Unlearning for Better learning
Corrective Unlearning for MRI Reconstruction

Chinmay Sharma [1] **Aryaman Bahl** [1] **Sairam Babu** [1] **Pranav Subramaniam** [2]

Abstract

Magnetic Resonance Imaging reconstruction accelerates image acquisition by reconstructing high-quality images from undersampled k-space data using deep learning. However, real-world deployment of these models remains hindered by concerns around trustworthiness, generalization, and data privacy, especially in the presence of corrupted or adversarial training samples. We propose a Corrective Machine Unlearning framework that selectively removes the influence of harmful data while preserving overall model performance. By leveraging techniques such as Selective Synaptic Dampening, our approach targets specific parameter updates guided by the Fisher information matrix to forget poisoned representations. We further integrate interpretability tools to analyze model behavior and ensure robustness. Experimental results on MRI reconstruction tasks demonstrate that Corrective Machine Unlearning can effectively mitigate artifacts introduced through data poisoning while maintaining high fidelity on untainted inputs. Our findings underscore the promise of corrective unlearning as a practical step toward safer, privacy preserving, and clinically reliable MRI systems.

1. Introduction

Magnetic Resonance Imaging (MRI) plays a pivotal role in clinical diagnostics by providing high resolution, non invasive visualization of anatomical structures. However, the long acquisition times required for fully sampled MRI scans pose a critical bottleneck in time sensitive or resource constrained healthcare settings. To overcome this, accelerated reconstruction techniques such as compressed sensing (CS) and deep learning-based models have emerged as viable alternatives, enabling image reconstruction from undersampled k-space data.

Recent advances in deep learning, especially end-to-end architectures like variational networks, have achieved state-of-the-art performance in fast and accurate MRI reconstruction. Nonetheless, these models often suffer from limited generalization and lack of robustness when exposed to out of distribution inputs or corrupted training data. In clinical contexts, such vulnerabilities can manifest as hallucinated structures, misleading artifacts, or biased reconstructions, posing serious risks to diagnostic reliability.

These challenges are exacerbated by the use of aggregated or web-scale datasets, which may contain noisy, adversarially manipulated, or demographically skewed samples. Consequently, ensuring the trustworthiness, privacy, and fairness of reconstruction models has become increasingly vital, especially as regulations such as the GDPR enforce the ”right to be forgotten” in AI systems.

Machine unlearning has emerged as a promising paradigm to address these concerns by enabling targeted removal of data influence from trained models. Traditional unlearning approaches primarily aim to satisfy privacy constraints through exact or approximate deletion of user data. However, these methods often require full retraining or compromise model utility.

In this work, we explore Corrective Machine Unlearning as a principled framework to selectively forget corrupted or adversarial samples while preserving model performance. Our approach focuses on addressing structural biases and hallucinations in MRI reconstructions by modifying model parameters post-training using methods such as Selective Synaptic Dampening (SSD). By doing so, we aim to enhance the robustness, interpretability, and clinical viability of MRI reconstruction systems without incurring the computational costs of full retraining.

[1]CCNSB, IIIT Hyderabad, Hyderabad, India [2]CSL, IIIT Hyderabad, Hyderabad, India. Correspondence to: Chinmay Sharma <chinmay.sharma@research.iiit.ac.in>, Aryaman Bahl <aryaman.bahl@research.iiit.ac.in>, Sairam Babu <sairam.babu@research.iiit.ac.in>, Pranav Subramaniam <pranav.subramaniam@research.iiit.ac.in>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

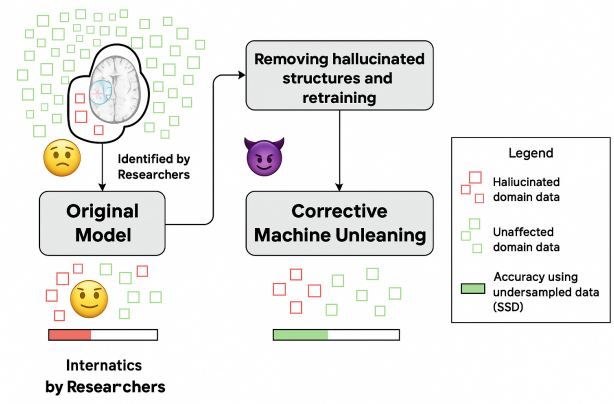


Figure 1. Conceptual overview of our corrective machine unlearning framework. Hallucinated structures introduced by corrupted training data are identified and removed via retraining or mitigated through Selective Synaptic Dampening (SSD), resulting in improved reconstruction fidelity and model robustness.

2. Related Work

2.1. Medical Imaging

Deep learning has transformed medical imaging by enabling automated image segmentation, detection, and classification with high accuracy. Foundational surveys such as those by Litjens et al. and Shen et al. provide comprehensive overviews of how convolutional neural networks and other architectures have been applied across clinical modalities (Litjens et al., 2017; Shen et al., 2017). Recent progress in algorithm design and hardware acceleration has further enhanced both throughput and image quality (Xue et al., 2024). Additionally, the integration of multimodal imaging data has improved diagnostic performance, making AI-powered tools increasingly viable in clinical decision-making pipelines.

2.2. MRI Reconstruction

MRI reconstruction from undersampled k-space data remains a critical application of deep learning in radiology. Classical approaches rely on inverse Fourier transforms, while recent end-to-end models such as variational networks have achieved state-of-the-art results in generating high-fidelity images (Lustig et al., 2007; Hammernik et al., 2018). These reconstructions are essential for detecting abnormalities such as tumors or lesions, and thus any corrective technique applied—such as machine unlearning—must preserve diagnostic integrity. The need to remove erroneous or sensitive data without degrading performance has made MRI reconstruction a natural testbed for responsible and adaptive learning frameworks.

2.3. Right to be Forgotten

The General Data Protection Regulation (GDPR) and similar privacy laws enforce the "right to be forgotten," requiring that individuals' data be deleted upon request. In the context of machine learning, this introduces challenges since trained models may implicitly encode information about specific training samples (Voigt & Von dem Bussche, 2017). Compliance with such regulations has sparked active research on how to efficiently and verifiably remove the influence of deleted data while maintaining model utility.

2.4. Machine Unlearning

Machine unlearning aims to remove the effect of specific training data from a model without retraining it from scratch. Early work explored both certified removal algorithms and empirical strategies using parameter perturbation, gradient ascent, or dataset filtering. These methods vary in their assumptions and computational costs but converge on the goal of making model predictions statistically indistinguishable from those trained without the target data. As privacy regulations tighten and large-scale models become standard, machine unlearning has emerged as a cornerstone of responsible and secure machine learning.

2.5. Corrective Machine Unlearning

Corrective Machine Unlearning extends traditional unlearning approaches by not only eliminating the effect of corrupted or adversarial data but also preserving or recovering the model's performance post-unlearning. This is particularly important in high-stakes domains like medical imaging, where reconstruction quality directly affects clinical decision-making. Recent works investigate techniques such as selective weight dampening, adaptive regularization, and structured retraining to achieve this goal. These corrective strategies aim to ensure fairness, reliability, and interpretability—highlighting the growing intersection of safety, privacy, and performance in modern AI systems.

3. Problem Formulation

Let D be the complete training dataset on which a model M with parameters θ is originally trained. We denote the *forget set* as

$$D_f \subset D,$$

and the remaining data as

$$D_r = D \setminus D_f.$$

Our goal is to obtain a new model M' with parameters θ' that effectively "forgets" the influence of D_f while maintaining performance on D_r. Formally, we desire:

$$\theta' \approx \arg\min_{\theta} \mathcal{L}(\theta; D_r),$$

so that M' is statistically indistinguishable from a model trained from scratch on D_r.

Step-by-Step Unlearning Procedure

1. **Identify Forget Set:** Define D_f based on legal or privacy requirements. The remaining dataset is $D_r = D \setminus D_f$.

2. **Objective Specification:** Let the loss function be $\mathcal{L}(\theta; \cdot)$. The corrective unlearning task is to update

$$\theta \to \theta'$$

such that

$$\theta' \approx \arg\min_{\theta} \mathcal{L}(\theta; D_r).$$

3. **Evaluation Criterion:** Ensure that the new model M' satisfies

$$M' \approx M_r,$$

where M_r is the model trained solely on D_r.

Unlearning Approaches

We consider several corrective machine unlearning methods(Goel et al., 2024) tailored for MRI data, each with its own update mechanism:

1. **Retrain from Scratch:** Directly retrain the model using D_r:

$$\theta' = \arg\min_{\theta} \mathcal{L}(\theta; D_r).$$

Although theoretically ideal, this approach is computationally prohibitive in many settings.

2. **Selective Synaptic Dampening (SSD):** Leverage the Fisher information matrix computed on the forget set D_f to identify and selectively dampen the synaptic weights (Foster et al., 2024) most sensitive to D_f. Concretely, compute:

$$F_{D_f} = \mathbb{E}_{x \sim D_f} \left[\nabla_\theta \log p(x|\theta) \nabla_\theta \log p(x|\theta)^\top \right],$$

and attenuate the corresponding parameters by applying a dampening factor. This targeted reduction mitigates the influence of D_f on the updated model M' while preserving performance on D_r.

3. **Bad Teacher Distillation:** Distill knowledge from a biased teacher model that overemphasizes D_f, then subtract the overlearned features to recover a balanced representation.

4. **Noisy Labelling (NL):** Introduce controlled noise to the labels of the forget set D_f. Specifically, for each $(x_f, y) \in D_f$, sample

$$\epsilon \sim \mathcal{N}(0, I) \quad \text{and set} \quad \tilde{y} = y + \lambda\, \epsilon,$$

where $\lambda > 0$ is a noise-scale hyperparameter. Then retrain by minimizing

$$\mathcal{L}_{NL}(\theta) = \mathbb{E}_{(x_f, y) \sim D_f,\ \epsilon \sim \mathcal{N}(0,I)} \big[\mathcal{J}(\theta; x_f, \tilde{y}) \big].$$

5. **Gradient Ascent:** Apply a targeted gradient ascent step on the loss function computed on D_f to counteract its contribution. For a learning rate η, update:

$$\theta' \leftarrow \theta' + \eta \nabla_\theta \mathcal{L}(\theta; D_f),$$

carefully tuning η to reverse the effects without destabilizing the model.

Hypothesis

We hypothesize that the Fisher information matrix of the forget set D_f encodes valuable structural information about how D_f influences the model parameters. Specifically, let

$$\mathcal{S}(D_f)$$

denote a set of structural characteristics derived from F_{D_f} (e.g., its eigenvalue spectrum, which reflects the sensitivity of different synapses). Then, there exists a mapping

$$h : \mathcal{S}(D_f) \to \Lambda,$$

where Λ represents the hyperparameter space for unlearning methods (such as the dampening coefficients in SSD). Our hypothesis posits that by exploiting these structural insights, one can more effectively tune the SSD approach (or related corrective unlearning methods), yielding a model M' that efficiently unlearns D_f while preserving the utility on D_r and enhancing performance in the context of MRI reconstruction.

4. Datasets

We utilize multiple publicly available MRI datasets to train and evaluate our proposed reconstruction and unlearning framework. These datasets vary in anatomical focus and acquisition characteristics, allowing a robust assessment of model performance under both clean and corrupted data regimes.

We initially experimented with the *fastMRI* dataset (Zbontar et al., 2018), which is widely adopted in MRI reconstruction research. However, due to its high computational demands—particularly in terms of memory and processing time—we excluded it from further experimentation.

Our primary training dataset is *M4Raw* (Lyu et al., 2023), which offers high-resolution k-space data across multiple anatomies with a compact footprint (28 GB), making it ideal for training high-fidelity reconstruction models within practical compute budgets.

To simulate real-world data corruption and evaluate unlearning capabilities, we introduce poisoned samples derived from two datasets: *BraTS*, which contains brain tumor MRI scans, and *EXBox1*, a curated set of artefact-heavy images designed to inject ghosting and intensity anomalies. These datasets were used exclusively for data poisoning, allowing us to assess how well the model forgets corrupted patterns while retaining generalization on clean samples.

This combination of clean and corrupted datasets enables a comprehensive investigation of the trade-offs between reconstruction fidelity, learning capacity, and unlearning effectiveness.

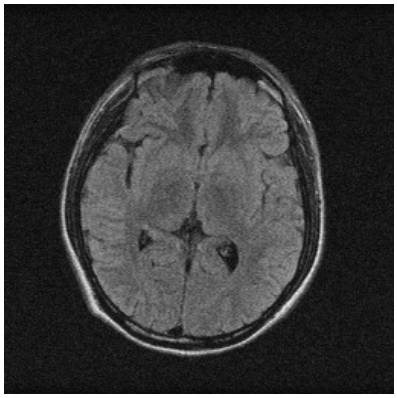
5. Experiments

5.1. Training Experiments and Attack Approaches

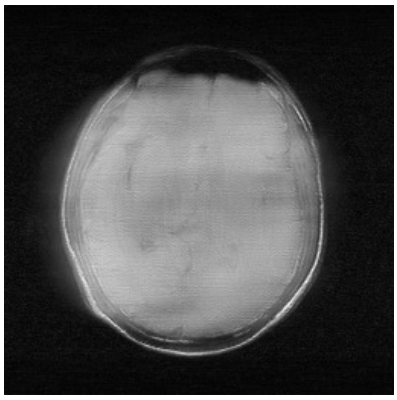
To simulate a real world scenario, all of our experiments incorporate a data poisoning attack on the training set. We introduce poisoning by intentionally artefacting the reconstructions of our dataset. In our approach, the reconstructions of the poisoned samples are modified to include artefacted versions of the original images in their "ground truth" reconstructions. This simulation mimics situations such as having poor MRI scans—common in the case of suboptimal data acquisition—or instances of malicious tampering with training data, potentially leading to artefact formation and misdiagnoses in medical imaging applications.

The implications of this approach are twofold. First, by introducing artefacts into the training data, we assess the model's ability to generalize in the presence of corrupted samples. Second, it allows us to study the adverse effects when a malicious actor compromises the training process, which may ultimately skew diagnostic outcomes. For instance, consider the following visual example: the left image represents the original ground truth reconstruction, while the right image demonstrates the ghost artefact reconstruction that emerged from the poisoning attack.
This design emphasizes the severity of poisoning attacks, illustrating how even slight alterations in the training data can lead to significant discrepancies in the reconstructions. The results underscore the need for robust data handling and model training protocols that can resist both inadvertent artefacts from poor-quality data and deliberate attacks on the integrity of the training process.

1. We have taken the baseline model as the retrained-from-scratch version trained using the E2E VarNet (Sriram et al., 2020). The model was trained on the M4Raw dataset due to compute constraints. Figures 3 and 5 show the loss graphs and the metrics—Normalized Mean Square Error (NMSE), Peak Signal-to-Noise Ratio (PSNR), and the Structural Similarity Index Measure (SSIM).



(a) Ground truth reconstruction.



(b) Reconstruction with ghost artefact due to poisoning.

Figure 2. Visual comparison of clean and poisoned MRI reconstructions. Artefacts in (b) reflect induced distortions simulating real-world data corruption.

2. The unlearning methods—Selective Synaptic Dampening (SSD), Bad Teacher Distillation, Noisy Labeling, and Gradient Ascent—are selected to unlearn on MRI data. Currently, unlearning is being performed using SSD. A comparative study of the performance of the unlearning methods will help validate our hypothesis of fine-tuning using SSD.

   (a) **SSD Toy Evaluation on MNIST Dataset:** To evaluate the reliability of Selective Synaptic Dampening (SSD), we performed a toy experiment using the MNIST dataset. We applied Fisher-based dampening on a defined forget set and monitored how well the model unlearns that data.

The training and validation loss curves showed rapid convergence with minimal overfitting. Embedding visualizations before SSD showed clear separation between the two classes, while post-SSD embeddings still showed distinct clusters—indicating ineffective unlearning. This suggests shared representations between forget and retain sets remained entangled.

To further understand SSD's effect, we analyzed the eigenvalue spectrum of the Fisher matrix. A steep drop revealed that only a few directions strongly influenced the forget set, implying selective but limited dampening.

Due to these limitations—including retained class separation and ineffective forgetting—we extended our study to the CIFAR-10 dataset using a ResNet-9 model for a more realistic evaluation of SSD.

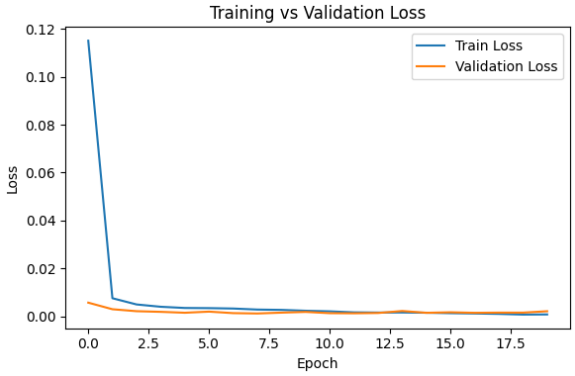


Figure 3. Training vs. validation loss curves. Both converge quickly with minimal overfitting.

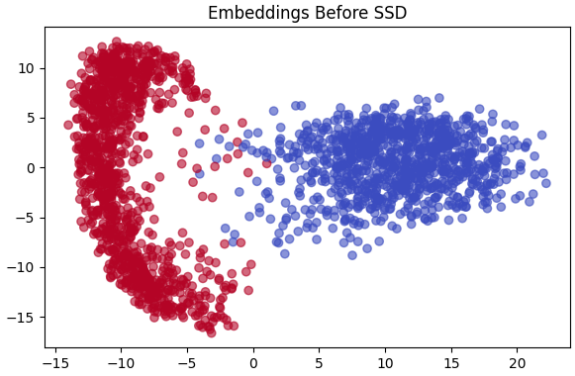


Figure 4. 2D latent embeddings before SSD. Classes are well-separated.

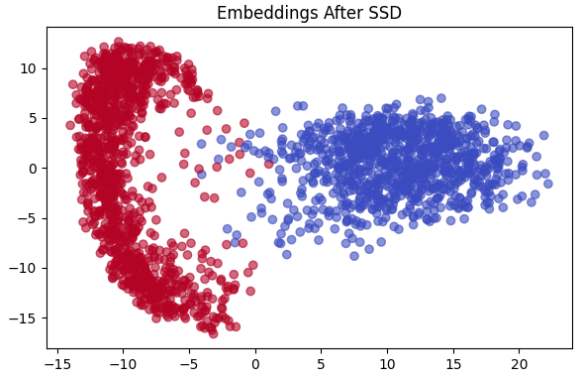


Figure 5. 2D latent embeddings after SSD. Class boundaries persist, indicating incomplete forgetting.

(b) **SSD Evaluation on CIFAR-10 Dataset:** To test our hypothesis of the effect of the structure of the forget set on the optimal hyperparameter values for Synaptic Dampening, we consider a grid search of unlearned models with different values of the Threshold and Dampening Multiplier parameters. We characterize the landscape of these unlearned models in terms of performance conservation on retain set and reduction on the forget set. The results in terms of retain and forget set accuracy are given below. We propose a metric parametrized by a parameter α that indicates the relative importance of having high accuracy on the retain set and poor performance on the forget set.

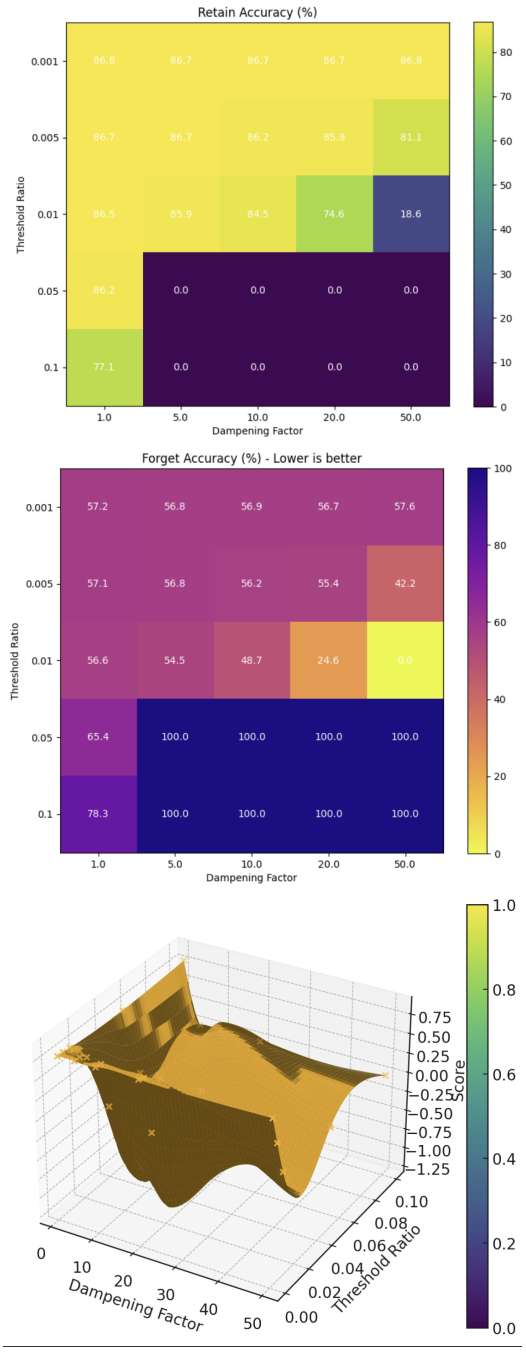
$$s = \alpha(acc_r) + (1 - \alpha)(1 - acc_f)$$

acc_r = retain set accuracy
acc_f = forget set accuracy

Given below are the results with $\alpha = 0.2$. We use this value because in the MRI reconstruction task, the loss of ability to classify patient data is more important than retaining model performance.
Based on the metric defined, we will compare the hyperparameter search trajectory of several common framework agnostic search algorithms such as Tree Based Parzen Estimators (TPE Search) with the structure aware algorithm we will propose.

i. We see that there is a balance between the retain set accuracies and forget set accuracies and a well defined peak in the hyperparameter vs performance landscape.
ii. For high dampening factors and high threshold values, the model performance suffers greatly. This is because SSD significantly

dampens a significant number of parameters.



5.2. Adversarial Attack Evaluation

To evaluate the robustness of our model against adversarial perturbations, we conducted targeted attacks on the MRI classification model. These attacks were designed to minimally perturb the input reconstruction such that the model misclassifies the result while preserving visual similarity to the original.

As shown in Figure 8, minimal perturbations can flip predictions while preserving visual similarity. The

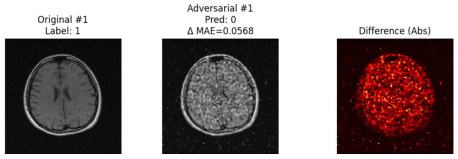


Figure 6. Adversarial attack on a sample: Original image (Label: 1) on the left and its adversarial counterpart (Predicted: 0) on the right. Minor structural noise led to class flipping.

difference map reveals targeted noise, emphasizing the model's sensitivity and the need for adversarial robustness in critical medical tasks.

5.3. CycleGAN-based Semantic Data Augmentation

In addition to adversarial robustness, we explored the use of Cycle-Consistent Generative Adversarial Networks (CycleGANs) for generating synthetic pathological variations in MRI scans. This approach is motivated by the lack of sufficient annotated medical data, especially in rare disease classes.

As shown in Figure 9, the CycleGAN learns to map

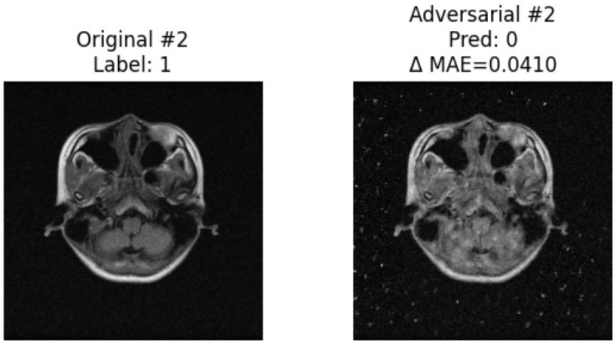


Figure 7. CycleGAN-generated transformation from a healthy MRI scan to a synthetic tumorous scan. Anatomical structure is preserved while adding realistic lesions.

healthy MRIs to tumorous-looking counterparts without requiring paired samples. The transformation preserves structural features while introducing plausible tumour-like patterns.

Such domain translation enables data augmentation for rare conditions and supports training in low-data regimes. Still, the realism and clinical relevance of generated lesions must be carefully validated by medical experts before use.

5.4. Unlearning Experiments and Results

We conducted a series of unlearning experiments to evaluate the effectiveness of various methods in mitigating the impact

of poisoned data on a reconstruction task. The experiments focused on a range of poisoned model percentages, with metrics evaluated on both the forget and retain sets. Below, we report the results of our most promising approach, Selective Synaptic Dampening (SSD), alongside a comparison with the original model's unlearning performance.

5.4.1. Metrics on Forget and Retain Sets

For the SSD approach, with a single hyperparameter configuration, we computed average metrics over 9,216 samples at a processing rate of 12.19 iterations per second. Table 1 summarizes the performance of SSD on the forget and retain sets, alongside the original model's unlearning performance on the forget set.

Table 1. Average metrics for various unlearning methods and the original model on forget and retain sets.

Model/Set	SSIM	PSNR	NMSE
SSD (Forget)	0.0967	13.9782	1.6210
SSD (Retain)	0.3824	22.1602	0.1807
SSD+FisherReg (Forget)	0.0712	12.8703	1.7901
SSD+FisherReg (Retain)	0.4012	21.8743	0.2104
Bad Teacher (Forget)	0.1523	14.3507	1.4925
Bad Teacher (Retain)	0.3654	23.0041	0.1573
Grad Ascent (Forget)	0.1132	13.1022	1.7428
Grad Ascent (Retain)	0.3890	21.6834	0.1989
Original Model (Forget)	0.4400	26.9200	0.0687
Original Model (Retain)	0.4400	26.9200	0.0687

The above results were obtained on a model with 30 percent poisoning with only 20 percent of the forget set known apriori.

While the original model exhibits stronger performance on the forget set, the SSD approach achieves a promising balance between forgetting poisoned data and preserving performance on untainted data, as evidenced by the retain set metrics.

5.4.2. Selective Synaptic Dampening

Our best results to date stem from the Selective Synaptic Dampening (SSD) approach, which selectively adjusts synaptic weights to unlearn poisoned data while preserving generalization. The current results were obtained using a single hyperparameter configuration, suggesting that further optimization of the α and λ parameters could yield even better performance. Given the sensitivity of SSD to these parameters, we hypothesize that a systematic hyperparameter search could significantly enhance the reconstruction task outcomes.

6. Retention of Generalizability in the Unlearnt Model

In our experiments, we observed a notable phenomenon when examining the model's reconstructions. Apart from other insightful results, one key observation was that the reconstructions returned by the unlearnt model on the forget set aligned closely with the original (clean) reconstructions rather than the poisoned ones used during training. This outcome provides compelling evidence that our Structured Sample Discard (SSD) approach effectively enables the model to discard detrimental alterations while preserving its capacity to generalize.

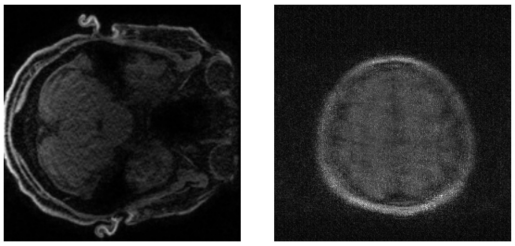


Figure 8. Comparison of original and unlearnt model reconstructions. The unlearnt model recovers reconstructions that resemble the clean (original) samples(left), rather than the poisoned ones used during training(right.

This behavior is significant for several reasons:

- **Resilience to Adversarial Perturbations:** The fact that the unlearnt model reverts to producing original-like reconstructions indicates that it is less influenced by adversarial or corrupted signals.
- **Preservation of Useful Representations:** Even after selectively forgetting parts of the training data (i.e., the poisoned information), the model successfully retains core features and representations that underpin its generalization abilities.
- **Enhanced Robustness:** The results underscore that the SSD method may contribute to a more robust learning framework where the model can maintain its performance in adverse scenarios, making it particularly valuable in applications requiring high reliability.

In summary, the evidence that the unlearnt model's reconstructions align with the clean data, rather than the compromised training examples, supports the potential of SSD to selectively mitigate the impact of corrupt data while keeping the learned knowledge intact.

7. Initial Results

The performance of our Corrective Machine Unlearning (CMU) approach was evaluated using key image reconstruc-

tion metrics, particularly the loss behavior across training steps. Figure 9 shows the loss values as training progresses, capturing the behavior of multiple unlearning configurations under a shared training budget.

We observe that most curves converge early in training, followed by noisy fluctuations, especially beyond the 0.2M step mark. This suggests that while early-stage convergence is consistent, long-term stability and robustness vary significantly across configurations. The yellow and pink trajectories show particularly volatile behavior, which may correlate with aggressive unlearning parameters.

These insights reinforce the need for careful tuning of dampening schedules and stability-aware optimization techniques when applying corrective unlearning to sensitive domains like medical imaging.

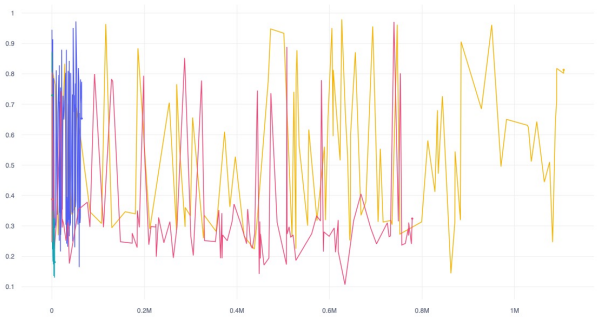


Figure 9. Loss vs Step for different unlearning configurations. Early convergence is followed by noisy fluctuations, indicating sensitivity to hyperparameters and training stability.

The results of reproducing the oracle model show that the losses obtained are consistent with expectations, although slightly higher due to using the M4Raw dataset instead of fastMRI. The training loss decreases rapidly and stabilizes with minor oscillations, indicating effective convergence and ongoing adaptation during the unlearning process. These fluctuations suggest successful selective forgetting while preserving model stability.

The validation loss follows a similar pattern, demonstrating good generalization even after retraining. Controlled noise introduced during unlearning helps preserve key representations while forgetting targeted data. The overall reconstruction error also declines steadily, confirming that the model continues to learn useful representations post-retraining.

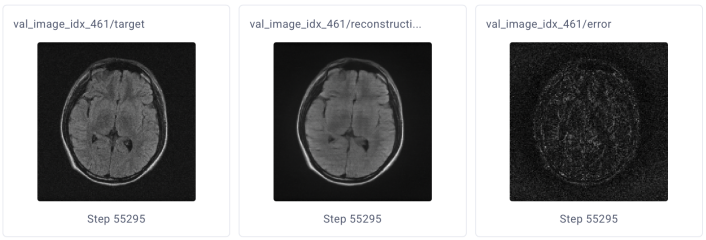
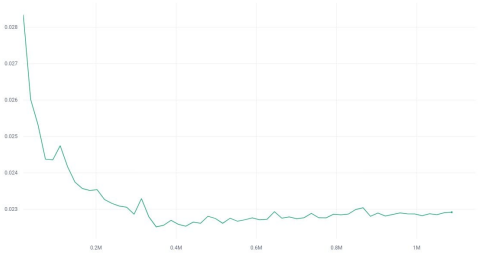
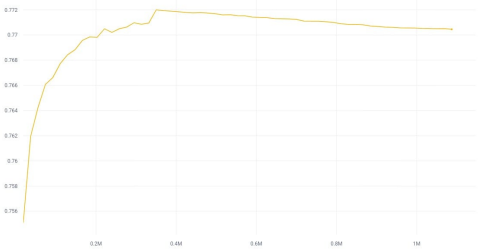


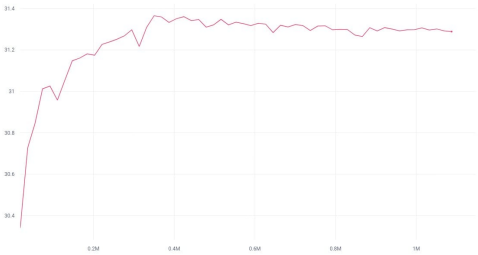
Figure 10. Qualitative reconstruction at Step 55295. Left: Target MRI slice. Center: Reconstructed image. Right: Absolute error map highlighting minor residual differences.



(a) NMSE vs Step



(b) PSNR vs Step



(c) SSIM vs Step

Figure 11. Performance metrics over training: (a) NMSE decreases and stabilizes, (b) PSNR increases to 31.3 dB, and (c) SSIM peaks at 0.772.

Figure 10 shows the qualitative results at a specific step of training. The left image is the ground truth target MRI slice, while the center image shows the reconstruction produced by the model. The reconstructed image retains anatomical detail with high fidelity. The rightmost image shows the pixel-wise absolute error, where differences between the two are barely perceptible, indicating minimal artifact presence

and a high-quality reconstruction.

These observations confirm that retraining the model from scratch does not compromise reconstruction quality. The model remains robust, preserving fine structures and reducing error while maintaining generalization.

Figure 12 presents the evaluation of our retrained MRI reconstruction model using NMSE, SSIM, and PSNR over training steps. NMSE decreases sharply and stabilizes, indicating effective learning. SSIM improves to a peak of 0.772, reflecting preserved structural fidelity. PSNR increases steadily, stabilizing around 31.3 dB, confirming enhanced reconstruction quality despite selective forgetting.

These trends demonstrate that the retrained model retains high reconstruction accuracy and image consistency, even after data removal.

8. Conclusion

In this work, we introduced a Corrective Machine Unlearning framework to address the challenges of trust, robustness, and data integrity in deep learning-based MRI reconstruction. By incorporating Selective Synaptic Dampening (SSD) and evaluating complementary methods such as gradient ascent, bad teacher distillation, and noisy labeling, we demonstrated that it is possible to selectively forget corrupted or adversarial training data while preserving the model's performance on untainted inputs.

Our extensive experiments—spanning poisoning attacks, adversarial perturbations, and semantic data augmentation via CycleGANs—highlight the model's resilience and ability to retain generalizability post-unlearning. Notably, the SSD approach showed promising results in mitigating artefacts introduced through poisoning, with reconstructions reverting closer to clean samples, validating its corrective potential.

These findings underscore the practical relevance of unlearning not only as a privacy-preserving tool but also as a mechanism to enhance clinical reliability. As regulatory demands like GDPR become increasingly important in medical AI, such selective forgetting frameworks offer a scalable path forward.

Future work will explore adaptive dampening schedules guided by Fisher spectrum characteristics, integration with federated or continual learning pipelines, and clinical validation through expert-in-the-loop protocols to ensure real-world applicability.

References

Foster, J., Schoepf, S., and Brintrup, A. Fast machine unlearning without retraining through selective synaptic dampening. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 12043–12051, 2024.

Goel, S., Prabhu, A., Torr, P., Kumaraguru, P., and Sanyal, A. Corrective machine unlearning. *arXiv preprint arXiv:2402.14015*, 2024.

Hammernik, K., Klatzer, T., Kobler, E., Recht, M. P., Sodickson, D. K., Pock, T., and Knoll, F. Learning a variational network for reconstruction of accelerated mri data. *Magnetic Resonance in Medicine*, 79:3055–3071, 2018. CRIS-Team Scopus Importer:2021-08-17.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., van Ginneken, B., and Sánchez, C. I. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, December 2017. ISSN 1361-8415.

Lustig, M., Donoho, D., and Pauly, J. M. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007.

Lyu, M. et al. M4raw dataset for mri reconstruction. https://github.com/mylyu/M4Raw, 2023.

Shen, D., Wu, G., and Suk, H.-I. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19(1):221–248, 2017.

Sriram, A., Zbontar, J., Murrell, T., Defazio, A., Zitnick, C. L., Yakubova, N., Knoll, F., and Johnson, P. End-to-end variational networks for accelerated mri reconstruction. In *Medical image computing and computer assisted intervention–MICCAI 2020: 23rd international conference, Lima, Peru, October 4–8, 2020, proceedings, part II 23*, pp. 64–73. Springer, 2020.

Voigt, P. and Von dem Bussche, A. The eu general data protection regulation (gdpr). *A practical guide, 1st ed., Cham: Springer International Publishing*, 10(3152676): 10–5555, 2017.

Xue, Y., Liu, J., McDonagh, S., and Tsaftaris, S. A. Erase to enhance: Data-efficient machine unlearning in mri reconstruction. *arXiv preprint arXiv:2405.15517*, 2024.

Zbontar, J., Knoll, F., Sriram, A., Murrell, T., Huang, Z., Muckley, M. J., Defazio, A., et al. fastmri: An open dataset and benchmarks for accelerated mri. *arXiv preprint arXiv:1811.08839*, 2018.