Curia-2: Scaling Self-Supervised Learning for Radiology Foundation Models

Antoine Saporta*[1], Baptiste Callard*[1], Corentin Dancette[1], Julien Khlaut[1,2,3], Charles Corbière[1], Leo Butsanets[1], Amaury Prat[1], and Pierre Manceron[1]

[1] Raidium, 27 rue du Faubourg Saint-Jacques, 75014 Paris, France.
firstname.lastname@raidium.eu
[2] Department of Vascular and Oncological Interventional Radiology, Hôpital Européen Georges Pompidou, AP-HP, Paris, France
[3] Faculté de Santé, Université Paris-Cité, Paris, France.
corresponding author: antoine.saporta@raidium.eu

Abstract. The rapid growth of medical imaging has fueled the development of Foundation Models (FMs) to reduce the growing, unsustainable workload on radiologists. While recent FMs have shown the power of large-scale pre-training to CT and MRI analysis, there remains significant room to optimize how these models learn from complex radiological volumes. Building upon the Curia framework, this work introduces Curia-2, which significantly improves the original pre-training strategy and representation quality to better capture the specificities of radiological data. The proposed methodology enables scaling the architecture up to billion-parameter Vision Transformers, marking a first for multimodal CT and MRI FMs. Furthermore, we formalize the evaluation of these models by extending and restructuring CuriaBench into two distinct tracks: a 2D track tailored for slice-based vision models and a 3D track for volumetric benchmarking. Our results demonstrate that Curia-2 outperforms all FMs on vision-focused tasks and fairs competitively to vision-language models on clinically complex tasks such as finding detection. To foster further research, model weights are made publicly available at https://huggingface.co/raidium/curia-2.

Keywords: Foundation Models · Computational Radiology.

1 Introduction

Radiology is the cornerstone of modern diagnostic medicine, providing the essential visual data required to detect disease, monitor treatment efficacy, and guide surgical interventions across nearly every clinical specialty.

Foundation Models (FMs) represent a paradigm shift in radiological AI, moving away from task-specific architectures toward broad representations learned from unlabeled data. In the natural vision domain, self-supervised models such as

* Equal contribution.

DINO [17,21] have demonstrated that large-scale pre-training enables high performance on downstream tasks with minimal fine-tuning. Yet, adapting these methods requires handling the uniqueness of radiological data.

Recent advancements have begun to address these challenges, with numerous medical foundation models offering diverse solutions, including MedImageInsight [7], BioMedCLIP [25], MedGemma [19], Merlin [4], Pillar-0 [2], or Curia [9]. Among these, Curia leverages self-supervised pretraining on over 200 million CT and MR images and introducing CuriaBench, a 19-task benchmark designed to evaluate clinical versatility on these modalities. While Curia demonstrated the power of scale, the efficiency of the training strategy remains a challenge due to DINOv2 being fundamentally tailored for natural images.

In this work, we introduce several optimizations to refine Curia's pre-training strategy. More specifically, we propose multiple adaptions to the data augmentations and objective functions of DINOv2 to explicitly address the inherent characteristics of radiological data. Leveraging the efficiency and stability provided by our method, we introduce a new family of models, Curia-2, which scales up to billion-parameter architectures, a first for multi-modal CT and MRI FMs. This family consists of Curia-2 B (86M), Curia-2 L (303M), and Curia-2 g (1.3B). To ensure a rigorous assessment of these advancements, we reformulate CuriaBench into two distinct tracks: a 2D track dedicated to the evaluation of slice-based FMs, and a 3D track providing a standardized benchmark for all radiological FMs. Our results show that the Curia-2 family outperforms all existing slice-based FMs across both tracks and performs competitively with 3D-native FMs on the 3D track.

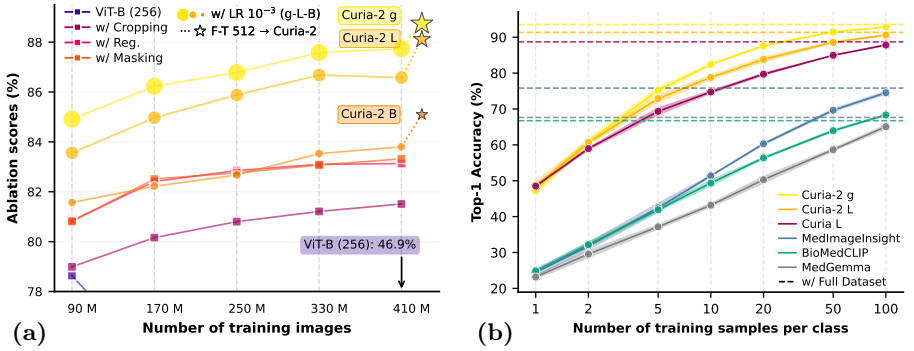

Fig. 1: (a) **Scaling Laws.** The training dynamic steadily improves after each modification (cf. Section 3.2). ☆: fine-tuning in 512 resolution. (b) **Few-shot Learning (A1).** Curia-2 g demonstrates quicker convergence and requires significantly fewer samples to achieve the same performance as other models.

2 Method

2.1 Large-scale Pre-training

Our method builds on Curia [9], which adapts the DINOv2 [17] framework to large-scale radiology data. DINOv2 algorithm combines two objectives: (1) DINO loss, which aligns global representations between a student and a momentum teacher via a cross-entropy self-distillation, (2) the iBOT loss, which enforces consistency between masked patch-level representations across views. Together, these losses enable learning both global and local visual representations from unlabeled data. This section outlines the methodological differences between our approach and both Curia and DINOv2, focusing on improved scalability for radiological data. These fundamental modifications significantly alter the training dynamics and the behavior of representation learning.

Low Pre-training Resolution. In contrast to Curia, which trains on 512-resolution square images, we train at 256 resolution. This configuration significantly lowers the computational cost per sample, enabling more optimization steps for an equivalent compute budget and improving scalability. For context, DINOv2 and its successor DINOv3 [21] are trained at resolutions of 224 and 256, respectively.

Content-Aware Cropping. DINOv2 generates multiple random crops of the input image, on which student and teacher representations are aligned using the DINO objective. While uniform random cropping is well-suited to natural images, it is less appropriate for radiological data, where a substantial portion of images consists of non-informative background. To ensure that feature alignment is conducted on meaningful content, we introduce a two-stage content-aware filtering during crop generation: First, input images that contain more than 50% uniform background are discarded. In the remaining samples, we generate global and local crops containing at most 70% background. This mechanism significantly improves DINOv2's stability on radiological images, especially in low-resolution training, where uninformative background crops are more prevalent.

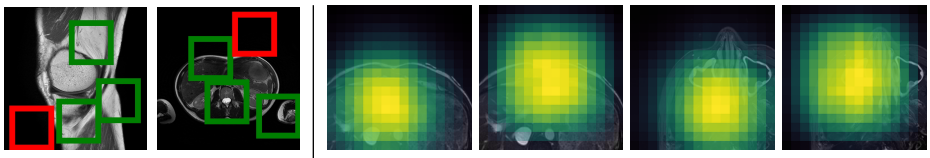

Fig. 2: (Left) **Content-Aware Cropping.** Comparison of valid (green) and filtered out (red) crop regions. (Right) **Anatomically-Guided Masking.** Each pair shows the average of 2000 mask samplings of our anatomically-guided masking against the blockwise masking with uniform prior strategy (DINOv2 [17]).

Anatomically-Guided Masking. The blockwise masking with a uniform prior used in iBOT [26] suffers from similar limitations when applied to radiological images. While the iBOT loss requires the model to reconstruct features of masked patches, a large fraction of the input typically consists of non-informative background; consequently, the uniform prior frequently targets these empty regions, rendering the task trivially simple. We improve upon this strategy by implementing a Gaussian-weighted prior centered on anatomical regions. To maintain spatial consistency, we adjust the Gaussian centers to account for the offsets introduced by global cropping. This biases the masking process toward informative structures, illustrated in Fig. 2, generating a more meaningful learning signal.

Clinical-Prior Regularization. KoLeo loss is the regularizer employed in DINOv2 to encourage a uniform span of features within a batch. Because KoLeo indiscriminately pushes nearest neighbors apart to satisfy an entropy requirement, it may force the model to artificially separate representations that are clinically similar, making it ill-suited for radiological images. Instead, we opt for the SigReg loss [3]. SigReg regularizes the global latent geometry by pulling the distribution toward an isotropic Gaussian. In the context of radiology, where the anatomical topology across different patients follows a highly consistent structural prior, it allows these points to remain naturally clustered while ensuring the overall embedding space remains dense. SigReg preserves the nuances necessary for clinical differentiation without inflating the variance of similar anatomical structures.

High-Resolution Fine-Tuning. To maintain comparability with Curia while pre-training at a lower resolution, we incorporated a final training stage to scale the input resolution to 512. We first evaluated a direct upscaling approach by interpolating the ViT's positional embeddings. This straightforward adaptation yielded performance gains across most tasks compared to the 256 baseline. To further optimize high-resolution performance, we conducted a brief fine-tuning phase at resolution 512, following established protocols [17,21]. This stage requires significantly fewer training steps and incurs minimal computational overhead, thereby preserving the efficiency gains of our low-resolution pre-training.

2.2 Data Preparation

The pre-training dataset, shared with Curia, consists of routine cross-sectional clinical examinations acquired from 2019 to 2022 through the collaboration with a private hospital. All data were fully anonymized, including removal of identifying metadata and defacing of head-containing scans. The initial dataset included 130 TB of data, with 164 million CT images and 64 million MR images. For quality control, only 3D CT and MR examinations with at least five images were retained, while low-quality localizer or scout sequences were excluded.

2.3 Training Details

We trained three variants of ViT architectures [11]: ViT-B (86M), ViT-L (303M), and ViT-g (1,3B). All model configurations were scaled to train on 400M images:

Curia-2 B was trained for 125,000 iterations across 32 NVIDIA A100 (64GB) GPUs with a batch size of 120, with 2 global crops and 8 local crops generated for each image in the batch; Curia-2 L for 156,300 iterations on 64 GPUs with a batch size of 48; and Curia-2 g for 156,300 iterations on 128 GPUs with a batch size of 20. High-resolution fine-tuning stages were run for 20,000 iterations across 8 GPUs, with batch sizes adjusted to hardware constraints. Total compute for the final Curia-2 models amounted to approximately 1,820, 3,270, and 9,680 GPU hours for the B, L, and g variants, respectively.

3 Experiments

3.1 Evaluation Protocol

All models were evaluated based on the protocol defined in Curia [9]. For each downstream task, we trained a simple classification, regression, or survival head on top of the model, without fine-tuning its encoder weights. For every model, we report the average performance over 5 runs. This strategy enables a lightweight, fair comparison of multiple foundation models.

We constructed our evaluation benchmark following CuriaBench [9] tasks and datasets for both ablation studies and comparative evaluations, described in Table 1. To ensure a rigorous comparison, the benchmark incorporates two primary methodological principles: (1) evaluation is conducted across two parallel tracks, a 2D track and a 3D track, where each task may reside in either or both, depending on the dataset structure. The 2D track is dedicated to 2D FMs, while the 3D track serves as a universal evaluation benchmark across all models – for slice-based models, features are averaged on the volume for all 3D tasks. (2) Ten tasks from the 2D track are designated for the development and ablation of Curia-2, while the remaining tasks are reserved to evaluate the generalization of our learned representations. These tasks were selected based on their discriminative power as reported in the CuriaBench study, while ensuring diverse coverage of task categories, modalities, and anatomical regions. Furthermore, we incorporate a multi-finding screening task as proposed by Pillar-0 [2].

3.2 Ablation Study

We conduct an incremental ablation study to evaluate our methodology, beginning with a standard DINOv2 baseline using a ViT-B backbone at 256 resolution. Note this differs a bit from Curia B, which was trained at 512 resolution.

As shown in Table 2 and in Fig. 1a, the ViT-B model initially suffers from training collapse. Our cropping strategy, requiring crops to contain at least 30% foreground signal, is critical for stability (+34.6 points). Since Curia B did not report such instability, we hypothesize that working with crops of half the resolution inherently leads to a higher frequency of empty samples.
Our results confirm that the standard KoLeo loss is ill-suited for the structural homogeneity of medical data, as removing this local distance penalty yields a

Table 1: **Evaluation Benchmark.** The tasks are split into 7 categories: Anatomy (A), Oncology (O), Musculoskeletal (M), Emergency (E), Infectious (I), Neurodegenerative (N), Multi-Finding Screening (F). †: Development task.

A1	†	2D/3D	**CT Organ Recognition** in CT scans from TotalSegmentator [23] dataset. Note that we used 104 organ labels compared to 54 in [9].
A2	†	2D/3D	**MRI Organ Recognition** from TotalSegmentator MRI [10] dataset.
A3	†	2D/3D	**Neuroimaging Age Estimation** from T1-weighted MRI (IXI dataset [1]).
O1	†	2D/3D	**Kidney Lesion Malignancy** classification in CT (KITS23 [5] dataset).
O2		2D/3D	**Lung Nodule Malignancy** classification in CT (LUNA16 dataset [20]).
O3		2D	**Tumor Anatomical Localization** prediction in CT (DeepLesion [24]).
O4		3D	**Kidney Cancer Survival** in contrast-enhanced CT from TCIA[6] cohort.
M1	†	2D	**Subarticular Stenosis** severity classification in axial T2WI [18].
M2		2D	**Foraminal Narrowing** severity classification in sagittal T1WI [18].
M3		2D	**Spinal Canal Stenosis** severity classification in sagittal T2WI & STIR [18].
M4		3D	**Anterior Cruciate Ligament (ACL) Tear** severity classification - injury, complete rupture, or absence - in knee MRI exams [22].
E1	†	2D/3D	**Intracranial Hemorrhage** detection in cranial CT (RSNA 2019 [12]).
E2	†	2D/3D	**Abdominal Trauma** detection in contrast CT (RSNA 2023 challenge [8]).
E3	†	2D/3D	**Myocardial Infarction** signs detection in cardiac MRI (EMIDEC [14]).
E4	†	2D	**Stroke**-resulting brain lesions in T1-weighted MRI (ATLAS R2.0 [15]).
I	†	2D	**Pulmonary Infections** classification (COVIDx CT [13]).
N		3D	**Alzheimer's Disease** detection in brain MRIs (Oasis-1 dataset [16]).
F		3D	**Findings Detection** of around 200 clinical entities in images from the Stanford Merlin Abdominal CT Dataset [4] extracted with RATE [2].

+1.6 points increase. We instead adopt SigReg, which promotes a dense embedding space without sacrificing performance.

Replacing uniform masking with Gaussian sampling further refines the learning signal (+0.2 points). This focuses the iBOT reconstruction task on the central anatomical regions where clinical signal is most concentrated.

Furthermore, we found that increasing the learning rate to 10^{-3} from the original $5 \cdot 10^{-4}$ further optimizes training dynamics (+0.5 points), leveraging the increased stability afforded by our architectural and sampling modifications.

Finally, we evaluated the staged transition to 512 resolution: Positional embedding interpolation alone outperforms evaluation at the native 256 resolution (+1.1 points), high-resolution fine-tuning further refining the features (+0.2).

As shown in the final rows of Table 2, scaling model size to ViT-L and g yields substantial improvements, reaching 88.1% and 88.7%, respectively, confirming that our modifications provide a robust foundation for larger-scale ViTs.

Table 2: **Ablation Study.** Training collapses when lowering the resolution to 256. Our refinements stabilize convergence on radiological data and consistently improve performance while facilitating the expansion to larger ViTs.

Strategy	**Average**	A1 acc	A2 acc	A3 r2	O1 AUC	M1 AUC	E1 AUC	E2 AUC	E3 AUC	E4 AUC	I bacc
Curia B (512)	84.8	85.4	82.3	75.8	74.5	87.8	93.7	82.6	84.7	89.5	91.5
ViT-B (256)	46.9 ↓37.9	22.3	17.7	0.7	49.2	68.8	69.9	64.6	64.6	74.0	37.2
+ Cropping	81.5 ↑34.6	85.4	81.4	64.8	71.0	87.2	92.1	78.8	86.8	87.1	80.5
− KoLeo	83.1 ↑1.6	84.6	81.4	74.8	70.3	87.2	93.3	82.7	82.4	87.3	87.1
+ SigReg	83.1 =	84.7	81.3	75.2	70.4	87.3	92.9	83.2	83.4	85.9	86.9
+ Masking	83.3 ↑0.2	84.2	78.2	75.3	67.9	87.1	93.1	84.1	88.7	87.5	87.2
+ LR 10^{-3}	83.8 ↑0.5	85.1	81.7	76.0	70.6	86.7	93.0	84.7	84.8	88.0	87.4
+ Interp. 512	84.9 ↑1.1	86.7	84.2	76.0	75.5	87.2	92.6	85.7	84.4	88.3	88.2
+ F-T 512 = Curia-2 B	85.1 ↑0.2	87.3	85.2	75.4	75.3	87.3	92.6	82.7	89.0	88.8	87.3
+ ViT-L = Curia-2 L	88.1 ↑3.0	91.4	89.8	77.2	84.7	87.9	93.3	87.4	89.1	89.9	90.4
+ ViT-g = Curia-2 g	88.7 ↑0.6	93.6	90.6	79.0	84.7	88.1	94.0	86.9	87.6	90.7	92.3

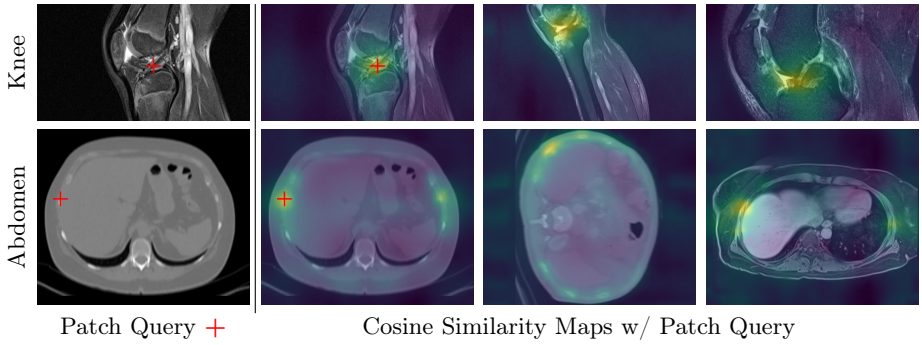

Fig. 3: **Dense Features.** Cosine similarity maps between a query patch + and all patches in multiple images. Curia-2 g demonstrates strong semantic understanding of anatomical structures and cross-modality alignment capabilities, mapping structures between CT and MRI domains, even under rotations.

3.3 Results on the 2D and 3D Tracks

On the 2D track (Table 3), Curia-2 L and Curia-2 g set a new state-of-the-art with an average of 88.5%, significantly exceeding competing models and demonstrating superior robustness across all categories. MedImageInsight [7], best non-Curia model, notably performs well in specific emergency tasks (E2, E3) but falters in anatomical tasks. Curia-2 consistently improves upon its predecessor (+0.6% for ViT-B and +0.9% for ViT-L). Notably, while Curia was already boasting high performance on anatomical tasks (A1-3), the Curia-2 g advances anatomical understanding with peak performance in all of them, a capability further evidenced by the robust, cross-modal alignments shown in Fig. 3. In particular, in a few-shot scenario (Fig. 1b), Curia-2 achieves equivalent per-

Table 3: **2D Track Results.** Curia-2 outperforms 2D FMs, including MedImageInsight (MI2), BioMedCLIP (BMCLIP), MedGemma (MGemma) and Curia.

Model	Avg.	A1 acc	A2 acc	A3 r2	O1 AUC	O2 AUC	O3 bacc	M1 AUC	M2 AUC	M3 AUC	E1 AUC	E2 AUC	E3 AUC	E4 AUC	I bacc
MI2	84.4	75.8	63.2	72.4	68.1	**92.9**	88.9	85.6	86.3	93.0	90.1	**93.2**	**94.0**	88.5	90.0
BMCLIP	79.1	66.7	63.2	69.4	62.7	84.8	86.5	83.9	84.5	92.4	87.8	79.2	71.5	85.1	89.3
MGemma	80.1	67.6	63.4	57.8	63.8	91.2	89.9	83.6	84.7	91.0	87.7	85.1	77.0	88.7	90.4
Curia B	86.2	85.4	82.3	75.8	74.4	86.6	91.9	87.8	86.4	**94.6**	93.7	82.6	84.7	89.5	91.5
Curia L	87.9	88.7	89.1	75.6	80.0	84.8	91.7	87.4	86.3	93.9	93.7	87.1	89.1	89.6	**93.4**
Curia-2 B	86.8	87.2	85.2	75.4	75.3	89.6	**93.5**	87.3	86.9	94.5	92.6	82.7	89.0	88.8	87.3
Curia-2 L	**88.5**	91.4	89.8	77.2	**84.7**	84.0	91.8	87.9	**87.4**	94.5	93.3	87.4	89.1	89.9	90.4
Curia-2 g	**88.5**	**93.6**	**90.6**	**79.0**	**84.7**	80.1	90.6	**88.1**	86.6	94.5	**94.0**	86.9	87.6	**90.7**	92.3

Table 4: **3D Track results.** Curia-2 scores the best results on this generalist benchmark against both 3D volumetric FMs and 2D slice-based FMs.

Model	Avg.	A1 acc	A2 acc	A3 r2	O1 AUC	O2 AUC	O4 cindex	M4 AUC	E1 AUC	E2 AUC	E3 AUC	N bacc	F acc
CT-CLIP	51.5	38.1	20.9	22.0	63.8	56.8	58.1	43.5	69.3	42.6	69.7	69.9	63.4
Merlin	65.4	45.4	38.5	53.3	55.7	64.9	66.3	50.4	85.1	92.5	75.6	76.0	81.2
Pillar-0	71.5	50.7	40.0	53.5	70.2	84.9	68.1	69.2	92.4	**98.2**	56.9	91.7	**82.4**
MI2	83.6	86.3	83.0	80.4	79.8	**89.8**	59.3	71.2	91.5	94.4	94.6	94.4	79.0
BMCLIP	78.3	77.8	71.9	85.0	64.1	81.3	62.3	72.8	89.2	90.1	79.3	91.7	74.1
MGemma	81.8	76.2	79.5	77.7	77.9	88.7	75.3	70.1	87.9	94.0	89.7	91.2	72.7
Curia B	86.0	88.2	92.7	87.8	82.6	77.8	74.3	78.2	**95.1**	93.5	93.8	89.0	78.8
Curia L	84.5	89.4	95.7	86.7	84.0	77.6	60.6	74.4	94.1	91.9	87.9	92.4	79.2
Curia-2 B	87.7	90.9	95.0	87.4	84.5	78.5	72.5	81.1	93.7	94.7	**99.2**	**94.7**	80.2
Curia-2 L	**88.6**	93.2	95.7	88.9	87.1	79.3	**75.5**	**84.9**	94.2	93.2	98.6	92.6	79.7
Curia-2 g	87.8	**94.2**	**96.5**	**89.7**	**87.9**	75.3	74.6	79.4	94.6	93.8	96.1	92.8	79.3

formance milestones much earlier in the training process than competing models, highlighting its superior data efficiency.

On the 3D track (Table 4), the Curia-2 family outperforms all competing FMs, with Curia-2 L securing the top overall position at an 88.6% average, while Curia-2 g achieves remarkable performance on anatomy-related tasks (A1-3). Notably, this represents a substantial leap from its predecessor, improving upon Curia L by +4.1%. Finally, while findings detection (F) is traditionally thought to require the domain-specific vision-language knowledge of architectures like Merlin [4] and Pillar-0 [2], the slice-based Curia-2 models remain competitive on this task with a peak AUC of 80.2% compared to Pillar-0's 82.4%.

4 Conclusion

In this work, we introduced a refined pre-training recipe to account for the unique characteristics of radiological imaging. The proposed method, leading to Curia-2, stabilizes training and achieves superior performance over the original Curia.

Notably, Curia-2 demonstrates a consistent scaling benefit from ViT-B to ViT-L, a trend that remained inconclusive in Curia. Yet, the performance of Curia-2 g remains close to that of Curia-2 L, indicating that the challenge of scaling to billion-parameter models in medical imaging is not yet fully resolved. Curia-2 establishes a new state-of-the-art in vision-focused radiological tasks and bridges the gap with vision-language models on complex tasks like findings detection.

Future work will explore ultra-large-scale strategies from the natural image community, such as drastic increases in data, batch size and training duration, alongside post-training refinements like Gram anchoring from DINOv3. By releasing our weights as open-source, we aim to provide a robust backbone for the community and accelerate research in general-purpose computational radiology.

Acknowledgments. We acknowledge the EuroHPC Joint Undertaking for awarding this project access to the EuroHPC supercomputer LEONARDO, hosted by CINECA (Italy) and the LEONARDO consortium through an EuroHPC AI Factory Access call.

Disclosure of Interests. The authors are affiliated with Raidium.

References

1. Ixi dataset, `https://brain-development.org/ixi-dataset`
2. Agrawal, K.K., Liu, L., Lian, L., Nercessian, M., Harguindeguy, N., Wu, Y., Mikhael, P., Lin, G., Sequist, L.V., Fintelmann, F., et al.: Pillar-0: A new frontier for radiology foundation models. arXiv preprint arXiv:2511.17803 (2025)
3. Balestriero, R., LeCun, Y.: Lejepa: Provable and scalable self-supervised learning without the heuristics. arXiv preprint arXiv:2511.08544 (2025)
4. Blankemeier, L., Cohen, J.P., Kumar, A., Van Veen, D., Gardezi, S.J.S., Paschali, M., Chen, Z., Delbrouck, J.B., Reis, E., Truyts, C., et al.: Merlin: A vision language foundation model for 3d computed tomography. Research Square pp. rs–3 (2024)
5. Cardoso, M.J., Arbel, T., Carneiro, G., Syeda-Mahmood, T., Tavares, J.M.R., Moradi, M., Bradley, A., Greenspan, H., Papa, J.P., Madabhushi, A., et al.: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings, vol. 10553. Springer (2017)
6. Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al.: The cancer imaging archive (tcia): maintaining and operating a public information repository. Journal of Digital Imaging **26**, 1045–1057 (2013)
7. Codella, N.C.F., Jin, Y., Jain, S., Gu, Y., Lee, H.H., Ben Abacha, A., Santamaria-Pang, A., Guyman, W., Sangani, N., Zhang, S., Poon, H., Hyland, S., Bannur, S., Alvarez-Valle, J., Li, X., Garrett, J., McMillan, A., Rajguru, G., Maddi, M., Vijayrania, N., Bhimai, R., Mecklenburg, N., Jain, R., Holstein, D., Gaur, N., Aski, V., Hwang, J.N., Lin, T., Tarapov, I., Lungren, M., Wei, M.: Medimageinsight: An open-source embedding model for general domain medical imaging (Oct 2024)
8. Colak, E., Lin, H.M., Ball, R., Davis, M., Flanders, A., Jalal, S., Magudia, K., Marinelli, B., Nicolaou, S., Prevedello, L., Rudie, J., Shih, G., Vazirabad, M., Mongan, J.: Rsna 2023 abdominal trauma detection. `https://kaggle.com/competitions/rsna-2023-abdominal-trauma-detection` (2023)

9. Dancette, C., Khlaut, J., Saporta, A., Philippe, H., Ferreres, E., Callard, B., Danielou, T., Alberge, L., Machado, L., Tordjman, D., et al.: Curia: A multi-modal foundation model for radiology. arXiv preprint arXiv:2509.06830 (2025)
10. D'Antonoli, T.A., Berger, L.K., Indrakanti, A.K., Vishwanathan, N., Weiß, J., Jung, M., Berkarda, Z., Rau, A., Reisert, M., Küstner, T., et al.: Totalsegmentator mri: Robust sequence-independent segmentation of multiple anatomic structures in mri. arXiv preprint arXiv:2405.19492 (2024)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (ed.) International Conference on Learning Representations (2021)
12. Flanders, A.E., Prevedello, L.M., Shih, G., Halabi, S.S., Kalpathy-Cramer, J., Ball, R., Mongan, J.T., Stein, A., Kitamura, F.C., Lungren, M.P., et al.: Construction of a machine learning dataset through collaboration: the rsna 2019 brain ct hemorrhage challenge. Radiology: Artificial Intelligence **2**(3), e190211 (2020)
13. Gunraj, H., Sabri, A., Koff, D., Wong, A.: Covid-net ct-2: Enhanced deep neural networks for detection of covid-19 from chest ct images through bigger, more diverse learning. Frontiers in Medicine **8**, 729287 (2022)
14. Lalande, A., Chen, Z., Decourselle, T., Qayyum, A., Pommier, T., Lorgis, L., de la Rosa, E., Cochet, A., Cottin, Y., Ginhac, D., Salomon, M., Couturier, R., Meriaudeau, F.: Emidec: A database usable for the automatic evaluation of myocardial infarction from delayed-enhancement cardiac mri. Data **5** (09 2020)
15. Liew, S.L., Lo, B.P., Donnelly, M.R., Zavaliangos-Petropulu, A., Jeong, J.N., Barisano, G., Hutton, A., Simon, J.P., Juliano, J.M., Suri, A., et al.: A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. Scientific data **9**(1), 320 (2022)
16. Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L.: Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. Journal of Cognitive Neuroscience **19**(9), 1498–1507 (2007)
17. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., 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., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024)
18. Richards, T., Talbott, J., Ball, R., Colak, E., Flanders, A., Kitamura, F., Mongan, J., Prevedello, L., Vazirabad, M.: Rsna 2024 lumbar spine degenerative classification. `https://kaggle.com/competitions/rsna-2024-lumbar-spine-degenerative-classification` (2024)
19. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
20. Setio, A.A.A., Traverso, A., De Bel, T., Berens, M.S.N., van den Bogaard, C., Cerello, P., Chen, H., Dou, Q., Fantacci, M.E., Geurts, B., et al.: Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. Medical Image Analysis **42**, 1–13 (2017)
21. 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)

22. Štajduhar, I., Mamula, M., Miletić, D., Uenal, G.: Semi-automated detection of anterior cruciate ligament injury from mri. Computer Methods and Programs in Biomedicine **140**, 151–164 (2017)
23. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., Bach, M., Segeroth, M.: Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5) (Sep 2023)
24. Yan, K., Wang, X., Lu, L., Summers, R.M.: Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. Journal of Medical Imaging **5**(3), 036501–036501 (2018)
25. 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., Poon, H.: Biomedclip: A multimodal biomedical foundation model trained from fifteen million image–text pairs. NEJM AI **2**(1) (Jan 2025)
26. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832 (2021)