



OPEN Task based evaluation of sparse view CT reconstruction techniques for intracranial hemorrhage diagnosis using an AI observer model

Matthew Tivnan^{1,3}, Irene Désirée Kikkert^{2,3}✉, Dufan Wu¹, Kai Yang¹, Jelmer M. Wolterink², Quanzheng Li¹ & Rajiv Gupta¹

Sparse-view computed tomography (CT) holds promise for reducing radiation exposure and enabling novel system designs. Traditional reconstruction algorithms, including Filtered Backprojection (FBP) and Model-Based Iterative Reconstruction (MBIR), often produce artifacts in sparse-view data. Deep Learning Reconstruction (DLR) offers potential improvements, but task-based evaluations of DLR in sparse-view CT remain limited. This study employs an Artificial Intelligence (AI) observer to evaluate the diagnostic accuracy of FBP, MBIR, and DLR for intracranial hemorrhage detection and classification, offering a cost-effective alternative to human radiologist studies. A public brain CT dataset with labeled intracranial hemorrhages was used to train an AI observer model. Sparse-view CT data were simulated, with reconstructions performed using FBP, MBIR, and DLR. Reconstruction quality was assessed using metrics such as Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). Diagnostic utility was evaluated using Receiver Operating Characteristic (ROC) analysis and Area Under the Curve (AUC) values for One-vs-Rest and One-vs-One classification tasks. DLR outperformed FBP and MBIR in all quality metrics, demonstrating reduced noise, improved structural similarity, and fewer artifacts. The AI observer achieved the highest classification accuracy with DLR, while FBP surpassed MBIR in task-based accuracy despite inferior image quality metrics, emphasizing the value of task-based evaluations. DLR provides an effective balance of artifact reduction and anatomical detail in sparse-view CT brain imaging. This proof-of-concept study highlights AI observer models as a viable, cost-effective alternative for evaluating CT reconstruction techniques.

Computed tomography (CT) is a widely used imaging modality in diagnostic radiology, producing volumetric images given measured X-ray projections from multiple angles¹. Since the introduction of the first CT scan of the brain in 1971, continuous innovations have improved both acquisition speed and image quality². Head CT is now the standard of care for many applications, including detection and diagnosis of intracranial hemorrhage, which we focus on in this work^{2–5}. Modern third generation CT scanners with a rotating gantry can acquire over one thousand projection views per rotation, enabling the generation of high-resolution images^{1,2,6}. However, there are several limitations to these conventional CT systems. Firstly, ionizing radiation necessary for high-quality imaging poses significant public health risks, due to an increased risk of cancer^{7,8}. Secondly, these systems have a bulky design, substantial weight, and high power consumption, making them immobile and impractical for deployment in resource-limited environments⁹.

One emerging technology to address these issues is static CT, which replaces a rotating gantry with a fixed circular array of electronically controlled X-ray sources. This design reduces the system's size, weight, and power efficiency¹⁰. Fewer projection views and lower power sources naturally lend the system to low-radiation-dose applications¹¹. Despite the advantages, the sparse-view nature of static CT measured data presents significant

¹Radiology Department, Harvard Medical School & Massachusetts General Hospital, Boston, MA, USA. ²Department of Applied Mathematics, Technical Medical Center, University of Twente, Enschede, The Netherlands. ³Matthew Tivnan and Irene Désirée Kikkert contributed equally to this work. ✉email: i.d.kikkert@student.utwente.nl

challenges in reconstructing diagnostically accurate images¹⁰. Traditionally, CT reconstruction algorithms, such as filtered-back projection (FBP) and model-based iterative reconstruction (MBIR), have been employed.

However^{13,14}, FBP can introduce substantial noise and artifacts, especially with sparse-view data, while MBIR, although more effective in reducing artifacts, often compromises spatial resolution^{13–17}. Alternatively, deep learning based reconstruction (DLR) techniques have emerged, leveraging neural networks to enhance image quality by learning patterns from paired low- and high-quality images^{18–20}. Given the challenges of reconstructing accurate images from sparse-view CT data, DLR is considered a promising technique for sparse-view CT reconstruction. However, many clinicians have expressed concern about the use such artificial intelligence (AI) algorithms in radiology, citing possible issues with fidelity and diagnostic quality^{21–23}.

While studies highlight the capability of DLR to improve standard image quality metrics, there has been little research on the comparative diagnostic performance relative to FBP or MBIR. This may be due to the high cost of a proper observer study with trained radiologists. This study aims to bridge this gap by using an AI observer, trained for automatic diagnosis of intracranial hemorrhages, as a surrogate for evaluating diagnostic utility of different reconstruction techniques for sparse view CT. We use simulations of sparse-view CT sinogram data and apply FBP, MBIR, and DLR for reconstruction. The primary objective of this study is to compare the diagnostic performance of different reconstruction techniques in the context of sparse-view CT measurements. Our main innovation is the use of an AI observer model as a low-cost alternative to evaluate diagnostic impact of reconstruction techniques for this specific clinical task. Our work has resulted in a new tool to verify that proposed reconstruction algorithms truly translate to better outcomes for patients.

Methods

Hemorrhage diagnosis dataset

We used an open-source dataset from the RSNA Intracranial Hemorrhage Detection Challenge (2019) for this study¹². This dataset contains over 750,000 2D axial slices from head CT scans, each labeled by an expert into six categories: (1) no hemorrhage, (2) epidural hemorrhage, (3) intraparenchymal hemorrhage, (4) intraventricular hemorrhage, (5) subarachnoid hemorrhage, and (6) subdural hemorrhage. To simplify the analysis, cases with multiple hemorrhage types were excluded. The images were down-sampled to 256×256 with $1.0 \text{ mm} \times 1.0 \text{ mm}$ voxel spacing to reduce computational demands on time and memory. A comparative analysis demonstrating the negligible performance difference between 256×256 and 512×512 resolutions is provided in the supplementary information, Figure S1. We split the data into 719,729 images for training and validation and 1,000 images for a 3 clinical evaluation. The evaluation subset consists of 500 healthy patients and 100 patients from each hemorrhage type. Figure 1 block 1 shows examples from each hemorrhage type.

Sparse-view CT simulations

We used a stochastic model to simulate a sparse-view CT imaging system. The image of attenuation coefficients is represented by a flattened column vector $\mathbf{x} \in \mathbb{R}^{(N \times 1)}$, where N denotes the total number of voxels. The corresponding sinogram measurements are expressed as $\mathbf{y} \in \mathbb{R}^{(M \times 1)}$, where M represents the total number of projection rays across all views, and were treated as a stochastic function of $\mathbf{x}$. Specifically, we used the conditional distribution $p(\mathbf{y}|\mathbf{x}) \sim \mathcal{P}(I_0 \exp(-\mathbf{A}\mathbf{x}))$, where $\mathcal{P}(\lambda)$ represents a jointly independent Poisson random vector with expectation vector λ , $I_0 \in \mathbb{R}$ is the expected number of photons per pixel per view, and $\mathbf{A} \in \mathbb{R}^{(M \times N)}$ the system response matrix, otherwise known as a forward projector. For all experiments, we set $I_0 = 10^6$ photons per detector pixel per view, consistent with a low-dose acquisition protocol inspired by a prototype static CT system under development in our lab. The forward projector is defined by elements $A_{i,j}$, which represent the effective length traveled by the i -th ray and j -th voxel. The system was configured according to a proposed static CT design with 72 sources uniformly spaced in a circular array at a 500 mm source-to-object distance, using a 2D fan beam projection⁹. The detector setup consisted of 40 modules arranged in a regular polygon with an inscribed radius of 500 mm and 1 mm spacing. The LEAP CT projector followed by a re-binning step was utilized for generating projections²⁴. For this simulation, the system response matrix, defined as $N \times M$ where $N = 256 * 256$ and $M = 400 * 72$, was stored in memory, consuming approximately 6 GB.

Image reconstruction methods

Filtered back projection

Our FBP algorithm is based on the Moore-Penrose pseudoinverse. This approach adapts the traditional FBP algorithm with a shift-variant linear filter specific to the sparse-view geometry. The mathematical expression for our FBP reconstructor is

$$\hat{\mathbf{x}} = \mathbf{A}^+ \ell = \underbrace{(\mathbf{A}^T \mathbf{A})^+}_{\text{Filter}} \underbrace{\mathbf{A}^T}_{\text{Back Projection}} \ell,$$

where $\ell = -\log\left(\frac{\mathbf{y}}{I_0}\right)$ are the log-corrected estimates of line integrals, $\mathbf{A}^+$ denotes the Moore-Penrose pseudoinverse of $\mathbf{A}$, with the shift-variant image-domain filter defined as $(\mathbf{A}^T \mathbf{A})^+$, and ℓ represents the measured projection data. This formulation inverts the nonzero eigenvalues of $\mathbf{A}^T \mathbf{A}$ while zero eigenvalues are retained. To mitigate amplification of small but nonzero eigenvectors, we used only the highest 3000 eigenvalues and set the rest to zero to effectively balance detail recovery with artifact reduction.

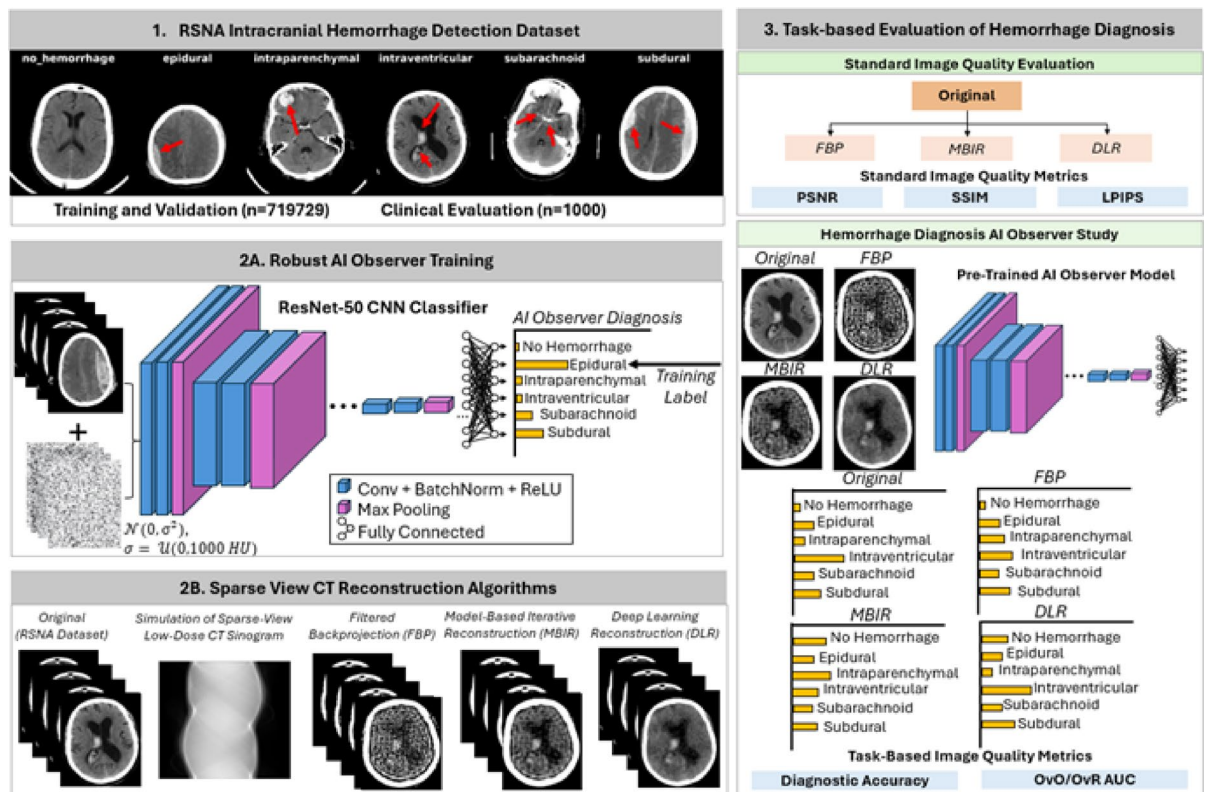


Fig. 1. Workflow of the presented study. The first block illustrates different intracranial hemorrhage types in the RSNA Intracranial Hemorrhage Detection Challenge¹². Each image utilized in the dataset is labeled with one of these hemorrhage types. While the examples shown in the figure are easily identifiable, many images in the dataset feature smaller, more subtle hemorrhages that are more challenging to detect. The second block illustrates the architecture of the modified ResNet-50 utilized to train a supervised classifier on predicting the hemorrhage type. On a clinical validation set of 1,000 images, FBP, MBIR, and DLR was applied, as presented in block 2B, yielding 4 different quality versions of the validation set. Each of these 4 datasets was evaluated with standard image quality metrics, and was presented to the trained AI observer for a hemorrhage diagnosis AI observer study.

Model-based iterative reconstruction

Our MBIR algorithm is based on a quadratic penalized likelihood estimator. This approach iteratively optimizes an objective function that balances data fidelity and prior information to reduce noise and artifacts. The first term enforces a fit between the modeled projections, $A\mathbf{x}$, and the observed data ℓ , using statistical weights, $\mathbf{W} = \mathcal{D}\{I_0 \exp(-\ell)\}$. The second term uses the discrete Laplacian operator, ∇^2 , to encourage smooth results. The gradient descent update equation is

$$\hat{\mathbf{x}}^{[n+1]} = \hat{\mathbf{x}}^{[n]} - \alpha (A^T \mathbf{W} (A \hat{\mathbf{x}}^{[n]} - \ell) + \beta \nabla^2 \hat{\mathbf{x}}^{[n]})$$

Where α is a step size, and β a regularization strength. To prevent non-physical results, we also included another penalty term on the L2 norm of negative attenuation values. To select an appropriate value for β , we performed a parameter sweep over a range of candidate values. For each setting, we reconstructed a representative subset of the dataset and evaluated both image quality and downstream classification performance. The final choice of β reflects a trade-off between image smoothness and task-relevant detail.

Deep learning reconstruction

A deep convolutional U-Net model was implemented to perform DL image reconstruction by processing pseudoinverse reconstructions and predicting high-quality reference images. The U-Net model architecture follows an encoder-decoder structure with skip connections, featuring four down-sampling layers with channel sizes (16,32,64,128) and two convolutional blocks per depth. Key architectural enhancements include channel-wise cross attention in the deepest layers and group normalization. The model training was conducted using a mean squared error loss function:

$$\ell_\theta = E \left[(\mathcal{D}_\theta (A^+ \ell) - \mathbf{x})^T (\mathcal{D}_\theta (A^+ \ell) - \mathbf{x}) \right]$$

Where $\mathbf{x}$ represents high quality ground truth images sampled from the training dataset, $\mathbf{A}^+\ell$ are pseudoinverse reconstructions, $\mathcal{D}_\theta$ is the deep learning reconstructor, parameterized by the weights θ , and $E[\bullet]$ represents the expectation or mean. Following existing methods, the DLR output was incorporated as a prior in an MBIR framework^{25,26}. The gradient descent update equation, as a function of regularization strength γ is:

$$\hat{\mathbf{x}}^{[n+1]} = \hat{\mathbf{x}}^{[n]} - \alpha (\mathbf{A}^T \mathbf{W} (\mathbf{A} \hat{\mathbf{x}}^{[n]} - \ell) + \gamma (\hat{\mathbf{x}}^{[n]} - \mathcal{D}_\theta (\mathbf{A}^+ \ell)))$$

In this way, the DLR technique balances the influence of the same data fidelity term used in MBIR with a prior based on the deep learning model.

AI model observer

Task-based image quality evaluation in clinical applications typically relies on costly and challenging human observer studies with trained radiologists. In this work, a convolutional neural network-based AI model observer was utilized as a surrogate. The ResNet-50 architecture was employed for this purpose, trained on the original images of the RSNA dataset²⁷. To enhance the robustness of the model, white noise was added to the input images, with standard deviations uniformly ranging from 0 to 1,000 HU, as illustrated in Fig. 1 block 2B. The model's output consists of a vector of six real numbers, representing the probability distribution across disease categories. A weighted sampler was used to rebalance the dataset, ensuring that the model was exposed to all six classes of images with equal probability during training. The training process consisted of 1,000 epochs, with 1,000 iterations per epoch and a batch size of 128. To examine the potential impact of domain shift, we conducted an additional fine-tuning experiment on a small sparse-view training set. Results confirm that the AI observer can adapt to the image characteristics of each reconstruction method (Supplementary information, Figure S2), supporting the robustness of our comparative analysis.

Image reconstruction quality evaluation

Reconstruction evaluations were conducted on the clinical validation set of 1,000 patients, which were not included in the training or validation phases. After applying the reconstruction algorithms, standard paired metrics were utilized to assess image quality, including peak signal-to-noise ratio (PSNR), structural similarity index metric (SSIM), and Learned Perceptual Image Patch Similarity 5 (LPIPS). These metrics were computed for all images in the evaluation dataset, comparing the simulated sparse-view CT reconstructions with the original high-quality images from the RSNA dataset. Evaluations were performed across two specific windows: a brain window ranging from 0 to 80 HU and a bone window spanning from −1,000 to 2,000 HU. For each reconstruction method, the mean and standard deviation of the PSNR, SSIM, and LPIPS were computed to evaluate image quality.

Task-based image quality evaluation

A task-based evaluation of all discussed reconstruction algorithms for intracranial hemorrhage detection was conducted using the AI model observer for automated hemorrhage diagnosis from FBP, MBIR, and DLR images for all 1000 images in the evaluation subset. We also repeated the same analysis for the original images as a baseline without sparse-view artifacts. After obtaining the classifier outputs for the full evaluation cohort of 1,000 patients, Receiver Operating Characteristic (ROC) curve analysis was performed to assess the trade-off between the false positive rate and true positive rate. Both One vs. One (OvO) and One vs. Rest (OvR) classification performances were evaluated, and the Area Under the Curve (AUC) was computed to summarize the classifier's performance. To gain further insight into the decision-making process of the AI observer, we additionally applied Grad-CAM saliency mapping to visualize which image regions contributed most to the model's predictions.

Results

For each β , the evaluation subset of images was reconstructed and evaluated both standard image quality metrics (PSNR, SSIM) and downstream classification performance (AUC per hemorrhage type). Results are shown in Fig. 2. We observed that increasing β leads to smoother reconstructions, which improves PSNR and SSIM up to a point, but can negatively impact diagnostic performance due to over-smoothing. A value of $\beta = 100$ was selected as it offers a favorable balance between image fidelity and task performance.

Representative examples of each sparse-view CT reconstruction technique and each hemorrhage type are shown in Fig. 3. In the FBP reconstructions, view sparsity manifests as starshaped streak artifacts which could significantly hinder the identification of anatomical structures and the differentiation of hemorrhage types, particularly in the brain window. MBIR exhibits similar artifacts, but they are less pronounced and appear smoother in both brain and bone windows. The DLR shows markedly fewer streaking artifacts compared to both FBP and MBIR, resulting in clearer visualization of brain anatomy and hemorrhage details.

Standard image quality evaluation

Results on the quantitative image quality metrics, including PSNR, SSIM, and LPIPS, are presented in Fig. 4. Across the clinical evaluation of 1,000 samples, DLR consistently outperforms both MBIR and FBP across all metrics, indicating greater structural similarity and perceptual quality. MBIR surpasses FBP in PSNR, SSIM, and LPIPS metrics. However, the PSNR in the bone window is an exception where FBP outperforms MBIR.

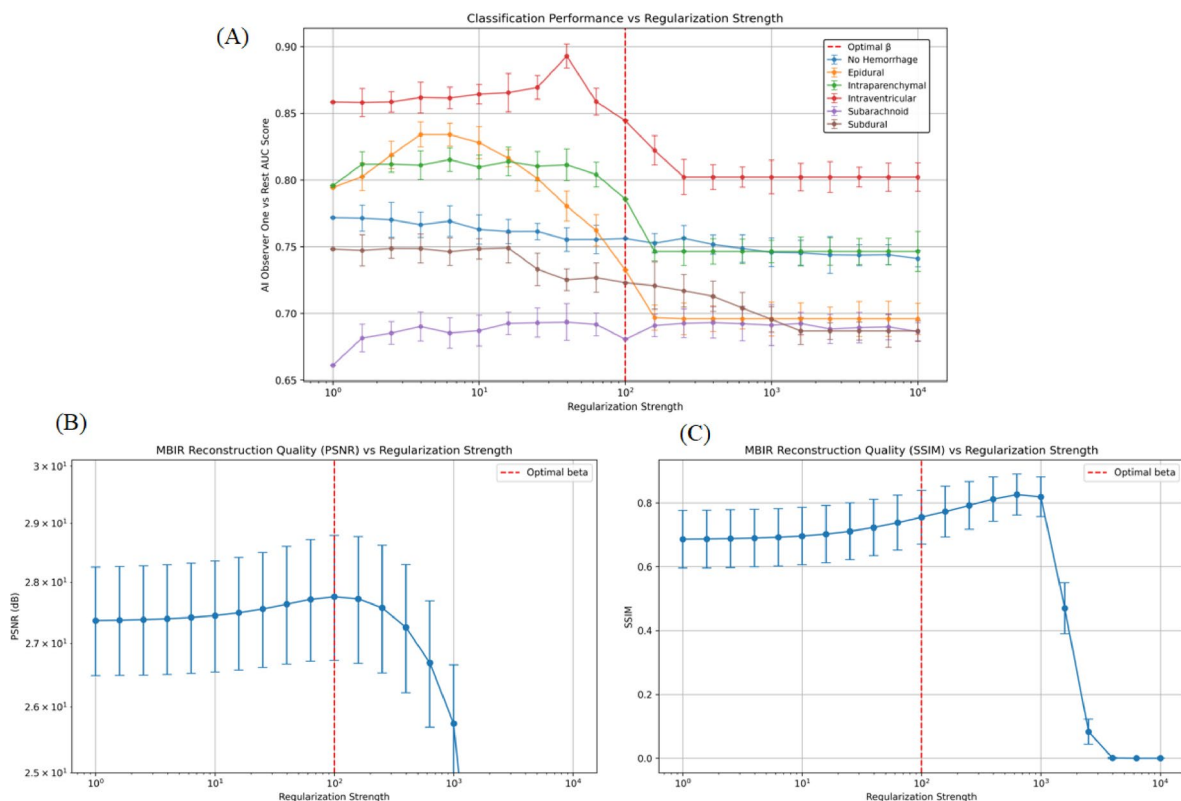


Fig. 2. Impact of MBIR regularization strength (β) on classification performance (A), PSNR (B), and SSIM (C). The vertical dashed red line indicates the selected $\beta = 100$, which represents a favorable trade-off between image quality and task performance.

Task-based image quality evaluation

The ROC AUC values for both One vs. Rest and One vs. One classification with the AI observer are summarized in Table 1. DLR achieved the highest diagnostic accuracy at 71.7% compared to FBP at 69.7%. Surprisingly, MBIR had worse task based performance than FBP, with 65.6% diagnostic accuracy. Applying the AI observer model to the original images, with no view sparsity artifacts, resulted in 73.2% accuracy, so these results also validate the robustness of our classifier to artifacts. For the One vs. Rest classification, DLR outperformed both FBP and MBIR across all hemorrhage types. FBP outperformed MBIR in all hemorrhage classifications except for subarachnoid hemorrhage, where MBIR showed superior performance. The largest performance gap was observed in the detection of epidural hemorrhage, where DLR achieved an AUC of 0.8787 compared to 0.7943 for FBP. The One vs. One classification performance mirrors the trends identified in the One vs. Rest analysis, with DLR outperforming FBP and MBIR in nearly all classifications. Exceptions include the differentiation between healthy versus intraparenchymal hemorrhage and intraventricular versus subarachnoid hemorrhage, where FBP demonstrated superior performance. FBP outperformed MBIR on most classifications, except for differentiating between healthy vs. subarachnoid and intraparenchymal vs. subarachnoid, where MBIR outperforms FBP. However, this improvement comes at the expense of reduced performance in other hemorrhage classifications. Overall, while all sparse-view reconstruction techniques fall short on the performance compared to the original images with no artifacts, the image quality of DLR results in the highest accuracy for most hemorrhage diagnosis and classification tasks.

Discussion

From the results of our study, we conclude that DLR image quality is superior to that of FBP or MBIR for this sparse view CT for intracranial hemorrhage diagnosis task. By visual inspections of the images in Fig. 3, we see that DLR achieves similar or better noise reduction compared to MBIR without over-smoothing edges and small features. The quantitative metrics (PSNR, SSIM, and LPIPS) demonstrated that DLR outperformed both FBP and MBIR, exhibiting superior noise reduction, structural similarity, and perceptual quality with reduced artifacts. Most importantly, DLR achieved the highest classification accuracy in detecting hemorrhages. These findings are consistent with the widespread conclusion in the field of tomographic reconstruction that DLR is the current state-of-the-art for sparse view CT^{28–30}.

This study demonstrates how an AI observer can be used to evaluate diagnostic utility of medical images. The gold standard for this type of evaluation would be a real human observer study where radiologists attempt to classify hemorrhage types based on sparse-view CT reconstructions. However, this type of study is costly and takes a great deal of time and money, especially for large sample sizes. Our proposed approach with an AI observer can

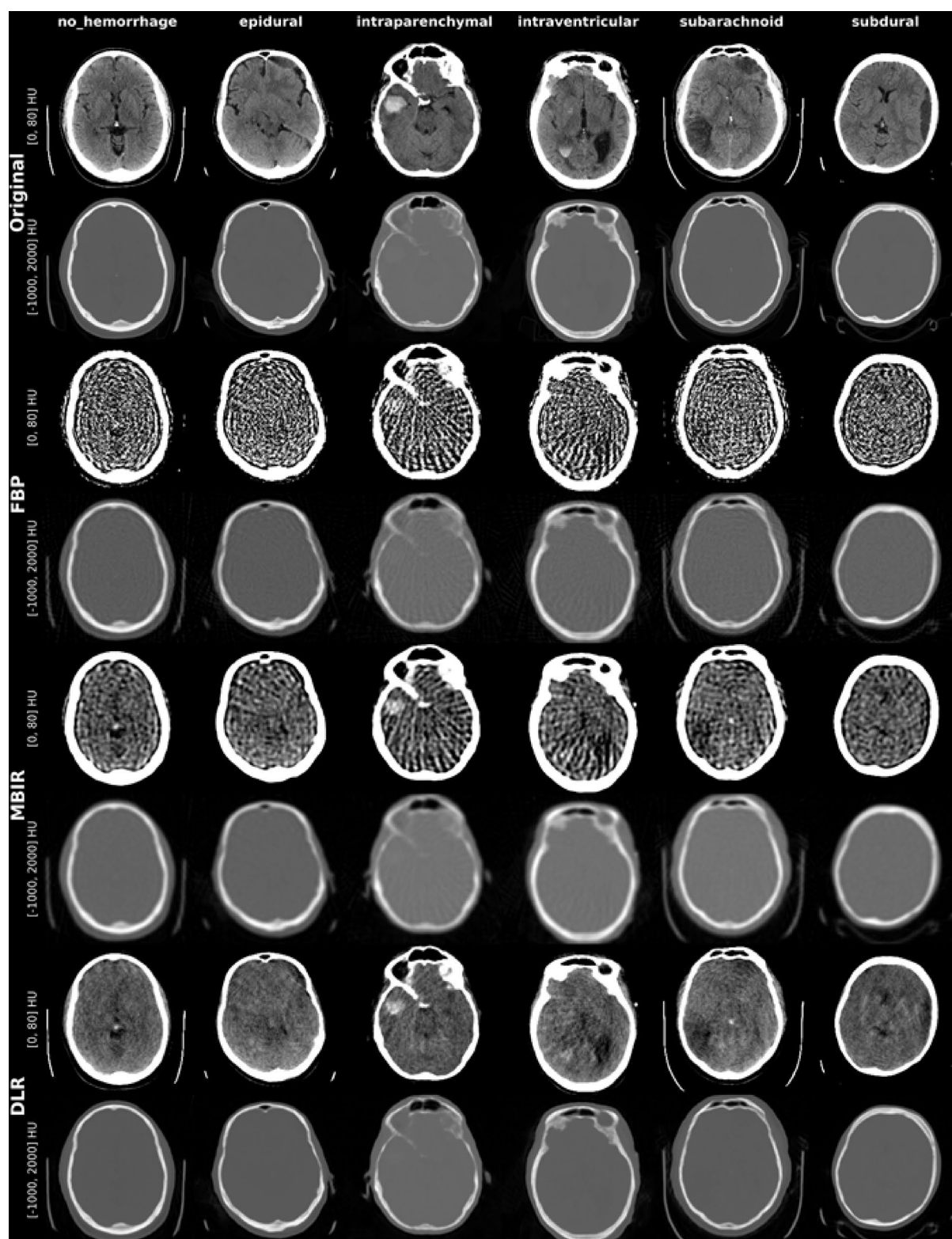


Fig. 3. Image results for FBP, MBIR, and DLR reconstruction. Each column represents a case example for each label, i.e. no hemorrhage, epidural hemorrhage, intraparenchymal hemorrhage, intraventricular hemorrhage, subarachnoid hemorrhage, and subdural hemorrhage.

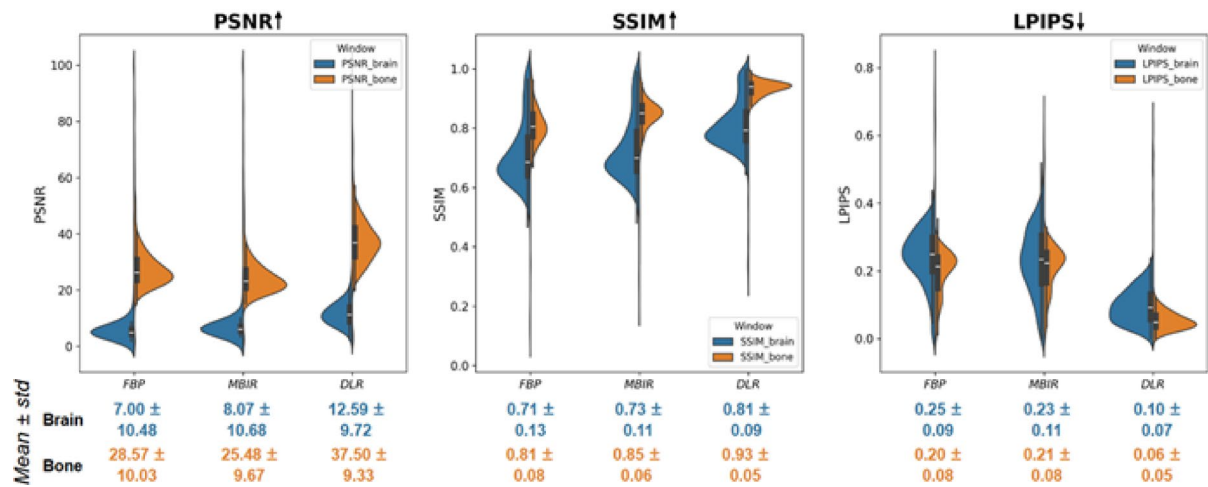


Fig. 4. Violin plots displaying the PSNR, SSIM, and LPIPS across the three reconstruction types for the brain window ([0, 80] HU) and bone window ([−1000, 2000] HU), accompanied with their mean and standard deviation (std) across all images in the evaluation dataset.

	Original	FBP	MBIR	DLR
Overall accuracy	73.2%	69.7%	65.6%	71.7%
OvR AUC				
No hemorrhage	0.8582	0.7717	0.7561	0.8083
Epidural	0.9385	0.7943	0.7325	0.8787
Intraparenchymal	0.8479	0.7957	0.7857	0.8018
Intraventricular	0.9317	0.8585	0.8444	0.8666
Subarachnoid	0.7756	0.6611	0.6806	0.6915
Subdural	0.8224	0.7482	0.7229	0.7533
OvO AUC				
No hemorrhage vs. epidural	0.9427	0.8292	0.7597	0.8914
No hemorrhage vs. intraparenchymal	0.8653	0.8356	0.8338	0.8277
No hemorrhage vs. intraventricular	0.9365	0.8633	0.8510	0.8741
No hemorrhage vs. subarachnoid	0.7416	0.6249	0.6601	0.6607
No hemorrhage vs. subdural	0.8191	0.7785	0.7520	0.7633
Epidural vs. intraparenchymal	0.9034	0.7673	0.7268	0.7876
Epidural vs. intraventricular	0.9777	0.9239	0.9006	0.9299
Epidural vs. subarachnoid	0.9025	0.7073	0.6837	0.7563
Epidural vs. subdural	0.8460	0.6603	0.6094	0.6788
Intraparenchymal vs. intraventricular	0.8900	0.8045	0.8028	0.8200
Intraparenchymal vs. subarachnoid	0.8210	0.7326	0.7452	0.7375
Intraparenchymal vs. subdural	0.8086	0.7506	0.7310	0.7435
Intraventricular vs. subarachnoid	0.8184	0.7512	0.7507	0.7461
Intraventricular vs. subdural	0.9007	0.7652	0.6738	0.8400
Subarachnoid vs. subdural	0.7512	0.6648	0.6420	0.7005

Table 1. Observer performance (AUC) across various hemorrhage types presenting the overall accuracy, One vs. Rest AUC, and One vs. One AUC.

be used to rapidly process thousands of cases at low cost. If an AI observer consistently performs poorly after a certain change to the reconstruction algorithm, that is an indication that the change should not be included, even if it improves classical image quality metrics. We believe this is a valuable research tool to evaluate new reconstruction techniques before proceeding to a full human observer study. Figure 5 presents a representative case study illustrating the influence of reconstruction quality on hemorrhage classification. In this example, the model observer correctly identifies the presence of an epidural hemorrhage in both the Original and DLR images. However, when interpreting the FBP and MBIR reconstructions, the model erroneously classifies the lesion as an intraparenchymal hemorrhage. This misclassification can be attributed to the presence of

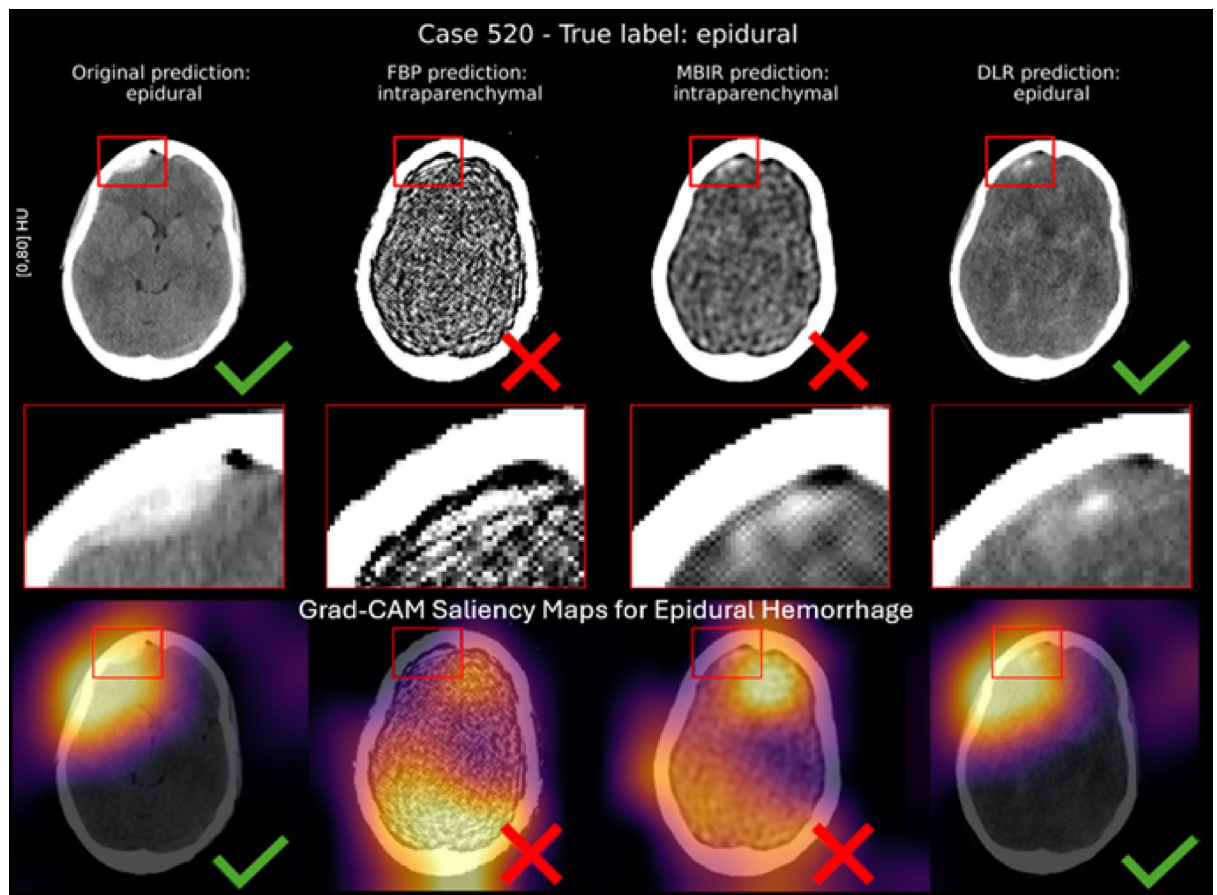


Fig. 5. Example of patient 520 with epidural hemorrhage. The model observer accurately identifies the epidural hemorrhage in the original CT data and DLR images. However, classification fails on FBP and MBIR images, likely due to view sparsity artifact.

pronounced artifacts in the FBP and MBIR images, which obscure key anatomical boundaries and give rise to the illusion of a parenchymal-dural separation. These observations are further supported by Grad-CAM saliency maps, which reveal that, in correctly classified cases (Original and DLR), the model's attention is concentrated on the actual hemorrhagic region³¹. Conversely, in misclassified reconstructions (FBP and MBIR), the model's attention appears spatially diffuse or displaced from the true site of pathology. These findings suggest that image degradation, particularly over-smoothing in MBIR, may compromise the model's ability to localize diagnostically relevant features, thereby impairing classification performance.

Our study highlights the impact of image quality beyond standard quantitative metrics like PSNR, SSIM. Even LPIPS, which uses a well-validated perceptual similarity metric predicts the wrong result in terms of diagnostic accuracy. FBP reconstructions exhibited significant streaking artifacts, compromising the interpretation of hemorrhage types, as indicated by lower SSIM values compared to MBIR and DLR.

The paradox arises in the model observer's classification performance, where FBP images yielded better results than MBIR (for all except subarachnoid hemorrhages). This suggests that the AI model can identify certain features in FBP images that aid in hemorrhage detection, despite the presence of artifacts. This result echoes common feedback from radiologists who complain MBIR is over-smoothed, with many preferring FBP image quality. It also agrees with well an existing mathematical proof that quadratic penalized likelihood reconstruction cannot improve the detectability index for an ideal Bayesian observer³².

MBIR produces smoother images with fewer artifacts, achieving higher SSIM values, particularly in bone window imaging. However, this smoothing can obscure essential diagnostic features, hindering the AI model's classification performance. Interestingly, this over-smoothing may also explain why MBIR performs better in detecting subarachnoid hemorrhages compared to FBP, as displayed in Table 1. Subarachnoid hemorrhages, which often appear as diffuse blood within the sulci and cisterns, may benefit from the reduced noise and clearer delineation of cerebrospinal fluid spaces in smoother reconstructions. However, for other hemorrhage types, the excessive smoothing can obscure critical features, leading to misclassification, also seen in Fig. 5. Additionally, case studies reveal that MBIR often predicts no hemorrhage when other methods detect one, indicating excessive smoothing that compromises hemorrhage detection. These findings underscore the need to balance noise reduction and diagnostic feature preservation in MBIR. We observe that increasing the regularization strength initially improves reconstruction quality (PSNR), but beyond a certain point leads to over-smoothing

and reduced classification performance for most hemorrhage types. This trade-off is particularly evident for intraventricular and intraparenchymal hemorrhages, where AUC scores drop sharply beyond the selected regularization strength (Fig. 2), likely because excessive smoothing obscures subtle yet diagnostically important features. It is possible that replacing the quadratic penalty in MBIR with an edge-preserving penalty term such as a total-variation or Huber penalty would improve performance³³.

DLR reconstructions demonstrated a promising balance of artifact reduction and anatomical detail, especially in brain imaging, achieving high noise suppression and structural accuracy. This made DLR particularly effective for intracranial hemorrhage detection, though some limitations were observed in distinguishing specific hemorrhage types, as indicated by lower AUC values (Table 1). This discrepancy might stem from the AI model's limited exposure to smoothed images, as the training data included random noise for robustness against FBP artifacts but not simulated smoothing for MBIR and DLR appearances. Incorporating smoothness variations during training could help the model better recognize features across different reconstruction methods. While DLR's performance metrics suggest it could enhance clinical diagnostics, its dependency on deep learning introduces potential biases, especially if clinical features diverge from those in the training data. To support adoption in practice, rigorous validation with a diverse and representative dataset, along with radiologist assessment, will be essential to confirm DLR's clinical effectiveness and ensure its utility across varied imaging scenarios.

Conclusion

This work presents an AI observer that can be used as a surrogate for human radiologists to evaluate the diagnostic quality of CT reconstruction techniques. Our study reveals that standard quality metrics like PSNR, SSIM, and LPIPS do not always correlate with a more accurate diagnosis. While many previous studies have shown DLR achieves high performance on those metrics, our results go further, confirming that DLR also translates to improvements in diagnostic accuracy for this hemorrhage detection and diagnosis task. Our results also highlight a risk that MBIR may result in worse clinical outcomes than FBP despite improvements in classical image quality metrics. We believe our AI observer studies have potential as a low-cost alternative to human observer studies for evaluating medical image reconstruction techniques to improve the quality of patient care.

Data availability

The dataset analyzed in this study is publicly available, as described in the Methods section. For any queries or data requests, the corresponding author may be contacted.

Received: 20 February 2025; Accepted: 8 July 2025

Published online: 17 July 2025

References

- Al-Naser, Y. A., Tafti, D. C. T. & Instrumentation and Physics. Nov 2. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing. PMID: 37983332. (2023).
- Raymond, A., Schulz, J. A., Stein & Pelc, N. J. How CT happened: the early development of medical computed tomography. In: *J. Med. Imaging* 8.05 (Oct. 2021). ISSN: 2329–4302. <https://doi.org/10.1117/1.JMI.8.5.052110>
- Parizel, P. et al. Intracranial hemorrhage: principles of CT and MRI interpretation. In: *European Radiology* 11.9Sept. pp. 1770–1783. ISSN: 0938–7994. (2001). <https://doi.org/10.1007/s003300000800>
- Chelsea, S. & Kidwell Comparison of MRI and CT for Detection of Acute Intracerebral Hemorrhage. In: *JAMA* 292.15Oct. p. 1823. ISSN: 0098–7484. (2004). <https://doi.org/10.1001/jama.292.15.1823>
- Lenhart, M., Bretschneider, T., Gmeinwieser, J., Ullrich, O. W. & Schlaier, J. S. Feuerbach. Cerebral CT angiography in the diagnosis of acute subarachnoid hemorrhage. In: *Acta Radiologica* 38.5 (Sept. 1997), pp. 791–796. ISSN: 0284–1851. <https://doi.org/10.1080/02841859709172412>
- Balint Botz and Aditya Shetty. CT scanner (evolution). In: Radiopaedia.org. Radiopaedia.org, June 2014. <https://doi.org/10.5334/rID-29733>
- Lipson, R. S. B. J. et al. Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer. In: *Archives of internal medicine* 169.22 (Dec. 2009), pp. 2078–86. ISSN: 1538–3679. 10.1001/archinternmed.2009.427
- David, J., Brenner & Hall, E. J. Computed Tomography — An increasing source of radiation exposure. *New. Engl. J. Med.* 357 22 (Nov. ISSN, 2277–2284. 10.1056/NEJMra072149 (2007).
- Avilash Cramer, J. et al. Sept. Stationary Computed Tomography for Space and other Resource-constrained Environments. In: *Scientific Reports* 8.1 p. 14195. ISSN: 2045–2322. (2018). <https://doi.org/10.1038/s41598-018-32505-z>
- Junguo Bian, J. H. et al. Evaluation of sparse-view reconstruction from flat-panel-detector cone-beam CT. In: *Physics in medicine and biology* 55.22 (Nov. 2010), pp. 6575–99. ISSN: 1361–6560. DOI: 10.1088/0031-9155/55/22/001
- Li, Y., Fu, X., Zhao, S., Jin, R. & Kevin Zhou, S. Sparse-view CT Reconstruction with 3D Gaussian Volumetric Representation. In: Dec. (2023). <https://doi.org/10.48550/arXiv.2312.15676>
- Adam, E. et al. Construction of a Machine Learning Dataset through Collaboration: The RSNA 2019 Brain CT Hemorrhage Challenge. In: *Radiology: Artificial Intelligence* 2.3 (May 2020), e190211. ISSN: 2638–6100. <https://doi.org/10.1148/ryai.2020190211>
- Feldkamp, L. A., Davis, L. C. & Kress, J. W. Practical cone-beam algorithm. In: *Journal of the Optical Society of America A* 1.6 (June 1984), p. 612. ISSN: 1084–7529. <https://doi.org/10.1364/JOSAA.1.000612>
- Emil, Y. S. & Pan, X. Image reconstruction in circular cone-beam computed tomography by 9 constrained, total-variation minimization. In: *Physics in medicine and biology* 53.17 (Sept. 2008), pp. 4777–807. ISSN: 0031–9155. <https://doi.org/10.1088/0031-9155/53/17/021>
- Taubmann, O. et al. Computed tomography. In: (eds Maier, A., Steidl, S., Christlein, V. & Horneegger, J.) *Medical Imaging Systems. Lecture Notes in Computer Science()*, vol 11111. Springer, 147–189. DOI: https://doi.org/10.1007/978-3-319-96520-1_8. (2018). 8(\}
- Shanzhou Niu, Y. et al. Sparse-view x-ray CT reconstruction via total generalized variation regularization. In: *Physics in medicine and biology* 59.12 (June 2014), pp. 2997–3017. ISSN: 1361–6560. <https://doi.org/10.1088/0031-9155/59/12/2997>

17. Perry, J. et al. Abdominal CT With ModelBased Iterative Reconstruction (MBIR): Initial Results of a Prospective Trial Comparing Ultralow-Dose With Standard-Dose Imaging. In: American Journal of Roentgenology 199.6 (Dec. 2012), pp. 1266–1274. ISSN: 0361-803X. <https://doi.org/10.2214/AJR.12.9382>
18. Kerstin Hammernik and Florian Knoll. Machine learning for image reconstruction. In: Handbook of Medical Image Computing and Computer Assisted Intervention. Elsevier, 25–64. DOI: <https://doi.org/10.1016/B978-0-12-816176-0.00007-7>. (2020).
19. Lennart, R. et al. Mar. Deep Learning Image Reconstruction for CT: Technical Principles and Clinical Prospects. In: Radiology 306.3 ISSN: 0033-8419. (2023). <https://doi.org/10.1148/radiol.221257>
20. Anirudh, R. et al. Lose the Views: Limited Angle CT Reconstruction via Implicit Sinogram Completion, *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 6343–6352, Salt Lake City, UT, USA, 2018, pp. 6343–6352, (2018). <https://doi.org/10.1109/CVPR.2018.00664>. J.
21. Raymond Geis, A. P. et al. Ethics of Artificial Intelligence in Radiology: Summary of the Joint European and North American Multisociety Statement. In: Radiology 293.2 (Nov. 2019), pp. 436–440. ISSN: 0033-8419. <https://doi.org/10.1148/radiol.2019191586>
22. Adrian, P. B. & Neri, E. Artificial Intelligence in Radiology-Ethical Considerations. In: Diagnostics (Basel, Switzerland) 10.4 (Apr. 2020). ISSN: 2075–4418. <https://doi.org/10.3390/diagnostics10040231>
23. Erik, R. et al. Advantages, challenges, and risks of artificial intelligence for radiologists. In: Artificial Intelligence in Medical Imaging. Cham: Springer International Publishing, 329–346. DOI: <https://doi.org/10.1007/978-3-319-94878-2>. (2019).
24. Hyojin Kim and Kyle Champley. Differentiable Forward Projector for X-ray Computed Tomography. In: (July 2023). <https://doi.org/10.48550/arXiv.2307.05801>
25. Aggarwal, H. K., Mani, M. P. & Jacob, M. Dec. MoDL: model based deep learning architecture for inverse problems. In: DOI: 10. (2017). 1109/TMI.2018.2865356.
26. Kofler, A. et al. Neural networks-based regularization for large-scale medical image reconstruction. In: *Phys. Med. Biol.* 65.13 (July 2020), p. 135003. ISSN: 0031–9155. <https://doi.org/10.1088/1361-6560/ab990e>
27. He, K., Zhang, X. & Ren, S. Jian Sun. Deep Residual Learning for Image Recognition. In: Dec. (2015). <https://doi.org/10.48550/arXiv.1512.03385>
28. Di Jianglei, J., Lin, L., Zhong, K., Qian & Qin, Y. Review of Sparse-View or limitedangle CT reconstruction based on deep learning. *Laser Optoelectron. Progress.* 60, 0811002. 10.3788/ LOP230488 (2023).
29. Emil, Y., Sidky & Pan, X. Report on the AAPM deep-learning sparse-view CT grand challenge. In: Medical Physics 49.8 (Aug. 2022), pp. 4935–4943. ISSN: 0094-2405. <https://doi.org/10.1002/mp.15489>
30. Ge Wang, J. C., Ye & De Man, B. Deep learning for tomographic image reconstruction. In: Nature Machine Intelligence 2.12 (Dec. 2020), pp. 737– 748. ISSN: 2522–5839. DOI: 10. 1038 / s42256-020-00273-z.
31. Ramprasaath, R. S. et al. Grad-CAM: why did you say that? (Nov <https://doi.org/10.48550/arXiv.1611.07450> (2016).
32. Harrison, H., Barrett, Kyle, J. & Myers *Foundations of Image Science* (John Wiley & Sons, 2013).
33. Jeffrey Fessler. Statistical Image Reconstruction Methods for Transmission Tomography. In: Handbook of Medical Imaging, Volume 2. Medical Image Processing and Analysis. 1000 20th Street, Bellingham, WA 98227- 0010 USA: SPIE, pp. 1–70. <https://doi.org/10.1117/3.831079.ch1>

Author contributions

I.D.K. was responsible for drafting the introduction, results, and discussion, while M.T. was responsible for drafting the method. All other authors contributed to reviewing the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-11089-5>.

Correspondence and requests for materials should be addressed to I.D.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025