

IMAGING FOUNDATION MODELS

Pretraining and Adapting a 2D Mammography Foundation Model

Multi-teacher distillation, multi-granularity inference, and **embedding-space evaluation of transferable representations**

AUTHOR

Neurologic AI

MODALITY

2D screening mammography (FFDM and synthetic 2D)

ABSTRACT

Foundation models offer a route to reusable, transferable representations for medical imaging, but progress in mammography has been constrained on two fronts: public screening corpora are small and demographically narrow relative to the variability of real screening practice, and the prevailing pattern remains one narrow classifier per finding, which forfeits shared visual structure across findings and is acutely sample-inefficient for uncommon ones. We describe a 2D

mammography foundation model: a Vision Transformer encoder pretrained on a proprietary corpus of more than four million de-identified screening studies comprising more than 24 million full-field digital mammography (FFDM) and synthetic 2D images acquired at multiple clinical sites across the United States. Pretraining follows DOME (Distillation of Multi-teacher Expertise), a self-distillation framework that augments an online, DINO-style teacher-student objective with a second branch supervised by a frozen expert teacher, so the student inherits knowledge already embedded in a general-purpose medical imaging encoder alongside knowledge extracted directly from large-scale unlabeled data. Downstream adaptation is deliberately parameter-efficient: low-rank adaptation (LoRA) adapters are injected into the encoder and a lightweight multilayer perceptron head is trained over the encoded outputs, and the model is exposed through three inference variants that share a backbone and differ only in the level at which per-image predictions are aggregated: image, laterality, and study. Across eight held-out evaluations spanning four-category BI-RADS breast density, biopsy-confirmed cancer, negative triage, and cardiac device detection (each measured separately on FFDM and on synthetic 2D reconstructions), fine-tuned classifiers achieved a median test ROC-AUC of 0.91 (range 0.87 to 0.99). On a pathology-proven, laterality-level cancer benchmark built by linking BI-RADS 0 screening exams to biopsy outcomes within a 90-day follow-up window, a fine-tuned classifier reached ROC-AUC 0.747 (95% CI 0.700 to 0.792) on a held-out split of 2,400 lateralities, a harder discrimination than it appears, because every examination in that cohort had already been recalled by a radiologist. On the public RSNA challenge corpus, evaluated on that dataset's own held-out split, a classifier fine-tuned from the same encoder reached 0.90. To characterise the representation independently of any classifier head, we analysed embedding geometry in a matched longitudinal cohort of 1,412 case-control pairs: index-exam embeddings of women subsequently biopsied separated from those of matched screen-negative controls far beyond a label-permutation null, and embeddings drifted toward a cohort-derived cancer direction significantly faster in cases than in controls in the years preceding biopsy (paired one-sided Wilcoxon $p = 6 \times 10^{-4}$ at breast level; median paired difference $+0.14 \text{ yr}^{-1}$), with the effect concentrated in the final screening interval before biopsy. Biopsy-negative women traversed the same direction, and malignant and biopsy-negative trajectories did not separate from one another, indicating that the direction encodes biopsy-worthy

mammographic appearance rather than malignancy as such. Validation is predominantly internal, external evaluation is limited, and the cancer benchmark is prevalence-enriched with report-derived labels; the model is accordingly positioned as a developer tool rather than a medical device, requiring fine-tuning, threshold selection, population-specific validation, and independent regulatory authorisation before any clinical or patient-facing use.

0.91

Median test ROC-AUC across eight evaluations

0.747

Pathology-proven, laterality-level ROC-AUC

0.90

ROC-AUC on the public RSNA held-out split

1,412

Matched case-control pairs in the drift cohort

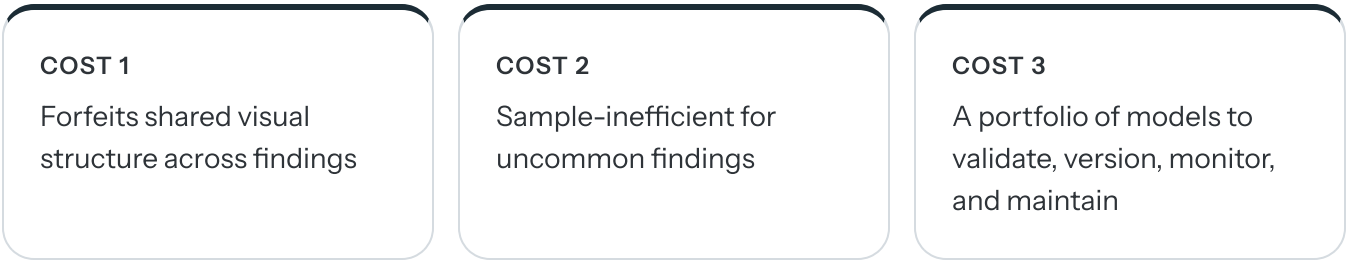
— 01

Introduction

Breast cancer is the most frequently diagnosed malignancy among women worldwide, and population screening with mammography remains the principal instrument for reducing mortality from it [1], [2]. Screening operates at extraordinary volume and under a demanding statistical regime: the overwhelming majority of examinations are normal, the findings that matter are often subtle, and interpretation must sustain high sensitivity without generating recall rates that erode the programme's net benefit. Inter-reader variability under these conditions is well documented [3], and screening volumes continue to grow against a constrained radiology workforce. The resulting pressure has made mammography one of the most intensively studied applications of medical imaging AI, with several systems now reporting stand-alone performance comparable to human readers on retrospective cohorts [4], [5].

Most of that work, however, shares a common architecture of effort: a model is built for one task, on one label, with one dataset, and its representation is discarded when the next task arrives. This pattern has three costs. First, it forfeits shared visual structure: the textural, morphological, and positional regularities that a spiculated mass, an architectural distortion, and a cluster of amorphous calcifications partly hold in common. Second, it is sample-inefficient in exactly the regime where mammography is hardest, namely uncommon

findings for which positive examples are scarce. Third, it produces a portfolio of models that must each be separately validated, versioned, monitored, and maintained, a significant burden for organisations building clinical software at scale.



The foundation-model paradigm proposes a different division of labour [6], [7]. A single encoder is pretrained once on a large, diverse, largely unlabeled corpus, and downstream tasks are addressed by lightweight adaptation of that shared representation. In medical imaging, where annotation is expensive, expert-dependent, and often unavailable at the scale modern architectures reward, the argument is particularly forceful, and self-supervised pretraining has been shown to yield label-efficient, robust representations across modalities including retinal imaging, chest radiography, and pathology [8], [9], [10]. Mammography has begun to follow: contrastive vision-language models trained on public corpora have demonstrated meaningful data-efficiency gains [11], and public challenges have supplied common evaluation ground [12].

Two gaps motivate the work reported here. The first is data. Public mammography corpora, valuable as they are, are small relative to the variability a screening model must absorb: vendor and detector differences, acquisition and post-processing conventions, FFDM versus synthetic 2D reconstructions, positioning and compression practice, and the demographic composition of the screened population. A representation pretrained only on public data is, for most deployment settings, out of distribution. The second gap is more subtle. The field has accumulated a substantial stock of expert models, each trained on distinct data with distinct objectives, and each therefore encoding generic knowledge that a new foundation model is otherwise obliged to rediscover from raw pixels. Standard self-supervised recipes make no use of this stock. Both gaps point the same way: pretraining should draw on large-scale, clinically representative, in-domain data, and it should be able to inherit from existing expert encoders rather than starting from nothing.

This paper describes the model and the evidence assembled around it. Our contributions are as follows. First, we describe a large-scale, clinically sourced pretraining corpus of more than four million de-identified screening studies, comprising more than 24 million FFDM and synthetic 2D images from multiple United States clinical sites, with structured labels

derived from the associated radiology reports through a documented, auditable extraction pipeline. Second, we adapt DOME to mammography; DOME retains the online teacher-student self-distillation objective of DINO and adds a second branch in which a frozen expert teacher supplies additional supervision, letting the student combine data-driven and expert-derived knowledge within a single end-to-end objective. Third, we describe a parameter-efficient, multi-granularity adaptation scheme in which downstream tasks are addressed by LoRA adapters together with a small multilayer perceptron head, and in which the same encoder is exposed through three aggregation granularities, so the unit of prediction matches the unit at which a finding is clinically asserted. Fourth, we evaluate along three complementary axes: fine-tuned classification performance across a panel of findings, performance on a pathology-proven cancer benchmark with biopsy-linked ground truth, and a direct analysis of embedding geometry in a matched longitudinal cohort that characterises the representation independently of any task head. Fifth, we give an explicit account of what the evidence does not establish, and situate the model within a developer-tool rather than a medical-device posture, with the obligations that entails for downstream integrators.

The remainder of this paper is organized as follows. Section 2 reviews related work; Section 3 describes the data; Section 4 details preprocessing, the DOME objective, downstream adaptation, aggregation granularities, and deployment; Section 5 sets out the evaluation protocol; Section 6 reports results; Sections 7 and 8 discuss interpretation and limitations; and Section 9 concludes.

02

Related work

2.1 Deep learning for mammographic interpretation

Convolutional systems for screening mammography have progressed from finding-specific detectors to exam-level cancer classifiers evaluated against reader panels and, in several cases, against subsequent clinical outcomes [4], [5]. A parallel line of work has targeted risk rather than detection, learning to predict future cancer from a currently negative mammogram; Mirai and its predecessors established that mammographic appearance carries risk information beyond conventional density and questionnaire-based models, and that this information is at least partly recoverable by learned representations [13], [14]. The premise of that literature (that subclinical structure precedes clinical detectability) directly

motivates the embedding-drift analysis of Section 6.3. Open artefacts such as the RSNA Screening Mammography Breast Cancer Detection challenge [12] and released ensemble models [15] provide reproducible reference points, and we use both as external comparators.

2.2 Self-supervised pretraining and knowledge distillation

Self-distillation methods learn representations by enforcing agreement between differently augmented views of the same image under a teacher-student pairing in which the teacher is an exponential moving average of the student. DINO [16] established the recipe used here (multi-crop augmentation, a projection head over the class token, and centring with temperature-based sharpening to avoid collapse), and DINOv2 [17] demonstrated that with sufficient scale and curation the resulting features are competitive without any fine-tuning. Related non-contrastive approaches such as BYOL [18] share the momentum-teacher structure. Classical knowledge distillation, by contrast, transfers from a fixed, already-trained teacher [19]. DOME's contribution is to combine the two regimes: the online moving-average teacher supplies scale-driven supervision while a frozen expert teacher supplies transferred supervision, and the student is optimised against both simultaneously. For the expert teacher we use MedImageInsight [20], a general-domain medical imaging embedding model, though the framework is agnostic to the choice and extends to multiple expert teachers.

2.3 Parameter-efficient adaptation

Full fine-tuning of a large encoder for each downstream finding reintroduces precisely the cost that pretraining was meant to eliminate, and in low-data regimes it invites overfitting. Low-rank adaptation [21] freezes the pretrained weights and injects trainable low-rank factors into selected projections, so adaptation scales with the chosen rank rather than the full parameter count. We adopt this approach for all downstream adaptation, with a small multilayer perceptron head over the encoded representation. This has an operational consequence beyond compute: because the backbone is invariant across tasks, a deployed encoder can serve many adapters, and version control reduces to pinning one backbone checkpoint together with a set of small adapter artefacts.

2.4 Documentation and governance of imaging models

Model cards [22] and datasheets for datasets [23] have become the expected vehicle for communicating intended use, training-data provenance, evaluation conditions, and known limitations. We follow both conventions: each released inference variant ships a structured

model card, and each training or benchmark dataset ships a dataset card documenting composition, labelling methodology, annotation quality, and known biases. This matters for a foundation model in particular, because the entity released is not a finished clinical function but an input to someone else’s device, and the recipient’s regulatory obligations depend on information that only the developer of the representation holds.

03

Data

3.1 Pretraining corpus

The pretraining corpus comprises de-identified 2D screening mammography acquired at multiple clinical sites across several regions of the United States, together with the associated radiology reports. It includes both FFDM and synthetic 2D reconstructions, restricted to the standard screening projections, craniocaudal (CC) and mediolateral oblique (MLO), and covers predominantly adult screening populations. Table 1 summarises the corpus and the subsets used at each pretraining stage.

Quantity	Value	Notes
Screening studies (corpus)	> 4,000,000	De-identified; multi-site United States
2D images (corpus)	> 24,000,000	FFDM and synthetic 2D combined
Projections retained	CC, MLO	Bilateral standard screening views
DINO stage (images)	≈ 9,000,000	Online self-distillation, mammography only
DOME stage (images)	≈ 2,000,000	Expert-augmented distillation, from DINO
Population	Adult screening	Predominantly female; multi-site US
Report linkage	Available for corpus	Source of structured findings labels

Table 1. Pretraining corpus and staged subsets. Study and image counts describe the full available training corpus; the self-supervised stages were run on the subsets indicated.



Figure 1. From corpus to released encoder: the two staged self-supervised subsets (Table 1 and Section 4.2).

3.2 Structured label extraction from reports

Ground truth for supervised stages derives from the radiology reports authored by interpreting radiologists at the time of acquisition. Findings are extracted from free text using a combination of large language model prompting, regular-expression matching, and radiologist review, and the extraction method, version, label definitions, permitted values, audit procedure, and error-correction plan are documented as a condition of release. This design reflects a lesson from a parallel chest-radiography programme, where a two-tier extraction system reached 96.9% agreement at full coverage on a core label set, against 79.0% agreement at 16.3% coverage for an earlier rule-based approach, an improvement in both accuracy and, more consequentially, in the fraction of the corpus for which any label exists at all. The mammography label taxonomy spans several families: suspected malignancy; imaging anomalies (mass, calcification, asymmetry, architectural distortion); external objects such as clips, markers, and pacemakers; image-quality attributes covering exposure, motion, positioning, and breast coverage; and BI-RADS breast density category [24].

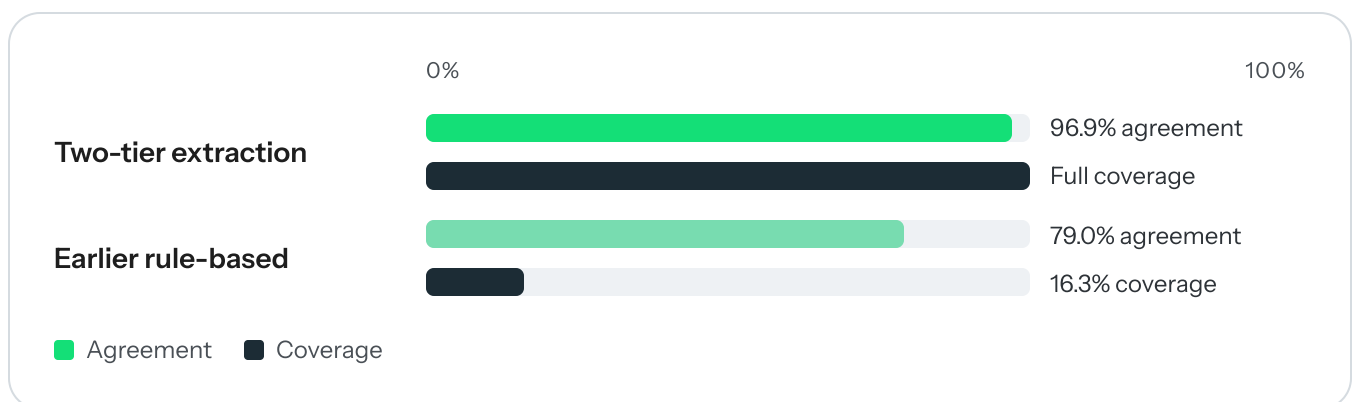


Figure 2. The parallel chest-radiography programme: agreement and coverage on a core label set, drawn to scale.

3.3 Pathology-proven cancer benchmark

Report-derived malignancy labels are necessary at pretraining scale but insufficient for a defensible cancer benchmark, because a screening report records suspicion rather than outcome. We therefore constructed a separate dataset in which laterality-level malignancy is anchored to tissue diagnosis. Only screening examinations assessed BI-RADS 0 (incomplete, requiring additional imaging) were eligible, and each was linked to any pathology from biopsy occurring within a 90-day clinical follow-up window. A breast was labelled positive if it was flagged for further evaluation and had a concordant follow-up biopsy demonstrating malignancy, and negative if it was flagged and had at least one follow-up biopsy from that breast, all benign. Breasts without follow-up biopsy, or with discordant findings, were excluded. Because all images from the same study and laterality inherit the same label, the effective unit of analysis is the breast, mapped down to individual SOPInstanceUID s. Table 2 reports the composition.

LABELLED POSITIVE

Flagged for further evaluation, with a concordant follow-up biopsy demonstrating malignancy

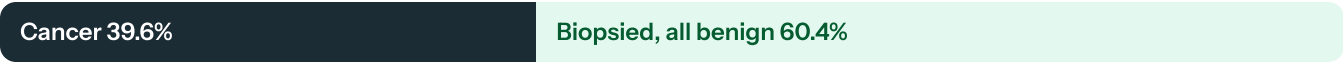
LABELLED NEGATIVE

Flagged, with at least one follow-up biopsy from that breast, all benign

Attribute	Value	Detail
Patients	2,028	Screening exams only, BI-RADS 0
Studies	2,053	Bilateral CC and MLO required
Series / Instances	2,326 / 5,163	Instance-level label mapping
Labels	5,163	No missing labels
Cancer = True	2,047 (39.6%)	Concordant malignant biopsy ≤ 90 d
Cancer = False	3,116 (60.4%)	Biopsied, all benign
Age (years)	60.2 ± 11.8	Range 30–90
Sex	99.8% female	0.2% other or unknown
Race / ethnicity	Not available	Rarely present in source reports

Attribute	Value	Detail
Geography	United States	Multi-site
Minimum resolution	448 × 448	Lower-resolution images excluded

Table 2. Composition of the pathology-proven laterality-level cancer benchmark.

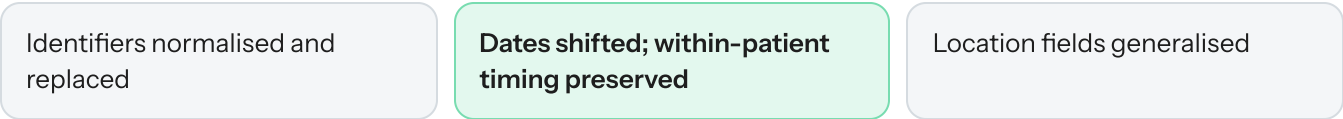


Label balance across 5,163 labels, drawn to scale

Label quality was assessed rather than assumed. A multi-stage extraction pipeline parsed the screening assessment category, identified the suspicious laterality for BI-RADS 0 exams, interpreted follow-up pathology to recover biopsied lesions with their laterality and diagnostic class, and linked screening to outcome within the fixed window. An MQSA-qualified radiologist with six years of experience reviewed a 1.1% sample of reports, verifying final extracted labels together with the intermediate outputs of each extraction step; on that sample, macro-averaged F1 for lesion diagnostic classification and for left-versus-right assignment were both 1.0. The review was of report text and extracted labels only; no direct image adjudication was performed, and we treat this as a material limitation rather than a footnote (Section 8).

3.4 De-identification

All data enter the corpus through a fixed de-identification pipeline. Identifiers are normalised and replaced with internally generated ones. Dates are transformed so as to obscure absolute calendar dates while preserving within-patient relative timing, so that longitudinal intervals and sequencing survive intact. Location-related fields are generalised to reduced granularity. Protected health information is removed from free text by a machine-learning detector, followed by a hidden-in-plain-sight transformation that replaces detected entities with synthetic surrogates. Text burned into image pixels is detected and redacted. Metadata is minimised against an approved attribute allow-list, malformed or low-quality objects are quarantined, and only de-identified images and minimal derived metadata are retained in primary storage. The preservation of within-patient intervals is what makes the longitudinal analysis of Section 6.3 possible at all.



Free-text PHI replaced with synthetic surrogates

Burned-in pixel text redacted

Metadata minimised to an allow-list

04

Model and training

4.1 Preprocessing

Inference and training share a single preprocessing chain, enforced identically at both stages so that deployment-time inputs match training-time statistics. Each image is cropped to the breast region to maximise the fraction of the field occupied by tissue; resized with anti-aliasing to a fixed 896×896 resolution; normalised per channel using standard ImageNet statistics (means 0.485, 0.456, 0.406; standard deviations 0.229, 0.224, 0.225); and then tiled, yielding four 448×448 tiles together with a 448×448 downsampled view of the whole image. The tiling scheme is the key resolution decision. Mammographic malignancy is frequently signalled by structures whose diagnostic content survives poorly under aggressive global downsampling (calcification clusters, subtle spiculation, focal architectural change), yet naive full-resolution transformer inference is prohibitive at corpus scale. Tiling preserves local high-frequency detail in the tiles while retaining global context in the downsampled view, and the model consequently underperforms on inputs whose native resolution falls below 896×896 .

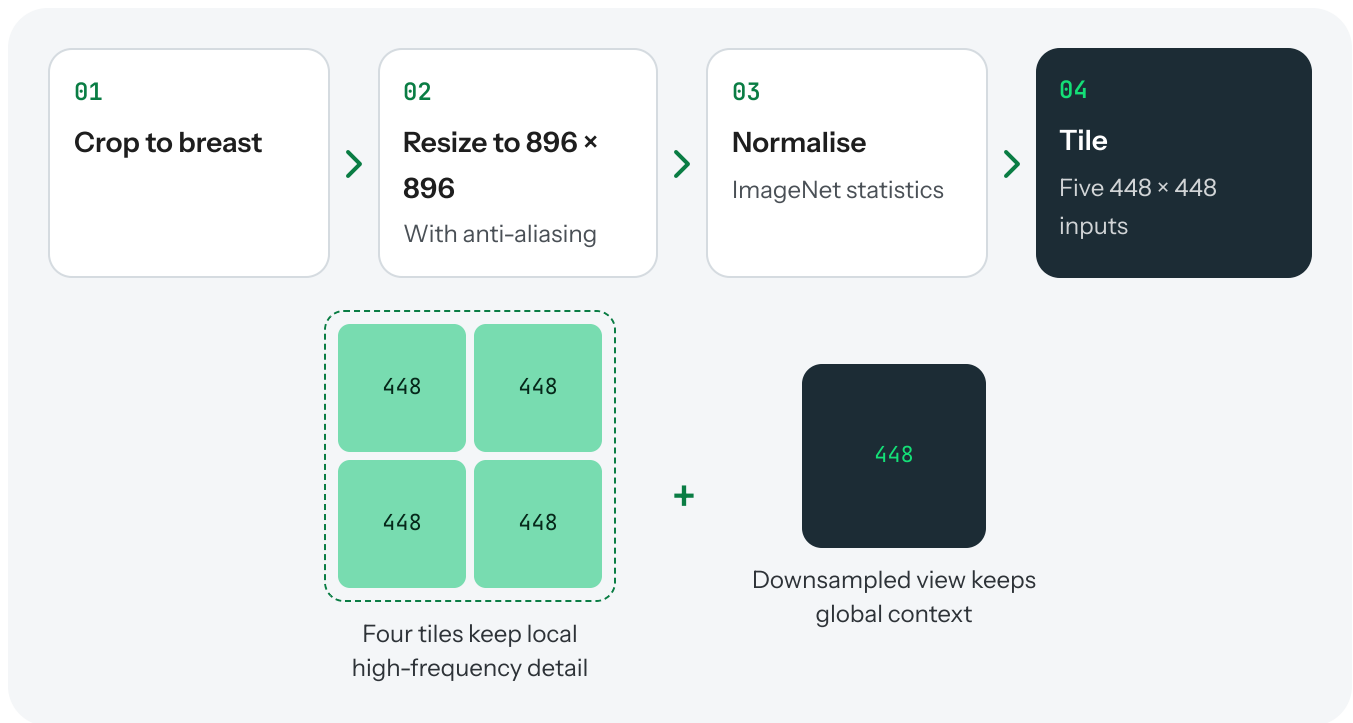


Figure 3. The shared preprocessing chain and the tiling scheme, the key resolution decision.

Eligibility is enforced on DICOM metadata rather than inferred from pixels. Seven attributes are required: `ViewPosition`, `PerformedProcedureStepDescription`, `SeriesDescription`, `ImageLaterality`, `SOPInstanceUID`, `StudyInstanceUID`, and `Modality`. The two description fields may be null but must be present. Studies are excluded when `PerformedProcedureStepDescription` indicates diagnostic rather than screening intent, or an interventional context, through terms such as `diag`, `spot`, `bx`, `biopsy`, `specimen`, or `ultrasound` (with `us` admitted only when `scr` also appears); and when `SeriesDescription` indicates a stereotactic or post-intervention series, through terms such as `STEREO SCOUT`, `POSTBIOPSY`, `POSTMARKER`, `PREFIRE MINUS`, `PREFIRE PLUS`, or `POSTFIRE MINUS`. These filters keep the screening distribution clean; without them, post-biopsy and stereotactic images leak markers and clips that a cancer classifier will learn in preference to tissue.

4.2 Pretraining: expert-augmented self-distillation

The encoder is pretrained with DOME, which optimises a shared student network against two teachers simultaneously.

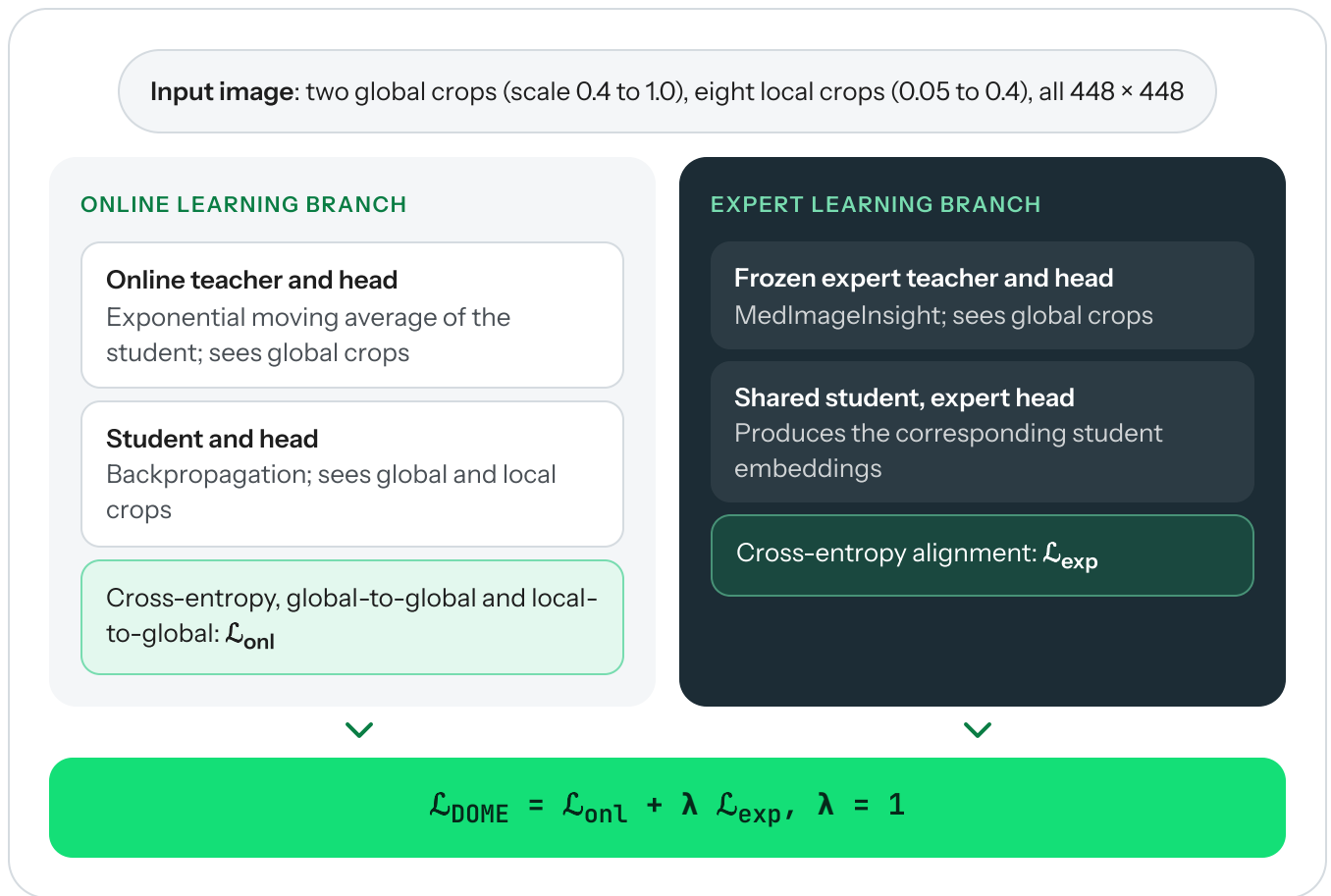


Figure 4. DOME optimises one shared student against an online teacher and a frozen expert teacher simultaneously.

Online learning branch. The online branch comprises a student, an online teacher, and a projection head for each. The student and its head are optimised by backpropagation, while the online teacher and its head are updated as an exponential moving average of the corresponding student parameters. For an input image we extract two global crops and a set of local crops at smaller scale. Global crops pass through the teacher and its head to give one set of embeddings, while global and local crops both pass through the student and its head to give another. Following DINO, consistency is enforced both global-to-global and local-to-global via a cross-entropy objective, so the student must recover from a local patch a representation consistent with the teacher’s view of the whole breast, an inductive bias that suits a modality in which a small region routinely determines the interpretation of the entire image.

Expert learning branch. The expert branch shares the student and adds a frozen expert teacher with its own projection head. Global crops pass through the expert teacher to produce target embeddings, while the student with its expert head produces the corresponding student embeddings, and a cross-entropy loss aligns them. Because the

expert teacher is frozen, this branch is a stable transfer channel rather than a second moving target: it injects knowledge accumulated in a model trained on different data under different objectives, without the instability that co-adapting teachers would introduce. We use MedImageInsight [20] as the expert teacher.

Objective and schedule. The total objective combines the two branches,

$$\mathcal{L}_{\text{DOME}} = \mathcal{L}_{\text{onl}} + \lambda \mathcal{L}_{\text{exp}}, \quad (1)$$

where the weighting coefficient λ balances expert supervision against online self-distillation and is set to 1 in the configurations reported here. Augmentation extracts two global crops at scale range 0.4 to 1.0 and eight local crops at 0.05 to 0.4, all resized to 448×448 .

Optimisation uses AdamW at a learning rate of 2×10^{-4} with a two-epoch warm-up, and projection heads are three-layer multilayer perceptrons. Mammography pretraining proceeded in two stages: a DINO checkpoint was first trained on the internal mammography corpus (≈ 9 million images), and DOME was then trained on an ≈ 2 -million-image subset initialised from that checkpoint. Two initialisation regimes were evaluated: in the first, the DINO checkpoint was trained from random initialisation on mammography alone; in the second, it was itself initialised from a chest-radiography DINO checkpoint before mammography pretraining, testing whether cross-modality radiographic transfer confers additional benefit. The released encoder produces 1,536-dimensional image embeddings.

INITIALISATION REGIME 1

DINO checkpoint trained from random initialisation on mammography alone

INITIALISATION REGIME 2

DINO checkpoint initialised from a chest-radiography DINO checkpoint first

4.3 Downstream adaptation

Adaptation freezes the pretrained encoder and injects LoRA adapters into it, training them jointly with a lightweight multilayer perceptron head over the encoded outputs. For a frozen weight matrix $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, the method introduces trainable matrices $A \in \mathbb{R}^{r \times d_{\text{in}}}$ and $B \in \mathbb{R}^{d_{\text{out}} \times r}$ with rank r much smaller than either dimension, so the effective weight becomes

$$W' = W + BA, \quad (2)$$

and trainable parameters scale as $r(d_{in} + d_{out})$ rather than as the product of the two dimensions. The output head is determined by the label type: Boolean labels yield a binary classifier and a single score, while string or integer labels yield a multiclass classifier with one score per class. All labels in a fine-tuning dataset must sit at a single level of the hierarchy, and mixed-level label sets are rejected. Datasets are supplied as images together with JSONL keyed on `S0PInstanceUID` for image- and laterality-level tasks, or on `StudyInstanceUID` for study-level tasks, and all returned scores lie in the unit interval.

4.4 Aggregation granularity

Findings differ in the unit at which they are clinically asserted. A pacemaker is a property of an image; malignancy is asserted of a breast; breast density is reported for the examination. Forcing all three onto one granularity either discards information or fabricates it. The released variants therefore share the encoder and differ only in aggregation, as summarised in Table 3. At image level, training and scoring operate per image, and one score is returned for every image, together with both the internal identifier and the supplied reference; all seven required DICOM attributes must be populated. At laterality level, training operates per image but predictions are aggregated over images sharing an `ImageLaterality`, returning one score per breast; if one breast is absent its output is null, if both are absent inference fails, and at least one image per laterality must carry the required attributes. At study level, predictions are aggregated across all images in the study regardless of laterality, with at least one image per study required to carry the attributes.

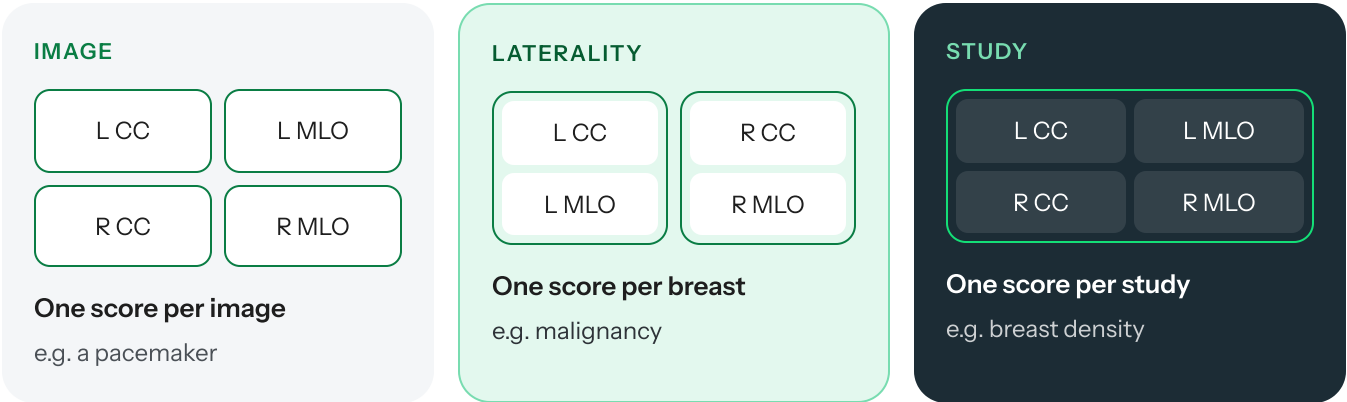


Figure 5. One encoder, three aggregation granularities, matched to the unit at which a finding is clinically asserted.

Variant	JSONL key	Prediction unit	Representative task and output
Image	SOPInstanceUID	Per image	External objects; scores array with image reference and identifier
Laterality	SOPInstanceUID	Per breast	Cancer; e.g. left 0.92, right 0.07; null for an absent breast
Study	StudyInstanceUID	Per study	Density; per-class scores over categories A to D

Table 3. Released inference variants. All variants share backbone, preprocessing, and training corpus.

The choice of granularity is a substantive modelling decision rather than a formatting one, and Section 6.3 provides direct evidence for its consequences: restricting analysis to the affected breast yields a materially stronger signal than pooling both breasts, because the unaffected breast contributes variance without signal.

4.5 Deployment

The model is served over a secure API returning results in a fixed JSON schema, with versioned checkpoints so that a specific model version can be frozen for traceability, and with logging sufficient for usage accounting. The service target is that more than 99% of inference requests complete end-to-end within 120 seconds for up to four high-resolution images, inclusive of de-identification, embedding generation, and inference, and exclusive of customer-side transmission and result download. Integrators remain responsible for the data-privacy and security obligations applying to their own datasets.

> 99%
 Of inference requests
 complete end-to-end

120 s
 Service target

4
 High-resolution images per
 request, at most

Evaluation protocol

A foundation model cannot be evaluated in the way a finished classifier is, because it has no intrinsic task. We therefore evaluate along three axes that answer different questions.

AXIS 1

Transferability

Fine-tuned classifiers across a panel of findings

AXIS 2

Outcome-anchored performance

The pathology-proven benchmark

AXIS 3

Representation geometry

The embedding space, with no trained classifier

The first axis is transferability. For each finding, adapters and an output head are fine-tuned on the frozen encoder, with checkpoint selection on validation ROC-AUC and all reported metrics computed on a held-out test partition disjoint from training and validation. Independence of training and test partitions is enforced procedurally and documented as a release condition. Discrimination is summarised by ROC-AUC, which is threshold-free and therefore appropriate for a component whose operating point is chosen downstream by the integrator. Sensitivity, specificity, and F1 are additionally reported at the operating point appropriate to each task, since a threshold-free summary alone would hide the sharp trade-off a high-sensitivity triage configuration imposes on specificity.

The second axis is outcome-anchored performance. On the pathology-proven benchmark, a classifier fine-tuned on that dataset is compared against a classifier fine-tuned only on public RSNA challenge data [12], with both evaluated on the same internal held-out split. This isolates the contribution of in-domain, outcome-linked supervision from that of the encoder itself.

The third axis is representation geometry. Task metrics conflate the quality of a representation with the quality of the head trained on it, and they say nothing about what the representation encodes. We therefore analyse the embedding space directly in a matched longitudinal cohort, using no trained classifier at all. Women with at least three screening-eligible examinations at annual intervals (consecutive exams 365 to 425 days apart) were assembled; cases had a biopsy linked to a screening exam, while controls had none and were assigned a synthetic index at their most recent BI-RADS 1 or 2 exam.

Matching was one-to-one without replacement, exact on breast density category and on number of priors, and within three years on age at index, giving 1,412 matched pairs from 1,412 cases (of which 534 were malignant and 878 biopsy-negative) and 1,412 controls.

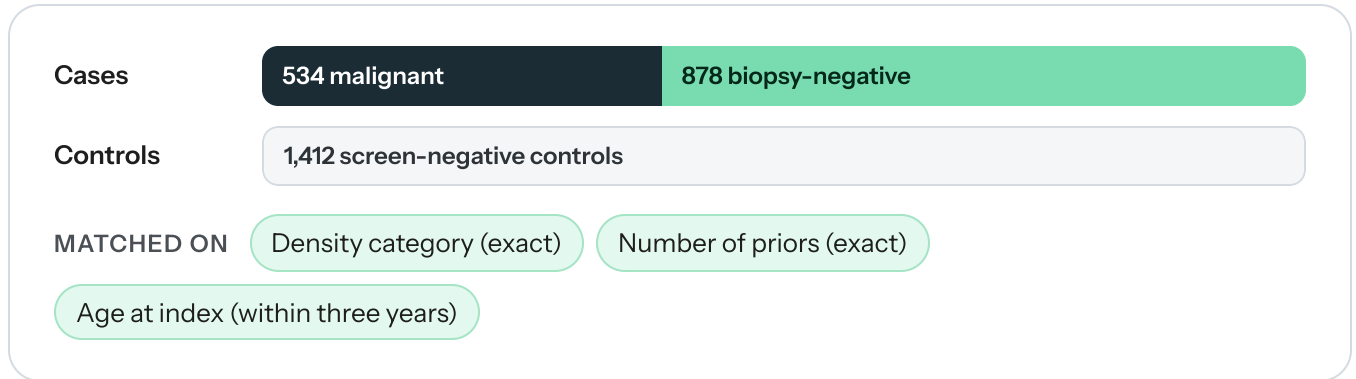


Figure 6. The matched longitudinal cohort: 1,412 one-to-one pairs, drawn to scale.

Within this cohort we define a *cancer direction* as the L_2 -normalised difference between group-mean index embeddings, computed leave-one-out per patient so that no patient's own index embedding contributes to both the direction and their measured displacement; without this correction, the interval closest to the index exam is optimistically biased. Per-exam embeddings are formed at breast level as the mean over images of the affected side (inherited from the matched case for controls), and at study level as the mean over all images. For consecutive exams separated by an interval of Δt years, the primary metric is the signed velocity

$$v_{\text{signed}} = \frac{\langle e_{\text{newer}} - e_{\text{older}}, \hat{u} \rangle}{\Delta t}, \quad (3)$$

expressed in units of yr^{-1} and positive toward the cancer direction. The primary test collapses this quantity to one value per patient and applies a one-sided paired Wilcoxon signed-rank test on matched pairs. Exploratory per-interval tests use Holm-Bonferroni correction [25], and a confirmatory linear mixed-effects (LME) model regresses signed velocity on time-to-index interacted with group, with pair and nested patient random effects; its interaction coefficient (yr^{-2}) measures differential acceleration of drift. Static separation of group-mean index embeddings is assessed against a label-permutation null of 1,000 shuffles. The entire pipeline was additionally run with an external, public-data-

trained comparator model [11] on the identical cohort, so that any finding attributable to one model’s training distribution can be distinguished from a general property of mammographic embeddings.

Results

6.1 Transfer across the finding panel

Four representative tasks spanning the finding panel were evaluated in full: four-category BI-RADS breast density, biopsy-confirmed cancer, negative triage, and cardiac device detection. Each was evaluated separately on FFDM and on synthetic 2D reconstructions, giving eight evaluations in total, on balanced held-out test sets disjoint from the corresponding training and validation partitions. Checkpoints were selected on highest validation AUC, and all reported figures are computed on the test partition. Table 4 gives the breakdown. Across the eight evaluations the median test ROC-AUC is 0.91, with a range of 0.87 to 0.99.

Task	Image type	ROC-AUC	PR-AUC	Sensitivity	Specificity	F ₁
Density	FFDM	0.93	0.81	0.75	0.92	0.75
Density	Synthetic 2D	0.95	0.84	0.79	0.93	0.79
Cancer	FFDM	0.88	0.84	0.80	0.80	0.80
Cancer	Synthetic 2D	0.88	0.85	0.80	0.81	0.80
Triage	FFDM	0.89	0.84	0.99	0.63	0.80
Triage	Synthetic 2D	0.87	0.88	0.99	0.54	0.82
Cardiac device	FFDM	0.99	0.93	0.96	0.99	0.96
Cardiac device	Synthetic 2D	0.99	0.99	0.95	1.00	0.97

Table 4. Held-out test performance of classifiers fine-tuned from the encoder. Density is the four-category BI-RADS classification; cancer is biopsy-confirmed malignancy; triage is the high-sensitivity negative-triage configuration.

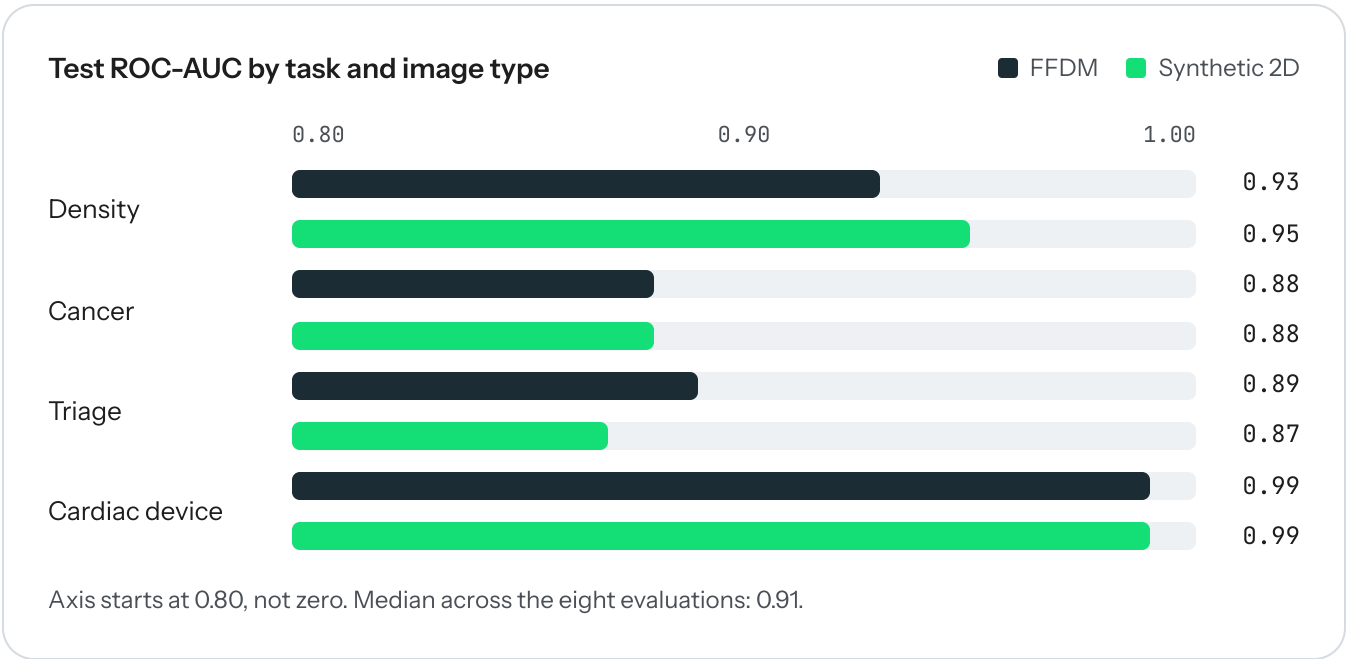


Figure 7. Test ROC-AUC from Table 4, bars sized by the reported values.

Density metrics are macro-averaged across the four BI-RADS categories, with one-versus-rest ROC-AUC; the remaining tasks are binary. Operating points differ by task and were chosen for the intended use: cancer at an 80% sensitivity threshold, triage at 99.9% sensitivity, cardiac device at 95% sensitivity. Test partitions are 1,500 studies for density (7,506 FFDM and 6,616 synthetic 2D images), 175 studies for cancer and triage (358 and 363 images), and 180 and 168 studies for cardiac device (889 and 727 images). All test sets are balanced by construction, so specificity and F1 will not carry over to a screening-prevalence distribution.

The spread across tasks is the informative part. Cardiac device detection reaches 0.99 on both image types, which is what a spatially compact, high-contrast, unambiguously labelled finding should yield, and it sets a practical ceiling for the encoder rather than a claim about clinical difficulty. Breast density follows at 0.93 and 0.95 one-versus-rest, though the macro sensitivity of 0.75 and 0.79 shows that four-category discrimination is considerably harder than the AUC alone suggests, and that errors concentrate in the adjacent-category confusions that dominate reader disagreement. Cancer sits at 0.88 on both image types, the genuinely difficult soft-tissue discrimination. The negative-triage configuration is the clearest illustration of why a threshold-free summary is insufficient: at 99% sensitivity it

retains only 0.63 specificity on FFDM and 0.54 on synthetic 2D, so it would clear roughly half to two-thirds of negative examinations rather than the large majority a screening programme would need.

Two cautions attach to these numbers. The cancer and triage test sets are small (175 studies and around 360 images per image type), so their confidence intervals are correspondingly wide, and the agreement between FFDM and synthetic 2D at 0.88 should not be read as evidence of precision. And every test set here is balanced by construction rather than drawn at screening prevalence, so specificity and F1 will not transfer to a deployment distribution without recalibration.

6.2 Pathology-proven cancer benchmark

Beyond the internal panel, cancer detection was assessed on two further corpora. A classifier fine-tuned on the pathology-proven dataset achieved a laterality-level ROC-AUC of 0.747 (95% CI 0.700 to 0.792) on a held-out test split of 2,400 lateralities. A classifier fine-tuned from the same encoder on public RSNA challenge data and evaluated on that dataset’s own held-out split reached ROC-AUC 0.90. Table 5 records both.

Fine-tuning corpus	Evaluation set	Test ROC-AUC
Pathology-proven dataset	Internal held-out, BI-RADS 0 only (n = 2,400 lateralities)	0.747 (95% CI 0.700–0.792)
RSNA challenge data	RSNA held-out, unselected screening population	0.90

Table 5. Cancer detection by fine-tuning corpus and evaluation set. The two rows are not comparable to one another: they are evaluated on different populations, as discussed below.

These two figures must not be read as a like-for-like comparison; the difference between them is a property of the evaluation populations rather than of the encoder. The internal split contains only examinations already assessed BI-RADS 0, so the task there is malignant-versus-benign discrimination within a population a radiologist has already recalled, with positives at 39.6% prevalence. The RSNA split is an unselected screening population, in which the large majority of examinations are unambiguously normal; those easy negatives raise ROC-AUC substantially without the underlying discrimination having improved. The 0.90 figure is nonetheless the single most useful external result available for this encoder,

since both the images and the labels originate outside the proprietary corpus, and it is the strongest evidence reported here that the representation transfers beyond its training distribution.

Three features of the benchmark govern how the 0.747 should be read, and none is incidental. First, cohort construction restricted eligibility to BI-RADS 0 examinations, so every case had already been judged suspicious by a radiologist; the task is malignant-versus-benign discrimination within an already-recalled population, on which an AUC near 0.75 means something quite different from the same figure on an unselected cohort. Second, positive prevalence is 39.6%, far above population prevalence, so calibration will not transfer and thresholds must be re-derived on the deployment distribution. Third, the 90-day follow-up window will miss slow-growing cancers, care delivered outside the contributing network, and patients lost to follow-up, each of which pushes true positives into the negative class and depresses measured performance.

Already recalled

Eligibility restricted to BI-RADS 0 examinations

Enriched prevalence

39.6% positive; thresholds must be re-derived

90-day window

Pushes some true positives into the negative class

6.3 Representation geometry in a matched longitudinal cohort

6.3.1 Static separation at the index exam

Before testing any temporal hypothesis we verified that the cancer direction is meaningful at all. Group-mean index embeddings separate well beyond the label-permutation null for both case subgroups, at both granularities (Table 6). Absolute distances are not comparable across models, since the embedding spaces differ in dimension and feature magnitude, but permutation *p*-values are.

Model	Comparison	n_1/n_2	$\ \Delta_{\text{mean}}\ $ (null mean $\pm$ SD)	Perm. <i>p</i>
Internal	Malignant vs control	274/861	1.49 (0.60 $\pm$ 0.15)	1×10^{-3}
External	Malignant vs control	274/862	0.78 (0.39 $\pm$ 0.08)	2×10^{-3}
Internal	Biopsy-neg. vs control	591/861	0.92 (0.47 $\pm$ 0.11)	3×10^{-3}
External	Biopsy-neg. vs control	592/862	0.61 (0.30 $\pm$ 0.06)	1×10^{-3}

Model	Comparison	n_1/n_2	$\ \Delta_{\text{mean}}\ $ (null mean $\pm$ SD)	Perm. p
Internal	Malignant vs biopsy-neg.	274/591	0.88 (0.64 $\pm$ 0.16)	0.08
External	Malignant vs biopsy-neg.	274/592	0.53 (0.41 $\pm$ 0.09)	0.09

Table 6. Static separation of group-mean index embeddings at breast level, against a 1,000-shuffle label-permutation null. Internal embeddings are 1,536-dimensional; the external comparator is 2,048-dimensional.

Two observations follow. The internal encoder produces group-mean separations roughly 1.5 to 2 times larger than the external comparator at every comparison and granularity, consistent with its in-distribution status relative to this cohort; because the permutation nulls scale proportionally with feature magnitude, the p -values remain in the same significance class. And the malignant-versus-biopsy-negative comparison fails to reach significance in both models at breast level: the first indication that the direction is not a malignancy axis.

6.3.2 Longitudinal drift

Embeddings of malignant cases drift toward the cancer direction significantly faster than those of matched controls, and the effect is concentrated in the year immediately preceding biopsy. Table 7 reports the primary, exploratory, and confirmatory tests side by side for both encoders and both granularities.

Test	Breast, internal	Breast, external	Study, internal	Study, external
Patient Wilcoxon p	6×10^{-4}	9×10^{-6}	0.02	0.04
Median paired diff. (yr^{-1})	+0.14	+0.08	+0.06	+0.03
Interval-1 Holm p	4×10^{-3}	2×10^{-5}	0.04	0.01
LME interaction β (yr^{-2})	+0.17	+0.12	+0.04	+0.03
LME interaction p	0.04	2×10^{-3}	0.27	0.22

Table 7. Primary drift analysis, malignant versus matched controls. Wilcoxon tests are one-sided paired on per-patient mean signed velocity. Interval 1 is the screening interval closest to the index exam, Holm-corrected across intervals. The mixed-effects rows report the time-by-group interaction.

The temporal profile is the most informative feature of this result. A static, time-invariant difference between groups (an incompletely controlled density confound, for instance) would elevate signed velocity across all intervals. Instead the paired-difference distributions collapse toward zero by intervals two and three in both models. That is the pattern expected of precursor-stage imaging change: with typical tumour doubling times of roughly 100 to 300 days [26], clinically relevant subclinical features should be detectable in the final screening interval before biopsy and largely absent two or more years out. Statistical power also falls sharply at earlier intervals, from 271 breast-level pairs at interval one to 85 at interval three, so the attenuation is partly a power effect and should not be over-read. Breast-level analysis outperforms study-level analysis in both models, supporting the interpretation that the drift is localised to suspicious tissue rather than being a global property of case patients, and vindicating the decision to expose laterality-level aggregation as a first-class variant.

6.3.3 The direction is not malignancy-specific

Biopsy-negative women also drift toward the cancer direction (at breast level with $p = 2 \times 10^{-3}$ for the internal encoder and 9×10^{-7} for the external comparator), and the direct malignant-versus-biopsy-negative comparison of drift trajectories reaches significance in neither model at either granularity (Mann-Whitney $p = 0.27$ at breast level for both encoders, and 0.56 and 0.10 respectively at study level). The direction therefore encodes biopsy-worthy mammographic appearance (suspicious masses, architectural distortion, and calcification) rather than malignancy as such. This is a negative result about specificity, and a useful one: it identifies the ceiling a triage application built on this axis would encounter, and it locates the limitation in the biopsy-driven enrichment of the cohort rather than in the encoder, since both an in-distribution and an out-of-distribution model agree.

WHAT THE DIRECTION ENCODES

Biopsy-worthy mammographic appearance (suspicious masses, architectural distortion, and calcification) rather than malignancy as such.

6.3.4 Robustness

Four checks support the primary finding. Replacing the leave-one-out direction with a globally computed one inflates effects by 14% to 31% in median difference and by roughly half an order of magnitude in p -value, confirming that the leave-one-out correction removes real optimism and that the reported figures are the conservative ones. The permutation-null separation test was fixed before drift testing and passed for both models on every case-arm

comparison. Velocity tests are scale-invariant, so the choice of annual units is presentational only. Most importantly, agreement between an in-distribution and an out-of-distribution model on the primary finding (paired-Wilcoxon significance with a positive median difference, concentration in the first interval, and a significant breast-level mixed-effects interaction) argues against a single-model artefact.

CHECK 1

Leave-one-out direction gives the conservative figures

CHECK 2

Permutation-null separation fixed before drift testing

CHECK 3

Velocity tests are scale-invariant

CHECK 4

In-distribution and out-of-distribution models agree

07

Discussion

7.1 What the results establish, and what they do not

Taken together the three evaluation axes support a specific and bounded claim: the encoder produces representations that transfer efficiently to a broad panel of mammographic findings, that carry outcome-relevant information anchored to tissue diagnosis, and that encode temporally structured signal detectable before clinical diagnosis. They do not establish that the model detects cancer in the sense that a cleared device detects cancer. The distinction is not rhetorical. Detection claims require a defined intended use, a fixed operating point, prospective or otherwise unbiased evaluation on the deployment population, and regulatory authorisation. What we report is that a shared representation, adapted cheaply, reaches performance levels that make it a credible substrate for building such claims, and that the representation's structure has properties, both encouraging and limiting, that a developer should know before building on it.

SUPPORTED

- Transfers efficiently to a broad panel of findings
- Carries outcome-relevant information anchored to tissue diagnosis
- Encodes temporally structured signal detectable before clinical diagnosis

NOT ESTABLISHED

That the model detects cancer in the sense that a cleared device detects cancer.

7.2 In-distribution pretraining amplifies signal and variance together

The comparison between the in-distribution internal encoder and the out-of-distribution external comparator produced a result worth stating plainly, because it inverts a common expectation. In-distribution status delivered larger effects: group-mean separations 1.5 to 2 times larger, median paired velocity differences of $+0.14$ against $+0.08 \text{ yr}^{-1}$, and a steeper mixed-effects interaction of 0.17 against 0.12 yr^{-2} . It did not deliver uniformly smaller p -values; the primary breast-level Wilcoxon p was in fact larger for the internal encoder (6×10^{-4} against 9×10^{-6}). The explanation is a signal-to-noise trade-off rather than a defect: an encoder that projects cohort-specific information more faithfully also amplifies the irreducible patient-to-patient heterogeneity of mammographic appearance, and rank-based tests depend on the ratio of median displacement to spread, not on absolute magnitude. Where in-distribution status helped unambiguously was the mixed-effects interaction coefficient (which by construction isolates time-dependent change from constant offsets) and the subgroup with longer surveillance histories, where the study-level test for patients with three or more priors was markedly stronger for the internal encoder ($p = 2 \times 10^{-3}$) than for the comparator (0.01). The practical implication for anyone benchmarking foundation models is that effect size and significance can move in opposite directions across encoders, and reporting only one of them will mislead.

Effect size and significance can move in opposite directions across encoders, and **reporting only one of them will mislead.**

7.3 Why multi-granularity aggregation earns its complexity

Shipping three variants of one model imposes real cost in documentation, testing, and integration surface. The evidence indicates the cost is justified. Breast-level analysis dominated study-level analysis on every drift test in both models, because pooling the unaffected breast into a case patient’s representation adds variance without adding signal: exactly the failure mode a single fixed granularity would impose on any laterality-specific finding. Conversely, breast-level analysis is unavailable when laterality is unknown; in the drift cohort, 38% of cases lacked a recorded biopsy side and were excluded from breast-level analysis together with their matched controls. And image-level scoring is the only sensible granularity for per-image attributes such as positioning quality or the presence of a marker. No single choice serves all three cases, and the alternative to exposing the choice is making it silently on the developer’s behalf.

7.4 Report-derived labels as a scaling instrument

Deriving labels from reports is what makes corpus-scale supervision tractable, and the chest-radiography experience quantifies the gain: full coverage at 96.9% agreement, against 16.3% coverage at 79.0% for a rule-based predecessor. Coverage is the more consequential axis, because a label that exists for one sixth of the corpus cannot supervise the rest of it. But the instrument has characteristic failure modes that survive any improvement in extraction accuracy. Radiologists hedge, and phrases such as *possible* or *cannot exclude* admit no clean binary reading. Reports record what was described, not everything that was present, so absence of a label is not evidence of absence of a finding. Each study is labelled independently, so information from a patient’s prior examinations does not inform the label. And documentation conventions drift over time, so a corpus spanning years is not distributionally homogeneous. The pathology-proven benchmark exists precisely because these limitations are unacceptable in an outcome-anchored evaluation, though it inherits its own, since its labels too are report-derived, with only 1.1% of cases reviewed by a radiologist and no image adjudication at all.

Hedging

No clean binary reading

Omission

Absence of a label is not absence of a finding

Independence

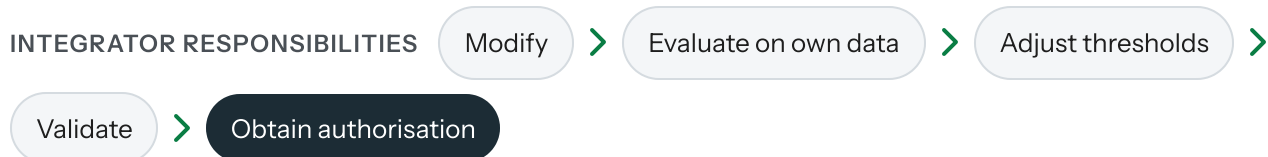
Prior examinations do not inform the label

Drift

Conventions change over time

7.5 Positioning as a developer tool

The model is positioned as a developer tool rather than a medical device. It is not intended for diagnostic use and must be modified toward a customer-defined intended use before it can support one; the applicable regulatory pathway depends on the integrator's intended use and final software application. Concretely, integrators are responsible for making necessary modifications to the model, evaluating it on their own data and target population, adjusting decision thresholds to obtain the sensitivity and specificity their use case requires, validating the final product, and obtaining any required marketing authorisations. This division of responsibility is only defensible if the developer discloses enough for the integrator to discharge it, which is the reason for the structured model and dataset cards, the documented preprocessing and eligibility rules, the traceable training-data record, and the explicit limitation inventory of Section 8. A foundation model shipped without that disclosure transfers risk without transferring the information needed to manage it.



— 08

Limitations

We group limitations by what they threaten rather than by where they arise.

8.1 Generalisation

Experiments and metrics were obtained predominantly on proprietary datasets. External validation rests on a single public dataset, the RSNA challenge corpus, on which a fine-tuned classifier reached ROC-AUC 0.90; one external dataset from one country does not establish generalisation, so results on other external data may differ, and independent evaluation on the target population is a precondition of use rather than an optional confirmation.

Demographic diversity is bounded by the available proprietary corpus, drawn from several regions of the United States; international and underserved populations, and workflows and equipment differing from United States practice, are not represented. Race and ethnicity information was rarely present in the source data, which materially constrains subgroup fairness assessment: we cannot rule out differential performance across groups we cannot

measure, and we do not interpret the absence of measured disparity as evidence of its absence. Exposure to rare disease cohorts is likewise limited, and performance on rare pathologies may be lower than the panel figures suggest.

8.2 Inputs and failure modes

Performance degrades on images below 896×896 native resolution, a direct consequence of the tiling scheme of Section 4.1, and low-quality images may be misclassified or rejected outright. Eligibility depends on DICOM metadata being populated and conventionally used, so sites whose `PerformedProcedureStepDescription` or `SeriesDescription` conventions differ from those the filters assume may see series incorrectly included or excluded.

8.3 Benchmark and cohort design

The pathology-proven benchmark is prevalence-enriched at 39.6% positive and restricted to BI-RADS 0 examinations, so it does not mirror screening prevalence and calibration will not transfer without recalibration. Its labels are report-derived by automated pipeline; a radiologist reviewed 1.1% of cases, with one reviewer per case, and no direct image review was performed. The 90-day follow-up window may misclassify slow-growing cancers, care delivered outside the contributing network, and patients lost to follow-up as negatives. The drift cohort is single-source: both encoders were evaluated on the same 1,412 matched pairs, so external replication at an independent site remains outstanding. Thirty-eight percent of drift-cohort cases lacked a recorded biopsy laterality and were excluded from breast-level analysis along with their matched controls, reducing power for the strongest granularity; and approximately 0.6% of cases had both a malignant and a benign biopsy linked to the index study and were classified as malignant under an any-positive rule, with a sensitivity analysis dropping them deferred. Finally, the study-level mixed-effects interaction was non-significant in both models despite significant patient-level Wilcoxon results, indicating that the pooled time-by-group trend is noisier than per-patient averaged velocity.

8.4 Interpretation

The cancer direction is not malignancy-specific in either model, and no significant separation of malignant from biopsy-negative trajectories was demonstrated, so claims about the representation should be framed in terms of biopsy-worthy appearance rather than malignancy. Absolute embedding distances and velocities are not comparable across models with different dimensions and feature magnitudes; only p -values and effect

directions are. And per-finding performance metrics, confusion matrices, and full training-data characterisation are held in the release record rather than reproduced here, which limits independent scrutiny of this document in isolation.

09

Conclusion

This paper described a 2D mammography foundation model pretrained on more than four million de-identified screening studies using expert-augmented self-distillation, adapted downstream through low-rank adaptation and a lightweight classification head, and released in three aggregation granularities matched to the units at which mammographic findings are clinically asserted. Fine-tuned classifiers reached a median held-out ROC-AUC of 0.91 across eight evaluations, ranging from 0.87 for negative triage on synthetic 2D to 0.99 for cardiac device detection on both image types; a classifier fine-tuned on biopsy-anchored labels reached 0.747 on a laterality-level cancer benchmark drawn from an already-recalled population, and 0.90 on the public RSNA corpus; and direct analysis of the embedding space demonstrated temporally structured, pre-diagnostic signal concentrated in the final screening interval and localised to the affected breast, while also showing that the axis carrying that signal encodes biopsy-worthy appearance rather than malignancy as such.

Several directions follow from the evidence and the current requirement set. A vision-language successor is specified to accept prior reports and text prompts alongside images and to generate a draft findings section for a screening report at the laterality level, with entity-level acceptance targets for mass, calcification, architectural distortion, and asymmetry, together with BI-RADS 0-versus-1/2 discrimination and four-category density accuracy. Localisation is a further requirement axis: given a bounding box, resolving side, view, quadrant, and depth, evaluated in weighted F_1 . Extension beyond 2D to digital breast tomosynthesis has been identified as a priority modality gap requiring dedicated ingestion, labelling, training, and optimisation. On the research side, the analysis here points to specific next experiments: deriving a direction from the malignant-versus-biopsy-negative contrast and matching those arms on density, age, and prior count to power the comparison properly; stratifying by time-to-index beyond a binary prior-count split; replicating on a third encoder with independent training data; and, most importantly, replicating the drift finding at an external site.

NEXT

A vision-language successor that drafts a laterality-level findings section

NEXT

Localisation: side, view, quadrant, and depth, in weighted F_1

NEXT

Digital breast tomosynthesis, a priority modality gap

The broader methodological claim we advance is that embedding-space analysis belongs in the standard evaluation repertoire for imaging foundation models. Headline task metrics tell an integrator how well one head performs on one label. Analyses of the kind reported in Section 6.3 tell them what the representation has learned, where its signal is concentrated in time and space, and what it cannot distinguish; and it was such an analysis, not the classification metrics, that identified the specificity ceiling most likely to constrain a triage application built on this model.

THE METHODOLOGICAL CLAIM

Embedding-space analysis belongs in the standard evaluation repertoire for imaging foundation models.

ETHICS

Ethical considerations

All model development and validation used de-identified data processed through the pipeline described in Section 3.4. The model is provided as a developer tool and not as a medical device; it is not intended for diagnostic use, and any clinical or patient-facing application requires fine-tuning, threshold selection, population-specific validation, and independent regulatory authorisation by the integrator. Integrators are responsible for meeting the data-privacy and security obligations applicable to their own datasets.

Data and code availability

The pretraining corpus and the pathology-proven cancer dataset are proprietary. Training-data descriptions, per-finding performance metrics, and confusion matrices are available on request under agreement. Analysis code for the longitudinal embedding-drift study is maintained in an internal repository, with per-step scripts and cluster job specifications; the aggregation and statistics stages are model-agnostic and accept an embedding-dimension argument, which is how a single pipeline served both encoders reported here.

References

1. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 74(3):229-263, 2024.
2. Lehman CD, Arao RF, Sprague BL, et al. National performance benchmarks for modern screening digital mammography: update from the Breast Cancer Surveillance Consortium. *Radiology*, 283(1):49-58, 2017.
3. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. *New England Journal of Medicine*, 331(22):1493-1499, 1994.
4. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788):89-94, 2020.
5. Rodriguez-Ruiz A, Lang K, Gubern-Merida A, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *Journal of the National Cancer Institute*, 111(9):916-922, 2019.
6. Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models. arXiv:2108.07258, 2021.
7. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259-265, 2023.
8. Zhou Y, Chia MA, Wagner SK, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156-163, 2023.
9. Azizi S, Culp L, Freyberg J, et al. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nature Biomedical Engineering*, 7(6):756-779, 2023.
10. Chen RJ, Ding T, Lu MY, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850-862, 2024.
11. Ghosh S, Poynton CB, Visweswaran S, Batmanghelich K. Mammo-CLIP: a vision language foundation model to enhance data efficiency and robustness in mammography. In *MICCAI*, 2024.
12. Radiological Society of North America. RSNA Screening Mammography Breast Cancer Detection AI Challenge. Kaggle competition, 2023.

13. Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology*, 292(1):60-66, 2019.
14. Yala A, Mikhael PG, Strand F, et al. Toward robust mammography-based models for breast cancer risk. *Science Translational Medicine*, 13(578):eaba4373, 2021.
15. Pan I. MammoScreen: an open-source ensemble for breast cancer probability and density estimation from screening mammography. Software artefact, Apache-2.0 licence.
16. Caron M, Touvron H, Misra I, et al. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650-9660, 2021.
17. Oquab M, Darcet T, Moutakanni T, et al. DINOv2: learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.
18. Grill JB, Strub F, Altché F, et al. Bootstrap your own latent: a new approach to self-supervised learning. In *NeurIPS*, 2020.
19. Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv:1503.02531, 2015.
20. Codella NCF, Jin C, Munasinghe S, et al. MedImageInsight: an open-source embedding model for general domain medical imaging. arXiv:2410.06542, 2024.
21. Hu EJ, Shen Y, Wallis P, et al. LoRA: low-rank adaptation of large language models. In *ICLR*, 2022.
22. Mitchell M, Wu S, Zaldívar A, et al. Model cards for model reporting. In *FAT**, pages 220-229, 2019.
23. Gebru T, Morgenstern J, Vecchione B, et al. Datasheets for datasets. *Communications of the ACM*, 64(12):86-92, 2021.
24. American College of Radiology. *ACR BI-RADS Atlas: Breast Imaging Reporting and Data System*, 5th edition. Reston, VA, 2013.
25. Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65-70, 1979.
26. Peer PGM, van Dijck JAAM, Hendriks JHCL, Holland R, Verbeek ALM. Age-dependent growth rate of primary breast cancer. *Cancer*, 71(11):3547-3551, 1993.
27. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale. In *ICLR*, 2021.
28. US Food and Drug Administration. Clinical Decision Support Software: Guidance for Industry and Food and Drug Administration Staff. 2022.