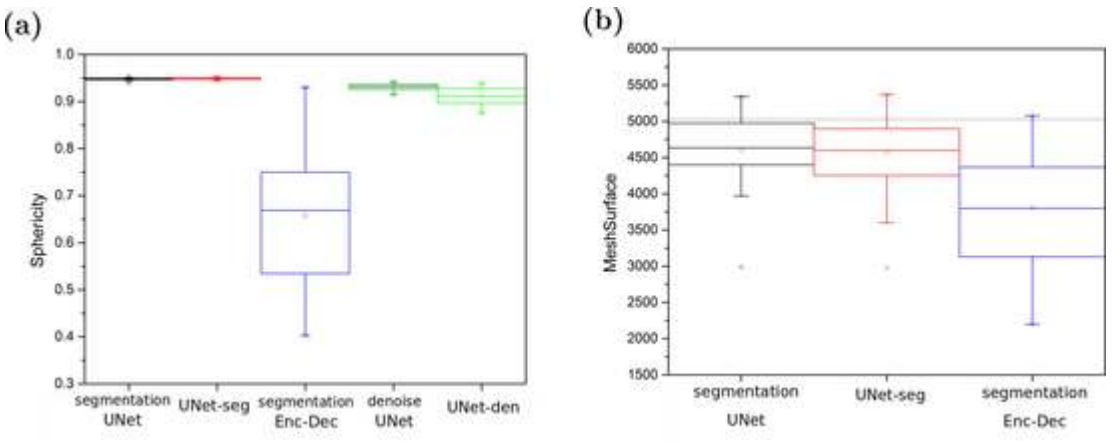

Fig. 15. Distributions of *Sphericity* (a) and *MeshSurface* (b) features extracted from segmented images, corresponding to *CTDI* = 4.4 mGy and 7 mm diameter insert with C1 contrast, FBP reconstruction: comparison among different models output. The horizontal dashed line in panel (*b*) represents the *MeshSurface* value computed on the ground truth images for segmentation task. See fig. 12 for the graphical representation of the values.

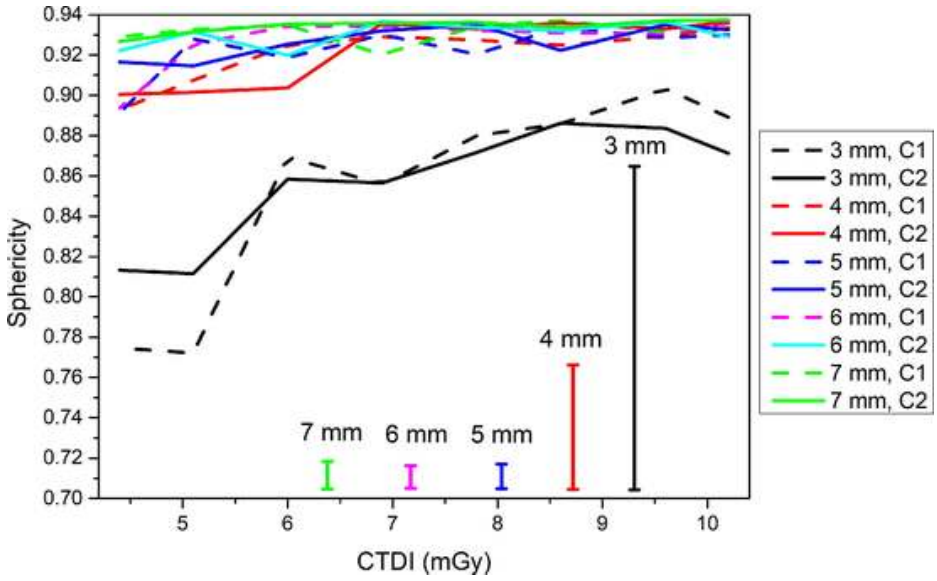

Fig. 16. *Sphericity* as a function of noise from trained UNet model, computed on segmented images. Solid lines indicate inserts of C2 contrast of different diameters while dotted lines indicates inserts of C1 contrast. Average error bars (equal to the standard deviation of the average) for each curve are shown in the bottom, indicating larger errors for small inserts diameters.

unwanted way. A quantitative explanation of this result is found in the NPS evaluation and in the radiomic features analysis reported below.

AUC as a function of CTDI (FIg. 7) in case of UNet model show, as expected, increased performances as a function of CTDI, with saturation trend observed above 8 mGy. It is worth noticing that slightly better performance were obtained with IR reconstruction technique rather than with FBP. This last results is of particular interest considering that the quantification of CT image quality has become a non-trivial issue after the introduction of IR techniques, which alter the images in non-linear way and have shown noise dependent spatial resolution ([91–95]) respect to the more traditional FBP technique.

Comparison between different models allows to understand the influence of the architectures on their behavior and to address the contribution of the model layers in the training performance. Even if the autoencoder scheme is the starting point for both the Enc-Dec and the UNet model (see schemes in Fig. S2 of S.I. section), in the first case the low variability of the layers types (only convolutional and fully connected) and the small number of layers were the main reasons for the lower performance observed. Furthermore, the Enc-Dec suffered from overfitting, and its generalization ability was not satisfactory (Enc-Dec losses trend in Fig. S4 in S.I.).

On the contrary, in case of the UNet model, the greater depth (*i.e.* the large number of concatenated layers) and the presence of multiple skip connections, which prevent the losses of fundamental information during the encoding step, give rise to enhanced performances (see metrics comparison in Fig. 8) and and reduce the overfitting issue (UNet losses trend in Fig. S3 in S.I.). The trained UNet model not only locates the insert correctly in most cases, but also associates the correct diameter and shape, as deductible from Fig. 6, where DSC scores are plotted as a function of CTDI and confirmed by the *Sphericity* feature analysis in Fig. 16. DSC curves also indicate, as expected, that the trained neural network performs better in case of low noise and for images containing inserts with larger diameters and more pronounced contrast.

The experiments performed by training one task at a time (either denoise or segmentation), by minimizing the two losses separately, allows to evaluate the reciprocal influence of the two branches during training. The comparison of AUC (Fig. 8) shows that denoise performances of the UNet model, when optimized for both tasks, are slightly superior respect to the model trained for the denoise task (UNet-den) alone. On

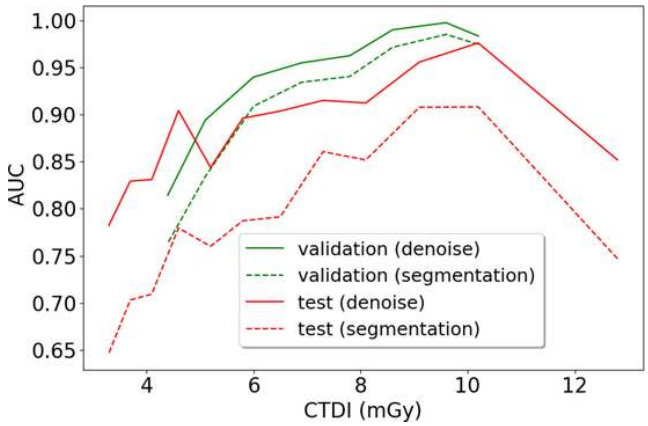

Fig. 17. AUC as a function of CTDI computed to compare the performances of the UNet model in the processing of images from two different datasets. One (acquired on CT scanner of Siemens vendor - green lines) used for validation after training, and a second one (acquired in a CT scanner from Philips vendor - red lines) used only to test the domain generalization. The plots indicate the performance in the denoise (solid line) and segmentation (dashed line) tasks. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the contrary, segmentation performances improve with the single task training: this means that in our experiments segmentation task is useful to improve denoise but denoise doesn't help segmentation, contrary to previous studies ([49,96]). The reciprocal reinforcement of the two tasks is confirmed by the evaluation, shown in Fig. S10 of S.I., of the performances of the UNet-den and UNet-seg in the task for which they were not trained, *i.e.* denoise for UNet-seg and segmentation for UNet-den.

The tuning of the relative weights seem to confirm the previous result. The evaluated AUC as a function of $\mathscr{L}_{CE}$ (cross entropy loss for segmentation) weights (Fig. S4 of S.I.) suggests that the two tasks are correlated: the denoise performances indeed are strongly quenched not only at large $\mathscr{L}_{CE}$ weight, which is expected considering that $\mathscr{L}_{MSE}$ (mean square error loss for denoise) becomes negligible, but also at very low $\mathscr{L}_{CE}$ weights, indicating that the optimization of model weights to improve segmentation is useful also to achieve better denoise performance.

Interesting results were found when testing the deep learning algorithm on a dataset of images acquired by a different CT scanner. The metrics evaluated, reported in Section S7 of S.I., evidence DSC score mostly above 50%, except the smaller inserts (3 mm diameter), where the model lacks of spatial identification ability. AUC values are also above 50%. This is a surprisingly good result when considering the lack of specific training of the CNN. It is noticeable that CNN performances increase with CTDI until 10 mGy (the largest CTDI of the training dataset), then they start to decrease even if the larger SNR should facilitate the inserts detection, possibly because no training at all was performed in that range of CTDIs. This behavior confirms a well-known generalization issue associated to CNNs: their optimization involve a combination of multiple factors. Disentanglement of such factors is a complex subject that is often overcome by increasing the dataset variability in the training step.

In order to perform a proper comparison between the performances of the trained CNN on such different datasets, several factors should be taken into account. First of all, SNR, noise intensity and texture are influenced not only by CTDI, but also by the reconstruction algorithm, proper of each CT scanner, and by the concentrations of the iodinated contrast media. The complexity of the generalization aspects requires a dedicated, extended, investigation which is outside the goal of the present paper.

4.2. Quality evaluation of the CNNs performance by means of conventional metrics

The results shown in Fig. 9 demonstrate the very high performance of the CNNs in reducing the images noise. Actually in the processed images the SNR increases by factor 20–30 respect to the original ones, well above $SNR = 5$, the reference value of the Rose detectability criterion [97] and therefore providing a quantitative explanation of the sharp visual detectability of the low SNR insert shown in Fig. 5. While the SNRs computed on original images increase with CTDI and contrast, as expected, in almost all the denoised images, the SNRs are essentially independent from CTDI and contrast. This can be considered a further valuable performance of the CNNs. Both UNet and UNet-den perform better than the Encoder-Decoder, the latter showing however a little lower dependence from the contrast of the inserts. The SNR of the single task UNet-den slightly overtakes the one of the combined denoise-segmentation UNet.

The NPS curves (Fig. 10) show the expected shape when computed on original images. A very high zero-frequency peak is clearly visible in the NPS obtained from the UNet and UNet-den output images, while the ones obtained from Enc-Dec output images (as well as from ground truth images) show a smaller zero peak according to its weaker capability of reducing noise (see also Fig. 5 and Fig. 12). This result is in agreement with the already stated behavior of the considered CNNs.

UNet denoised images provide an MTF curve almost identical to the ground truth images unlike common noise reduction methods that can result in a loss of spatial resolution (e.g. a typical smoother kernel) [98–102]. This ability to preserve spatial resolution has already been observed in case of UNet-based models and it has been attributed to the presence of global skip connections ([33]), that we have employed in our UNet model as well. Despite Ref. [33], we avoided using max pooling not to lose details, implementing instead stride in convolutions to reduce feature maps size in the contracting path. In the expanding path features of contracting path are concatenated. This improves reconstruction, mostly the concatenation of first layers features to the last ones, that has been shown to preserve in a better way spatial resolution.

In the single task UNet-den processed images the spatial resolution is slightly negatively affected, confirming that denoise can benefit from the presence of the simultaneous segmentation task. For the Enc-Dec the calculation of spatial resolution is obfuscated by the noticeable geometrical distortion of the edge of the inserts as previously discussed.

4.3. Quality evaluation of the CNNs performance by means of radiomic features

The 1D and 2D texture features, listed in Table 2 and reported in Fig. 12, show that for all the trained models the noise reduction process does not produce a texture similar to that of the ground truth images (especially for the UNet models). In addition, such alteration is dependent on the CNN model.

The values of the four selected features computed on the denoised images suggest a finer noise texture and a greater spatial frequency of intensity changes, respect to the ground truth images.

Furthermore, the distributions of the features values obtained from the denoised images are much wider than those obtained from the ground truth and original images, suggesting that CNNs produce variability in the properties of the noise texture.

The denoising process of Enc-Dec is very different from that of the UNet, since alteration of the texture is less pronounced and the values of all the estimated features are more similar to the ground truth images respect to UNet (see Fig. 14). On the other hand, UNet greatly reduces noise by producing a much more homogeneous texture: this result reflect the NPS analysis and it is apparent also by looking at Fig. 5. This characteristic should be taken under consideration when dealing with

the applications of CNNs processing to clinical images: such remarkable alteration of the background pattern could produce unusual images for the experienced eye of the clinical staff which may compromise their interpretation and therefore may not be easily accepted.

UNet and UNet-den behave very similarly to each other, with a slightly more pronounced alteration of the texture for the first model, as shown in Fig. 12 for *ShortRunEmphasis* and *ZonePercentage* features.

The dependence of the *ShortRunEmphasis* feature on *CTDI* (Fig. 13) is less significant for CNNs denoised images respect to original images: this suggests that the background noise pattern in the denoised images remains quite the same regardless of the dose index of input images, in analogy with the trend found for the SNR (Fig. 9).

The features that are not altered by the CNNs images processing, identified by the robustness analysis reported in Fig. 14, should be taken under consideration for a potential clinical application of the AI algorithms: those features are indeed the best suitable for radiomic texture investigation of CT images under different acquisition conditions.

The distributions of the geometrical features (Fig. 15) suggest that both denoise and segmentation processes alter the shape of the contrast object. The Enc-Dec neural network alters the shape of the inserts more than the UNet does, by decreasing their area and by degrading their roundness, which makes it difficult to evaluate the MTF curve, as already discussed in Section 3.3. No significant performance differences were observed between the UNet and the UNet-den, that presents slightly lower performances at high noise level.

The *Sphericity* feature, computed on UNet segmented images and plotted in Fig. 16 as a function of CTDI for different inserts, indicates that for small objects diameters less than 4 mm the performances significantly worsen.

5. Conclusions

We developed, trained and tested two CNNs, namely a standard encoder-decoder and its extension, the Unet, in the tasks of segmentation and denoise of simple CT images obtained by scanning a specifically designed PMMA phantom, with the aim of determine in a quantitative way the different behavior of the two models and, at the same time, the different textures they induced into the processed images.

The alterations produced by the CNN algorithms on the original images was thoroughly studied by combining deep learning metrics, such as Dice Similarity Coefficient (DSC) and area under Receiving Operating Characteristic (ROC) curves, conventional metrics (SNR, NPS, MTF) and radiomics features.

As expected UNet attains superior results in the segmentation task however in the denoise task, in spite of an apparent much higher definition of the test object, it introduced also evident alterations of the background, whereas the enc-dec was able to reproduce the ground truth more faithfully.

It is worth noticing that when ideally extrapolating this proposition towards the clinical applications of images denoised via CNNs, an alert occurs, that is to pay attention to the selection of the appropriate CNN not to introduce heavy alterations into the noise texture that could mislead the proper judgment of the radiologist.

The highly controlled conditions of our experiment allowed us to give some reliable contribution to the sometimes debated problem of the possible reciprocal influence of the model branches, each designed for one individual tasks, when combined them together for multi-tasks learning. We reported a slight increase in denoise performance, contrary to what previously reported [49,96], suggesting that there is still space for additional experiments in this field. In any case, either in the single or in the multi-task learning, images denoised by means of the UNet model preserved the original spatial resolution, despite the trade-off between denoise and spatial resolution commonly reported in literature [98–102].

The evaluation approach we adopted resulted effective, even if in simplified images, to accurately detect and quantify the differences in CNNs behavior for both the attained result of the tasks and for the alterations introduced in the processed images.

On the other hand by using the radiomic approach we were able to select a number of robust features, i.e. features insensitive to the CNN images processing with both models. For this reason they constitute potential candidates for a conventional radiomic analysis of images processed with diverse CNNs.

The extremely good results of the UNet model in the segmentation task of the CT images of the phantom persuaded us about the actual feasibility of a CNN based model observer [103,27,104–108,57,29] behaving like a human observer for the highly needed task of optimizing (detectability vs patient dose) current CT protocols.

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

The authors acknowledge the support and the CT resources provided by the Radiology department of the Careggi University Hospital in Florence, Italy.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.ejmp.2021.02.022.

References

[1] European Commission. Medical Radiation Exposure of the European Population. Rad Prot 2015;180.
[2] International Commission On Radiological Protection (ICRP). ICRP PUBLICATION 26: 1977 Recommendations of the International Commission on Radiological Protection. Ann ICRP 1977;26(1(3)).
[3] Causey J, Guan Y, Dong W, Walker K, Qualls J, Prior F, et al. Lung cancer screening with low-dose CT scans using a deep learning approach. arXiv: 190600240 2019;.
[4] Wong KK, Fortino G, Abbott D Deep learning-based cardiovascular image diagnosis: A promising challenge. Future Gener Comp Sy 2020;110:802–11. https://doi.org/10.1016/j.future.2019.09.047.
[5] Thrall JH, Li X, Li Q, Cruz C, Do S, Dreyer K, et al. Artificial intelligence and machine learning in radiology: Opportunities, challenges, pitfalls, and criteria for success. J Am Coll Radiol 2018;15(3, Part B):504–508. doi: 10.1016/j.jacr.2017.12.026.
[6] Mamoshina P, Vieira A, Putin E, Zhavoronkov A Applications of deep learning in biomedicine. Mol Pharm 2016;13(5):1445–54. https://doi.org/10.1021/acs.molpharmaceut.5b00982.
[7] Johnson KW, Torres SJ, Glicksberg BS, Shameer K, Miotto R, Ali M, et al. Artificial intelligence in cardiology. J Am Coll Cardiol 2018;71(23): 2668–79. https://doi.org/10.1016/j.jacc.2018.03.521.
[8] Sumida I, Magome T, Das IJ, Yamaguchi H, Kizaki H, Aboshi K, et al. A convolution neural network for higher resolution dose prediction in prostate volumetric modulated arc therapy. Phys Med 2020;72:88–95. https://doi.org/10.1016/j.ejmp.2020.03.023.
[9] Mori S Deep architecture neural network-based real-time image processing for image-guided radiotherapy. Phys Med 2020;40:79–87. https://doi.org/10.1016/j.ejmp.2017.07.013.
[10] Ker J, Wang L, Rao J, Lim T Deep learning applications in medical image analysis. IEEE Access 2018;6:9375–89. https://doi.org/10.1109/ACCESS.2017.2788044.
[11] Chang H, Han J, Zhong C, Snijders AM, Mao J Unsupervised transfer learning via multi-scale convolutional sparse coding for biomedical applications. IEEE T Pattern Anal 2018;40(5):1182–94. https://doi.org/10.1109/TPAMI.2017.2656884.
[12] Zemouri R, Zerhouni N, Racoceanu D Deep learning in the biomedical applications: Recent and future status. Appl Sci 2019;9:1526. https://doi.org/10.3390/app9081526.
[13] Cao C, Liu F, Tan H, Song D, Shu W, Li W, et al. Deep learning and its applications in biomedicine. Genom Proteom Bioinf 2018;16(1):17–32. https://doi.org/10.1016/j.gpb.2017.07.003.

[14] Chen H, Zhang Y, Zhang W, Liao P, Li K, Zhou J, et al. Low-Dose CT via Deep Neural Network. Biomed Opt Express 2017;8(2):679–94. https://doi.org/10.1364/BOE.8.000679.

[15] Humphries T, Si D, Coulter S, Simms M, Xing R. Comparison of deep learning approaches to low dose CT using low intensity and sparse view data. Proc SPIE 2019;10948. doi: https://doi.org/10.1117/12.2512597.

[16] He K, Zhang X, Ren S, Sun J Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. p. 770–8. https://doi.org/10.1109/CVPR.2016.90.

[17] Kim H, Park CM, Lee M, Park SJ, Song YS, Lee J, et al. Impact of Reconstruction Algorithms on CT Radiomic Features of Pulmonary Tumors: Analysis of Intra- and Inter-Reader Variability and Inter-Reconstruction Algorithm Variability. PLoS One 2016;11. doi: https://doi.org/10.1371/journal.pone.0164924.

[18] Willemink M, Noël P The evolution of image reconstruction for CT - from filtered back projection to artificial intelligence. Eur Radiol 2019;29:2185–95. https://doi.org/10.1007/s00330-018-5810-7.

[19] Higaki T, Nakamura Y, Zhou J, Yu Z, Nemoto T, Tatsugami F, et al. Deep Learning Reconstruction at CT: Phantom Study of the Image Characteristics. Acad Radiol 2020;27:82–7. https://doi.org/10.1016/j.acra.2019.09.008.

[20] Dutta S, Fan J, Chevalier D. Study of CT image texture using deep learning techniques. Proc SPIE 2018;10577. doi: https://doi.org/10.1117/12.2292560.

[21] Greffier J, Hamard A, Pereira F, Barrau C, Pasquier H, et al. Image quality and dose reduction opportunity of deep learning image reconstruction algorithm for CT: a phantom study. Eur Radiol 2020;30:3951–3959. doi: https://doi.org/10.1007/s00330-020-06724-w.

[22] Shin YJ, Chang W, Ye JC, Kang E, Oh DY, Lee YJ, et al. Low-Dose Abdominal CT Using a Deep Learning-Based Denoising Algorithm: A Comparison with CT Reconstructed with Filtered Back Projection or Iterative Reconstruction Algorithm. Korean J Radiol 2020;21(3):356–64. https://doi.org/10.3348/kjr.2019.0413.

[23] Samei E, Richard S. Assessment of the dose reduction potential of a model-based iterative reconstruction algorithm using a task-based performance metrology. Med Phys 2015;42. doi: https://doi.org/10.1118/1.4903899.

[24] Solomon J, Lyu P, Marin D, Samei E Noise and spatial resolution properties of a commercially available deep-learning based CT reconstruction algorithm. Med Phys 2020. https://doi.org/10.1002/MP.14319.

[25] Lee D, Choi S, Kim HJ High quality imaging from sparsely sampled computed tomography data with deep learning and wavelet transform in various domains. Med Phys 2019;46(1). https://doi.org/10.1002/mp.13258.

[26] Nakamura Y, Higaki T, Tatsugami F, Honda Y, Narita K, Akagi M, et al. Possibility of Deep Learning in Medical Imaging Focusing Improvement of Computed Tomography Image Quality. J Comput Assist Tomogr 2015;44(2):314. https://doi.org/10.1097/RCT.0000000000000928.

[27] Gong H, Yu L, Leng S, Dilger S, 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). https://doi.org/10.1002/mp.13500.

[28] Urikura A, Yoshida T, Nakaya Y, Nishimaru E, Hara T, Endo M Deep learning-based reconstruction in ultra-high-resolution computed tomography: Can image noise caused by high definition detector and the miniaturization of matrix element size be improved?. Phys Med 2021;81(4):121–9. https://doi.org/10.1016/j.ejmp.2020.12.006.

[29] Verdun F, Racine D, Ott J, Tapiovaara M, Toroi P, Bochud F, et al. Image quality in ct: From physical measurements to model observers. Physica Medica 2015;31:823–43. https://doi.org/10.1016/j.ejmp.2015.08.007.

[30] Dickmann J, Sarosiek C, Rykalin V, Pankuch M, Coutrakon G, Johnson R, et al. Proof of concept image artifact reduction by energy-modulated proton computed tomography (empct). Phys Med 2021;81(4):237–44. https://doi.org/10.1016/j.ejmp.2020.12.012.

[31] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Cham: Springer International Publishing. p. 234–241, ISBN 978-3-319-24574-4; 2015.

[32] Zhang K, Liu X, Shen J, Li Z, Sang Y, Wu X, et al. Clinically Applicable AI System for Accurate Diagnosis, Quantitative Measurements, and Prognosis of COVID-19 Pneumonia Using Computed Tomography. Cell 2020. https://doi.org/10.1016/j.cell.2020.04.045.

[33] Kim B, Han M, Shim H, Baek J A performance comparison of convolutional neural network-based image denoising methods: The effect of loss functions on low-dose ct images. Med Phys 2019;46(9):3906–23. https://doi.org/10.1002/mp.13713.

[34] Heinrich MP, Stille M, Buzug TM Residual u-net convolutional neural network architecture for low-dose ct denoising. Current Directions Biomed Eng 2018;4(1):297–300. https://doi.org/10.1515/cdbme-2018-0072.

[35] Yu S, Park B., Jeong J. Deep iterative down-up cnn for image denoising. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. 2019.

[36] Park B, Yu S, Jeong J. Densely connected hierarchical network for image denoising. In: Proc. of the IEEE/CVF Conf. CVPR. 2019,.

[37] Komatsu R, Gonsalves T Comparing u-net based models for denoising color images. AI 2020;1(4):465–86. https://doi.org/10.3390/ai1040029.

[38] Yuan H, Jia J, Zhu Z. Sipid: A deep learning framework for sinogram interpolation and image denoising in low-dose ct reconstruction. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). 2018, p. 1521–1524. doi: https://doi.org/10.1109/ISBI.2018.8363862.

[39] Bao L, Yang Z, Wang S, Bai D, Lee J. Real image denoising based on multi-scale residual dense block and cascaded u-net with block-connection. In: Proc. of the IEEE/CVF Conf. CVPR. 2020.

[40] Gu S, Li Y, Gool LV, Timofte R. Self-guided network for fast image denoising. In: Proc. of the IEEE/CVF Conf. CVPR. 2019,.

[41] Maharjan P, Li L, Li Z, Xu N, Ma C, Li Y Improving extreme low-light image denoising via residual learning. 2019 IEEE International Conference on Multimedia and Expo (ICME); 2019. p. 916–21. https://doi.org/10.1109/ICME.2019.00162.

[42] Yi X, Babyn P. Sharpness-aware low-dose ct denoising using conditional generative adversarial network. J Digit Imaging 2018;5:655–669. doi: https://doi.org/10.1007/s10278-018-0056-0.

[43] Chen C, Chen Q, Xu J, Koltun V Learning to see in the dark. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2018.

[44] Chen H, Zhang Y, Zhang W, Liao P, Li K, Zhou J, et al. Low-dose ct denoising with convolutional neural network. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). 2017, p. 143–146. doi: https://doi.org/10.1109/ISBI.2017.7950488.

[45] Nasrin S, Alom MZ, Burada R, Taha TM, Asari VK Medical image denoising with recurrent residual u-net (r2u-net) base auto-encoder. 2019 IEEE National Aerospace and Electronics Conference (NAECON); 2019. p. 345–50. https://doi.org/10.1109/NAECON46414.2019.9057834.

[46] Moradi S, Oghli MG, Alizadehasl A, Shiri I, Oveisi N, Oveisi M, et al. Mfp-unet: A novel deep learning based approach for left ventricle segmentation in echocardiography. Phys Med 2019;67:58–69. https://doi.org/10.1016/j.ejmp.2019.10.001.

[47] Komatsu R, Gonsalves T Comparing u-net based models for denoising color images. AI 2020;1(4):465–86. https://doi.org/10.3390/ai1040029.

[48] Quattoni A, Collins M, Darrell T. Transfer learning for image classification with sparse prototype representations. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. 2008, p. 1–8. doi: 10.1109/CVPR.2008.4587637.

[49] Buchholz T, Prakash M, Schmidt D, Krull A, Jug F. Denoiseg: Joint denoising and segmentation. In: Computer Vision - ECCV 2020 Workshops; vol. 12535. 2020, p. 143–146. doi: 10.1007/978-3-030-66415-2_21.

[50] Henkelman RM, Kay I, Bronskill MJ Receiver operator characteristic (roc) analysis without truth. Med Decis Mak 1990;10(1):24–9. https://doi.org/10.1177/0272989X9001000105.

[51] Hajian-Tilaki K. Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation. Caspian J Intern Med 2013;4(2):627–635. doi: PMID:24009950; PMCID: PMC3755824.

[52] Lusted LB Signal detectability and medical decision-making. Science 1971;171(3977):1217–9. https://doi.org/10.1126/science.171.3977.1217.

[53] C.E. M.. Evaluation of Medical Images. Springer; 1992. doi: https://doi.org/10.1007/978-3-642-77888-9_10.

[54] Sharp P, Barber DC, Brown DG, Burgess AE, Metz CE, Myers KJ, et al. 4. quality of the observed image. ICRU Report 1996;os-28(1):23–30. doi: https://doi.org/10.1093/jicru_os28.1.23.

[55] Brickley MR, Prytherch IM, Kay EJ, Shepherd JP A new method of assessment of clinical teaching: Roc analysis. Med Educat 1995;29(2):150–3. https://doi.org/10.1111/j.1365-2923.1995.tb02819.x.

[56] Martin C, Sharp P, Sutton D Measurement of image quality in diagnostic radiology. Appl Radiat Isot 1999;50(1):21–38. https://doi.org/10.1016/S0969-8043(98)00022-0.

[57] Noferini L, Taddeucci A, Bartolini M, Bruschi A, Menchi I CT image quality assessment by a Channelized Hotelling Observer (CHO): Application to protocol optimization. Phys Med 2016;32:1717–23. https://doi.org/10.1016/j.ejmp.2016.11.002.

[58] Rizzo S, Botta F, Raimondi S, Origgi D, Fanciullo C, Morganti AG, et al. Radiomics: the facts and the challenges of image analysis. Eur Radiol Exp 2018;2:36. https://doi.org/10.1186/s41747-018-0068-z.

[59] Park BW, Kim JK, Heo C, Park KJ Reliability of CT radiomic features reflecting tumour heterogeneity according to image quality and image processing parameters. Nature Sci Rep 2020;10:3852. https://doi.org/10.1038/s41598-020-60868-9.

[60] Midya A, Chakraborty J, Gönen M, Do RKG, Simpson AL. Influence of CT acquisition and reconstruction parameters on radiomic feature reproducibility. J of Med Imaging 2018;011020. doi: https://doi.org/10.1117/1.JMI.5.1.011020.

[61] van Griethuysen J, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res 2017;77(21):e104–7. https://doi.org/10.1158/0008-5472.CAN-17-0339.

[62] von Falck C, Bratanova V, Rodt T, Meyer B, Waldeck S, Wacker F, et al. Influence of sinogram affirmed iterative reconstruction of ct data on image noise characteristics and low-contrast detectability: An objective approach. PLOS ONE 2013;8:1–10. https://doi.org/10.1371/journal.pone.0056875.

[63] Khobragade P, Fan J, Rupcich F, Crotty D, TG TS. Application of fractal dimension for quantifying noise texture in computed tomography images. Med Phys 2018;8:1–10. doi: https://doi.org/10.1002/mp.13040.

[64] Christianson O, Chen JJS, Yang Z, Saiprasad G, Dima A, Filliben JJ, et al. An improved index of image quality for task-based performance of ct

iterative reconstruction across three commercial implementations. Radiology 2015;275(3):725–34. https://doi.org/10.1148/radiol.15132091.
[65] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R Dropout: A simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15(1):1929–58. https://doi.org/10.5555/2627435.2670313.
[66] cogswell M, ahmed F, Girshick BR, Zitnick LC, Batra D. Reducing overfitting in deep networks by decorrelating representations. international conference on learning representations; 2015.
[67] Zhao W, Liu L, Xiao J, Ke J Research on the deep learning of the small sample data based on transfer learning. AIP Conference Proceedings 2017; 1864(1):020018. https://doi.org/10.1063/1.4992835.
[68] Perez L, Wang J. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. arXiv:171204621 2017;.
[69] Shorten C, Khoshgoftaar T A survey on image data augmentation for deep learning. J Big Data 2019;6(60). https://doi.org/10.1186/s40537-019-0197-0.
[70] Taylor L, Nitschke G Improving deep learning with generic data augmentation. 2018 IEEE Symposium Series on Computational Intelligence (SSCI); 2018. p. 1542–7. https://doi.org/10.1109/SSCI.2018.8628742.
[71] Salman S, Liu X. Overfitting Mechanism and Avoidance in Deep Neural Networks. arXiv:190106566 2019;.
[72] Bradski G. The OpenCV Library. Dr Dobb's Journal of Software Tools 2000; doi: https://docs.opencv.org/.
[73] Kanakis M, Bruggemann D, Saha S, Georgoulis S, Obukhov A, Gool LV. Reparameterizing convolutions for incremental multi-task learning without task interference. In: Proc. ECCV 2020; vol. 12365. Springer; 2020, p. 689–707. doi: 10.1007/978-3-030-58565-5_41.
[74] Ubernet KI Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. Procof IEEE Conf CVPR 2017.
[75] Ponomarenko N, Krivenko S, Egiazarian K, Astola J, Lukin V. Weighted mse based metrics for characterization of visual quality of image denoising methods. 2009.
[76] Krull A, Buchholz TO, Jug F Noise2void - learning denoising from single noisy images. Proc IEEE/ CVPR. 2019.
[77] Lee S, Negishi M, Urakubo H, Kasai H, Ishii S Mu-net: Multi-scale u-net for two-photon microscopy image denoising and restoration. Neural Networks 2020;125:92–103. https://doi.org/10.1016/j.neunet.2020.01.026.
[78] Liu D, Wen B, Liu X, Huang TS. When image denoising meets high-level vision tasks: A deep learning approach. CoRR 2017;abs/1706.04284. URL http://arxiv.org/abs/1706.04284.
[79] Yang L, Shangguan H, Zhang X, Wang A, Han Z High-frequency sensitive generative adversarial network for low-dose ct image denoising. IEEE Access 2020;8:930–43. https://doi.org/10.1109/ACCESS.2019.2961983.
[80] Chi J, Wu C, Yu X, Ji P, Chu H Single low-dose ct image denoising using a generative adversarial network with modified u-net generator and multi-level discriminator. IEEE Access 2020;8:133470–87. https://doi.org/10.1109/ACCESS.2020.3006512.
[81] Dong C, Loy CC, He K, Tang X Image super-resolution using deep convolutional networks. IEEE Trans Pattern Analysis Mach Intell 2016;38 (2):295–307. https://doi.org/10.1109/TPAMI.2015.2439281.
[82] Ge YQW, Pingkun Y, Mannudeep K Ct image denoising with perceptive deep neural networks. Fully3D 2017 Proc.; 2017. p. 858–63. https://doi.org/10.12059/Fully3D.2017-11-3202015.
[83] Zhao H, Gallo O, Frosio I, Kautz J Loss functions for image restoration with neural networks. IEEE Trans Comput Imag 2017;3(1):47–57. https://doi.org/10.1109/TCI.2016.2644865.
[84] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: Large-scale machine learning on heterogeneous systems. 2015. doi: http://tensorflow.org/.
[85] Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative - Reference Manual. arXiv 2019;1612.07003. doi: https://doi.org/10.1148/radiol.2020191145.
[86] Traverso A, Wee L, Dekker A, Gillies R Repeatability and Reproducibility of Radiomic Features: A Systematic Review. Int J Radiation Oncol Biol Phys 2018;102(4):1143–58. https://doi.org/10.1016/j.ijrobp.2018.05.053.
[87] Myers L, Sirois MJ. Spearman Correlation Coefficients, Differences between. Am Canc Soc. ISBN 9780471667193; 2006, doi: 10.1002/0471667196.ess5050.pub2.
[88] Sheen H, Kim W, Byun B, Kong C, Song WS, Cho WH, et al. Metastasis risk prediction model in osteosarcoma using metabolic imaging phenotypes: A multivariable radiomics model. PLoS One 2019;14(11). https://doi.org/10.1371/journal.pone.0225242.
[89] Oliver J, Budzevich M, Hunt D, Moros EG, Latifi K., TJ T.J.D., et al. Sensitivity of Image Features to Noise in Conventional and Respiratory-Gated PET/CT Images of Lung Cancer: Uncorrelated Noise Effects. Technol Cancer Res Treat 2017;16(5) 595–608. doi: https://doi.org/10.1177/1533034616661852.
[90] Hunter JD Matplotlib: A 2d graphics environment. Computing Sci Eng 2007; 9(3):90–5. https://doi.org/10.1109/MCSE.2007.55.
[91] Baker M, Dong F, Primak A, Obuchowski N, Einstein D, Gandhi N, et al. Contrast-to-Noise Ratio and Low-Contrast Object Resolution on Full- and Low-Dose MDCT: Safire Versus Filtered Back Projection in a Low-Contrast Object Phantom and in the Liver. AJR Am J Roentgenol 2012;199. doi: https://doi.org/DOI:10.2214/AJR.11.7421.
[92] Fessler J, Rogers W Spatial Resolution Properties of Penalized-Likelihood Image Reconstruction: Space-Invariant Tomographs. IEEE Trans Im Proc 1996;5(9). https://doi.org/10.1109/83.535846.
[93] Fletcher J, Yu L, Li Z, Manduca A, Blezek D, Hough D, et al. Observer Performance in the Detection and Classification of Malignant Hepatic Nodules and Masses with CT Image-Space Denoising and Iterative Reconstruction. Radiol 2015;276(15). https://doi.org/10.1148/radiol.2015141991.
[94] Evans J, Politte D, Whiting BR, O'Sullivan J, Williamson J Noise-resolution tradeoffs in x-ray CT imaging: A comparison of penalized alternating minimization and filtered backprojection algorithms. Med Phys 2011;38(3). https://doi.org/10.1118/1.3549757.
[95] Geyer L, Schoepf UJ, Meinel F, Nance J, Bastarrika G, Leipsic J, et al. State of the Art: Iterative CT Reconstruction Techniques. Radiol 2015;276(2). https://doi.org/10.1148/radiol.2015132766.
[96] Huerga C, Glarıa L, Castro P, Alejo L, Bayón J, Guibelalde E Segmentation improvement through denoising of pet images with 3d-context modelling in wavelet domain. Phys Med 2018;52:62–71. https://doi.org/10.1016/j.ejmp.2018.08.008.
[97] Boone J, Brink J, Edyvean S, Huda W, Leitz W, McCollough C, et al. Radiation dose and image-quality assessment in computed tomography. J ICRU 2012;12:9–149. https://doi.org/10.1093/jicru/ndt007.
[98] Ehman E, Yu L, Manduca A, Hara A, Shiung M, Jondal D, et al. Methods for clinical evaluation of noise reduction techniques in abdominopelvic ct. Radiographics 2014;34(4):849–62. https://doi.org/10.1148/rg.344135128.
[99] Artmann U, Wueller D. Interaction of image noise, spatial resolution, and low contrast fine detail preservation in digital image processing. In: Digital Photography V; vol. 7250. SPIE; 2009, p. 154–162. doi: 10.1117/12.805927.
[100] Mohan J, Krishnaveni V, Guo Y A survey on the magnetic resonance image denoising methods. Biomed Signal Process Control 2014;9:56–69. https://doi.org/10.1016/j.bspc.2013.10.007.
[101] Artmann U, Wueller D. Noise reduction versus spatial resolution. In: Digital Photography IV; vol. 6817. SPIE; 2008, p. 71–80. doi: 10.1117/12.765887.
[102] Li Z, Yu L, Trzasko J, Lake D, Blezek D, Fletcher J, et al. Adaptive nonlocal means filtering based on local noise level for ct denoising. Med Phys 2014; 41(1):56–69. https://doi.org/10.1118/1.4851635.
[103] Gong H, Hu Q, Walther A, Koo C, Takahashi E, Levin D, et al. Deep-learning-based model observer for a lung nodule detection task in computed tomography. Proc SPIE 2020;042807. doi: https://doi.org/10.1117/1.JMI.7.4.042807.
[104] Kopp F, Catalano M, Pfeiffer D, Fingerle A, Rummeny E, Noel PB CNN as model observer in a liver lesion detection task for x-ray computed tomography: A phantom study. Med Phys 2018;45(10). https://doi.org/10.1002/mp.13151.
[105] Reith F, Wandell B. Comparing pattern sensitivity of a convolutional neural network with an ideal observer and support vector machine. arXiv: 191105055 2019;.
[106] Zhou W, Li H, Anastasio M Approximating the Ideal Observer and Hotelling Observer for binary signal detection tasks by use of supervised learning methods. IEEE Trans Med Im 2019;38:3142456–68. https://doi.org/10.1109/TMI.2019.2911211.
[107] Alnowami M, Mills G, Awis M, Elangovanr P, Patel M, Halling-Brown M, et al. A deep learning model observer for use in alterative forced choice virtual clinical trials. Proc SPIE 2018;10577. doi: https://doi.org/10.1117/12.2293209.
[108] Massanes F, Brankov J Evaluation of CNN as anthropomorphic model observer. Proc SPIE 2017;10136. https://doi.org/10.1117/12.2254603.