# Reconstructing Training Images from Foundation Model Parameters in the Healthcare Domain: Privacy Risks and Defences

**Anonymous Authors**[1]

## Abstract

Recent studies have shown that it is possible to reconstruct training images from finetuned foundation model parameters alone: first by reconstructing the embeddings, and then by using model inversion to invert these embeddings to the image domain. This could pose a privacy risk for healthcare applications. Yet whether this risk actually transfers to the healthcare domain, where images differ substantially from the general images used in previous works, remains unknown. In this work, we systematically evaluate this risk across three healthcare domains, seven datasets, and sixteen foundation models (twelve medically specialized and four general models). We find that embedding reconstruction attack success is strongly domain-dependent: on average, X-ray images are more vulnerable to embedding reconstruction attacks than pathology or ophthalmology images. However, unlike in the general image domain, current inversion techniques do not lead to recognizable images reconstructed from these embeddings for any of the considered healthcare domains. We thus estimate the current risk to be low. Crucially, we also find that linear probing consistently neutralizes the complete attack, thereby suggesting a readily deployable and future-proof defense.

## 1. Introduction

In healthcare applications, it is becoming more and more common to finetune foundation models pretrained on large, open-source general or medical datasets with privacy-sensitive, institutional datasets (Kim et al., 2022; Huang et al., 2023; Khan et al., 2025). If sensitive patient images used during finetuning can be recovered from the resulting model parameters, then sharing or deploying finetuned models may inadvertently expose private data. But while the possibility of this kind of attacks has been demonstrated in general settings (Haim et al., 2022; Buzaglo et al., 2023; Oz et al., 2024) (i.e., using general models such as DINO (Caron et al., 2021) on general image datasets such as Imagenet (Deng et al., 2009)), it is unclear how high the risk is when using medical images and medically specialized foundation models. Medical images of the same modality are namely less diverse in terms of colors and structure than natural images (figure 2), and often the important parts are small abnormalities and very local patterns, such as bleeding and inconsistent structures (Huang et al., 2023; Xu et al., 2024). In this work, we therefore systematically evaluate the success of the attacks over different healthcare domains, using seven datasets and sixteen foundation models in total. The proposed attack itself was developed over a series of papers (Haim et al., 2022; Buzaglo et al., 2023; Oz et al., 2024) based on theoretical results about the implicit bias of gradient descent (see methods). The most recent work, by Oz et al., focuses on large foundation models. For foundation models, the attack consists of two stages (see fig. 1): first, embeddings are reconstructed directly from the parameters of the fine-tuning head; and second, the reconstructed embeddings are inverted back to the pixel domain via model inversion, yielding approximations of the original images. It is crucial to note that this attack only uses knowledge of the model parameters, not of the model outputs or training data itself. We here compare the obtained reconstructions to the dataset to assess the attack, but in (Oz et al., 2024), the authors also propose a method to identify good reconstructions without access to the original trainset. We first focus on the reconstruction of the embeddings, and find that the success of this part of the attack:

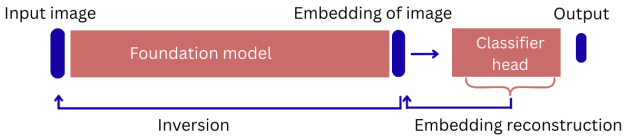



*Figure 1.* The attack consists of two stages: the embedding reconstruction from the parameters of the head, and the subsequent reconstruction of the input image through model inversion starting from the reconstructed embedding.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by *The Impact of Memorization on Trustworthy Foundation Models* Workshop @ ICML. Do not distribute.

- depends on the characteristics of the fine-tuning procedure and classification head (training time, number of samples, weight decay, number of neurons and initialization scale) in the same way as in the original setting and in line with theory on memorization.
- depends on the use of nonlinearities in the head. For linear probing (thus the use of a linear head), the attack does not succeed.
- differs consistently between domains, where the average number of reconstructed embeddings is higher in the X-ray than in the other considered domains (ophthalmology and pathology). This conclusion largely holds even if we split models between foundation models specialized for the domain and general image foundation models. There is no clear relationship between success of the attack and image resolution, embedding size or generalization gap (train - test).

Finally, we find that the inversion of reconstructed embeddings to the image domain is less successful when using healthcare data, when compared to the image reconstructions of general images in the original studies.

**Related Work** Depending on the information an adversary possesses, such as shared gradients, model architectures, model outputs, or even training data, several different privacy attacks have been explored. Federated learning techniques allow for attacks where an adversary can exploit shared gradients to reconstruct patient data (Zhao et al., 2024; Fan et al., 2025; Petrov et al., 2024), or use model inversion to reverse-engineer original samples from outputs or internal representations (Peng et al., 2024; Islam et al., 2025; Ye et al., 2025). Membership inference attacks can determine whether a medical exam was included in a model's training set (Ye et al., 2025; Wang et al., 2026; Nguyen et al., 2025; Hagestedt et al., 2020). We here focus on input image reconstruction attacks as developed and applied for general images in (Haim et al., 2022; Buzaglo et al., 2023; Oz et al., 2024). These attacks extract general training images from the parameters of neural networks alone.

## 2. Experimental Setup

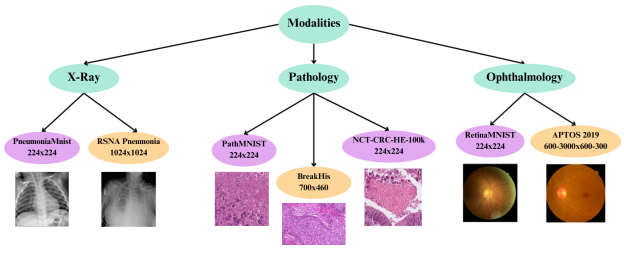



*Figure 2.* Overview of datasets

### 2.1. Overview

The considered setup consists of a frozen foundation model, pre-trained on general or medical data (see 'models'), together with a task-specific head for binary classification. Only the parameters of the head are updated during finetuning. After finetuning on a specific task, we applied an embedding reconstruction attack as described below; and for a small subset of successfully reconstructed embeddings, we also performed the inversion back to the image domain.

**Embedding Reconstruction Attack** To reconstruct embeddings, we use the method originally proposed in (Haim et al., 2022). This method is based on a theoretical relationship between the input dataset and the model parameters of a homogeneous network trained with binary classification: $\tilde{\theta} = \sum_{i=1}^{n} \lambda_i y_i \nabla_\theta \Phi(\tilde{\theta}; x_i)$ that can be used to optimize randomly initialized $\{x_i, \lambda_i\}$ to correspond to the initial model training data, according to a loss given by $\mathcal{L}_{\text{rec}}(\hat{x}_{1:m}, \lambda_{1:m}) := \left\|\theta - \sum_{i=1}^{m} \lambda_i y_i \nabla_\theta \left(\phi(\hat{x}_i, \theta)\right)\right\|_2^2$, In (Oz et al., 2024), this is applied to the reconstruction of embeddings in the context of foundation models.

**Inversion** The reconstructed embeddings can then be inverted back to the image space. In (Oz et al., 2024), the authors utilized the approach proposed in (Tumanyan et al., 2022) for foundation models, such as DINOv2 and ViT, and unCLIP (Donghoon Lee & Kim, 2022), for the CLIP model, where DIP was not effective. The former treats inversion as an optimization problem, utilizing the [CLS] token's powerful representation of visual appearance and global information, such as object parts. Following the deep image prior (DIP) technique the method optimizes the parameters of a U-net model to generate an image which, when passed through the foundation model, produces an embedding vector maximally similar to the reconstructed candidate embedding. The diffusion decoder unCLIP implementation of (Donghoon Lee & Kim, 2022) was deployed, to create an image that matches the target embedding. However, due to poor performance, no results are included.

**Datasets** We used 7 image datasets over three distinct medical imaging domains (X-Rays, Ophthalmology images, and Pathology images of different dimensionalities), shown in fig. 2. Importantly, these do not overlap with the pre-training dataset of any foundation model we used. More information, see appendix A.

**Models** We used 3 categories of models: *Multi-modal models* that were trained on various distinct medical modalities (including MedSigLIP (Sellergren et al., 2026), Biomed-CLIP (Zhang et al., 2025), and LVM-Med (Nguyen et al., 2023)), *Modality-specific models* tailored for Chest X-rays (BioViL-T (Bannur et al., 2023), CheXzero (Tiu et al., 2022), MedCLIP (Wang et al., 2022)), for Pathology (UNI (Chen et al., 2024), Hibou-b (Nechaev et al., 2024), H0-mini (Filiot et al., 2025), Virchow2 (Zimmermann et al., 2024)), and for Ophthalmology (RETFound (Zhou et al., 2023)

and RET-CLIP (Du et al., 2024)). We will refer to these models, together with the specific multi-modal models, as 'specialized models'. The last category are the *natural image models* (ViT (Dosovitskiy et al., 2021), DinoV2 (Oquab et al., 2024), CLIP (Radford et al., 2021) and MAE (He et al., 2021)) that were trained on natural image datasets to serve as non-medical baselines. We will refer to these as the general models.

### 2.2. Experiment Details

**Finetuning** Unless stated otherwise, experiments utilized datasets of 100 samples, preprocessed to align with each foundation model's requirements. Feature embeddings were extracted as [CLS] token for ViT, CLIP, DINOv2, Biomed-CLIP, Chexzero, MedCLIP, RetCLIP, RETFound, H0-mini, Hibou-b, UNI, and Virchow2, or as global average pooling (GAP) embeddings for MAE, Biovil-T, LVM-MED, and MedSigLip. Unless stated otherwise, the classifier models (MLP of one 500-neuron hidden layer) were trained for 10,000 epochs for different values of weight decay (1e-5, 1e-4, 0.001, 0.01, 0.1, 0.13). Over these values of weight decay, the models that achieved the highest testing accuracy were selected for the reconstruction attack.

**Embedding Reconstruction** Each reconstruction attack was repeated 100 times. We used $m = 500$ for datasets containing $N = 100$ samples; for larger datasets we used $m = 2N$. Additionally, the number of successfully reconstructed training samples was calculated as the number of unique training samples for which a reconstructed embedding achieved a cosine similarity greater than 0.75 (information over the complete distributions is added in the appendix F). Cosine similarity scores were calculated using the FAISS library (Douze et al., 2025).

**Inversion** For the inversion, a U-Net model was optimized using the Adam optimizer (lr=0.01) for 20.000 iterations. Following the methods proposed in(Oz et al., 2024), the model's input was a fixed random tensor augmented with gaussian noise ($\sigma$ scaled by 10, 2, and 0.5 at iteration thresholds 10k, 15k, and 20k, respectively), while the output was preprocessed to align with the foundation model before feature extraction. The loss function deployed is the cosine similarity as the authors observed that it performs better across varying embedding scales. For the quantitative inversion comparison, SSIM (Wang et al., 2004) and PSNR were used. More information can be found in the appendix G.

## 3. Results and Discussion

### 3.1. Characteristics of the Finetuning Procedure

We first determined the influence of several hyperparameters used during the finetuning procedure (training time, number of samples, and number of neurons) on the success of the embedding reconstruction attack for a subset of models. The results are in line with previous findings in the original studies (Oz et al., 2024): **Training time:** there is a minimal amount of training time needed to allow for reconstructions, although for very long training time the success of the attacks declines. **Number of samples:** the attack remains succesful when larger datasets (we tested up to N=10000), but the relative number of succesful reconstructions goes down. **Number of neurons:** the more neurons in the classification head, the more succesful reconstructions. More details can be found in the appendix C.

### 3.2. Linear vs. Nonlinear Head

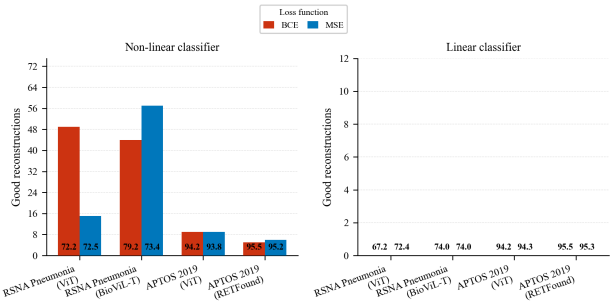



*Figure 3.* Number of reconstructions for nonlinear heads (left) vs. linear heads (right), for different datasets and models (see labels) and two different losses. The numbers above the x-axis correspond to classification accuracy.

It has been shown, also in the healthcare context, that using a linear head (thus removing the activation functions) instead of a nonlinear head can yield satisfying results (Ambsdorf et al., 2025; Zhang et al., 2025; Nguyen et al., 2023). As linear models have a lower capacity to memorize, we tested to what degree we can reconstruct embeddings from linear heads for a subset of datasets and models. We have repeated this experiment for MSE and cross-entropy loss, as the latter adds a non-linear softmax function to the head. Fig. 3 shows that using a linear head eliminates the ability to reconstruct good embeddings, while at the same time showing a classification performance on par with the nonlinear heads.

### 3.3. Characteristics of the Foundation Model & Domain

In fig. 4, we show the number of good reconstructions (cosine similarity ¿0.75) grouped per domain (X-ray vs. Pathology vs. Ophthalmology). Information on the complete distribution of the cosine similarities over all experiments can be found in the appendix F. The average number of good reconstructions differs significantly between the domains when we consider all models together (left plot, Kruskal-Willis test: p=0.0563 for different means). A difference in means can still be observed when we consider the specialized models (middle plot, p=0.0597) and the general models (right plot, p=0.1883) separately. This result

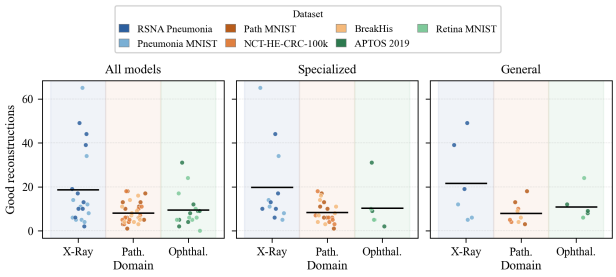



*Figure 4.* Number of good reconstructions grouped per data domain (X-ray, pathology, ophthalmology) for all models combined (left), models specialized in the given domain (middle), and general foundation models (right).

is mainly driven by the high number of good reconstructions for the specialized models BioViL-T and ChexZero and general models DinoVIT and ViT on the 2 used X-ray datasets (RSNA Pneunominia and Pneumonia MNIST). The result indicates that the ability to reconstruct embeddings is, on average, tied to the dataset domain. To make sure the result cannot be explained by another factor (e.g., the dimension of the X-ray dataset as compared to the other datasets), we identified a number of model properties that might influence the embedding reconstruction risk: performance on the test set and generalization gap (overfitting), embedding dimension, model input resolution and selected weight decay value (detailed results see appendix D). We did not find a clear relationship between any of these properties and the risk. The high risk for the aforementioned combination of specific models and dataset domains can thus not be explained by these properties.

### 3.4. Inversion

Before assessing the inversion techniques on reconstructed embeddings, the techniques were tested on the embeddings of the original training images first. Tests across four datasets (RSNA Pneumonia, Penumonia MNIST, APTOS 2019, and Retina MNIST), as Figure 5 depicts, revealed that the inversion scheme deployed by Oz et al. fails to invert medical image embeddings in high quality for most of the models. However, the embedding inversion of the RETFound model produces more visually similar results. Figure 6 then compares the inversions obtained for the original training sample embeddings and the inversions obtained from their reconstructed embeddings, for two samples with highly succesful embedding reconstructions(from the APTOS 2019 and Retina MNIST datasets using the RETFound model, 0.91 and 0.94 cosine similarity scores, respectively, see section 3.3). Quantitatively, there is a noticeable decrease in pixel and structural accuracy, as both PSNR and SSIM values drop, despite the high cosine similarity between the original and reconstructed embeddings. More inversion can be found in the appendix G.

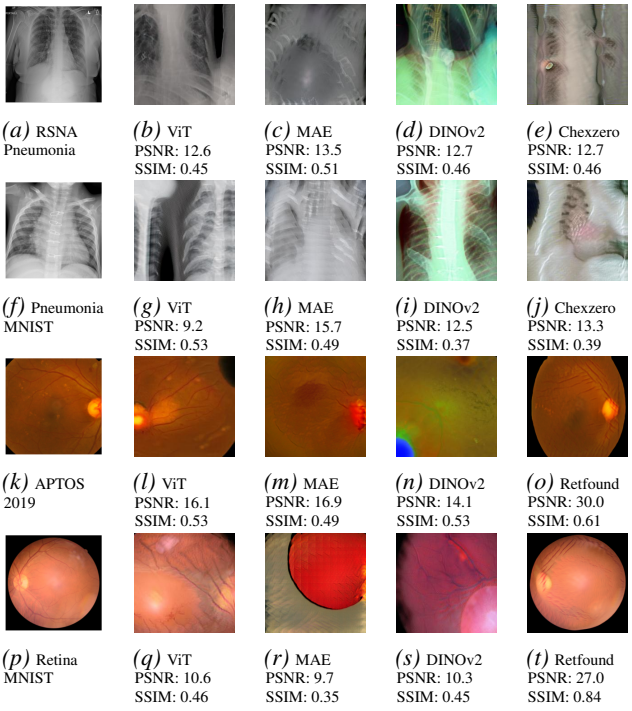



*Figure 5.* Inversions of embeddings from original training images for different datasets and models.

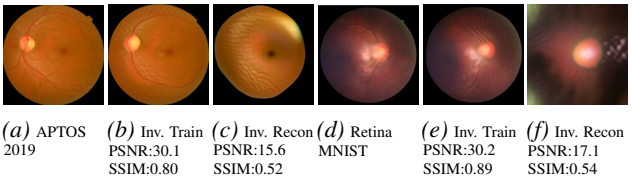



*Figure 6.* Comparison between original images, inversions of their embeddings, and inversions from their reconstructed embeddings.

## 4. Conclusion

Adapting the embedding reconstruction attack to the medical domain demonstrates its success across various foundation models and three tested medical imaging subdomains, with X-ray images showcasing higher vulnerability. But while embeddings can be reconstructed with relatively high cosine similarity, the overall privacy risk, in most cases, seems low, as the current inversion techniques fail to produce images of high visual quality. We thus estimate the risk of the total privacy attack succeeding in the context of healthcare domain, with current techniques, to be relatively low. Crucially, we also determined a pathway to defend against the attack that is not tied to current methods: at least in our set of experiments, deploying linear probing instead of using nonlinear classifiers eliminates the possibility of embedding reconstructions completely. Our work thus serves as a starting point for further studies on the risk of and defense against this type of privacy attack, both in the healthcare as in other domains.

## References


Ambsdorf, J., Munk, A., Llambias, S., Christensen, A. N., Mikolaj, K., Balestriero, R., Tolsgaard, M. G., Feragen, A., and Nielsen, M. General Methods Make Great Domain-specific Foundation Models: A Case-study on Fetal Ultrasound . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15966. Springer Nature Switzerland, September 2025.

Bannur, S., Hyland, S., Liu, Q., Pérez-García, F., Ilse, M., Castro, D. C., Boecking, B., Sharma, H., Bouzid, K., Thieme, A., Schwaighofer, A., Wetscherek, M., Lungren, M. P., Nori, A., Alvarez-Valle, J., and Oktay, O. Learning to exploit temporal structure for biomedical vision-language processing, 2023. URL https://arxiv.org/abs/2301.04558.

Buzaglo, G., Haim, N., Yehudai, G., Vardi, G., Oz, Y., Nikankin, Y., and Irani, M. Deconstructing data reconstruction: Multiclass, weight decay and general losses, 2023. URL https://arxiv.org/abs/2307.01827.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers, 2021. URL https://arxiv.org/abs/2104.14294.

Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F., Jaume, G., Chen, B., Zhang, A., Shao, D., Song, A. H., Shaban, M., et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024.

Dekhil, O., Naglah, A., Shaban, M., Ghazal, M., Taher, F., and Elbaz, A. Deep learning based method for computer aided diagnosis of diabetic retinopathy. In *2019 IEEE International Conference on Imaging Systems and Techniques (IST)*, pp. 1–4, 2019. doi: 10.1109/IST48021.2019.9010333.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Denker, A., Schmidt, M., Leuschner, J., and Maass, P. Conditional invertible neural networks for medical imaging, 2021. URL https://arxiv.org/abs/2110.14520.

Donghoon Lee, Jiseob Kim, J. C. J. K. M. B. W. B. and Kim, S. Karlo-v1.0.alpha on coyo-100m and cc15m. https://github.com/kakaobrain/karlo, 2022.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., and Jégou, H. The faiss library, 2025. URL https://arxiv.org/abs/2401.08281.

Du, J., Guo, J., Zhang, W., Yang, S., Liu, H., Li, H., and Wang, N. RET-CLIP: A Retinal Image Foundation Model Pre-trained with Clinical Diagnostic Reports . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15012. Springer Nature Switzerland, October 2024.

Fan, M., Wang, F., Chen, C., and Zhou, J. Boosting gradient leakage attacks: Data reconstruction in realistic fl settings. In *34th USENIX Security Symposium (USENIX Security 25)*. USENIX Association, 2025.

Filiot, A., Dop, N., Tchita, O., Riou, A., Dubois, R., Peeters, T., Valter, D., Scalbert, M., Saillard, C., Robin, G., and Olivier, A. Distilling foundation models for robust and efficient models in digital pathology, 2025. URL https://arxiv.org/abs/2501.16239.

Hagestedt, I., Humbert, M., Berrang, P., Lehmann, I., Eils, R., Backes, M., and Zhang, Y. Membership inference against dna methylation databases. In *2020 IEEE European Symposium on Security and Privacy (EuroSP)*, pp. 509–520, 2020. doi: 10.1109/EuroSP48549.2020.00039.

Haim, N., Vardi, G., Yehudai, G., Shamir, O., and Irani, M. Reconstructing training data from trained neural networks, 2022. URL https://arxiv.org/abs/2206.07758.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners, 2021. URL https://arxiv.org/abs/2111.06377.

Huang, S.-C., Pareek, A., Jensen, M., Lungren, M. P., Yeung, S., and Chaudhari, A. S. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *NPJ Digital Medicine*, 6(1): 74, 2023.

Islam, M. M., Neupane, S., Faruk, M. J. H., Shahriar, H., and Cuzzocrea, A. Exposing privacy vulnerabilities in federated learning: A gan-based model inversion attack. In *2025 IEEE International Conference on Big Data (BigData)*, pp. 4227–4236, 2025. doi: 10.1109/BigData66926.2025.11401939.

Kather, J. N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.-A., Gaiser, T., Marx, A., Valous, N. A., Ferber, D., Jansen, L., Reyes-Aldasoro, C. C., Zörnig, I., Jäger, D., Brenner, H., Chang-Claude, J., Hoffmeister, M., and Halama, N. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLOS Medicine*, 16(1):1–22, 01 2019. doi: 10.1371/journal.pmed.1002730. URL https://doi.org/10.1371/journal.pmed.1002730.

Khan, W., Leem, S., See, K. B., Wong, J. K., Zhang, S., and Fang, R. A comprehensive survey of foundation models in medicine, 2025. URL https://arxiv.org/abs/2406.10729.

Kim, H. E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M. E., and Ganslandt, T. Transfer learning for medical image classification: a literature review. *BMC Medical Imaging*, 22(1):69, 2022. doi: 10.1186/s12880-022-00793-7.

Nechaev, D., Pchelnikov, A., and Ivanova, E. Hibou: A family of foundational vision transformers for pathology, 2024. URL https://arxiv.org/abs/2406.05074.

Nguyen, D. M. H., Nguyen, H., Diep, N. T., Pham, T. N., Cao, T., Nguyen, B. T., Swoboda, P., Ho, N., Albarqouni, S., Xie, P., Sonntag, D., and Niepert, M. Lvm-med: Learning large-scale self-supervised vision models for medical imaging via second-order graph matching, 2023. URL https://arxiv.org/abs/2306.11925.

Nguyen, K., Kerkouche, R., Fritz, M., and Karatzas, D. Docmia: Document-level membership inference attacks against docvqa models, 2025. URL https://arxiv.org/abs/2502.03692.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Oz, Y., Yehudai, G., Vardi, G., Antebi, I., Irani, M., and Haim, N. Reconstructing training data from real world models trained with transfer learning, 2024. URL https://arxiv.org/abs/2407.15845.

Peng, X., Han, B., Liu, F., Liu, T., and Zhou, M. Pseudo-private data guided model inversion attacks. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 33338–33375. Curran Associates, Inc., 2024. doi: 10.52202/079017-1051. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/3a797b10ff20562b1ecee0d4e914c1c7-Paper-Conference.pdf.

Petrov, I., Dimitrov, D. I., Baader, M., Müller, M. N., and Vechev, M. Dager: Exact gradient inversion for large language models. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 87801–87830. Curran Associates, Inc., 2024. doi: 10.52202/079017-2787. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/9ff1577a1f8308df1ccea6b4f64a103f-Paper-Conference.pdf.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., Chen, J., Mahvar, F., Yatziv, L., Chen, T., Sterling, B., Baby, S. A., Baby, S. M., Lai, J., Schmidgall, S., Yang, L., Chen, K., Bjornsson, P., Reddy, S., Brush, R., Philbrick, K., Asiedu, M., Mezerreg, I., Hu, H., Yang, H., Tiwari, R., Jansen, S., Singh, P., Liu, Y., Azizi, S., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Riviere, M., Rouillard, L., Mesnard, T., Cideron, G., bastien Grill, J., Ramos, S., Yvinec, E., Casbon, M., Buchatskaya, E., Alayrac, J.-B., Lepikhin, D., Feinberg, V., Borgeaud, S., Andreev, A., Hardin, C., Dadashi, R., Hussenot, L., Joulin, A., Bachem, O., Matias, Y., Chou, K., Hassidim, A., Goel, K., Farabet, C., Barral, J., Warkentin, T., Shlens, J., Fleet, D., Cotruta, V., Sanseviero, O., Martins, G., Kirk, P., Rao, A., Shetty, S., Steiner, D. F., Kirmizibayrak, C., Pilgrim, R., Golden, D., and Yang, L. Medgemma technical report, 2026. URL https://arxiv.org/abs/2507.05201.

Shih, G., Wu, C. C., Halabi, S. S., Kohli, M. D., Prevedello, L. M., Cook, T. S., Sharma, A., Amorosa, J. K., Arteaga, V., Galperin-Aizenberg, M., Gill, R. R., Godoy, M. C. B., Hobbs, S., Jeudy, J., Laroia, A., Shah, P. N., Vummidi, D., Yaddanapudi, K., and Stein, A. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041, 2019. doi: 10.1148/ryai.2019180041.

Spanhol, F. A., Oliveira, L. S., Petitjean, C., and Heutte, L. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7):1455–1462, 2016.

Tiu, E., Talius, E., Patel, P., Langlotz, C. P., Ng, A. Y., and Rajpurkar, P. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12):1399–1406, 2022. doi: 10.1038/s41551-022-00936-9.

Tumanyan, N., Bar-Tal, O., Bagon, S., and Dekel, T. Splicing vit features for semantic appearance transfer, 2022. URL https://arxiv.org/abs/2201.00424.

Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13 (4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Wang, Z., Wu, Z., Agarwal, D., and Sun, J. Medclip: Contrastive learning from unpaired medical images and text, 2022. URL https://arxiv.org/abs/2210.10163.

Wang, Z., Khatibi, E., Sharma, A., Chakrabarty, K., Moosavi, S. R., Firouzi, F., and Rahmani, A. Membership inference attacks expose participation privacy in ecg foundation encoders, 2026. URL https://arxiv.org/abs/2604.10424.

Weber, T., Ingrisch, M., Bischl, B., and Rügamer, D. Implicit embeddings via GAN inversion for high resolution chest radiographs. In *MICCAI Workshop on Medical Applications with Disentanglements*, pp. 22–32. Springer, 2022. doi: 10.1007/978-3-031-25046-0_3.

Xia, W., Zhang, Y., Yang, Y., Xue, J.-H., Zhou, B., and Yang, M.-H. Gan inversion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3121–3138, 2023. doi: 10.1109/TPAMI.2022.3181070.

Xu, Y., Shen, Y., Fernandez-Granda, C., Heacock, L., and Geras, K. J. Understanding differences in applying detr to natural and medical images. *arXiv preprint arXiv:2405.17677*, 2024.

Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., and Ni, B. Medmnist v2 - a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1), January 2023. ISSN 2052-4463. doi: 10.1038/s41597-022-01721-8. URL http://dx.doi.org/10.1038/s41597-022-01721-8.

Ye, D., Zhu, T., Wang, S., Liu, B., Zhang, L. Y., Zhou, W., and Zhang, Y. Data-Free Model-Related attacks: Unleashing the potential of generative AI. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 1709–1727, Seattle, WA, August 2025. USENIX Association. ISBN 978-1-939133-52-6. URL https://www.usenix.org/conference/usenixsecurity25/presentation/ye-attacks.

Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Tupini, A., Wang, Y., Mazzola, M., Shukla, S., Liden, L., Gao, J., Crabtree, A., Piening, B., Bifulco, C., Lungren, M. P., Naumann, T., Wang, S., and Poon, H. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, 2025. URL https://arxiv.org/abs/2303.00915.

Zhao, J. C., Sharma, A., Elkordy, A. R., Ezzeldin, Y. H., Avestimehr, S., and Bagchi, S. Loki: Large-scale Data Reconstruction Attack against Federated Learning through Model Manipulation . In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 1287–1305, Los Alamitos, CA, USA, May 2024. IEEE Computer Society. doi: 10.1109/SP54263.2024.00030. URL https://doi.ieeecomputersociety.org/10.1109/SP54263.2024.00030.

Zhou, Y., Chia, M. A., Wagner, S. K., Ayhan, M. S., Williamson, D. J., Struyven, R. R., Liu, T., Xu, M., Lozano, M. G., Woodward-Court, P., Kihara, Y., Allen, N., Gallacher, J. E. J., Littlejohns, T., Aslam, T., Bishop, P., Black, G., Sergouniotis, P., Atan, D., Dick, A. D., Williams, C., Barman, S., Barrett, J. H., Mackie, S., Braithwaite, T., Carare, R. O., Ennis, S., Gibson, J., Lotery, A. J., Self, J., Chakravarthy, U., Hogg, R. E., Paterson, E., Woodside, J., Peto, T., McKay, G., McGuinness, B., Foster, P. J., Balaskas, K., Khawaja, A. P., Pontikos, N., Rahi, J. S., Lascaratos, G., Patel, P. J., Chan, M., Chua, S. Y. L., Day, A., Desai, P., Egan, C., Fruttiger, M., Garway-Heath, D. F., Hardcastle, A., Khaw, S. P. T., Moore, T., Sivaprasad, S., Strouthidis, N., Thomas, D., Tufail, A., Viswanathan, A. C., Dhillon, B., MacGillivray, T., Sudlow, C., Vitart, V., Doney, A., Trucco, E., Guggenheim, J. A., Morgan, J. E., Hammond, C. J., Williams, K., Hysi, P., Harding, S. P., Zheng, Y., Luben, R., Luthert, P., Sun, Z., McKibbin, M., O'Sullivan, E., Oram, R., Weedon, M., Owen, C. G., Rudnicka, A. R., Sattar, N., Steel, D., Stratton, I., Tapp, R., Yates, M. M., Petzold, A., Madhusudhan, S., Altmann, A., Lee, A. Y., Topol, E. J., Denniston, A. K., Alexander, D. C., Keane, P. A., and Consortium, U. B. E. . V. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06555-x. URL https://doi.org/10.1038/s41586-023-06555-x.

Zimmermann, E., Vorontsov, E., Viret, J., Casson, A., Zelechowski, M., Shaikovski, G., Tenenholtz, N., Hall, J., Klimstra, D., Yousfi, R., Fuchs, T., Fusi, N., Liu, S.,

and Severson, K. Virchow2: Scaling self-supervised mixed magnification models in pathology, 2024. URL https://arxiv.org/abs/2408.00738.

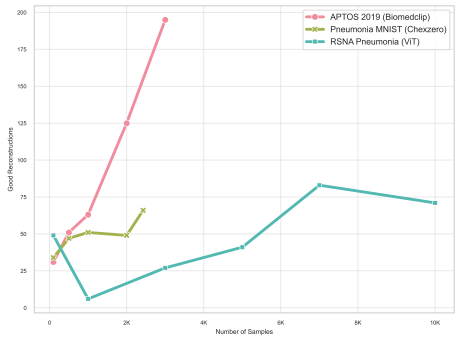
# A. Datasets

- MedMNIST (Pneumonia, Retina, Path) (Yang et al., 2023): 224x224 images enabling the exploration of reconstruction attacks in low-scale data. The datasets are already split into train-test sets, which were maintained in this project.
- RSNA Pneumonia (Shih et al., 2019): 30,000 high-resolution (1024x1024) frontal-view chest radiographs curated from the NIH CXR8 dataset. Curated for binary classification of whether pneumonia is present or absent in the X-ray. 15% of the whole dataset was used as a test set (4000 samples).
- NCT-CRC-HE-100K (Kather et al., 2019): 100,000 H&E-stained pathology images (224x224) representing nine tissue categories. 2 of the classes were explored in this project, colorectal adenocarcinoma epithelium (TUM) and normal colon mucosa (NORM). 2000 samples were used as a test set.
- BreaKHis (Spanhol et al., 2016): Over 9,000 breast tumor tissue images (700x460) collected from 82 patients. Obtained from the P&D Laboratory Pathological Anatomy and Cytopathology in Paraná, Brazil. The dataset includes 9 classes out of those 2 were used in this project to formulate a binary classification problem: ductal carcinoma and fibroadenoma. The patients were split to avoid overlapping between patients in the train and test set. 300 samples were used as a test set.
- APTOS 2019 (Dekhil et al., 2019): Retina fundus photographs of varying resolutions for diabetic retinopathy severity classification obtained from the Aravind Eye Hospital in India. 600 samples were used as a test set.

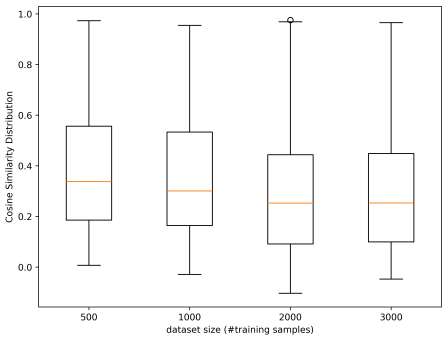
# B. Models

# C. Characteristics of finetuning procedure

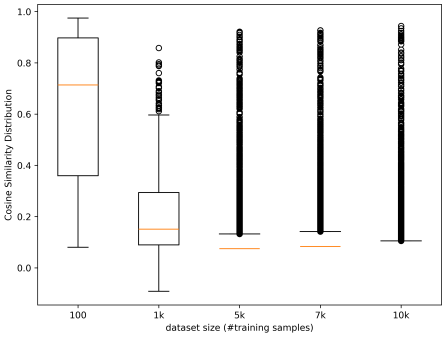
### C.0.1. Dataset size



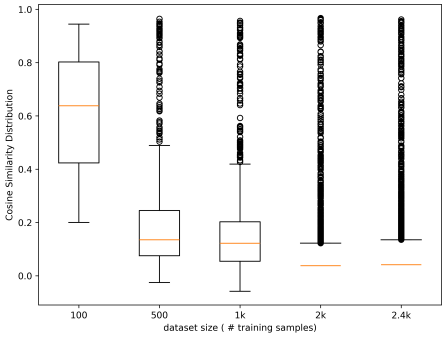

*(a)* Embeddings vs Dataset Size




*(b)* Aptos 2019 - Biomedclip: Cosine Similarity Distributions




*(c)* RSNA Pneumonia - ViT: Cosine Similarity Distributions




*(d)* Pneumonia MNIST - Chexzero: Cosine Similarity Distributions

*Figure 7.* Effect of the dataset size on the number of reconstruction and cosine similarity distributions for each dataset experiment.

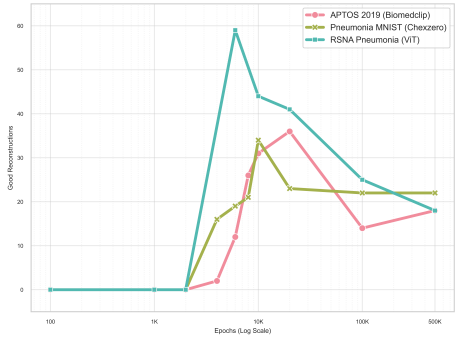
The dataset size directly affects the attack as, according to existing literature, the more unknown variables (training samples), the more known parameters (model weights) are required for the attack to reconstruct a significant percentage of samples.

Figure 7 demonstrates that the attack is applicable in realistic scenarios, as it remained successful when performed on larger datasets of up to 10.000 samples. As expected, while the percentage of reconstructed training samples decreases as the dataset size increases, the absolute number of reconstructed samples can still pose a privacy threat.

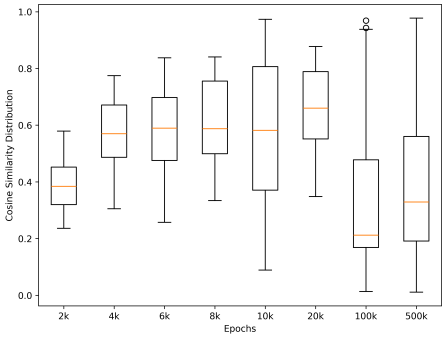
*Table 1.* Model Performance and Reconstructions by Number of Training Samples

| Dataset | Foundation Model | # Samples | Train Loss | Train Accuracy | Test Accuracy | # Good Reconstructions |
|---|---|---|---|---|---|---|
| Pneumonia MNIST | CheXzero | 100 | 0.11560160 | 0.9800 | 0.9279 | 34 |
| Pneumonia MNIST | CheXzero | 500 | 0.15338624 | 0.9600 | 0.9327 | 47 |
| Pneumonia MNIST | CheXzero | 1000 | 0.16711205 | 0.9500 | 0.9375 | 51 |
| Pneumonia MNIST | CheXzero | 2000 | 0.16692871 | 0.9495 | 0.9295 | 49 |
| Pneumonia MNIST | CheXzero | 2428 | 0.16227745 | 0.9522 | 0.9327 | 66 |
| APTOS 2019 | BiomedCLIP | 100 | 0.12971833 | 0.9800 | 0.9510 | 31 |
| APTOS 2019 | BiomedCLIP | 500 | 0.17073770 | 0.9620 | 0.9467 | 51 |
| APTOS 2019 | BiomedCLIP | 1000 | 0.16510810 | 0.9660 | 0.9500 | 63 |
| APTOS 2019 | BiomedCLIP | 2000 | 0.17194358 | 0.9555 | 0.9500 | 125 |
| APTOS 2019 | BiomedCLIP | 3000 | 0.18099311 | 0.9489 | 0.9500 | 195 |
| RSNA Pneumonia | ViT | 100 | 0.17268561 | 0.9900 | 0.7417 | 49 |
| RSNA Pneumonia | ViT | 1000 | 0.45500976 | 0.7980 | 0.7494 | 6 |
| RSNA Pneumonia | ViT | 3000 | 0.48831135 | 0.7743 | 0.7474 | 27 |
| RSNA Pneumonia | ViT | 5000 | 0.49887672 | 0.7678 | 0.7494 | 41 |

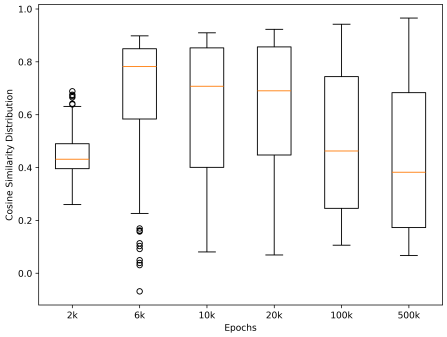
### C.0.2. EPOCHS



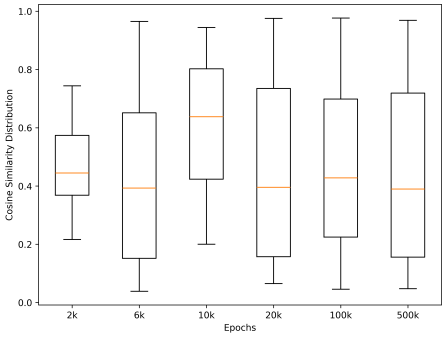

*(a)* Embeddings vs Epochs




*(b)* Aptos 2019 - Biomedclip: Cosine Similarity Distributions




*(c)* RSNA Pneumonia - ViT: Cosine Similarity Distributions




*(d)* Pneumonia MNIST - Chexzero: Cosine Similarity Distributions

*Figure 8.* Effect of the training time on the number of reconstruction and cosine similarity distributions for each dataset experiment.

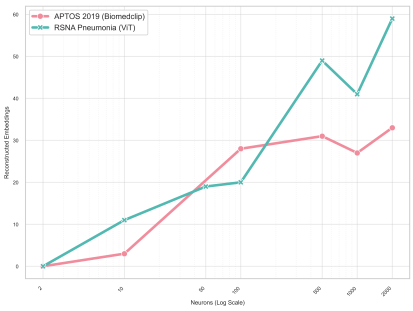
Figure 8 illustrates the training time's effect on the attack. Specifically, experiments were conducted on three different datasets using three distinct models for a ranging number of epochs of 100 up to 500.000 while using weight decay. Results reveal that the highest number of training samples was reconstructed when models were trained between 10.000 and 40.000 epochs. Training models for longer than this threshold reduced the attack's efficiency for these experiments.

*Table 2.* Model Performance and Reconstructions Over Training Epochs

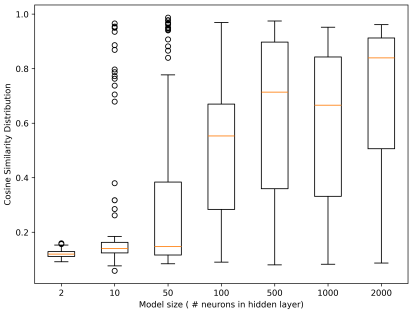
| Dataset | Foundation Model | Epochs | Train Loss | Train Accuracy | Test Accuracy | # Reconstructions |
|---|---|---|---|---|---|---|
| Pneumonia MNIST | CheXzero | 100 | 0.68726623 | 0.7500 | 0.7644 | 0 |
| Pneumonia MNIST | CheXzero | 1000 | 0.66262674 | 0.8800 | 0.8285 | 0 |
| Pneumonia MNIST | CheXzero | 2000 | 0.63723892 | 0.9100 | 0.8590 | 0 |
| Pneumonia MNIST | CheXzero | 4000 | 0.56950170 | 0.9500 | 0.8958 | 16 |
| Pneumonia MNIST | CheXzero | 6000 | 0.52361250 | 0.9600 | 0.9038 | 19 |
| Pneumonia MNIST | CheXzero | 8000 | 0.50752902 | 0.9600 | 0.9054 | 21 |
| Pneumonia MNIST | CheXzero | 10000 | 0.50318152 | 0.9600 | 0.9087 | 34 |
| Pneumonia MNIST | CheXzero | 20000 | 0.50173408 | 0.9600 | 0.9087 | 23 |
| Pneumonia MNIST | CheXzero | 100000 | 0.50173014 | 0.9600 | 0.9087 | 16 |
| Pneumonia MNIST | CheXzero | 500000 | 0.50172937 | 0.9600 | 0.9087 | 22 |
| APTOS 2019 | BiomedCLIP | 100 | 0.68918437 | 0.9200 | 0.9165 | 0 |
| APTOS 2019 | BiomedCLIP | 1000 | 0.64756966 | 0.9200 | 0.9060 | 0 |
| APTOS 2019 | BiomedCLIP | 2000 | 0.51912719 | 0.9100 | 0.9110 | 0 |
| APTOS 2019 | BiomedCLIP | 4000 | 0.26242569 | 0.9300 | 0.9310 | 2 |
| APTOS 2019 | BiomedCLIP | 6000 | 0.19330683 | 0.9300 | 0.9410 | 12 |
| APTOS 2019 | BiomedCLIP | 8000 | 0.16688113 | 0.9500 | 0.9440 | 26 |
| APTOS 2019 | BiomedCLIP | 10000 | 0.15336691 | 0.9600 | 0.9465 | 31 |
| APTOS 2019 | BiomedCLIP | 20000 | 0.13358364 | 0.9800 | 0.9510 | 36 |
| APTOS 2019 | BiomedCLIP | 100000 | 0.12942556 | 0.9800 | 0.9510 | 14 |
| APTOS 2019 | BiomedCLIP | 500000 | 0.12941958 | 0.9800 | 0.9510 | 18 |
| RSNA Pneumonia | ViT | 100 | 0.50599730 | 0.7600 | 0.7187 | 0 |
| RSNA Pneumonia | ViT | 2000 | 0.19319724 | 0.9700 | 0.6727 | 0 |
| RSNA Pneumonia | ViT | 6000 | 0.17864968 | 0.9800 | 0.6805 | 59 |
| RSNA Pneumonia | ViT | 10000 | 0.17000000 | 0.9900 | 0.7417 | 49 |
| RSNA Pneumonia | ViT | 20000 | 0.16737442 | 0.9900 | 0.7117 | 41 |
| RSNA Pneumonia | ViT | 100000 | 0.16832678 | 0.9900 | 0.7100 | 25 |
| RSNA Pneumonia | ViT | 500000 | 0.16732837 | 0.9900 | 0.7120 | 18 |

### C.0.3. Number of Neurons

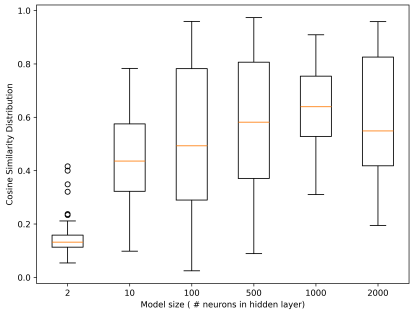
Figure 9 presents the effect of model sizes compared to the number of reconstructed training embeddings. The experiments were performed on two datasets (100 samples from the RSNA Pneumonia and 100 samples from the APTOS 2019) using two foundation models, one general foundation model (ViT) and one medical foundation model (Biomedclip). Results indicate that larger models, with more weights, are more vulnerable against the reconstruction attack. Smaller MLP models are more robust to the attack, a phenomenon that aligns with the non-reconstructible nature of linear models.




*(a)* Reconstructions vs Model Size




*(b)* RSNA Pneumonia - ViT: Cosine Similarity Distributions




*(c)* APTOS 2019 - BiomedCLIP: Cosine Similarity Distributions

*Figure 9.* Number of reconstructions across different datasets and model sizes accompanied by their cosine similarity distributions.

*Table 3.* Performance and Reconstructions by Model Size Across Datasets

| Dataset | Foundation Model | Model | Train Loss | Train Accuracy | Test Accuracy | # Good Reconstructions |
|---|---|---|---|---|---|---|
| RSNA Pneumonia | ViT | d-2-c | 0.46228263 | 0.9400 | 0.7207 | 0 |
| RSNA Pneumonia | ViT | d-10-c | 0.20460334 | 0.9200 | 0.7724 | 11 |
| RSNA Pneumonia | ViT | d-50-c | 0.16768199 | 0.9800 | 0.7197 | 19 |
| RSNA Pneumonia | ViT | d-100-c | 0.21864095 | 0.9400 | 0.6293 | 20 |
| RSNA Pneumonia | ViT | d-500-c | 0.17268561 | 0.9700 | 0.7417 | 49 |
| RSNA Pneumonia | ViT | d-1000-c | 0.19576223 | 0.9500 | 0.7709 | 41 |
| RSNA Pneumonia | ViT | d-2000-c | 0.21870601 | 0.9100 | 0.7767 | 59 |
| APTOS 2019 | BiomedCLIP | d-2-c | 0.39258221 | 0.9200 | 0.8380 | 0 |
| APTOS 2019 | BiomedCLIP | d-10-c | 0.08902651 | 0.9800 | 0.9510 | 3 |
| APTOS 2019 | BiomedCLIP | d-100-c | 0.15545484 | 0.9600 | 0.9460 | 28 |
| APTOS 2019 | BiomedCLIP | d-500-c | 0.12971833 | 0.9800 | 0.9510 | 31 |
| APTOS 2019 | BiomedCLIP | d-1000-c | 0.13334933 | 0.9800 | 0.9510 | 27 |
| APTOS 2019 | BiomedCLIP | d-2000-c | 0.13324386 | 0.9800 | 0.9510 | 33 |

# D. Embedding Reconstruction Risk in Terms of Generalization Gap, Embedding Dimension, Input Resolution and Weight Decay Value

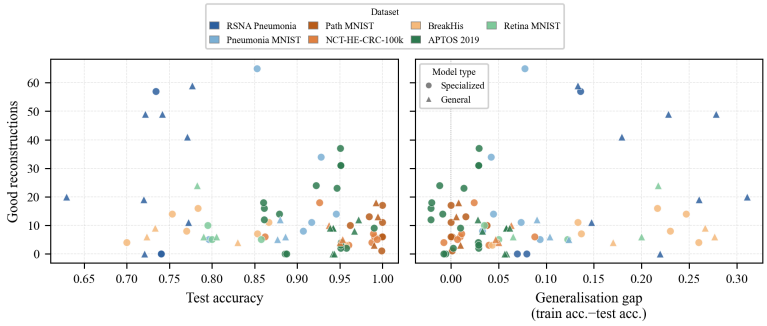



*Figure 10.* Number of good reconstructions in function of the test accuracy (left) and generalization gap (right).

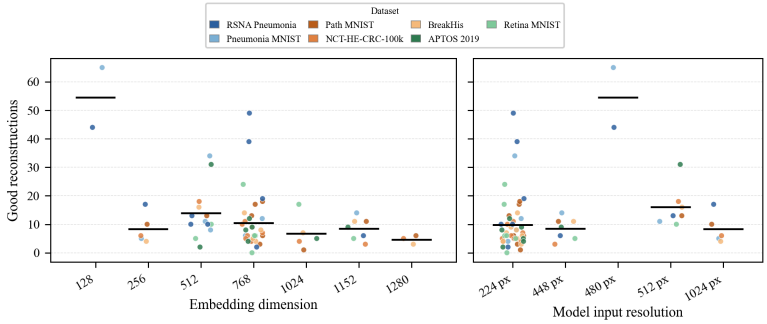



*Figure 11.* Number of good reconstructions in grouped by model embedding dimension (left) and model input resolution (right).

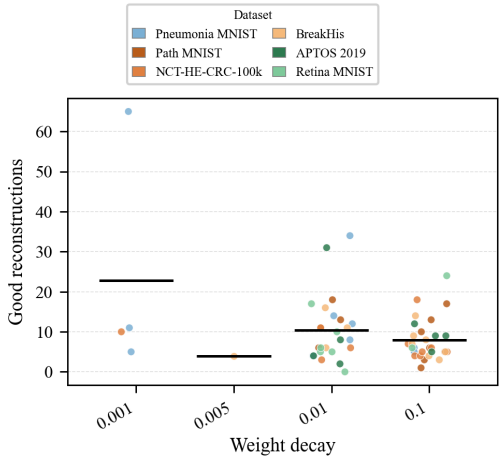
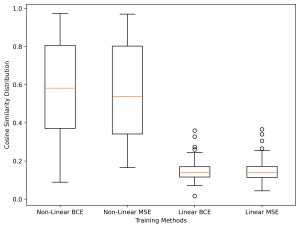



*Figure 12.* Number of good reconstructions in grouped by weight decay value.

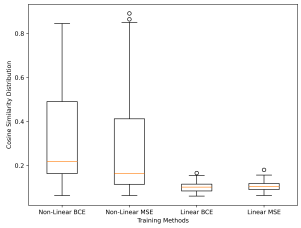
# E. Characteristics of the Classification Head - Models Performance

*Table 4.* Impact of Architecture and Loss on Model Performance and Reconstructions

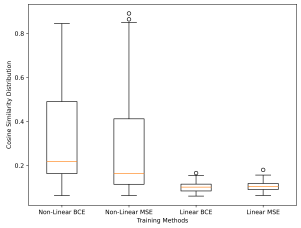
| Dataset | Foundation Model | Architecture | Loss | Weight Decay | Train Loss | Train Acc | Test Acc | # Reconstructions |
|---|---|---|---|---|---|---|---|---|
| RSNA Pneumonia | ViT | Non-linear | BCE | 0.1 | 0.1700 | 0.9900 | 74.17 | 49 |
| RSNA Pneumonia | ViT | Linear | BCE | 0.1 | 0.2960 | 1.0000 | 72.00 | 0 |
| RSNA Pneumonia | ViT | Linear | MSE | 0.1 | 0.0799 | 0.9600 | 72.40 | 0 |
| RSNA Pneumonia | ViT | Non-linear | MSE | 0.1 | 0.0896 | 0.9600 | 72.45 | 15 |
| RSNA Pneumonia | BioViL-T | Non-linear | BCE | 0.001 | - | - | 79.20 | 44 |
| RSNA Pneumonia | BioViL-T | Linear | BCE | 0.01 | 0.5288 | 0.8200 | 0.7404 | 0 |
| RSNA Pneumonia | BioViL-T | Linear | MSE | 0.01 | 0.1949 | 0.8100 | 0.7404 | 0 |
| RSNA Pneumonia | BioViL-T | Non-linear | MSE | 0.01 | 0.1479 | 0.8700 | 0.7339 | 57 |
| RSNA Pneumonia | BiomedCLIP | Non-linear | BCE | 0.1 | 0.6820 | 0.8000 | 0.8196 | 13 |
| RSNA Pneumonia | BiomedCLIP | Linear | BCE | 0.1 | 0.6052 | 0.8100 | 0.7594 | 0 |
| RSNA Pneumonia | BiomedCLIP | Linear | MSE | 0.1 | 0.2247 | 0.8100 | 0.7642 | 0 |
| RSNA Pneumonia | BiomedCLIP | Non-linear | MSE | 0.01 | 0.1354 | 0.8200 | 0.7479 | 12 |
| APTOS 2019 | ViT | Non-linear | BCE | 0.1 | 0.03958056 | 1.0000 | 0.9420 | 9 |
| APTOS 2019 | ViT | Linear | BCE | 0.1 | 0.03959610 | 1.0000 | 0.9415 | 0 |
| APTOS 2019 | ViT | Linear | MSE | 0.1 | 0.01450443 | 1.0000 | 0.9435 | 0 |
| APTOS 2019 | ViT | Non-linear | MSE | 0.1 | 0.02357103 | 1.0000 | 0.9385 | 9 |
| APTOS 2019 | RETFound | Non-linear | BCE | 0.1 | 0.06784815 | 1.0000 | 0.9555 | 5 |
| APTOS 2019 | RETFound | Linear | BCE | 0.1 | 0.07433688 | 1.0000 | 0.9550 | 0 |
| APTOS 2019 | RETFound | Linear | MSE | 0.1 | 0.02857309 | 0.9800 | 0.9535 | 0 |
| APTOS 2019 | RETFound | Non-linear | MSE | 0.1 | 0.04067658 | 0.9800 | 0.9520 | 6 |
| APTOS 2019 | BiomedCLIP | Non-linear | BCE | 0.01 | 0.12971833 | 0.9800 | 0.9510 | 31 |
| APTOS 2019 | BiomedCLIP | Linear | BCE | 0.01 | 0.37192231 | 0.9100 | 0.9230 | 0 |
| APTOS 2019 | BiomedCLIP | Linear | MSE | 0.01 | 0.14285316 | 0.9000 | 0.9180 | 0 |
| APTOS 2019 | BiomedCLIP | Non-linear | MSE | 0.01 | 0.07085040 | 0.9400 | 0.9425 | 31 |



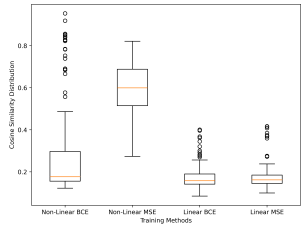

*(a)* APTOS 2019 - Biomedclip: Cosine Similarity Distributions



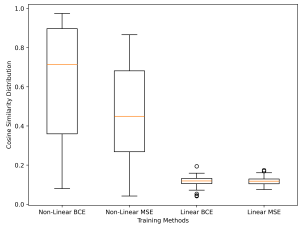

*(b)* APTOS 2019 - ViT: Cosine Similarity Distributions



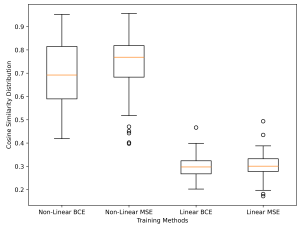

*(c)* APTOS 2019 - Retfound: Cosine Similarity Distributions




*(d)* RSNA Pneumonia - Biomedclip: Cosine Similarity Distributions




*(e)* RSNA Pneumonia - ViT: Cosine Similarity Distributions




*(f)* RSNA Pneumonia - Biovil-T: Cosine Similarity Distributions

*Figure 13.* Cosine Similarity Distributions

# F. Characteristics of the Finetuning Procedure - Models Performance

| Model | Decay | Train Loss | Train Acc | Test Acc | Recons |
|---|---|---|---|---|---|
| CheXzero | 0.001 | 0.07758304 | 0.9900 | 0.7921 | 0 |
| BiomedCLIP | 0.1 | 0.68200141 | 0.7800 | 0.8196 | 13 |
| MedSigLIP | 0.16 | 0.69269830 | 0.8000 | 0.8176 | 6 |
| LVM-Med | 0.01 | 0.50062346 | 0.7600 | 0.7550 | 17 |
| BioViL-T | 0.01 | 0.39684641 | 0.8900 | 0.7354 | 44 |
| MedCLIP | 1e-4 | 0.27398041 | 0.9200 | 0.7434 | 10 |
| ViT | 0.1 | 0.17000000 | 0.9900 | 0.7417 | 49 |
| DinoV2 | 0.1 | 0.23536007 | 0.9000 | 0.7919 | 39 |
| CLIP | 0.01 | 0.40167022 | 0.8200 | 0.7192 | 19 |
| MAE | 0.1 | 0.07317843 | 0.8200 | 0.7190 | 2 |

*Table 5.* RSNA Pneumonia - Performance Metrics and Good Reconstructions

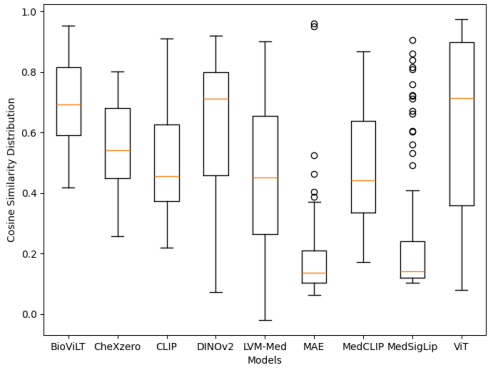



*Figure 14.* Cosine Similarity Distributions - RSNA Pneumonia

| Model | Decay | Train Loss | Train Acc | Test Acc | Recons |
|---|---|---|---|---|---|
| CheXzero | 0.01 | 0.11560160 | 0.9700 | 0.9279 | 34 |
| BiomedCLIP | 0.001 | 0.07176498 | 0.9900 | 0.9167 | 11 |
| MedSigLIP | 0.01 | 0.08290758 | 0.9900 | 0.9455 | 14 |
| LVM-Med | 0.001 | 0.31384239 | 0.8900 | 0.7965 | 5 |
| BioViL-T | 0.001 | 0.18184599 | 0.9300 | 0.8526 | 65 |
| MedCLIP | 0.01 | 0.26241457 | 0.9400 | 0.9071 | 8 |
| ViT | 0.1 | 0.06165028 | 0.9900 | 0.8862 | 6 |
| DinoV2 | 0.1 | 0.04646669 | 1.0000 | 0.8766 | 5 |
| CLIP | 0.01 | 0.21948516 | 0.9700 | 0.8798 | 12 |
| MAE | 0.01 | 0.04612757 | 1.0000 | 0.8974 | 15 |

*Table 6.* Pneumonia MNIST - Performance Metrics and Good Reconstructions

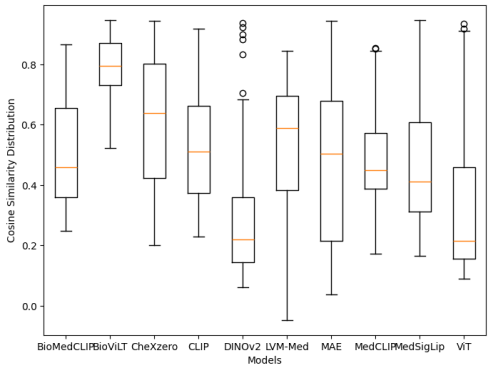



*Figure 15.* Cosine Similarity Distributions - Pneumonia MNIST

| Model | Decay | Train Loss | Train Acc | Test Acc | Recons |
|---|---|---|---|---|---|
| MedSigLIP | 0.01 | 0.03881542 | 1.0000 | 1.0000 | 11 |
| BiomedCLIP | 0.01 | 0.03959277 | 1.0000 | 0.9844 | 13 |
| UNI | 0.1 | 0.38115355 | 1.0000 | 0.9989 | 1 |
| Hibou-b | 0.01 | 0.00095042 | 1.0000 | 0.9989 | 6 |
| H0-mini | 0.1 | 0.00873268 | 1.0000 | 1.0000 | 17 |
| LVM-Med | 0.1 | 0.33096585 | 1.0000 | 0.9622 | 10 |
| DinoV2 | 0.1 | 0.00875209 | 1.0000 | 0.9944 | 13 |
| CLIP | 0.01 | 0.05115738 | 1.0000 | 0.9922 | 18 |
| MAE | 0.1 | 0.12620483 | 1.0000 | 0.9778 | 5 |
| VIRCHOW 2 | 0.1 | 0.00843160 | 1.0000 | 1.0000 | 6 |
| ViT | 0.1 | 0.01303031 | 1.0000 | 0.9900 | 3 |

*Table 7.* Path MNIST - Performance Metrics and Good Reconstructions

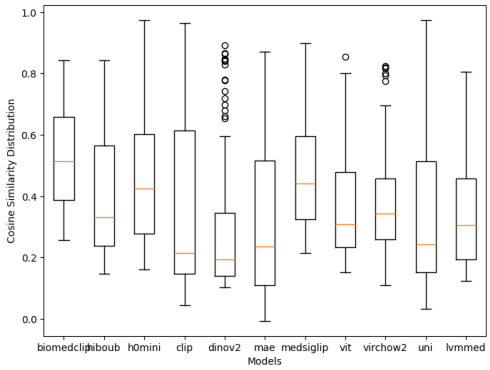



*Figure 16.* Cosine Similarity Distributions - Path MNIST

| Model | Decay | Train Loss | Train Acc | Test Acc | Recons |
|---|---|---|---|---|---|
| MedSigLIP | 0.01 | 0.06991281 | 1.0000 | 0.9605 | 3 |
| BiomedCLIP | 0.1 | 0.58957273 | 0.9500 | 0.9260 | 18 |
| UNI | 0.1 | 0.34819928 | 0.9800 | 0.9875 | 4 |
| Hibou-b | 0.1 | 0.01422368 | 1.0000 | 0.9905 | 6 |
| H0-mini | 0.1 | 0.01101026 | 1.0000 | 0.9890 | 7 |
| LVM-Med | 0.01 | 0.28367668 | 0.9500 | 0.8620 | 6 |
| DinoV2 | 0.1 | 0.03209658 | 1.0000 | 0.9500 | 4 |
| CLIP | 0.001 | 0.02475935 | 1.0000 | 0.9370 | 10 |
| MAE | 0.01 | 0.03205822 | 1.0000 | 0.9100 | 11 |
| VIRCHOW 2 | 0.1 | 0.00930226 | 1.0000 | 0.9935 | 5 |
| ViT | 0.1 | 0.04145234 | 1.0000 | 0.9530 | 5 |

*Table 8.* NCT-HE-CRC-100k - Performance Metrics and Good Reconstructions

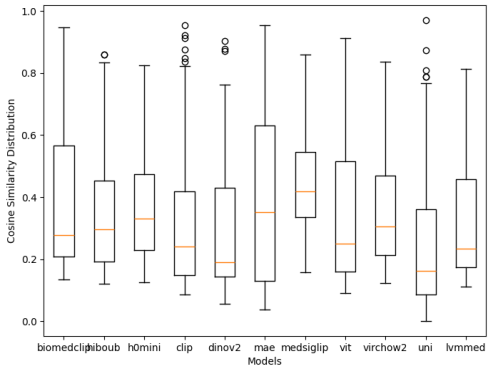



*Figure 17.* Cosine Similarity Distributions - NCT-CRC-HE-100k

| Model | Decay | Train Loss | Train Acc | Test Acc | Recons |
|---|---|---|---|---|---|
| MedSigLIP | 0.01 | 0.05682835 | 1.0000 | 0.8667 | 11 |
| BiomedCLIP | 0.01 | 0.10695966 | 1.0000 | 0.7833 | 16 |
| UNI | 0.1 | 0.34879142 | 0.9900 | 0.8533 | 7 |
| Hibou-b | 0.1 | 0.01256767 | 1.0000 | 0.7533 | 14 |
| H0-mini | 0.1 | 0.01703314 | 1.0000 | 0.7700 | 8 |
| LVM-Med | 0.005 | 0.13172665 | 0.9600 | 0.7000 | 4 |
| VIRCHOW 2 | 0.1 | 0.00913725 | 1.0000 | 0.9567 | 3 |
| DinoV2 | 0.1 | 0.03267370 | 1.0000 | 0.8300 | 4 |
| CLIP | 0.01 | 0.12919739 | 1.0000 | 0.7233 | 6 |
| MAE | 0.1 | 0.08940076 | 0.9900 | 0.8067 | 5 |
| ViT | 0.1 | 0.03060337 | 1.0000 | 0.7333 | 9 |

*Table 9.* BreakHis - Performance Metrics and Good Reconstructions

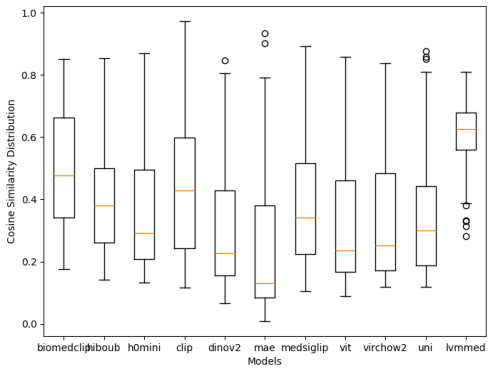



*Figure 18.* Cosine Similarity Distributions - Breakhis

| Model | Decay | Train Loss | Train Acc | Test Acc | Recons |
|---|---|---|---|---|---|
| MedSigLIP | 0.1 | 0.27485275 | 1.0000 | 0.9900 | 9 |
| BiomedCLIP | 0.01 | 0.12971833 | 0.9800 | 0.9510 | 31 |
| RETFound | 0.1 | 0.06784815 | 1.0000 | 0.9555 | 5 |
| RET-CLIP | 0.01 | 0.02254355 | 0.9600 | 0.9575 | 2 |
| DinoV2 | 0.1 | 0.02236914 | 1.0000 | 0.9720 | 12 |
| CLIP | 0.01 | 0.15130764 | 1.0000 | 0.9670 | 8 |
| MAE | 0.01 | 0.02726631 | 1.0000 | 0.9325 | 4 |
| ViT | 0.1 | 0.03958056 | 1.0000 | 0.9420 | 9 |

*Table 10.* APTOS 2019 - Performance Metrics and Good Reconstructions

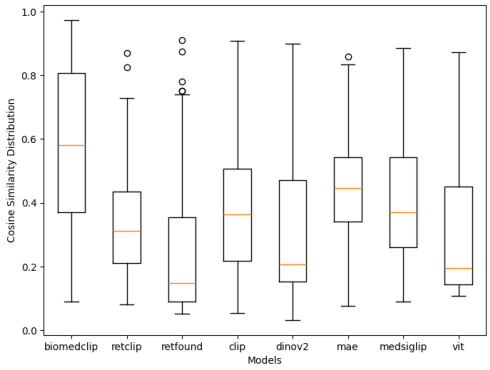



*Figure 19.* Cosine Similarity Distributions - APTOS 2019

| Model | Decay | Train Loss | Train Acc | Test Acc | Recons |
|---|---|---|---|---|---|
| MedSigLIP | 0.01 | 0.11818469 | 0.9800 | 0.8575 | 5 |
| BiomedCLIP | 0.01 | 0.39236304 | 0.8300 | 0.7950 | 10 |
| RETFound | 0.01 | 0.22596300 | 0.9200 | 0.8175 | 17 |
| RET-CLIP | 0.01 | 0.45515382 | 0.8500 | 0.8000 | 5 |
| DinoV2 | 0.1 | 0.05822228 | 1.0000 | 0.7825 | 24 |
| CLIP | 0.01 | 0.35271138 | 0.8700 | 0.8050 | 6 |
| MAE | 0.01 | 0.06790738 | 1.0000 | 0.7725 | 0 |
| ViT | 0.1 | 0.08934769 | 0.9900 | 0.7900 | 6 |

*Table 11.* Retina MNIST - Performance Metrics and Good Reconstructions

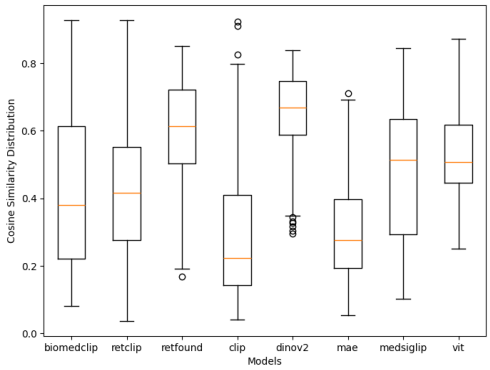



*Figure 20.* Cosine Similarity Distributions - Retina MNIST

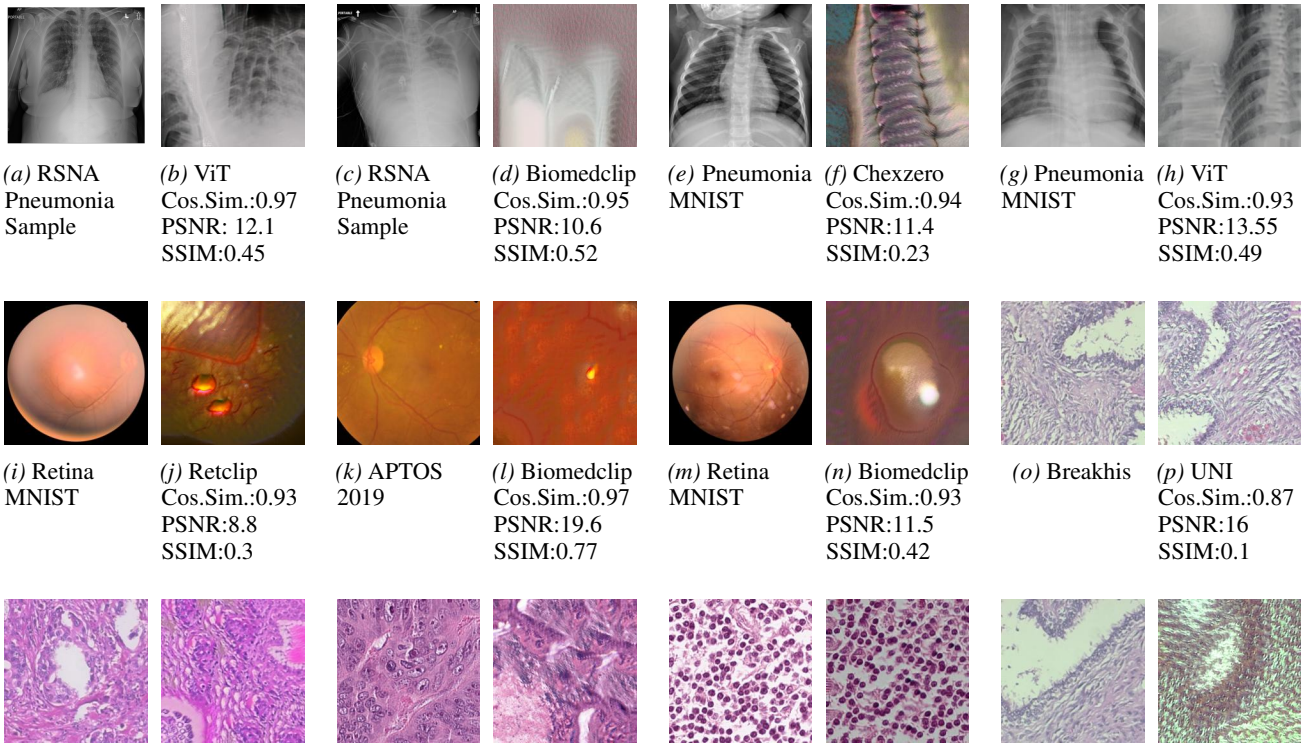
# G. Inversion

### G.0.1. Inversion Comparison Metrics

The metrics chosen for the quantitative image comparison of the inversions are PSNR and SSIM (Wang et al., 2004), as they are the most used for image similarity (Xia et al., 2023) and often referenced in medical image reconstruction comparisons (Weber et al., 2022; Denker et al., 2021). PSNR is a log-scale version of the Mean Squared Error (MSE), and expresses the ratio of the maximum value of the signal to the noise (loss) between images. SSIM (Wang et al., 2004), on the other hand, expresses visual similarity as it could have been perceived by the human eye. Specifically, it accounts for luminance, contrast, and the overall structure of the images, comparing overall image attributes instead of pixel-to-pixel differences.

### G.0.2. Reconstructed Embedding Inversions

Figure 21 depicts pairs of reconstructed embedding inversions, from reconstructed embeddings of section 3.3, accompanied with their original training samples, the cosine similarity scores between reconstructed and training embeddings, as well as their PSNR and SSIM scores.



*(a)* RSNA Pneumonia Sample | *(b)* ViT Cos.Sim.:0.97 PSNR: 12.1 SSIM:0.45 | *(c)* RSNA Pneumonia Sample | *(d)* Biomedclip Cos.Sim.:0.95 PSNR:10.6 SSIM:0.52 | *(e)* Pneumonia MNIST | *(f)* Chexzero Cos.Sim.:0.94 PSNR:11.4 SSIM:0.23 | *(g)* Pneumonia MNIST | *(h)* ViT Cos.Sim.:0.93 PSNR:13.55 SSIM:0.49

*(i)* Retina MNIST | *(j)* Retclip Cos.Sim.:0.93 PSNR:8.8 SSIM:0.3 | *(k)* APTOS 2019 | *(l)* Biomedclip Cos.Sim.:0.97 PSNR:19.6 SSIM:0.77 | *(m)* Retina MNIST | *(n)* Biomedclip Cos.Sim.:0.93 PSNR:11.5 SSIM:0.42 | *(o)* Breakhis | *(p)* UNI Cos.Sim.:0.87 PSNR:16 SSIM:0.1

*(q)* Breakhis | *(r)* h0-mini Cos.Sim.:0.87 PSNR:15.5 SSIM:0.1 | *(s)* NCT-CRC-HE-100k | *(t)* ViT Cos.Sim.:0.91 PSNR:14.6 SSIM:0.07 | *(u)* Path MNIST | *(v)* h0-mini Cos.Sim.:0.97 PSNR:9.3 SSIM:0.03 | *(w)* Breakhis | *(x)* MAE Cos.Sim.:0.93 PSNR:13.3 SSIM:0.07

*Figure 21.* Reconstructed embedding inversions