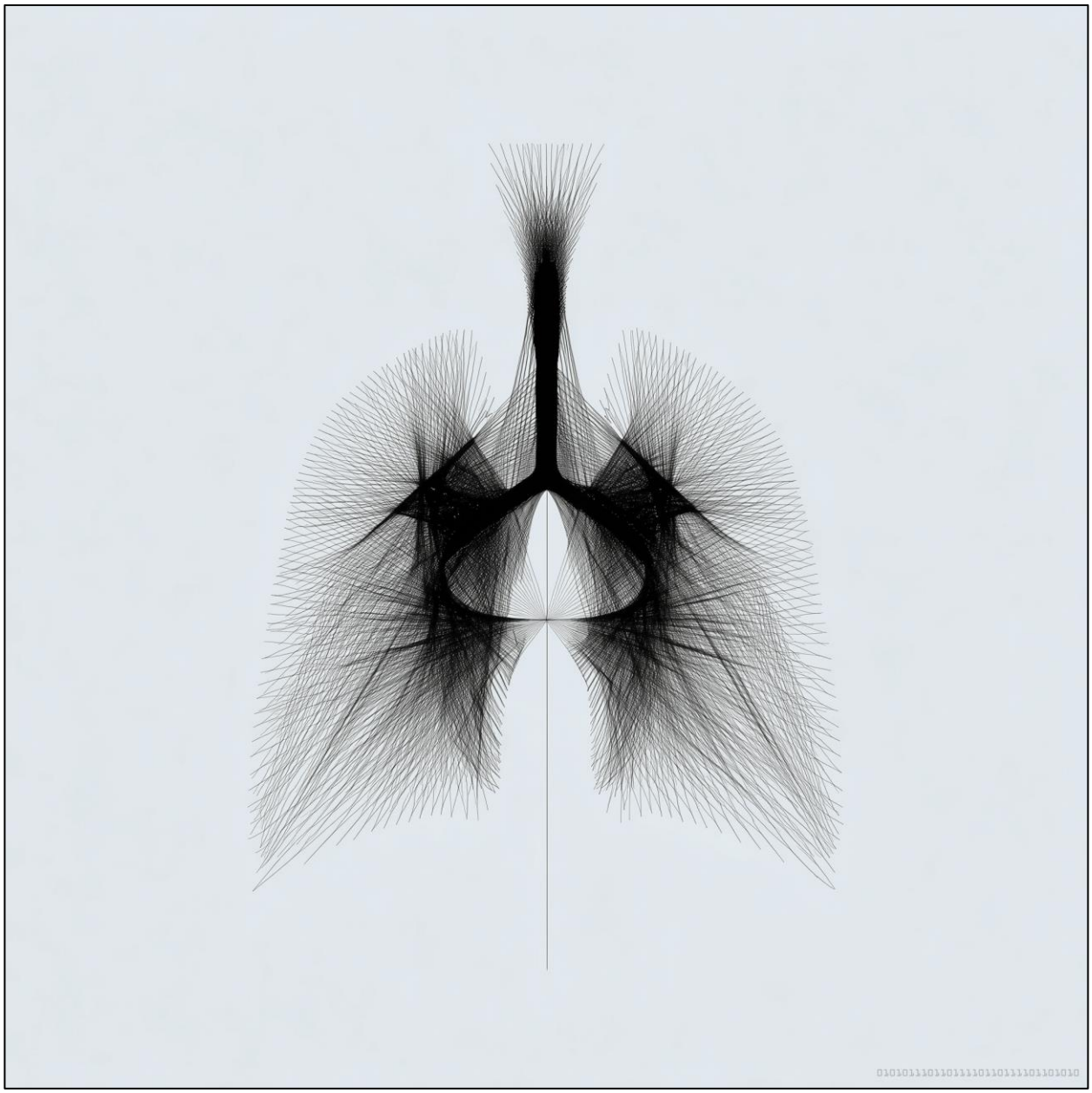
HOPPR

WHITEPAPER

Fine-Tuning Foundation Models for Medical Imaging

Akash Ranjan Pattnaik, PhD, Julie Bauml, MD, Melanie Traughber, DSc, Kalina Polet Slavkova, PhD, Ali Ganjizadeh, MD, Dave Van Veen, PhD, Dan Harms, Robert Bakos, Bradley Erickson, MD, PhD, Khan Siddiqui, MD



Cover art by Woojin Kim, MD, “The Art of Fine-Tuning” - created using Midjourney - 2025

Table of Contents

Abstract

Building high-performing medical imaging AI models typically requires labeled datasets, significant computing infrastructure, and specialized AI expertise—resources that are out of reach for many clinical teams.

The HOPPR™ AI Development Platform circumvents these barriers by enabling healthcare teams to fine-tune powerful pre-trained foundation models on their own data. The platform offers secure data upload, configurable training, real-time experiment monitoring, and API-based tools to enable seamless workflow integration.

Using this platform and the open-source VinDr-CXR dataset, we fine-tuned the HOPPR™ MC Chest Radiography Foundation Model to detect 17 findings on chest X-rays. We observe strong classification performance from fine-tuned models with as few as 400 labeled studies. Our analysis reveals linear scaling of computational requirements, with training duration ranging from 26 seconds per epoch for 100 samples to 12 minutes per epoch for larger datasets, enabling healthcare teams to balance between performance requirements and training time.

Introduction

Over the past decade, artificial intelligence (AI) models in medical imaging have enabled new approaches towards improving diagnostic accuracy and supporting clinical decision-making. AI models can help reduce radiology workload by automating repetitive and time-consuming tasks such as case triage and generating reports, resulting in enhanced efficiency, consistency, and throughput across care teams[1,2].

Traditionally, each AI model is trained end-to-end, which requires large amounts of annotated data, training time, and computational resources. The advent of foundation models (FMs), however, has introduced a major shift from this traditional model development strategy. FMs are large-scale, task-

[1] Yoon, S. H. *et al*. Use of artificial intelligence in triaging of chest radiographs to reduce radiologists' workload. *Eur. Radiol.* **34**, 1094–1103 (2024).
[2] Sloan, P., Clatworthy, P., Simpson, E. & Mirmehdi, M. Automated Radiology Report Generation: A Review of Recent Advances. *IEEE Rev. Biomed. Eng.* **18**, 368–387 (2025).

agnostic models that are pre-trained on vast and diverse datasets[3] and can be trained using self-supervision, which does not require labeled training data.

Once trained, FMs can be fine-tuned for specific downstream tasks such as classification, segmentation, and feature extraction using relatively small amounts of site-specific or task-specific labeled data. This capability is well-suited for medical applications where labeled data is often limited.

Foundation models represent a new paradigm where AI models are trained on broad datasets at scale, developing emergent capabilities that transfer across diverse tasks. This approach enables powerful leverage through model reuse, allowing a single pre-trained model to be adapted for multiple downstream applications[4].

In medical imaging, this means healthcare teams can fine-tune models that already understand anatomical structures and pathological patterns, rather than training from scratch. The development of foundation models is challenging as they demand specialized AI expertise, significant computational resources, and high-quality labeled data.

Also, in medical imaging, there is often a gap between the AI developers building FMs and the radiologists who ultimately use them. AI developers have access to technical skills and compute infrastructure to build models, and radiologists can define relevant use cases for AI to improve their workflow, have access to large amounts of data, and hold the technical expertise to annotate data.

To help bridge these gaps, we introduce the HOPPR™ AI Development Platform. This platform enables users to fine-tune and validate FMs on their own data and desired use cases without requiring a high degree of technical expertise, computer resources, and large, high-quality data repositories.

In this paper, we describe the Platform's fine-tuning tool and provide examples of models that were fine-tuned and validated using the HOPPR™ MC Chest Radiography Foundation Model.

---

[3] Zhou, C. *et al.* A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT. Preprint at https://doi.org/10.48550/arXiv.2302.09419 (2023).

[4] Bommasani, R. *et al.* On the Opportunities and Risks of Foundation Models. Preprint at https://doi.org/10.48550/arXiv.2108.07258 (2022).

Background

HOPPR™ Foundation Models

The HOPPR™ Foundation Models have learned general representations of anatomical structures and pathological patterns through pre-training on millions of medical imaging studies, but they lack task-specific specialization. During pre-training, the models use unsupervised learning objectives to identify patterns and structures in medical images without requiring labeled data. This process creates a high-dimensional embedding space that captures diverse anatomical and pathological features. The HOPPR™ MC Chest Radiography Foundation Model, used in the experiments described in this paper, was pre-trained on over 2.5 million chest X-ray exams.

The Fine-Tuning Process

Fine-tuning adapts pre-trained representations for specific clinical tasks using curated datasets that are typically 100-1000 times smaller than pre-training datasets. During fine-tuning, a subset of the foundation model's parameters is updated, while most weights remain fixed to preserve learned representations. A task-specific classification head is added to receive the foundation model's embeddings, and both this new component and selected foundation model layers are trained on labeled examples to produce task-specific outputs such as binary classification scores.

Data Requirements and Considerations

Fine-tuning performance depends heavily on data quantity and quality. While diverse demographics and scanner types improve model generalizability, models trained on homogeneous data can perform well for specific patient populations[5]. For binary classification tasks, data must be labeled as positive (finding present) or negative (finding absent) and split into training, validation, and test sets. To prevent data leakage that could artificially inflate performance, patient data is stratified so that all images from a single patient appear in only one dataset split. The training set updates model weights, the validation set guides training decisions and determines optimal stopping points, and the test set provides unbiased performance evaluation.

[5] Ricci Lara, M. A., Echeveste, R. & Ferrante, E. Addressing fairness in artificial intelligence for medical imaging. *Nat. Commun.* **13**, 4581 (2022).

Existing tools for fine-tuning

The growing availability of foundation models has led to the development of various fine-tuning tools spanning code-based and interface-based options. Code-based platforms offer customizable fine-tuning frameworks built with libraries like PyTorch that integrate into developer environments[6], while no-code interfaces allow users to define tasks and datasets without programming. For proprietary models, API-based approaches enable fine-tuning without exposing sensitive model components, providing healthcare teams with expert-optimized parameters while maintaining security. However, medical imaging applications require specialized considerations including regulatory compliance, data security, and clinical validation that general-purpose platforms may not address comprehensively.

Fine-tuning with the HOPPR™ AI Development Platform

The fine-tuning tool allows users to develop task-specific solutions on top of the HOPPR™ suite of foundation models. The HOPPR™ Foundation Models are pre-trained on millions of medical images and developed within a secure, HIPAA-compliant AI platform supported by a Quality Management System[7]. The tool allows users to define a task and stratify their data into training, validation, and test sets for fine-tuning and to evaluate the final fine-tuned models.

Data uploading

Training and validation data are uploaded to a secure Amazon Web Services S3 storage bucket, where they are immediately de-identified by the HOPPR™ platform. Data is filtered for non-radiograph tags and artifactual scans that are abnormally small. The user defines a training split and sets a split seed to determine how uploaded data is assigned to training and validation splits. The training and validation splits are maintained across all epochs, and varying the training split seed can provide a user with confidence intervals on training and validation metrics across fine-tuning jobs, akin to cross-validation.

---

[6] Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. Preprint at https://doi.org/10.48550/arXiv.1912.01703 (2019).
[7] Slavkova, K. P. *et al.* HOPPR Medical-Grade Platform for Medical Imaging AI. Preprint at https://doi.org/10.48550/arXiv.2411.17891 (2024).

Fine-tuning

Once the uploaded data has been de-identified and assigned to training and validation splits, the user defines the name of the resultant model and hyperparameters of interest, including the number of epochs before the fine-tuning job should conclude. The model's trainable layers are updated during each epoch using information from both positive and negative cases.

Real-time monitoring and validation

During the fine-tuning process, the user can monitor training and validation loss and metrics for each epoch using the MLFlow dashboard *via* the URL provided when a user checks the status of the fine-tuning job. Receiving real-time feedback allows users to adjust their input data and hyperparameters to achieve satisfactory metrics of interest or terminate training jobs early if they do not observe ideal model performance. Once a foundation model has been fine-tuned to the desired performance, the fine-tuned model is promoted to the HOPPR™ platform for inference.

Inference and API integration

The newly fine-tuned model can be used to make predictions on any unseen data, including the held-out test for evaluation. API-based inference allows a user to incorporate inference with the fine-tuned model into an existing workflow. For example, an AI developer who has built a web app can use the REST API along with JavaScript to create a widget of HOPPR™ model inference results, or a data scientist can use the HOPPR™ Software Development Kit (SDK)[8] to perform experiments and compare HOPPR™ model performance with other baselines.

Platform Performance and Resource Requirements

A critical consideration for fine-tuning FMs is determining the minimum labeled dataset size required to achieve clinically meaningful performance. While traditional deep learning models often require thousands to millions of labeled examples, foundation models leverage pre-trained representations to achieve strong performance with significantly reduced data requirements. To provide evidence-

[8] hopprai: SDK for interacting with the HOPPR API.

based guidance for data collection strategies, we systematically evaluated fine-tuning performance across multiple dataset sizes using the HOPPR™ AI Development Platform.

Experimental Design

We evaluated fine-tuning performance across dataset sizes using 17 chest radiography findings from the VinDr-CXR dataset, with training sets ranging from 100 to 4,000 balanced samples per finding[9,10,11]. Due to varying data availability, 17 findings had sufficient positive cases for 100-sample models, while 10 findings had enough data for 400 sample models. Each dataset followed a nested design where larger sets contained all samples from smaller sets.

For each finding-sample size combination, we fine-tuned the HOPPR™ MC Chest Radiography Foundation Model for 20 epochs using default hyperparameters, selecting weights from the epoch with highest validation AUROC. Performance was evaluated on consistent test sets of 477-514 studies per finding, maintaining the original VinDr-CXR data distribution.

Data Volume Requirements

The improvement in test-set classification performance from increasing the amount of fine-tuning data varied for each finding *(Figure 1 below)*.

[9] Goldberger, A. L. *et al*. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* **101**, E215-220 (2000).

[10] Nguyen, H. Q., Pham, H. H., Tuan Linh, Le, Dao, M. & Khanh, Lam. VinDr-CXR: An open dataset of chest X-rays with radiologist annotations. PhysioNet https://doi.org/10.13026/3AKN-B287.

[11] Nguyen, H. Q. *et al*. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. Preprint at https://doi.org/10.48550/arXiv.2012.15029 (2022).

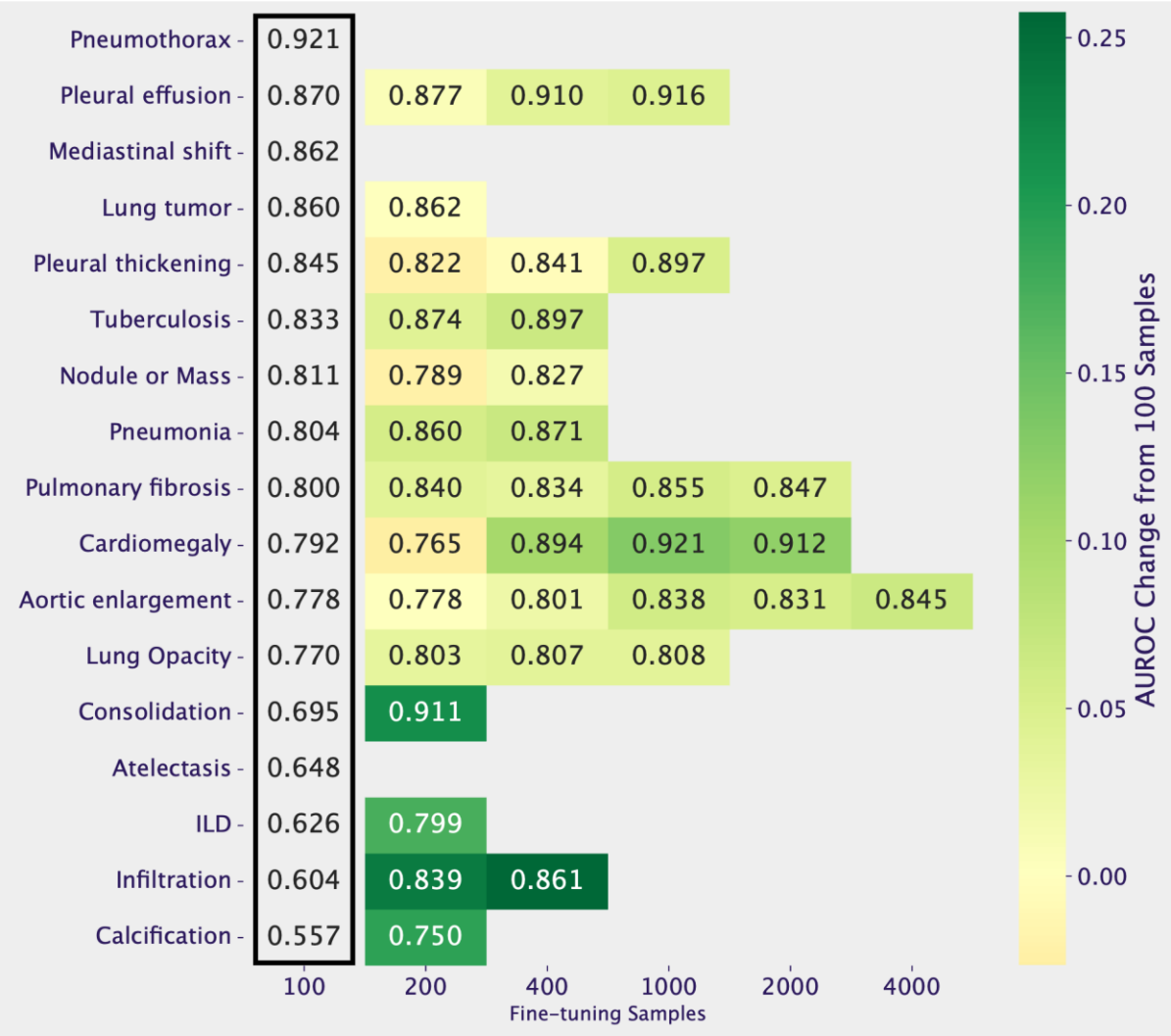


Figure 1. ***AUROC performance across dataset sizes for 17 chest radiography findings.*** *Each cell displays AUROC values for fine-tuned models trained on the indicated number of samples. Color intensity represents the magnitude of performance improvement from the 100-sample baseline. Findings are ordered by baseline performance (100 samples) from highest to lowest. Test sets are consistent within all models for a given finding.*

Mean AUROC increased from 0.769 for 100-sample datasets to 0.863 for 2000-sample datasets. Infiltration and Consolidation demonstrated the largest improvements in AUROC from 100 to 200 samples (0.235 and 0.215, respectively).

Conversely, findings such as Pneumothorax and Pleural effusion achieved strong baseline performance relative to other findings (AUROC > 0.87) even with minimal training data. Detailed results for all models including the AUROC, area under the precision-recall curve (AUPRC), sensitivity, specificity, and F1 score are provided *(Table 1 below)*.

Finding	Fine-tuning Studies	Inference Studies	AUROC	AUPRC	Sensitivity	Specificity	F1 Score
Aortic enlargement	100	496	0.78	0.21	0.87	0.60	0.26
	200	496	0.78	0.22	0.84	0.64	0.27
	400	496	0.80	0.28	0.95	0.60	0.28
	1000	496	0.84	0.29	0.87	0.72	0.33
	2000	496	0.83	0.27	0.87	0.74	0.35
	4000	496	0.84	0.32	0.89	0.74	0.35
Atelectasis	100	514	0.65	0.05	0.64	0.72	0.11
Calcification	100	511	0.56	0.18	0.28	0.93	0.26
	200	511	0.75	0.19	0.89	0.56	0.23
Cardiomegaly	100	477	0.79	0.27	0.87	0.68	0.36
	200	477	0.77	0.22	0.91	0.54	0.30
	400	477	0.89	0.55	0.98	0.67	0.39
	1000	477	0.92	0.61	0.98	0.74	0.45
	2000	477	0.91	0.58	0.91	0.81	0.50
Consolidation	100	510	0.70	0.05	0.93	0.52	0.11
	200	510	0.91	0.17	1.00	0.82	0.26
ILD	100	509	0.63	0.16	0.80	0.47	0.22
	200	509	0.80	0.32	0.91	0.59	0.29
Infiltration	100	503	0.60	0.03	0.89	0.42	0.05

	200	503	0.84	0.08	0.89	0.72	0.10
	400	503	0.86	0.09	1.00	0.64	0.09
Lung Opacity	100	508	0.77	0.09	0.94	0.58	0.13
	200	508	0.80	0.11	0.88	0.69	0.15
	400	508	0.81	0.11	0.94	0.68	0.16
	1000	508	0.81	0.11	0.81	0.76	0.18
Lung tumor	100	513	0.86	0.24	1.00	0.61	0.15
	200	513	0.86	0.26	1.00	0.69	0.18
Mediastinal shift	100	512	0.86	0.06	0.83	0.88	0.14
Nodule or Mass	100	508	0.81	0.22	0.77	0.74	0.29
	200	508	0.79	0.19	0.91	0.61	0.25
	400	508	0.83	0.22	0.91	0.62	0.26
Pleural effusion	100	505	0.87	0.37	0.95	0.68	0.21
	200	505	0.88	0.29	0.86	0.76	0.24
	400	505	0.91	0.46	0.86	0.83	0.30
	1000	505	0.92	0.45	0.95	0.75	0.26
Pleural thickening	100	510	0.85	0.28	0.94	0.64	0.25
	200	510	0.82	0.24	0.74	0.79	0.29
	400	510	0.84	0.31	0.74	0.81	0.32
	1000	510	0.90	0.40	0.90	0.80	0.36
Pneumonia	100	508	0.80	0.26	0.78	0.76	0.37
	200	508	0.86	0.44	0.91	0.72	0.38
	400	508	0.87	0.44	0.96	0.73	0.40

Pneumothorax	100	514	0.92	0.36	1.00	0.75	0.06
Pulmonary fibrosis	100	500	0.80	0.30	0.65	0.80	0.33
	200	500	0.84	0.40	0.73	0.82	0.39
	400	500	0.83	0.36	0.68	0.87	0.43
	1000	500	0.86	0.39	0.88	0.70	0.33
	2000	500	0.85	0.43	0.78	0.78	0.36
Tuberculosis	100	509	0.83	0.31	0.74	0.87	0.36
	200	509	0.87	0.39	0.81	0.79	0.29
	400	509	0.90	0.45	0.78	0.84	0.33

Table 1: Area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), sensitivity, specificity, and F1 score for all test sets across findings and fine-tuning sample sizes.

These results demonstrate that different findings have varying data requirements to achieve target performance levels, with some conditions reaching plateau performance earlier than others.

Training Duration Scaling

Understanding computational time requirements is important for planning model development timelines and resource allocation. In our experiments, training duration scales linearly with dataset size *(Figure 2 below)*.

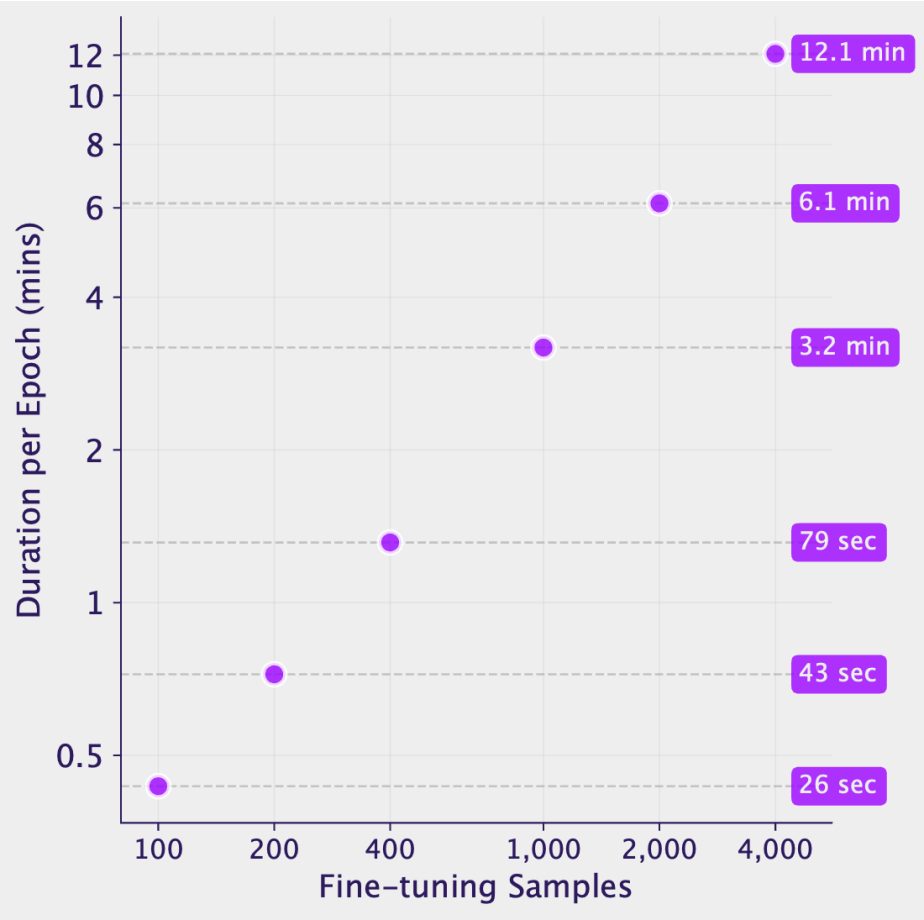


Figure 2. ***Training duration per epoch versus dataset size.*** *Duration measurements from fine-tuning experiments demonstrate linear scaling from 26 seconds (100 samples) to 12.1 minutes (4000 samples). Each data point represents the mean duration across completed training jobs. Both axes are logarithmic scaled since fine-tuning data sizes ranged from 100 to 4,000 studies with varying increments.*

Increasing the dataset size by 100 fine-tuning samples increased the training time per epoch by 18 seconds.

For experiments with 1000 samples over 20 epochs, fine-tuning took approximately 64 minutes (mean = 63.64 min; standard deviation = 0.55 min; $n_{\text{experiments with 1000 samples}} = 6$) to complete.

This predictable scaling relationship enables healthcare teams to estimate computational requirements and project timelines based on their available labeled data, supporting informed decisions about the balance between dataset size, performance targets, and computational

resources. With the HOPPR™ AI Development Platform, users can build models with less than one thousand labeled studies and within a few hours of fine-tuning.

Conclusion

Foundation models represent a powerful shift in medical imaging AI, making it possible to develop high-performing point solutions from limited, site-specific data. The HOPPR™ AI Development Platform bridges the gap between advanced AI and real-world integration by empowering healthcare teams to quickly and securely adapt foundation models to their specific workflows with minimal technical burden. Our results show that even small datasets can produce reliable models that generalize well, offering a scalable path to faster diagnoses, streamlined triage, and more responsive care delivery.

Citations

1. Yoon, S. H. *et al.* Use of artificial intelligence in triaging of chest radiographs to reduce radiologists' workload. *Eur. Radiol.* **34**, 1094–1103 (2024).
2. Sloan, P., Clatworthy, P., Simpson, E. & Mirmehdi, M. Automated Radiology Report Generation: A Review of Recent Advances. *IEEE Rev. Biomed. Eng.* **18**, 368–387 (2025).
3. Zhou, C. *et al.* A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT. Preprint at https://doi.org/10.48550/arXiv.2302.09419 (2023).
4. Bommasani, R. *et al.* On the Opportunities and Risks of Foundation Models. Preprint at https://doi.org/10.48550/arXiv.2108.07258 (2022).
5. Ricci Lara, M. A., Echeveste, R. & Ferrante, E. Addressing fairness in artificial intelligence for medical imaging. *Nat. Commun.* **13**, 4581 (2022).
6. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. Preprint at https://doi.org/10.48550/arXiv.1912.01703 (2019).
7. Slavkova, K. P. *et al.* HOPPR Medical-Grade Platform for Medical Imaging AI. Preprint at https://doi.org/10.48550/arXiv.2411.17891 (2024).
8. hopprai: SDK for interacting with the HOPPR API.
9. Goldberger, A. L. *et al.* PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* **101**, E215-220 (2000).
10. Nguyen, H. Q., Pham, H. H., Tuan Linh, Le, Dao, M. & Khanh, Lam. VinDr-CXR: An open dataset of chest X-rays with radiologist annotations. PhysioNet https://doi.org/10.13026/3AKN-B287.

11. Nguyen, H. Q. *et al*. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. Preprint at https://doi.org/10.48550/arXiv.2012.15029 (2022).