Preliminary Exploration Using MNIST Dataset Highlights Autoencoders Limitations in Denoising Medical Images

Nicolas Friley
nfriley@stanford.edu

Abstract - Computed tomography (CT) imaging is widely used for diagnosing medical conditions, but it exposes patients to ionizing radiation. To minimize radiation risk, low-dose CT scans are recommended, but they often suffer from higher noise levels, compromising their diagnostic value. This paper evaluates the performance of an autoencoder-based denoising method for enhancing the quality of very-low-dose CT images. The autoencoder is trained on a simplified dataset of grayscale images, leveraging the similarities between handwritten digits and anatomical structures in chest CT scans. The performance of the autoencoder is evaluated under various noise levels using the Mean Squared Error (MSE) loss function and the Peak Signal to Noise Ratio (PSNR) metric. Comparative analysis with other denoising methods is also performed. The results show that the autoencoder outperforms other methods quantitatively and qualitatively at medium and high noise levels. However, the decoded images occasionally exhibit shape discrepancies, limiting the trust and interpretability of the encoded space. Alternative evaluation metrics, such as the Structural Similarity Index Metric (SSIM) and Vessel Sharpness Metric, are suggested for medical imaging tasks. Additionally, exploring loss functions sensitive to Poisson noise characteristics is recommended. The study highlights the importance of qualitative checks and the challenges associated with fully controlling and trusting the encoded space in autoencoders for medical imaging applications.

---

1 - INTRODUCTION

CT imaging plays a crucial role in diagnosing internal injuries, trauma, clots, or other medical conditions. However, the use of CT scans exposes patients to ionizing electromagnetic radiation. The radiation dose is determined by tuning the x-ray tube current according to the size and weight of the patient. While CT scans provide valuable diagnostic information, the potential risk associated with radiation exposure should be carefully considered and minimized. Although the risk of developing cancer from x-ray radiation remains relatively small, it is essential to mitigate this risk as much as possible. Additionally, low-dose CT scans are recommended for screening patients at high risk of developing lung cancer based on factors such as age, medical history, and smoking habits. However, decreasing the radiation dose significantly can introduce higher noise levels in CT scan images, compromising their accuracy and diagnostic value. Therefore, there is a need to develop an efficient denoising method that can enhance the quality of very-low-dose CT scan images.

Various architectures have been proposed for denoising tasks on medical images, including autoencoders, U-nets, and Conditional Generative Adversarial Networks (C-GANs) [2][3]. For the scope of this project, we have chosen to implement an autoencoder model using a simplified dataset of grayscale images and evaluate its performance in denoising tasks.

To avoid value jumps between the area outside of the body (in black in the CT scan images displayed in fig. 1) and the rest of the image, a research team [7] suggested cropping a square at the center of the image. Considering the idea of feeding small patches of this cropped square into the autoencoder model, we obtained sample patches of size 28x28 pixels from various 512x512 pixels CT scan images. To simulate this "crop-and-patchify" approach, we will train the autoencoder on a simplified dataset that shares similarities with these patches. By training the autoencoder on this dataset, we aim to analyze its performance in denoising tasks and assess its potential for enhancing very-low-dose CT image quality.

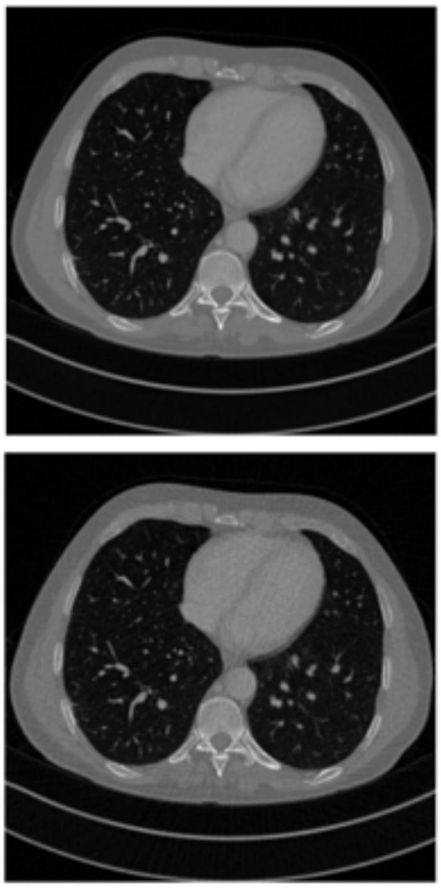

Fig. 1: Chest CT-scan images with full-dose radiation (top) and quarter-dose (bottom)

2 - RELATED WORK

In the field of medical image denoising, researchers have explored various approaches to achieve accurate results while minimizing radiation dose in CT examinations. Simulated low-dose CT scans with realistic noise levels have been used, where a simple Poisson model is often sufficient for generating accurate low-dose images [1]. One approach that has shown promise is the use of autoencoders [5]. This work specifically focuses on the development of a denoising convolutional autoencoder. The autoencoder generates encoded representations of medical images and reconstructs them to remove noise [6]. Another research team managed to effectively remove CT noise patterns using a modified U-net architecture trained on extracted patches [2]. Conditional generative adversarial networks (CGAN) have been proposed as another method for denoising CT scan images. This approach outperforms other techniques like total-variation minimization and non-local means [3].

3 – PROPOSED METHOD

3.1 - Dataset

The MNIST dataset is a collection of 70,000 grayscale images of handwritten digits ranging from 0 to 9 (fig. 2). Each image has a size of 28x28 pixels. This dataset was originally designed for handwritten digit recognition. Yet, it provides a suitable starting point for our exploratory analysis. We will leverage the similarities between the shapes of the handwritten digits and anatomical structures to approximate the denoising task on CT images. While the MNIST dataset does not directly represent medical images, it offers a simplified representation of anatomical structures that can be found in CT scans, such as pulmonary bronchi, arteries, veins, or nodules (fig.3).

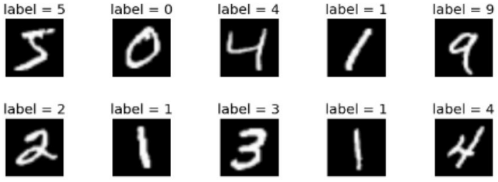

Fig. 2: MNIST sample images of size 28x28

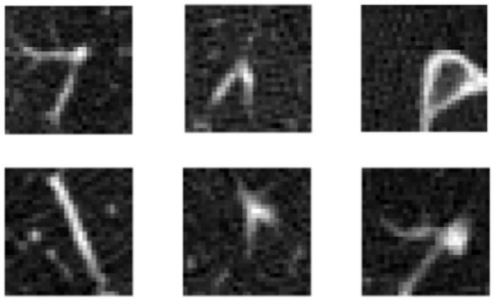

Fig. 3: Sample patches of size 28x28 from chest CT scan images

3.2 - Noise Addition

In our approach, we introduced noise to the MNIST dataset and trained an autoencoder model from scratch using the noisy images. We conducted separate training runs for different types of noise and noise levels to evaluate the autoencoder's performance under various settings. Specifically, we added Gaussian and Poisson noise at low, medium, and high levels.

3.3 - Loss Function

To train the autoencoder model, we employed the Mean Square Error (MSE) loss function. This loss function compared the reconstructed (decoded) image and the ground truth image, which represents the original image before the addition of noise.

3.4 - Evaluation Metrics

We chose the Peak Signal to Noise Ratio (PSNR) as the evaluation metric to assess the performance of the autoencoder. The PSNR metric allowed us to compare the denoising performance of the model across different noise levels. We also used the PSNR to compare the autoencoder's performance against other denoising methods on the same noise level.

3.5 - Comparison Methods

In addition to the autoencoder, we used several other denoising methods for comparative analysis. These methods included Median Blur, Gaussian Blur, Average Blur and Bilateral Filter. To provide a comprehensive evaluation, we also calculated the PSNR between the original image and noisy image, which served as a baseline for comparison.

3.6 - Autoencoder architecture

An autoencoder is a neural network architecture made of two parts, an encoder, and a decoder [fig.4]. The encoder takes as input the noisy image and converts it to a lower-dimensional representation called the encoded space. This first part contains a convolutional block and a linear block. In the convolutional block, the noisy image is processed through a series of convolutional layers with ReLU activation. The output of the convolutional block is flattened into a 1-dimensional vector and passed through a linear layer with ReLU to reduce the dimensionality of the representation even more. Lastly, the last linear layer maps the low-dimensional vector obtained to the encoded space.

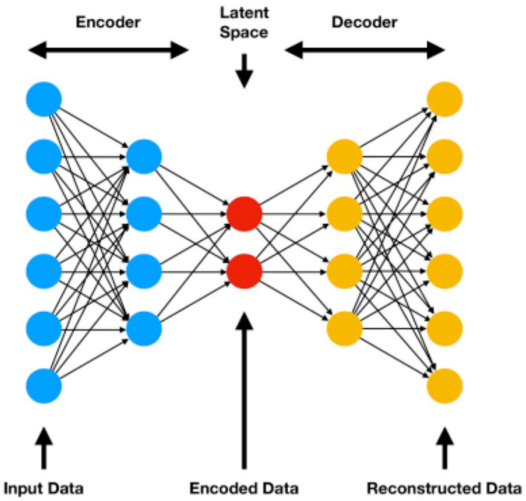

Fig. 4: General architecture of an autoencoder model

The decoder takes the encoded representation as input and tries to reconstruct the original image by mirroring the conversions done by the encoder, in reverse. It starts with a linear block where the encoded representation is passed through a linear layer with ReLU activation to increase the dimensionality of the representation back to the 1-dimensional vector. This vector is then reshaped into a 3-dimensional tensor. The convolutional block passes the tensor through deconvolutional layers, batch normalization and ReLU activations to up-sample the tensor. We used one last deconvolutional layer to generate the decoded image, which is then passed through a sigmoid activation function to force the pixel values in the interval [0,1].

4 – RESULTS AND ANALYSIS

4.1 - Loss Plots Analysis

In all six experiments, the plots of training and validation loss against the number of epochs exhibited a similar profile, resembling figure 5 and 6. The smooth convergence and the very small variance between the training and validation loss can be attributed to the regularization techniques and the choice of optimizer used during the training of the autoencoder. The incorporation of batch normalization, weight decay and use of the ADAM optimizer helped ensure that the model generalizes well on unseen data and does not overfit to the noise in the data.

Additionally, the consistency of the MNIST dataset played a crucial role in the smooth learning. The dataset exhibits no class imbalance and is well-structured, which ensured that all data points used during learning were of high quality. The synthetic noise that was added before training followed a known distribution. These elements contributed to helping the model learn and capture the underlying patterns in the data more efficiently.

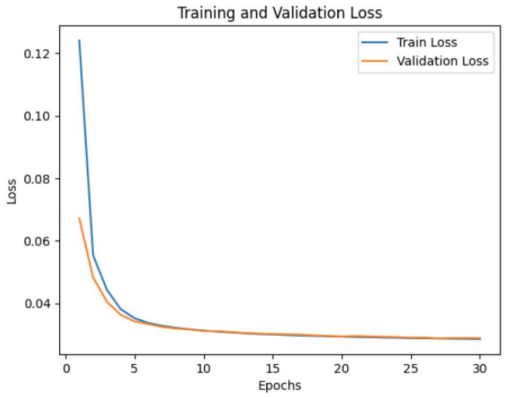

Fig. 5: Training and Validation Loss for Gaussian noise with medium intensity

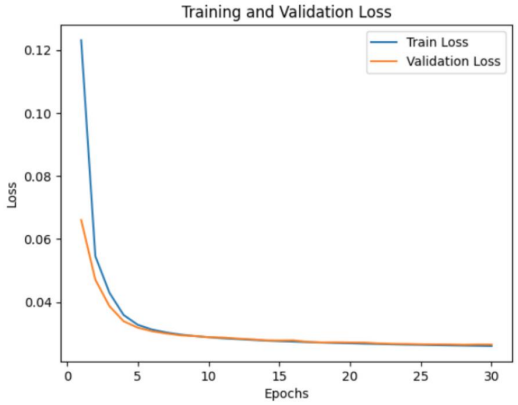

Fig. 6: Training and Validation Loss for Poisson noise with medium intensity

4.2 - Quantitative Analysis

Quantitatively, the autoencoder demonstrates superior performance compared to other denoising techniques at medium and high noise levels of both Poisson noise and Gaussian noise (table 1, table 2, fig. 7 and fig. 8 in appendix). However, when evaluating the PSNRs individually for randomly picked test images, the PSNR values across all methods varied significantly depending on the specific input image. For instance, more complex shapes tend to exhibit lower PSNR values across the board.

4.3 - Qualitative Analysis

Qualitatively, the autoencoder model outperforms other techniques in all 6 experiments. This distinction is particularly apparent when dealing with medium and high noise levels. However, it is worth mentioning that in the context of medical imaging analysis, many medical professionals prefer images with some level of noise and a perceived sharpness rather than images where shapes appear slightly blurred, even if the edges and contours of anatomical structures remain clearly identifiable [4]. In other words, although the autoencoder may visually outperform other techniques in terms of denoising, the added blur in the decoded image makes it less desirable compared to an image with less denoising but reduced apparent blur.

Figures 9 and 10 below compare the outputs of all denoising methods for different noise types and intensities. The first row from left to right showcases the ground truth image, the noisy image, the decoded image. The second row from left to right showcases the median blur, gaussian blur, bilateral filter, and average blur.

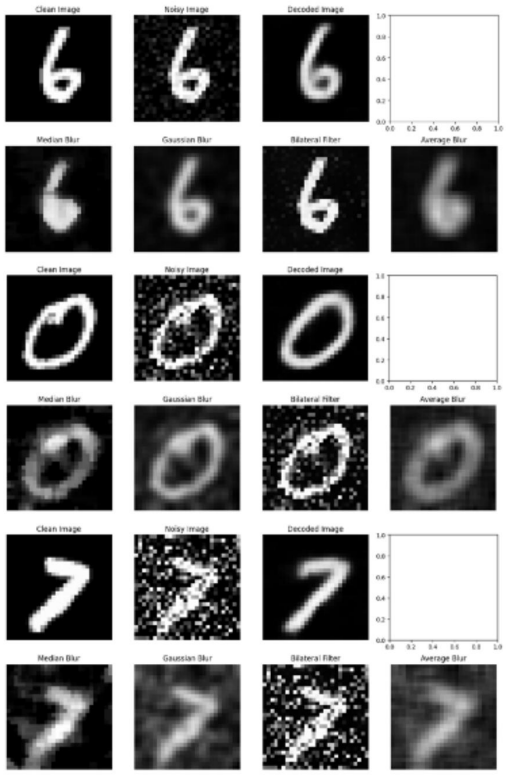

Fig. 9: Sample outputs from all denoising methods with different Gaussian noise levels ('6': low noise level, '0': medium noise level, '7': high noise level)

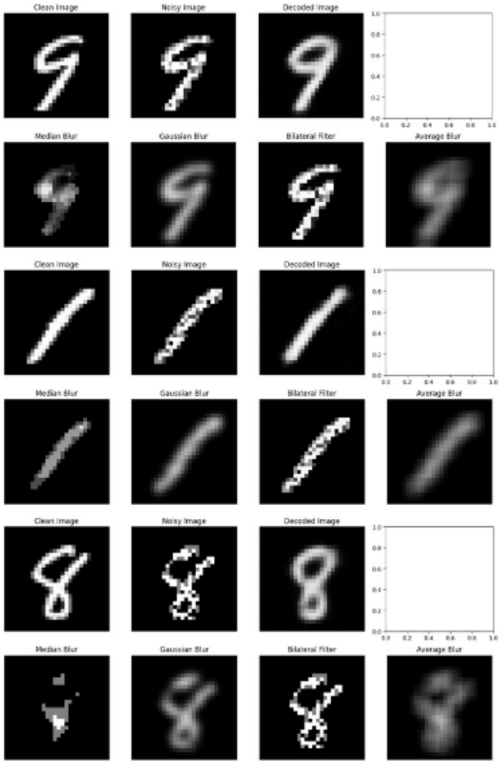

Fig. 10: Sample outputs from all denoising methods with different Poisson noise levels ('9': low noise level, '1': medium noise level, '8': high noise level)

4.4 - Shape Discrepancies

The decoded images occasionally exhibit shape discrepancies. This behavior is due to the nature of the autoencoder, which generates the decoded image statistically, based on the mapping it learned between the input image and the encoded space. In the provided examples in fig. 9, the decoded '7' displays a slight hook on the upper end of the digit, while the decoded '0' loses the details of the strokes extending inside the handwritten '0' in the input image. Similarly, the decoded '8' fails to capture the detail of the stroke discontinuity between its start and end point.

4.5 - Reconstruction Errors

In certain instances, the autoencoder had trouble mapping the input image to its corresponding encoded representation, resulting in an output that significantly differed

from the original input (fig.11). Surprisingly, despite these significant shape discrepancies, the calculated PSNR remained high. These reconstruction errors or misclassifications can be attributed to the limitations of the latent space within the autoencoder. But they highlight the potential risks associated with the autoencoder's ability to introduce alterations that completely undermine the diagnostic value of medical images. Radiologists, in particular, rely on detecting subtle details and identifying elements that may appear in certain CT scan slices but do not continue to others [4]. The presence of this kind of "inserted error" may confuse clinicians and erode their trust in the images.

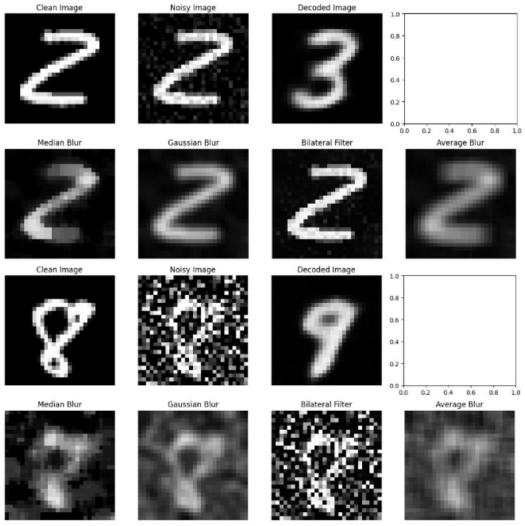

Fig. 11: Sample outputs with reconstruction errors

5 - DISCUSSION AND CONCLUSION

5.1 - Metrics

We found that the PSNR metric does not always accurately reflect the validity of the outputs. For evaluating the performance of the autoencoder on denoising tasks in medical imaging, other metrics like the Structural Similarity Index Metric (SSIM) and Vessel Sharpness Metric might be more suitable. The SSIM measures the similarity between the output and ground truth images and quantifies how well edges and spatial relationships are preserved. The Vessel Sharpness Metric, specifically used in medical imaging, measures the clarity of vessel boundaries or edges in images with blood vessels or vascular structures. In the context of chest CT scan, this metric could help evaluate the trade-off between sharpness and denoising effectiveness.

5.2 - Loss Functions

The autoencoder was trained using the MSE loss function. However, the MSE is more sensitive to Gaussian noise due to its assumption of symmetrically distributed errors around the true values. Since very-low-dose CT scan images are always simulated using Poisson noise, it might be useful to explore loss functions such as the Poisson negative log likelihood loss, which would be more sensitive to Poisson noise characteristics.

5.3 - Loss of diagnostic value

In the medical imaging field, the accurate representation of anatomical structures is crucial, and shape discrepancies can significantly affect the diagnostic value of the images. While misclassifications causing major discrepancies are primarily due to the limitations in the encoded space, it is very difficult to ascertain in practice whether or not the model's encoded space will introduce discrepancies in the image. Thorough qualitative checks are necessary to assess the risks and extent of these discrepancies. Autoencoders usually employ non-linear activation functions and several layers, making the mapping between the input and the encoded space very complex. As a result, the encoded representations learned by the model are difficult to interpret for humans. Specifically, the encoded features may not have a direct semantic meaning, making it challenging to interpret and control the encoded space in a meaningful way. This means that while an autoencoder may perform well for some tasks, its inherent complexity makes it almost impossible to have full control and trust over the encoded space, and therefore, the decoded output.

APPENDIX

Noise factor	0.1	0.3	0.5
Decoded Image	16.338	15.253	16.106
Noisy Image	22.44	13.309	9.452
Median Blur	17.653	14.151	15.469
Gaussian Blur	18.457	14.447	12.868
Average Blur	15.479	12.522	12.273
Bilateral Filter	25.407	13.761	9.61

Table 1: Comparison of PSNRs for Gaussian noise with different intensities

Noise factor	0.1	0.3	0.5
Decoded Image	17.865	23.291	16.082
Noisy Image	21.001	18.993	15.256
Median Blur	13.829	16.35	11.265
Gaussian Blur	17.576	18.537	15.565
Average Blur	14.678	16.349	13.655
Bilateral Filter	20.821	18.983	15.256

Table 2: Comparison of PSNRs for Poisson noise with different intensities

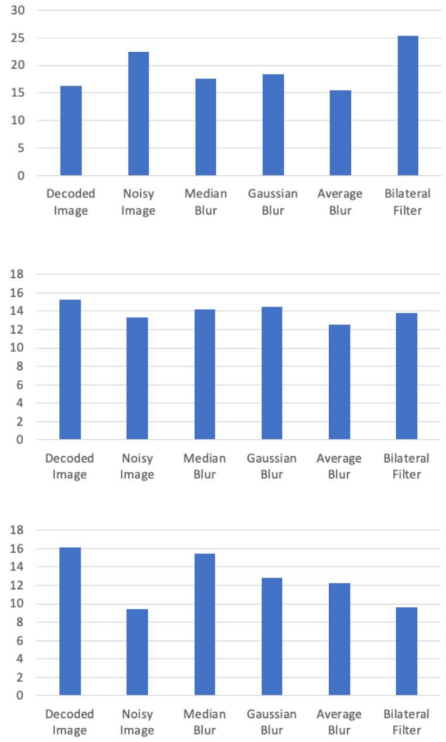

Fig. 7: Comparison of PSNRs for Gaussian noise with different intensities (from top to bottom: low, medium, high, from left to right: decoded image, noisy image, median blur, gaussian blur, average blur and bilateral filter)

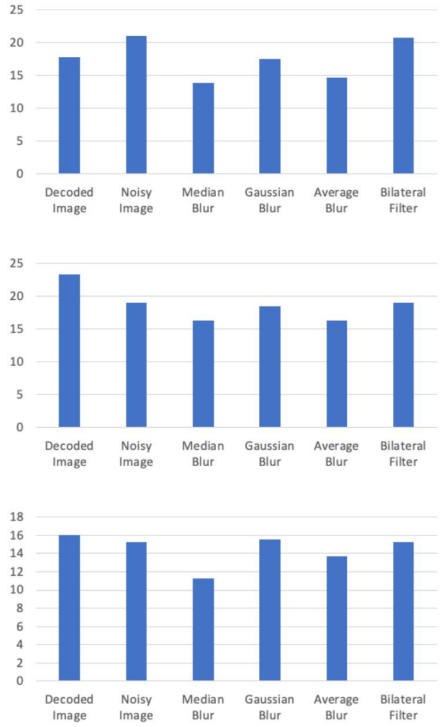

Fig. 8: Comparison of PSNRs for Poisson noise with different intensities (from top to bottom: low, medium, high, from left to right: decoded image, noisy image, median blur, gaussian blur, average blur and bilateral filter)

References

[1] Yu, Lifeng, et al. "Radiation Dose Reduction in Computed Tomography: Techniques and Future Perspective." Imaging in Medicine, vol. 1, no. 1, Oct. 2009, pp. 65–84, www.ncbi.nlm.nih.gov/pmc/articles/PMC3271708/, https://doi.org/10.2217/iim.09.5.

[2] Chulkyun Ahn, Changyong Heo, and Jong Hyo Kim "Combined low-dose simulation and deep learning for CT denoising: application in ultra-low-dose chest CT", Proc. SPIE 11050, International Forum on Medical Imaging in Asia 2019, 110500E (27 March 2019); https://doi.org/10.1117/12.2521539

[3] Hee-Joung Kim a b, et al. "Image Denoising with Conditional Generative Adversarial Networks (CGAN) in Low Dose Chest Images." Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, North-Holland, 28 Feb. 2019, www.sciencedirect.com/science/article/abs/pii/S0168900219302293.

[4] Krupinski, E. A. "Current Perspectives in Medical Image Perception." Attention, Perception & Psychophysics, vol. 72, no. 5, 30 June 2010, pp. 1205–1217, www.ncbi.nlm.nih.gov/pmc/articles/PMC3881280/, https://doi.org/10.3758/app.72.5.1205.

[5] J. M. Thomas and A. P. E, "Bio-medical Image Denoising using Autoencoders," 2022 Second International Conference on Next Generation Intelligent Systems (ICNGIS), Kottayam, India, 2022, pp. 1-6, doi: 10.1109/ICNGIS54955.2022.10079813.

[6] L. Gondara, "Medical Image Denoising Using Convolutional Denoising Autoencoders," 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), Barcelona, Spain, 2016, pp. 241-246, doi: 10.1109/ICDMW.2016.0041.

[7] Leuschner, Johannes, et al. "LoDoPaB-CT, a Benchmark Dataset for Low-Dose Computed Tomography Reconstruction." Scientific Data, vol. 8, no. 1, 16 Apr. 2021, p. 109, www.nature.com/articles/s41597-021-00893-z#citeas, https://doi.org/10.1038/s41597-021-00893-z. Accessed 8 Apr. 2022.