

— CHEST RADIOGRAPH AI

A foundation model approach to chest radiograph classification

A single, reusable visual representation learned from
a very large and heterogeneous corpus of real-world
chest radiographs

AUTHOR

Neurologic AI

PUBLISHED

July 2025

DATASET

6.2 million chest X-ray studies

ABSTRACT

Chest radiography (CXR) is the single highest-volume diagnostic imaging examination performed worldwide, yet the great majority of computer-aided detection systems built for it are trained narrowly: one labeled dataset, one architecture, and one clinical finding at a time. This paper describes a foundation-

model approach to CXR image classification that instead learns a single, reusable visual representation from a very large and heterogeneous corpus of real-world chest radiographs, and then specializes that representation to a broad panel of clinically relevant findings through a staged training pipeline. We detail the composition and curation of a dataset of approximately 6.2 million chest X-ray studies (over 12.2 million images) drawn from both outpatient and inpatient care settings across multiple U.S. regions, and describe a four-stage training methodology comprising self-supervised pretraining, contrastive text-image alignment, multi-class pretraining, and per-finding binary fine-tuning. We further present a controlled, head-to-head benchmark in which the resulting model is compared against two other publicly documented medical vision-language backbones under identical training and evaluation conditions across 28 findings, using area under the receiver operating characteristic curve (AUROC) as the primary metric. The results indicate that a well-curated, CXR-specific pretraining strategy is competitive with, and in several findings superior to, general-purpose medical foundation models, while offering a single reusable backbone that amortizes representation-learning cost across an entire suite of downstream classifiers.

01

Introduction

Chest X-ray remains the most frequently ordered radiological examination in both inpatient and outpatient care, used to screen for and monitor conditions ranging from pneumothorax and pleural effusion to cardiomegaly and device placement. Despite this ubiquity, interpretation still depends almost entirely on manual visual review by radiologists, who must recognize dozens of overlapping and sometimes subtle patterns across a study, often under significant time pressure and rising case volumes. This creates a structural mismatch: the diagnostic bandwidth of a single reader is finite, but the number of findings that must be actively searched for on every study keeps growing as clinical practice incorporates more device monitoring, comorbidity tracking, and incidental-finding detection. Two consequences follow directly from this mismatch. First, reader fatigue and time pressure are well-documented contributors to missed or delayed findings, particularly for less common

conditions that a radiologist may encounter only occasionally, where pattern recognition is weaker simply due to lower exposure. Second, most existing computer-aided detection tools are built to compensate for this on a finding-by-finding basis: a team trains a dedicated classifier for pneumothorax, another for cardiomegaly, another for tube and line placement, each starting largely from scratch or from a generic (non-radiological) pretrained backbone such as ImageNet weights. This approach requires a large, separately labeled dataset for every new finding, duplicates representation-learning effort across projects, and scales poorly as the list of clinically desired findings grows.

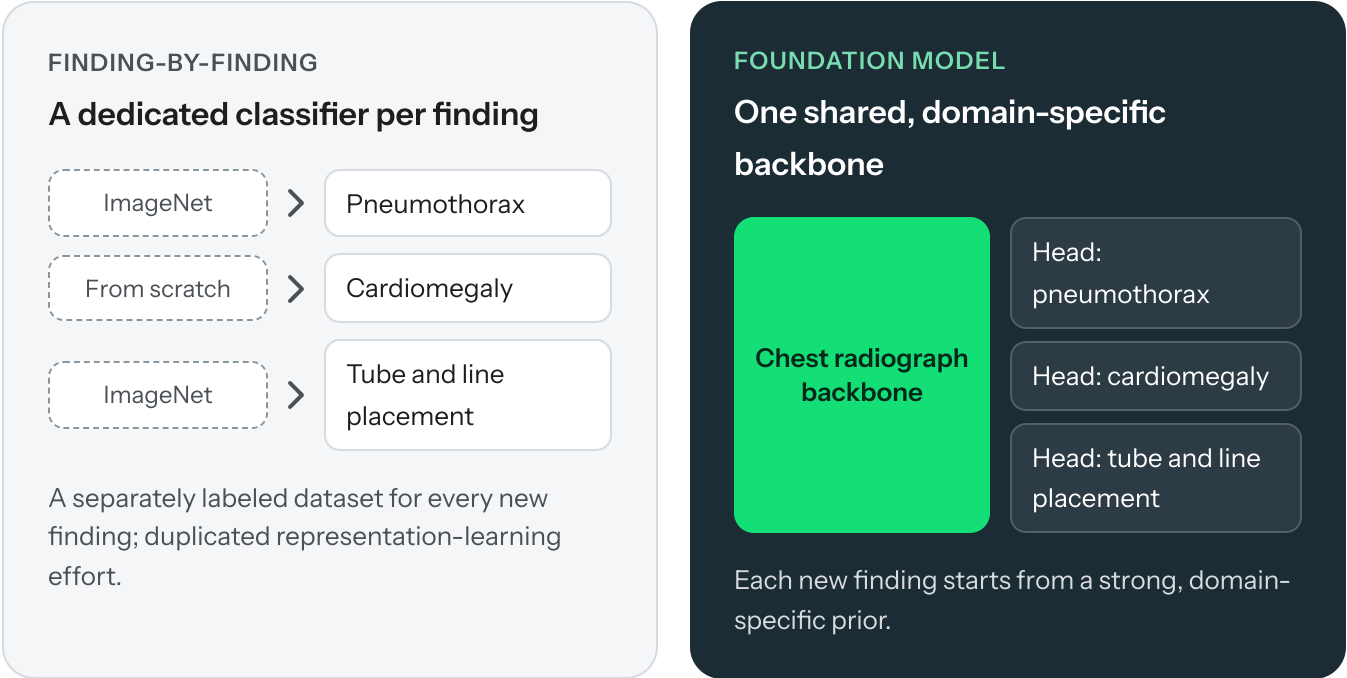


Figure 1. Training each classifier independently versus reusing a single learned representation underneath many task-specific classification heads.

Rather than training each classifier independently, we first learn a single, general-purpose visual representation of chest radiographs.

The work described in this paper takes a different starting point. Rather than training each classifier independently, we first learn a single, general-purpose visual representation of chest radiographs from a very large and diverse pool of real-world studies, using self-supervised and weakly-supervised objectives that do not require exhaustive manual labeling. This representation is then reused as a frozen or lightly-adapted backbone underneath many different task-specific classification heads, so that each new finding starts

from a strong, domain-specific prior rather than from a generic image-recognition prior or from nothing at all. Concretely, the pipeline proceeds through four progressively more specialized training stages: self-supervised pretraining on unlabeled images to learn general radiological structure; contrastive image-text alignment to ground the visual representation in clinical label semantics; multi-class pretraining to sharpen the representation against an explicit panel of findings; and finally per-finding binary fine-tuning, which specializes a shared backbone into a family of independent, high-accuracy classifiers. Because the early stages are self-supervised or weakly supervised, they can exploit a much larger pool of studies than any single labeled classification task would otherwise provide, and the resulting backbone is reused across the full downstream suite rather than re-learned for every finding. The remainder of this paper describes the dataset that underlies this pipeline, the training methodology in detail, and a comparative benchmark against other current medical imaging foundation models.

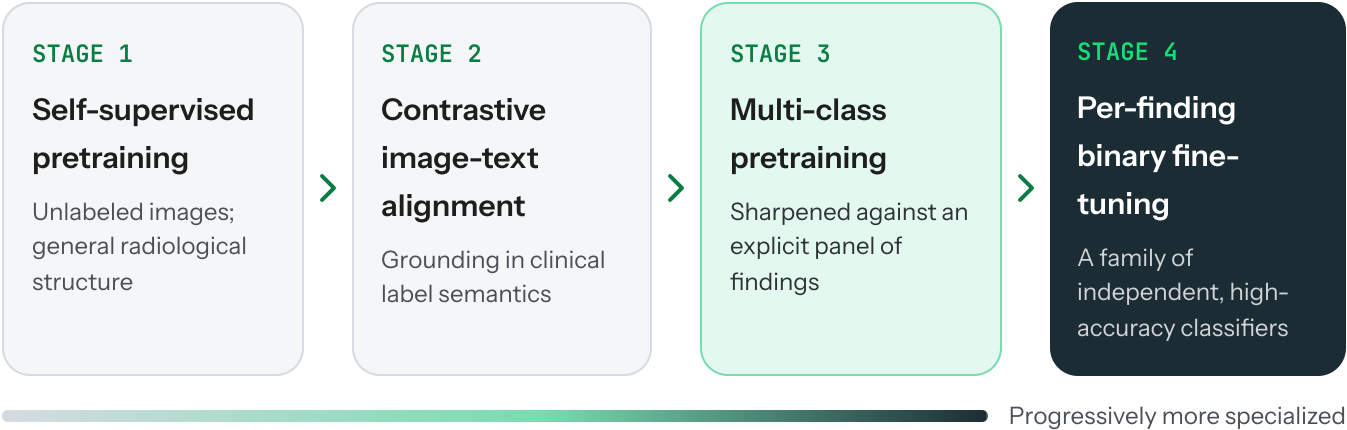


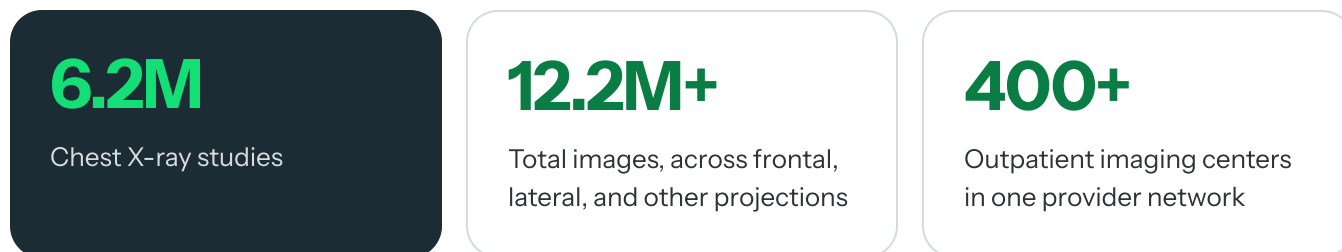
Figure 2. The four progressively more specialized training stages.

02

Data

2.1 Description and size

The dataset underlying this work consists of approximately 6.2 million chest X-ray studies, comprising over 12.2 million total images. Each study may include multiple images captured from different views, including frontal, lateral, and other projections.



2.2 Source

The dataset is a real-world dataset originating from two U.S.-based data providers. One provider owns and operates a network of over 400 outpatient imaging centers spanning the East, West, and Southeastern United States. The second provider is a data broker supplying inpatient data from the East, Central, and Southern United States. Data were collected over the period from October 2001 through March 2025.

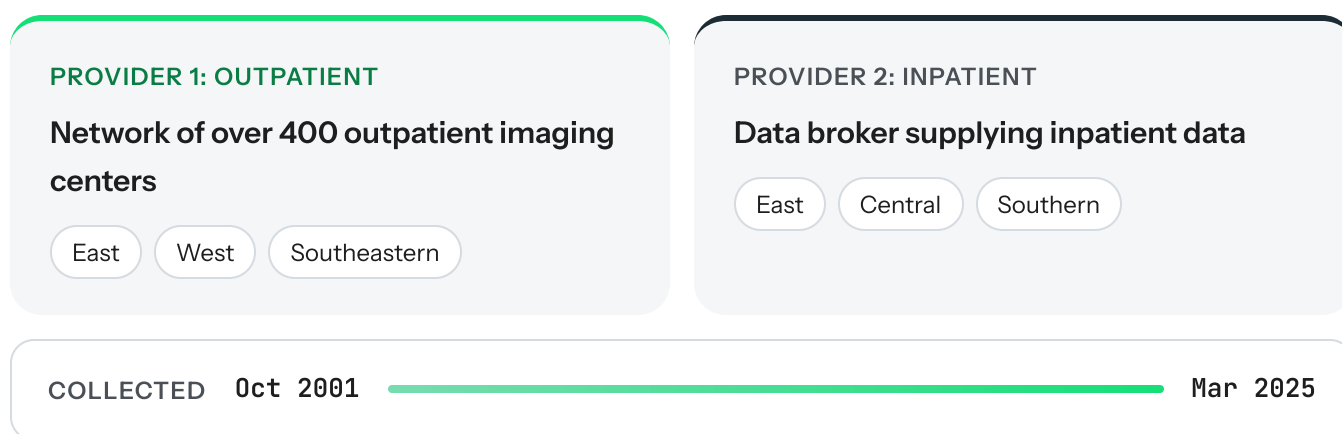


Figure 3. The two U.S.-based data providers and the collection period.

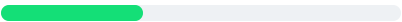
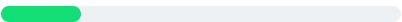
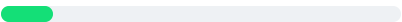
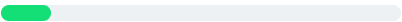
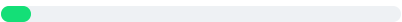
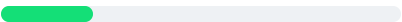
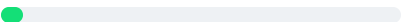
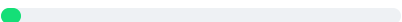
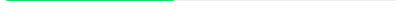
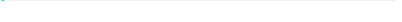
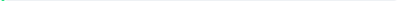
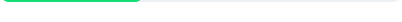
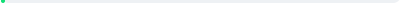
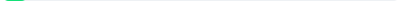
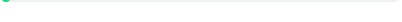
2.3 Acquisition method

The dataset is composed of proprietary, retrospective data obtained through direct data-use agreements with healthcare providers and data brokers. It includes imaging studies collected during the course of routine clinical care and acquired across a broad array of scanner hardware vendors, rather than from a single controlled acquisition protocol.

2.4 Disease conditions

The underlying representation-learning backbone is trained on a broad range of radiological chest findings and features so that it generalizes across many downstream classification tasks. The classification panel built on top of this backbone targets 24 key findings

prioritized for their clinical relevance; studies containing additional findings beyond this panel were grouped under an “Other” category during pretraining. Table 1 summarizes the training data distribution across these findings.

Finding		Data distribution (studies)
Pleural effusion		280,873
Normal		160,942
Nonsurgical internal foreign body		102,561
Hyperinflation		99,466
Hiatus hernia		59,899
Pulmonary congestion / venous congestion		185,166
Interstitial thickening		44,017
Peribronchial cuffing		39,158
Pacemaker / electronic cardiac device or wires		34,039
Lobar / segmental collapse		62,568
Cardiomegaly		728,591
Air space opacity		793,172
Bronchial wall thickening		10,199
Intercostal drain		37,617
Atelectasis		354,913
Pneumothorax		27,324
Tracheal deviation		3,515
Subcutaneous emphysema		2,126
Hilar lymphadenopathy		1,872
Lung nodule or mass		281,940
Shoulder dislocation		838
Pulmonary artery enlargement		58,004
Bullous disease		20,987
Implantable electronic device		522

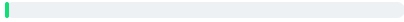
Finding		Data distribution (studies)
Whole lung / majority collapse		3,340

Table 1. Training data distribution by finding.

2.5 Patient demographics

Table 2 summarizes the available patient demographic distribution across the dataset.

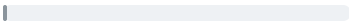
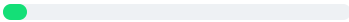
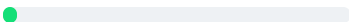
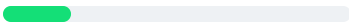
Demographic	Category		Data distribution (studies)
Sex	Male		2,885,842
	Female		3,573,775
	Other / Unknown		28,892
Age	0–18 years		149,930
	18–30 years		80,810
	30–50 years		406,697
	50–70 years		1,659,752
	70–100 years		2,082,040
	Other / Unknown		2,109,150

Table 2. Patient demographic data distribution.

Structured data for race, ethnicity, height, and weight were not available for most studies and are therefore not included in this analysis. While these characteristics may influence disease prevalence or image acquisition in certain contexts, their impact on model performance could not be formally assessed given the missing metadata. The dataset does, however, comprise more than 6.2 million unique chest X-ray studies drawn from both inpatient and outpatient settings across diverse U.S. regions, spanning all adult age groups and sexes, which supports substantial demographic and clinical variability even though structured subgroup auditing on race, ethnicity, height, and weight is not currently possible.

2.6 Acquisition sites and equipment

Studies were labeled with scanner manufacturer information derived from DICOM metadata, spanning ten major hardware vendors; unknown, mislabeled, or less common scanners were grouped into an “Other” category, which represents 51% of all data (96% of that subset having unknown manufacturer information). Table 3 summarizes acquisition equipment distribution, and Table 4 summarizes the two contributing acquisition sites (outpatient and inpatient), whose specific identities have been withheld pursuant to confidentiality obligations.

Scanner manufacturer		Data distribution (studies)
Carestream Health		714,417
Fujifilm		694,681
Medlink Imaging		652,000
GE Healthcare		340,941
Konica Minolta Healthcare		375,486
Kodak		270,391
Canon Medical Systems		273,151
Siemens Healthineers		265,905
Samsung Medison		171,741
Agfa		78,832
Other / Unknown		2,362,168

Table 3. Data acquisition equipment distribution.

Acquisition site		Data distribution (studies)
Data Site 1 (outpatient)		5,608,150
Data Site 2 (inpatient)		591,926

Table 4. Data acquisition sites.

2.7 Preprocessing

The following preprocessing steps are applied to all images prior to training:

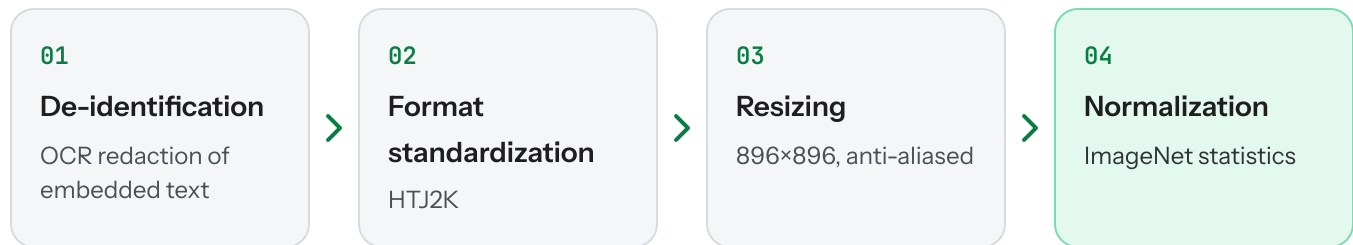
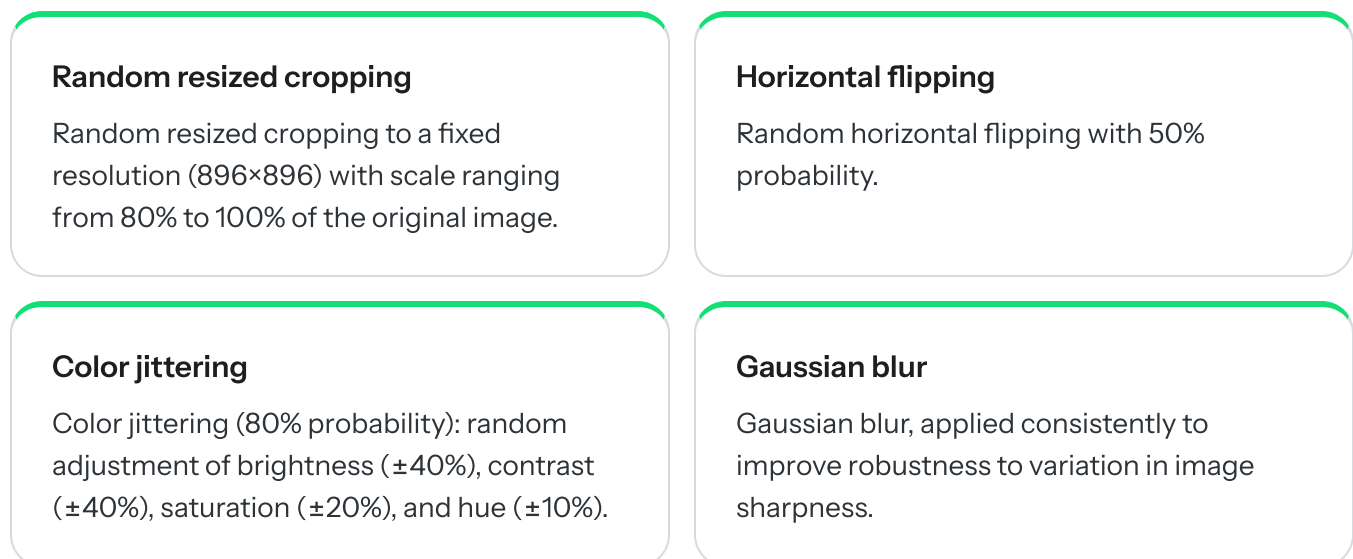


Figure 4. Preprocessing steps applied to all images prior to training.

- **De-identification:** images are de-identified using customized third-party OCR solutions to identify and redact embedded text that has not been explicitly whitelisted.
- **Format standardization:** image data is standardized to High-Throughput JPEG 2000 (HTJ2K) format.
- **Resizing:** each image is resized to a fixed resolution (default 896×896 pixels) using anti-aliasing to preserve image quality.
- **Normalization:** pixel values are normalized using standard ImageNet statistics (mean [0.485, 0.456, 0.406]; standard deviation [0.229, 0.224, 0.225]) for consistency across the dataset.

During training, additional augmentations are applied to improve model robustness and generalization:



2.8 Dataset splitting

All training data were split into training, validation, and test sets using an 80/10/10 ratio, with partitioning performed at the patient level so that all studies and images belonging to a given patient remain within a single subset. This prevents data leakage across splits and avoids scenarios where a single patient’s data could appear in more than one partition, which would compromise the independence of model evaluation.

During training, both the training and validation datasets were curated to maintain a balanced distribution of positive and negative cases per finding, allowing the model to learn effectively across the full spectrum of conditions. The test dataset, by contrast, was constructed to reflect the real-world prevalence of findings, providing a more clinically realistic assessment of generalization. Table 5 summarizes the total number of studies assigned to each partition for the final fine-tuning stage.

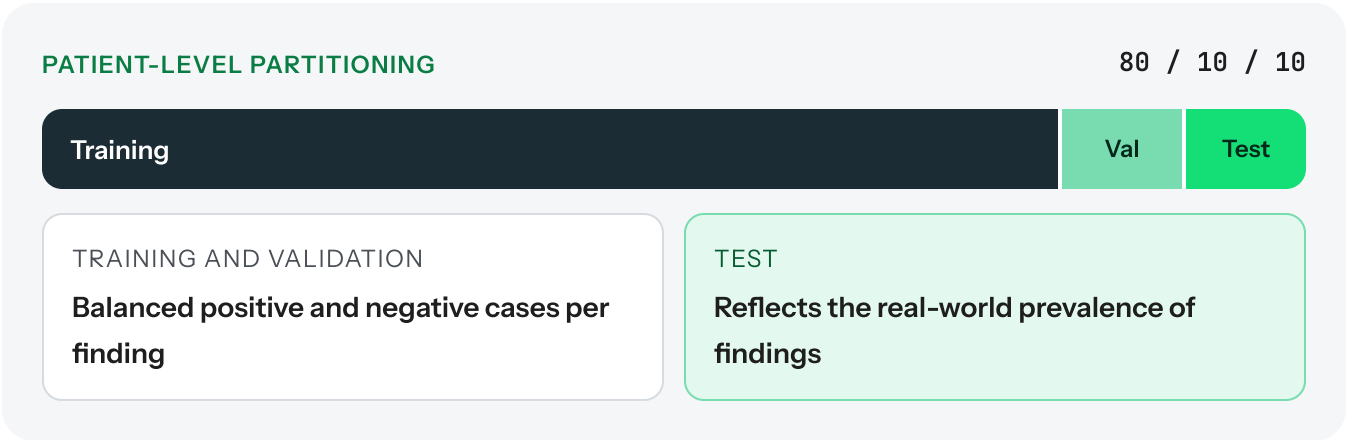


Figure 5. The 80/10/10 patient-level split, with balanced training and validation sets and a prevalence-matched test set.

Partition	Total studies
Training	5,001,283
Validation (hyperparameter tuning)	626,859
Test (final model evaluation)	627,046

Table 5. Training / validation / test split for the final fine-tuning stage.

Classification tasks in this pipeline are currently restricted to frontal chest X-ray images; lateral and other views are excluded. Each study may correspond to multiple frontal images, and for training and validation, only a single representative frontal image per study is used to

ensure consistency. During testing, all frontal images linked to a study are used, and classification outputs are aggregated by taking the maximum predicted probability across all images for that study, so that the final prediction reflects the most confident finding-level signal available at evaluation time.



Figure 6. Test-time aggregation: the study-level prediction is the maximum predicted probability across all frontal images linked to the study.

03

Model training

The classification system was trained in a staged fashion designed to balance generalization, representation learning, and task-specific performance. Each stage progressively enhances the model’s ability to extract clinically meaningful features from chest radiographs, culminating in a suite of per-finding binary classifiers built on a single shared visual backbone.

3.1 Stage 1: Self-supervised learning (SSL)

The first stage pretrains a Vision Transformer (ViT) backbone on unlabeled chest X-ray images using a self-supervised learning method based on the DINO framework, so that the model learns general-purpose, radiologically meaningful representations prior to any supervised fine-tuning.

Stage 1 specification		DINO ViT, ~0.25B parameters
Architecture	A DINO Vision Transformer with approximately 0.25B parameters, using a shared student/teacher configuration in which the teacher is updated via an exponential moving average of the student's weights, with momentum increased over training from 0.99 to 0.9999.	
Projection head	A multi-layer perceptron (MLP) image-level head produces prototype scores from the [CLS] token of the ViT.	
Training approach	Pretraining proceeds over three phases with an increasing number of studies and a curated balance of findings, using image resizing to 448×448, two global crops (scale 0.4 to 1.0) to encourage coarse-grained holistic learning, and eight local crops (scale 0.05 to 0.4) to encourage fine-grained local feature learning.	
Objective	The DINO loss computes the cross-entropy between the softmax-normalized outputs of the teacher and student networks for different augmented views of the same image.	

3.2 Stage 2: Contrastive text pretraining

The second stage aligns chest X-ray image representations with weakly labeled, text-derived findings using a contrastive, CLIP-like training strategy at the study level, grounding the visual representation in clinical label semantics across a large vocabulary of findings.

Stage 2 specification		LoRA $r = 256$, $\alpha = 16$
Architecture	A LoRA-adapted Vision Transformer with study-level embedding aggregation, using parameter-efficient fine-tuning (rank $r = 256$, $\alpha = 16$), with image and study embedding dimensionality reduced from 2048 to 512.	
Augmentation	Each image is divided into five tiles, passed through the encoder, and aggregated into a single study-level representation.	
Training approach	An initial phase trains on 1M studies across 50 findings, with oversampling of minority conditions such as scoliosis, followed by a second phase on 1M studies across 33 findings.	

3.3 Stage 3: Multi-class pretraining (MCPT)

The third stage incorporates explicit multi-label classification supervision by adding a classification head and continuing to fine-tune the LoRA-adapted backbone, aligning the embedding space with clinically relevant class labels and enhancing generalization across findings.

Stage 3 specification		15 classes, then 33 classes
Architecture	The ViT encoder pretrained in the prior stages, together with its LoRA adapters, feeds into an MLP classification head that produces logits for each visual finding.	
Training approach	The model is first trained to convergence on a 15-class configuration (10,000 studies per finding) before expanding to a 33-class configuration (800 studies per finding).	

3.4 Stage 4: Binary classification fine-tuning

The final stage fine-tunes the multi-class pretrained model for per-finding binary classification, treating each finding as an independent positive-versus-negative task while retaining the shared representation learned in the earlier stages.

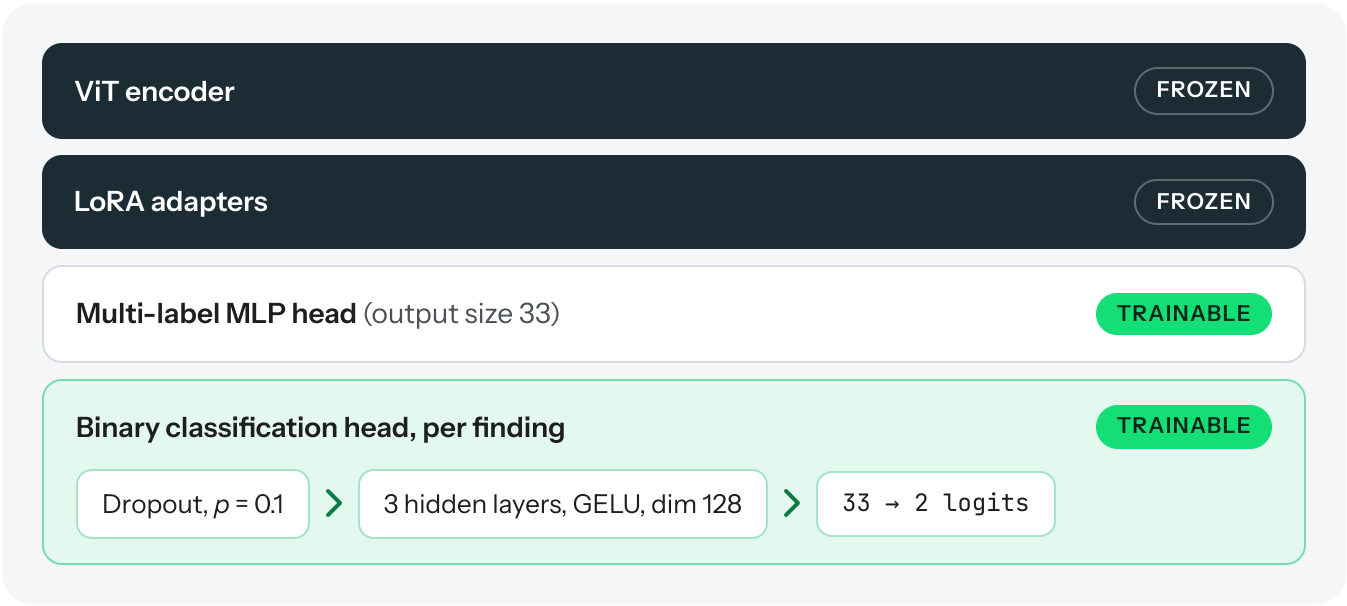


Figure 7. Stage 4 model: the ViT backbone and LoRA adapters remain frozen; only the MLP classifier layers are trained.

- **Base model:** a frozen ViT encoder and frozen LoRA adapters from the earlier stages, together with a trainable multi-label MLP head (output size 33).
- **Classification head:** a final binary classification layer added on top of the pretrained MLP head for each finding, structured as a dropout layer ($p = 0.1$), three hidden layers with GELU activations (intermediate dimension 128), and a final two-layer mapping from 33 to 2 binary logits per finding.
- **Training scope:** only the MLP classifier layers are trained in this stage; the ViT backbone and LoRA adapters remain frozen, so that each binary classifier specializes without disturbing the shared learned features.
- **Approach:** each binary classifier is trained independently for its corresponding finding, taking the same preprocessed embeddings as input and producing class logits interpreted as positive versus negative probabilities.

3.5 Backbone reuse and deployment

The self-supervised backbone produced by this pipeline is released as a versioned artifact and reused across the full family of downstream classification and segmentation tasks, rather than retrained for each one. In deployment, backbone weights are loaded from object storage and a modular, configurable MLP head is attached dynamically for each task, supporting both binary and multi-class outputs from the same frozen backbone, with an optional low-rank adaptation (LoRA) path available for domain-shift adaptation. The resulting system is served behind a FastAPI application with autoscaling infrastructure, allowing the same backbone to support many concurrent finding-specific classifiers in production. This design amortizes the cost of representation learning across the entire downstream product suite: every new classifier or segmenter starts from a strong, chest-radiograph-specific prior rather than from a generic image-recognition prior.

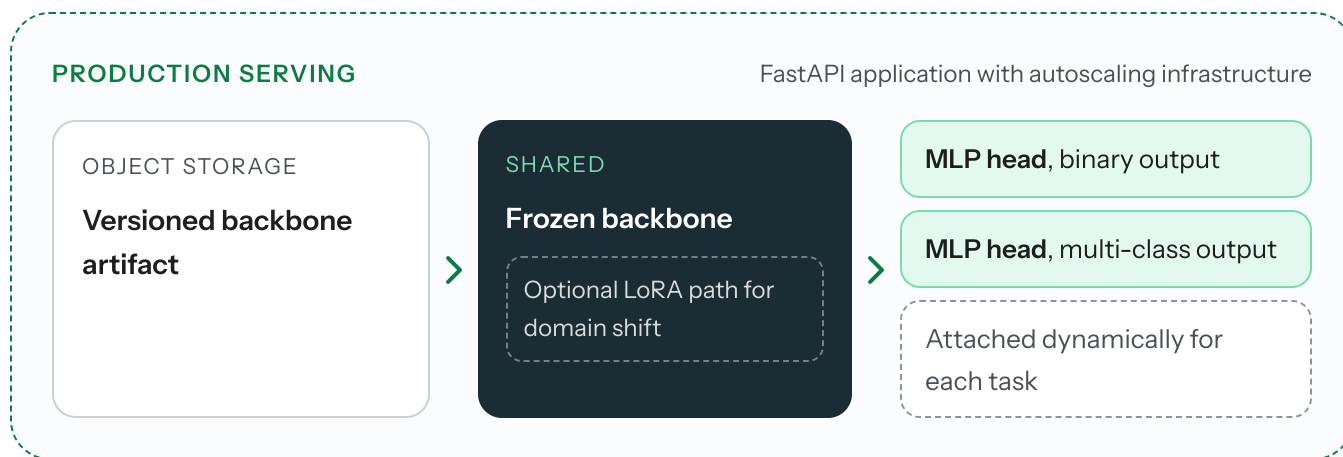


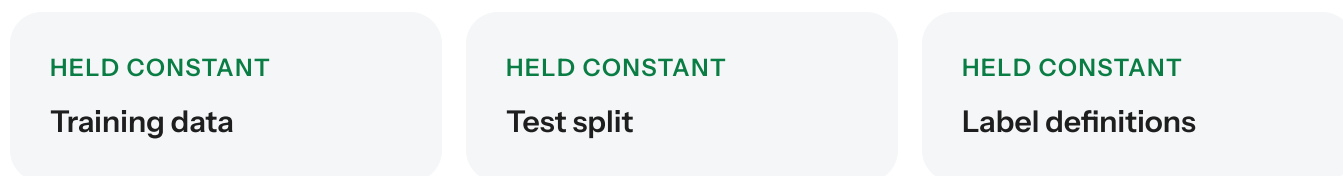
Figure 8. Backbone reuse in deployment: one frozen backbone serves many concurrent finding-specific classifiers.

04

Metrics and results

4.1 Benchmark objective and design

To evaluate how the resulting representation compares against other current medical-imaging foundation models, a controlled head-to-head benchmark was conducted. All three encoders under comparison were fine-tuned from their respective pretrained weights on the same internal chest X-ray dataset and evaluated on the same held-out test set, using AUROC per finding across 28 clinically relevant findings as the primary metric. Unlike many published benchmarks, which compare models evaluated under heterogeneous protocols, training data, and label sets, this comparison holds training data, test split, and label definitions constant across all three models, making the resulting differences directly interpretable rather than an artifact of differing evaluation conditions.



4.2 Results

Table 6 reports AUROC per finding for the CXR-specific foundation model described in this paper alongside two other medical imaging foundation models, MedImageInsight and MedGemma, under the identical evaluation protocol described above.

Finding	CXR Foundation (Ours)	MedImageInsight	MedGemma
Normal	0.88	0.85	0.84
NG tube	0.95	0.95	0.95
Lobar collapse	0.72	0.68	0.64
Interstitial thickening	0.60	0.63	0.58
Internal foreign body	0.80	0.79	0.76
Hyperinflation	0.78	0.76	0.74
Hilar LAD	0.74	0.74	0.69
Hiatal hernia	0.94	0.95	0.90
Gastric band	0.97	0.97	0.90
ET tube	0.99	0.99	0.99
Clavicle fracture	0.82	0.84	0.80
Chest tube	0.96	0.97	0.96
Cardiomegaly	0.88	0.89	0.87
Bullous disease	0.80	0.81	0.78
Bronchial wall thickening	0.72	0.69	0.76
Atelectasis	0.73	0.76	0.69
Airspace opacity	0.80	0.79	0.75
Vertebral fracture	0.72	0.74	0.68
Tracheal deviation	0.74	0.78	0.60
Subcutaneous emphysema	0.98	0.99	0.98
Rib fracture	0.90	0.89	0.74
Pulmonary venous congestion	0.90	0.89	0.86
Pneumothorax	0.90	0.90	0.80
Pleural thickening	0.84	0.85	0.73
Pleural effusion	0.90	0.90	0.89
Peribronchial cuffing	0.77	0.75	0.69
Pacemaker or leads	0.95	0.95	0.95

Finding	CXR Foundation (Ours)	MedImageInsight	MedGemma
PA catheter	0.90	0.91	0.89

Table 6. AUROC per finding across 28 clinically relevant findings, evaluated under identical training data, test split, and label set. Highest value per finding shown in bold on a tinted cell.

4.3 Discussion

Across the 28 evaluated findings, the CXR-specific foundation model performs competitively with, and in a number of findings exceeds, both comparator models, particularly for findings such as lobar collapse, hyperinflation, rib fracture, pneumothorax, pleural thickening, and tracheal deviation. Performance is closely matched across all three models on high-prevalence or visually distinctive findings such as ET tube, NG tube, pacemaker or leads, and subcutaneous emphysema, where all three encoders reach AUROC values at or above 0.95. The largest performance gaps favoring the CXR-specific backbone tend to occur on findings that are more subtle or less visually stereotyped, such as bronchial wall thickening, tracheal deviation, and rib fracture, suggesting that pretraining on a very large, diverse, chest-radiograph-native corpus provides a meaningful advantage for exactly the findings that are hardest to learn from limited labeled data alone.

Taken together, these results support the central premise of this work: a single, well-pretrained, chest-radiograph-specific visual backbone can match or exceed the performance of broader medical foundation models on this modality, while offering the practical advantage of a shared representation that can be reused across an entire suite of downstream classifiers rather than re-learned for each one.

THE CENTRAL PREMISE

A single, well-pretrained, chest-radiograph-specific visual backbone can match or exceed the performance of broader medical foundation models on this modality.

