Benchmarking the Robustness of Foundation Models for Mammography under Domain Shift

Giang Nguyen[1⋆], Raghav Mehta[2⋆], Emma A.M. Stanley[2], Tian Xia[2], Thi Hao Nguyen[3], Hieu Pham[1,4,5], and Ben Glocker[2]

[1] College of Engineering and Computer Science, VinUniversity, Hanoi, Vietnam
[2] Imperial College London, London, UK
[3] Radiology Department, Vietnam National Cancer Hospital, Hanoi, Vietnam
[4] VinUni-Illinois Smart Health Center, VinUniversity, Hanoi, Vietnam
[5] The Computer Vision and Medical AI Lab, VinUniversity, Hanoi, Vietnam
raghav.mehta@imperial.ac.uk

Abstract. Foundation models are increasingly used as image feature extractors for mammography, but their robustness under external domain shift remains unclear. We benchmark 15 foundation-model backbones across breast density, BI-RADS severity, and cancer status using a unified frozen-backbone linear-probe protocol, training on 3 source datasets and evaluating on 12 task-compatible out-of-distribution (OOD) datasets after label harmonization. Mammography-specific vision-language models (`Mammo-FM` and `MaMA`) provide the strongest mean OOD performance, but robustness is not explained by mammography exposure alone. `DINOv3` remains a competitive vision-only baseline, and mammography-adapted pretraining does not consistently improve generalization. Dataset-level analysis further shows that even leading models show heterogeneous performance across datasets. Feature-space inspection reveals that useful representations can preserve clinical signal while retaining dataset and acquisition structure. These findings highlight dataset-level OOD evaluation as a central criterion for assessing mammography representations[1].

Keywords: Mammography · Foundation models · Out-of-distribution generalization · Domain shift · Vision-language models

1 Introduction

Foundation models (FMs) offer a practical route towards broadly applicable mammography representations. A typical workflow freezes a pretrained backbone, trains a lightweight classification head on a labeled source dataset, and applies the classifier to external data. A single backbone may support multiple downstream mammography tasks while reducing task-specific architectural design. However, the utility of an FM depends on whether its pretrained representation remains useful under distribution shift.

⋆ Equal contribution.
[1] Our code is publicly available at: https://github.com/biomedia-mira/mammo-ood.

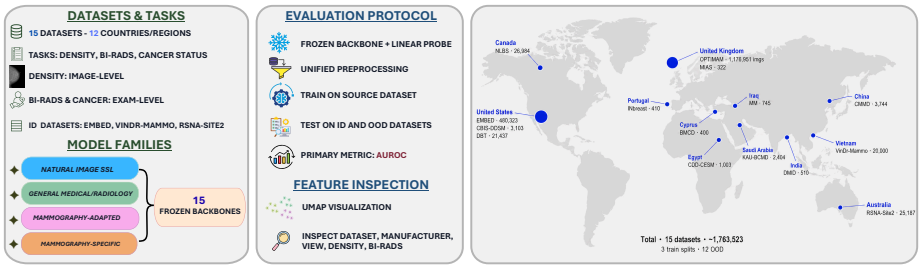

Fig. 1. Overview of the mammography foundation-model benchmark.

Prior work [4,9,12,23,29] has shown that large-scale or modality-specific pretraining does not by itself guarantee robust out-of-distribution (OOD) generalization. This issue is amplified in mammography by heterogeneous clinical labels: density, BI-RADS, and cancer status differ in granularity, prevalence, and annotation protocols. Thus, in-distribution (ID) performance may not translate to OOD generalization. To evaluate whether mammography representations generalize beyond their source data, we establish a benchmark for FM encoders across heterogeneous clinical tasks. We compare natural-image self-supervised learning (SSL) [27,33], general medical/radiology [19,25,30,32,35], mammography adapted SSL [14,27,33], and mammography specific [7,8,10,11,15] models under frozen-backbone linear-probe protocol. Our contributions are threefold:

- We construct a harmonized benchmark from 15 public mammography datasets spanning 12 countries/regions. It covers three different downstream tasks: image-level density, exam-level BI-RADS, and exam-level cancer status.
- We evaluate 15 foundation-model backbones under a unified frozen-backbone source-to-external ID/OOD protocol across these three tasks.
- We analyze OOD behavior beyond aggregate scores by combining model-family comparisons, per-dataset evaluation, and feature-space inspection.

Our results show that mammography pretraining alone is insufficient for robust OOD generalization. Instead, OOD generalization depends on the pretraining objective, model type (vision-only vs. vision-language), and pretraining dataset diversity. Mammography-specific vision-language models (VLMs) – `MaMA` [7] and `Mammo-FM` [11] – achieve the strongest overall performance, `DINOv3` [33] remains a competitive vision-only baseline, and continued pretraining on a single mammography dataset does not improve OOD generalization. Together, these findings highlight the importance of OOD evaluation as a central criterion for assessing the clinical utility of FM representations in mammography.

2 Benchmark Setup

2.1 Benchmark Datasets and Tasks

We study three clinically distinct mammography prediction tasks: breast-density classification, BI-RADS assessment prediction [2], and cancer-status classification. These tasks differ in their clinical meaning and label structure: breast

Table 1. Benchmark datasets after task-specific label filtering.Values are task-specific sample counts: exams for BI-RADS/cancer and images for density. * marks the source dataset for each task; parentheses show held-out ID test units for source datasets. † marks BI-RADS targets with available ordinal labels but a missing class; these targets are excluded from the final OOD AUROC result to avoid biasing the benchmark.

	VinDr-Mammo[26]	EMBED[16]	RSNA-site2[31]	OPTIMAM[13]	CBIS-DDSM[20]	CDD-CESM[18]	INbreast[24]
Region	Vietnam	USA	Australia	UK	USA	Egypt	Portugal
BI-RADS 1	*2,515 (494)	8,897	–	–	1	47	10
BI-RADS 2	*1,568 (319)	2,064	–	–	101	41	44
BI-RADS 3	*436 (91)	723	–	–	185	60	9
BI-RADS 4	*368 (73)	227	–	–	850	103	20
BI-RADS 5	*113 (23)	33	–	–	309	73	25
Density A	20	*31,320 (6,674)	–	9,448	425	8	136
Density B	380	*115,182 (22,565)	–	46,492	1,207	328	146
Density C	3,060	*107,243 (22,545)	–	26,580	952	514	98
Density D	540	*14,062 (2,727)	–	11,100	517	70	28
Non-cancer	4,852	648	*5,861 (1,156)	46,569	814	147	–
Cancer	113	298	*234 (44)	1,284	752	179	–

	DMID[28]	BMCD[21]	KAU-BCMD[1]	CMMD[6]	DBT[5]	MIAS[34]	MM[3]	NLBS[17]
Region	India	Cyprus	Saudi Arabia	China	USA	UK	Iraq	Canada
BI-RADS 1	209	†22	†460	–	–	–	–	–
BI-RADS 2	25	†28	†0	–	–	–	–	–
BI-RADS 3	121	†0	†102	–	–	–	–	–
BI-RADS 4	135	†48	†32	–	–	–	–	–
BI-RADS 5	19	†2	†9	–	–	–	–	–
Density A	79	20	680	–	–	–	–	–
Density B	203	114	1,053	–	–	–	–	–
Density C	186	56	383	–	–	–	–	–
Density D	40	10	126	–	–	–	–	–
Non-cancer	180	–	–	465	5,389	270	620	5,848
Cancer	130	–	–	1,310	88	52	125	149

density characterizes breast tissue composition, BI-RADS provides an ordinal radiological assessment, and cancer status is a clinically important binary outcome with substantial class imbalance. As public datasets differ in their native annotation schemes, we harmonize labels before evaluation. BI-RADS and cancer status are evaluated per exam, reflecting clinical reading workflow, while breast density is evaluated per image.

The benchmark covers 15 datasets from 12 countries across North America, Europe, Asia, the Middle East, Africa, and Oceania. For each task, the source (ID) dataset acts as the training data for the task-specific linear classifier on frozen backbone features: `VinDr-Mammo` for BI-RADS, `EMBED` for density, and `RSNA-site2`[2] for cancer status[3]. As each dataset has different available labels, only task-compatible samples are included after filtering in Table 1.

2.2 Foundation Models

We evaluate 15 foundation-model backbones grouped into four non-overlapping families (Table 2). The mammography-adapted family consists of three vision-

[2] We only use `RSNA-site2` (the Australian cohort) and exclude `RSNA-site1` as it overlaps with `EMBED`, avoiding leakage across datasets.

[3] Since `EMBED`, `VinDr`, and `RSNA` were used as unsupervised pretraining datasets for some foundational models (see Sec. 2.2), using them for OOD evaluation may introduce a small but non-negligible effect.

Table 2. Foundation-model families used throughout the paper. Family refers to pretraining/domain source; † denotes VLM or language-aligned encoders.

Family	Models
Natural image SSL	DINOv2 [27], DINOv3 [33]
General medical/radiology	BioMedCLIP† [35], UniMedCLIP† [19], MedSigLIP† [32], RAD-DINO [30], RayDINO [25]
Mammography-adapted	DINOv2-EMBED [27], DINOv3-EMBED [33], MAE-EMBED [14]
Mammography-specific	MaMA† [7], GLAM† [8], Mammo-CLIP† [10], Mammo-FM† [11], VersaMammo [15]

only SSL models that we further pretrained on the EMBED [16] dataset. All other models are publicly available, which we directly used for feature extraction without any further modification. For VersaMammo, we evaluate the stage-1 backbone pretrained using the DINOv2 framework, rather than its later distilled variant supervised by a CNN backbone. Reported mammography-specific training sources for various publicly available models are: EMBED for MaMA and GLAM; UMPC (private dataset) and VinDr-Mammo for Mammo-CLIP; EMBED, UMPC, and Mayo Clinic (private dataset) for Mammo-FM; and EMBED, VinDr-Mammo, and RSNA for VersaMammo.

2.3 Evaluation Protocol

All models are evaluated as frozen image encoders. For ViT-style models, we use the final classification-token or pooled representation as the frozen image feature; for CNN-based models, we use global-average-pooled features. A task-specific linear classifier is trained on the source (ID) dataset[4] and evaluated on the held-out ID test set and compatible external OOD datasets. All datasets are standardized through a common preprocessing pipeline, and images are resized to 1024×768, preserving the global breast context. For BI-RADS and cancer status, image-level model probabilities are aggregated to the exam level by per-class max pooling followed by renormalization. Similarly, exam-level *ground-truth* labels are defined by the maximum image-level BI-RADS or cancer label within each exam. Density predictions are evaluated directly at the image level. We use macro-averaged one-vs-rest AUROC for BI-RADS and density, and binary AUROC for cancer. Unless stated otherwise, OOD scores are averaged over task-compatible external datasets with finite AUROC. We calculate 95% confidence intervals (CIs) using 1000 bootstrap runs (with replacement) for each dataset/model; $\pm$ denotes CI half-width.

3 Results

3.1 ID–OOD Performance Alignment and Gaps

Fig. 2 plots ID AUROC against mean OOD AUROC for density, BI-RADS, and cancer status; the diagonal indicates equivalent ID and OOD performance. For

[4] For a fixed random seed, models were trained using the AdamW optimizer, a learning rate of 0.001, and binary cross-entropy loss for 20 epochs. For cancer classification, we used a weighted batch sampler with $\alpha = 0.3$ to tackle the high class imbalance.

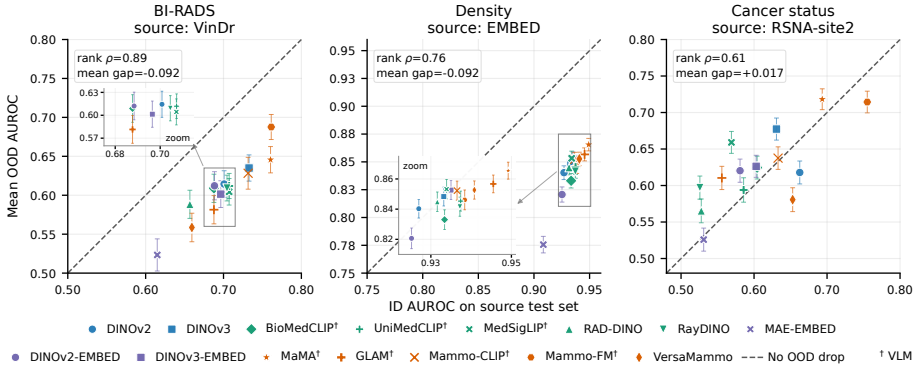

Fig. 2. ID-to-OOD generalization of foundation-model representations. Each point is one frozen backbone. Axes show ID AUROC (x-axis) and mean OOD AUROC (y-axis) across compatible external datasets. Error bars shows 95% confidence intervals across OOD datasets. Colors indicate pretraining/domain family and † marks VLM encoders. The dashed diagonal marks equivalent ID and OOD performance.

BI-RADS, source and external performance are strongly rank-aligned (Spearman $\rho = 0.89$), but models still fall below the diagonal on average, with a mean OOD gap of -0.092. `Mammo-FM` and `MaMA` occupy the high-ID/high-OOD region, while `DINOv3` is the strongest vision-only model. Density shows a higher source-domain ceiling, with many models reaching ID AUROC between 0.93 and 0.95. However, the mean OOD gap remains the same (-0.092) with a lower rank correlation ($\rho = 0.76$), reflecting a narrow, high-performing cluster led by `MaMA`. Cancer has the weakest ID/OOD rank relationship ($\rho = 0.61$) and a slightly positive mean gap ($+0.017$). These results are noisy because cancer-status targets differ in label definition across datasets: from BI-RADS-derived labels in `VinDr-Mammo` (1–2 negative, 5 positive) to coarse normal/benign/abnormal annotations and screening versus cancer-enriched cohorts. The averaged score, therefore, mixes targets with different effective task difficulty.

3.2 Model-Level OOD Performance

We next focus on external performance and ask which models remain strong after a domain shift. Mammography-specific VLMs achieve the highest performance, with `MaMA`, `Mammo-FM`, and `Mammo-CLIP` among the strongest models. The leading model is task-dependent rather than universal: `Mammo-FM` achieves the best BI-RADS OOD AUROC ($0.688_{\pm 0.016}$), whereas `MaMA` reaches the best performance for density ($0.865_{\pm 0.006}$) and cancer status ($0.718_{\pm 0.014}$). `DINOv3` is the most important general-purpose vision-only model. Despite no mammography-specific pretraining, it reaches $0.635_{\pm 0.017}$ on BI-RADS, $0.848_{\pm 0.006}$ on density, and $0.677_{\pm 0.015}$ on cancer, outperforming several mammography-adapted or mammography specific models. These results suggest two conclusions: mammography-specific VLMs provide the strongest OOD performance overall, but a strong natural-image vision-only encoder remains a difficult baseline to beat.

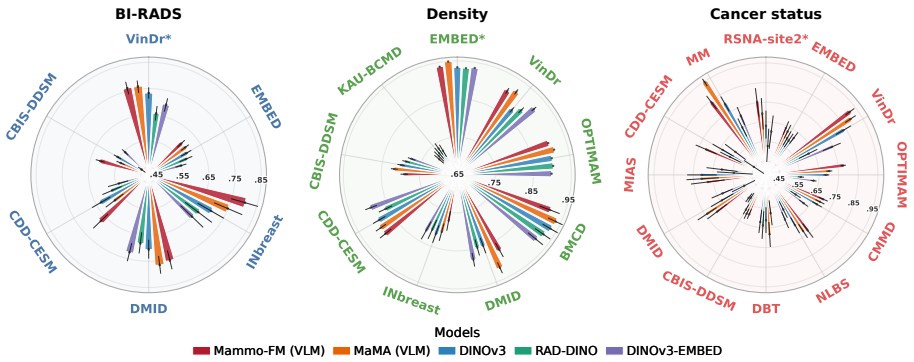

Fig. 3. Dataset-level AUROC performances (with 95% confidence interval error-bar) for five representative backbones across BI-RADS, breast density, and cancer status. The ID datasets are marked with an asterisk (∗); all other datasets are used as OOD.

3.3 Pretraining Factors in OOD Generalization

Across model families, OOD performance varies with clinical-language alignment, pretraining source, and pretraining objective. VLMs achieve higher mean OOD AUROC than vision-only models for BI-RADS ($0.624_{\pm 0.007}$ vs $0.593_{\pm 0.006}$), density ($0.850_{\pm 0.002}$ vs $0.835_{\pm 0.002}$), and cancer status ($0.651_{\pm 0.006}$ vs $0.601_{\pm 0.006}$). However, this advantage is not uniform: `DINOv3` remains competitive with `MaMA` on BI-RADS ($0.635_{\pm 0.017}$ vs $0.646_{\pm 0.017}$) and is close to the leading VLMs on cancer status ($0.677_{\pm 0.015}$ vs $0.718_{\pm 0.014}$/$0.714_{\pm 0.015}$ for `MaMA`/`Mammo-FM`). Thus, although VLMs perform better on average, strong SSL vision-only models remain competitive for OOD generalization.

When grouped by broad pretraining domain (see Table 2), natural-image models achieve mean OOD AUROCs of $0.625_{\pm 0.012}$, $0.844_{\pm 0.004}$, and $0.648_{\pm 0.011}$ for BI-RADS, density, and cancer status, respectively, compared with $0.605_{\pm 0.008}$, $0.844_{\pm 0.003}$, and $0.608_{\pm 0.007}$ for general medical/radiology models and $0.605_{\pm 0.006}$, $0.840_{\pm 0.002}$, and $0.629_{\pm 0.005}$ for mammography-adapted models. Although the strongest individual models are mammography-specific VLMs, the mammography-adapted vision-only model performance remains heterogeneous, indicating that mammography exposure alone does not guarantee robust OOD generalization.

Backbone- and objective-level comparisons show a similar pattern. `DINOv3-EMBED` provides only a marginal density gain over `DINOv3` ($0.853_{\pm 0.006}$ vs $0.848_{\pm 0.006}$) while reducing BI-RADS ($0.601_{\pm 0.017}$ vs $0.635_{\pm 0.017}$) and cancer performance ($0.626_{\pm 0.015}$ vs $0.677_{\pm 0.015}$). `VersaMammo` is strong only for density ($0.853_{\pm 0.006}$) despite multi-source mammography pretraining. These results indicate that OOD generalization depends on the interaction between source domain, pretraining objective, and model type, rather than on mammography exposure alone.

3.4 Dataset-Level OOD Heterogeneity

To more closely assess how different backbones generalize across dataset shifts, we examine per-dataset performance for five representative models: `DINOv3`,

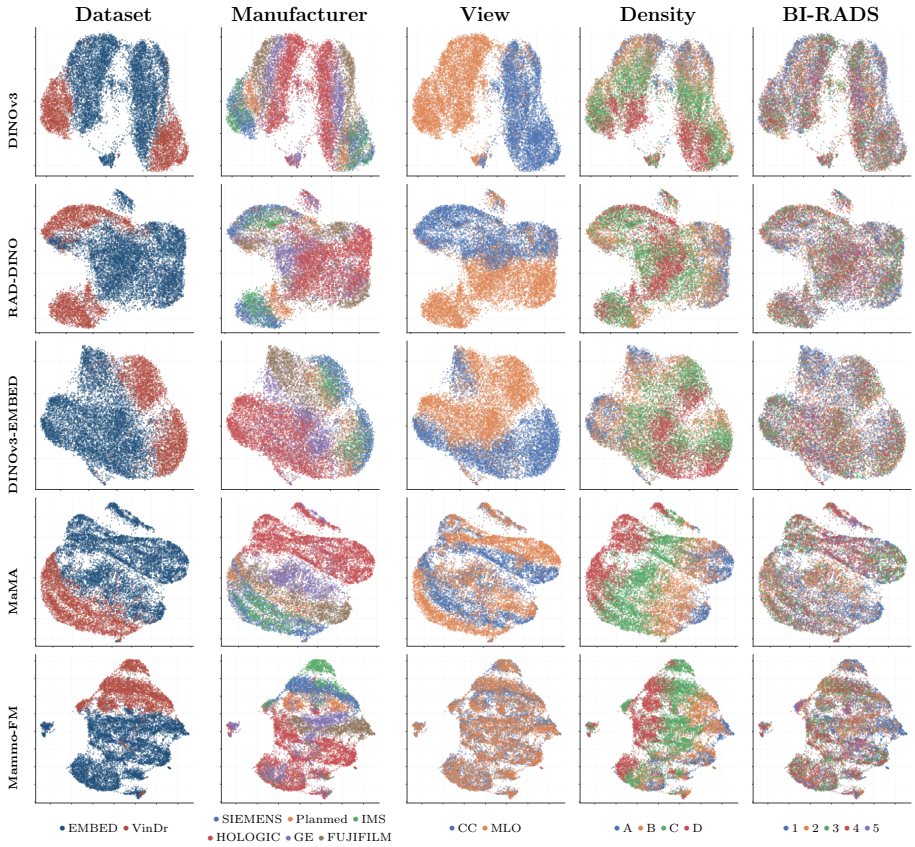

Fig. 4. UMAP inspection grid for five representative backbones. Rows are `DINOv3`, `RAD-DINO`, `DINOv3-EMBED`, `MaMA`, and `Mammo-FM`. Columns color the embedding by dataset, manufacturer, view position, breast density, and BI-RADS.

`RAD-DINO`, `DINOv3-EMBED`, `MaMA`, and `Mammo-FM`. The first three provide a `DINO`-family contrast spanning natural-image, general-radiology, and mammography-adapted pretraining, while `MaMA` and `Mammo-FM` represent the strongest mammography specific VLMs. Fig. 3 shows that strong average OOD performance does not necessarily correspond to uniformly strong behavior across OOD datasets. All three `DINO`-family vision models exhibit heterogeneous performance across datasets and tasks. However, their relative ranking remains consistent, with `RAD-DINO` showing the lowest performance and `DINOv3` achieving the highest.

For BI-RADS, `Mammo-FM` shows a higher mean OOD performance compared to `MaMA` ($0.688_{\pm 0.016}$ vs. $0.646_{\pm 0.017}$). However, Fig. 3 reveals that `Mammo-FM` demonstrates substantially higher performance on the `CDD-CESM` and `CBIS-DDSM` datasets, while performing similarly to `MaMA` on the other three datasets (`EMBED`, `DMID`, and `INbreast`). This performance disparity indicates `Mammo-FM`'s better overall generalization to variability in image acquisition parameters, as both `CDD-CESM` (contrast-enhanced mammography) and `CBIS-DDSM` (scanned film mammogra-

phy) represent non-standard mammography modalities. However, for density, MaMA consistently outperforms Mammo-FM across all datasets (mean OOD $0.865_{\pm 0.006}$ vs $0.846_{\pm 0.007}$). We hypothesise that this is due to the explicit use of density information during MaMA's pretraining, as it relied on tabular metadata (e.g., density, BI-RADS) to generate text reports for language pretraining. In contrast, Mammo-FM used actual clinical reports, which do not explicitly encode density information, focusing instead on clinical outcomes like BI-RADS and cancer status. This difference in pretraining supervision is also reflected in cancer status. Mammo-FM performs better on datasets where cancer status is closer to an exam-level clinical outcome or proxy label (RSNA, OPTIMAM, VinDr-Mammo, and CDD-CESM), consistent with its report-based pretraining. Conversely, MaMA is stronger on datasets with more direct or curated cancer annotations (DBT, DMID, MIAS, and MM), suggesting that metadata-derived report supervision may better preserve explicit visual abnormality cues.

3.5 Clinical and Domain Structure in Feature Space

Following the dataset-level analysis, we use UMAP [22] as a qualitative inspection of what information remains organized in the same five representations. Fig. 4 compares their joint EMBED/VinDr embeddings. Across models, the dataset structure remains visible, indicating that strong external generalization does not require complete domain invariance. The DINO-family comparison shows that changing the pretraining source reshapes the embedding geometry but does not necessarily improve clinical organization. DINOv3 preserves strong view structure and some density structure, while RAD-DINO and DINOv3-EMBED do not show clear density separation, consistent with their weaker OOD gains.

The two strongest mammography-specific VLMs show different qualitative biases. MaMA has the clearest density organization, consistent with its leading density OOD AUROC, and with report-aligned pretraining capturing density semantics. However, MaMA still preserves visible view separation. This was surprising, as MaMA explicitly aligned different views of the same patient during its image pre-training. We hypothesize that this is a byproduct of using tabular image information (e.g., view, side) in the language pre-training of the MaMA model. In contrast, Mammo-FM shows better alignment between different views, despite not having any explicit multi-view alignment in its pre-training. We hypothesize that this is due to the use of actual clinical reports during the language pre-training of Mammo-FM, which do not have image information encoded in them.

4 Conclusion

This benchmark shows that strong source-domain performance and mammography exposure alone do not guarantee robust OOD generalization. Mammography-specific VLMs achieve the strongest mean OOD performance, with MaMA and Mammo-FM forming the most consistent top pair, but their gains are task- and dataset-dependent. DINOv3 remains a competitive vision-only baseline, while

general-radiology pretraining and single-source mammography adaptation do not consistently improve transfer. These results suggest that robust mammography representations depend on how pretraining objectives and data sources align with each clinical task and its label provenance. Dataset-level and feature-space analyses further show that external robustness is heterogeneous rather than uniform. Therefore, mammography foundation models should be assessed using task-specific OOD validation rather than ID performance.

A limitation of this study is that our linear-probe protocol measures representation quality rather than the best achievable absolute performance for different tasks. Future work should evaluate full fine-tuning and other model adaptation strategies to explore the full potential of different foundation models. In the future, it would also be interesting to evaluate uncertainty estimation and confidence calibration of foundational models across OOD datasets.

Acknowledgement

This project was supported by the Foreign, Commonwealth & Development Office (FCDO), Natural Sciences and Engineering Research Council (NSERC) of Canada, the Royal Academy of Engineering as part of the Kheiron/RAEng Research Chair, the Imperial College London UKRI Impact Acceleration Account `EP/X52556X/1`, and VinUniversity's Seed Grant Program under Project `VUNI.2425.EME.005`.

Disclosure of interests

B.G. is a part-time employee of DeepHealth. No other competing interests.

References

1. Alsolami, A.S., Shalash, W., Alsaggaf, W., et al.: King Abdulaziz University Breast Cancer Mammogram Dataset (KAU-BCMD). Data **6**(11), 111 (2021)
2. American College of Radiology: ACR BI-RADS Atlas: Breast Imaging Reporting and Data System. 5th edn. American College of Radiology, Reston (2013)
3. Aqdar, K.B., Mustafa, R.K., Abdulqadir, Z.H., et al.: Mammogram mastery: A robust dataset for breast cancer detection and medical education. Data in Brief **55**, 110633 (2024)
4. Bommasani, R., Hudson, D.A., Adeli, E., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)
5. Buda, M., Saha, A., Walsh, R., et al.: A data set and deep learning algorithm for the detection of masses and architectural distortions in digital breast tomosynthesis images. JAMA Network Open **4**(8), e2119100 (2021)
6. Cai, H., Wang, J., Dan, T., et al.: An online mammography database with biopsy confirmed types. Scientific Data **10**(1), 123 (2023)
7. Du, Y., Onofrey, J., Dvornek, N.C.: Multi-view and multi-scale alignment for contrastive language-image pre-training in mammography. arXiv preprint arXiv:2409.18119 (2024)

8. Du, Y., Chen, L., Dvornek, N.C.: GLAM: Geometry-guided local alignment for multi-view VLP in mammography. arXiv preprint arXiv:2509.10344 (2025)
9. Germani, E., Türk, I.S., Zeineddine, F., Mourad, C., Albarqouni, S.: Bias and generalizability of foundation models across datasets in breast mammography. arXiv preprint arXiv:2505.10579 (2025)
10. Ghosh, S., Poynton, C.B., Visweswaran, S., Batmanghelich, K.: Mammo-CLIP: A vision language foundation model to enhance data efficiency and robustness in mammography. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024, pp. 632–642. Springer Nature Switzerland, Cham (2024)
11. Ghosh, S., Joshi, V.P., Syed, R., et al.: Mammo-FM: Breast-specific foundational model for integrated mammographic diagnosis, prognosis, and reporting. arXiv preprint arXiv:2512.00198 (2025)
12. Glocker, B., Jones, C., Roschewitz, M., Winzeck, S.: Risk of bias in chest radiography deep learning foundation models. Radiology: Artificial Intelligence **5**(6), e230060 (2023)
13. Halling-Brown, M.D., Warren, L.M., Ward, D., et al.: OPTIMAM Mammography Image Database: A large-scale resource of mammography images and clinical data. Radiology: Artificial Intelligence **3**(1), e200103 (2021)
14. He, K., Chen, X., Xie, S., et al.: Masked autoencoders are scalable vision learners. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15979–15988. IEEE (2022)
15. Huang, F., Zhu, J., Yu, Y., et al.: A versatile foundation model for AI-enabled mammogram interpretation. arXiv preprint arXiv:2509.20271 (2025)
16. Jeong, J.J., Vey, B.L., Bhimireddy, A., et al.: The EMory BrEast imaging Dataset (EMBED): A racially diverse, granular dataset of 3.4 million screening and diagnostic mammographic images. Radiology: Artificial Intelligence **5**(1), e220047 (2023)
17. Kendall, E., Hajishafiezahramini, P., Hamilton, M., et al.: Full field digital mammography dataset from a population screening program. Scientific Data **12**(1), 1479 (2025)
18. Khaled, R., Helal, M., Alfarghaly, O., et al.: Categorized contrast enhanced mammography dataset for diagnostic and artificial intelligence research. Scientific Data **9**(1), 122 (2022)
19. Khattak, M.U., Kunhimon, S., Naseer, M., Khan, S., Khan, F.S.: UniMed-CLIP: Towards a unified image-text pretraining paradigm for diverse medical imaging modalities. arXiv preprint arXiv:2412.10372 (2024)
20. Lee, R.S., Gimenez, F., Hoogi, A., Miyake, K.K., et al.: A curated mammography data set for use in computer-aided detection and diagnosis research. Scientific Data **4**(1), 170177 (2017)
21. Loizidou, K., Skouroumouni, G., Pitris, C., Nikolaou, C.: Digital subtraction of temporally sequential mammograms for improved detection and classification of microcalcifications. European Radiology Experimental **5**(1), 40 (2021)
22. McInnes, L., Healy, J., Melville, J.: UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018)
23. Moor, M., Banerjee, O., Abad, Z.S.H., et al.: Foundation models for generalist medical artificial intelligence. Nature **616**, 259–265 (2023)
24. Moreira, I.C., Amaral, I., Domingues, I., et al.: INbreast: Toward a full-field digital mammographic database. Academic Radiology **19**(2), 236–248 (2012)
25. Moutakanni, T., Bojanowski, P., Chassagnon, G., et al.: Advancing human-centric AI for robust X-ray analysis through holistic self-supervised learning. arXiv preprint arXiv:2405.01469 (2024)

26. Nguyen, H.T., Nguyen, H.Q., Pham, H.H., et al.: VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. Scientific Data **10**(1), 277 (2023)
27. Oquab, M., Darcet, T., Moutakanni, T., et al.: DINOv2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
28. Oza, P., Oza, U., Oza, R., et al.: Digital mammography dataset for breast cancer diagnosis research (DMID) with breast mass segmentation analysis. Biomedical Engineering Letters **14**(2), 317–330 (2024)
29. Paschali, M., Chen, Z., Blankemeier, L., et al.: Foundation models in radiology: What, how, why, and why not. Radiology **314**(2), e240597 (2025)
30. Pérez-García, F., Sharma, H., Bond-Taylor, S., et al.: Exploring scalable medical image encoders beyond text supervision. Nature Machine Intelligence **7**(1), 119–130 (2025)
31. Radiological Society of North America: RSNA Screening Mammography Breast Cancer Detection AI Challenge, https://www.kaggle.com/competitions/rsna-breast-cancer-detection, last accessed 2026/06/30
32. Sellergren, A., Kazemzadeh, S., Jaroensri, T., et al.: MedGemma Technical Report. arXiv preprint arXiv:2507.05201 (2025)
33. Siméoni, O., Vo, H.V., Seitzer, M., et al.: DINOv3. arXiv preprint arXiv:2508.10104 (2025)
34. Suckling, J., Parker, J., Dance, D., Astley, S., et al.: Mammographic Image Analysis Society (MIAS) database v1.21. Apollo - University of Cambridge Repository (2015)
35. Zhang, S., Xu, Y., Usuyama, N., et al.: BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)