Clinical Evaluation of Deep Denoising Filter applied on Bone Scan

Si Young Yie[1,2], Joon Hyung Gil[3], Jin Cheol Paeng[3], Jae Sung Lee[1, 2, 3, 4*]

[1]Interdisciplinary Program in Bioengineering, Seoul National University, Seoul, Korea
[2]Integrated Major in Innovative Medical Science, Seoul National University, Seoul, Korea
[3]Department of Nuclear Medicine, Seoul National University Hospital, Seoul, Korea
[4]Brightonix Imaging Inc., Seoul, Korea
[*]Corresponding Author: jaes@snu.ac.kr

Objectives:

Denoising in nuclear medicine can be done in various ways for various purposes. The advancement of imaging hardware allows higher sensitivity, which allows a better count statistic. On the other hand, denoising by filters allow better post-processing of observed count statistics. Of these filters, analytical filters have a limitation of acceptable noise level due to the low complexity of the model. However, data-driven filters have shown a higher tolerance on the noise level and robustness on noise levels which is beyond filter design limits.

We have previously shown the high performance of deep learning-based denoising methods. Moreover, we have shown that using the interpolation of the short scan and the output of self-supervised denoising network can reduce discrepancy from the full scan. We have also shown the relation between the supervised denoising and self-supervised denoising.

Bone scan is a common clinical practice performed to evaluate skeletal lesions and metastases of tumor using gamma camera. Despite its benefits, reducing scan time or radiotracer dose without loss of critical information is challenging. We applied deep learning-based denoising filters on quarter-time bone scan images and evaluated the clinical value of these results. We analyzed the performance of these filters in the clinician's point of view.

Methods:

For the process of developing the deep denoising filter and evaluation, ^{99m}Tc-MDP or DPD bone scan data of 250 patients were acquired (200 for training and 50 for evaluation) using a GE Discovery 670 scanner. From the list-mode data, we generated 5~50% time scan of the full scan duration to train the deep-denoising filter. We trained two denoising networks in supervised manner, Noise2FullCount (N2F) and self-supervised manner, Noise2Noise (N2N).

For evaluation, we generated quarter-time scan (QS) from the list-mode data and rendered

the output of N2F, N2N and the interpolation of QS and N2N making interpolated N2N (iN2N). A board certified nuclear medicine physician and resident conducted the evaluation and compared the unfiltered full scan (FS), unfiltered QS, N2F, N2N, and iN2N filtered scan. Different processing for each subject's scan were given random indices (A~E) which was not open to the evaluators. The evaluation was done in order of group A, B, C, D, E, and group A was reevaluated for consistent result.

The evaluation focused on assessment of spatial resolution on large (ribs, femur) and small skeletal structures (T-L spine, radius, ulna, and fingers), noise level and clarity of the scan, and region, uptake and distinction of major findings. The spatial resolution, noise level, clarity and the distinction of findings was graded 1~4. Regarding the findings, the region was divided into head & neck, chest, T-L spine, pelvis, arm & leg, foot & hand and two findings were recorded at most. The uptake was graded -1~+3.

After the evaluation, the performance of denoising filters were compared to the full scan data. We performed a Wilcoxon test to check for significant difference of diagnostic performance between denoising methods and full scan. We also analyzed the inter-observer agreement by calculating Intraclass correlation (ICC) .

Results:

	Evaluator 1					Evaluator 2				
	QS	N2N	N2F	iN2N	FS	QS	N2N	N2F	iN2N	FS
Res. L.	2.79	3.78	3.49	3.66	3.56	2.52	3.26	3.09	3.13	3.16
Res. S.	2.60	3.42	3.25	3.42	3.26	2.11	2.90	2.61	2.80	2.80
Noise	2.50	4.00	3.76	4.00	3.76	1.74	3.98	2.98	3.08	2.68
Clarity	3.94	3.02	3.82	3.62	3.96	4.00	1.94	3.28	3.20	3.92
Uptake	1.34	1.69	1.64	1.67	1.59	1.53	1.63	1.58	1.60	1.56
Distinct	2.80	3.38	3.46	3.51	3.47	2.96	3.43	3.57	3.58	3.53

Table 1. Mean grade of QS, N2N, N2F, iN2N, and FS for each evaluator (1, 2) on spatial resolution (large, small), noise level and clarity, uptake and distinction of findings

In terms of spatial resolution of large and small structures, both evaluators gave the highest grade to the N2N filtered scan, and the following were iN2N or FS, N2F, and finally QS. In terms of noise level, both evaluators agreed on grading higher in order of N2N, iN2N, N2F, FS, and QS. In terms of blurriness, evaluators gave the highest grade to the FS or the QS, and the following were N2F, iN2N, and N2N. In terms of uptake in findings, both evaluators agreed on grading higher in order of N2N, iN2N, N2F, FS, and QS. In terms of distinction of findings, the evaluators gave the highest grade to iN2N filtered scan, and the following were N2F or FS, N2N, and finally the QS.

On performing a Wilcoxon test on denoising methods against the FS, the difference of spatial

resolution was significant only in QS. The difference of noise level was significant in N2N, iN2N and QS, and the difference of blurriness was significant in only N2N and partly N2F and iN2N. The uptake level and distinction in findings did not show significant difference from the FS. In terms of inter-observer agreement, the distinction of findings showed poor reliability and other evaluations showed moderate reliability.

Conclusion:

The deep denoising filters applied on quarter time scan showed similar or better performance in clinical diagnosis except blurriness. Of the deep denoising filters, N2N showed the best performance while presenting blurry features. The supervised denoising N2F and self-supervised denoising iN2N showed less blurriness while showing similar performance to N2N. Finally, the deep denoising filters potentially have the strength to be used in clinical settings.