Gerchberg-Saxton Spectral Preprocessing as Empirical Leakage Mitigation in Simulated Federated Medical Imaging

Seha Ay[1⋆], Wei Zhang[1,2], and Umit Topaloglu[1,3]

[1] Department of Biomedical Engineering, Wake Forest School of Medicine, Winston-Salem, NC 27101, United States (seha.ay@wfusm.edu)
[2] Department of Cancer Biology, Wake Forest School of Medicine, Winston-Salem, NC 27101, United States (wei.zhang@advocatehealth.org)
[3] Center for Biomedical Informatics & Information Technology, National Cancer Institute, Rockville, MD 20850, United States (umit.topaloglu@nih.gov)

Abstract. Federated learning (FL) keeps training data local, but it does not by itself prevent attacks against the shared model. Reconstruction and inference attacks can still recover information from model parameters and learned representations, which is a serious concern for medical imaging. We study Gerchberg-Saxton (GS) spectral preprocessing, a stochastic phase-randomization and frequency-masking transform applied on each client before training. We frame GS strictly as an empirical leakage-mitigation step and make no formal privacy claims. Using a simulated three-client FL setup on brain MRI under IID and non-IID partitions, we evaluate GS with three attacks: a blind feature-inversion attack with a matched-versus-wrong separability protocol; a query-only black-box semantic-leakage attack; and a membership-inference check. DP-SGD is included only as a formal-privacy reference point. Baseline models leak measurable per-sample structure and recoverable class signal. GS reduces both toward chance while preserving classification utility at low and moderate masking. The most protective masking level is attack-dependent, and the response is non-monotonic, so we report effects without claiming a dose-response relationship. We also implement GS as a client-side package in a standard FL framework to show deployment feasibility. GS does not provide an (ε, δ) guarantee and is presented as an empirical, low-cost preprocessing step rather than a formal privacy mechanism.

Keywords: Federated learning · Privacy · Medical imaging · Empirical leakage mitigation · Feature inversion.

1 Introduction

Deep learning is now central to medical image analysis [9,18], yet clinical images are sensitive and subject to strict regulation [22]. Federated learning (FL)

⋆ Corresponding author.

addresses this by keeping data at each participating client's site and sharing only model updates [11]. Decentralization alone, however, is not privacy. Shared parameters, gradients, and learned representations can leak training information through model inversion [5], membership inference [8,19], and collaborative-learning attacks [7,13,16]. Formal defenses such as differential privacy [1,4] and secure aggregation [3] carry strong guarantees, but their utility and systems costs are particularly acute in medical imaging [12,20].

We investigate Gerchberg-Saxton (GS) spectral preprocessing as a lightweight, client-side, model-agnostic transform applied to images before local training. GS combines Fourier-domain phase randomization with optional iterative frequency masking. We frame GS as an empirical leakage-mitigation strategy and make no formal privacy claims. Our aim is to characterize empirically how GS affects attack success.

Our contributions are fourfold. First, we show that GS integrates into a standard FedAvg pipeline as a drop-in client preprocessing step that preserves classification utility under both IID and non-IID partitions. Second, we evaluate leakage under three attacks: a blind feature-inversion attack with a matched-versus-wrong separability protocol; a query-only black-box semantic-leakage attack; and a membership-inference check, with DP-SGD as a reference. Third, we find that GS reduces leakage relative to the baseline across attacks. The most protective masking level is attack-dependent and lies in the low-to-moderate range, and the response is non-monotonic rather than a clean dose response. Fourth, we report deployment feasibility through an end-to-end implementation of GS as a client-side preprocessing component for NVFlare [15], a widely used FL framework.

2 Background, Setup, and Threat Models

2.1 Gerchberg-Saxton Spectral Transformation

The full algorithmic specification and structural analysis of the GS transform used here are given in a companion study that characterizes GS in a re-identification setting [2]. We summarize only what is needed for this work. GS is a stochastic, image-level preprocessing operation applied independently to each image, without using labels, anatomical priors, or model information. Starting from a random phase field, it iterates a simple loop: modulate the image by the current phase, move to the Fourier domain, normalize all magnitudes to a unit spectrum, randomly mask a fraction p of frequency coefficients, return to the image domain, and update the phase from the result. After I iterations the released image is the magnitude of the final complex field. We write GS $p\%$ for the variant with masking probability p; GS 0% applies phase randomization and magnitude normalization without frequency masking.

Three properties make this construction relevant to leakage. The magnitude normalization step discards the spectral magnitude of each intermediate field, so that profile is not carried forward [6]. The transform is stochastic for a fixed

input, because both the initial phase and the per-iteration masks are random; repeated application yields different outputs. The final step keeps only magnitude and discards phase, so many distinct complex fields map to the same released image. These are structural observations, not formal guarantees. They do not establish that an input is unrecoverable, which is why the empirical attacks below probe what is actually recoverable.

2.2 Federated Setup

We simulate FedAvg [11] with three clients on a public brain MRI tumor classification dataset [14] with four classes: glioma, meningioma, no tumor, and pituitary. Raw images are standardized and normalized to $[0, 1]$, and GS is then applied on each client. We deliberately do not renormalize after GS. Re-scaling the GS output to a fixed range would partially undo the spectral perturbation that the transform introduces, so each client trains on the native post-GS intensities. As a result, GS outputs occupy a wider intensity range than the raw inputs, and the model learns on that range. We use an IID partition and a non-IID partition in which class proportions differ across clients. Federated training uses 10 communication rounds with 3 local epochs per round. The DP-SGD reference is trained and evaluated on the same non-IID client splits and the same held-out test set, so its results are directly comparable to the non-IID GS and baseline results.

2.3 Threat Models

We consider an honest-but-curious adversary with white-box access to the trained model, but without access to the training data. The adversary follows standard practice and assumes inputs in the conventional $[0, 1]$ range, the same assumption used for every condition so that comparisons are controlled.

Feature inversion. The adversary optimizes a synthetic input so that its intermediate feature representation, taken from the convolutional stack, matches that of a target [5,10]. We use representation matching rather than gradient matching for a concrete reason. Trained classifiers in this regime are overconfident, and correctly classified samples yield near-zero loss gradients; this makes gradient-based inversion ineffective. The objective combines a feature-matching term with total-variation regularization. Per-sample leakage is judged by a matched-versus-wrong protocol. For each reconstruction we compute structural similarity (SSIM) [23] and a structure score (SSIM times an edge-similarity term [21]), measured against the true target and against random other targets. We then report the advantage, defined as matched minus wrong, with bootstrap 95% confidence intervals. An interval that excludes zero indicates leakage of individual structure rather than only class-level structure.

Black-box semantic leakage. A query-only adversary reconstructs inputs from the served classifier without reference images. The reconstructions are then classified by a separate evaluation model that shares the baseline architecture and is trained on the raw, untransformed data splits. Semantic leakage

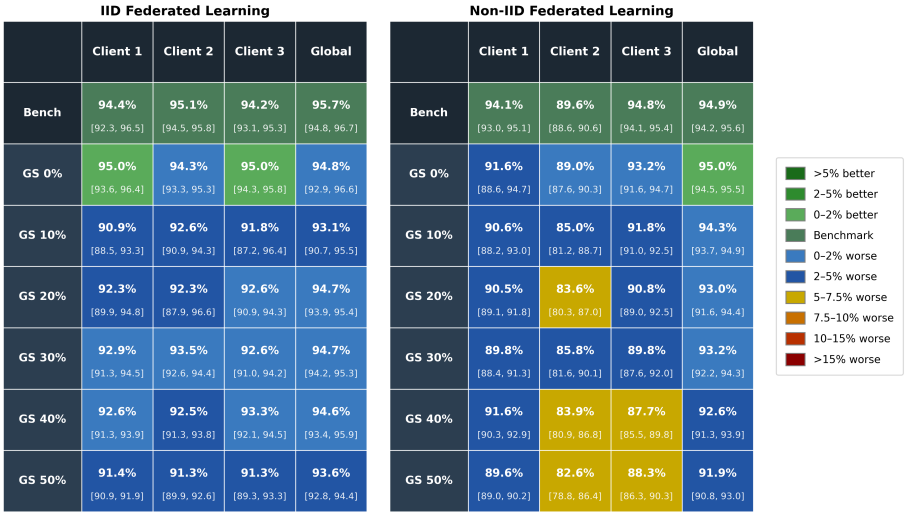

Fig. 1. Per-client and global classification accuracy (mean with 95% confidence intervals) under IID (left) and non-IID (right) partitions across GS masking levels. The global model stays close to baseline at low and moderate masking. Client-level sensitivity is larger under non-IID, particularly for the minority-heavy Client 2.

is measured as macro-F1 and balanced accuracy on the reconstructions, with a four-class random floor at balanced accuracy 0.25.

Membership inference. We include a standard confidence-based membership-inference attack [17,19] as a standard check, reporting attack F1 averaged over training epochs to characterize stability [8].

DP-SGD reference. DP-SGD [1] serves as a formal-privacy reference across the utility, membership-inference, and feature-inversion analyses. It is not presented as a method that GS replaces. For the differentially private model we use a DP-compatible backbone with group normalization in place of batch normalization, which differs from the GS and baseline architecture. Because feature inversion is architecture-sensitive, we treat the DP-SGD feature-inversion result as an observation about a deployed DP model and discuss this caveat in Section 3.2.

3 Results

3.1 Utility under GS in Federated Learning

Figure 1 reports per-client and global classification accuracy with 95% confidence intervals under IID and non-IID partitions. Under IID, the global model stays close to baseline across masking levels, moving from 95.7% at baseline to 94.8% at GS 0% and 93.6% at GS 50%. Under non-IID, it is similarly stable, from 94.9% at baseline to 95.0% at GS 0% and 91.9% at GS 50%. Client-level accuracy varies more under non-IID. The minority-heavy client (Client 2) is the most sensitive:

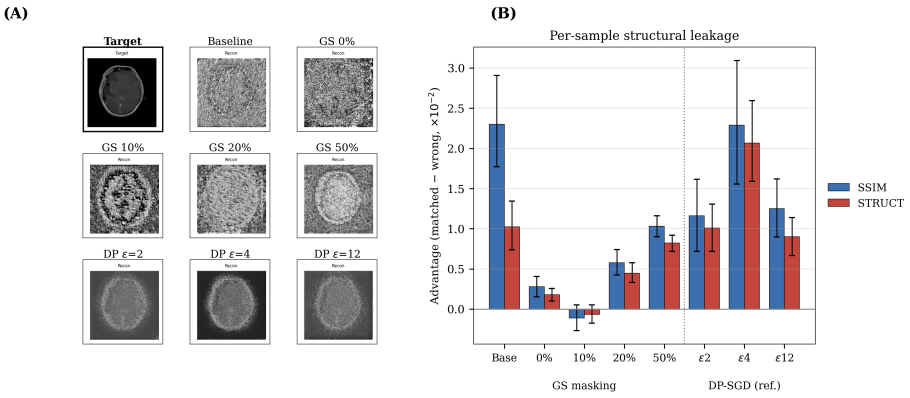

Fig. 2. Feature inversion on the global model. (A) Reconstructions of a shared target under baseline, GS masking levels, and DP-SGD references. (B) Matched-minus-wrong advantage for SSIM and structure score in units of 10^{-2} (mean with 95% confidence intervals); the dotted divider separates the GS masking sweep from the DP-SGD references. The baseline leaks per-sample structure, GS reduces it at low masking, and the masking response is non-monotonic.

it begins at 89.6% at baseline and declines to 83.6% at GS 20% and 82.6% at GS 50%. The global model absorbs this variation and remains above 91% throughout. Utility is therefore preserved at low and moderate masking.

3.2 Feature Inversion

Figure 2 summarizes the feature-inversion analysis. Panel A shows global-model reconstructions of a shared target across conditions. Panel B shows the matched-minus-wrong advantage for SSIM and the structure score, expressed in units of 10^{-2}. Baseline models leak statistically significant per-sample structure, with an SSIM advantage of about 2.3×10^{-2} and a structure advantage of about 1.0×10^{-2}, both with intervals excluding zero. This holds for the global model and for all three client models. GS reduces the advantage by roughly an order of magnitude at low masking: GS 0% and GS 10% bring the SSIM and structure advantages close to zero, with the qualitatively noisiest reconstructions in the figure. The response is non-monotonic, since the advantage partially returns at GS 20% and GS 50% as coarse edge similarity rises. We therefore do not claim a dose-response relationship with masking. Instead, GS reduces per-sample leakage relative to the baseline, with the most protective level lying in the low-to-moderate range and depending on the attack.

Under the same blind attack, the DP-SGD reference model produces more coherent coarse anatomical reconstructions, and its per-sample advantage is comparable to or above the baseline. We report this as an observation about a deployed DP model rather than a property of the DP mechanism, and we note the architecture difference. The observation is consistent with the design scope

Table 1. Black-box semantic leakage on the global model. Reconstruction-side macro-F1 and balanced accuracy (lower is more private; four-class random floor at balanced accuracy 0.25), with the clean-image reference. The baseline leaks; GS reduces recoverable signal toward the floor with a non-monotonic response.

	Reconstruction side		Clean reference	
Condition	Macro-F1	Balanced acc.	Macro-F1	Balanced acc.
Baseline	0.645	0.644	0.838	0.961
GS 0%	0.132	0.265	0.861	0.977
GS 10%	0.394	0.470	0.827	0.959
GS 20%	0.109	0.250	0.823	0.955
GS 50%	0.174	0.301	0.883	0.955

of DP-SGD, which bounds the influence of individual training examples and so targets membership-style leakage rather than feature reconstruction. Across all conditions, reconstructions recover at most coarse anatomical structure, such as the skull and brain envelope, rather than fine identifying detail.

3.3 Black-box Semantic Leakage

Table 1 reports macro-F1 and balanced accuracy of an independent classifier on reconstructions from the query-only attack, alongside the clean-image reference. The reference classifier is competent, with clean macro-F1 about 0.83 to 0.88 across conditions. Baseline reconstructions carry strong recoverable class signal (macro-F1 0.645, balanced accuracy 0.644). GS reduces this toward the four-class random floor: GS 0% collapses to macro-F1 0.132 and balanced accuracy 0.265, and GS 20% to 0.109 and 0.250. As in feature inversion, the response is non-monotonic, with GS 10% retaining more signal than the surrounding levels. The two reconstruction-based attacks therefore agree that GS substantially reduces recoverable content relative to the baseline and that intermediate masking does not help monotonically.

3.4 Membership Inference and DP-SGD Reference

Table 2 reports the membership-inference check and the DP-SGD reference for the global model, with non-IID global accuracy and its 95% confidence interval. We report attack F1 and the membership advantage, defined as the gap between the true-positive and false-positive rates. Baseline membership-inference F1 is modestly above chance at 0.568, with an advantage of 0.051. GS moves the attack toward chance at low and moderate masking, reaching F1 0.485 at GS 0% and 0.503 at GS 10%. It then rebounds partially to 0.589 at GS 50%, matching the non-monotonic behavior of the reconstruction attacks. DP-SGD holds membership-inference F1 near chance across ε, but at a substantial utility cost:

Table 2. Membership inference (global model, mean attack F1 over epochs and membership advantage; F1 closer to 0.5 and lower advantage are more private) with non-IID global accuracy and 95% confidence interval. DP-SGD shares the same non-IID splits and test set.

Condition	MIA F1	MIA advantage	Global accuracy [95% CI]
Baseline	0.568	0.051	94.9 [94.2, 95.6]
GS 0%	0.485	0.044	95.0 [94.5, 95.5]
GS 10%	0.503	0.040	94.3 [93.7, 94.9]
GS 20%	0.527	0.047	93.0 [91.6, 94.4]
GS 50%	0.589	0.052	91.9 [90.8, 93.0]
DP-SGD ε=2	0.518	0.029	70.7 [68.3, 73.1]
DP-SGD ε=4	0.438	0.023	71.4 [68.9, 73.6]
DP-SGD ε=8	0.494	0.030	71.7 [69.5, 74.0]
DP-SGD ε=12	0.311	0.027	73.3 [71.0, 75.7]

global accuracy is 71–73%, compared with about 94% for the baseline and low-masking GS. These reference numbers situate GS against a formal mechanism without implying that the two are combined.

3.5 Deployment Feasibility

To assess practical integration, we implemented GS as a client-side preprocessing component for NVFlare [15], one of the standard FL frameworks, and validated it end-to-end on two datasets, including the brain MRI set used here. The algorithm, its logic, and its implementation were kept identical to the legacy transform. The packaged version only improves time and computational efficiency through parallelized computation. All quantitative results in this paper were produced with the legacy GS implementation, and the packaged outputs were verified to match it. These experiments establish deployment feasibility only and are not the source of the results above. The implementation is publicly available at `https://github.com/seha-ay/GSTransformFL`.

4 Discussion and Limitations

Two reconstruction-based attacks, feature inversion and black-box semantic leakage, agree on the central finding. Baseline FL models leak recoverable per-sample structure and class signal. GS reduces both toward chance while preserving utility at low and moderate masking. The agreement across attacks of different adversary classes strengthens this conclusion. The membership-inference check points the same way at low masking, and the global model stays stable under non-IID partitioning, which matters for heterogeneous federations.

The masking response deserves emphasis. Leakage does not decrease monotonically with masking. At low masking the reconstructions are closest to noise,

while heavier masking reintroduces coarse shared structure such as the skull and brain envelope. This coarse structure raises edge-based similarity without recovering fine detail, which is why the per-sample advantage partially returns at higher masking. A practical reading is that the protective effect of GS is concentrated at low-to-moderate masking, where utility is also best. In this setting a single moderate masking level can serve both goals.

Several limitations bound these claims. GS provides no (ε, δ) guarantee and is not a formal privacy mechanism. The adversary we study is non-adaptive and assumes the conventional $[0, 1]$ input range. A GS-aware adversary that also estimates the transformed intensity range is a stronger threat. Such an adversary would need additional knowledge of the transform and its parameters, which raises the practical bar. We view it as less likely, though not impossible, for a colluding participant in FL. The recommendation toward low-to-moderate masking holds for the threat models and dataset studied here and should not be generalized without further evaluation. Reconstructions recover coarse structure rather than fine detail, so our results characterize structural leakage; the re-identification setting is studied separately in the companion work [2]. Some attacks are evaluated as single runs, and our study uses one modality and a single three-client setup. Broader datasets, larger federations, adaptive adversaries, and the interaction of GS with fairness-aware and privacy-preserving collaborative learning [24] are important future work.

5 Conclusion

We presented an empirical evaluation of Gerchberg-Saxton spectral preprocessing as a client-side, model-agnostic leakage-mitigation step in simulated federated medical imaging. Under a blind feature-inversion attack and a query-only black-box attack, GS reduced per-sample structural leakage and recoverable class signal toward chance. Utility was preserved at low and moderate masking, and the masking response was non-monotonic rather than a clean dose response. Framed honestly as an empirical preprocessing step rather than a formal privacy mechanism, GS is a low-cost and deployable option for privacy-conscious federated medical imaging.

Acknowledgments. This work was partially supported by the Wake Forest Baptist Comprehensive Cancer Center Cancer Center Support Grant (P30CA012197) and NIH/NCI grant 3R01CA272627-01A1S1 (W.Z.); W.Z. is supported by the Hanes and Willis Family Professorship in Cancer. The authors thank the Bioinformatics Shared Resources of the Wake Forest Baptist Comprehensive Cancer Center for their input and suggestions, and Deha Ay for reviewing the experimental design and analysis pipelines and providing methodological feedback.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. pp. 308–318. Association for Computing Machinery (2016). https://doi.org/10.1145/2976749.2978318
2. Ay, S., Zhang, W., Topaloglu, U.: Empirical characterization of re-identification risk and diagnostic utility under Gerchberg–Saxton spectral transformation in medical imaging. Applied Sciences **16**(14), 6850 (2026). https://doi.org/10.3390/app16146850
3. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H.B., Patel, S., Ramage, D., Segal, A., Seth, K.: Practical secure aggregation for privacy-preserving machine learning. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. Association for Computing Machinery (2017). https://doi.org/10.1145/3133956.3133982
4. Dwork, C., Roth, A.: The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science **9**(3–4), 211–407 (2014). https://doi.org/10.1561/0400000042
5. Fredrikson, M., Jha, S., Ristenpart, T.: Model inversion attacks that exploit confidence information and basic countermeasures. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. pp. 1322–1333. Association for Computing Machinery (2015). https://doi.org/10.1145/2810103.2813677
6. Gerchberg, R.W.: A practical algorithm for the determination of phase from image and diffraction plane pictures. Optik **35**, 237–246 (1972)
7. Hitaj, B., Ateniese, G., Perez-Cruz, F.: Deep models under the GAN: Information leakage from collaborative deep learning. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. pp. 603–618. Association for Computing Machinery (2017). https://doi.org/10.1145/3133956.3134012
8. Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P.S., Zhang, X.: Membership inference attacks on machine learning: A survey. ACM Computing Surveys **54**(11s), Article 235 (2022). https://doi.org/10.1145/3523273
9. Lundervold, A.S., Lundervold, A.: An overview of deep learning in medical imaging focusing on MRI. Zeitschrift für Medizinische Physik **29**(2), 102–127 (2019). https://doi.org/10.1016/j.zemedi.2018.11.002
10. Mahendran, A., Vedaldi, A.: Understanding deep image representations by inverting them. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5188–5196 (2015)
11. McMahan, B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.y.: Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. vol. 54, pp. 1273–1282. PMLR (2017), https://proceedings.mlr.press/v54/mcmahan17a.html
12. Mehmood, M.H., Khan, M., Ibrahim, A.: Balancing privacy and accuracy: Federated learning with differential privacy for medical image data. IEEE (2025). https://doi.org/10.1109/DSIT61374.2024.10880906
13. Nasr, M., Shokri, R., Houmansadr, A.: Comprehensive privacy analysis of deep learning: Stand-alone and federated learning under passive and active white-box inference attacks (2018). https://doi.org/10.48550/arXiv.1812.00910

14. Nickparvar, M.: Brain tumor MRI dataset. Kaggle (2021), https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset
15. NVIDIA: NVIDIA FLARE: Federated learning application runtime environment. https://github.com/NVIDIA/NVFlare (2022), open-source federated learning framework
16. Phong, L.T., Aono, Y., Hayashi, T., Wang, L., Moriai, S.: Privacy-preserving deep learning: Revisited and enhanced. In: Applications and Techniques in Information Security. pp. 100–110. Springer Singapore (2017)
17. Salem, A., Zhang, Y., Humbert, M., Fritz, M., Backes, M.: ML-Leaks: Model and data independent membership inference attacks and defenses on machine learning models (2018). https://doi.org/10.48550/arXiv.1806.01246
18. Shen, D., Wu, G., Suk, H.I.: Deep learning in medical image analysis. Annual Review of Biomedical Engineering **19**, 221–248 (2017). https://doi.org/10.1146/annurev-bioeng-071516-044442
19. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: Proceedings of the IEEE Symposium on Security and Privacy. pp. 3–18. IEEE (2017). https://doi.org/10.1109/SP.2017.41
20. Singh, S., Sikka, H., Kotti, S., Trask, A.: Benchmarking differentially private residual networks for medical imagery (2020)
21. Smith, S.M., Brady, J.M.: SUSAN: A new approach to low level image processing. International Journal of Computer Vision **23**(1), 45–78 (1997). https://doi.org/10.1023/A:1007963824710
22. U.S. Department of Health and Human Services: Health insurance portability and accountability act of 1996 (1996), https://www.hhs.gov/hipaa/
23. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing **13**, 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
24. Zhang, F., Zhai, D., Bai, G., Jiang, J., Ye, Q., Ji, X., Liu, X.: Towards fairness-aware and privacy-preserving enhanced collaborative learning for healthcare. Nature Communications **16**(1), 2852 (2025). https://doi.org/10.1038/s41467-025-58055-3