Implementing Natural Image Quality Evaluator for Performance Indicator on Noise Artefacts Recovery in CT Scan

Rudy Gunawan, Student Member, IEEE, Yvonne Tran, Jinchuan Zheng, Senior Member, IEEE
Hung Nguyen, Senior Member, IEEE, and Rifai Chai, Senior Member, IEEE

***Abstract* — The two most common evaluators for CT scan denoising are Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index (SSIM). This paper offers an alternative evaluator by utilizing of Natural Image Quality Evaluator (NIQE) assessment to determine the performance of denoising work on noise artefact. The noise artefact was obtained during the cancer screening process and had a particular noise density pattern across the image. NIQE is one of the blind image assessments which rely on the measurable deviation of image patch as a reference; it can determine the improved quality of denoising image. Due to the method of comparison in NIQE, the two parameters: patch size and sharpness threshold, will play an essential part in getting the score compared with the result from the other evaluators (PSNR and SSIM).**

I. Introduction

Cancer screening (low dose CT scan/ LDCT) can increase patient survivability by initiating an early treatment [1]; however, less radiation causes an increased noise level on the image. This noise can lead to incorrect assessment, in which one research showed a correct 24.2% of cancer from 53,454 patients [2]. Reducing the tube current, collimation factor, bowtie filter, and exposure time can reduce the number of photons entering the target area, thus is called low dose radiation. The effect of photon absorption and scattering causes fewer photons to be detected, especially on a thicker target. The path with few photons may appear as a weak streak line over a solid tissue (muscle) in the image; others may appear as shot noise over non-solid tissue (lung). In general, the noise artefact in CT scans does not have uniformity across different tissue in the image.

There are many attempts to remove the noise artefact, from classical denoising to the advanced Convolutional Neural Network (CNN) method. And in those attempts, the evaluator primarily consists of two methods. The Peak Signal to Noise Ratio (PSNR) measures the ratio of signal (pixel in the image) between the target image and the original image. The improvement is shown when comparing the ratio from two targets, the image with noise, and after denoising. The second method is the Structural Similarity Index (SSIM) which measures three image parameters (brightness, contrast, and structural) between the target image and the original image [3]. The denoising improvement is similar to the PSNR method, in which a better score shows better denoising.

In the previous paper, a third method was offered as an alternative for the evaluator, Natural Image Quality Evaluator (NIQE), to measure the denoising improvement [4]. The NIQE method was initially used for a photo-like image [5] as a blind/ no reference image evaluator. This paper discusses the usage of the NIQE evaluator; three CNN models for denoising are used in conjunction with the scoring from PSNR and SSIM for comparison to NIQE. It will allow observing the NIQE parameters against the scoring result from PSNR and SSIM.

II. Methodology

A. Denoising data and the Evaluators

The image data for denoising work comes from the public library of the Cancer Imaging Archive (TCIA) [6] under the research title of LDCT-and-Projection-data. The dataset has two variants, the first one is the original CT scan, and the second one is the noise simulation on the projection data. The result resembled the LDCT scan image with a variation of around 6% [7]. With two sets of images, it was possible to conduct a CNN training. Then the obtained weight of CNN after training is used to perform the denoising work with a better improvement than the other four CNN models (U-Net, RedNet, Seg-Net, Deconv-Net). Further testing on this weight was performed on the actual LDCT scan images; the proposed Double U-Net model could improve the quality based on the PSNR and SSIM scores. The fact was further enhanced with the NIQE scoring comparison [4].

The 44 test images were taken from a dataset consisting of the resultant denoising images, the noisy images, and the standard dose images, For further exploring the NIQE method, especially on the patch size and the sharpness threshold. Figure 1 shows the scoring sources from three evaluator methods; while the higher scores in PSNR and SSIM indicate a better image (compared to the noisy score), the lower score in NIQE suggests a closeness with the reference model and thus has better quality (could be compared to the noisy score).

Rudy Gunawan, Jinchuan Zhang, Hung T. Nguyen and Rifai Chai are with School of Science, Computing and Engineering Technologies, Hawthorn, VIC 3122, Australia (e-mail: rgunawan@swin.edu.au, jzheng@swin.edu.au; hungnguyen@swin.edu.au; rchai@swin.edu.au)

Yvonne Tran is with Australian Institute of Health Innovation, Macquarie University, Macquarie Park NSW 2109 (e-mail: yvonne.tran@mq.edu.au).

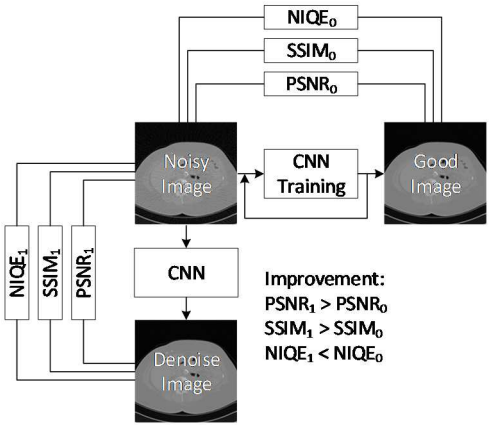

Figure 1. Denoising and three evaluator methods

The NIQE scores will be able to tell the improvement on one denoising method directly; however, in doing so, we cannot tell the difference in a quality pick-up with the other two evaluators. Thus, this research will compare three denoising methods concerning the score obtained from PSNR and SSIM. The three denoising methods would be U-Net, RedNet, and Double U-Net/ DU-Net (Figure 2).

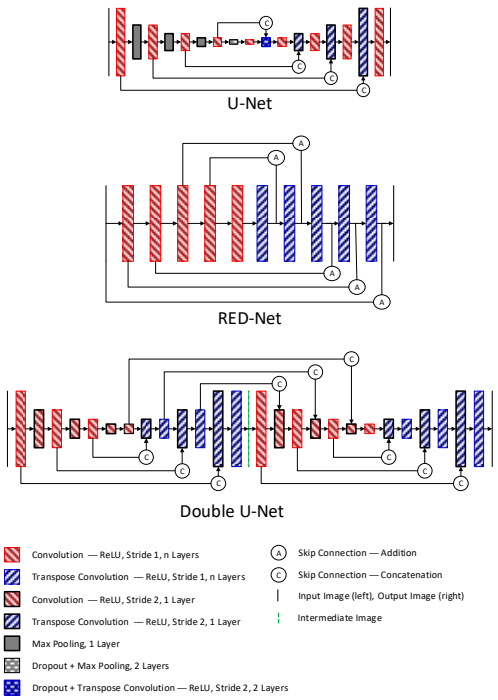

Figure 2. Three CNN models in which the data would be compared through NIQE score.

B. NIQE Evaluator

NIQE is a blind image quality assessment utilizing measurable deviation from statistical regularity on the image patches [5]. The NIQE scoring has a basis of comparison from the target image and the 'good' image model. The model is calculated by observing the image patches of the Natural Scene Statistic (NSS) for the statistical distribution (Gaussian Distribution – GD, Generalize GD, and Asynchronous GGD) and deviation (Mean Subtracted Contrast Normalized – MSCN) of images in the library [8].

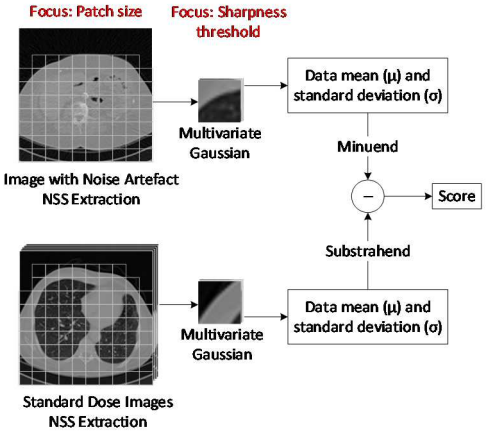

Figure 3. NIQE scoring

The NSS targets the area of interest (body) in the CT scan images with statistically significant features (Figure 3), extracted as a smaller patch. When calculating the Gaussian model, the sharpness threshold will provide a filter control on selecting the patch to be used. The focus indicates the parameter setting for NIQE on that process. A lower NIQE score indicates the target image closely resembles the model; therefore, it is considered better quality.

$$NIQE = SQRT((v_1-v_2)^{\tau}(\Sigma_1+\Sigma_2/2)^{-1}(v_1-v_2)) \quad (1)$$

The NIQE calculation is based on the square root (SQRT) of vector and covariance of the NIQE model (v_1 and Σ_1) and the target image (v_2 and Σ_2), with the sharpness threshold of the patch τ (1). The NIQE improvement score is obtained by reducing the noisy score to the denoise counterpart (2).

$$NIQE_{im} = NIQE_0 - NIQE_1 \quad (2)$$

with $NIQE_{im}$ as the improvement score, $NIQE_0$ as the score of a noisy image and $NIQE_1$ as the score of the image after going through the denoising process.

C. Scoring Comparison

PSNR and SSIM scores are used to determine the performance of the NIQE evaluator on different patch sizes and sharpness thresholds. Peak Signal to Noise Ratio (PSNR) measures the ratio between peak signal against noise in the image. This ratio can determine the amount of noise before and after denoising related to the maximum value of the image pixel.

$$PSNR = 20 \log_{10} (f_{max}/\sqrt{mse}) \quad (3)$$

The PSNR calculation uses a logarithmic scale between the maximum signal (f_{max}) and the mean square error (mse) between the target and reference image (3).

$$SSIM_{(t,r)} = [l_{(t,r)}]^{\alpha} + [c_{(t,r)}]^{\beta} + [s_{(t,r)}]^{\gamma} \quad (4)$$

The Structural Similarity (SSIM) measures the brightness/ luminance ($l_{(t,r)}$), contrast ($c_{(t,r)}$), and structural information ($s_{(t,r)}$) from the target (t) compared to the reference (r) image

(4). Brightness refers to the difference between the target image's signal mean intensity against the reference, contrast refers to the difference in the image's standard signal deviation, and structural information refers to the difference in the image's signal normalization. α, β, γ are the parameter controls that precede the instability when calculating the difference. This instability may occur when both images' signals are 0. Mostly these parameters are set into 1 to simplify the equation [3].

A higher PSNR or SSIM scores indicate the closeness of the image with the reference/ less pixel difference or having less noise, and score improvement before and after denoising can show the effectiveness of the denoising process.

$$\mathrm{PSNR_{im}} = \mathrm{PSNR_1} - \mathrm{PSNR_0} \quad (5)$$

$$\mathrm{SSIM_{im}} = \mathrm{SSIM_1} - \mathrm{SSIM_0} \quad (6)$$

The improvement score (PSNR_{im} or SSIM_{im}) is the difference between the score of the denoise image ($\mathrm{PSNR_1}$ or $\mathrm{SSIM_1}$) and the corresponding score of noisy images ($\mathrm{PSNR_0}$ or $\mathrm{SSIM_0}$) (5) (6).

The next step would be getting a normalization for the three evaluators; this step is essential since each score has a different value range. It will calculate the distance between the first and second ranks based on the full range (difference between ranks 1 and 3) (7).

$$\mathrm{Score_{sc}} = (\mathrm{Score_1} - \mathrm{Score_2}) / (\mathrm{Score_1} - \mathrm{Score_3}) \quad (7)$$

with $\mathrm{Score_{SC}}$ as a scaled score, then $\mathrm{Score_1}$, $\mathrm{Score_2}$, and $\mathrm{Score_3}$ are the improvement score on ranks 1 to 3.

III. Results

A. The dataset and PSNR and SSIM Score

Table I shows the average evaluator scores of PSNR and SSIM from the 44 images taken at random. It shows the Red-Net and U-Net have an almost similar score, with a big jump on the DU-Net. There is a difference in the ranking of Red-Net and U-Net within PSNR and SSIM scoring; therefore, the test relies on the first rank DU-Net as a comparison.

TABLE I. PSNR and SSIM Evaluator Scores on 1st Test with a Random 44 Images

CNN model	*Evaluator Improvement Score*	
	PSNR (rank)	**SSIM (rank)**
DU-Net	10.61 (1)	0.00962 (1)
Red-Net	9.91 (3)	0.00938 (2)
U-Net	10.06 (2)	0.00937 (3)

DU-Net has double the number of layers compared to Red-Net or U-Net, which makes the improvement quality score stand out in both PSNR and SSIM. Another scoring was taken place with the second test on 78 images to ensure the score and rank do not have a relation with test image characteristics. Table II shows the same score ranking on both PSNR and SSIM.

TABLE II. PSNR and SSIM Evaluator Scores on 2nd Test with a Random 78 Images

CNN model	*Evaluator Improvement Score*	
	PSNR (rank)	**SSIM**
DU-Net	10.59 (1)	0.00921 (1)
Red-Net	9.87 (3)	0.00898 (2)
U-Net	10.03 (2)	0.00896 (3)

These two tests indicate the scoring process on the dataset is correct; it showed a higher difference between the first and second rank: 0.55 (1st test), 0.56 (2nd test) for PSNR, and 0.00015 (1st test), 0.00023 (2nd test) for SSIM.

B. The score of different NIQE settings

The result shows NIQE score can have a negative improvement score because the modeling from the image library consists of different image characteristics (i.e., contrast based on tissue type). It means the NIQE quality score on the denoising image is lower than the noisy image. It can give an incorrect impression of the improvement; this research takes only the positive value of the improvement score To reduce this impression impact. The test observes 63 different settings, a combination of patch sizes from 16 to 112, and a sharpness threshold from 0.1 to 0.9.

Table III shows the score ranking of DU-Net as a guide for selecting suitable NIQE settings. Only the setting where DU-Net comes in first place will be set, the score value should not be negative, and there should be a substantial difference between DU-Net and other CNN scores.

TABLE III. Top Scoring on Patch Size and Threshold, DU means DU-Net as the first rank, X means outside the criteria.

Patch Size [n × n]	*CNN on Sharpness Threshold*								
	0.1	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
16	DU	DU	DU	DU	DU	DU	x	DU	x
32	x	DU	x	x	x	x	x	x	x
48	x	x	x	x	x	x	x	DU	x
64	x	x	x	x	x	x	x	x	x
80	x	x	x	x	x	x	x	x	x
96	x	DU	x	x	x	x	x	x	x
112	x	x	x	x	x	x	x	x	x

The initial result indicates patch size [16×16] has multiple threshold available to use, then [32×32], [48×48], [80×80], and [96×96] has one possible threshold. Table IV shows the average score and the variation between the three CNNs. These settings show that DU-Net attains the first rank compared to other CNN. Then the normalization will get a comparable value to the previous evaluator scores.

After normalization on PSNR and SSIM, the indicator for a better setting relies on a bigger value.

TABLE IV. NIQE SCORES ON VARIOUS NIQE SETTINGS

[n × n] - Threshold	*Scores on CNN Model*		
	DU-Net	Red-Net	U-Net
[16×16] – 0.1	1.369	0.974	1.259
[16×16] – 0.2	1.769	1.406	1.691
[16×16] – 0.3	2.400	2.159	2.396
[16×16] – 0.4	2.535	2.335	2.532
[16×16] – 0.5	2.784	2.593	2.764
[16×16] – 0.6	3.381	2.972	3.339
[16×16] – 0.8	7.464	7.103	6.971
[32×32] – 0.2	0.606	0.358	0.536
[48×48] – 0.8	25.082	24.031	24.409
[96×96] – 0.2	8.390	8.379	7.802

As per Table V, PSNR and SSIM get a normalization score of 0.7713 and 0.9223 between the first and second ranks of CNN quality improvement. The best NIQE normalization score (comparable to PSNR and SSIM has [16×16] patch size and 0.8 sharpness threshold, followed by [48×48] - 0.8 at 0.6401; [32×32] - 0.2 at 0.2827; [16×16] – 0.1 at 0.2797; and [16×16] - 0.2 at 0.2155.

TABLE V. SCORE COMPARISON AFTER NORMALIZATION

Scoring Type	Normalization Score
PSNR	0.7713
SSIM	0.9223
NIQE [16×16] – 0.1	0.2797
NIQE [16×16] – 0.2	0.2155
NIQE [16×16] – 0.3	0.0148
NIQE [16×16] – 0.4	0.0121
NIQE [16×16] – 0.5	0.1095
NIQE [16×16] – 0.6	0.1021
NIQE [16×16] – 0.8	0.7325
NIQE [32×32] – 0.2	0.2827
NIQE [48×48] – 0.8	0.6401
NIQE [96×96] – 0.2	0.0177

The rest has a normalization score of 0.1 or less, which indicate almost similar quality between the first and second rank. Although it is usable as a quality evaluator, it does not have similar traits with PSNR or SSIM, which show a greater distance.

IV. CONCLUSION

The NIQE method is not straightforward when directly evaluating quality improvement when comparing different techniques (in this case, CNN denoising). It will need assistance from the other known methods, PSNR and SSIM. However, it can provide extra data into the performance indicator matrix once the settings are known. As the NIQE calculation depends on a library of images, it has a different approach to the other method, which resorts one on one comparison.

The patch size depends on the nature of the artefacts found in the image; as the CT scan noise artefacts have a non-uniform noise pattern, the smaller patch seems to perform better than the bigger patch. The initial processing is also crucial as there is a sharpness filter to construct the Gaussian modeling. The threshold depends on the patch size selection.

For a future direction, observation of different noise densities (radiation level) and another patch size (not within factor 16) greater than 112 could be possible. And because this NIQE model is one of the blind image quality evaluators, it will be interesting to explore other blind evaluators, such as Perception based Image Quality Evaluator (PIQE) and Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) for the same purpose of finding the improvement quality of denoising CT scan images.

REFERENCES

[1] H. J. de Koning, C. M. van der Aalst, P. A. de Jong, E. T. Scholten, K. Nackaerts, M. A. Heuvelmans, J. J. Lammers, C. Weenink, U. Yousaf-Khan, N. Horeweg, S. van 't Westeinde, M. Prokop, W. P. Mali, F. A. A. Mohamed Hoesein, P. M. A. van Ooijen, J. Aerts, M. A. den Bakker, E. Thunnissen, J. Verschakelen, R. Vliegenthart, J. E. Walter, K. Ten Haaf, H. J. M. Groen, and M. Oudkerk, "Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial," *N Engl J Med,* vol. 382, no. 6, pp. 503-513, Feb 6, 2020, Jan.

[2] D. R. Aberle, F. Abtin, and K. Brown, "Computed Tomography Screening for Lung Cancer: Has It Finally Arrived? Implications of the National Lung Screening Trial," *Journal of Clinical Oncology,* vol. 31, no. 8, pp. 1002-1008, 2013, Feb.

[3] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing,* vol. 13, no. 4, pp. 600-12, Apr, 2004.

[4] R. Gunawan, Y. Tran, J. Zheng, H. Nguyen, and R. Chai, "Image Recovery from Synthetic Noise Artifacts in CT Scans Using Modified U-Net," *Sensors (Basel),* vol. 22, no. 18, Sep 16, 2022.

[5] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "Completely Blind" Image Quality Analyzer," *IEEE Signal Processing Letters,* vol. 20, no. 3, pp. 209-212, 2013.

[6] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior, "The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository," *J Digit Imaging,* vol. 26, no. 6, pp. 1045-57, Dec, 2013, Jul.

[7] C. H. McCollough, B. Chen, D. R. I. Holmes, X. Duan, Z. Yu, L. Yu, S. Leng, and J. G. Fletcher, "Low Dose CT Image and Projection Data (LDCT-and-Projection-data) (Version 4) [Data set]," *The Cancer Imaging Archive*, 2020.

[8] S. Athar, and Z. Wang, "A Comprehensive Performance Evaluation of Image Quality Assessment Algorithms," *IEEE Access,* vol. 7, pp. 140030-140030, 2019.