Medical Data Under Shadow Attacks via Hybrid Model Inversion

Asfandyar Azhar
Stanford University
Carnegie Mellon University

Paul Thielen
Technische Universiteit Eindhoven

Curtis Langlotz
Stanford University

Abstract

We introduce MEDUSA (Medical Data Under Shadow Attacks), a novel hybrid model inversion framework that leverages gradient-based optimization and TCNNs to reconstruct high-fidelity medical images from model outputs in a gray-box setting. Unlike traditional attacks requiring full model details, MEDUSA uses surrogate shadow models trained on publicly available data, simulating limited-information scenarios often encountered in practice. Our approach shows that even with restricted access, quality image reconstructions are possible, raising serious privacy concerns for patient data. Contributions include demonstrating that a combination of gradient-based methods and TCNNs yields potent reconstructions, even with limited model access, and providing a detailed analysis of how different input configurations impact reconstruction quality. We also evaluate the reconstructions as viable training data, finding that they can approximate real images well enough to use for model training. Finally, we propose robust defensive mechanisms such as output vector truncation, Gaussian noise, and a new *k-NN smearing* technique to tackle privacy risks.

1 INTRODUCTION

The integration of deep learning into medical imaging has provided substantial improvements in diagnostic accuracy and efficiency (Litjens et al., 2017). However, these advances come with significant privacy risks, especially when models are deployed in environments where patient data security is paramount (Shokri et al., 2017). The vulnerability of machine learning models to attacks that can reconstruct training data—commonly known as model inversion (MI) attacks—is increasingly concerning, particularly for sensitive domains like healthcare (Fredrikson et al., 2015b). Such attacks exploit model outputs to recover information about private inputs, which can transform supposedly de-identified medical data into identifiable patient information, thereby compromising patient confidentiality (Zhang et al., 2020).

In this paper, we introduce MEDUSA (Medical Data Under Shadow Attacks), a novel hybrid model inversion framework that leverages both gradient-based optimization and transposed convolutional neural networks (TCNNs) to reconstruct high-fidelity medical images from output vectors. Unlike traditional MI attacks that rely on a white-box setup, where complete model details are accessible, our work is situated within a gray-box setting—a scenario common in practical deployments (He et al., 2019). In the gray-box setting, the model architecture is known, but the weights are not disclosed, and the attacker must rely on surrogate, or shadow, models trained on publicly available data from the same distribution (Shokri et al., 2017). MEDUSA shows that even in this limited-information context, attackers can still extract highly detailed image reconstructions, posing a significant risk to patient data security.

Our main contributions are that:

- We demonstrate, for the first time, that combining gradient-based reconstruction techniques with TCNNs in a gray-box setup can lead to high-fidelity reconstructions of medical images, which underscores the underestimated vulnerability of output vectors in deployed models
- We explore various input configurations, including shadow model outputs, gradient-based reconstructions, and linear activations, to maximize reconstruction quality, providing a thorough analysis of how an attacker can effectively leverage these inputs.

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

- We show that these reconstructed images have utility as synthetic training samples, further elevating the potential damage if such data were to be used maliciously.
- We evaluate standard defense mechanisms, such as output vector rounding, truncation, Gaussian noise, and propose "k-NN smearing", to understand their effectiveness in mitigating the risk of model inversion attacks.

Our results are particularly concerning for healthcare institutions deploying AI models in diagnostic workflows. Hospitals often assume that by securing model weights or restricting access to training data, patient privacy can be preserved (Kaissis et al., 2020; Choi et al., 2017). However, MEDUSA demonstrates that even the output vectors generated during model inference can be powerful enough to reconstruct sensitive images, effectively making private data public. To address these vulnerabilities, we emphasize the importance of robust defensive strategies, particularly in gray-box scenarios where full model protection may be impractical (Wang et al., 2019). By advancing our understanding of how output vectors contribute to potential privacy breaches and evaluating practical defenses, our work aims to enhance the safety and trustworthiness of machine learning models used in medical settings (Abadi et al., 2016).

2 RELATED WORKS

Privacy in Medical Imaging. The intersection of deep learning and medical imaging has brought significant advancements in diagnostic capabilities, but it has also introduced new privacy challenges. (Kaissis et al., 2020) provide a comprehensive overview of the privacy risks in medical imaging AI, highlighting the need for secure and privacy-preserving techniques in healthcare. (Choi et al., 2017) demonstrated the potential for generating synthetic medical records that maintain statistical properties of real data while preserving patient privacy, a concept that could be extended to medical imaging.

Model Inversion Attacks. Model inversion attacks have evolved significantly since their introduction. (Fredrikson et al., 2015b) pioneered the concept, showing how an attacker could reconstruct recognizable facial images from a facial recognition system using only the model's output and some demographic information. (Zhang et al., 2020) advanced this field by introducing a generative model-inversion attack that could produce high-fidelity reconstructions of private training data, demonstrating the increasing sophistication of these attacks. (Zhao et al., 2021) highlights the privacy threats of explanations by showing the ability to reconstruct private image data from model explanations with transposed convolutional neural networks. Within the context of medical imaging, this poses a threat (Wu et al., 2020).

Shadow Models and Gray-box Attacks. The concept of shadow models, crucial to our work, was introduced by (Shokri et al., 2017) for membership inference attacks. This approach has been adapted for various privacy attacks in machine learning. (Salem et al., 2019) further showed how these attacks could be generalized with fewer assumptions about the target model, making them more applicable in real-world scenarios. (He et al., 2019) extended this concept to model inversion attacks in collaborative inference settings, demonstrating how an adversary can reconstruct sensitive inference inputs by exploiting intermediate outputs in distributed deep learning models. This work is particularly relevant to our gray-box setting.

Transposed Convolutional Neural Networks. TCNNs have emerged as a powerful tool for image reconstruction tasks, capable of generating high-quality images from output vectors and outperforming traditional gradient-based methods that often struggle to preserve spatial details (Dumoulin, Visin, 2016; Dosovitskiy, Brox, 2016). The TCNN architecture consists of transposed convolutional layers that upsample the input, batch normalization layers to stabilize and improve training (Ioffe, Szegedy, 2015), and ReLU activation functions to introduce non-linearity for learning complex patterns (Nair, Hinton, 2010). Essentially, transposed convolutional layers perform an approximate inverse operation of standard convolutional layers, enabling TCNNs to generate high-resolution images from low-resolution inputs, which makes them particularly effective for model inversion attacks (Dosovitskiy, Brox, 2016; Yang et al., 2019; Zhao et al., 2021).

Further Enhancing Reconstructions The integration of additional information beyond output vectors has shown promise in improving the quality of image reconstructions in model inversion attacks. (Zhao et al., 2021) demonstrated that incorporating gradient-weighted class activation mappings (Grad-CAMs) (Selvaraju et al., 2019), alongside output vectors significantly enhances reconstruction quality. Grad-CAMs use the gradients flowing into the final convolutional layer to produce a localization map that highlights important regions for predicting a given concept. In their approach, they flattened the 2D Grad-CAMs into 1D arrays and concatenated them with output vectors, leveraging the spatial information captured by Grad-CAMs to preserve structural details during reconstruction. However, this method requires access to gradients with respect to specific model layers, limiting its applicability in gray-box scenarios where such access is restricted.

Our work builds upon these insights but adapts to the constraints of a gray-box setting. Instead of relying on Grad-CAMs, which are not accessible without full model access, we explore alternative methods to enhance reconstruction quality, such as utilizing shadow models and combining gradient-based reconstructions with output vectors. This approach allows us to maintain the benefits of additional input information while working within the limitations of a more realistic attack scenario.

3 METHODS

3.1 The Target & The Shadow

Currently, developers of state-of-the-art AI systems often keep most details of their models private (Bommasani et al., 2023), making the gray-box setting increasingly common. To get around the gray-box model one can utilize a shadow model, $\mathcal{M}_S$, which will mimic the behavior of the target model, $\mathcal{M}_T$, and hence act as a surrogate model which is fully accessible. $\mathcal{M}_S$ learns from and trains on image data, $\mathcal{D}^S_T \subset \mathcal{D}^S$, drawn from the same distribution as the data used to train the target model, $\mathcal{D}^T_T \subset \mathcal{D}^T$. Here, $\mathcal{D}^S$ and $\mathcal{D}^T$ represent the full datasets for $\mathcal{M}_S$ and $\mathcal{M}_T$ respectively. Importantly, $\mathcal{D}^S$ and $\mathcal{D}^T$ are disjoint ($\mathcal{D}^S \cap \mathcal{D}^T = \emptyset$), ensuring no data leakage. This setup allows $\mathcal{M}_S$ to learn relevant features and patterns applicable to $\mathcal{M}_T$'s domain. Additionally, a validation set for $\mathcal{M}_T$, $\mathcal{D}^T_V \subset \mathcal{D}^T$, and a validation set for $\mathcal{M}_S$, $\mathcal{D}^S_V \subset \mathcal{D}^S$, monitor the training process and prevent overfitting, where $\mathcal{D}^T_V \cap \mathcal{D}^T_T = \emptyset$ and $\mathcal{D}^S_V \cap \mathcal{D}^T_T = \emptyset$.

3.2 Datasets

The MNIST dataset (LeCun et al., 2010) is used to quickly refine image reconstruction settings, including the initialization method for gradient reconstructions (Appendix 1A) and the input configurations for final image reconstructions (Section 3.4). In this case, $\mathcal{M}_S$ and $\mathcal{M}_T$ follow a simple CNN architecture.

MeDUSA is evaluated on increasingly complex and various (Appendix 1B) biomedical imaging datasets from the MedMNIST-V2 collection (Yang et al., 2023). We utilize the following 2D datasets within the collection: ChestMNIST, OCTMNIST, OrganAMNIST, PathMNIST, DermaMNIST, and RetinaMNIST (Wang et al., 2017; Kermany et al., 2018; Bilic et al., 2019; Kather et al., 2019; Tschandl et al., 2018; Liu et al., 2022), each of which presents unique medical imaging classification challenges. All images are of size 224 × 224 pixels, and dataset pre-processing, along with training, validation, and test splits, are provided by (Yang et al., 2023). The training and validation sets are further divided in half[1] respectively to provide $\mathcal{M}_T$ with $\mathcal{D}^T_T$ and $\mathcal{D}^T_V$, and $\mathcal{M}_S$ with $\mathcal{D}^S_T$ and $\mathcal{D}^S_V$, such that

$$(\mathcal{D}^T_T \cup \mathcal{D}^T_V) \cup (\mathcal{D}^S_T \cup \mathcal{D}^S_V) = \mathcal{D}^T \cup \mathcal{D}^S = \mathcal{D}$$

, where $\mathcal{D}$ represents a full MedMNIST-V2 dataset. Both $\mathcal{M}_S$ and $\mathcal{M}_T$ employ the ResNet-50 architecture (He et al., 2016), and we denote the ResNet-50 benchmark model from MedMNIST-V2 as $\mathcal{M}_{MM}$, which is trained on the full dataset D for all datasets.

3.3 Gradient Reconstructions

Gradient reconstructions have been used in an multilayer perceptron setting (Fredrikson et al., 2015a), and our work will extend on this by creating reconstructions from a CNN model's outputs. These reconstructions will then be flattened and passed as such into the TCNN, $\mathcal{M}_I$, to obtain synthetic reconstructions that aim to be indistinguishable from the ground-truth image.

The image set to be optimized, $\Phi = \{\phi_1, \phi_2, ..., \phi_n\}$, can be initialized in 4 different ways: randomly, from a specific instance of a class, from the class average, or from a k-NN average. A k-NN average image is calculated by finding the k closest output vectors in $\mathcal{D}^S_T$ and retrieving the corresponding images and averaging them. Class averages and class instances are also calculated from $\mathcal{D}^S_T$. Appendix 1A shows that the best initialization method is k-NN averaging, where k was optimized on each dataset.

Φ will be updated at each time step t to converge to the corresponding target images, Φ^T. Φ will be updated 500 times. y represents the target labels of Φ. Furthermore, let $\mathcal{M}_S$ be the shadow CNN model.

$$\Phi_t = \Phi_{t-1} - \eta \cdot \nabla_\Phi \mathcal{L}_{\text{total}}(\Phi_{t-1})$$

$$\Phi_t = \text{clamp}(\Phi_t, 0, 1) = \max(0,\ \min(\Phi_t, 1))$$

Φ is updated based on the gradient of the total loss, and their values are clamped to ensure they remain within a valid image pixel range. The loss function used is defined as:

$$\mathcal{L}_{\text{total}}(\Phi, y) = L_{\text{CE}}(\Phi, y) + R_{L2}(\phi) + R_{\text{TV}}(\phi) \qquad (1)$$

The combination of cross-entropy loss, L_2 regularization, and total variation regularization results in a comprehensive overall loss function that balances classification accuracy, perturbation magnitude, and visual coherence. Each of the loss function's components

[1] Done randomly over five seeds.

serves as a very distinct and vital function to optimize and enhance the reconstruction of Φ^T.

$$L_{\text{CE}}(\Phi, y) = -\mathcal{M}_{c=1}^{C} y_c \log(\text{softmax}(\mathcal{M}_S(\phi))_c) \tag{2}$$

$$R_{L2}(\phi) = \lambda_{\text{reg-l2}} \mathcal{M}_{i,j} \pi_{i,j}^2 \tag{3}$$

$$R_{\text{TV}}(\phi) = \lambda_{\text{reg-tv}} \mathcal{M}_{i,j \in \mathcal{N}} \|\pi_{i,j} - \pi_{i+1,j}\| + \|\pi_{i,j} - \pi_{i,j+1}\| \tag{4}$$

In Equation 2, C represents the number of classes, and $y_c \in \mathbb{B}$, indicating whether the class label is the correct classification. $\mathcal{M}_S(\phi)$ represents passing image ϕ through the model $\mathcal{M}_S$. Equation 3 and Equation 4 use $\pi_{i,j}$ notation; where (i, j) represent the indices of the pixel π. Lastly $\lambda_{\text{reg-l2}}$ and $\lambda_{\text{reg-tv}}$ are the two hyper-parameters that will be tuned in Section 4.3.

L_{CE} measures the discrepancy between the predicted labels and the true labels. It effectively guides the model towards more accurate predictions by penalizing deviations from the true class labels (Ma et al., 2021). R_{L2} and R_{TV} are both used to impose smoothness in optimization problems, but they do so in fundamentally different ways. R_{L2} penalizes large values in ϕ to ensure smaller perturbations, and it encourages the image to have a globally low dynamic range, making it uniformly smooth without necessarily considering the relationships between adjacent pixels (Yuying et al., 2022). R_{TV} targets the spatial variation between adjacent pixels. It penalizes the sum of the absolute differences between neighboring pixel values. This approach encourages spatial coherence by making the value of a pixel close to that of its neighbors. Overall, it encourages local smoothness while still allowing for sharp transitions, which is crucial in maintaining edge integrity in images (Rudin et al., 1992).

3.4 Unifying Inputs for Inversion Attacks

In the gray box setting, access to the target model is restricted, making it impossible to obtain Grad-CAMs and use them as inputs to $\mathcal{M}_I$, as done in (Zhao et al., 2021). Instead, we incorporate components such as the last linear activation, A^L, the output vector, O_S, and corresponding flattened gradient reconstructions, $\tilde{X}$, from $\mathcal{M}_S$ as additional inputs to achieve higher-fidelity final reconstructions via $\mathcal{M}_I$. The primary baseline uses $\tilde{X}$ alone, while the secondary baseline uses O_S from the CNN as input to $\mathcal{M}_I$. However, other input variations may enhance reconstruction quality.

Each $\mathcal{M}_I$ was trained for 100 epochs using SSIM as the loss function ($\mathcal{L} = 1 - \text{SSIM}$). As shown in Table 1, the input combination of $O_S + \tilde{X}$ provided the best reconstruction performance on the test set. Contrary to expectations, increasing the number of input features did not improve SSIM or reduce pixel-wise similarity (1 - MSE). This may be due to the added complexity introducing noise and redundancy, making it harder for the model to learn effectively. When multiple inputs are combined, $\mathcal{M}_I$ may struggle to identify the most relevant features, leading to diminished performance.

Table 1: Simple ablation on MNIST to select best input configuration for final reconstructions. See Section 4.1 for the evaluation metrics' definitions.

Method	**SSIM**	**1-MSE**	$\mathbf{\Gamma}(\cdot)$
$\tilde{X}$	0.695 ± 0.003	0.993 ± 0.002	0.855 ± 0.002
$\mathcal{M}_I(O_S)$	0.701 ± 0.004	0.993 ± 0.001	0.859 ± 0.001
$\mathcal{M}_I(A^L)$	0.641 ± 0.004	0.980 ± 0.001	0.732 ± 0.005
$\mathcal{M}_I(\tilde{X})$	0.703 ± 0.003	0.995 ± 0.001	0.863 ± 0.002
$\mathcal{M}_I(O_S + A^L)$	0.685 ± 0.003	0.985 ± 0.002	0.788 ± 0.004
$\mathcal{M}_I(O_S + \tilde{X})$	$\mathbf{0.710 \pm 0.001}$	$\mathbf{0.997 \pm 0.001}$	$\mathbf{0.869 \pm 0.002}$
$\mathcal{M}_I(A^L + \tilde{X})$	0.689 ± 0.005	0.987 ± 0.002	0.790 ± 0.006
$\mathcal{M}_I(O_S + A^L + \tilde{X})$	0.697 ± 0.006	0.993 ± 0.005	0.856 ± 0.005

3.5 MEDUSA: Medical Data Under Shadow Attacks

Training consists of two stages: (1) training the CNN models, $\mathcal{M}_T$ and $\mathcal{M}_S$; (2) training the TCNN model, $\mathcal{M}_I$. Models $\mathcal{M}_T$ and $\mathcal{M}_S$ are trained on their respective training and validation datasets, as defined in Section 3.2, to obtain the output vectors O_T and O_S. To train $\mathcal{M}_I$, gradient reconstructions, $\tilde{X}$, are computed for each output vector.

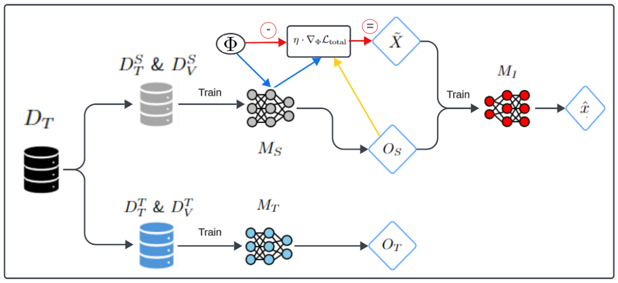

Figure 1: **Training Phase.** This part of the MEDUSA framework shows the training procedure involving $\mathcal{M}_T$, $\mathcal{M}_S$, and $\mathcal{M}_I$ model training.

For gradient reconstructions, an image batch Φ is initialized at $t = 0$ using k-NN initialization (see Section 3.3 and Appendix 1A). The image batch Φ is then updated based on the gradient of $\mathcal{L}_{\text{total}}$ (shown by the red arrows, see Equation 1), which uses the target output vector (yellow arrow) and the current output vector state $\mathcal{M}_S(\Phi)$ (blue arrows). This process is repeated for 500 iterations until the gradient-based reconstructions $\tilde{X}$ are obtained. $\mathcal{M}_I$ is then trained using the combined input of O_S and $\tilde{X}$ to produce the synthetic reconstructions, $\hat{x}$.

The testing procedure begins by passing the test dataset, $\mathcal{D}_F$, through $\mathcal{M}_T$ to obtain O_T. To derive reconstructions from $\mathcal{M}_I$, the gradient-based reconstructions, $\tilde{X}$, must be computed. Since access to $\mathcal{M}_T$

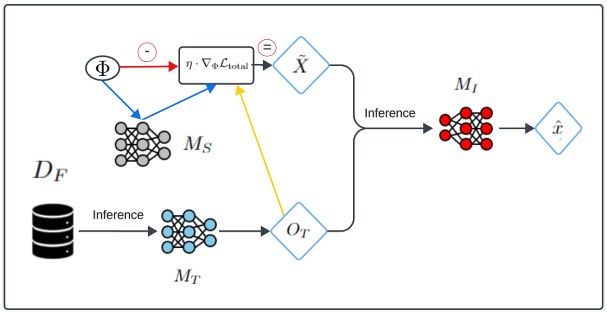

Figure 2: **Inference Phase.** This part of the MEDUSA framework shows the testing procedure involving $\mathcal{M}_T$, $\mathcal{M}_S$, and $\mathcal{M}_I$.

(and $\mathcal{M}_{MM}$) is restricted in our study, the trained model $\mathcal{M}_S$ is used to compute $\tilde{X}$. The key difference in this process is that, instead of using the O_S vectors, the O_T vectors are utilized, as the goal is to reconstruct images from the output vectors of $\mathcal{M}_T$. Finally, the combination of the output vector O_T and gradient-based reconstructions $\tilde{X}$ is fed into the trained $\mathcal{M}_I$ to produce the final reconstructions of $\mathcal{D}_F$, denoted $\hat{x}$.

4 EXPERIMENTS

For technical specifications and training information related to our experiments, see Appendix 2.

4.1 Evaluation Metrics

Pixelwise Similarity. Mean Squared Error (MSE) is commonly used to evaluate how well a reconstructed image matches the original at the pixel level. We normalize the pixel values of both images to the [0, 1] range and compute the MSE. The similarity metric, defined as 1 - MSE, provides a score between 0 and 1, where higher values indicate closer pixelwise correspondence. This measure is simple and size-invariant, focusing purely on numerical pixel differences between the original and reconstructed images.

Structural Similarity Index (SSIM). While MSE captures pixel-level differences, it does not reflect perceptual quality. SSIM is a perception-based metric that evaluates image similarity by comparing luminance, contrast, and structure between the original and reconstructed images (Wang et al., 2004). By focusing on how humans perceive visual quality, SSIM provides a more nuanced evaluation. It ranges from -1 to 1, where 1 indicates perfect similarity, factoring in the perceptual quality of the reconstruction. However, SSIM is highly sensitive to the factors it considers, which are quite variable in medical imaging, making it a potentially flawed metric in our case (Maruyama, 2023; Pambrun, Noumeir, 2015).

Geometric Hybrid Similarity. Hence, we propose a Geometric Hybrid Similarity metric, denoted as $\Gamma(\cdot)$, as a measure for assessing the fidelity of reconstructed images in attack scenarios, particularly when using $\mathcal{M}_I$ for reconstruction. Unlike conventional similarity metrics that focus solely on either pixel-level differences or directional alignment, $\Gamma(\cdot)$ combines the strengths of both euclidean distance and cosine similarity. This allows it to simultaneously account for the magnitude and the directional alignment of the reconstructed and original image embeddings. By incorporating both aspects, $\Gamma(\cdot)$ provides a more nuanced evaluation of how well an attack has preserved not only the overall structural details but also the finer, identity-preserving features of the reconstructed images. This is especially important in settings like ours, where shadow model output vectors and gradient-based reconstructions are used, as attackers may succeed in maintaining key features even if some image details are distorted. Therefore, $\Gamma(\cdot)$ attempts to offer a holistic and conservative measure of attack success in reconstructing sensitive medical images. It is defined as:

$$\Gamma(E(\phi), E(\hat{x})) = \left(e^{-\|E(\phi)-E(\hat{x})\|_2^2}\right)^{\gamma} \cdot \left(\frac{1 + \frac{E(\phi)\cdot E(\hat{x})}{\|E(\phi)\|\|E(\hat{x})\|}}{2}\right)^{1-\gamma}$$

where γ is a weight parameter (defaulted to 0.5), $\Gamma(\cdot) \in [0, 1]$, and $(E(\phi), E(\hat{x}))$ is the tuple that contains the original and reconstructed image embeddings.

To compute the embeddings for $\Gamma(\cdot)$, we rely on the penultimate fully connected layer of the shadow model $\mathcal{M}_S$. This choice aligns with the gray-box setting, where $\mathcal{M}_T$ is not fully accessible, but $\mathcal{M}_S$ serves as an effective surrogate by learning from the same domain distribution. Using $\mathcal{M}_S$'s embeddings thus preserves attack relevance: we measure similarity in the very feature space that an adversary would realistically exploit. Furthermore, because $\mathcal{M}_S$ closely approximates $\mathcal{M}_T$ (as shown in Section 4.2), its feature representations remain indicative of $\mathcal{M}_T$'s behavior. We also evaluated a separate ImageNet-pretrained ResNet-50 for perceptual embeddings, but observed only marginal differences (1–2% on average), likely due to the architectural similarity among the models. Therefore, using $\mathcal{M}_S$ not only maintains consistency with the gray-box threat model but also provides a task-specific and contextually appropriate embedding space for measuring reconstructed image fidelity.

4.2 Forever Surrogates: $\mathcal{M}_\mathbf{T} \approx \mathcal{M}_\mathbf{S}$

Given that $\mathcal{D}_T^T \cap \mathcal{D}_T^S = \emptyset$ and $\frac{1}{2}(\mathcal{K}(\mathcal{D}_T^T, \mathcal{D}_T^S) + \mathcal{K}(\mathcal{D}_T^S, \mathcal{D}_T^T)) < 0.2^2$, allows $\mathcal{M}_S$ to learn a similar decision boundary as $\mathcal{M}_T$. As a result, the AUC and accuracy on $\mathcal{D}_F$ are similar for the two models, demonstrating that $\mathcal{M}_S$ replicates $\mathcal{M}_T$'s behavior (Table 2).

[2] $\mathcal{K}(\cdot)$ is the Kullback-Leibler (KL) Divergence score that considers both pixel intensity histograms and extracted features from a pre-trained encoder (see Appendix 1D).

Table 2: Performance of $\mathcal{M}_T$ and $\mathcal{M}_S$ are approximately the same proving $\mathcal{M}_S$'s surrogacy. This is a trivial concept that serves as a starting point for our experiments.

Dataset	$\mathcal{M}_T$		$\mathcal{M}_S$	
	AUC	ACC	AUC	ACC
ChestMNIST	0.676	0.829	0.673	0.826
OCTMNIST	0.838	0.679	0.831	0.672
OrganAMNIST	0.873	0.829	0.878	0.834
PathMNIST	0.865	0.780	0.860	0.775
DermaMNIST	0.798	0.640	0.800	0.642
RetinaMNIST	0.626	0.447	0.628	0.449
Mean Scores	0.779	0.701	$0.778_{\downarrow 0.01}$	$0.700_{\downarrow 0.01}$

4.3 Grid-Search for Gradient Reconstructions

Gradient reconstructions use two hyperparameters, $\lambda_{\text{reg-l2}}$ and $\lambda_{\text{reg-tv}}$, tuned via random grid search. This approach samples random combinations instead of exhaustively searching all options. The search spaces are defined as $\lambda_{\text{reg-l2}} \in [1 \times 10^{-5}, 3 \times 10^{-2}]$ and $\lambda_{\text{reg-tv}} \in [1 \times 10^{-6}, 3 \times 10^{-2}]$, with values sampled using a log-uniform distribution to ensure better coverage across a wide range of magnitudes.

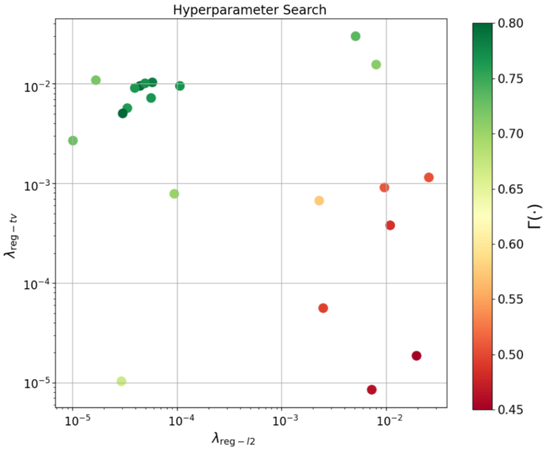

Figure 3: Random grid search for $\lambda_{\text{reg-l2}}$ and $\lambda_{\text{reg-tv}}$.

The tuning process involves two stages. First, a coarse search is conducted, consisting of 20 iterations across the entire range, with each iteration involving 320 gradient reconstructions. The mean $\Gamma(\cdot)$ values are recorded to identify promising regions. Once the coarse search identifies suitable ranges, a fine search follows, focusing on narrower ranges around the best coarse values, denoted as $\lambda_{\text{best-l2}}$ and $\lambda_{\text{best-tv}}$. These refined ranges are defined as $[\frac{1}{2} \cdot \lambda_{\text{best}}, \frac{3}{2} \cdot \lambda_{\text{best}}]$, and the fine search includes 10 iterations, each with 320 reconstructions, again tracking average $\Gamma(\cdot)$ scores. In both stages, gradient reconstructions are derived from O_S vectors, obtained by passing $\mathcal{D}_T^T$ through $\mathcal{M}_S$. The optimal hyperparameters determined through this process were $\lambda_{\text{reg-l2*}} = 4.43 \times 10^{-5}$ and $\lambda_{\text{reg-tv*}} = 9.49 \times 10^{-3}$ as displayed in Figure 3.

4.4 Reconstruction Quality from $\mathcal{M}_I$

As discussed in Section 3.4 and shown in Table 1, the combination of output vectors with their corresponding gradient reconstructions produces the highest quality final synthetic reconstructions. Table 3 displays the reconstruction qualities of the baseline reconstructions $\tilde{X}$ in comparison to $\mathcal{M}_I(O_T + \tilde{X})$ across all datasets.

Table 3: It is empirically shown that the reconstructions $\mathcal{M}_I$ outperform vanilla gradient reconstructions, reflecting significant gains in the image reconstruction quality evident in Figure 4. This underscores $\mathcal{M}_I$'s robustness across all the medical datasets.

Dataset	Method	SSIM	1-MSE	$\Gamma(\cdot)$
ChestMNIST	$\tilde{X}$	0.573	0.891	0.687
	$\mathcal{M}_I(O_T + \tilde{X})$	0.697	0.996	0.805
OCTMNIST	$\tilde{X}$	0.529	0.881	0.649
	$\mathcal{M}_I(O_T + \tilde{X})$	0.652	0.988	0.769
OrganAMNIST	$\tilde{X}$	0.550	0.888	0.667
	$\mathcal{M}_I(O_T + \tilde{X})$	0.673	0.990	0.776
PathMNIST	$\tilde{X}$	0.435	0.834	0.583
	$\mathcal{M}_I(O_T + \tilde{X})$	0.562	0.947	0.699
DermaMNIST	$\tilde{X}$	0.426	0.818	0.573
	$\mathcal{M}_I(O_T + \tilde{X})$	0.552	0.931	0.687
RetinaMNIST	$\tilde{X}$	0.517	0.880	0.611
	$\mathcal{M}_I(O_T + \tilde{X})$	0.639	0.982	0.730
Mean Scores	$\tilde{X}$	0.505	0.865	0.628
	$\mathcal{M}_I(O_T + \tilde{X})$	$0.629_{\uparrow 0.12}$	$0.972_{\uparrow 0.11}$	$0.744_{\uparrow 0.12}$

4.5 Reconstructions as Synthetic Training Samples: Do They Have Any Utility?

We evaluate the performance of three ResNet-50 models: $\mathcal{M}_{MM}$, $\mathcal{M}_{syn}$, and $\mathcal{M}_{syn+}$. The model $\mathcal{M}_{MM}$ is trained on the dataset D, while $\mathcal{M}_{syn}$ is trained on synthetic data, $\mathcal{D}^{syn} = \mathcal{M}_I(\mathcal{D}_T^S)$. Although the AUC and accuracy of $\mathcal{M}_{syn}$ do not reach the benchmark, $\mathcal{M}_{MM}$, they indicate that synthetic reconstructions can be effective as training samples. Notably, $\mathcal{M}_{syn+}$, trained on a combined dataset $\mathcal{D}^{syn+} = \mathcal{D}_T^S \cup \mathcal{D}^{syn}$, achieves performance within 5% of $\mathcal{M}_{MM}$, highlighting the advantages of integrating synthetic data. Furthermore, the improved performance of $\mathcal{M}_{syn+}$ compared to $\mathcal{M}_S$ (Table 2) suggests that using synthetic reconstructions enhances the training process, leading to better generalization. This improvement is consistent across all datasets, implying that synthetic reconstructions contribute valuable information that complements the original data. Also, we report that on average $\frac{1}{2}(\mathcal{K}(\mathcal{D}^T, \mathcal{D}_{syn}) + \mathcal{K}(\mathcal{D}_{syn}, \mathcal{D}^T)) = 0.419$, indicating that the distribution of synthetic images is progres-

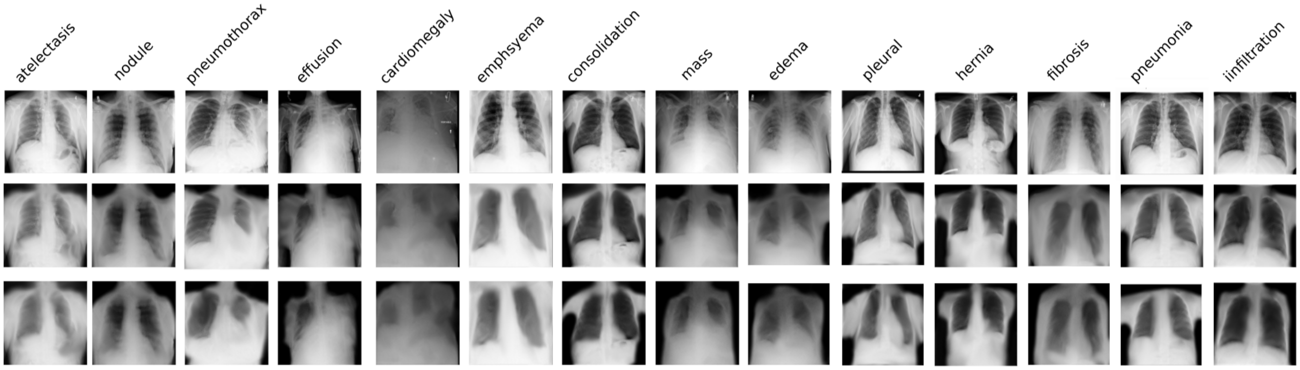

Figure 4: The first row shows the ground truth images, the second row shows their final synthetic reconstructions, $\hat{x} = \mathcal{M}_I(O_T + \tilde{X})$, and the third row shows the corresponding gradient reconstructions, $\tilde{X}$, for ChestMNIST.

sively converging to that of the real training data. This convergence enables an adversary to approximate and nearly replicate the deployed clinical model, underscoring the risk not only of data leakage but also of potential model theft or replication.

Table 4: Performance comparison for $\mathcal{M}_{MM}$, $\mathcal{M}_{syn}$, and $\mathcal{M}_{syn+}$.

	$\mathcal{M}_{\mathbf{MM}}$		$\mathcal{M}_{\mathbf{syn}}$		$\mathcal{M}_{\mathbf{syn+}}$	
Data	**AUC**	**ACC**	**AUC**	**ACC**	**AUC**	**ACC**
ChestMNIST	0.773	0.948	0.404	0.496	0.710	0.885
OCTMNIST	0.958	0.776	0.499	0.403	0.933	0.727
OrganAMNIST	0.998	0.947	0.527	0.500	0.962	0.899
PathMNIST	0.989	0.892	0.516	0.465	0.921	0.813
DermaMNIST	0.912	0.731	0.480	0.385	0.845	0.694
RetinaMNIST	0.716	0.511	0.377	0.269	0.655	0.505
Mean	0.891	0.801	$0.467_{\downarrow 0.42}$	$0.420_{\downarrow 0.38}$	$0.838_{\downarrow 0.05}$	$0.754_{\downarrow 0.05}$

4.6 How Much Should We Hide From The Shadows?

We further explore how varying the amount of training data provided to $\mathcal{M}_{syn+}$ affects the balance between transparency and privacy. While sharing images can enhance transparency, it also poses a risk of exploitation by attackers. The goal is to identify the optimal level of data disclosure that promotes transparency without excessively aiding adversaries. The results in Figure 5 show a clear decline in performance as the training data for $\mathcal{M}_{syn+}$ decreases. At 50% data availability, $\mathcal{M}_{syn+}$ sees AUC and accuracy scores about 5% lower than the benchmark $\mathcal{M}_{MM}$ (Table 4). At 25%, the AUC drops to 0.715 and accuracy to 0.617, reflecting decreases of 6.4% and 8.4%, respectively, compared to $\mathcal{M}_T$. Despite $\mathcal{D}^{syn+} = \mathcal{D}_T^S \cup \mathcal{D}^{syn}$, where $|\mathcal{D}_T^S| = |\mathcal{D}^{syn}|$, the synthetic data still supports $\mathcal{M}_{syn+}$'s performance. However, at just 5% data availability, AUC falls by 63.6% and accuracy by 57.1%, demonstrating that the more data withheld, the greater the reduction in $\mathcal{M}_{syn+}$'s performance.

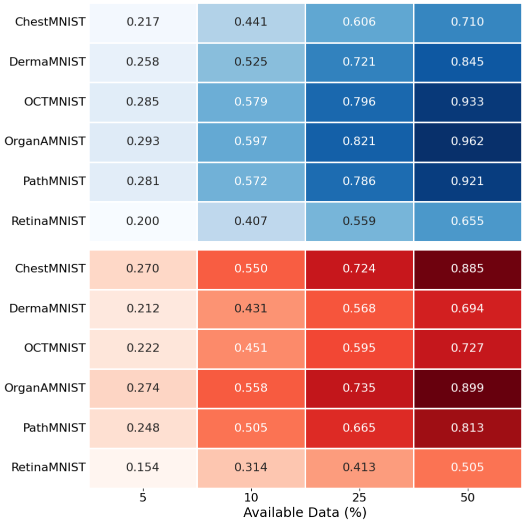

Figure 5: Performance of $\mathcal{M}_{syn+}$ on $\mathcal{D}_F$ given varied $|\mathcal{D}_T^S|$. The blue heatmap displays the AUC while the red heatmap displays the accuracy.

Figure 6 visually illustrates how reduced data availability affects the quality of synthetic reconstructions. As data decreases, image quality declines, which explains the drop in $\mathcal{M}_{syn+}$'s performance. With less data to train on, the synthetic reconstructions provide insufficient information, reducing their value as training images and leading to performance losses. This serves as a defense mechanism among others that we investigate in Section 4.7.

4.7 If We Can't Hide, Then We Must Defend

Output Vector Rounding. Output vector rounding is a potential defense against model inversion attacks by rounding each value in the output vector to the nearest fifth or tenth, limiting the information available to the user. Appendix 1C examines its impact

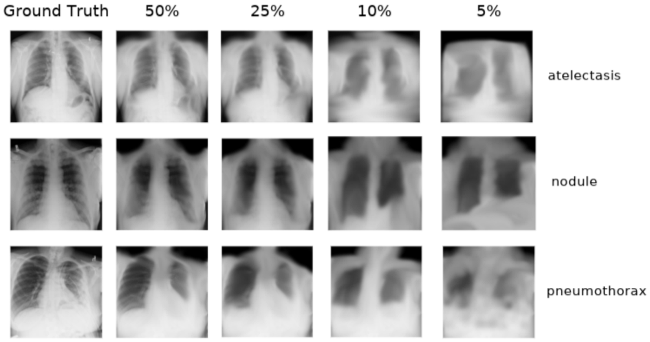

Figure 6: $\mathcal{M}_{syn+}$ reconstructions with varied $|\mathcal{D}_T^S|$.

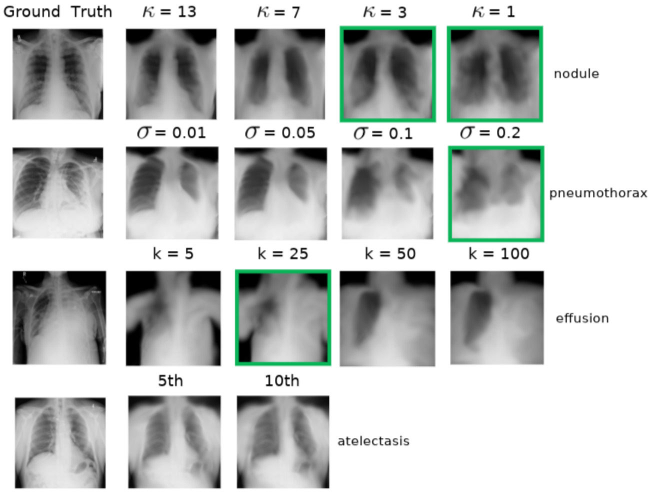

Figure 7: Synthetic reconstructions with varying defense strategies. From top to bottom: truncation, Gaussian noise, k-NN smearing, and output vector rounding. Green indicates the best setting for each method.

on the inverse TCNN model, $\mathcal{M}_I$, with gradient reconstructions based on the rounded output. Despite this rounding, it has no significant effect on reconstruction quality, as $\mathcal{M}_I$ is robust to small input perturbations. The noise introduced by rounding remains within a tolerable range, preserving the overall structure of the reconstructions. Given the minimal impact of rounding, we explore more advanced defensive methods.

Truncation. We examined how truncating output vectors to different values, κ, impacts the quality of $\mathcal{M}_I(O_T + \tilde{X})$. As κ decreases, the loss of class-discriminatory features increases, leading to lower quality reconstructions. This trend is seen in both the metrics (first row in Figure 8) and the visual degradation of images (Figure 7). The optimal κ value can vary by dataset, balancing security and transparency. For ChestMNIST, $\kappa = 3$ could retain key decision-making information while obscuring other details, effectively guarding against model inversion attacks.

Gaussian Noise. We added Gaussian noise to output vectors with varying standard deviations (σ) to assess its impact on synthetic reconstruction quality. As σ increased, the reconstructed image quality degraded, shown by a drop in all metrics. Figure 7 and the second row of Figure 8 illustrate how higher noise levels result in increasingly blurred reconstructions. This demonstrates that adding Gaussian noise is an effective defense against model inversion attacks.

k-NN Smearing. We introduce a novel approach called *k-NN smearing* to construct a mixed output vector by averaging the output of a model with its k-nearest neighbors. The core idea of k-NN smearing is to generate a new output vector by blending the original output with a set of its k-nearest neighbors in a weighted manner, where the weights are selected randomly subject to a unity constraint.

More formally, we have O_0 and its k-nearest neighbors, $O_1, O_2, \ldots, O_k$. We denote all output vectors by: $O_i$for $i = 0, 1, 2, \ldots, k$. Let w_i represent the weights for each vector O_i, where: $\mathcal{M}_{i=0}^{k} w_i = 1, \text{and} w_i \geq 0 \forall i$. The weights can be represented as: $(w_0, w_1, w_2, \ldots, w_k) \sim \text{Dirichlet}(\alpha)$ where $\alpha = (1, 1, \ldots, 1)$ is a vector of ones of length $k + 1$. Alternatively, if we generate u_i from a uniform distribution, we normalize them as follows: $w_i = \frac{u_i}{\mathcal{M}_{j=0}^{k} u_j}$, for $i = 0, 1, \ldots, k$ where $u_i \sim \mathcal{U}(0, 1)$. Finally, the k-NN smeared output vector can be either $O_{\text{smeared}} = \mathcal{M}_{i=0}^{k} \left(\frac{u_i}{\mathcal{M}_{j=0}^{k} u_j} \right) O_i, u_i \sim \mathcal{U}(0, 1)$ or $O_{\text{smeared}} = \mathcal{M}_{i=0}^{k} w_i O_i, w_i \sim \text{Dirichlet}(\alpha)$[3]. This approach had the most efficacy as a defensive strategy as the mean scores for SSIM, 1 - MSE, and $\Gamma(\cdot)$ were $(0.236, 0.364, 0.279)$ at $k^* = 25$ (Figure 9).

5 CONCLUSION

This paper presented MeDUSA, a hybrid model inversion framework that exploits vulnerabilities in medical image models deployed in gray-box settings. By using gradient-based optimization and TCNNs, MeDUSA reconstructs high-quality medical images from model outputs, which can expose confidential patient data and risk proprietary model theft. Our experiments demonstrated the effectiveness of defense mechanisms such as truncation, Gaussian noise, and the k-NN smearing technique, emphasizing the need for layered defenses to safeguard patient data in clinical AI applications.

Limitations & Future Work. The effectiveness of the defenses may vary depending on the dataset and model architecture, requiring further exploration to assess their generalizability across clinical settings. Additionally, the use of surrogate shadow models as-

[3] Both weight generation methods performed similarly.

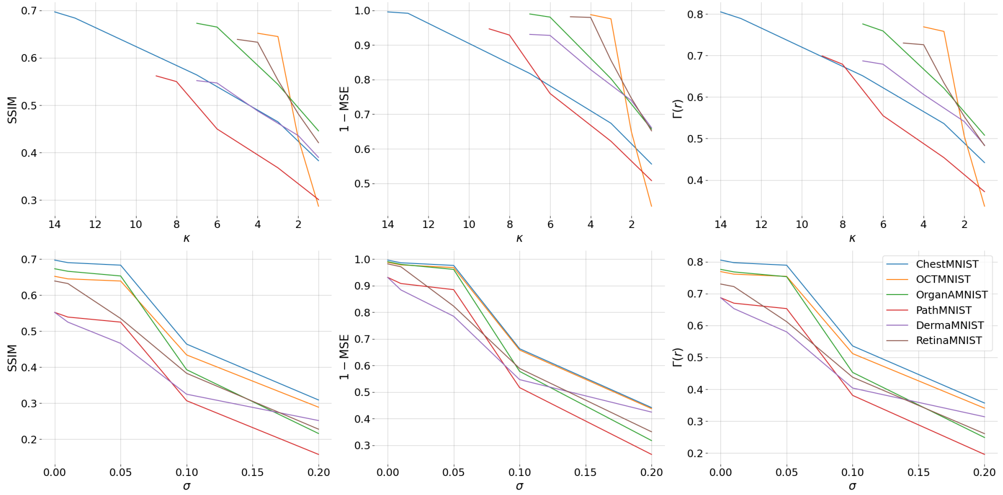

Figure 8: O_T with high truncation (top) or sufficient Gaussian noise (bottom) lead to bad reconstructions.

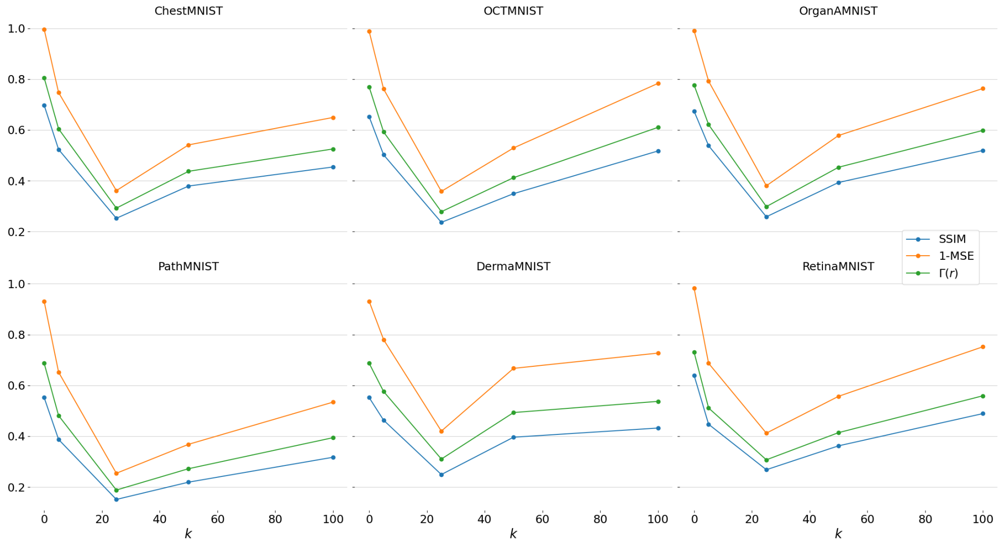

Figure 9: k-NN smearing has an optimal $k^* = 25$. At higher k the reconstruction quality goes back up as Φ^T starts to approximate a set of class average samples.

sumes similar distributions between target and shadow data, which may not always be practical in real-world scenarios.

Training the target and shadow models on data from the same distribution is ideal for producing a shadow model that closely approximates the target. However, domain shifts are common in medical imaging (e.g., data from different centers or devices), and in such cases the performance of both gradient-based and TCNN-based reconstructions is likely to degrade. One avenue to address this limitation is domain adaptation: for instance, adversarial domain adaptation techniques can help align feature spaces between the shadow and target domains, even under significant distribution shifts. Similarly, transfer learning from related datasets or using synthetic data that approximates the target domain can improve the shadow model's fidelity. Addressing these issues is critical for developing stronger defenses and that systematically studying their effectiveness is an important direction for future work.

Ethical Considerations. Our findings highlight significant ethical concerns. Reconstructing sensitive medical images poses serious privacy risks, underscoring the need for stricter guidelines in healthcare model deployment. We advocate for continuous assessment of AI models against inversion attacks and the integration of privacy-preserving methods at every stage of AI development. Protecting patient confidentiality must remain a priority as AI use in healthcare grows.

References

Abadi Martin, Chu Andy, Goodfellow Ian, McMahan H Brendan, Mironov Ilya, Talwar Kunal, Zhang Li. Deep learning with differential privacy // Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. 2016. 308–318.

Bilic Patrick, Christ Patrick Ferdinand, others . The Liver Tumor Segmentation Benchmark (LiTS) // CoRR. 2019. abs/1901.04056.

Bommasani Rishi, Klyman Kevin, Longpre Shayne, Kapoor Sayash, Maslej Nestor, Xiong Betty, Zhang Daniel, Liang Percy. The Foundation Model Transparency Index. 2023.

Choi Edward, Biswal Siddharth, Malin Bradley, Duke Jon, Stewart Walter F, Sun Jimeng. Generating multi-label discrete patient records using generative adversarial networks // Machine learning for healthcare conference. 2017. 286–305.

Dosovitskiy Alexey, Brox Thomas. Inverting Visual Representations with Convolutional Networks. 2016.

Dumoulin Vincent, Visin Francesco. A guide to convolution arithmetic for deep learning // arXiv preprint arXiv:1603.07285. 2016.

Fredrikson Matt, Jha Somesh, Ristenpart Thomas. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures // Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. New York, NY, USA: Association for Computing Machinery, 2015a. 1322–1333. (CCS '15).

Fredrikson Matt, Jha Somesh, Ristenpart Thomas. Model inversion attacks that exploit confidence information and basic countermeasures // Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. 2015b. 1322–1333.

He Kaiming, Zhang Xiangyu, Ren Shaoqing, Sun Jian. Deep residual learning for image recognition // Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. 770–778.

He Zecheng, Zhang Tianwei, Lee Ruby B. Model inversion attacks against collaborative inference // Proceedings of the 35th Annual Computer Security Applications Conference. 2019. 148–162.

Ioffe Sergey, Szegedy Christian. Batch normalization: Accelerating deep network training by reducing internal covariate shift // International conference on machine learning. 2015. 448–456.

Kaissis Georgios A, Makowski Marcus R, Rückert Daniel, Braren Rickmer F. Secure, privacy-preserving and federated machine learning in medical imaging // Nature Machine Intelligence. 2020. 2, 6. 305–311.

Kather Jakob Nikolas, Krisam Johannes, others . Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study // PLOS Medicine. 01 2019. 16, 1. 1–22.

Kermany Daniel S., Goldbaum Michael, others . Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning // Cell. 2018. 172, 5. 1122 – 1131.e9.

LeCun Yann, Cortes Corinna, Burges CJ. MNIST handwritten digit database // ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist. 2010. 2.

Litjens Geert, Kooi Thijs, Bejnordi Babak Ehteshami, Setio Arnaud Arindra Adiyoso, Ciompi Francesco, Ghafoorian Mohsen, Van Der Laak Jeroen AWM, Van Ginneken Bram, Sánchez Clara I. A survey on deep learning in medical image analysis // Medical image analysis. 2017. 42. 60–88.

Liu Ruhan, Wang Xiangning, Wu Qiang, Dai Ling, Fang Xi, Yan Tao, Son Jaemin, Tang Shiqi, Li Jiang, Gao Zijian, Galdran Adrian, Poorneshwaran J.M., Liu Hao, Wang Jie, Chen Yerui, Porwal Prasanna, Wei Tan Gavin Siew, Yang Xiaokang, Dai Chao, Song Haitao, Chen Mingang, Li Huating, Jia Weiping, Shen Dinggang, Sheng Bin, Zhang Ping. DeepDRiD: Diabetic Retinopathy—Grading and Image Quality Estimation Challenge // Patterns. 2022. 100512.

Ma Jun, Chen Jianan, Ng Matthew, Huang Rui, Li Yu, Li Chen, Yang Xiaoping, Martel Anne L. Loss odyssey in medical image segmentation // Medical Image Analysis. 2021. 71. 102035.

Maruyama S. Properties of the SSIM metric in medical image assessment: correspondence between measurements and the spatial frequency spectrum // Physical and Engineering Sciences in Medicine. 2023. 46. 1131–1141.

Nair Vinod, Hinton Geoffrey E. Rectified linear units improve restricted boltzmann machines // Proceedings of the 27th international conference on machine learning (ICML-10). 2010. 807–814.

Pambrun Jean-François, Noumeir Rita. Limitations of the SSIM quality metric in the context of diagnostic imaging // 2015 IEEE International Conference on Image Processing (ICIP). 2015. 2960–2963.

Rudin Leonid I., Osher Stanley, Fatemi Emad. Nonlinear total variation based noise removal algorithms // Physica D: Nonlinear Phenomena. 1992. 60, 1. 259–268.

Salem Ahmed, Zhang Yang, Humbert Mathias, Berrang Pascal, Fritz Mario, Backes Michael. ML-leaks:

Model and data independent membership inference attacks and defenses on machine learning models // Network and Distributed Systems Security Symposium. 2019.

Selvaraju Ramprasaath R., Cogswell Michael, Das Abhishek, Vedantam Ramakrishna, Parikh Devi, Batra Dhruv. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization // International Journal of Computer Vision. X 2019. 128, 2. 336–359.

Shokri Reza, Stronati Marco, Song Congzheng, Shmatikov Vitaly. Membership inference attacks against machine learning models // 2017 IEEE Symposium on Security and Privacy (SP). 2017. 3–18.

Tschandl Philipp, Rosendahl Cliff, Kittler Harald. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions // Scientific data. 2018. 180161.

Wang Qiang, Du Mu, Chen Xin, Chen Yanjiao, Zhou Pan, Chen Xiaofei, Huang Xiaojiang. Privacy-preserving deep learning via additively homomorphic encryption // IEEE Transactions on Information Forensics and Security. 14, 11. 2019. 2991–3006.

Wang Xiaosong, Peng Yifan, others . ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases // CVPR. 2017. 3462–3471.

Wang Zhou, Bovik A.C., Sheikh H.R., Simoncelli E.P. Image quality assessment: from error visibility to structural similarity // IEEE Transactions on Image Processing. 2004. 13, 4. 600–612.

Wu Maoqiang, Zhang Xinyue, Ding Jiahao, Nguyen Hien, Yu Rong, Pan Miao, Wong Stephen T. Evaluation of Inference Attack Models for Deep Learning on Medical Data. 2020.

Yang Jiancheng, Shi Rui, Wei Donglai, Liu Zequan, Zhao Lin, Ke Bilian, Pfister Hanspeter, Ni Bingbing. MedMNIST v2-A large-scale lightweight benchmark for 2D and 3D biomedical image classification // Scientific Data. 2023. 10, 1. 41.

Yang Ziqi, Shao Jiyi, Yao Bowen, Cai Yiran, Zhang Wen, Sedoc João. Neural network inversion in adversarial setting via background knowledge alignment // Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. 2019. 225–240.

Yuying Gou, Gui-cang Zhang, Genlian Han. L2 Regularization Model with Removal of Gaussian Noise // Journal of mathematics and informatics. 2022. 23. 41–50.

Zhang Yuheng, Jia Ruoxi, Pei Hengzhi, Wang Wenxiao, Li Bo, Song Dawn. The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks. 2020.

Zhao Xuejun, Zhang Wencan, Xiao Xiaokui, Lim Brian. Exploiting Explanations for Model Inversion Attacks // Proceedings of the IEEE International Conference on Computer Vision. 4 2021. 662–672.

Appendix 1: Additional Experiments

A. Initialization Method Selection

Table 1: Simple ablation on MNIST to select best initialization method for gradient reconstructions. The optimal k-values for each dataset in MedMNIST-V2 are $[k^*_{chest}, k^*_{oct}, k^*_{organa}, k^*_{path}, k^*_{derma}, k^*_{retina}] = [13, 6, 7, 22, 5, 10]$.

Method	**SSIM**	**1-MSE**	$\mathbf{\Gamma}(\cdot)$
Random	0.248 ± 0.014	0.931 ± 0.008	0.795 ± 0.011
Specific Instance	0.302 ± 0.079	0.940 ± 0.004	0.805 ± 0.006
Class Average	0.608 ± 0.133	0.965 ± 0.007	0.828 ± 0.012
k-NN Average	$\mathbf{0.695 \pm 0.098}$	$\mathbf{0.969 \pm 0.005}$	$\mathbf{0.835 \pm 0.008}$

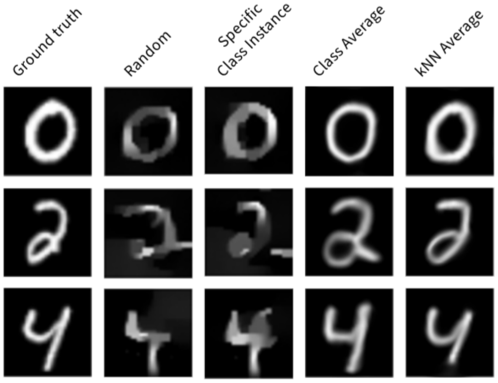

Figure 1: Visual comparison of gradient-based reconstructions using different initialization methods on MNIST.

B. Measuring Dataset Complexity

Table 2: Equal weight is given to each *normalized* complexity metric to compute the final complexity score.

Dataset	**Entropy**	**Edge Density**	**Spatial Frequency**	**Fractal Dimension**	**Color Variance**	**Complexity**
MNIST	0.35	0.20	0.30	0.25	0.15	0.250 **(7)**
ChestMNIST	0.40	0.25	0.35	0.30	0.20	0.300 **(6)**
OCTMNIST	0.45	0.40	0.50	0.38	0.25	0.396 **(4)**
OrganAMNIST	0.42	0.35	0.45	0.33	0.22	0.354 **(5)**
PathMNIST	0.65	0.60	0.75	0.45	0.65	0.620 **(2)**
DermaMNIST	0.60	0.55	0.70	0.43	0.70	0.596 **(3)**
RetinaMNIST	0.68	0.65	0.80	0.48	0.85	0.692 **(1)**
Mean ± St. Dev	0.51 ± 0.13	0.43 ± 0.18	0.55 ± 0.20	0.37 ± 0.08	0.43 ± 0.27	0.458 ± 0.162

We quantified and compared the complexity of various medical imaging datasets from the MedMNIST collection using a set of standard image complexity metrics. These include entropy, edge density, spatial frequency, fractal dimension, and color variance defined as follows:

Entropy. Measures the randomness in pixel intensity values, indicating the amount of information or detail in an image: $H = -\sum_{i=0}^{L-1} P(i) \log_2 P(i)$, where L is the number of possible intensity levels (e.g., 256 for 8-bit images), and $P(i)$ is the probability of pixel intensity i.

Edge Density. Represents the proportion of edge pixels within the image, reflecting structural complexity: $\frac{|\text{Edge Pixels}|}{|\text{ Pixels}|}$ (we used the Sobel edge detection algorithm).

Spatial Frequency. This metric captures the rate of intensity changes, indicating the level of texture and fine-grained details: $\sqrt{F_r^2 + F_c^2}$, where F_r is the row frequency and F_c is the column frequency derived from the image's 2D Fast Fourier Transform (FFT).

Fractal Dimension. This metric quantifies the complexity of self-similar patterns, useful for analyzing natural textures: $D = \lim_{\epsilon \to 0} \frac{\log N(\epsilon)}{\log(1/\epsilon)}$, where $N(\epsilon)$ is the number of boxes of size ϵ needed to cover the structure in the image (box-counting method).

Color Variance. This measures the variability in color or brightness across the image: $\text{Var} = \frac{1}{3}(\sigma_R^2 + \sigma_G^2 + \sigma_B^2)$, where $\sigma_R^2, \sigma_G^2, \sigma_B^2$ are the variances of the RGB channels, respectively. For grayscale images, it is computed as the brightness variance $\sigma_{\text{Gray}}^2 = \frac{1}{N}\sum_{i=1}^{N}(I_i - \frac{1}{N}\sum_{i=1}^{N} I_i)^2$, where I_i is the intensity value of the i-th pixel.

By assessing these factors, we developed a weighted complexity score to evaluate and rank the datasets based on their visual complexity. This analysis is crucial as more complex datasets, with higher entropy and texture variance, pose a greater challenge for both tasks (attacking and defending). In Table 3, the most notable finding is a strong negative linear correlation between dataset complexity and the $\Gamma(\cdot)$ metric for the $M_I(O_T + \tilde{X})$ method, which was statistically significant ($p = 0.024$). This suggests that higher dataset complexity may lead to lower $\Gamma(\cdot)$ values, indicating a potential decrease in reconstruction quality. However, no significant correlations were found between dataset complexity and the classification performance for the M_{syn} model, despite moderate negative trends. Additionally, there was no significant correlation between dataset complexity and the best $\Gamma(\cdot)$ values from the k-NN smearing defensive method.

Table 3: Correlation analysis between dataset complexity and various metrics for attacking and defending.

Complexity vs.	**Test**	**Coeff.**	***p*-value**
$\Gamma(\cdot)$ for $M_I(O_T + \tilde{X})$	Pearson	$r = -0.872$	**0.024**
$\Gamma(\cdot)$ for $M_I(O_T + \tilde{X})$	Spearman	$\rho = -0.771$	0.072
M_{syn}'s accuracy	Pearson	$r = -0.730$	0.100
M_{syn}'s accuracy	Spearman	$\rho = -0.771$	0.072
$\Gamma(\cdot)$ for k^*-NN smearing	Pearson	$r = -0.210$	0.690
$\Gamma(\cdot)$ for k^*-NN smearing	Spearman	$\rho = 0.143$	0.787

C. Output Vector Rounding

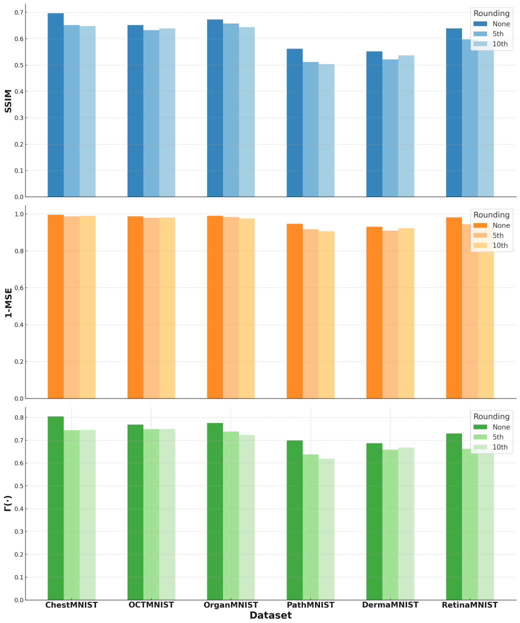

Figure 2: Bar plots showing no significant drop in reconstruction performance across the three evaluation metrics when using output vector rounding. On average, (SSIM, 1-MSE, $\Gamma(\cdot)$) dropped by (0.031, 0.016, 0.045).

D. Image Features Distribution Analysis

To evaluate how closely the synthetic data generated by the `MEDUSA` framework mimics the real datasets, we utilize a symmetric Kullback-Leibler (KL) Divergence score, denoted $\mathcal{K}(\cdot)$. This score provides a balanced measure of the divergence between two probability distributions, enabling us to assess the similarity between the real and synthetic data distributions. We incorporate both low-level pixel intensity histograms and high-level feature embeddings extracted from an ImageNet initialized ResNet-50 encoder, providing a comprehensive assessment of the fidelity of the reconstructed images.

Derivation. Given two datasets, the real dataset D^T and the corresponding synthetic dataset D_{syn}, we denote their probability distributions as P and Q respectively. We extract two sets of distributions:

- *Pixel Intensity Histograms*: P_π and Q_π, representing the normalized histograms of pixel intensities.
- *Feature Distributions*: P_f and Q_f, obtained by passing images through the pre-trained encoder and generating histograms over the extracted feature embeddings.

The KL Divergence between two distributions P and Q is defined as: $\mathrm{K}(P \parallel Q) = \sum_i P(i) \log \left(\frac{P(i)}{Q(i)} \right)$, where i indexes over the bins or features. Since KL Divergence is not inherently symmetric, we calculate the score in both directions and average them to obtain a symmetric KL Divergence. However, before doing so, the pixel and feature distribution scores need to be combined:

$$\mathcal{K}(D^T, D_{syn}) = \alpha \cdot \mathrm{K}(P_\pi \parallel Q_\pi) + (1 - \alpha) \cdot \mathrm{K}(P_f \parallel Q_f)$$

$$\mathcal{K}(D_{syn}, D^T) = \alpha \cdot \mathrm{K}(Q_\pi \parallel P_\pi) + (1 - \alpha) \cdot \mathrm{K}(Q_f \parallel P_f)$$

where $0 \leq \alpha \leq 1$ controls the emphasis on low-level pixel information against high-level semantic features. In our experiments, we set $\alpha = 0.5$ to give equal weighting to both components. The symmetric score follows as:

$$\mathcal{K}(D^T \simeq D_{syn}) = \frac{1}{2}(\mathcal{K}(D^T, D_{syn}) + \mathcal{K}(D_{syn}, D^T))$$

Interpretation. As $\mathcal{K}(\cdot) \to 0$, the synthetic data is closely aligned with the real data, suggesting that the model inversion process was effective in recreating the original images. Conversely, a higher score reflects a greater divergence, indicating that the synthetic data fails to accurately replicate the real dataset's characteristics. This metric can therefore be crucial for evaluating the privacy risks posed by potential data leakage through hybrid model inversion attacks. In Table 4, on average, we see that the synthetic data does relatively well at approximating the image feature distributions of the real dataset.

Table 4: Symmetric KLD scores between real and synthetic data across all datasets.

Dataset	$\mathcal{K}(D^T \simeq D_{syn})$
ChestMNIST	0.418
OCTMNIST	0.372
OrganAMNIST	0.390
PathMNIST	0.465
DermaMNIST	0.452
RetinaMNIST	0.417
Mean	$\mathbf{0.419 \pm 0.032}$

Appendix 2: Technical Specifications

All experiments were performed using NVIDIA A100-80GB and V100-32GB configurations. We employ the ResNet-50 architecture as the backbone for M_S and M_T, with M_I using a modified transposed ResNet architecture to handle inverse gradient reconstructions. The input tensor shape for each model is configured to be $3 \times 224 \times 224$, ensuring compatibility with any pre-trained weights and consistent normalization.

Table 5: Summary of Training Details

General Specifications	
GPU Configuration	NVIDIA A100-80GB (SXM) / V100-32GB
Model Backbone	ResNet-50 (M_S and M_T) Transposed ResNet (M_I)
Input Tensor Shape	$3 \times 224 \times 224$
Normalization	$\mu = [0.485, 0.456, 0.406],\ \sigma = [0.229, 0.224, 0.225]$
Training Configuration for M_S and M_T	
Loss Function	Cross-Entropy w/ L_2 and total variation regularizations
Max Epochs	100
Batch Size	128
Optimizer	Adam
Max Learning Rate	0.001
Learning Rate Scheduler	Step LR (decay by 0.1 at 50, 75 epochs)
Training Configuration for M_I	
Loss Function	1 - SSIM
Max Epochs	100
Warm-Up Epochs	10
Batch Size	128
Optimizer	AdamW
Max Learning Rate	0.008
Learning Rate Scheduler	Cosine Annealing
Training Configuration for M_{syn+}	
Loss Function	Cross-Entropy
Max Epochs	100 (w/ early stopping)
Batch Size	128
Optimizer	Adam
Max Learning Rate	0.001
Learning Rate Scheduler	Step LR (decay by 0.1 at 50, 75 epochs)

Appendix 3: Auxiliary Images

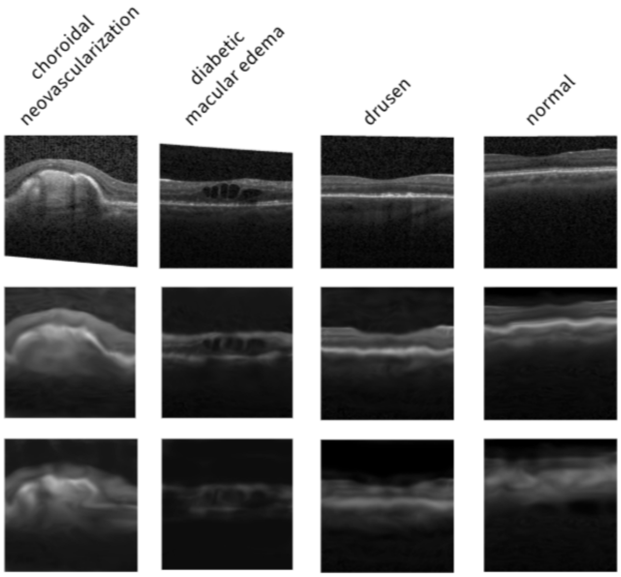

Figure 3: The first row shows the ground truth images, the second row shows their final synthetic reconstructions, $\hat{x} = M_I(O_T + \tilde{X})$, and the third row shows the corresponding gradient reconstructions, $\tilde{X}$, for OCTMNIST.

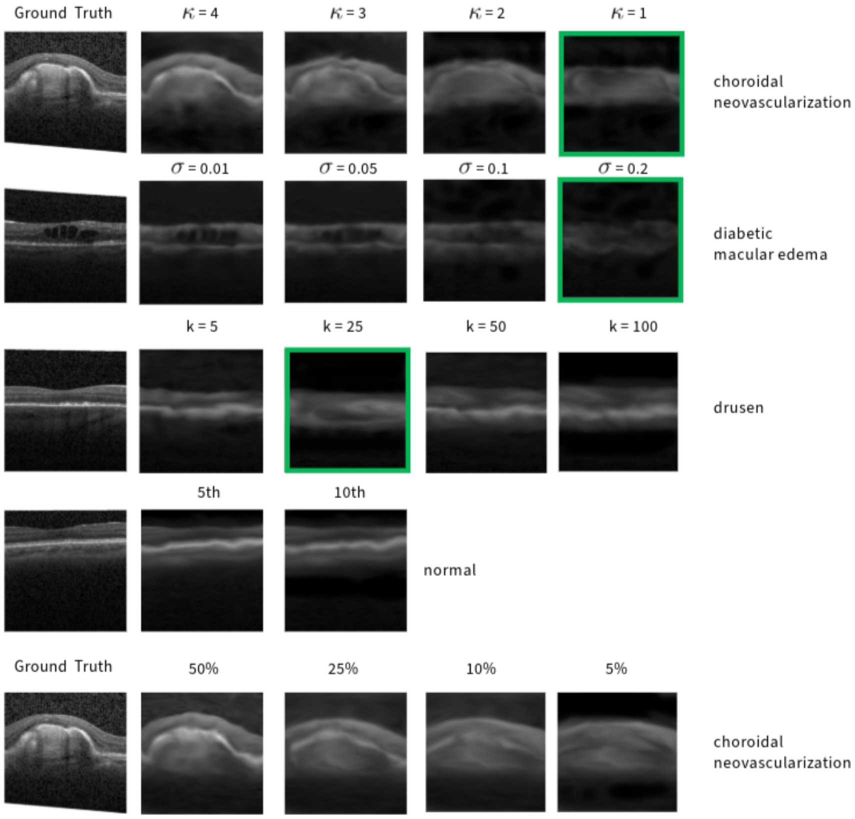

Figure 4: Synthetic reconstructions with varying defense strategies. From top to bottom: truncation, Gaussian noise, k-NN smearing, and output vector rounding. Green indicates the best setting for each method. Last row are M_{syn+} reconstructions with varied $|D_T^S|$.

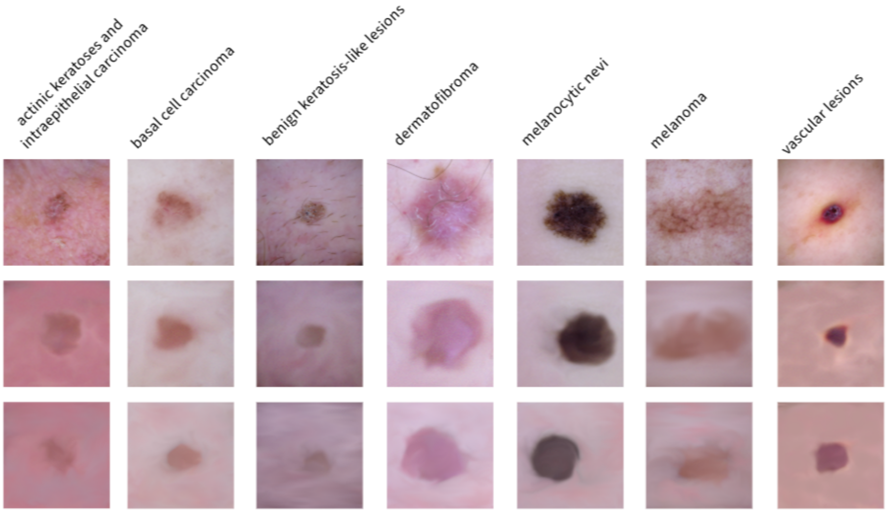
actinic keratoses and intraepithelial carcinoma
basal cell carcinoma
benign keratosis-like lesions
dermatofibroma
melanocytic nevi
melanoma
vascular lesions

Figure 5: The first row shows the ground truth images, the second row shows their final synthetic reconstructions, $\hat{x} = M_I(O_T + \tilde{X})$, and the third row shows the corresponding gradient reconstructions, $\tilde{X}$, for DermaMNIST.

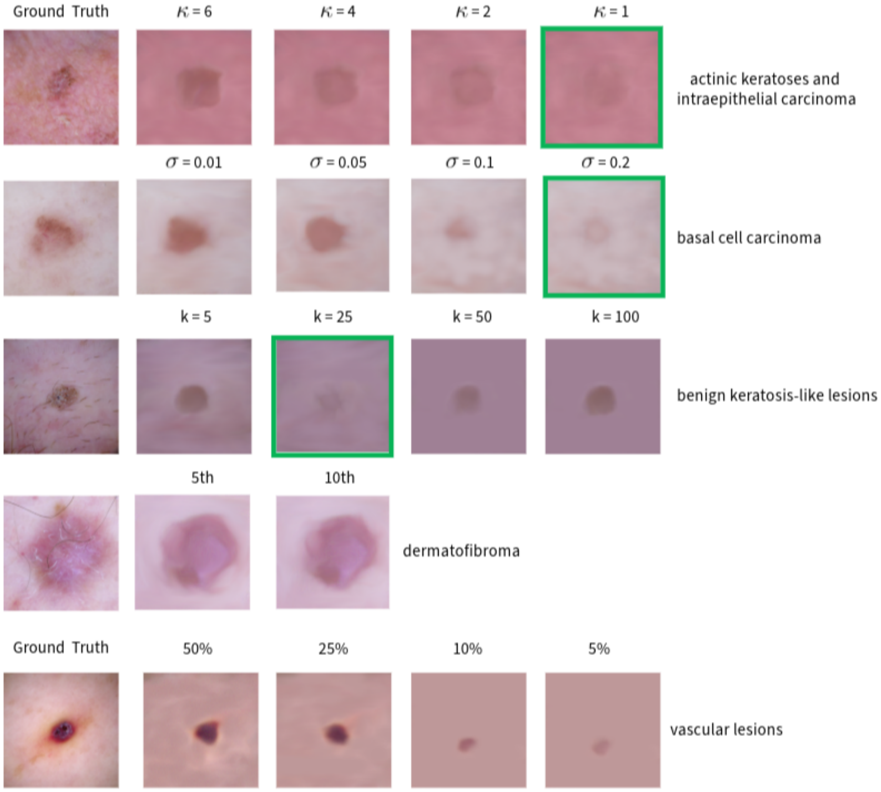

Figure 6: Synthetic reconstructions with varying defense strategies. From top to bottom: truncation, Gaussian noise, k-NN smearing, and output vector rounding. Green indicates the best setting for each method. Last row are M_{syn+} reconstructions with varied $|D_T^S|$.

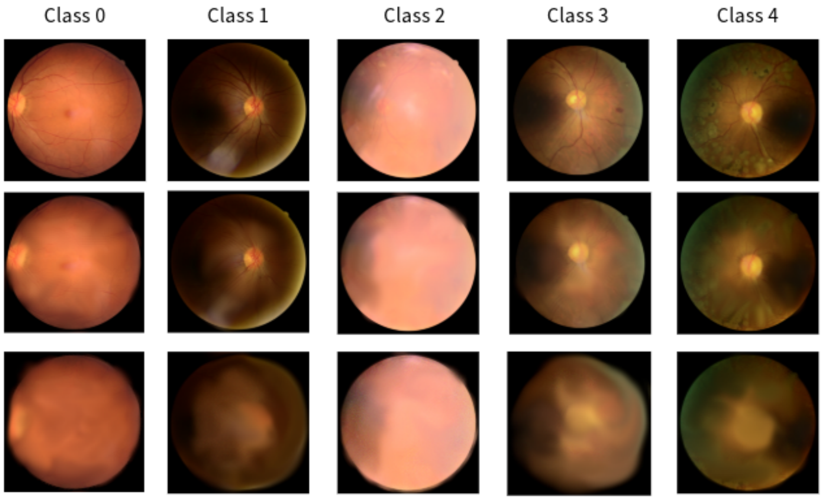

Figure 7: The first row shows the ground truth images, the second row shows their final synthetic reconstructions, $\hat{x} = M_I(O_T + \tilde{X})$, and the third row shows the corresponding gradient reconstructions, $\tilde{X}$, for RetinaMNIST.

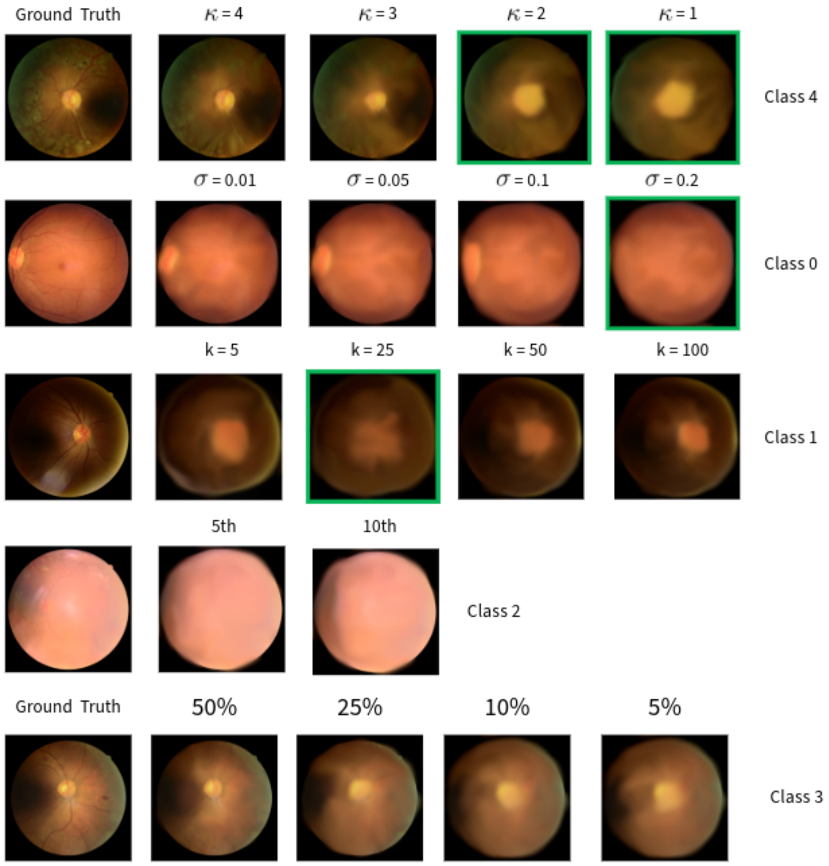

Figure 8: Synthetic reconstructions with varying defense strategies. From top to bottom: truncation, Gaussian noise, k-NN smearing, and output vector rounding. Green indicates the best setting for each method. Last row are M_{syn+} reconstructions with varied $|D_T^S|$.