Differential privacy representation geometry for medical image analysis

Soroosh Tayebi Arasteh[1,2][0000−0003−1015−7733], Marziyeh Mohammadi[3], Sven Nebelung[1,2], and Daniel Truhn[1,2]

[1] Lab for AI in Medicine, RWTH Aachen University, Aachen, Germany
[2] Department of Diagnostic and Interventional Radiology, University Hospital RWTH Aachen, Aachen, Germany
[3] INDA Institute, RWTH Aachen University, Aachen, Germany
soroosh.arasteh@rwth-aachen.de

Abstract. Differential privacy (DP)'s effect in medical imaging is typically evaluated only through end-to-end performance, leaving the mechanism of privacy-induced utility loss unclear. We introduce Differential Privacy Representation Geometry for Medical Imaging (DP-RGMI), a framework that interprets DP as a structured transformation of representation space and decomposes performance degradation into encoder geometry and task-head utilization. Geometry is quantified by representation displacement from initialization and spectral effective dimension, while utilization is measured as the gap between linear-probe and end-to-end utility. Across over 594,000 images from four chest X-ray datasets and multiple pretrained initializations, we show that DP is consistently associated with a utilization gap even when linear separability is largely preserved. At the same time, displacement and spectral dimension exhibit non-monotonic, initialization- and dataset-dependent reshaping, indicating that DP alters representation anisotropy rather than uniformly collapsing features. Correlation analysis reveals that the association between end-to-end performance and utilization is robust across datasets but can vary by initialization, while geometric quantities capture additional prior- and dataset-conditioned variation. These findings position DP-RGMI as a reproducible framework for diagnosing privacy-induced failure modes and informing privacy model selection.

Keywords: Deep learning · Differential privacy · Representation learning · Medical image analysis · Privacy-preserving AI.

1 Introduction

Deep neural networks in medical image analysis are trained on highly sensitive patient data [13]. Although such models achieve state-of-the-art diagnostic performance, they may memorize individual-specific patterns, raising concerns about membership inference, reconstruction attacks, and regulatory compliance [10,18,19]. Differential privacy (DP) [7] provides a formal guarantee that limits the influence of any single patient on the learned model. A randomized algorithm

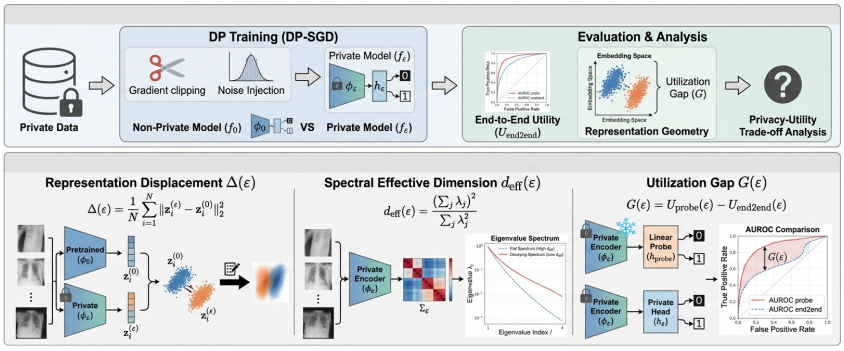

Fig. 1. Overview of DP-RGMI framework decomposing DP training into representation displacement $\Delta(\varepsilon)$, spectral structure $d_{\text{eff}}(\varepsilon)$, and utilization gap $G(\varepsilon)$.

$\mathcal{A}$ satisfies (ε, δ)-DP if for all neighboring datasets $\mathcal{D}, \mathcal{D}'$ differing in one sample and all measurable outputs $\mathcal{S}$,

$$\Pr[\mathcal{A}(\mathcal{D}) \in \mathcal{S}] \leq e^{\varepsilon} \Pr[\mathcal{A}(\mathcal{D}') \in \mathcal{S}] + \delta. \tag{1}$$

Smaller ε implies stronger privacy. In deep learning, DP is typically implemented [10,18,22,13] via DP-SGD [1], which clips per-sample gradients and injects Gaussian noise. While this ensures provable privacy, it perturbs optimization dynamics and often reduces predictive performance.

In medical imaging, this privacy-utility trade-off is almost exclusively evaluated through end-to-end task metrics such as AUROC or Dice [10,13,18,16,17]. Yet models are rarely used once; they are fine-tuned, transferred across institutions, or deployed as frozen feature extractors. End-to-end performance alone does not reveal whether privacy noise reduces linear separability, reshapes representation geometry, or primarily impairs optimization of the task head [5], so privacy model selection remains empirical rather than diagnostic. Representation geometry provides a principled perspective on this problem. The covariance spectrum of embeddings characterizes intrinsic dimensionality and anisotropy, and privacy-constrained optimization can induce structured spectral reshaping rather than uniform collapse [2]. What is missing is a framework connecting such geometric changes to downstream utility under DP.

We address this gap by introducing the **D**ifferential **P**rivacy **R**epresentation **G**eometry for **M**edical **I**maging (DP-RGMI) framework. DP-RGMI interprets DP training as a transformation of representation space and separates geometric change of the encoder from utilization by task head. Concretely, it quantifies (i) representation displacement from a shared pretrained initialization, (ii) spectral effective dimension of the learned embeddings, and (iii) a utilization gap defined as the difference between linear-probe AUROC and end-to-end private AUROC.

Algorithm 1 DP-RGMI workflow

Require: Probe regularization λ; utility metric $U(\cdot)$
Ensure: For each ε: $(U_{\text{end2end}}, U_{\text{probe}}, G, \Delta, d_{\text{eff}})$
for $\varepsilon \in \mathcal{E} \cup \{\infty\}$ **do**
  Train $(\phi_\varepsilon, h_\varepsilon)$ using DP-SGD with budget ε
  $U_{\text{end2end}}(\varepsilon) \leftarrow U(h_\varepsilon \circ \phi_\varepsilon; \mathcal{D}_{\text{test}})$
  $Z_\varepsilon \leftarrow [\phi_\varepsilon(x_i)]$, $Z_0 \leftarrow [\phi_0(x_i)]$ on $\mathcal{D}_{\text{test}}$
  $\Delta(\varepsilon) \leftarrow \frac{1}{N}\sum_{i=1}^{N} \|Z_\varepsilon[i] - Z_0[i]\|_2^2$
  $\Sigma_\varepsilon \leftarrow \frac{1}{N}(Z_\varepsilon - \mu_\varepsilon)^\top (Z_\varepsilon - \mu_\varepsilon)$
  $d_{\text{eff}}(\varepsilon) \leftarrow \frac{\text{tr}(\Sigma_\varepsilon)^2}{\text{tr}(\Sigma_\varepsilon^2)}$
  Train linear probe $\hat{h}_\varepsilon$ on frozen ϕ_ε
  $U_{\text{probe}}(\varepsilon) \leftarrow U(\hat{h}_\varepsilon \circ \phi_\varepsilon; \mathcal{D}_{\text{test}})$
  $G(\varepsilon) \leftarrow U_{\text{probe}}(\varepsilon) - U_{\text{end2end}}(\varepsilon)$
end for
return geometric diagnostic profile of DP training

2 DP-RGMI framework

We formalize DP as a transformation of representation space rather than only a scalar constraint on predictive performance. Given a pretrained encoder ϕ_0 and its differentially private counterpart ϕ_ε, DP-RGMI decomposes the analysis into three components (Fig. 1): representation displacement, spectral structure, and utilization, separating what remains encoded in ϕ_ε from how effectively it is exploited during private training. We consider a model factorized as $f_\varepsilon(x) = h_\varepsilon(\phi_\varepsilon(x))$, where $\phi_\varepsilon : \mathcal{X} \to \mathbb{R}^d$ is the encoder and h_ε a linear head. All geometric quantities are defined in the embedding space of ϕ_ε and compared to the fixed initialization ϕ_0, giving a shared coordinate system across privacy regimes. Algorithm 1 summarizes the workflow.

Representation displacement. Let $z_i^{(\varepsilon)} = \phi_\varepsilon(x_i)$ and $z_i^{(0)} = \phi_0(x_i)$ denote embeddings of the same test samples under private and initial encoders. We quantify representation displacement as:

$$\Delta(\varepsilon) = \frac{1}{N}\sum_{i=1}^{N} \|z_i^{(\varepsilon)} - z_i^{(0)}\|_2^2. \tag{2}$$

This measures how strongly DP-constrained optimization deviates from pretrained prior. Crucially, $\Delta(\varepsilon)$ captures geometric movement independently of task labels and isolates privacy-induced change from task-specific fitting. We use a Euclidean metric to keep Δ orthogonal to the spectral diagnostic: a covariance-aware metric (e.g., Mahalanobis) would whiten by the same covariance whose anisotropy d_{eff} characterizes, conflating the two.

Spectral structure. Let $\Sigma_\varepsilon = \frac{1}{N}\sum_{i=1}^{N}(z_i^{(\varepsilon)} - \mu_\varepsilon)(z_i^{(\varepsilon)} - \mu_\varepsilon)^\top$ denote embedding covariance with eigenvalues $\{\lambda_j\}$. We compute the effective dimension as:

$$d_{\text{eff}}(\varepsilon) = \frac{\left(\sum_j \lambda_j\right)^2}{\sum_j \lambda_j^2}. \tag{3}$$

This quantity summarizes spectral concentration and anisotropy. Changes in d_{eff} reflect how DP reshapes variance distribution across principal directions rather than merely translating embeddings [2].

Utilization. To decouple intrinsic separability from private joint optimization, we freeze ϕ_ε and train a regularized linear probe. Probe utility U_{probe} measures linear recoverability of class structure in the embedding. The utilization gap is defined as:

$$G(\varepsilon) = U_{\text{probe}}(\varepsilon) - U_{\text{end2end}}(\varepsilon), \tag{4}$$

which quantifies performance loss attributable to optimization under DP rather than representational collapse. In this study $U = \text{AUROC}$, but the definition is metric-agnostic. A large $G(\varepsilon)$ indicates that discriminative structure persists in ϕ_ε but is not fully exploited during private training.

DP-RGMI is model-agnostic and dataset-agnostic: it requires only access to embeddings and standard evaluation metrics.

3 Experimental setup

Data. We study multi-label chest X-ray (CXR) classification on PadChest [4] (110,525 frontal images from 67,205 patients) as the primary dataset, and use an additional 269,796 images from other public CXR datasets for generalization. PadChest was chosen for primary analysis as it provides binary presence/absence annotations, the most radiologist-annotated labels among available datasets, and sufficient size for stable geometric evaluation. As no official split exists, we construct a fixed *patient-stratified* partition into training, validation, and test sets. All geometric and utility analyses are performed exclusively on the held-out test set (22,045 images). We focus on five common findings: atelectasis, cardiomegaly, pleural effusion, pneumonia, and no finding. Images are resized to 224×224, intensity-normalized, and contrast-standardized following prior work [16,3]. Class imbalance is handled via label-wise loss weighting. Code and pretrained weights are publicly available [4].

Model and training. We use ConvNeXt-Small [11] (49M parameters, embedding dimension $d = 768$) with a linear multi-label head. ConvNeXt avoids

[4] https://github.com/tayebiarasteh/CXR-adaptation, and weights are available from HuggingFace.

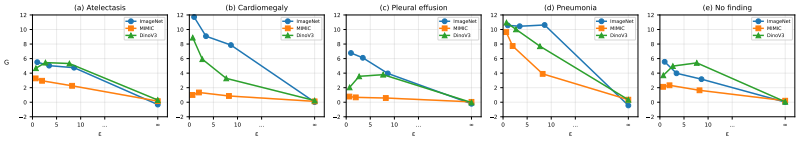

Fig. 2. Per-label utilization gaps $G(\varepsilon)$ for different ε on the PadChest dataset.

batch normalization, which is incompatible with the per-sample gradients required by DP-SGD, and remains stable under gradient clipping and additive noise; convolutional backbones have predominated in DP-SGD medical imaging, whereas transformers [20] show unstable DP optimization [13]. Models are optimized with AdamW (weight decay 0.01). Non-private training uses learning rate 10^{-5}, batch size 128, weighted binary cross-entropy loss, and light augmentation (random horizontal flips and small rotations).

DP training. Private runs use DP-SGD without data augmentation [13,18]. Per-sample gradients $g_i = \nabla_\theta \ell(\theta; x_i, y_i)$ are clipped to ℓ_2 norm C via $\bar{g}_i = g_i \cdot \min\left(1, \frac{C}{\|g_i\|_2}\right)$ and perturbed with Gaussian noise, $\tilde{g} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \bar{g}_i + \mathcal{N}(0, \sigma^2 C^2 I)$, followed by the update $\theta \leftarrow \theta - \eta \tilde{g}$, with clipping norm $C = 4$. Training uses Poisson subsampling: each example is independently included in a batch with probability $q = 128/|\mathcal{D}_{\text{train}}|$, consistent with privacy accounting. A Rényi DP accountant [12] tracks (ε, δ) with $\delta = 6 \times 10^{-6}$ fixed, which satisfies $\delta < 1/|\mathcal{D}_{\text{train}}|$ for all datasets; ε is controlled by adjusting the noise multiplier σ. Each initialization branch includes a non-private baseline ($\varepsilon = \infty$) and 3 decreasing privacy budgets, all within $\varepsilon < 10$, a commonly adopted privacy range in medical imaging studies as a private model [13]. All private runs use a maximum budget of 150 epochs; since the accountant accumulates privacy expenditure per step, each run is trained until convergence and the reported ε corresponds to its selected checkpoint, which differs across runs. Privacy guarantees are applied at the image level; since training samples correspond to individual radiographs, the formal guarantee applies per image rather than per patient [16,18,13].

Initialization regimes. Recent studies consistently highlight the critical role of initialization in DP-SGD training for medical imaging [3,13]. To analyze initialization-dependent geometric responses under DP, we consider three pretrained encoders: (i) supervised ImageNet [6] initialization as a generic baseline, (ii) self-supervised DINOv3 [15] initialization representing modern foundation models, and (iii) domain-specific initialization pretrained on MIMIC-CXR [9] (213,921 frontal images), the largest publicly available CXR dataset to date, using identical preprocessing and label space as the downstream task. The architecture and all non-privacy hyperparameters (optimizer, learning rate schedule, batch size, epochs) are fixed across privacy levels and initializations.

Table 1. Overall results on the PadChest dataset, computed by paired bootstrap on the test set (1000 resamples), reported as mean ± standard deviation.

Initialization	ε	$\text{AUROC}_{\text{end2end}}$	$\text{AUROC}_{\text{probe}}$	G	Δ	d_{eff}
ImageNet	∞	88.8 ± 0.2	88.6 ± 0.2	-0.2 ± 0.3	1.9 ± 0.0	7.8 ± 0.1
	8.6	76.6 ± 0.3	82.7 ± 0.2	6.1 ± 0.4	1.1 ± 0.0	3.4 ± 0.0
	3.5	74.9 ± 0.3	81.9 ± 0.3	6.9 ± 0.4	1.0 ± 0.0	4.7 ± 0.0
	1.0	74.5 ± 0.3	82.5 ± 0.3	8.0 ± 0.4	1.1 ± 0.0	9.2 ± 0.1
MIMIC	∞	90.0 ± 0.2	90.2 ± 0.2	0.2 ± 0.3	0.1 ± 0.0	3.3 ± 0.0
	8.2	85.8 ± 0.2	87.6 ± 0.2	1.9 ± 0.3	1.4 ± 0.0	4.4 ± 0.0
	2.0	84.4 ± 0.3	87.4 ± 0.2	3.0 ± 0.3	1.4 ± 0.0	5.4 ± 0.0
	0.7	83.9 ± 0.2	87.2 ± 0.2	3.4 ± 0.3	1.3 ± 0.0	5.5 ± 0.0
DINOv3	∞	89.5 ± 0.2	89.7 ± 0.2	0.2 ± 0.3	0.7 ± 0.0	2.8 ± 0.0
	7.7	77.4 ± 0.3	82.5 ± 0.3	5.1 ± 0.4	1.9 ± 0.0	5.1 ± 0.0
	2.7	76.0 ± 0.3	81.9 ± 0.3	6.0 ± 0.4	1.9 ± 0.0	4.6 ± 0.0
	0.7	75.6 ± 0.3	81.6 ± 0.3	6.1 ± 0.4	1.6 ± 0.0	3.9 ± 0.0

Statistical estimation. Uncertainty is estimated via nonparametric bootstrap over test samples ($B = 1000$) [14]. Within each initialization branch, we compute rank correlations between $\text{AUROC}_{\text{end2end}}$ and geometric statistics across privacy budgets. For each configuration this yields ($\text{AUROC}_{\text{end2end}}$, $\text{AUROC}_{\text{probe}}$, $\Delta(\varepsilon)$, $d_{\text{eff}}(\varepsilon)$), forming the basis of the representation-level analysis. All reported classification metrics are expressed in percent.

4 Results

DP-RGMI decomposes privacy degradation into separability and utilization. Table 1 summarizes the results. As expected, $\text{AUROC}_{\text{end2end}}$ decreases under privacy across all initializations (ImageNet: 88.8 → 76.6 → 74.5; DINOv3: 89.5 → 77.4 → 75.6; MIMIC: 90.0 → 85.8 → 83.9 as ε decreases). However, DP-RGMI asks *where* the degradation arises.

Under non-DP, $G(\infty) \approx 0$ for all initializations, indicating that joint training largely realizes the linearly recoverable structure in ϕ_∞. Under DP, probe AUROC remains consistently higher than $\text{AUROC}_{\text{end2end}}$, yielding large gaps at strong privacy: $G = 8.0$ (ImageNet, $\varepsilon = 1.0$), 3.4 (MIMIC, $\varepsilon = 0.7$), and 6.1 (DINOv3, $\varepsilon = 0.7$). This implies that DP can preserve substantial linear separability in ϕ_ε while impairing its utilization during joint DP training.

To test whether the utilization gap reflects a coherent failure mode rather than random degradation, Fig. 2 shows per-label AUROC results. DP shifts AUROC downward across labels while broadly preserving relative difficulty. This supports a global DP-induced transformation rather than selective erasure of a single label-specific axis. In contrast, the utilization gap becomes sharply label-dependent. Under ImageNet at $\varepsilon = 1.0$, pneumonia exhibits a +10.6 AUROC gap, whereas no finding has a smaller +5.6 gap. Under MIMIC at $\varepsilon = 0.7$, gaps are smaller for most labels overall, although pneumonia still shows a relatively large gap (+9.6). Under DINOv3 at $\varepsilon = 0.7$, pneumonia again shows a large gap (+11.0), indicating that utilization failure is not specific to supervised or self-supervised pretraining, but depends on how DP-constrained optimization interacts with initialization and label geometry.

Table 2. Spearman rank correlation ρ with $\text{AUROC}_{\text{end2end}}$ for DP models ($\epsilon < 10$). Across inits: $n = 3 \times 3$ (ϵ, datasets); across datasets: $n = 3 \times 3$ (ϵ, inits).

Setting	n	$\rho(\text{AUROC}_{\text{end2end}}, G)$	$\rho(\text{AUROC}_{\text{end2end}}, \Delta)$	$\rho(\text{AUROC}_{\text{end2end}}, d_{\text{eff}})$
ImageNet init	9	-0.78	-0.43	-0.33
MIMIC init	9	-0.31	+0.81	+0.49
DINOv3 init	9	+0.55	+0.15	+0.52
PadChest dataset	9	-0.95	+0.32	-0.07
CheXpert dataset	9	-0.86	-0.39	-0.02
ChestX-ray14 dataset	9	-0.98	-0.23	-0.43
Overall	**27**	**-0.61**	**-0.12**	**-0.07**

Geometry under DP. DP-RGMI attributes the remaining variation in performance to changes in representation geometry. Table 1 shows that DP induces measurable displacement $\Delta(\varepsilon)$ and reshaping of spectral structure through $d_{\text{eff}}(\varepsilon)$, with patterns that depend on initialization.

Displacement. Under DP, all initializations move away from their pretrained prior, but to different extents. DINOv3 exhibits the largest drift (e.g., $\Delta = 1.9$ at $\varepsilon = 7.7$), ImageNet shows moderate but consistent displacement ($\Delta \approx 1.0$–1.1), and MIMIC transitions from near-zero movement without privacy ($\Delta = 0.1$) to substantial displacement under DP ($\Delta \approx 1.3$–1.4). Importantly, displacement magnitude does not map monotonically to utility. For example, configurations with similar AUROC can correspond to different Δ values, indicating that geometric departure from initialization alone does not determine task performance.

Spectral reshaping. Changes in $d_{\text{eff}}(\varepsilon)$ are non-monotonic and initialization-dependent. Under ImageNet, d_{eff} decreases at moderate privacy (3.4 at $\varepsilon = 8.6$) but increases at stronger privacy (9.2 at $\varepsilon = 1.0$). In contrast, DINOv3 trends toward lower effective dimension as privacy strengthens ($5.1 \rightarrow 3.9$), while MIMIC exhibits a gradual increase under DP. These heterogeneous trajectories argue against a uniform representation collapse. Instead, DP induces structured spectral transformations whose direction depends on the pretrained prior. Geometry therefore provides context for how privacy reshapes embeddings, while the utilization gap identifies where performance is lost.

Cross-dataset generalization and correlation structure. We next examine the generalization beyond PadChest. The full protocol is repeated on CheXpert [8] (total of 157,676 frontal images) and ChestX-ray14 [21] (total of 112,120 frontal images) datasets under identical data partitioning ($15 - 20\%$ patient-stratified held-out test sets), δ, initialization regimes, training settings, and statistical estimation, with comparable privacy budgets. Datasets are evaluated as multi-label classification for the same 5 labels, with macro-averaged analysis.

Fig. 3 shows the trajectories of $G(\varepsilon)$, $\Delta(\varepsilon)$, and $d_{\text{eff}}(\varepsilon)$ for both generalization datasets. Despite differences in dataset size and baseline AUROC, a consistent pattern emerges: under stronger privacy, probe AUROC remains higher than $\text{AUROC}_{\text{end2end}}$ across initializations, yielding a positive $G(\varepsilon)$. In contrast, $\Delta(\varepsilon)$ and $d_{\text{eff}}(\varepsilon)$ follow dataset- and initialization-specific trajectories rather than exhibiting uniform degradation. To quantify these relationships, we compute Spear-

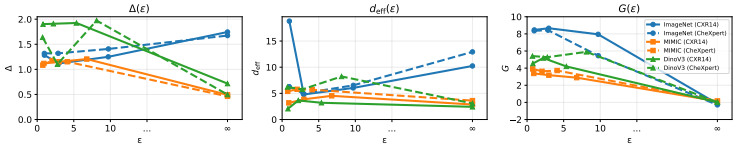

Fig. 3. Generalization results on CheXpert and ChestX-ray14 datasets.

man rank correlations between $\text{AUROC}_{\text{end2end}}$ and DP-RGMI quantities for DP models (Table 2). Across initializations, the association between AUROC and G is negative for ImageNet ($\rho = -0.78$) but becomes weak or reverses sign for MIMIC ($\rho = -0.31$) and DINOv3 ($\rho = +0.55$), indicating that the monotonic relationship is initialization-dependent within the DP regime. In contrast, geometric associations can be substantial for some priors (e.g., MIMIC: $\rho = +0.81$ with Δ), suggesting that geometry captures prior-conditioned variation not explained by G alone. Across datasets, AUROC remains negatively associated with G (PadChest: $\rho = -0.95$, CheXpert: $\rho = -0.86$, ChestX-ray14: $\rho = -0.98$), while correlations with Δ and d_{eff} remain dataset-specific. Overall, the association between AUROC and G is moderate ($\rho = -0.61$), whereas correlations with Δ ($\rho = -0.12$) and d_{eff} ($\rho = -0.07$) are weak. Note that the association with G partly reflects its definition relative to AUROC and is interpreted descriptively rather than causally.

5 Discussion and conclusion

We reframed DP evaluation in CXR classification as a representation-level diagnostic problem. Instead of relying solely on end-to-end performance, DP-RGMI separates encoder geometry from downstream utilization. Strong privacy is consistently associated with a utilization gap, while the strength of this relationship can vary by initialization and geometry can explain additional prior- and dataset-conditioned variation.

This separation supports concrete deployment decisions. If two privacy budgets yield similar AUROC but one exhibits a larger G, DP-RGMI indicates that recoverable signal persists, pointing to optimization-side remedies (freezing the encoder, retraining only the head, or adjusting clipping for head parameters) as candidate interventions to be tested rather than relaxing privacy. Our linear-probe measurement already realizes a controlled instance of this: $U_{\text{probe}} > U_{\text{end2end}}$ under DP (Table 1) shows the recoverable structure is exploitable once the head is trained on a frozen private encoder, and quantifying the privacy cost of such head-only re-optimization is direct future work. If Δ is large while probe performance remains stable, representation has moved substantially from its pretrained prior, which may affect transfer or reuse across institutions even when classification performance appears acceptable. Conversely, marked reductions in d_{eff} indicate increased spectral concentration and reduced representational diversity, potentially limiting adaptation to new tasks. In such

cases, revisiting pretraining or privacy strength may be more appropriate than head-level adjustments.

In this study, all experiments use multi-label chest X-ray classification with a single ConvNeXt-Small backbone. This choice is constrained rather than free: batch normalization is incompatible with the per-sample gradients of DP-SGD, and transformers exhibit unstable DP optimization [13], while cross-architecture consistency of DP trade-offs has been reported separately [13,16,18,22]. The framework is model-agnostic by construction; its behavior in other tasks such as segmentation remains to be validated, and we expect similar geometry-utilization interactions wherever representations are reused or fine-tuned.

Overall, DP-RGMI provides a reproducible framework for diagnosing privacy-induced failure modes and guiding principled privacy model selection in cases with cross-institutional reuse, transfer learning, or frozen-feature deployment.

Acknowledgments. STA is supported by the Excellence Strategy of the German Federal Government, the Länder, and RWTH ERS (START_526-26). SN is supported by the DFG (701010997, 517243167). DT is supported by the German Ministry of Research, Technology and Space (TRANSFORM LIVER - 031L0312C, DECIPHER-M - 01KD2420B), DFG (515639690), and the European Union (Horizon Europe, ODELIA - GA 101057091, ERC Starting Grant SAGMA – GA 101222556).

Disclosure of Interests. DT received honoraria for lectures by Bayer, GE, Roche, AstraZeneca, and Philips and holds shares in StratifAI GmbH and Synagen GmbH, Germany. The other authors declare no competing interests relevant to this work.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: SIGSAC 2016. pp. 308–318
2. Ansuini, A., Laio, A., Macke, J.H., Zoccolan, D.: Intrinsic dimension of data representations in deep neural networks. In: NeurIPS 2019. vol. 32
3. Arasteh, S.T., Farajiamiri, M., Lotfinia, M., et al.: The role of self-supervised pretraining in differentially private medical image analysis. arXiv preprint arXiv:2601.19618 (2026)
4. Bustos, A., Pertusa, A., Salinas, J.M., De La Iglesia-Vaya, M.: Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical image analysis **66**, 101797 (2020)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML 2020
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR 2009. pp. 248–255
7. Dwork, C., Roth, A.: The algorithmic foundations of differential privacy. foundations and trends® in theoretical computer science 9 (3-4), 211–407 (2014)
8. Irvin, J., Rajpurkar, P., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019)

9. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data **6**, 317 (2019)
10. Kaissis, G., Ziller, A., Passerat-Palmbach, J., Ryffel, T., Usynin, D., Trask, A., Lima Jr, I., Mancuso, J., Jungmann, F., Steinborn, M.M., et al.: End-to-end privacy preserving deep learning on multi-institutional medical imaging. Nat Mach Intell **3**(6), 473–484 (2021)
11. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: CVPR 2022. pp. 11976–11986
12. Mironov, I.: Rényi differential privacy. In: 2017 IEEE 30th CSF. pp. 263–275
13. Mohammadi, M., Vejdanihemmat, M., Lotfinia, M., Rusu, M., Truhn, D., Maier, A., Tayebi Arasteh, S.: Differential privacy for medical deep learning: methods, tradeoffs, and deployment implications. npj Digit. Med. **9**, 93 (2026)
14. Mooney, C.Z., Duval, R.D.: Bootstrapping: A Nonparametric Approach to Statistical Inference. Quantitative Applications in the Social Sciences (1993)
15. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
16. Tayebi Arasteh, S., Lotfinia, M., Nolte, T., Sähn, M.J., Isfort, P., Kuhl, C., Nebelung, S., Kaissis, G., Truhn, D.: Securing collaborative medical ai by using differential privacy: Domain transfer for classification of chest radiographs. Radiology: Artificial Intelligence **6**(1), e230212 (2023)
17. Tayebi Arasteh, S., Lotfinia, M., Perez-Toro, P.A., et al.: Differential privacy enables fair and accurate ai-based analysis of speech disorders while protecting patient data. npj Artif. Intell. **1**, 37 (2025)
18. Tayebi Arasteh, S., Ziller, A., Kuhl, C., Makowski, M., Nebelung, S., Braren, R., Rueckert, D., Truhn, D., Kaissis, G.: Preserving fairness and diagnostic accuracy in private large-scale ai models for medical imaging. Commun Med **4**(1), 46 (2024)
19. Usynin, D., Ziller, A., Makowski, M., Braren, R., Rueckert, D., Glocker, B., Kaissis, G., Passerat-Palmbach, J.: Adversarial interference and its mitigations in privacy-preserving collaborative machine learning. Nat Mach Intell **3**(9), 749–758 (2021)
20. Vaswani, A.: Attention is all you need. NeurIPS 2017 **30**
21. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: CVPR 2017. pp. 2097–2106
22. Ziller, A., Mueller, T.T., Stieger, S., Feiner, L.F., Brandt, J., Braren, R., Rueckert, D., Kaissis, G.: Reconciling privacy and accuracy in ai for medical imaging. Nat Mach Intell **6**(7), 764–774 (2024)