High-Capacity Robust Medical Image Exfiltration via Neural Network Weight Replacement

Elie Thellier*, Huiyu Li, Nicholas Ayache, and Hervé Delingette

Université Côte d'Azur, Inria, Epione Team, Sophia Antipolis, France

Abstract. Collaborative medical AI platforms allow researchers to train models on sensitive imaging data while restricting data export. However, trained models can serve as covert carriers of patient information: medical images may be encoded within model parameters and reconstructed outside the secure environment. Existing defenses rely on lightweight sanitization (e.g., fine-tuning, pruning, quantization) and limited statistical auditing, creating a realistic insider exfiltration risk. We introduce a high-capacity neural steganography attack that encodes medical images as continuous latent representations embedded into model initialization. A StyleGAN2-based adversarial autoencoder learns compact latent codes regularized to match standard weight initialization statistics, keeping embedded parameters statistically aligned with clean models. Noise injection during training improves robustness to export-time mitigation. The carrier model remains functional on its intended task and hidden images can be reconstructed directly from the model weights after export. This continuous encoding enables robust and scalable exfiltration, allowing up to 99 brain MRI volumes to be embedded within a 30MB model, and remains recoverable under mitigations that disrupt prior bit-level schemes. While reconstructions are approximate rather than lossless, embedded content remains anatomically recognizable and recoverable at scale, exposing a privacy risk distinct from prior bit-level approaches. Experiments on MIMIC-CXR, BraTS, and LiTS demonstrate effectiveness across modalities, tasks, and architectures, highlighting the need for structural defenses beyond parameter-level sanitization.

Keywords: Medical AI security · Image exfiltration · Steganography

1 Introduction

Collaborative medical AI platforms, often referred to as data lakes [1], allow researchers to train models on sensitive imaging data while preventing the export of raw images. Trained models, however, are exportable after validation, under the assumption that they consist only of numerical parameters. This creates a privacy risk: a malicious user can encode medical images inside model weights and reconstruct them after export, potentially recovering clinically relevant structures. Export-time defenses are lightweight and utility-preserving, where utility

* Corresponding author: elie.thellier@inria.fr

denotes performance on the intended clinical task. Common mitigations such as fine-tuning, pruning, or quantization [2,3,4] aim to remove hidden information while maintaining accuracy. Auditing remains limited, as architectural inspection is rare and operators often lack trusted benign reference models. Under this threat model, an attack that survives mild sanitization may remain undetected.

Image exfiltration relates to neural steganography, where secret information is hidden inside neural networks to produce a stego model that appears benign. Existing methods follow two main directions, both with limitations. Some embed a secondary task within a legitimate model [5,6,7,8,9,10]; for example, the Transpose attack [5] jointly trains a classifier to memorize images. Because recovery depends on precise parameter configurations, standard mitigation often disrupts these attacks [2], and their reliance on modified training or architectures limits compatibility with standard medical models. Others encode images directly into weights as binary streams using least significant bit or related encodings [11,12,13,14]. The Data Exfiltration by Compression (DEC) attack [12] improves capacity but relies on bit-level encoding, which is inherently fragile: small weight perturbations induce bit flips that severely degrade reconstruction [2]. Adversarial variants improve stability [15], yet secret remains constrained to bit messages. Watermarking approaches [16,17,18] offer only a few thousand bits, far below what is needed for full medical images, while implicit neural representation methods [19,20] enable continuous encoding but require non-standard architectures. Current approaches trade compatibility, capacity, and robustness: task-based methods are mitigation-sensitive, bit-level schemes are fragile and capacity-limited, and continuous implicit methods reduce compatibility.

We address this gap by introducing a continuous, high-capacity image exfiltration attack compatible with standard medical imaging models. Images are compressed into compact latent codes via a StyleGAN2-based [21] adversarial autoencoder, replacing previous bit-level encoding with a continuous representation that is inherently more robust to weight perturbations. Export-time modifications are simulated during training to improve robustness, and latent codes are constrained to match standard weight initialization statistics for stealthiness. The compressed images are embedded into the initialization of a conventional model that remains fully functional. We evaluate across datasets, modalities, and architectures, showing that dozens of volumetric scans can be embedded without degrading task performance and that reconstruction survives sanitization, motivating stronger structural auditing in collaborative medical AI systems.

2 Methodology

We propose a continuous neural steganography framework that embeds compressed images as model parameters while preserving downstream utility and statistical plausibility. As illustrated in Fig. 1, the method consists of three stages: (1) learning a robust latent representation of medical images, (2) embedding these latent codes into a standard medical model before training, and (3) extracting and reconstructing images after export and potential mitigation.

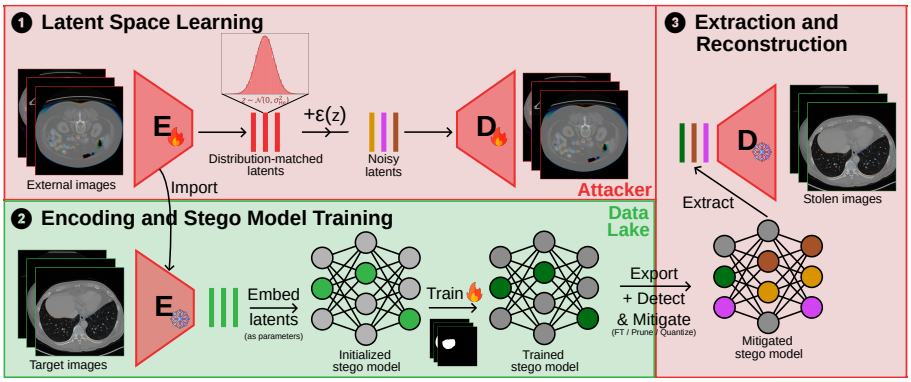

Fig. 1. Overview of the proposed medical image exfiltration framework.

Threat model. We assume an insider adversary with authorized access to a collaborative medical platform. The adversary can train models on sensitive data and export validated models, but cannot directly export raw images. Before release, models may undergo lightweight sanitization such as fine-tuning, pruning, or quantization, intended to preserve utility while disrupting hidden information. The defender performs limited statistical inspection of model parameters. The adversary's objective is to embed a large number of medical images into a legitimate model so that they remain recoverable after mitigation, while ensuring the model remains functional and statistically indistinguishable from a benign one.

2.1 Learning a Robust and Compact Latent Space

To achieve high capacity, images are compressed into compact continuous latent codes. We use a StyleGAN2-based adversarial autoencoder trained to reconstruct medical images from a related external dataset to reduce domain shift. Given an image x, the encoder E produces a latent vector $w = E(x) \in \mathbb{R}^d$ with $d = 512$, providing strong compression while preserving perceptual quality. The generator G reconstructs the image using adversarial, ℓ_2, and LPIPS [22] losses.

Continuous latents provide inherent tolerance to small perturbations. To further improve robustness to mitigation, we inject scale-adaptive Gaussian noise during training, $\tilde{w} = w + \lambda_{\text{noise}}\sigma\epsilon, \epsilon \sim \mathcal{N}(0, I)$, where σ is the minibatch standard deviation. This simulates parameter drift caused by optimization or sanitization.

To ensure stealthiness, defined as statistical consistency with a benign model, we align latent statistics with Kaiming (He) initialization [23], commonly used for non-symmetric activation functions. Rather than matching layer-specific statistics, we estimate global reference moments (mean, variance, skewness, kurtosis) from Kaiming-initialized layers of the target architecture. Let μ and σ denote the mean and standard deviation of a latent batch, and $(\mu^{\text{He}}, \sigma^{\text{He}})$ the reference

statistics from Kaiming-initialized weights. We minimize

$$\mathcal{L}_{\text{dist}} = \left(\frac{\mu - \mu^{\text{He}}}{\sigma^{\text{He}}}\right)^2 + \left(\frac{\sigma - \sigma^{\text{He}}}{\sigma^{\text{He}}}\right)^2 + \lambda_{\text{skew}}(\mathbb{E}[\bar{w}^3])^2 + \lambda_{\text{kurt}}(\mathbb{E}[\bar{w}^4] - 3)^2,$$

where $\bar{w} = (w - \mu)/\sigma$. This encourages latent distributions to be consistent with standard model initialization.

2.2 Embedding Latent Codes into the Model

Using the trained encoder, each target image x_i is mapped to a latent vector $w_i \in \mathbb{R}^d$. Embedding N images therefore requires storing $N \times d$ real-valued parameters, which directly determines capacity.

Latent vectors are inserted into selected convolutional or fully connected weight tensors of a standard medical imaging model at its initialization. Biases and normalization parameters remain unchanged and follow Kaiming initialization [23]. Layers are selected based on payload size N and expected Kaiming variance: candidate tensors are sorted by descending expected Kaiming variance, and latent codes are inserted sequentially into higher-variance tensors first, since these better tolerate perturbations [15], until the desired capacity is reached or all tensors are filled.

The resulting stego model is trained normally on its intended task without freezing embedded parameters. Because latent statistics match benign initialization, training proceeds as in a standard model and preserves downstream utility.

2.3 Extraction and Reconstruction

After mitigation and export, the attacker retrieves the parameters from the embedding layers using the known insertion order and splits them into latent vectors $\hat{w}_i$ of size d. Reconstructed images are obtained through the pretrained generator: $\hat{x}_i = G(\hat{w}_i)$. Robustness is achieved if latent perturbations caused by mitigation result only in gradual degradation rather than catastrophic failure.

3 Experimental Setup

3.1 Evaluation Protocol

We evaluate the attack along four criteria: utility, capacity, robustness, and stealthiness, and compare against two baselines, Transpose [5] and DEC [12]. Utility measures performance on the legitimate medical task, reported as Dice (segmentation) and AUC/accuracy (classification). Capacity corresponds to the number of embedded images N, equivalently the proportion of model parameters replaced by latent codes. Reconstruction quality is measured by PSNR, SSIM, and LPIPS between reconstructed and original images. Stealthiness quantifies statistical detectability via Jensen–Shannon divergence (JSD) between stego and

benign models, averaged across layers. Benign references are trained under similar settings without latent embedding. This assumes access to a benign reference beyond what our threat model allows, but provides a worst-case audit, since failure here implies weaker reference-free detectors are unlikely to succeed.

3.2 Datasets and Models

We evaluate on three public medical benchmarks: MIMIC-CXR [24] (54,038 chest X-rays for multi-label classification), BraTS 2021 [25] (1,251 MRI for brain tumor segmentation), and LiTS [26] (130 abdominal CT volumes for liver tumor segmentation). To simulate a realistic attack scenario where the compression model is trained outside the secure environment, we use external datasets: FLARE 2021 [27] for CT, BraTS-Reg 2022 [28] for MRI, and a disjoint MIMIC-CXR split for X-ray. Volumes are processed slice-wise using three adjacent slices concatenated, following [12]. For segmentation, we use U-Net; for classification, DenseNet121. Both are 30MB (≈8M parameters), reflecting realistic deployment settings.

3.3 Baselines and Mitigation

We use official implementations of Transpose [5] and DEC [12], adapted to our architectures. Unless stated otherwise, our method embeds 1,641 images (1641×512 parameters, approximately 10% of a 30MB model). Transpose only applies to classification and is evaluated on MIMIC-CXR with a capacity of 100 images. DEC is evaluated on segmentation tasks at its maximum feasible capacity, allowing exfiltration of 67 images on BraTS and 15 on LiTS. We compare by exfiltrated payload rather than parameter percentage, since DEC's lower compression efficiency would make percentage-matched comparison even more degenerate (1–6 images for DEC vs. 1,600+ for ours at 10% capacity).

To assess robustness, we apply mitigation strategies with hyperparameters tuned to preserve task utility: Super-FT (2 epochs, base lr 10^{-4} with peak lr $10^{-2}/10^{-3}$, phase ratio 0.1) [3], LWLRD fine-tuning (3 epochs) [2], unstructured magnitude pruning (40% of weights removed), Fine-Pruning (8% allowed accuracy drop) [4], and post-training 8-bit and 4-bit quantization.

We set $\lambda_{\text{noise}} = 0.2$, $\lambda_{\text{skew}} = 0.5$, and $\lambda_{\text{kurt}} = 0.2$.

4 Results

4.1 Capacity Trade-offs

We analyze the trade-offs induced by a change in attack capacity. Fig. 2 reports segmentation Dice, reconstruction SSIM, and stealthiness (JSD) as capacity increases from 0% to 100% of model parameters. As capacity increases, Dice and SSIM remain stable, JSD stays close to the benign reference and is comparable to benign models trained with different weight decay (WD), indicating statistical consistency. At full capacity, a 30MB model (≈7.8M parameters) stores up to 99 MRI volumes of 153 slices each ($99 \times 153 \times 512$ latent parameters), showing most parameters can be repurposed without compromising utility or stealthiness.

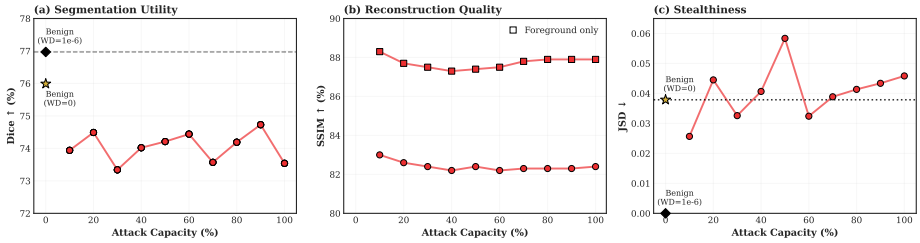

Fig. 2. Effect of attack capacity on utility, reconstruction quality, and stealthiness. "Foreground only" SSIM focuses on anatomically relevant structures (background thresholded at 5th percentile). Benign models (0% capacity) are shown for two weight decay (WD) values.

4.2 Comparison with Baseline Attacks

Table 1 compares our method with Transpose and DEC across datasets. Our approach preserves task performance close to the benign model while achieving much higher capacity (1,641 images at 10% capacity in a 30MB model vs. 100 for Transpose and up to 67/15 for DEC). It also maintains competitive utility and reconstruction fidelity with low JSD, indicating statistical consistency.

Table 1. Comparison with baseline attacks on classification and segmentation tasks.

Method	MIMIC-CXR					BraTS				LiTS			
	AUC	Acc	SSIM	LPIPS	JSD	Dice	SSIM	LPIPS	JSD	Dice	SSIM	LPIPS	JSD
Benign	75.9	72.4	—	—	0	77.0	—	—	0	81.5	—	—	0
Transpose [5]	48.6	64.6	68.6	0.48	0.08	—	—	—	—	—	—	—	—
DEC [12]	—	—	—	—	—	77.0	95.4	0.01	0	81.5	98.2	0.01	0
Ours	73.6	71.2	64.4	0.32	0.01	73.9	83.2	0.08	0.02	81.0	53.2	0.35	0.06

Although DEC achieves higher raw SSIM before mitigation, its capacity is substantially lower and reconstruction quality deteriorates sharply under mitigation (Fig.3). In contrast, our method prioritizes robustness and scalability over peak pre-mitigation fidelity. Notably, pixel-level reconstruction quality is not critical, as moderate reconstructions may retain clinical information.

4.3 Robustness to Mitigation

Fig. 3 evaluates robustness to fine-tuning, pruning, and quantization, showing the utility–reconstruction trade-off under different mitigations. Our method maintains high reconstruction quality and downstream utility, whereas baselines degrade sharply, especially under pruning and quantization. Under stronger perturbations (e.g., 4-bit quantization), baselines collapse toward low-utility, low-reconstruction regions, whereas our method remains stable.

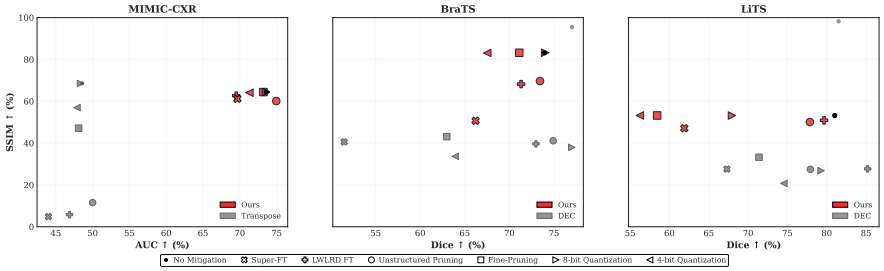

Fig. 3. Robustness to mitigation: reconstruction quality vs. utility.

4.4 Visual Reconstruction Quality

Pre- and post-mitigation reconstructions are shown in Fig. 4. The largest degradation is observed pre-mitigation due to domain shift in compression model training. After mitigation, reconstructions remain largely consistent across modalities, with only a small residual gap even under strong shifts (e.g., BraTS), indicating that the attack remains effective and robust to the applied mitigation strategies.

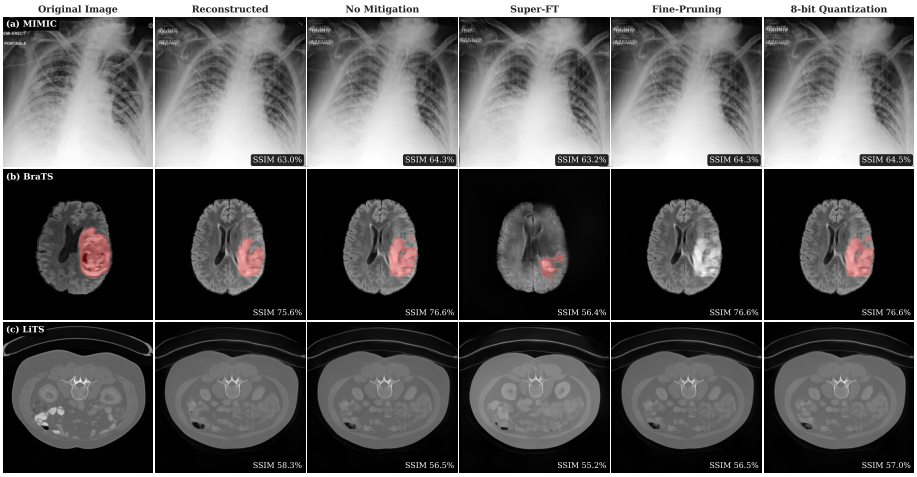

Fig. 4. Qualitative reconstruction examples on MIMIC-CXR, BraTS, and LiTS. Originals (left) and reconstructions before and after mitigation (fine-tuning, pruning, quantization), with SSIM indicated. Predicted segmentation masks are shown in red.

4.5 Ablation Study

Table 2 evaluates the contribution of each component. Removing noise injection reduces post-mitigation SSIM, indicating lower robustness under mitigation.

Table 2. Ablation study. Variants are evaluated in terms of utility (Dice), reconstruction (SSIM/LPIPS), and stealthiness (JSD), before and post LWLRD FT mitigation.

Variant	Dice	SSIM	LPIPS	JSD	Dice (post)	SSIM (post)
Benign Model	77.0	—	—	0	75.8	—
Full Method	73.9	83.2	0.08	0.02	71.3	68.2
Full w/o Noise Injection	74.2	80.3	0.09	0.02	71.9	53.5
Full w/o Layer Selection	73.9	82.7	0.08	0.03	71.4	62.7
Full w/o Moment Loss	72.3	82.6	0.09	0.18	70.5	67.9
Full w/ 2-Moment Loss	72.6	82.8	0.09	0.17	70.8	67.5
Full w/ Frozen Latents	68.7	85.0	0.08	0.04	66.9	72.4

Dropping moment alignment increases JSD, weakening stealthiness, while the 2-moment variant also shows elevated JSD. Freezing latent parameters degrades task utility and balance. Overall, robustness and stealth emerge from the joint effect of noise simulation and moment alignment in continuous latent encoding.

5 Discussion and Conclusion

We introduced a continuous, high-capacity image exfiltration attack compatible with imaging architectures and training pipelines. By encoding images as continuous latent vectors and aligning their statistics with standard initialization, the method embeds a high number of volumetric scans within a 30MB model while preserving downstream utility. Unlike bit-level schemes, the approach relies on continuous distributional alignment rather than exact parameter preservation, enabling graceful degradation under fine-tuning, pruning, and quantization.

The results suggest that parameter-level mitigations alone are insufficient: because embedded latents remain within the same statistical regime as standard weights, lightweight sanitization degrades reconstruction gradually but does not eliminate recoverability. More effective defenses may require structural interventions, such as permutation-based transformations or symmetry-breaking mechanisms [29,30], combined with deeper per-layer or correlation-aware audits.

This work has several limitations. Reconstruction fidelity depends on compression model alignment and may degrade under domain shifts. Our implementation represents one operating point in the fidelity-utility–capacity trade-off. As shown in Table 2, higher reconstruction quality (SSIM >85%) can be achieved by freezing latent parameters, at the cost of reduced utility (Dice 68.7% vs. 73.9%). Higher-dimensional latent representations (e.g., W+ instead of W) can further improve reconstruction fidelity at the expense of capacity, indicating that the reported reconstruction quality is not an upper bound. Furthermore, we do not evaluate DP-SGD [31], which modifies training rather than applying post-hoc sanitization and therefore falls outside the scope of this work. In addition, our stealthiness evaluation assumes access to a benign reference model, which is stronger than the stated threat model, motivating reference-free auditing methods such as anomaly or outlier detection over imported models. Finally, we do

not directly measure whether reconstructed images preserve clinically identifiable or actionable content (e.g., through re-identification or downstream task transfer using only exfiltrated data); our reconstruction metrics (SSIM, LPIPS) serve as a proxy for this risk but are not a substitute for it, as even imperfect reconstructions may still preserve clinically relevant information.

Future work includes evaluation against DP-SGD and structural defenses, improved volumetric compression via 3D generative models [32], clinical and re-identification risk assessment, and sensitivity analysis of λ hyperparameters. Overall, our results show that high-capacity medical image exfiltration can be made stealthy, utility-preserving, and resilient to common post-hoc mitigations, highlighting the need for structural defenses and model auditing.

Code Availability The code will not be publicly released due to dual-use concerns.

Acknowledgments. This work was supported by the Région PACA and France 2030 initiative through the funding of the I-Démo project PLICIA, and by the French government through the National Research Agency (ANR), under the IA Cluster projects (ANR-23-IACL-0001). Experiments in this paper were carried out using the Jean Zay supercomputer, a national computing facility operated by IDRIS (CNRS) and provided through GENCI under the French Ministry of Higher Education and Research.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Hai, R., Koutras, C., Quix, C., Jarke, M.: Data lakes: A survey of functions and systems. IEEE Transactions on Knowledge and Data Engineering **35**(12), 12571–12590 (2023)
2. Thellier, E., Li, H., Ayache, N., Delingette, H.: Mitigating data exfiltration attacks through layer-wise learning rate decay fine-tuning. In: Bridging Regulatory Science and Medical Imaging Evaluation; and Distributed, Collaborative, and Federated Learning. pp. 80–90. Springer Nature Switzerland, Cham (2026)
3. Sha, Z., He, X., Berrang, P., Humbert, M., Zhang, Y.: Fine-tuning is all you need to mitigate backdoor attacks. arXiv preprint arXiv:2212.09067 (2022)
4. Liu, K., Dolan-Gavitt, B., Garg, S.: Fine-pruning: Defending against backdooring attacks on deep neural networks. In: International Symposium on Research in Attacks, Intrusions, and Defenses (RAID). pp. 273–294. Springer (2018)
5. Amit, G., Levy, M., Mirsky, Y.: Transpose attack: Stealing datasets with bidirectional training. In: Proceedings of the Network and Distributed System Security Symposium (NDSS) (2024)
6. Fan, Y., Di, F., Zhang, M., Wang, Z., Liu, J.: Lossless steganographic network via model arithmetic operations. Neural Networks p. 107918 (2025)
7. Guo, C., Wu, R., Weinberger, K.Q.: On hiding neural networks inside neural networks. arXiv preprint arXiv:2002.10078 (2020)
8. Li, G., Li, S., Li, M., Zhang, X., Qian, Z.: Steganography of steganographic networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 5178–5186 (2023)

9. Li, G., Li, S., Luo, Z., Qian, Z., Zhang, X.: Purified and unified steganographic network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 27569–27578 (2024)
10. Pan, X., Zhang, M., Yan, Y., Zhang, S., Yang, M.: Matryoshka: Exploiting the over-parametrization of deep learning models for covert data transmission. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(2), 663–678 (2024)
11. Agrawal, R., Jou, K., Obili, T., Parikh, D., Prajapati, S., Seth, Y., Sridhar, C., Zhang, N., Stamp, M.: On the steganographic capacity of selected learning models. In: Machine Learning, Deep Learning and AI for Cybersecurity, pp. 457–491. Springer (2025)
12. Li, H., Ayache, N., Delingette, H.: Data exfiltration by compression attack: Definition and evaluation on medical image data. Journal of Machine Learning for Biomedical Imaging **3**, 728–756 (December 2025)
13. Song, C., Ristenpart, T., Shmatikov, V.: Machine learning models that remember too much. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS). pp. 587–601 (2017)
14. Xu, N., Liu, Q., Liu, T., Liu, Z., Guo, X., Wen, W.: Stealing your data from compressed machine learning models. In: Proceedings of the 2020 57th ACM/IEEE Design Automation Conference (DAC). pp. 1–6. IEEE (2020)
15. Xu, C., Huang, L., Qin, C., Feng, G., Zhang, X.: Steganography with constructing neural networks. IEEE Transactions on Circuits and Systems for Video Technology (2025)
16. Li, Y., Tondi, B., Barni, M.: Spread-transform dither modulation watermarking of deep neural network. Journal of Information Security and Applications **63**, 103004 (2021)
17. Tondi, B., Costanzo, A., Barni, M.: Robust and large-payload dnn watermarking via fixed, distribution-optimized, weights. IEEE Transactions on Dependable and Secure Computing **22**(2), 1082–1097 (2024)
18. Zhao, G., Qin, C., Yao, H., Han, Y.: Dnn self-embedding watermarking: Towards tampering detection and parameter recovery for deep neural network. Pattern Recognition Letters **164**, 16–22 (2022)
19. Dong, W., Liu, J., Chen, L., Sun, W., Pan, X., Ke, Y.: Implicit neural representation steganography by neuron pruning. Multimedia Systems **30**(5), 266 (2024)
20. Luo, P., Liu, J., Ke, Y., Dang, Q., Zhang, M., Mu, D.: Hiding functions within functions: Steganography by implicit neural representations. Tsinghua Science and Technology **31**(2), 1058–1074 (2026)
21. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8110–8119 (2020)
22. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)
23. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 1026–1034 (2015)
24. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv:1901.07042 (2019)
25. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al.: The rsna-asnr-miccai

brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021)
26. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., et al.: The liver tumor segmentation benchmark (lits). Medical Image Analysis **84**, 102680 (2023)
27. Ma, J., Zhang, Y., Gu, S., An, X., Wang, Z., Ge, C., Wang, C., Zhang, F., Wang, Y., Xu, Y., et al.: Fast and low-gpu-memory abdomen ct organ segmentation: the flare challenge. Medical Image Analysis **82**, 102616 (2022)
28. Baheti, B., Waldmannstetter, D., Chakrabarty, S., Akbari, M., Bilello, M., Wiestler, B., Schwarting, J., Calabrese, E., Jeffrey, R., Abidi, S., et al.: The brain tumor sequence registration challenge: establishing correspondence between pre-operative and follow-up mri scans of diffuse glioma patients. arXiv preprint arXiv:2112.06979 (2021)
29. Gilkarov, D., Dubin, R.: Neuperm: Disrupting malware hidden in neural network parameters by leveraging permutation symmetry. arXiv:2510.20367 (2025)
30. Torpmann-Hagen, B., Riegler, M.A., Halvorsen, P., Johansen, D.: Defending against stegomalware in deep neural networks with permutation symmetry. arXiv preprint arXiv:2509.20399 (2025)
31. 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 (CCS). pp. 308–318 (2016)
32. Hong, S., Marinescu, R., Dalca, A.V., Bonkhoff, A.K., Bretzner, M., Rost, N.S., Golland, P.: 3d-stylegan: A style-based generative adversarial network for generative modeling of three-dimensional medical images. In: MICCAI Workshop on Deep Generative Models. pp. 24–34. Springer (2021)