Re-identification Risk of Medical Imaging-Based Deep Learning Models

Sadie Lee

Undergraduate Capstone Project
Bachelor of Arts in Cognitive Systems

Cognitive Systems Program
University of British Columbia
Vancouver, Canada
August 2025

Abstract

This report studies the risk of patient re-identification from the parameters of models trained on de-identified radiology images. Re-identification is examined through the lens of image reconstruction, from which identifiable information is extracted from reconstructed images. Previous reconstruction attacks have assumed access to additional information such as gradients or held-out training data, and generally do not consider re-identification. We present a two-stage reconstruction approach that requires only a trained model and its parameters, and additionally predict demographics from reconstructed images. Rather than reconstructing private training data by optimizing the images themselves, as prior methods do, we optimize a generator to produce images that lie along the target model's training data manifold, in which internal statistics and parameters from a frozen checkpoint are used as proxies for its true structure. Moreover, we demonstrate that model architecture and presence of memorization significantly contribute to re-identification, where an overfit vision transformer (ViT) outperforms overfit and non-overfit convolutional neural networks (CNNs) and a non-overfit ViT. Cases in which patient re-identification would be possible from reconstructed images and their predicted demographics, as well as potential mitigation strategies, are also discussed. Code is available at `https://github.com/leesadie/Re-id_Risk_Imaging`.

Contents

Terminology

De-identification The removal of individually identifiable information that may allow for re-identification, as defined by standards such as NIST [1] and DICOM PS3.15 [2], and the HIPAA Privacy Rule [3].

The terms 'de-identification', 'anonymization', and 'pseudonymization' are often used synonymously. 'Anonymization', specifically, is sometimes used to indicate complete de-identification with zero residual risk of re-identification, and is often considered separable from 'de-identification'. However, complete de-identification is not always possible [4]. This report will use the term 'de-identification' exclusively for clarity.

Re-identification The extent to which an image or its features can be traced back to a real patient following de-identification, which is consistent with definitions used in the literature [5–8], regulation [9], and industry [10].

Direct identifier Information that can uniquely identify an individual on its own, e.g. patient names [11], and is sometimes referred to as an explicit identifier in the literature [12].

Indirect identifier Information that can be used in combination with auxiliary information to re-identify an individual, and is sometimes referred to as a quasi-identifier in the literature [12, 13].

Acronyms and Abbreviations

AUC	Area Under the Curve.
BatchNorm	Batch Normalization layers of a convolutional neural network.
CNN	Convolutional Neural Network.
CT	Computed Tomography.
DICOM	Digital Imaging and Communications in Medicine.
DP	Differential Privacy.
FID	Fréchet Inception Distance.
GAN	Generative Adversarial Network.
HIPAA	Health Insurance Portability and Accountability Act.
LayerNorm	Layer Normalization layers of a vision transformer.
MAE	Mean Absolute Error.
MRI	Magnetic Resonance Imaging.
NIST	National Institute of Standards and Technology.
NM	Nuclear Medicine.
PET	Positron Emission Tomography.
PGGAN	Progressive Growing Generative Adversarial Network.
PHI	Protected Health Information.
PII	Personally Identifiable Information.
RMSE	Root Mean Squared Error.
SSIM	Structural Similarity Index Measure.
US	Ultrasound.
ViT	Vision Transformer.
XR	X-ray.

1. Introduction

Medical imaging is a domain in which deep learning models have the potential to be clinically meaningful, with significant effect on patient outcomes [14–16]. Training deep learning models using medical images necessarily requires the acquisition and processing of patient data, which has raised privacy concerns. Particularly, the current state of the art has shown that images from a model's training set can be reconstructed [17–20], which, in a medical context, could pose viable harm to patient privacy if identifiable information can be gleaned from reconstructed images.

De-identification is thus the standard when using medical images to train deep learning models, i.e. removing information perceived as useful to patient identification [5, 21]. However, de-identification techniques may not definitively remove every pixel or voxel of information that has the potential to identify a patient, which could be a vast amount of the image, and images must retain utility to train a model [22]. There may consequently be some level of risk, such that a patient could be re-identified from an image that has been de-identified.

This report reviews the risk of patient re-identification from de-identified medical images and primarily addresses the following questions:

> **Q1**: What patient re-identification risks to a de-identified dataset are present in training deep learning models on radiology image data?
> **Q2**: What is the magnitude of these risks according to model architectures, imaging modalities, and anatomical regions?
> **Q3**: What mitigations can be taken to reduce the risk of re-identification?

These questions are constrained under the assumptions that 1) we have access to a model's parameters alone, e.g. an exported checkpoint, and 2) we know the imaging modality and anatomical region of the model's training images.

The scope of this report is narrowed to radiology-specific modalities. Other modalities and their specifications are not directly considered, including but not limited to, histopathology images, biospecimens, physiological time-based waveforms, and image-related but non-image objects such as DICOM structured reports.

Further, this report focuses on the training and development of radiology-based deep learning models and does not discuss multimodal models, e.g. vision-language models, matters of ownership, or policy recommendations.

Thus far, no comprehensive understanding of the risks to patient re-identification from model parameters alone exists. We aim to address this by proposing a two-stage reconstruction approach, predicting patient demographics from reconstructions, determining cases in which reconstructions and their demographics could aid in patient re-identification, and presenting potential mitigation strategies.

2. Related Work

This section discusses the theoretical background and common methods for reconstructing training data from a deep learning model. Note that reconstruction is often referred to as 'inversion' or an 'inversion attack' in the literature.

2.1. Theoretical Preliminaries

Deep neural networks may memorize specific features from the training data instead of learning general patterns to perform tasks [23]. Regarding image models, a notion of approximate memorization has been defined as a sufficient level of image similarity between an image and its reconstruction [24], and has often been linked to privacy leakage. Overfitting is considered to be one indicator of a model's memorization and a sufficient condition for privacy leakage, although memorization has been found to exist without overfitting [25].

A relationship between model inversion (i.e. reconstructing training data) and the presence of memorization in a model exists, such that memorized training examples may be more distinctly encoded within the trained model's parameters [17, 18]. This relationship is founded upon the idea that a trained model's parameters are generally determined by its training data [26]. Training data has been inverted within various bounds, including access to some held-out set of training data [17], federated learning environments with access to training data gradients [27, 28], and access to only model parameters in pixel-space [20]. Inversion methods are discussed below, in brief.

2.2. Gradient-based inversion

Direct gradient-based attacks aim to approximate training data by leveraging a shared gradient [29, 30], which can lead to pixel-wise accurate recovery of images [19, 31]. Attacks typically optimize over the input space to search for training examples whose gradient matches that of the observed gradient [19, 32]. Inversion from a single gradient query at a randomly chosen parameter value was demonstrated by [28], and batches of training data were able to be reconstructed without prior knowledge as shown by [33].

Undoubtedly, the usage of gradients to reconstruct training data requires access to gradients; these attacks are often studied in federated learning environments or specific settings where gradients are appropriately available.

2.3. Pixel-space and latent-space inversion

Without access to gradients, inversion attacks have been performed in pixel-space and latent-space, where reconstructing training data is treated as an optimization problem.

Requiring primarily a trained classification model and its parameters, the batch normalization (BatchNorm) layers of CNN architectures have been utilized to reconstruct images in pixel-space, termed DeepInversion [20]. Given that BatchNorm layers store the running means and variances of the original training data activations, these values are presumed to store the model's 'history' or 'memory' of previously seen data at multiple levels of representation through multiple BatchNorm layers. The method of DeepInversion then assumes that these intermediate activations follow a Gaussian distribution with mean and variance equal to the running statistics. As opposed to training a new attacker model, [20] directly optimizes random noise in pixel-space using the BatchNorm statistics to guide generation of the original training images, and aims to maximize the trained model's confidence. Specifically, the trained model's confidence is maximized for a given target class using labels from the

original training set. This is considered 'class-conditional' in the literature, where the random noise input is optimized per-class, and has been found to improve the efficiency of optimization [20, 34].

Approaching inversion in latent-space requires a generator, e.g. generative adversarial network (GAN), to access the latent representations of images. [35] used a GAN to learn a distributional prior from a disjoint public dataset of the same scope (e.g. if the target model is trained on a private dataset of faces, the GAN is trained on a separate publicly available dataset of faces). The prior then guides the optimization process in latent space where the aim is to find latent vector z that generates an image of maximal likelihood under the trained target model while maintaining realism with class-conditional information and additional regularizers. Similarly, [36] trained a separate decoder to minimize the distance between an original training image and its reconstruction through their latent space representations. However, this assumes access to latent space representations of the original training set, either from a held-out subset or the entire training set itself.

2.4. Architectural differences

Within the domain of radiology imaging, deep learning models are generally task-specific across modalities, with common non-generative tasks including detection, classification, and segmentation [16, 37, 38]. Detection and classification tasks often use some variation of a CNN architecture (DenseNet, ResNet, etc), with ViT architectures becoming increasingly common. Segmentation tasks tend to similarly use variations of CNNs, particularly encoder-decoder architectures such as the U-Net [39], and to a greater extent ViTs and hybrid ViTs such as the Swin-UNETR [40].

Regarding privacy, ViTs have been shown to memorize more, comparative to CNNs, where the attention mechanism particularly exacerbates vulnerability to attacks [41]. In conducting an ablation study, [41] also demonstrated that the fewer activation layers of a ViT are another contributing factor to privacy vulnerability.

3. Methods

3.1. Overview

We reconstruct images by approximating the trained target model's data manifold with a progressive growing GAN (PGGAN), maximizing the likelihood a generated sample was part of the original training set, and identifying patient information by predicting demographics from the reconstructions deemed likely (Figure 1).

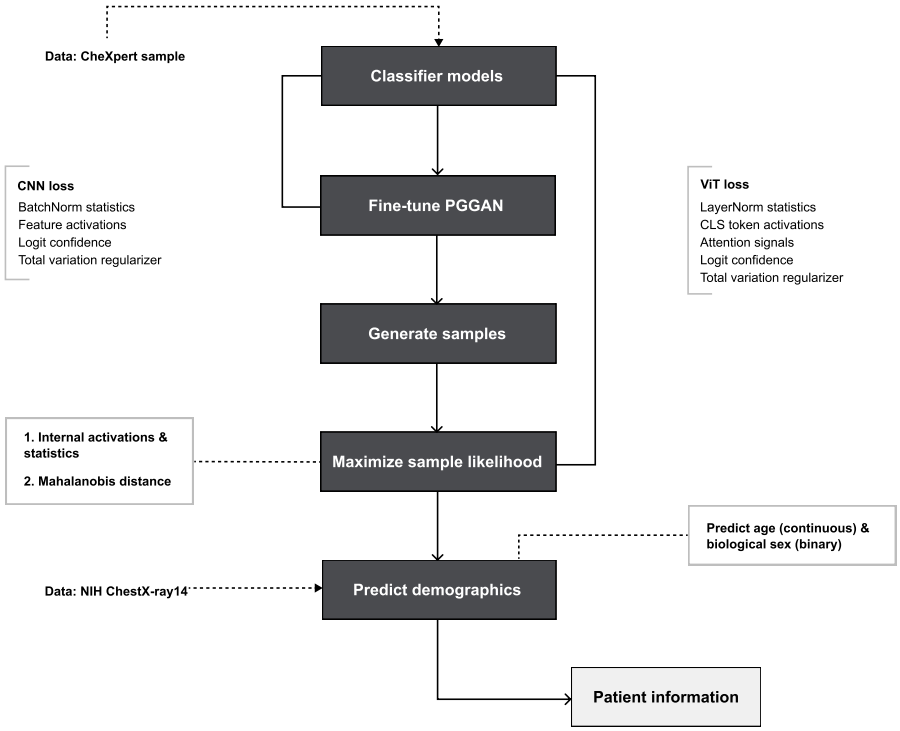

Figure 1: Overview of methods for re-identification from model parameters.

While these methods, i.e. an attack, employed by some unknown adversary may not appear to have a significant probability in practice, privacy is considered here in a worst-case scenario rather than an average-case one given the high-risk nature of deep learning models trained on sensitive medical data. We thus assume the probability of an attack attempting reconstruction and patient re-identification is non-zero. Realistically, medical data have been a target in past data breaches and attacks [42–44], and re-identifying patient information has been considered a lucrative target for health insurance companies [45] and data mining companies [46].

3.2. Materials

With our interest in understanding the effect of different model architectures and memorization, we consider 4 classification models: a CNN (`ResNet18` from `torchvision`), the CNN trained to overfit, a ViT (`vit_b_16` from `torchvision`), and the ViT trained to overfit. Each of these models were trained to predict cardiomegaly versus no finding from 2D chest x-rays as a binary classification task for

simplicity. For the overfit models, overfitting is confirmed with a training accuracy of 1.0 and a validation accuracy persisting at 0.75 (CNN) and 0.85 (ViT). Moving forward, we term each model a 'target model', i.e. the model from which we attempt to reconstruct training images and identify patient information.

The models were trained on 2D JPG chest x-rays from the CheXpert dataset [47]. The CheXpert dataset is publicly available and has been de-identified, which simulates the defined requirement of models trained on de-identified radiology images. Chest x-rays, specifically, are used given that structural images may allow for sufficiently clear per-patient textural differences or shape differences in comparison to functional images such as a cranial fMRI, as previously found in the literature [5, 48].

Additionally, we use the generator from a pre-trained PGGAN [49] that was trained on 2D PNG chest x-rays from the publicly available, de-identified NIH ChestX-ray14 dataset [50], instead of training a generator from scratch, due to compute limitations. The PGGAN was found to substantially improve reconstruction clarity compared to other generative models such as a variational autoencoder, likely due to its less restrictive latent space.

3.3. Stage 1: Approximate target model data manifold

If we assume that a target model's parameters are compressed representations of its training data, similar to [20], we can then fine-tune a pre-trained generator G to match those parameters, thus approximating the target model's training data manifold. Rather than optimizing on images or random Gaussian noise directly, as in previous methods, we optimize the generator to produce images that are more closely aligned to the target model's data manifold, according to its stored parameters. The pre-trained generator serves as an inductive prior, as it has already been trained on a different set of chest x-rays, which allows for a constrained search space to images considered realistic during optimization. By fine-tuning the generator such that its outputs induce similar internal activations in the target model as its original training data, we implicitly reconstruct the target model's data manifold, thereby compelling the generator to produce images that lie on or closer to the target model's true manifold. As opposed to learning the full probability density of the target model's training data, i.e. its 'distribution', fine-tuning the generator guides it towards regions of image space that look like the target model's training data, i.e. its 'manifold'.

This process varies between the CNN models and the ViT models, given differences in architecture. We leverage these architectural differences to sufficiently compare reconstruction performance per-architecture, which are discussed below. The fine-tuning process for each target model was completed on Google Colab with a T4 GPU.

For all target models, random noise latent vectors z are first sampled according to the generator's latent dimensionality (`dim=512`) and a batch size defined within compute limits (`batch_size=16`). Images are then produced from the latent vectors using the generator and passed through the target model to collect activations.

CNN models. Fine-tuning the generator towards the CNN models consists of 4 main components which are treated as loss terms:

1. **BatchNorm statistics**: Since BatchNorm layers learn population-level statistics (running mean and variance) during training, we use these statistics as proxies for the target model's data manifold.

2. **Feature activations**: A forward hook is first registered to retrieve activation maps from a desired feature layer, e.g. `layer4.1.conv2`, during forward passes. Matching feature activations between the generated image and the target model's frozen parameters encourages deep features to have certain statistics such as unit variance and non-zero mean, which aims to encourage reconstruction diversity that is semantically meaningful.

3. **Logits**: While we assume we do not have access to the real target model labels (e.g. cardiomegaly vs. no finding), the final `Linear` layer of the ResNet18 model's `state_dict` checkpoint contains the number of classes used to train the model through parameter `out_features`, where binary classification is designated by `Linear(in_features=512, out_features=1, bias=True)`. We can then use this information for class-conditional generation to encourage reconstruction of the representative data manifold per-class with logits.

4. **Total variation**: Total variation is a regularization term to support the plausibility of generated images by smoothing noise and artifacts common when generating images.

Each loss component measures the proximity between the generated images' activations in the target model to the target model's original frozen parameters. Total loss is computed as a weighted sum of the components, after which gradients are backpropagated through the frozen target model to the generator, updating its weights.

ViT models. Fine-tuning the generator towards the ViT models consists of 5 main components which we treat as loss terms:

1. **LayerNorm statistics**: In comparison to the global BatchNorm statistics of a CNN, the layer normalization (LayerNorm) layers of a ViT reflect local structure per-token, with no running mean or variance. However, we can aggregate across tokens and use the pre-normalized mean and variance to obtain more information about the true manifold.

2. **CLS token activations**: As the CLS token reflects the ViT's high-level representations of its training data, we extract its embedding from the final transformer encoder and regularize its magnitude to prevent random noise from being considered class-representative. While its 'counterpart' in this fine-tuning is the feature activations in a CNN, the CLS token embeddings afford global representations.

3. **Attention signals**: With the ViT, we take advantage of the attention mechanism to guide generated images towards the target model's internal processes. By measuring the entropy of attention distributions, we can minimize it to increase the confidence of attention maps, where each token attends strongly to only a few other tokens in the input sequence, i.e. ensuring that attention is focused, which encourages the generator to produce semantically structured images. Pairwise cosine similarity is also computed between attention heads (where ViT target models use multi-head attention) to encourage diverse and non-overlapping attention patterns.

4. **Logits**: The final `Linear` layer of the `vit_b_16` similarly contains the number of classes used to train the model, which we use in the same way as the CNN for class-conditional generation.

5. **Total variation**: The total variation regularizer is also used in the same way as the CNN.

Total loss is likewise computed as a weighted sum of the components, with gradients backpropagated to update the generator's weights.

3.4. Stage 2: Maximize sample likelihood

With the sufficiently large number of samples generated for each model, it is unlikely that every sample generated is part of the target model's original training set, given that fine-tuning the generator had the goal of approximating the target model's global training data manifold and not individual samples. In other words, each sample may lie along the target model's data manifold, depending on the performance of fine-tuning, but may not all be an individual sample from the target model's training set.

As our aim is to reconstruct images from the target model's original training set with its parameters alone, we aim to identify images that most likely to be part of the original training set. From all samples generated, the following are determined:

1. Images that closely match the target model's parameters, using the same components as stage 1 for each architecture, with the exception of the total variation regularizer.

2. Images that are close in Mahalanobis distance to the target model in feature-space, where low-distance is considered likely for a sample to be in-distribution. We assume that the distribution of the original training images are approximately multivariate Gaussian based on the model's BatchNorm (CNN) or aggregated LayerNorm (ViT) statistics. Under this assumption, Mahalanobis distance is considered appropriate, as it assumes the feature-space is approximately multivariate Gaussian. Equation 1 gives the formulation of Mahalanobis distance, where x is a feature vector, μ is the mean of the distribution, and C is a positive-definite covariance matrix.

$$D_M(x) = \sqrt{(x-\mu)^T C^{-1} (x-\mu)} \tag{1}$$

Specifically, we determine the samples that both closely match the target model's parameters and are close in distance to the target model in feature space. Such overlap may indicate that those samples are strong candidates for being in-distribution, relative to the target model. These 'likely' samples, then resemble the target model's data manifold from both a functional and distributional standpoint. Where stage 1 intends to learn the target model's overall data manifold, stage 2 attempts to recover the individual samples that best explain the target model's parameters.

3.5. Demographic prediction

A multi-task `DenseNet121` from `torchvision` was trained to predict patient age as a continuous value (regression task) and biological gender as male vs. female (binary classification task). Such demographic information has previously been predicted from chest x-rays in the context of bias and fairness [51].

The demographic model was trained on the full NIH ChestX-ray14 dataset, similar to the original PGGAN. Rare age bins were oversampled in the training set, given significant imbalance, using `WeightedRandomSampler` from PyTorch.

For age, the regression task achieved a validation mean absolute error (MAE) of 5.12, a validation root mean squared error (RMSE) of 6.63, and a validation R^2 of 0.845. For biological gender, the classification task achieved a validation accuracy of 0.95 and a validation area under the curve (AUC) of 0.991. Clearly, the model does not achieve perfect predictions of age or gender, but is able to predict rough estimates of each. See Figure 2 for the distribution of age predictions and ground truth and Figure 3 for the distribution and confusion matrix of gender, where both figures refer to predictions from the model's validation set.

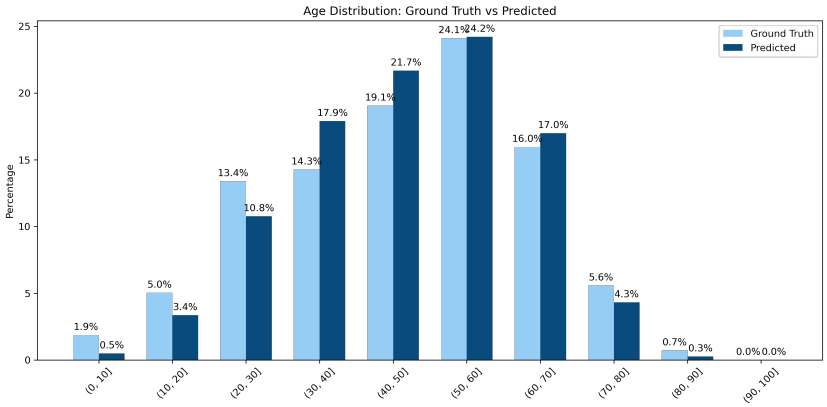

Figure 2: Distribution of predictions and ground truth from the validation set for age, binned only for visualization.

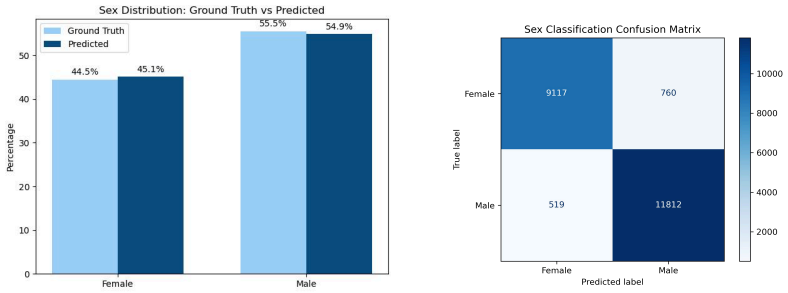

Figure 3: Distribution of predictions and ground truth, as well as the confusion matrix, from the validation set for biological gender.

4. Results

4.1. Stage 1

To validate this fine-tuning method, we measure distributional differences between the original generator and the fine-tuned generator, as well as between the fine-tuned generator and the target model, to examine if fine-tuning is shifting the overall distribution of the generator and if so, shifting towards the target model.

Per-image entropy is calculated in pixel-space with Shannon entropy, which measures the distribution of pixel intensities for an image. Higher entropy indicates greater variability, where lower entropy indicates a more uniform image, e.g. all grayscale noise. The average entropy is then taken across all generated images. Embedding-space is accessed with Principal Component Analysis (PCA), from which average kurtosis is calculated across principal components. Fréchet Inception Distance (FID) is generally used to evaluate the quality of images generated by a GAN and is used here to measure the distance between 2 distributions in feature space, where feature representations are extracted with a pre-trained InceptionV3 network from `torchvision`.

It is clear that fine-tuning the generator does, in fact, shift its distribution, which is seen in both pixel-space and embedding-space, as metrics for the fine-tuned generator are distinct from the original generator (Figure 4).

	CNN		Overfit CNN		ViT		Overfit ViT	
	Original → Fine-tuned	Target CNN	Original → Fine-tuned	Target Overfit CNN	Original → Fine-tuned	Target ViT	Original → Fine-tuned	Target Overfit ViT
Pixel entropy	7.402 → 6.486	7.963	7.402 → 6.176	7.965	7.402 → 6.176	7.963	7.402 → 6.125	7.962
PCA kurtosis	1.065 → 0.436	4.306	1.065 → 0.566	0.532	1.065 → 1.001	4.299	1.065 → 0.58	0.396
FID score	159.9088	178.9562	95.5836	198.1082	74.7530	104.9406	90.0923	171.7221

Figure 4: Distribution-level metrics in pixel-space and embedding-space from Stage 1 fine-tuning.

In embedding-space, fine-tuning leads the generator closer to the kurtosis distribution of the overfit CNN and overfit ViT, in comparison to the CNN and ViT that are not overfitting (Figure 5).

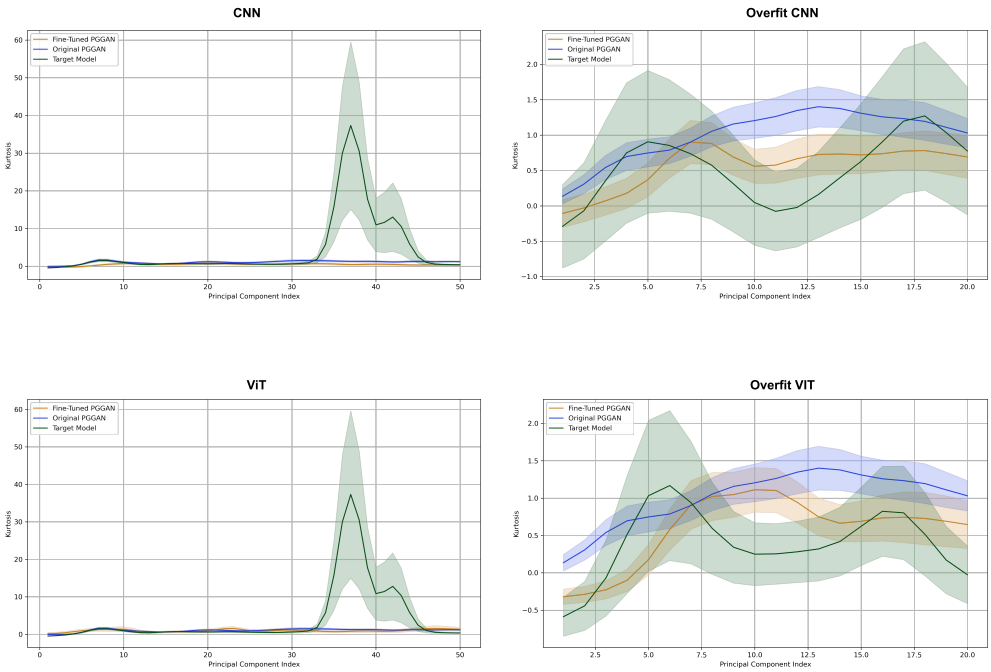

Figure 5: Smoothed kurtosis curves across principal components for the original generator, each fine-tuned generator, and each target model. Shading represents the bootstrapped standard error.

Regarding FID scores, the distributions of the fine-tuned generator on both the ViT and the overfit ViT are closer in distance to the original generator, but are also closer in distance to each target model, in comparison to the CNN and overfit CNN.

4.2. Stage 2

The number of samples considered likely varied per model: the CNN had 2 samples, ViT had 1 sample, overfit ViT had 2 samples, and the overfit CNN had the highest number with 23 samples.

To evaluate the efficacy of these methods, we conduct a nearest neighbor search in feature space between the likely samples and the real target model training images to examine if the likely generated samples are indeed close to images from the target model's training data. The Structural Similarity Index Measure (SSIM) is also computed between the sample and its nearest neighbor.

In practice, we assume we do not have access to the target model's training data, however, we do so here for validation purposes only. Figure 6 displays the likely samples and their nearest target neighbors for the overfit CNN, and Figure 7 displays the same for the CNN, ViT, and overfit ViT.

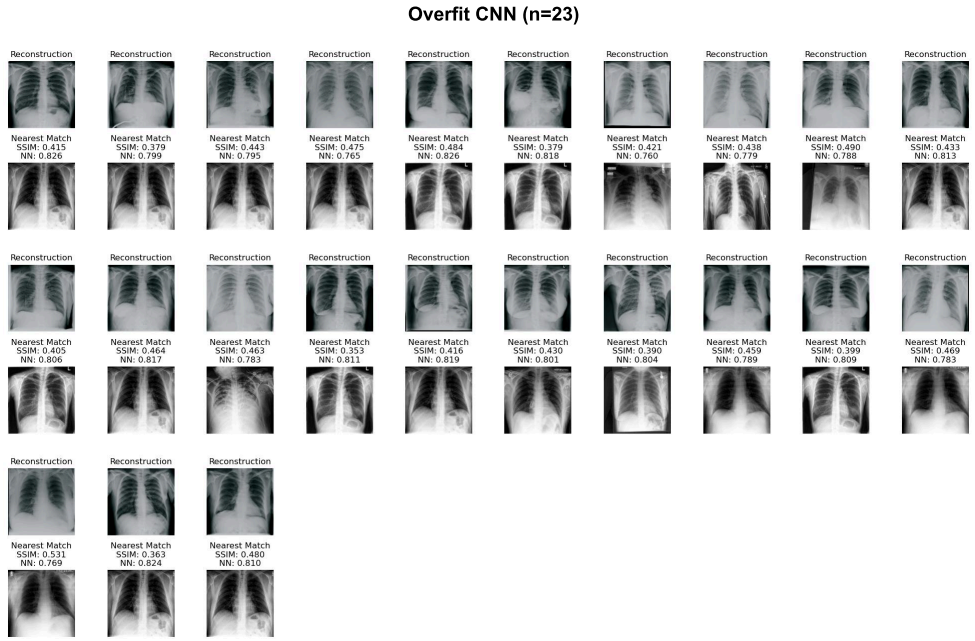

Figure 6: Reconstructed samples for the Overfit CNN with their nearest target neighbor in feature-space, and the SSIM score between them.

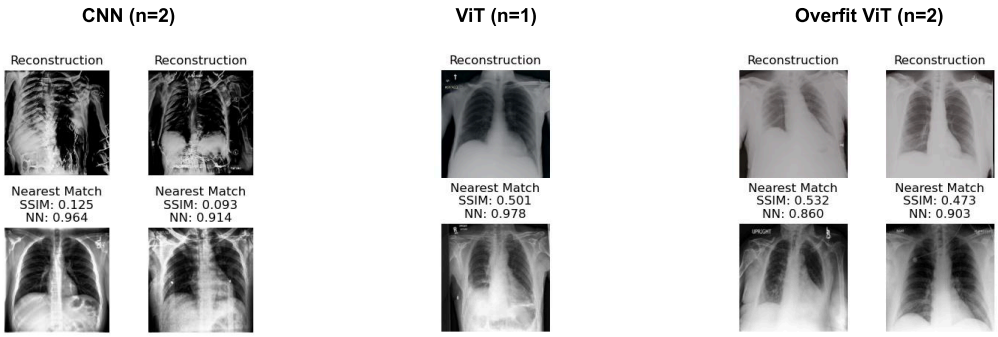

Figure 7: Reconstructed samples for the CNN, ViT, and overfit ViT with their nearest target neighbor in feature-space, and the SSIM score between them.

It should be noted that SSIM scores are relatively poor across all models. This may be in part due to stylistic differences between reconstructed samples and the target model's real training images, or perhaps because while reconstructed samples and their nearest neighbors are relatively close in shape, generally, they are not exact matches.

Particularly for the CNN, reconstructed samples are significantly worse in comparison to samples from the other models. Fine-tuning the generator towards the CNN may evidently shift its distribution substantially, yet do so at the expense of realistic chest x-rays. Whereas, the other models maintain greater realism in generated samples, which may affect image interpretability when aiming to extract identifiable information from them.

4.3. Demographic prediction on reconstructions

Figures 8 and 9 display the predictions of the reconstructed samples, the predictions of their nearest neighbors, and the ground truth of the nearest neighbors which was indicated in the original CheXpert dataset. The absolute error for age in years and a flag for correct gender prediction in comparison to the ground truth is also shown.

The overfit ViT appears to perform better than the other models, in terms of both error for age and accuracy for gender; it has the lowest absolute error for age and both reconstructed samples had accurate predictions for gender. However, it is important to note that the sample size is small (n=2 likely samples). While the overfit CNN has the greatest number of likely samples (n=23), predictions for both age and gender on these samples are comparatively deficient, with approximately 65% of samples accurately predicting gender and highest absolute error for age being around 30 years.

Overfit CNN

	PredAge_Recon	PredAge_NN	Age_GT	AbsoluteError_Recon	AbsoluteError_NN	PredSex_Recon	PredSex_NN	Sex_GT	Correct_Recon	Correct_NN
0	51.083	36.347	38.0	13.083	1.653	Male	Male	Male	True	True
1	40.074	36.347	38.0	2.074	1.653	Male	Male	Male	True	True
2	57.408	36.347	38.0	19.408	1.653	Male	Male	Male	True	True
3	51.036	36.347	38.0	13.036	1.653	Male	Male	Male	True	True
4	66.833	38.156	36.0	30.833	2.156	Female	Female	Female	True	True
5	56.344	38.156	36.0	20.344	2.156	Male	Female	Female	False	True
6	42.078	54.446	60.0	17.922	5.554	Male	Male	Male	True	True
7	51.957	48.092	49.0	2.957	0.908	Male	Female	Female	False	True
8	61.132	43.36	31.0	30.132	12.36	Male	Female	Female	False	True
9	61.883	36.347	38.0	23.882	1.653	Male	Male	Male	True	True
10	56.758	38.156	36.0	20.758	2.156	Male	Female	Female	False	True
11	56.860	36.347	38.0	18.860	1.653	Male	Male	Male	True	True
12	47.131	37.415	48.0	0.869	10.585	Female	Male	Male	False	True
13	61.876	38.156	36.0	25.876	2.156	Female	Female	Female	True	True
14	54.889	36.348	38.0	16.889	1.653	Female	Male	Male	False	True
15	49.812	44.66	32.0	17.712	12.66	Female	Male	Male	False	True
16	49.670	22.176	28.0	21.670	5.824	Male	Female	Female	False	True
17	63.451	55.218	44.0	19.451	11.218	Male	Male	Male	True	True
18	38.040	38.156	36.0	2.040	2.156	Female	Female	Female	True	True
19	42.856	55.218	44.0	1.144	11.218	Male	Male	Male	True	True
20	65.286	55.218	44.0	21.286	11.218	Male	Male	Male	True	True
21	45.753	36.347	38.0	7.753	1.653	Male	Male	Male	True	True
22	67.026	36.347	38.0	29.026	1.653	Male	Male	Male	True	True

Figure 8: Table of predicted demographics (age and gender) for reconstructions and their nearest neighbors, alongside the ground truth, for the overfit CNN.

CNN

	PredAge_Recon	PredAge_NN	Age_GT	AbsoluteError_Recon	AbsoluteError_NN	PredSex_Recon	PredSex_NN	Sex_GT	Correct_Recon	Correct_NN
0	53.775	27.856	21.0	32.775	6.856	Female	Male	Male	False	True
1	56.651	64.735	39.0	17.651	25.735	Female	Male	Male	False	True

ViT

	PredAge_Recon	PredAge_NN	Age_GT	AbsoluteError_Recon	AbsoluteError_NN	PredSex_Recon	PredSex_NN	Sex_GT	Correct_Recon	Correct_NN
0	45.282	73.385	69.0	23.718	4.385	Male	Male	Male	True	True

Overfit ViT

	PredAge_Recon	PredAge_NN	Age_GT	AbsoluteError_Recon	AbsoluteError_NN	PredSex_Recon	PredSex_NN	Sex_GT	Correct_Recon	Correct_NN
0	43.817	30.173	33.0	10.817	2.827	Female	Female	Female	True	True
1	57.571	60.202	54.0	3.571	6.202	Male	Male	Male	True	True

Figure 9: Table of predicted demographics (age and gender) for reconstructions and their nearest neighbors, alongside the ground truth, for the CNN, ViT, and overfit ViT.

5. Discussion

5.1. Re-identification cases

With reconstructed samples and their predicted demographics, it is suitable to observe that there exists some potential to re-identify a patient, although heavily dependent on the circumstances.

If, for example, a health insurance company aims to reduce their financial risk or build better risk models, the reconstructions and their demographics may significantly aid in re-identifying a patient, particularly for the overfit ViT where demographics tend to be more accurate. Since a health insurance company is more likely to have access to additional information such as medical records, it is also more likely that with a reconstructed chest x-ray, patient gender, and approximate age, the company would be able to re-identify a patient or at least reduce the population of potential patients down to small numbers. Reconstructed images may also contain visible hardware such as pacemakers, which would aid in re-identification.

However, without access to medical records, radiology reports, or other additional information, it appears infeasible, or at least very difficult, for an attacker to re-identify a patient from a reconstructed chest x-ray and approximate demographics alone. Even if, for the sake of argument, the attacker had prior knowledge of the institution site where the x-ray was acquired, the number of patients that fit this cohort could still be a population of thousands.

Thus, while reconstructed images and their approximate demographics alone are unlikely to be sufficient to re-identify a patient, there still exists some level of risk to patient privacy depending on additional information some attacker may have.

5.2. Effect of architecture and memorization

Distributionally, with the CNN models (overfit and non-overfit) as the target, there was generally greater distance between the original generator and the fine-tuned generator. It then seems likely that the CNN models enable the generator to leave its original manifold more aggressively during fine-tuning, but in doing so may sacrifice semantically meaningful and plausible images, particularly for the non-overfit CNN. In comparison, with the ViT models as the target, greater distance between the original generator and the fine-tuned generator, as well as between the fine-tuned generator and the target model, was observed. This could mean that while the ViT models do not enable the generator to leave its original manifold as aggressively as the CNN, the outputs produce more semantically valid chest x-rays.

One interpretation may be that although memorization does have an effect, how a model's memory is encoded in its parameters and architectures is also significant. In brief, a CNN's convolutional layers encode local features hierarchically, whereas a vision transformer captures global relationships encoded as distributed patterns and activations spread across multiple dimensions. Since fine-tuning is based on the intermediate activation layers of the target model, it may be more difficult for the generator to shift towards activations that are dispersed in the ViT and not in a designated feature map like the CNN.

This relationship is similarly reflected at the sample level, where the CNN models had more likely samples in comparison to their ViT counterparts, i.e. the CNN had more likely samples than the ViT (albeit only 1) and the overfit CNN had more likely samples than the overfit ViT. Yet, the ViT models, and specifically the overfit ViT, had samples that allowed for more accurate demographic predictions than the CNN models.

It is then apt to posit that with a model's parameters from a frozen checkpoint alone, the overfit ViT poses the greatest risk to patient re-identification. Although the generator fine-tuned towards the

overfit ViT does not produce images that are as closely aligned to the target model's internal statistics and parameters, the outputs are more semantically meaningful and stylistically similar to the target, upon which demographics can be more accurately predicted. Whereas, the generator fine-tuned towards the CNN models leads to images that are more closely aligned with the target model's parameters yet produces less semantically and stylistically similar images. In other words, while the CNN (specifically the overfit CNN) leads to more faithful reconstructions to the target model in shape, the overfit ViT allows for more accurate demographic predictions, i.e. identifiable information, and is thus a greater risk to patient re-identification.

5.3. Potential mitigations

Differential privacy (DP) is the canonical approach to mitigating reconstruction attacks and is common in privacy-preserving machine learning more generally [22, 52]. Within the image domain, DP typically adds calibrated noise, non-trivially tuned by hyperparameter ϵ, into model parameters during training which, in principle, should limit the possibility of reconstructing data by concealing or minimizing the relationship between training data and the model's response [53–55]. In practice, however, a tradeoff between total privacy and model utility exists similar to de-identification: too much noise added by DP leads to less accurate and generalizable models [22, 55].

It should also be noted that for some model reconstruction attacks, DP is an insufficient defense even with strong privacy budgets, e.g. $\epsilon = 0.1$ [35, 55]. This is due to the aim of DP not being to protect the entire data distribution, but rather hide the presence of a single sample in the training set [35, 54].

Based on the results of this research, a more practical and straightforward mitigation is to ensure that models are not overfitting, i.e. a simple metric as a proxy for memorization to some extent. Since the overfit ViT, specifically, led to the most accurate identifiable information (demographics) extracted from reconstructed images compared to the other models, it is not insignificant to use methods to prevent overfitting during training. Such methods are already commonly used in deep learning models for medical imaging, and include the use of regularization such as dropout [56] and various image augmentation techniques [57] which are particularly beneficial when the training dataset cannot itself be increased or diversified.

5.4. Limitations

This research has the following limitations that should be considered:

1. Only 2D images in PNG/JPG format were studied. Since medical images are primarily acquired and used in 3-dimensions, it would be valuable to examine if the results still hold with 3D images.

2. Chest x-rays were the only modality and anatomical region considered. While chest x-rays may be inherently better suited to patient re-identification, given that structural imaging allows for visibly unique per-patient differences, it would also be important to see if the methods used are feasible on other modalities and regions.

3. Classification was the only task used for the target models. Although class-conditional reconstruction has been previously established as beneficial, the prevalence of other tasks such as segmentation in medical imaging, renders those tasks similarly beneficial to understand their behavior with the methods used.

4. Mitigations such as DP were unable to be undertaken against the methods used in this research, due to time constraints, and should also be carried out to examine its validity as a robust defense.

5. This research does not attempt to link reconstructions and their demographics to real patient identities, in part due to data usage agreements from the publicly available datasets used. The re-identification scenarios in section 5.1. were given to demonstrate that re-identification is feasible under certain circumstances, in principle.

6. Conclusion

This report presents the use of reconstruction as a mechanism for patient re-identification, from a trained model's parameters. A two-stage reconstruction approach first approximates the trained target model's data manifold by fine-tuning a PGGAN generator pre-trained on images of the same modality and region (2D chest x-rays), and then identifies which of the reconstructed samples are most likely to fall under the target model's true training set. Age and biological gender are then predicted from the likely reconstructed samples.

Results show that an overfit ViT is the model most vulnerable to patient re-identification in comparison to a CNN, overfit CNN, and ViT, indicating that architecture and memorization are contributing factors to re-identification. We also outline cases in which re-identification would be possible from the reconstructed images and their predicted demographics, and discuss relevant mitigation strategies.

Given the constraint that we only have access to a trained model's parameters, e.g. through a frozen checkpoint, this research determines a realistic estimation of risk to patient re-identification for models trained on de-identified medical images.

Future research should examine the results of adversarial training with a full GAN (both generator and discriminator) instead of only using a pre-trained generator, which could aid in regularization and improve both the semanticity and visual style (e.g. contrast) of the reconstructed images. The demographic model could also be used as feedback during fine-tuning to encourage generation of images that can be more accurately predicted upon.

7. Acknowledgments

Thank you to Mayo Clinic Platform for the support of this research. Particularly, with gratitude, thank you to Adam Resnick and Rob Blundo for their supervision and continual guidance throughout the course of this work.

References

1. McCallister, E., Grance, T. & Scarfone, K. A. *Guide to protecting the confidentiality of Personally Identifiable Information (PII)* en. Tech. rep. (National Institute of Standards and Technology, Gaithersburg, MD, 2010). https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-122.pdf (2025).
2. National Electrical Manufacturers Association (NEMA). *E.3.10 Retain Safe Private Option* https://dicom.nema.org/medical/dicom/current/output/chtml/part15/sect_E.3.10.html (2025).
3. Moore, W. & Frye, S. Review of HIPAA, Part 1: History, Protected Health Information, and Privacy and Security Rules. en. *Journal of Nuclear Medicine Technology* **47,** 269–272. ISSN: 0091-4916, 1535-5675. http://tech.snmjournals.org/lookup/doi/10.2967/jnmt.119.227819 (2025) (Dec. 2019).
4. European Data Protection Supervisor. *AEPD-EDPS joint paper on 10 misunderstandings related to anonymisation — European Data Protection Supervisor* en. Apr. 2021. https://www.edps.europa.eu/data-protection/our-work/publications/papers/aepd-edps-joint-paper-10-misunderstandings-related_en (2025).
5. Clunie, D. A. *et al.* Report of the Medical Image De-Identification (MIDI) Task Group - Best Practices and Recommendations. *ArXiv,* arXiv:2303.10473v2. ISSN: 2331-8422. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10081345/ (2025) (Apr. 2023).
6. Xia, W. *et al.* Enabling realistic health data re-identification risk assessment through adversarial modeling. *Journal of the American Medical Informatics Association* **28,** 744–752. ISSN: 1527-974X. https://doi.org/10.1093/jamia/ocaa327 (2025) (Apr. 2021).
7. Fernandez, V. *et al. Privacy Distillation: Reducing Re-identification Risk of Diffusion Models* en. in *Deep Generative Models* (eds Mukhopadhyay, A., Oksuz, I., Engelhardt, S., Zhu, D. & Yuan, Y.) (Springer Nature Switzerland, Cham, 2024), 3–13. ISBN: 978-3-031-53767-7.
8. El Emam, K., Jonker, E., Arbuckle, L. & Malin, B. A Systematic Review of Re-Identification Attacks on Health Data. en. *PLoS ONE* **6** (ed Scherer, R. W.) e28071. ISSN: 1932-6203. https://dx.plos.org/10.1371/journal.pone.0028071 (2025) (Dec. 2011).
9. Rights (OCR), O. f. C. *The HIPAA Privacy Rule* en. Page. Last Modified: 2024-09-27T14:55:21-0400. May 2008. https://www.hhs.gov/hipaa/for-professionals/privacy/index.html (2025).
10. IBM Corporation. *Reidentification risk for AI* en. Publisher: IBM Corporation. Oct. 2015. https://dataplatform.cloud.ibm.com/docs/content/wsj/ai-risk-atlas/dataplatform.cloud.ibm.com/docs/content/wsj/ai-risk-atlas/reidentification.html (2025).
11. Oh, S., Sung, M., Rhee, Y., Hong, N. & Park, Y. R. Evaluation of the Privacy Risks of Personal Health Identifiers and Quasi-Identifiers in a Distributed Research Network: Development and Validation Study. EN. *JMIR Medical Informatics* **9.** Company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics Publisher: JMIR Publications Inc., Toronto, Canada, e24940. https://medinform.jmir.org/2021/5/e24940 (2025) (May 2021).
12. Osorio-Marulanda, P. A. *et al.* Privacy Mechanisms and Evaluation Metrics for Synthetic Data Generation: A Systematic Review. *IEEE Access* **12.** Conference Name: IEEE Access, 88048–88074. ISSN: 2169-3536. https://ieeexplore.ieee.org/document/10568134/?arnumber=10568134 (2025) (2024).
13. Dalenius, T. Finding a Needle In a Haystack or Identifying Anonymous Census Records. English. Num Pages: 329 Publisher: Statistics Sweden (SCB), 329. ISSN: 0282423X. https://www.proquest.com/docview/1266806751?pq-origsite=gscholar&fromopenview=true&sourcetype=Scholarly%20Journals (2025) (Sept. 1986).

14. Oren, O., Gersh, B. J. & Bhatt, D. L. Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints. *The Lancet Digital Health* **2,** e486–e488. ISSN: 2589-7500. https://www.sciencedirect.com/science/article/pii/S2589750020301606 (2025) (Sept. 2020).

15. Avanzo, M., Stancanello, J., Pirrone, G., Drigo, A. & Retico, A. The Evolution of Artificial Intelligence in Medical Imaging: From Computer Science to Machine and Deep Learning. *Cancers* **16,** 3702. ISSN: 2072-6694. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11545079/ (2025) (Nov. 2024).

16. Pinto-Coelho, L. How Artificial Intelligence Is Shaping Medical Imaging Technology: A Survey of Innovations and Applications. *Bioengineering* **10,** 1435. ISSN: 2306-5354. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10740686/ (2025) (Dec. 2023).

17. Haim, N., Vardi, G., Yehudai, G., Shamir, O. & Irani, M. Reconstructing Training Data From Trained Neural Networks. en. *Advances in Neural Information Processing Systems* **35,** 22911–22924. https://proceedings.neurips.cc/paper_files/paper/2022/hash/906927370cbeb53778110062 3cca6fa6-Abstract-Conference.html (2025) (Dec. 2022).

18. Suhail, P. & Sethi, A. *Network Inversion for Training-Like Data Reconstruction* arXiv:2410.16884 [cs]. Oct. 2024. http://arxiv.org/abs/2410.16884 (2025).

19. Zhu, L., Liu, Z. & Han, S. *Deep Leakage from Gradients* in *Advances in Neural Information Processing Systems* **32** (Curran Associates, Inc., 2019). https://proceedings.neurips.cc/paper/2019/hash/60a6c4002cc7b29142def8871531281a-Abstract.html (2025).

20. Yin, H. *et al. Dreaming to Distill: Data-free Knowledge Transfer via DeepInversion* arXiv:1912.08795 [cs]. June 2020. http://arxiv.org/abs/1912.08795 (2025).

21. Rempe, M., Heine, L., Seibold, C., Hörst, F. & Kleesiek, J. *De-Identification of Medical Imaging Data: A Comprehensive Tool for Ensuring Patient Privacy* arXiv:2410.12402 [eess]. Oct. 2024. http://arxiv.org/abs/2410.12402 (2025).

22. Shokri, R. & Shmatikov, V. *Privacy-Preserving Deep Learning* en. in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (ACM, Denver Colorado USA, Oct. 2015), 1310–1321. ISBN: 978-1-4503-3832-5. https://dl.acm.org/doi/10.1145/2810103.2813687 (2025).

23. Wei, J. *et al. Memorization in deep learning: A survey* arXiv:2406.03880 [cs]. June 2024. http://arxiv.org/abs/2406.03880 (2025).

24. Carlini, N. *et al. Extracting Training Data from Diffusion Models* en. in (2023), 5253–5270. ISBN: 978-1-939133-37-3. https://www.usenix.org/conference/usenixsecurity23/presentation/carlini (2025).

25. Carlini, N. *et al. Extracting Training Data from Large Language Models* en. in (2021), 2633–2650. ISBN: 978-1-939133-24-3. https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting (2025).

26. Tian, H. *et al. Simulating Training Dynamics to Reconstruct Training Data from Deep Neural Networks* en. in (Oct. 2024). https://openreview.net/forum?id=ZJftXKy12x (2025).

27. Enthoven, D. & Al-Ars, Z. *Fidel: Reconstructing Private Training Samples from Weight Updates in Federated Learning* arXiv:2101.00159 [cs]. Apr. 2022. http://arxiv.org/abs/2101.00159 (2025).

28. Wang, Z., Lee, J. & Lei, Q. *Reconstructing Training Data from Model Gradient, Provably* en. in *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics* ISSN: 2640-3498 (PMLR, Apr. 2023), 6595–6612. https://proceedings.mlr.press/v206/wang23g.html (2025).

29. Li, C., Song, Z., Wang, W. & Yang, C. *A Theoretical Insight into Attack and Defense of Gradient Leakage in Transformer* arXiv:2311.13624 [cs]. Nov. 2023. http://arxiv.org/abs/2311.13624 (2025).

30. Geiping, J., Bauermeister, H., Dröge, H. & Moeller, M. *Inverting Gradients - How easy is it to break privacy in federated learning?* in *Advances in Neural Information Processing Systems* **33** (Curran Associates, Inc., 2020), 16937–16947. https://proceedings.neurips.cc/paper_files/paper/2020/hash/c4ede56bbd98819ae6112b20ac6bf145-Abstract.html (2025).

31. Yin, H. *et al. See through Gradients: Image Batch Recovery via GradInversion* en. in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Nashville, TN, USA, June 2021), 16332–16341. ISBN: 978-1-6654-4509-2. https://ieeexplore.ieee.org/document/9577731/ (2025).

32. Wu, R., Chen, X., Guo, C. & Weinberger, K. Q. *Learning To Invert: Simple Adaptive Attacks for Gradient Inversion in Federated Learning* en. in *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence* ISSN: 2640-3498 (PMLR, July 2023), 2293–2303. https://proceedings.mlr.press/v216/wu23a.html (2025).

33. Qian, J. *et al. GI-SMN: Gradient Inversion Attack against Federated Learning without Prior Knowledge* en. arXiv:2405.03516 [cs]. May 2024. http://arxiv.org/abs/2405.03516 (2025).

34. Tian, Z. *et al.* The Role of Class Information in Model Inversion Attacks Against Image Deep Learning Classifiers. *IEEE Transactions on Dependable and Secure Computing* **21,** 2407–2420. ISSN: 1941-0018. https://ieeexplore.ieee.org/document/10225397/ (2025) (July 2024).

35. Zhang, Y. *et al. The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks* en. in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Seattle, WA, USA, June 2020), 250–258. ISBN: 978-1-7281-7168-5. https://ieeexplore.ieee.org/document/9156705/ (2025).

36. Subbanna, N., Wilms, M., Tuladhar, A. & Forkert, N. D. An Analysis of the Vulnerability of Two Common Deep Learning-Based Medical Image Segmentation Techniques to Model Inversion Attacks. en. *Sensors* **21.** Number: 11 Publisher: Multidisciplinary Digital Publishing Institute, 3874. ISSN: 1424-8220. https://www.mdpi.com/1424-8220/21/11/3874 (2025) (Jan. 2021).

37. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H. & Aerts, H. J. W. L. Artificial intelligence in radiology. *Nature reviews. Cancer* **18,** 500–510. ISSN: 1474-175X. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6268174/ (2025) (Aug. 2018).

38. Kelly, B. S. *et al.* Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE). en. *European Radiology* **32,** 7998–8007. ISSN: 1432-1084. https://doi.org/10.1007/s00330-022-08784-6 (2025) (Nov. 2022).

39. Ronneberger, O., Fischer, P. & Brox, T. *U-Net: Convolutional Networks for Biomedical Image Segmentation* arXiv:1505.04597 [cs]. May 2015. http://arxiv.org/abs/1505.04597 (2025).

40. Hatamizadeh, A. *et al. Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images* arXiv:2201.01266 [eess]. Jan. 2022. http://arxiv.org/abs/2201.01266 (2025).

41. Zhang, G. *et al. How Does a Deep Learning Model Architecture Impact Its Privacy? A Comprehensive Study of Privacy Attacks on {CNNs} and Transformers* en. in (2024), 6795–6812. ISBN: 978-1-939133-44-1. https://www.usenix.org/conference/usenixsecurity24/presentation/zhang-guangsheng (2025).

42. Gabriel, M. H., Noblin, A., Rutherford, A., Walden, A. & Cortelyou-Ward, K. Data breach locations, types, and associated characteristics among US hospitals. eng. *The American Journal of Managed Care* **24,** 78–84. ISSN: 1936-2692 (Feb. 2018).

43. Seh, A. H. *et al.* Healthcare Data Breaches: Insights and Implications. en. *Healthcare* **8.** Number: 2 Publisher: Multidisciplinary Digital Publishing Institute, 133. ISSN: 2227-9032. https://www.mdpi.com/2227-9032/8/2/133 (2025) (June 2020).

44. Munro, D. *Cyber Attack Nets 4.5 Million Records From Large Hospital System* en. Section: Pharma & Healthcare. Aug. 2014. https://www.forbes.com/sites/danmunro/2014/08/18/cyber-attack-nets-4-5-million-records-from-large-hospital-system/ (2025).

45. Jordon, J. *et al. Hide-and-Seek Privacy Challenge: Synthetic Data Generation vs. Patient Re-identification* en. in *Proceedings of the NeurIPS 2020 Competition and Demonstration Track* ISSN: 2640-3498 (PMLR, Aug. 2021), 206–215. https://proceedings.mlr.press/v133/jordon21a.html (2025).

46. Kaissis, G. A., Makowski, M. R., Rückert, D. & Braren, R. F. Secure, privacy-preserving and federated machine learning in medical imaging. en. *Nature Machine Intelligence* **2,** 305–311. ISSN: 2522-5839. https://www.nature.com/articles/s42256-020-0186-1 (2025) (June 2020).

47. Irvin, J. *et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison* arXiv:1901.07031 [cs]. Jan. 2019. http://arxiv.org/abs/1901.07031 (2025).

48. Packhäuser, K. *et al.* Deep learning-based patient re-identification is able to exploit the biometric nature of medical chest X-ray data. en. *Scientific Reports* **12,** 14851. ISSN: 2045-2322. https://www.nature.com/articles/s41598-022-19045-3 (2025) (Sept. 2022).

49. Segal, B., Rubin, D. M., Rubin, G. & Pantanowitz, A. Evaluating the Clinical Realism of Synthetic Chest X-Rays Generated Using Progressively Growing GANs. en. *SN Computer Science* **2,** 321. ISSN: 2661-8907. https://doi.org/10.1007/s42979-021-00720-7 (2025) (June 2021).

50. Wang, X. *et al. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases* in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* ISSN: 1063-6919 (July 2017), 3462–3471. https://ieeexplore.ieee.org/document/8099852 (2025).

51. Adleberg, J. *et al.* Predicting Patient Demographics From Chest Radiographs With Deep Learning. en. *Journal of the American College of Radiology* **19,** 1151–1161. ISSN: 15461440. https://linkinghub.elsevier.com/retrieve/pii/S1546144022005440 (2025) (Oct. 2022).

52. Fang, H. *et al. Privacy Leakage on DNNs: A Survey of Model Inversion Attacks and Defenses* arXiv:2402.04013 [cs]. Sept. 2024. http://arxiv.org/abs/2402.04013 (2025).

53. Dwork, C. *Differential Privacy* en. in *Automata, Languages and Programming* (eds Bugliesi, M., Preneel, B., Sassone, V. & Wegener, I.) (Springer, Berlin, Heidelberg, 2006), 1–12. ISBN: 978-3-540-35908-1.

54. Zhou, Z. *et al. Model Inversion Attacks: A Survey of Approaches and Countermeasures* arXiv:2411.10023 [cs]. Nov. 2024. http://arxiv.org/abs/2411.10023 (2025).

55. Yang, Y.-Y., Chou, C.-N. & Chaudhuri, K. *Understanding Rare Spurious Correlations in Neural Networks* arXiv:2202.05189 [cs]. Oct. 2022. http://arxiv.org/abs/2202.05189 (2025).

56. Raju, N. & Augustine, D. P. in *Machine Intelligence* Num Pages: 19 (Auerbach Publications, 2023). ISBN: 978-1-003-42455-0.

57. Chlap, P. *et al.* A review of medical image data augmentation techniques for deep learning applications. en. *Journal of Medical Imaging and Radiation Oncology* **65.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1754-9485.13261, 545–563. ISSN: 1754-9485. https://onlinelibrary.wiley.com/doi/abs/10.1111/1754-9485.13261 (2025) (2021).