UNet and MobileNet CNN-based model observers for CT protocol optimization: comparative performance evaluation by means of phantom CT images

Federico Valeri,[a,b] Maurizio Bartolucci,[c] Elena Cantoni,[a] Roberto Carpi,[d] Evaristo Cisbani,[e] Ilaria Cupparo,[a,b] Sandra Doria,[f,g,*] Cesare Gori,[a] Mauro Grigioni,[e] Lorenzo Lasagni,[a,b] Alessandro Marconi,[a] Lorenzo Nicola Mazzoni,[h] Vittorio Miele,[i] Silvia Pradella,[i] Guido Risaliti,[a] Valentina Sanguineti,[j] Diego Sona,[k] Letizia Vannucchi,[l] and Adriana Taddeucci[m,n]

[a]Università degli Studi di Firenze, Dipartimento di Fisica e Astronomia, Florence, Italy
[b]Università degli Studi di Firenze, Scuola di Scienze della Salute Umana, Florence, Italy
[c]Ospedale S. Stefano, Azienda USL Toscana Centro, SOC Radiodiagnostica, Prato, Italy
[d]Ospedale Santa Maria Nuova, Azienda USL Toscana Centro, SOC Radiologia, Florence, Italy
[e]Istituto Superiore di Sanità, Centro Nazionale Tecnologie Innvative in Sanità Pubblica, Rome, Italy
[f]Istituto di Chimica dei Composti OrganoMetallici, Consiglio Nazionale delle Ricerche, Florence, Italy
[g]Università degli Studi di Firenze, European Laboratory for Nonlinear Spectroscopy, Florence, Italy
[h]Ospedale San Jacopo, Azienda USL Toscana Centro, UO Fisica Sanitaria Prato e Pistoia, Pistoia, Italy
[i]Azienda Ospedaliero-Universitaria Careggi, SOD Radiodiagnostica di Emergenza-Urgenza, Florence, Italy
[j]Istituto Italiano di Tecnologia, Pattern Analysis & Computer Vision, Genoa, Italy
[k]Fondazione Bruno Kessler, Data Science for Health Unit, Trento, Italy
[l]Ospedale S. Jacopo, AUSL Toscana Centro, SOC Radiodiagnostica, Pistoia, Italy
[m]Azienda Ospedaliero-Universitaria Careggi, UO Fisica Sanitaria, Florence, Italy
[n]Istituto Nazionale di Fisica Nucleare - Sezione di Firenze, Sesto Fiorentino, Italy

Abstract

Purpose: The aim of this work is the development and characterization of a model observer (MO) based on convolutional neural networks (CNNs), trained to mimic human observers in image evaluation in terms of detection and localization of low-contrast objects in CT scans acquired on a reference phantom. The final goal is automatic image quality evaluation and CT protocol optimization to fulfill the ALARA principle.

Approach: Preliminary work was carried out to collect localization confidence ratings of human observers for signal presence/absence from a dataset of 30,000 CT images acquired on a PolyMethyl MethAcrylate phantom containing inserts filled with iodinated contrast media at different concentrations. The collected data were used to generate the labels for the training of the artificial neural networks. We developed and compared two CNN architectures based respectively on Unet and MobileNetV2, specifically adapted to achieve the double tasks of classification and localization. The CNN evaluation was performed by computing the area under localization-ROC curve (LAUC) and accuracy metrics on the test dataset.

Results: The mean of absolute percentage error between the LAUC of the human observer and MO was found to be below 5% for the most significative test data subsets. An elevated inter-rater agreement was achieved in terms of S-statistics and other common statistical indices.

Conclusions: Very good agreement was measured between the human observer and MO, as well as between the performance of the two algorithms. Therefore, this work is highly supportive of

*Address all correspondence to Sandra Doria, doria@lens.unifi.it

the feasibility of employing CNN-MO combined with a specifically designed phantom for CT protocol optimization programs.

 [DOI: 1.JMI.10.S1.S11904/1.JMI.10.S1.S11904]

Keywords: artificial intelligence; computed tomography; dose optimization; model observer; image quality evaluation.

Paper 22267SSR received Oct. 6, 2022; accepted for publication Feb. 9, 2023; published online Mar. 7, 2023.

1 Introduction

Computed tomography (CT) applications represent one of the most well-established diagnostic tools in current medical imaging; CT is capable of providing very detailed anatomical images of many biological tissues at one time due to its large dynamic range. With the widespread availability of CT equipment and the increasing number of patient examinations,[1] the issues of the quantification of risks related to X-ray exposure, and consequently the need for further optimization of CT protocols to fulfill the ALARA ("As Low As Reasonably Achievable") principle, have arisen.[2,3] The main international organizations dealing with ionizing radiation protection and safety standards, the International Commission For Radiological Protection (ICRP) and the International Atomic Energy Agency (IAEA), have provided patient dose management recommendations and have identified lacunae in justification and optimization, thus providing guidance and improving practice.[4–7] In the Directive 2013/59/EURATOM,[8] the European Union Council stated the need to develop and put into action optimization programs to achieve the best compromise between radiation dose and image quality, with the aim to reduce patients dose to the minimum level compatible with diagnostic accuracy.

The choice of the optimum dose level requires the evaluation of CT image quality, which can be measured by receiving operator characteristic (ROC) analysis in reader studies in which trained medical staff perform a specific clinical task. Such an approach is especially suitable in the case of iterative reconstruction techniques because the standard physical quantities are no longer suitable for a thorough image quality assessment.[9] However, the extremely high number of different CT protocols in use even within small radiological facilities makes *de facto* the evaluation of all necessary ROC curves almost impracticable as it would require too much observation time to be provided by medical staff. In the recent past, this fundamental limitation has been addressed by replacing human observers with algorithmic approaches (i.e., model observers); in particular, the channelized Hotelling observer (CHO) model[10–12] demonstrated great potential, but it is still limited by poor generalization capability to different CT settings.[13,14] The appreciable results obtained through such algorithmic methods have encouraged researchers to proceed by employing artificial intelligence (AI) algorithms, which are seemingly more powerful than CHO.[15–21] Recently, the increasing availability of computational resources has driven the scientific research toward the use of the latter approach, which has shown remarkable effectiveness in mimicking the human observers' performances in different diagnostic imaging evaluation tasks.[17,21–24] To take into account the inefficiency and variability of human responses, several strategies have been proposed. Previous adopted approaches consist of the introduction of an internal noise component in the output statistics of convolutional neural networks (CNNs)[13,16,23] and the use of human-labeled data for training.[17,21,25] When actual patient CT data are used, the dose level dependency is commonly studied by introducing appropriate noise into the images.[16,21–23,26–28]

Within this context, our goal is to build a solid model observer (MO) framework based on CNNs that is capable of reproducing the performances of human observers in the identification of low contrast-to-noise ratio (CNR) objects in reference phantom CT images. Compared with the current state-of-the-art methods, our work is characterized by the concurrence of large dataset variability in terms of size and CNR of the imaged objects, CT acquisitions at eight different dose indices, two reconstruction techniques, and labels by a large group of 30 human observers, including 19 radiologists from four different radiological departments.

The use of a specifically designed phantom allowed for the collection of a large dataset of 30,000 images at various dose indices and under controlled acquisition conditions.

Two intrinsically different CNN architectures were optimized for the double task of localization and classification of low CNR objects within the phantom CT images to get insight into the relation between CNN behavior and architectures.

We performed an extended statistical analysis of the results to address the overall observers performances in terms of localization-area under curve (L-AUC), which is expected to be more accurate than the conventional AUC metric because it takes into account both localization and classification capabilities.[29]

Several statistical indices and the accuracy metric were computed to obtain a better understanding of the CNNs response and the limitations of this AI approach.

The results are very promising: both approaches are capable of miming human detectability performance in phantom CT images. We believe that these CNN-based MOs, combined with specifically designed phantoms, may effectively support the optimization of CT protocols, avoiding the time-consuming limitations of medical staff evaluations.

2 Materials and Methods

2.1 *Image Dataset*

The annotated dataset used to train, validate, and test the proposed CNNs is a subset of the large dataset extensively described in our previous work.[30] The dataset consists of CT images of a specifically manufactured PolyMethyl MethAcrylate (PMMA) phantom (Fig. 1), containing 10 cylindrical inserts of different diameters (3, 4, 5, 6, and 7 mm); each couple of inserts with the same diameter provides two different contrast values (45 and 55 HU) with respect to the PMMA background obtained by filling the inserts with aqueous solutions of iodinated contrast media at two distinct concentrations. The phantom consists of three adjacent blocks, each with an ellipsoidal shape with a major axis of 31 cm, a minor axis of 21 cm, and a thickness of 7 cm: two blocks have five inserts each and the third, with no inserts, is finalized to obtain homogeneous background images.

Acquisition was performed with a 128 slice CT scanner (Somatom Definition Flash, Siemens Healthcare) at eight different volumetric CT Dose Index settings ($CTDI_{vol}$ [mGy] = 4.4, 5.1, 6.0, 6.9, 7.8, 8.6, 9.6, and 10.2), with the following protocol for abdomen selected: 120 kVp, AEC on, helical mode, pitch = 1, beam collimator = 38.4 mm, and slice thickness = 2 mm).

Both filtered back projection (FBP) and Iterative Reconstruction (IR, SAFIRE force 3) image reconstruction techniques were applied to the acquired data, with convolution kernels B41s and IF41s, respectively. The CT image reconstruction FoV (RFoV) was chosen to be 5 cm^2 (512×512 pixels per image) to produce reconstructed images containing one single insert each. Images without inserts were also similarly reconstructed and added to the dataset. To further increase the dataset variability, data augmentation techniques consisting of 90 degrees rotations and horizontal and vertical flips were applied on all images. Out of this large dataset, 30,000

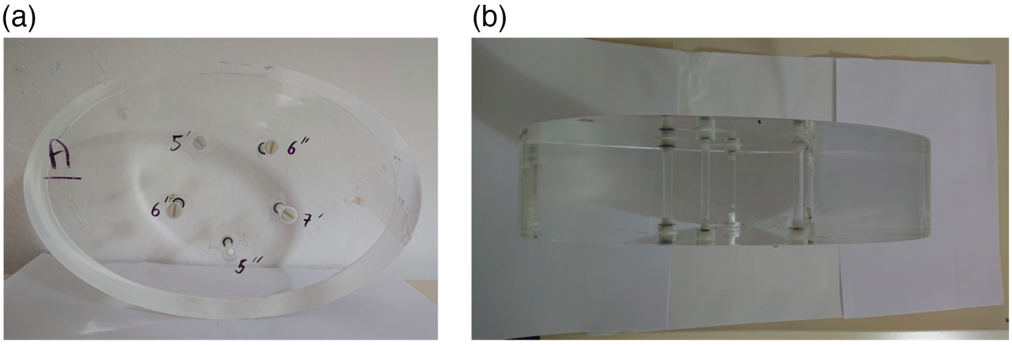

Fig. 1 Lateral view (a) and top view (b) of one of the two blocks containing five inserts filled with iodinated contrast media.

Table 1 Selected subensembles from the original image dataset characterized by insert (object) size and contrast.

No. images	object diameter	contrast
	d (mm)	C (HU)
10,000	Homogeneous images	—
3000	3	45
3000	3	55
3000	4	45
3000	4	55
2800	5	45
2800	5	55
1200	6	45
1200	7	45

images were selected with a tradeoff between having an adequate amount of training data for CNNs and an acceptable amount of time being spent to collect confidence scores and insert position coordinates (when detected) for each single image by visual inspection. On the basis of the knowledge acquired in a previous work,[30] the selected subensemble was chosen as described in Table 1. It is worth pointing out that the dataset is not balanced in terms of diameters and contrasts of the imaged objects: the abundance of imaged object types was selected according to their detectability (as quantified by human LAUC analysis in Sec. 3): the larger the detectability is, the smaller the subensemble is; moreover, images containing inserts of 6 and 7 mm in diameter at the higher CNR were excluded because the visibility of such objects was too elevated across the entire $CTDI_{vol}$ range. A reference subset was chosen to evaluate the observers performances; it consists of images containing 4-mm diameter objects, contrast $C = 45$ HU, and the iterative reconstruction (IR) algorithm: the CNR computed on such a subset monotonically increased with $CTDI_{vol}$ from 1.9 to 3.1.

For the human observers image visualization step and the subsequent steps of algorithm optimization, the dataset, originally reconstructed with a 512×512 RFoV, was reduced to 256×256 pixels per image to optimize computational resources, after testing to ensure that resizing the images did not affect the CNN performances.

Figure 2 shows an example set of reconstructed images of two inserts (4 and 7 mm diameter) at the lower contrast (45 HU) for different CTDI values. It is noticeable that the visibility of the insert decreases with $CTDI_{vol}$ due to the decrease of CNR.

2.2 *Confidence Scores Collection*

To collect the labels to train the MOs, the detection task was represented as a multiclass ranking task: an ordinal score was assigned by the human observer, corresponding to the confidence attributed to the presence (or absence) of the object, in a range from 0 to 3 (0 = object surely not present; 1 = object unlikely to be present; 2 = object likely to be present; 3 = object surely present). At the same time, the operator was asked to identify the location of the object (if assigned score is not 0). A graphical Python-based interface was developed to automatically save the score and the coordinates assigned to each identified object by the human observers. A representation of the screen window generated by the software and presented to the operator for image evaluation is reported in Fig. 3.

A total of 30 human observers contributed with the visual examination of 1000 images each; following a strategy already proposed in previous works,[12,16] both radiologists (20) and medical physicists (10) were included as evaluators to get a larger variability of evaluation performances

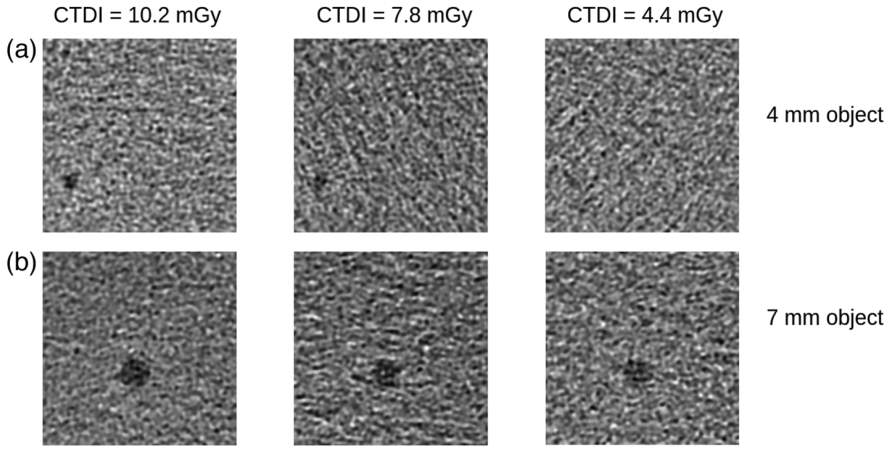

Fig. 2 Example of reconstructed images (iterative reconstruction technique) with (a) 4 mm and (b) 7 mm inserts at the lower contrast (45 HU); it is noticeable that object visibility depends both on size and $CTDI_{vol}$.

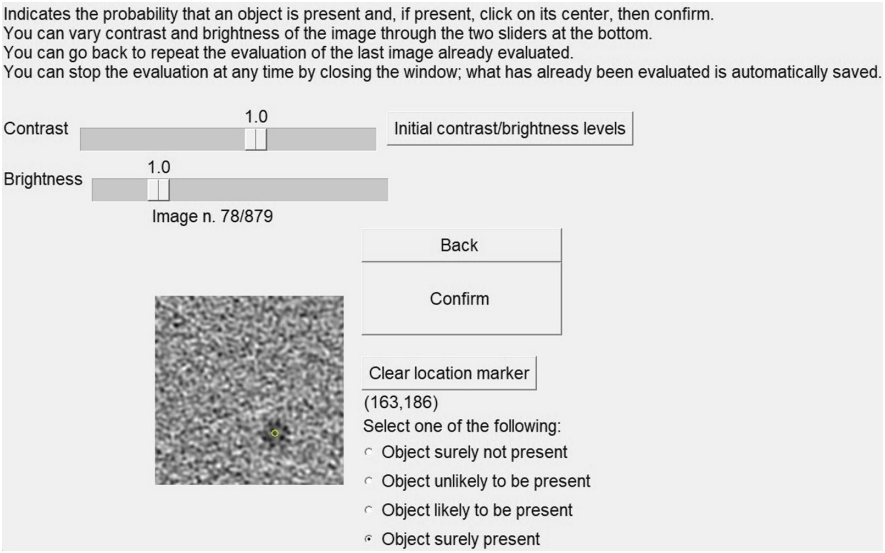

Fig. 3 Screenshot of the software interface developed to collect the human observer response to CT images visual inspection.

and make the CNN-MOs more reliable and robust. In consideration of the easy task and the simple content of the dataset, ingredients that risk inducing overfitting in CNN training, as well as the very time-consuming reader study, we decided to promote dataset size over multiple ratings of single images: there was no intersection between the subsets of 1000 images evaluated by each observers. However, for the same contrast, size, reconstruction technique, and $CTDI_{vol}$, a multitude of images with similar properties (noise pattern and signal detectability) were generated from different slices of the same CT scan, among which the signal location was varied by data augmentation (see also Sec. 2.1).

2.3 *Convolutional Neural Networks*

Two specific CNNs were developed and optimized to perform the MO task: a U-Net-based architecture and a MobileNetV2-based architecture.

Both CNNs were trained from scratch on the training dataset, which consisted of noisy images that previously underwent visual inspection, labeled with the corresponding confidence scores and coordinates assigned by the human observers. Despite a few recent applications,[25] this

labeling choice represents an alternative training strategy with respect to most of the MO algorithms reported in the literature,[16,16,24,31] often based on an *a posteriori* correction of the output statistics of CNNs, trained with impartial labels representing the actual presence and location of the object within the images.

To ensure a robust statistical analysis of results, a fivefold cross validation procedure was applied: five training experiments were carried out using randomly assigned train and test subsets (80% and 20%, respectively, for each experiment).

In the following, the developed CNNs are described in detail.

2.3.1 *UNet-based architecture*

The first architecture was designed for the MO tasks by customizing a UNet-based CNN, previously developed by the authors[30] for denoising and segmentation of phantom CT images. The double-task strategy, successfully employed in the cited work, was implemented in this context to achieve object localization and confidence score prediction at the same time.

The UNet is a CNN based on an autoencoder architecture already well documented in the literature[32–42] that, in particular, has been employed for segmentation and localization tasks in the postprocessing of medical images and has already demonstrated elevated performances as an MO.[42]

The original architecture, consisting of a combination of max pooling, convolution, and fully connected layers, was reduced to a total of nine layers and four skip connections. A scheme of the UNet used is reported in Fig. 4 (architecture details and layers sequence are reported in Fig. S1 in Supplementary Material). At the end of the encoder stream, a dense layer is connected to produce a scalar output representing the confidence score prediction (implemented as a multiclass classification task). A mean square error loss Loss_{MO} is implemented to let the CNN learn the scores given by the human observers (used as score labels for training).

The decoder stream is fully devoted to the localization task. The idea behind the implementation of the localization task originated from the CNN architecture proposed by Newell et al.,[43] in which concatenated autoencoders were used to estimate the pose of a human body through the generation of a series of heatmaps, one for each identified body joint. In our case, a single autoencoder is implemented to generate the heatmap corresponding to the object identified within the image. The heatmap is a 256×256 matrix with a maximum that is assumed to be the prediction of the object center. Following Newell et al.,[43] two additional losses are implemented and devoted to the localization tasks. The first one is a Kullback–Leibler divergence loss (Loss_{KLD})[44–48] between the heatmap produced by the network and the ground truth, represented by a 2D Gaussian (normalized to unity and with FWHM equal to the object diameter) centered in

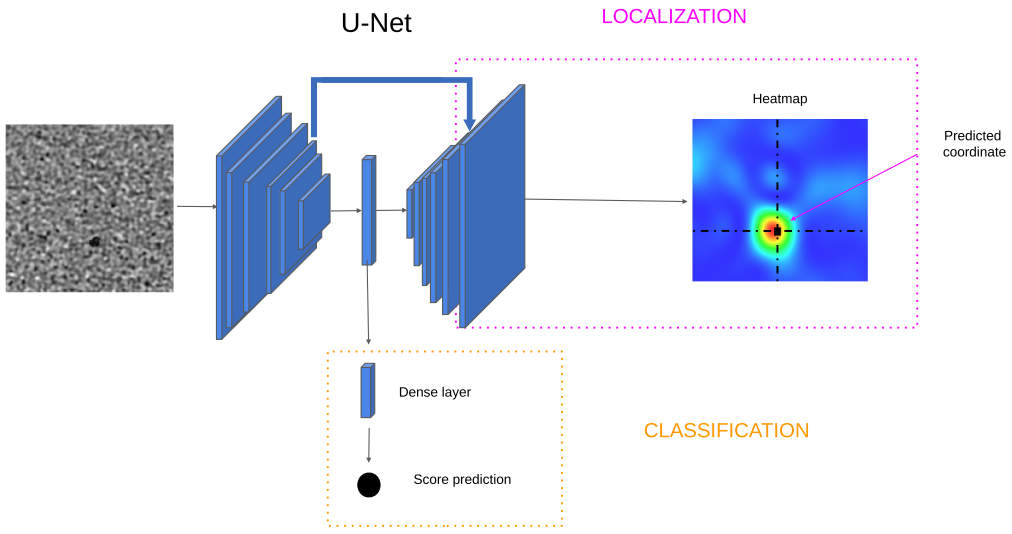

Fig. 4 Schematic illustration of the developed UNet-based CNN architecture.

the coordinates picked by the human observers. In the case of images classified as "object absolutely not present" by the human observer (score = 0), the ground truth consists of a matrix filled with zeros. The second loss consists of a mean square error loss (Loss_{LOC}) between the predicted coordinates and those actually picked by the human observers (the latter being used as coordinate labels for training). The contribution of this loss was set to zero for images classified as "object absolutely not present" by the human observer.

A weighted sum of the three losses was tuned during the optimization procedure for the final training as

$$\text{LOSS}_{\text{UNet}} = (\text{Loss}_{\text{MO}} + 100 \cdot \text{Loss}_{\text{KLD}} + 0.1 \cdot \text{Loss}_{\text{LOC}}). \quad (1)$$

It is worth noting that a specific function based on the softmax operation was built to compute the maximum of the heatmap by means of a differentiable function, an essential requirement for backpropagation to occur properly during CNN training.[49]

The batch size is 48, the learning rate is 0.0001, and the Adam algorithm is employed as the optimizer.

2.3.2 *MobileNetV2-based architecture*

The second strategy is based on the MobileNetV2 architecture,[50] the complexity of which was reduced. We used a MobileNetV2 architecture with fewer convolution layers than the original architecture, i.e., we only used the first 11 layers up to the layer called *block_3_depthwise_relu* during the optimization procedure to limit overfitting. The MobileNetV2 has already been exploited in the medical imaging field, mostly for classification and detection of lesions,[51–56] and recently it was successfully implemented for COVID-19 diagnosis.[57–60]

Two different MobileNetV2-based CNNs are implemented for prediction of the confidence score (represented as a multiclass classification task) and of the object coordinates; their architectures are reported in Figs. 5 and 6, respectively. Two distinct CNNs were built as their architectures are not exactly identical but differ for the final two layers and the input data in the training phase have different sizes in the two cases.

1. The CNN devoted to the classification task (Fig. 5) takes as input a CT image and, after the 11th layer of the original MobileNetV2, ends with a global average pooling layer followed by a densely-connected layer, consisting of four units and a softmax activation function to predict the confidence score of human observers. The sparse categorical cross-entropy function is used as the loss during the training phase.
2. The CNN devoted to the localization task (Fig. 6) is trained using 48×48 images, obtained by cropping the original 256×256 images around the coordinates picked by the human observers (or random coordinates when the assigned score is 0). The crop size is chosen to be large enough to include the largest insert diameter (7 mm = 36 pixels). After the 11th layer of the original MobileNetV2, an average pooling layer (pool size 4,

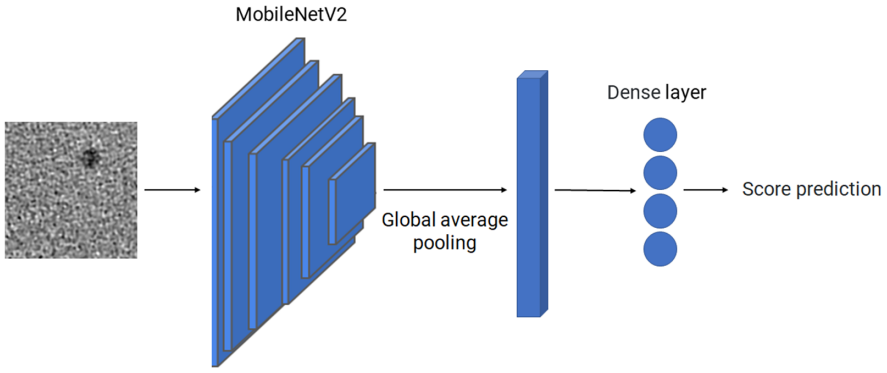

Fig. 5 Schematic illustration of the MobileNetV2-based CNN architecture used for the classification task.

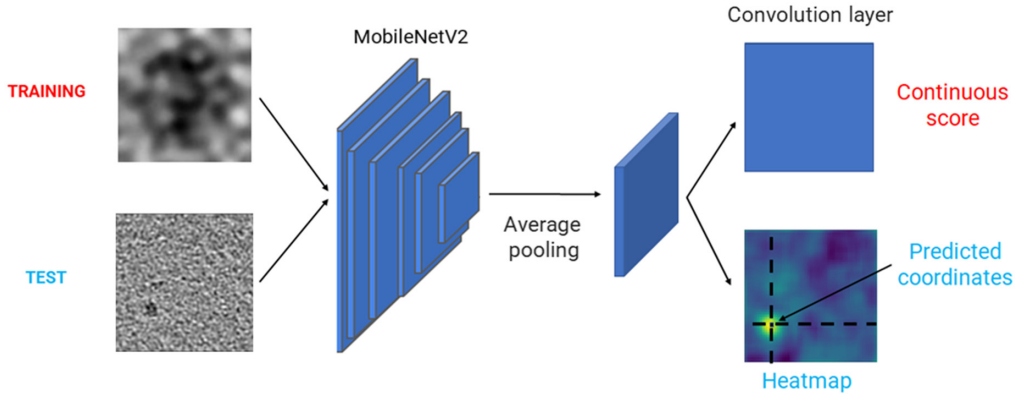

Fig. 6 Schematic illustration of the MobileNetV2-based CNN architecture developed for the localization task.

strides 1, no padding) followed by a convolutional layer (1 filter, 3×3 kernel size and linear activation function) produces a real number as output, which is then approximated to the closest integer number. In this training phase, a mean squared error loss is used to predict the confidence score.

Once the training is completed, a convolutional implementation of the sliding window approach[61] is implemented in the test phase to predict the object coordinates: the trained CNN takes as input the original 256×256 images and produces a 27×27 heatmap as output. Each pixel of the heatmap corresponds to a delimited region of the input test image and represents the probabilities that the object is located in the center of that region: the position of the probability maximum provides the predicted coordinates.

The batch size is 32, the learning rate is 0.001, and the Adam algorithm is employed as the optimizer.

2.4 *CNNs Evaluation*

Performance statistics were computed on each of the five experiments (fivefold cross validation) mentioned above and then averaged to get the final statistics and associated standard errors.

The performances of the human observer and MO in detecting and localizing the object in each image were evaluated by complementary approaches that emphasize different aspects of the CNNs behaviors. The most adopted method in clinical practices is the receiver-operating characteristic (ROC) analysis.[62,63]

The ROC curve shows the tradeoff between sensitivity (or TPR, true positive rate) and specificity (1 - FPR, false positive rate), thus measuring the performance of a classification model. In the case of a double detection-classification task, each image is classified as true positive (TP), true negative (TN), false positive (FP), or false negative (FN) by taking into consideration the localization accuracy. The resulting curve is the localization-ROC (LROC).

The computation of LROC requires choosing an upper threshold distance between the actual center of the contrast object and the location indicated by the observers (human and model) to discriminate between correct and incorrect localization.[12]

An accurate analysis on the distribution of the human observers' localization responses, reported in Fig. S2 in the Supplementary Material, was carried out to establish the threshold distance values for the different insert diameters (summarized in Table 2). The knee algorithm was used to accurately determine these threshold values.[64]

Table 2 Selected thresholds for the LROC curves computation for different insert diameters.

Insert diameter (mm)	3	4	5	6	7
Threshold (mm)	2.3	2.3	2.5	3.0	3.5

The LROC curve was calculated for different images subsets, each characterized by one fixed parameter to highlight the dependence of the observer capability related to that parameter (i.e. object contrast, diameter, image reconstruction technique, and $CTDI_{vol}$).

Therefore, the area under the LROC curve (LAUC), a measure of the overall detection performance of the observers, was calculated for each object size and contrast, reconstruction technique, and $CTDI_{vol}$ and was averaged over the five cross-validation experiments conducted on the train dataset (see Sec. 2.3) with the associated standard deviation.

The differences between the LAUC curve of the human observer and those of the MO are evaluated by the mean of the absolute percentage error (MAPE),[65] which is a measure of prediction accuracy. The MAPE is calculated according to the following formula:

$$\mathrm{MAPE} = \frac{1}{N}\sum_{i}^{N}\frac{|\mathrm{LAUC}_i^{\mathrm{Model}} - \mathrm{LAUC}_i^{\mathrm{Human}}|}{\mathrm{LAUC}_i^{\mathrm{Human}}} \cdot 100, \tag{2}$$

where i is an index for the $CTDI_{vol}$ level and $N = 8$ is the total number of $CTDI_{vol}$ levels.

The LAUC analysis has been complemented by the evaluation of inter-rater indices[66] that quantify the level of agreement between two or more evaluators of the same observed situations. The multiraters Krippendorff's Alpha[67] (an interval level of measurement), the intraclass correlation coefficient (ICC,[68] a random single rating), the widespread Cohen's Kappa, and the more robust S-statistics[69,70] have been estimated to compare the agreement between models and human observers. The first two indices are conventional statistical indices that, however, suffer from a limitation related to unbalanced datasets. Kappa and S-statistics are normalized at the baseline of random chance: they describe how much better a classifier performs than that of a classifier that simply guesses at random according to the frequency of each class. A reference table with interpretation guidelines for the considered indices is in Table S1 in the Supplementary Material.[71–73]

Moreover, the accuracy metric was computed to address the performance of the trained CNN for both confidential scores and localization prediction, separately. Accuracy, defined as the ratio between the number of correct predictions and the total number of human evaluated images,[74] was calculated as a function of the relevant image parameters (diameters, contrasts, $CTDI_{vol}$, and reconstruction techniques).

In the case of localization accuracy, only the images containing the low-contrast object and having a score >0 were analyzed. The same thresholds distance values used to discriminate the true positive localization in the ROC computation (see Table 2) were applied to evaluate the localization accuracy.

3 Results

As a preliminary analysis, the performance of the human observers in the task of identifying low-contrast objects within the images was evaluated in terms of the LAUCs reported in Fig. 7, as a function of $CTDI_{vol}$, in the case of the two reconstruction techniques (FBP top panel, IR bottom panel), for the different contrast values C expressed in terms of the HU difference from the PMMA background ($C = 45$ HU left panel, $C = 55$ HU right panel). Within each panel in Fig. 7, different curves refer to images with inserts of different diameters.

As expected, the human observer performance improves as $CTDI_{vol}$ increases, due to CNR increasing at larger radiation dose values. The detectability of the smallest objects (3 mm diameter) is poor for both contrasts C and remains below 90% even at high $CTDI_{vol}$. The noise in the CT images is correlated, which means that the noise in any point of the image is affected to some extent by the noise values of the neighboring points. Small objects are affected the most by noise correlations. The calculated correlation distance for the highest $CTDI_{vol}$ images in our dataset, following Refs. 75 and 76, is ~0.7 mm, which can significantly change the appearance of the 3 mm inserts (1.5 mm radius) and thus make them difficult to be detected even at high radiation doses.

The curve related to the 4 mm diameter objects shows a significant increase with $CTDI_{vol}$, and saturation of the human observer performances, corresponding to LAUC approaching 1, is reached above 6 to 7 mGy of $CTDI_{vol}$, though this slightly depends on the contrast and

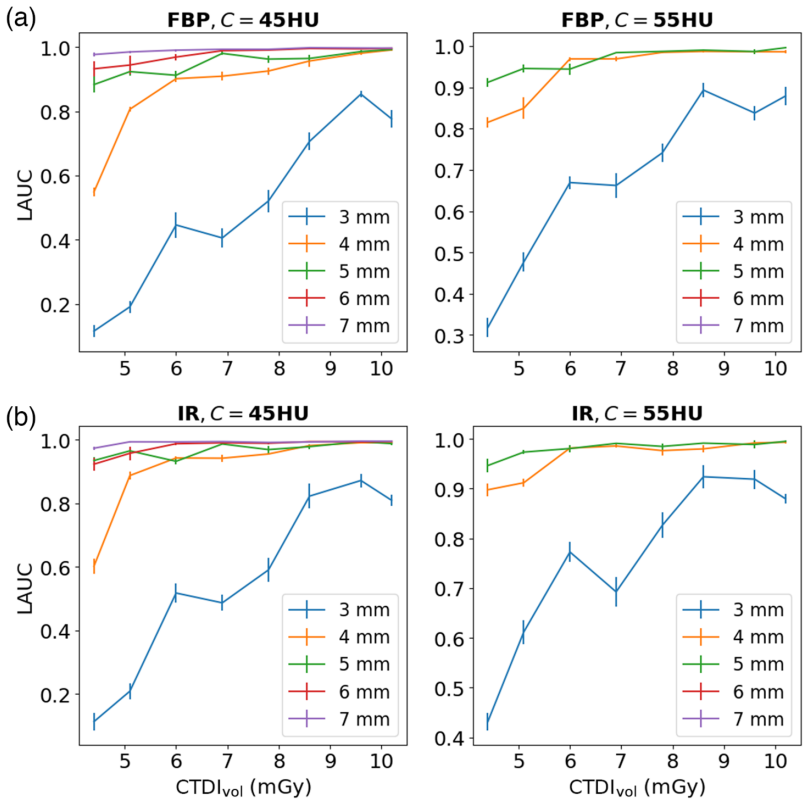

Fig. 7 Human observer performances quantified by LAUC versus $CTDI_{vol}$, for different object sizes, contrasts (left: 45 HU, right: 55 HU), and reconstruction techniques (a) FBP and (b) IR.

reconstruction technique: objects in IR reconstructed images are better recognized than in FBP reconstructed images.

In the case of inserts with diameters >4 mm, saturation of the human observer performance occurs even at low $CTDI_{vol}$, with LAUC values always above 80%. This result justifies the preliminary selection of the number of images for each contrast and size, as reported in Table 1, in which the more populated subsets are those relative to the inserts of 3 and 4 mm diameters.

Additional statistical analysis was performed to evaluate the difference among the human observers and between the two professional categories that took part in the visual inspection of the CT dataset: radiologists and medical physicists. The LAUC computed for the two classes, reported in Fig. S3 in the Supplementary Material, shows slightly higher performances of the radiologists, especially in the case of the less visible inserts (corresponding to 3 mm diameter inserts). This difference would be, of course, much more significant in the case of complex images, but the scope of this work lies outside the usage of diagnostic images: we aim to exploit the advantages given from using a simple phantom, which can be acquired under different user-defined CT settings (such as protocols and $CTDI_{vol}$) and in different CT scanners.

Given the above assertions, to achieve the optimization of CNNs, which are notoriously affected by overfitting and biases due to limited data selection, the increased dataset and label variability can be considered an added value, provided that the significant results of this research originates from the analysis of the inserts >3 mm (and, in particular, of the reference dataset).

The performances of trained convolutional neural networks were quantified by computing LAUC as well. The comparison of LAUCs extracted from the whole test dataset among the three observers (two CNNs and the human observer) is shown in Fig. 8, with associated standard errors, in the case of different reconstruction techniques (FBP left panel, IR right panel). A very good agreement is noticeable between the two CNNs and the human observer, especially in the case of IR reconstruction. It can be noticed that at the highest $CTDI_{vol}$ (10 m Gy) LAUC values of MobileNet show a decrease of performance, an anomalous trend that is present also for some LAUC curves of the human observer in the case of signal with 3 mm diameter (see Fig. 7).

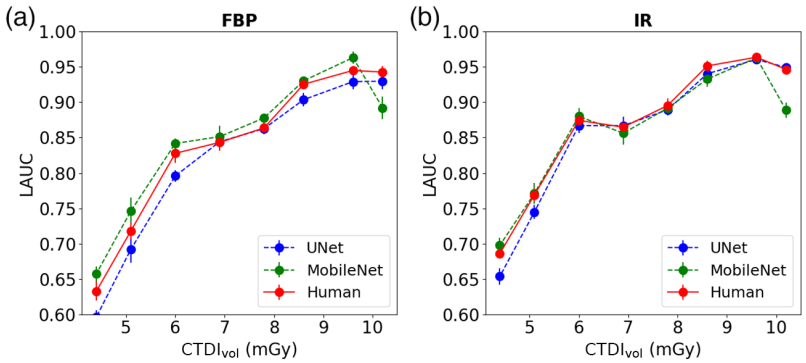

Fig. 8 Overall MO and human observer performances quantified by LAUC versus $CTDI_{vol}$, for the two reconstruction techniques (a) FBP and (b) IR, with associated standard errors.

This trend is under investigation, and further analysis are under way which include CNR evaluation of the images dataset. Our first hypothesis is that it can be related to the MobileNet noise modeling.

According to the previous consideration on the human observer performances, a reference LAUC was selected to evaluate the agreement between human observer and MO, as the one extracted from images containing 4 mm inserts with a lower concentration ($C = 45$ HU) was more explicative of the human observer behavior as a function of $CTDI_{vol}$. In addition, the iterative reconstruction technique, being the most common algorithm used by clinicians in CT protocols, was selected as the reference.

Figure 9 shows the LAUC values for the three observers as a function of $CTDI_{vol}$ in the case of the reference images subset (4 mm insert, $C = 45$ HU, IR reconstruction). The LAUC comparison in the case of the full dataset, i.e., at different diameters, contrasts, and reconstruction techniques, is reported in Figs. S4 and S5 in the Supplementary Materials. A very good agreement between the observers is qualitatively noticeable.

In order to address the utility of the full training dataset, we performed two additional CNN experiments by using reduced dataset, excluding the 6–7 mm inserts and the 3 mm inserts, respectively. The results of these new experiments are reported in the Supplementary Materials (Figs. S8–S10) and they support the evidence that the dataset variability and numerousness is essential, and it contributes to the success of the training in its totality.

In the following, the level of agreement among the observers is quantitatively addressed by means of appropriate metrics and statistical indices. The MAPE evaluated between the LAUC of

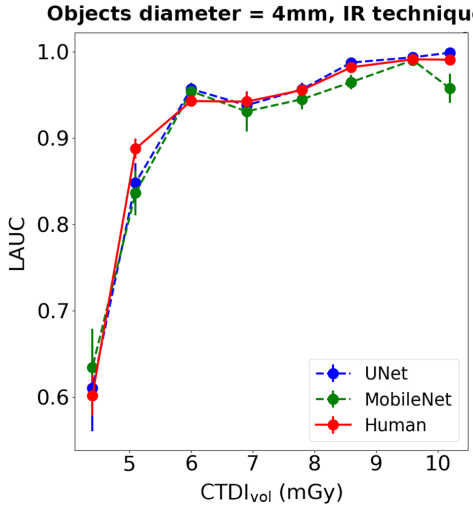

Fig. 9 Comparison of human observer and MO LAUCs versus $CTDI_{vol}$ for the images with an object size of 4 mm, $C = 45$ HU, and IR reconstruction, with associated standard errors.

Table 3 MAPE between human observer and MO LAUCs for full IR and FBP datasets and a representative case (Fig. 9).

CNN	FBP	IR	4 mm, IR, $C = 45$ HU
UNet	2.36	1.43	1.19
MobileNetV2	2.49	1.52	2.48

the human observer and the LAUC of the two MOs is summarized in Table 3 in the case of the full IR dataset, the full FBP dataset, and the reference subset. Excellent agreement is found between the trained CNNs and the human observer, with an MAPE below 2% in the case of IR reconstruction, slightly above 2% in the case of FBP reconstruction, and in general below 5% when considering all of the image subsets, each related to a different relevant parameter (diameter, contrast, and reconstruction technique), as reported in Tables S2 and S3 in the Supplementary Material. An exception is represented by the 3 mm inserts that are barely recognizable by either the human observer or the MO.

The accuracy metric, as defined in Sec. 2.4, was evaluated separately for the localization and score prediction tasks: the analysis results as a function of the different variables (inserts diameters, contrast C, $CTDI_{vol}$, and reconstruction technique) are reported in Figs. 10 and 11, respectively.

By looking at Figs. 10 and 11, it is noticeable that the UNet is able to localize slightly better than the MobileNetV2, whereas the latter classifies with a slightly higher overall accuracy than the UNet. This behavior appears consistent with the intrinsic character of the two CNNs, which were initially designed, as reported in the literature, for localization/segmentation and classification tasks, respectively.

Naively expected trends can be observed: the accuracy, in both tasks, increases with object size (insert diameter), contrast (C), $CTDI_{vol}$ (radiation intensity), and IR reconstruction.

In general, the localization accuracy is above 80%, and score prediction accuracy is well above 50%; once again an exception occurs for those images containing the 3 mm inserts.

Furthermore, other common multiclass statistical indices were computed to address the inter-rater agreement in the score prediction task. The values of Cohen kappa, S-statistics, Krippendorff's Alpha, and ICC, evaluated for the whole images dataset, are summarized in Table 4. The Cohen and S-statistics[77] indices show a fair to good agreement between the MOs and human observer score predictions, whereas Alpha and ICC indices show a good to excellent agreement (see also Table S1 in the Supplementary Material).

Moreover, consistent with the previous LROC analysis, a strong correlation between the S-statistics and $CTDI_{vol}$ is found for both CNNs, as shown in Figs. S6 and S7 in the

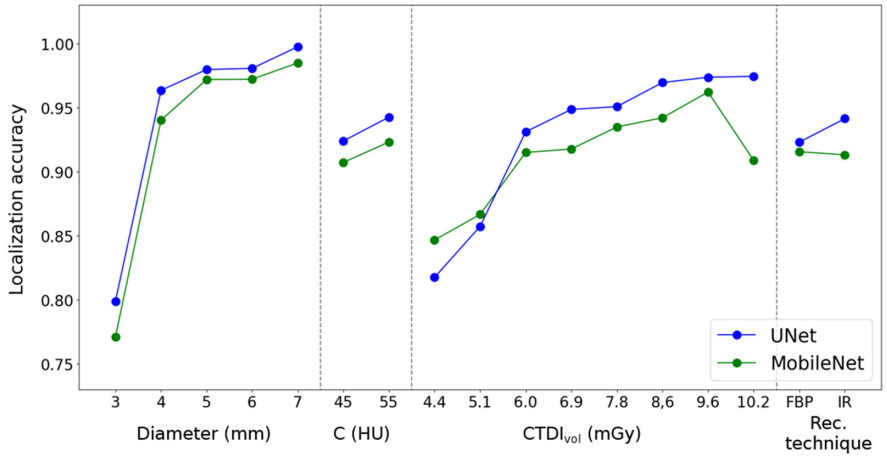

Fig. 10 MOs localization accuracy metric versus each of the independent parameters, from left to right: insert diameter, insert contrast, $CTDI_{vol}$, and reconstruction techniques.

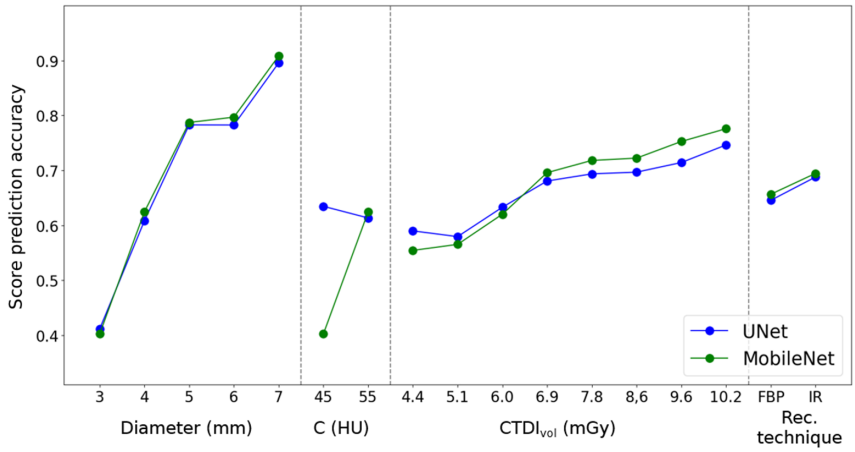

Fig. 11 MOs score prediction accuracy metric versus each of the independent parameters, as in Fig. 10.

Table 4 Human-model inter-raters statistical indices over the entire dataset.

CNN	Cohen kappa	S-statistics	Krippendorff's Alpha	ICC
UNet	0.5	0.56	0.77	0.77
MobileNetV2	0.53	0.57	0.83	0.83

Supplementary Material. These plots indicate that, when increasing $CTDI_{vol}$ and the insert diameter, the ability of the CNNs to predict the scores in agreement with the human observer increases as well. A very low value of the S-statistics and no correlation with $CTDI_{vol}$ are found in the case of the 3 mm insert: the very poor CNR in those images is such that the CNNs are mistaken because they cannot learn from the human observers answers, which are rather imprecise (see also Fig. 7). If the images containing 3 mm inserts are ruled out from the dataset to compute the S-statistics, index values are found to be 0.64 for the UNet and 0.66 for the MobileNnet, indicating a substantial agreement.

4 Discussion

In this work, we developed and characterized MOs based on artificial intelligence for automatic quality evaluation of phantom CT images. Two CNNs were trained to mimic human observer images assessment in terms of object detection and localization in CT images acquired on a specifically designed and manufactured phantom.

First, we collected a big dataset of phantom CT images containing objects of different sizes and contrasts, acquired at different $CTDI_{vol}$ settings and reconstructed by means of different techniques (FBP and IR). The dataset was initially submitted to the visual evaluation by human observers to collect the labels necessary for the algorithm training and testing. In this way, the labels fully reflect the human observers' interpretation of the CT images, regardless of the correctness of human images interpretation, and no internal noise component is necessary to calibrate the CNN-MO on the average human performance.

To verify the viability of our ultimate goal, which is the possibility of CT protocol optimization by means of CNN-MOs, in an almost independent way from the chosen CNN, we implemented two different architectures. UNet and MobileNetV2 were originally built and optimized for the tasks of segmentation and classification, respectively. To the above mentioned purpose, the relation between the CNN architecture and performance was also investigated. We found that both models performed quite similarly, suggesting that there are no critical aspects preventing MO application of them.

In the case of human observers, the confidence score (classification task) and the localization task are intrinsically interconnected and cannot be disentangled in the image evaluation process, whereas CNNs need to be specifically trained to carry out the two tasks, which are partially independent, using different loss functions. The accuracy metrics (Figs. 10 and 11) and the inter-rater agreement statistics (Table 4) show a trend in accordance with the above observation: the two CNNs have different performances in the two tasks of classification (MobileNet is slightly superior) and localization (UNet is slightly superior), as expected. However, the predictions of the two tasks, when combined together, for both CNN-MOs achieve very good overall performances, measured in terms of the LAUC metric. This result supports the robustness of the proposed approach and its being fairly independent from the CNN used. The quality of the trained CNNs was quantified by several statistical indices describing the inter-rater agreement between the MO and human observer in the confidence score task. The statistics computed on the full dataset, ruling out the 3 mm inserts with human detectability that is affected by strong noise correlation, give values of the robust S-statistics above 0.64 for both CNNs, indicating good general CNN performances.

The evaluation of the overall performance of the proposed algorithms in reproducing the human observer response was carried out by computing LAUC, a more accurate metric than AUC because it takes into consideration both localization and classification capability.[29] The MAPE calculated between LAUCs extracted from human observer and MO responses was found to be below 2.5%, with slightly higher performances in the case of IR reconstructed images. In addition to the LAUC averaged on the full dataset, we chose a reference subset of images (with a 4 mm insert, $C = 45$ HU, and IR reconstruction) reflecting a significant trend as a function of $CTDI_{vol}$: the LAUC extracted from human observer data of the reference subset (Fig. 9) covers a wide range of values, showing poor detectability performances at low $CTDI_{vol}$ and then rising until saturation. This curve is a suitable starting point for developing an optimization strategy for the current CT protocol: the value of 6 mGy can be considered the optimum $CTDI_{vol}$, above which there is no increase in detectability performances, and thus it is reasonable to expect a plateau of the diagnostic accuracy also.

In this work, we demonstrated the viability of an image quality assessment approach based on phantom acquisitions and CNN-MOs, which has a remarkable potential for improvements toward the final goal of dose optimization.

There are several pitfalls and perspectives to consider. We acknowledge several limitations in this study that we plan to address in future works. To finally achieve and implement a CT optimization program, a much more variable CT image dataset, acquired by different CT scanners with well defined setting parameters of a chosen CT protocol is needed. From this perspective, the proposed algorithms need to be retrained on the new dataset, after collection of new human-labeled data. However, we believe that, given the potential of the deep learning methods, the above mentioned effort, representing the next step of the ongoing research, enriched with elevated generalization capability of the algorithms, will be able to avoid the need to repeat the time-consuming reader studies for each protocol. Other limitations that we plan to address in future studies are the decrease of MobileNet performances at the highest $CTDI_{vol}$, which can be correlated to the CNR and/or noise modeling by the CNN, and the optimization of the phantom design in terms of CNR, which in turn affects the objects detectability.

5 Conclusion

In this work, we have developed and investigated the applicability of two MO algorithms based on CNN-MOs trained to mimic human observer performances in the phantom CT image detection task. We have demonstrated that two very different AI algorithms are both able to achieve very good results, thus indicating that the proposed approach is robust and fairly independent from the CNN used.

The positive results encourage continuing the exploitation of the proposed methodology toward an automatic image quality assessment based on the evaluation of CT images acquired on a specifically designed phantom; this should foster a systematic optimization, and possible standardization, of the large number of CT protocols currently used in radiological facilities,

with the final goal of reaching the best tradeoff between radiation dose and image quality, which is an issue of utmost relevance in diagnostic radiology as emphasized by international organizations (ICRP, IAEA, EURATOM) focused on ionizing radiation risk and radiological protection.

Disclosures

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

Acknowledgments

The authors are very grateful to the medical staff members of the Radiology Departments who evaluated the CT image datset: Careggi University Hospital (Director Dr. Vittorio Miele) and Santa Maria Nuova Hospital (Director Dr. Roberto Carpi) in Florence; San Jacopo Hospital in Pistoia (Director Dr. Letizia Vannucchi); Santo Stefano Hospital in Prato (Director Dr. Maurizio Bartolucci). The authors acknowledge the support and the CT resources provided by the Radiology department of the Careggi University Hospital in Florence, Italy. The authors also acknowledge the Physics Department of Florence University and the UNISER (polo pluridisciplinare, Pistoia e Pescia, Italy) for allowing the use of computational resources, essential for the neural network training and optimization tasks.

References

1. European Commission, "Medical radiation exposure of the European population," *Rad. Prot.* **180**, 1–181 (2015).
2. International Commission On Radiological Protection (ICRP), "ICRP PUBLICATION 26: 1977 recommendations of the international commission on radiological protection," *Ann. ICRP* **1**(3), 1–77 (1977).
3. W. R. Hendee and F. M. Edwards, "ALARA and an integrated approach to radiation protection," *Semin. Nucl. Med.* **16**(2), 142–150 (1986).
4. M. Rehani, "ICRP and IAEA actions on radiation protection in computed tomography," *Ann. ICRP* **41**(3-4), 154–160 (2012).
5. M. Rehani, "Managing patient dose in computed tomography," (2000).
6. M. Rehani, "Managing patient dose in multi-detector computed tomography (MDCT)," (2007).
7. M. Rehani, "Dose reduction in CT while maintaining diagnostic confidence: a feasibility/demonstration study," Int. Atomic Energy Agency, Vienna TECDOC-1621 (2009).
8. The Council of the European Union, "Council directive 2013/59/EURATOM," (2014).
9. J. Vaishnav et al., "Objective assessment of image quality and dose reduction in CT iterative reconstruction," *Med. Phys.* **41**(7), 071904 (2014).
10. L. Noferini et al., "CT image quality assessment by a channelized Hotelling observer (CHO): application to protocol optimization," *Phys. Med.* **32**, 1717–1723 (2016).
11. D. Racine et al., "Objective assessment of low contrast detectability in computed tomography with channelized Hotelling observer," *Phys. Med.* **32**, 76–83 (2016).
12. S. Leng et al., "Correlation between model observer and human observer performance in CT imaging when lesion location is uncertain," *Med. Phys.* **40**(8), 081908 (2013).
13. M. Han, B. Kim, and J. Baek, "Human and model observer performance for lesion detection in breast cone beam CT images with the FDK reconstruction," *PLoS One* **13**, 1–16 (2018).
14. M. Han et al., "Investigation on slice direction dependent detectability of volumetric cone beam CT images," *Opt. Express* **24**, 3749–3764 (2016).
15. H. Gong et al., "Deep-learning-based model observer for a lung nodule detection task in computed tomography," *J. Med. Imaging* **7**(4), 042807 (2020).

16. H. Gong et al., "A deep learning- and partial least square regression-based model observer for a low-contrast lesion detection task in CT," *Med. Phys.* **46**(5), 2052–2063 (2019).
17. F. Kopp et al., "CNN as model observer in a liver lesion detection task for x-ray computed tomography: a phantom study," *Med. Phys.* **45**(10), 4439–4447 (2018).
18. F. H. Reith and B. A. Wandell, "Comparing pattern sensitivity of a convolutional neural network with an ideal observer and support vector machine," ArXiv abs/1911.05055 (2019).
19. W. Zhou, H. Li, and M. Anastasio, "Approximating the ideal observer and Hotelling observer for binary signal detection tasks by use of supervised learning methods," *IEEE Trans. Med. Imaging* **38**, 3142456–2468 (2019).
20. M. Alnowami et al., "A deep learning model observer for use in alterative forced choice virtual clinical trials," *Proc SPIE* **10577**, 105770Q (2018).
21. F. Massanes and J. Brankov, "Evaluation of CNN as anthropomorphic model observer," *Proc SPIE* **10136**, 101360Q (2017).
22. C. Castella et al., "Mass detection on mammograms: influence of signal shape uncertainty on human and model observers," *J. Opt. Soc. Am. A* **26**, 425–436 (2009).
23. Y. Zhang, B. T. Pham, and M. P. Eckstein, "Evaluation of internal noise methods for Hotelling observer models," *Med. Phys.* **34**(8), 3312–3322 (2007).
24. M. Han and J. Baek, "A convolutional neural network-based anthropomorphic model observer for signal-known-statistically and background-known-statistically detection tasks," *Phys. Med. Biol.* **65**, 225025 (2020).
25. F. Kopp et al., "CNN as model observer in a liver lesion detection task for x-ray computed tomography: a phantom study," *Med. Phys.* **45**, 4439–4447 (2018).
26. G. Kim et al., "A convolutional neural network-based model observer for breast CT images," *Med. Phys.* **47**(4), 1619–1632 (2020).
27. R. D. Man et al., "Comparison of deep learning and human observer performance for detection and characterization of simulated lesions," *J. Med. Imaging* **6**(2), 025503 (2019).
28. I. Lorente, C. K. Abbey, and J. G. Brankov, "Understanding CNN based anthropomorphic model observer using classification images," *Proc. SPIE* **11599**, 115990C (2021).
29. R. Swensson, "Unified measurement of observer performance in detecting and localizing target objects on images," *Med. Phys.* **23**, 1709–1725 (1996).
30. S. Doria et al., "Addressing signal alterations induced in CT images by deep learning processing: a preliminary phantom study," *Phys. Med.* **83**, 88–100 (2021).
31. H. Gong et al., "Deep-learning model observer for a low-contrast hepatic metastases localization task in computed tomography," *Med. Phys.* **49**(1), 70–83 (2022).
32. R. Yang and Y. Yu, "Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis," *Front. Oncol.* **11**, 638182 (2021).
33. Y. Weng et al., "NAS-Unet: Neural architecture search for medical image segmentation," *IEEE Access* **7**, 44247–44257 (2019).
34. X. Li et al., "H-DenseUNet: Hybrid densely connected unet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imaging* **37**(12), 2663–2674 (2018).
35. S. Qamar et al., "A variant form of 3d-UNet for infant brain segmentation," *Future Gener. Comput. Syst.* **108**, 613–623 (2020).
36. J. Tian et al., "Automatic couinaud segmentation from CT volumes on liver using GLC-UNet," *Lect. Notes Comput. Sci.* **11861**, 274–282 (2019).
37. A. Lou, S. Guan, and M. Loew, "DC-UNet: rethinking the U-Net architecture with dual channel efficient CNN for medical image segmentation," *Proc. SPIE* **11596**, 758–768 (2021).
38. J. Dolz, C. Desrosiers, and I. Ben Ayed, "IVD-Net: intervertebral disc localization and segmentation in MRI with a multi-modal UNet," *Lect. Notes Comput. Sci.* **11397**, 130–143 (2019).
39. P. Ahmad et al., "Context aware 3D UNet for brain tumor segmentation," *Lect. Notes Comput. Sci.* **12658**, 207–218 (2021).
40. U. Latif et al., "An end-to-end brain tumor segmentation system using multi-inception-unet," *Int. J. Imaging Syst. Technol.* **31**(4), 1803–1816 (2021).
41. D. T. Kushnure and S. N. Talbar, "MS-UNet: a multi-scale unet with feature recalibration approach for automatic liver and tumor segmentation in CT images," *Comput. Med. Imaging Graph.* **89**, 101885 (2021).

42. I. Lorente et al., "Deep learning based model observer by U-Net," *Proc SPIE* **11316**, 113160F (2010).
43. A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," (2016).
44. T. van Erven and P. Harremos, "Rényi divergence and Kullback–Leibler divergence," *IEEE Trans. Inf. Theory* **60**(7), 3797–3820 (2014).
45. X. Yang et al., "Learning high-precision bounding box for rotated object detection via Kullback–Leibler divergence," CoRR abs/2106.01883 (2021).
46. S. Ji et al., "Kullback–Leibler divergence metric learning," *IEEE Trans. Cybern.* **52**(4), 2047–2058 (2022).
47. F. Martín et al., "Kullback–Leibler divergence-based global localization for mobile robots," *Rob. Auton. Syst.* **62**(2), 120–130 (2014).
48. D. I. Belov and R. D. Armstrong, "Distributions of the Kullback–Leibler divergence with applications," *Br. J. Math. Stat. Psychol.* **64**(2), 291–309 (2011).
49. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature* **521**, 436–444 (2015).
50. M. Sandler et al., "MobileNetV2: inverted residuals and linear bottlenecks," (2018).
51. M. Akay et al., "Deep learning classification of systemic sclerosis skin using the MobileNetV2 model," *IEEE Open J. Eng. Med. Biol.* **2**, 104–110 (2021).
52. C. Buiu, V.-R. Dănăilă, and C. N. Răduţă, "MobileNetV2 ensemble for cervical precancerous lesions classification," *Processes* **8**(5), 595 (2020).
53. A. Kanadath, J. A. A. Jothi, and S. Urolagin, "Histopathology image segmentation using MobileNetV2 based U-net model," in *Int. Conf. Intell. Technol. (CONIT)*, pp. 1–8 (2021).
54. R. Roslidar et al., "A study of fine-tuning cnn models based on thermal imaging for breast cancer classification," in *IEEE CybernéticsCom Int. Conf.*, pp. 77–81 (2019).
55. R. Indraswari, R. Rokhana, and W. Herulambang, "Melanoma image classification based on MobileNetV2 network," *Proc. Comput. Sci.* **197**, 198–207 (2022).
56. S. Taufiqurrahman et al., "Diabetic retinopathy classification using a hybrid and efficient MobileNetV2-SVM model," in *IEEE Region 10 Conf. (TENCON)*, pp. 235–240 (2020).
57. T. Kaur and T. K. Gandhi, "Automated diagnosis of Covid-19 from CT scans based on concatenation of MobileNetV2 and ResNet50 features," in *Computer Vision and Image Processing*, S. K. Singh et al., eds., pp. 149–160, Springer Singapore, Singapore (2021).
58. M. M. Ahsan et al., "Detection of Covid-19 patients from CT scan and chest x-ray data using modified MobileNetV2 and lime," *Healthcare* **9**(9), 1099 (2021).
59. S. Serte, M. A. Dirik, and F. Al-Turjman, "Deep learning models for Covid-19 detection," *Sustainability* **14**(10), 5820 (2022).
60. S. Aggarwal et al., "Automated Covid-19 detection in chest x-ray images using fine-tuned deep learning architectures," *Expert Syst.* **39**(3), e12749 (2022).
61. W. Zhiqiang and L. Jun, "A review of object detection based on convolutional neural network," in *36th Chin. Control Conf. (CCC)*, pp. 11104–11109 (2017).
62. T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.* **27**, 861–874 (2006).
63. F. R. Verdun et al., "Image quality in CT: from physical measurements to model observers," *Phys. Med.* **31**, 823–843 (2015).
64. V. Satopaa et al., "Finding a "kneedle" in a haystack: detecting knee points in system behavior," in *31st Int. Conf. Distrib. Comput. Syst. Workshops*, pp. 166–171 (2011).
65. U. Khair et al., "Forecasting error calculation with mean absolute deviation and mean absolute percentage error," *J. Phys. Conf. Ser.* **930**(1), 012002 (2017).
66. M. Warrens, "Inequalities between multi-rater kappas," *Adv. Data Anal. Classif.* **4**(4), 271–286 (2010).
67. K. Krippendorff, "Computing Krippendorff's alpha-reliability," (2011).
68. P. E. Shrout and J. L. Fleiss, "Intraclass correlations: uses in assessing rater reliability," *Psychol. Bull.* **86**(2), 420–428 (1979).
69. D. Marasini, P. Quatto, and E. Ripamonti, "Assessing the inter-rater agreement for ordinal data through weighted indexes," *Stat. Methods Med. Res.* **25**(6), 2611–2633 (2016).
70. D. J. Arenas, "Inter-rater: software for analysis of inter-rater reliability by permutating pairs of multiple users," (2018).
71. B. Dawson, *Basic & Clinical Biostatistics*, 4th ed., pp. 57–61, McGraw-Hill (2004).

72. J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics* **33**(1), 159–174 (1977).
73. P. Armitage, "The design and analysis of clinical experiments," *Biometrics* **43**(4), 1028–1028 (1987).
74. H. Gifford and M. King, "Implementing visual search in human-model observers for emission tomography," in *IEEE Nucl. Sci. Symp. Conf. Rec. (NSS/MIC)*, IEEE, pp. 2482–2485 (2009).
75. K. M. Hanson, "Detectability in computed tomographic images," *Med. Phys.* **6**(5), 441–451 (1979).
76. R. F. Wagner, "Fast Fourier digital quantum mottle analysis with application to rare earth intensifying screen systems," *Med. Phys.* **4**(2), 157–162 (1977).
77. C. H. Shweta and R. C. Bajpai, "Evaluation of inter-rater agreement and inter-rater reliability for observational data: an overview of concepts and methods," *J. Indian Acad. Appl. Psychol.* **41**(3), 20–27 (2015).

Federico Valeri received his MS degree in physics and his postgraduate diploma in medical physics from Florence University in 2019 and 2022, respectively. His research interests include CT physics, radiomics, computer vision, and machine learning.

Elena Cantoni is a student of Medical Physics Specialization School at the University of Bologna, Italy. In 2022, she was a research fellow at the University of Florence in the field of development and optimization of innovative AI methods on CT medical imaging. In particular, she focused on the statistical analysis of model observers and development of imaging software for early detection of hepatocellular cacarcinoma.

Evaristo Cisbani received his PhD in physics, research, and development of innovative instrumentations for nuclear medicine, radiation therapy, and nuclear experimental physics. He has experience in design and realization of Cherenkov detectors, gaseous chambers, and gamma imaging devices, combined to and supported by simulation, data acquisition, image processing, and data analysis. He is involved in the development of new approaches, based on artificial intelligence, for the optimization of instrumentation design, improvement of medical image quality, and performance evaluation of AI-based systems.

Ilaria Cupparo is a medical physicist. Currently, she is working as a research fellow on a radiation protection project in collaboration with the radioprotection expert at the University of Florence. She is involved in some research projects concerning the development of neural networks and machine learning algorithms for medical image analysis. Within the field of artificial intelligence, her work is particularly focused on optimizing CT protocols.

Sandra Doria received her PhD in physics in 2017 and the specialization in medical physics in 2021. She is now permanent researcher in the National Research Council of Italy. The main scientific activities are placed at LENS (European Laboratory for Nonlinear Spectroscopy), a center of excellence at the University of Florence. Her areas of expertise include deep learning in medical imaging, dose optimization of computed tomography protocols, radiomics, statistical image analysis, Monte Carlo modeling, and radioprotection.

Cesare Gori, formerly director of the Health Physics Department at the University Hospital "Careggi" in Firenze, Italy, is now appointed by the University of Firenze, Italy, as a radiation protection expert. He is founding partner and honorary member of the Italian Association of Medical Physics (AIFM) and he is also AIFM delegate to the International Organization for Medical Physics (IOMP) Council.

Lorenzo Lasagni is a research fellow at the University of Florence. He received his specialization in medical physics at the University of Florence in 2022. His main area of expertise is in developing software for medical image analysis, with a focus on segmentation, computer-aided diagnosis, and explainable artificial intelligence. He provides support for teaching at the University of Florence and co-supervises thesis activities.

Lorenzo Nicola Mazzoni is a medical physics expert at the Medical Physics Unit of Prato-Pistoia, AUSL Toscana Centro. His activity is mainly focused on medical imaging and radiation protection. He is member of the European Federation of Organisations for Medical Physics (EFOMP), European and international matter committee, and the current primary contact of EFOMP with the International Commission on Radiological Protection (ICRP).

Valentina Sanguineti received her BSc degree in electronic engineering and information technology and her MSc degree in internet and multimedia engineering from the University of Genoa, Italy, in 2016 and 2018, respectively. She received her PhD in computer vision, pattern recognition, and machine learning in 2022, which was done in collaboration between the University of Genoa and the Italian Institute of Techlology (IIT), where she has been involved in research on audio and video processing using deep neural networks.

Diego Sona received his PhD in computer science from the University of Pisa, Italy, in 2002. He joined the Adaptive Advisory Systems Group at the Istituto Trentino di Cultura. In 2008, he moved to the Neuroinformatics Laboratory in FBK. From 2011 to 2020, he joined as a visiting scientist in the Pattern Analysis and Computer Vision Department, IIT. In 2021, he moved to Data Science for Health Unit. His research has been always on machine learning.

Adriana Taddeucci is a medical physicist at Florence University Hospital, Italy. Her main activity concerns the implementation of optimization principle in diagnostic radiology. She started working with Model Observer (CHO) applied to CT protocols 10 years ago. She is also an adjunct professor at the University of Florence, where she teaches radiological equipments principles for medical imaging, including related quality assurance programs.

Biographies of the other authors are not available.