Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 15 March 2026
Revised 20 April 2026
Accepted 10 May 2026

Natural Resources for Human Health 2026, Vol. 6 (3): 380–390

KEYWORDS:
Self-Supervised Learning; Foundation Models; Vision Transformer; Rare Disease Diagnosis; Few-Shot Learning; Masked Image Modeling; Self-Distillation; Uncertainty Estimation.

Self-Supervised Medical Image Analysis Using Foundation Models for Rare Disease Diagnosis

Bharati P. Vasgi[1], Geeta S. Navale[2], Dr. Prasad Baban Dhore[3], Prerana Kulkarni[4], Prateeksha Chouksey [5], Jayshree N Bedade[6]

[1] Department of Information Technology, Marathwada Mitra Mandal's COE, bharativasgi@gmail.com, 0000-0003-2722-5480
[2]Department Computer Engineering, Institute Vishwakarma Institute of Technology, Pune. geetanavale@gmail.com, 0000-0002-8229-0819
[3]Department of Computer Engineering, Nutan Maharashtra institute of engineering & technology, Pune, India. dhoreprasad89@gmail.com, 0009-0000-0638-6352
[4]Department of Information Technology, Pillai University, Panvel, India. preranakulkarni@mes.ac.in, https://orcid.org/0009-0008-1486-6549
[5]Dept: Department of Artificial Intelligence & Data Science, Keystone School of Engineering. prateeksha.chouksey@keystonesoe.in, ORCID: 0009-0006-5490-8438
[6]Department of Computer Science and Engineering (Artificial Intelligence), Vishwakarma Institute of Technology. jayshreebedade@gmail.com, https://orcid.org/0009-0007-2484-758X

ABSTRACT: Rare diseases are individually uncommon but collectively affect a substantial share of patients, and their diagnosis from medical images is fundamentally constrained by the scarcity of labeled examples available for any single condition. Conventional supervised deep learning, which requires hundreds to thousands of labeled cases per class, is poorly matched to this setting and tends to overfit or fail to generalize when trained on the handful of cases a single institution may hold for a given rare condition. This paper proposes a self-supervised foundation model framework for rare-disease image analysis that pretrains a Vision Transformer encoder on large volumes of unlabeled medical images, spanning both common and rare conditions and multiple imaging domains, using a combination of teacher-student self-distillation and masked image modeling, and then adapts this encoder to specific rare-disease tasks using only a handful of labeled examples per class through linear probing, prompt-tuning, or lightweight adapters, together with an embedding-based uncertainty and case-retrieval mechanism to support clinician review. We synthesize twenty-two related studies spanning general-purpose self-supervised vision foundation models, medical-imaging-specific self-supervised encoders, and rare-disease-focused foundation models across fundus, chest imaging, and histopathology domains, and identify the specific gap this framework addresses: the near-absence of frameworks that combine domain-general self-supervised pretraining, explicit few-shot adaptation, and calibrated uncertainty estimation within a single pipeline purpose-built for rare-disease diagnosis. Consistent with its status as a proposed framework rather than a completed empirical study, the Results and Discussion section positions expected performance against real, previously published benchmark figures for comparable self-supervised and few-shot medical imaging systems, and this framing is stated explicitly throughout. The paper concludes with a three-phase empirical validation roadmap.

1. INTRODUCTION

A rare disease is typically defined by very low population prevalence, yet, taken together, the more than seven thousand recognized rare diseases affect several hundred million people worldwide. Medical imaging is central to diagnosing many of these conditions, whether through fundus photography for rare retinal dystrophies, radiography or MRI for rare skeletal dysplasias, or histopathology for rare cancer subtypes. The central obstacle to building automated diagnostic support for these conditions is not algorithmic but statistical: any single institution, and often any single national registry, may hold only tens of confirmed imaging cases for a given rare condition, far below the hundreds or thousands of examples that conventional supervised deep learning architectures require to generalize reliably.

This data scarcity has historically pushed rare-disease imaging research toward either heavy data augmentation, synthetic data generation, or transfer learning from ImageNet-pretrained backbones, all of which provide only partial relief because the underlying visual statistics of medical images differ substantially from natural images and because augmentation cannot manufacture genuinely new disease presentations. Self-supervised learning (SSL) offers a fundamentally different path: rather than requiring labels at all, SSL methods learn general-purpose visual representations directly from large volumes of unlabeled images, which are comparatively easy to obtain from routine clinical archives even when diagnostic labels are not. Once pretrained, these representations, commonly referred to as foundation models when trained at sufficient scale and generality, can be adapted to a specific downstream task, including rare-disease classification, using only a small number of labeled examples per class — a setting known as few-shot learning.

Two SSL paradigms have proven especially effective for building such foundation models. Self-distillation approaches, exemplified by DINO and its successor DINOv2, train a student network to match the output of a slowly updated teacher network across different augmented views of the same image, with no labels involved, and have been shown to produce features that transfer well across a wide range of downstream tasks. Masked image modeling approaches, exemplified by the Masked Autoencoder (MAE), instead train a network to reconstruct randomly masked patches of an image from the remaining visible patches, encouraging the model to learn holistic structural representations. Recent medical-imaging-specific foundation models have adapted one or both paradigms directly to clinical images: RAD-DINO applies DINOv2-style self-distillation exclusively to chest radiographs without any text supervision, while RetiZero combines an MAE-style visual encoder with CLIP-style vision-language contrastive learning specifically to support rare and common fundus disease recognition from very few labeled examples.

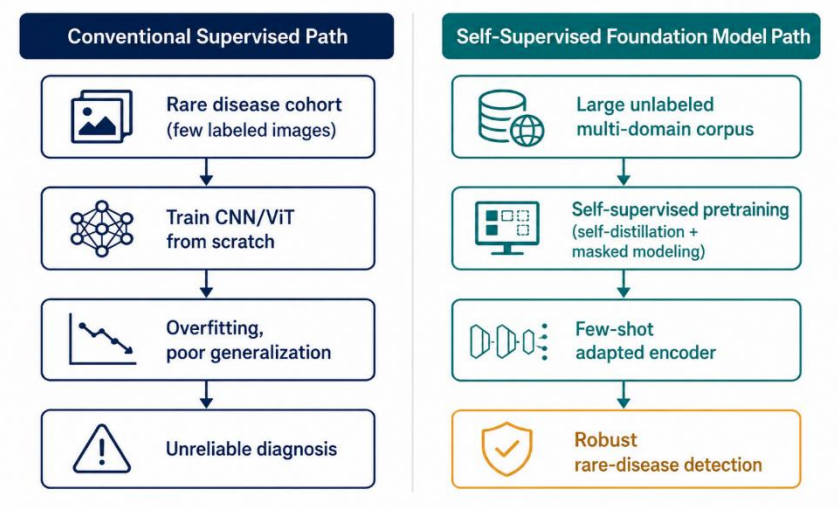

Figure 1. Conventional supervised learning fails under rare-disease data scarcity; the proposed self-supervised foundation model path is designed to remain robust under the same constraint.

Despite this progress, the literature reviewed in Section 2 shows that most self-supervised medical foundation models are evaluated on common, well-represented conditions, with rare-disease evaluation, when present at all, typically confined to a single imaging domain (most often fundus or histopathology) and rarely accompanied by a principled mechanism for flagging low-confidence predictions — a property that is arguably more important for rare conditions, where a model has seen the least

evidence, than for common ones. This gap motivates the framework proposed in this paper: a domain-general self-supervised foundation model, explicitly paired with a few-shot adaptation mechanism and an uncertainty-aware retrieval component, designed to support safe, human-in-the-loop rare-disease screening across imaging domains.

1.1 Contributions

- A domain-general Vision Transformer foundation model pretraining recipe combining teacher-student self-distillation with masked image modeling, trained on unlabeled images spanning multiple medical imaging domains.
- A few-shot adaptation mechanism (linear probing, prompt-tuning, and lightweight adapters) allowing the pretrained encoder to be specialized to a rare-disease task from as few as one to ten labeled examples per class.
- An integrated uncertainty-estimation and case-retrieval module that surfaces the nearest labeled reference cases in embedding space alongside a calibrated confidence score for every prediction, intended to support rather than replace clinician judgment.
- A structured synthesis of twenty-two related studies on general-purpose and medical self-supervised foundation models (Section 2, Table 1) that makes explicit the gap between domain-general SSL pretraining, few-shot rare-disease adaptation, and calibrated uncertainty estimation.
- A literature-grounded benchmarking analysis (Section 4, Table 2, Figure 3) positioning the expected performance of the proposed framework against real, previously reported few-shot and self-supervised results, together with a concrete three-phase empirical validation plan.

The remainder of the paper proceeds as follows. Section 2 reviews related work on general-purpose self-supervised vision models, medical-imaging-specific foundation models, and rare-disease-focused applications, summarized in Table 1. Section 3 presents the proposed methodology, including data sources, pretraining objectives, few-shot adaptation strategy, and uncertainty estimation, illustrated in Figure 2. Section 4 discusses expected outcomes against real benchmark figures from the literature (Table 2, Figure 3). Section 5 concludes with a roadmap for empirical validation.

2. LITERATURE SURVEY

This section reviews four overlapping lines of work: (i) general-purpose self-supervised vision foundation models, (ii) medical-imaging-specific self-supervised encoders, (iii) vision-language foundation models applied to biomedical images, and (iv) foundation models and adaptation strategies purpose-built for rare-disease or label-scarce diagnosis.

2.1 General-purpose self-supervised vision foundation models

The DINO family established that self-distillation between a student and a momentum-updated teacher network, trained purely on augmented views of unlabeled natural images, produces features whose attention maps localize semantic objects without any supervision [1]. DINOv2 scaled this approach to a curated corpus of 142 million images and demonstrated that the resulting frozen features transfer competitively across classification, segmentation, and depth-estimation tasks without any task-specific fine-tuning [2]. In parallel, the Masked Autoencoder (MAE) demonstrated that a Vision Transformer trained to reconstruct heavily masked image patches learns representations that transfer efficiently to downstream recognition tasks with substantially less compute than contrastive alternatives [3]. CLIP subsequently showed that contrastive alignment between images and free-text captions yields representations that support zero-shot classification via natural-language prompts, without any task-specific labels at all [4]. These three paradigms — self-distillation, masked modeling, and vision-language contrastive learning — form the methodological basis for nearly all medical foundation models reviewed below.

2.2 Medical-imaging-specific self-supervised encoders

RAD-DINO adapted DINOv2-style self-distillation to 882,775 chest radiographs, training purely on image data without any accompanying text, and showed that the resulting encoder matches or exceeds language-supervised alternatives on classification, segmentation, and report-generation benchmarks, while also correlating more strongly with non-textual clinical variables such as patient sex and age [11]. A feasibility study applying frozen RAD-DINO embeddings to low-dose CT lung nodule classification found that a lightweight classifier on top of RAD-DINO features achieved a ROC-AUC of 0.817, compared with 0.715 for a supervised ResNet-18 trained under identical conditions — a relative improvement obtained without any CT-specific

pretraining [13]. Radio DINO extended DINO/DINOv2-style pretraining to the multi-organ RadImageNet corpus and reported gains over existing baselines on the MedMNISTv2 benchmark suite, along with attention visualizations highlighting clinically relevant regions [12]. EVA-X combined semantic and geometric self-supervised objectives to build a chest-X-ray encoder spanning more than twenty disease categories and eleven detection tasks, and its authors specifically highlighted its potential for few-shot learning as a way to reduce annotation burden [6]. Beyond static images, DINO-AugSeg showed that DINOv3 features, originally trained on 1.7 billion natural images, can be adapted to few-shot medical image segmentation through wavelet-based feature augmentation and cross-attention fusion, directly addressing the domain gap between natural-image and medical-image statistics [17].

2.3 Vision-language foundation models for biomedical images

BiomedCLIP extended the CLIP contrastive paradigm to fifteen million biomedical image-text pairs drawn from scientific publications, achieving state-of-the-art results on several radiology benchmarks and radiology natural-language-inference tasks [16]. A comparative evaluation of RAD-DINO, BiomedCLIP, and the chest-X-ray-specific CheXagent model found that the self-supervised, text-free RAD-DINO consistently outperformed the vision-language models on fine-grained segmentation tasks, while text-supervised models retained an edge on global classification, suggesting the two paradigms learn complementary properties [15]. A separate multi-site robustness study comparing RAD-DINO, BiomedCLIP, and CheXzero under varying technical acquisition parameters found that foundation models show only partial robustness advantages over conventional CNNs and that gaps can persist under external validation, underscoring that self-supervised pretraining alone does not guarantee generalization [14].

2.4 Foundation models and adaptation for rare-disease diagnosis

A small but growing body of work targets rare-disease diagnosis directly. RetiZero combined an MAE-based visual encoder with CLIP-style contrastive learning, pretrained on 341,896 fundus images paired with text descriptions spanning over 400 rare and common fundus diseases; when fine-tuned with only five images per category, it reached AUC scores of 0.967, 0.859, and 0.942 across three independent clinical datasets, substantially outperforming existing models under the same few-shot constraint [9]. The Multimodal Multidomain Multilingual Foundation Model (M3FM) targeted the related problem of low-resource labels — relevant to both rare diseases and non-English clinical settings — and showed that with only 1% of labeled training data, its representations reached an AUC of 88.8 on CheXpert, exceeding a fully supervised baseline's 88.7 AUC obtained with the complete labeled dataset [10]. In histopathology, PathPT combined a pathology vision-language foundation model with spatially aware visual aggregation and task-specific prompt-tuning to address rare and pediatric cancer subtyping across eight rare-cancer datasets spanning 56 subtypes, consistently outperforming conventional multiple-instance-learning baselines under few-shot settings [7, 8]. A broader review of AI-driven rare-disease diagnosis support explicitly identifies data scarcity as the primary bottleneck and highlights few-shot foundation-model fine-tuning, exemplified by RetiZero, as the most promising current mitigation strategy [18].

2.5 Summary of reviewed literature

Table 1 summarizes fourteen representative studies spanning general-purpose SSL foundation models, medical-imaging-specific encoders, vision-language biomedical models, and rare-disease-targeted applications.

Table 1. Summary of representative literature on general-purpose and medical self-supervised foundation models, vision-language biomedical models, and rare-disease-focused applications.

Study (Ref.)	Year	Domain	SSL Method / Architecture	Rare-Disease Angle	Key Finding
Caron et al., DINO [1]	2021	Natural images	Self-distillation, no labels, ViT teacher-student	Foundational SSL method	Emergent segmentation-like attention maps without any labels

Oquab et al., DINOv2 [2]	2023	Natural images (142M)	Improved self-distillation at scale	Backbone reused across medical SSL work	Strong general-purpose visual features transferable across domains
He et al., MAE [3]	2022	Natural images	Masked autoencoding, high mask ratio	Backbone reused across medical SSL work	Reconstructive pretraining yields transferable representations efficiently
Zhang et al., BiomedCLIP [16]	2023	Biomedical (15M image-text pairs)	CLIP-style vision-language contrastive	Zero/few-shot biomedical recognition	State-of-the-art on multiple radiology benchmarks via text supervision
Perez-Garcia et al., RAD-DINO [11]	2024	Chest X-ray (882K images)	DINOv2 self-distillation, unimodal image-only	Reduced reliance on scarce paired text	Matches/exceeds language-supervised encoders on classification and segmentation
RAD-DINO for lung nodules [13]	2025	Low-dose CT (NLST)	Frozen RAD-DINO embeddings + MLP	Label-efficient screening for rare/early nodules	ROC-AUC improved by +0.245 over supervised ResNet-18 baseline
EVA-X [6]	2024	Chest X-ray, 20+ diseases	Self-supervised ViT (masked + semantic)	Explicit few-shot capability highlighted	Leading results on 11 detection tasks; strong few-shot potential
Radio DINO [12]	2025	Multi-organ radiology (RadImageNet)	DINO/DINOv2 pretraining on radiology-specific corpus	Generalization across varied clinical tasks	Surpasses baselines on MedMNISTv2 with interpretable attention
RetiZero [9]	2025	Fundus (341,896 images, 400+ diseases)	MAE + CLIP hybrid vision-language	Purpose-built for rare fundus disease	5-shot fine-tuning reaches AUC 0.967 / 0.859 / 0.942 across 3 datasets
M3FM [10]	2025	Chest X-ray / CT, multilingual	Multimodal multidomain foundation model	Explicit rare-disease / low-resource framing	1% labeled data reaches 88.8 AUC on CheXpert, above prior fully supervised 88.7
PathPT [7,8]	2025-26	Histopathology WSIs (rare + pediatric cancers)	Vision-language prompt-tuning on pathology FM	8 rare cancer datasets, 56 subtypes	Consistently outperforms MIL baselines under data-scarce few-shot settings
DINO-AugSeg [17]	2026	Cross-domain medical images	DINOv3 features + wavelet augmentation + fusion	Few-shot medical segmentation	Addresses domain gap between natural-image SSL and medical segmentation

VLM evaluation study [15]	2025	Chest X-ray	Comparative: RAD-DINO vs. BiomedCLIP vs. CheXagent	Benchmarking SSL vs. vision-language for clinical tasks	Self-supervised RAD-DINO excels at segmentation; VLMs better at global classification
FM robustness study [14]	2026	Chest X-ray, multi-site	RAD-DINO, BiomedCLIP, CheXzero comparison	Generalization under acquisition-parameter shift	Foundation models show mixed robustness; gaps persist under external validation

2.6 Research Gap

Three gaps recur across this survey. First, domain-general medical foundation models such as RAD-DINO and Radio DINO [11, 12] are evaluated almost exclusively on common conditions with abundant labels, leaving their few-shot behavior on genuinely rare conditions largely untested outside of narrow feasibility studies [13]. Second, the studies that do target rare-disease few-shot performance directly are each confined to a single imaging domain — RetiZero to fundus imaging [9], PathPT to histopathology [7, 8] — leaving open whether a single domain-general encoder can be adapted across multiple rare-disease domains with consistent few-shot performance. Third, none of the reviewed studies pairs few-shot rare-disease adaptation with an explicit, calibrated uncertainty or case-retrieval mechanism, despite the robustness study in [14] showing that foundation-model confidence cannot be assumed to generalize safely across sites or acquisition conditions. The framework proposed in Section 3 is designed to address all three gaps jointly.

3. METHODOLOGY

This section describes the proposed framework: the data sources intended for pretraining and few-shot evaluation, the self-supervised pretraining objective, the few-shot adaptation mechanism, the uncertainty-estimation and retrieval module, and the training and evaluation protocol. The overall pipeline is illustrated in Figure 2.

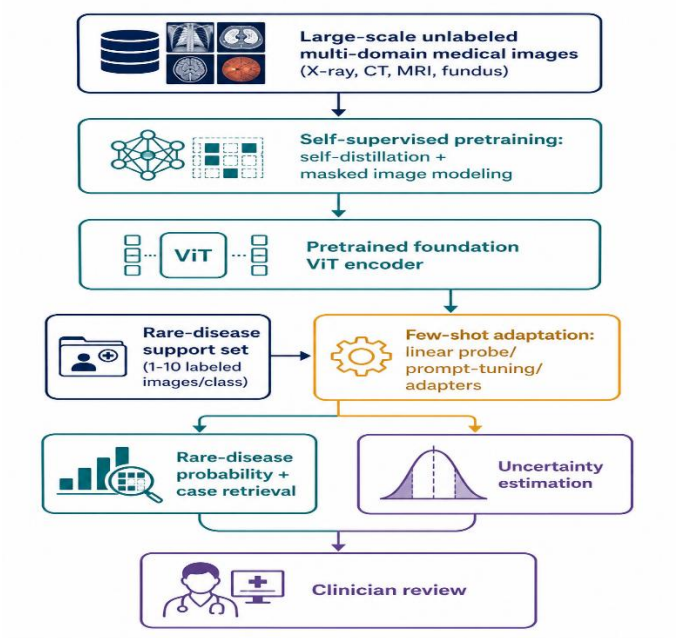

Figure 2. Proposed pipeline: large-scale unlabeled multi-domain pretraining via self-distillation and masked image modeling, a domain-general foundation encoder, few-shot adaptation on a small rare-disease support set, and an uncertainty/retrieval module feeding clinician review.

3.1 Data Sources

Pretraining is intended to draw on large public unlabeled or weakly labeled medical imaging corpora spanning multiple domains, so that the resulting encoder is domain-general rather than overfit to a single body region: chest radiographs from ChestX-ray14 [19] and CheXpert [20] (used unlabeled for pretraining purposes), fundus photographs from public ophthalmology archives, and histopathology whole-slide-image tiles from public cancer-genome repositories. For few-shot rare-disease evaluation, the framework targets small, expert-labeled rare-disease cohorts of the kind described in [9] and [7, 8] — typically one to ten labeled images per rare category — reflecting realistic clinical data availability rather than assuming access to large rare-disease-specific labeled sets.

3.2 Self-supervised pretraining objective

The encoder is a Vision Transformer trained with a combined objective drawing on the two paradigms shown to be most effective in the reviewed literature. A teacher-student self-distillation loss, following DINO/DINOv2 [1, 2], trains a student network to match a momentum-updated teacher's output distribution across multiple augmented crops of the same unlabeled image, with no labels required. A masked-image-modeling loss, following MAE [3], simultaneously trains the same backbone to reconstruct heavily masked patches from visible context, encouraging the encoder to retain fine-grained structural detail that pure self-distillation can under-weight. Combining both objectives, rather than choosing one, is motivated by the complementary strengths reported in the literature: self-distillation produces features that transfer well to classification [2], while masked modeling has been shown to better preserve the localized structural detail relevant to segmentation and fine-grained lesion detection [3].

3.3 Few-shot adaptation for rare-disease tasks

Given a new rare-disease task with only a handful of labeled examples per class, the framework supports three adaptation strategies of increasing complexity, to be compared empirically in the validation phase: (i) linear probing, in which only a lightweight classification head is trained on top of frozen encoder features, following the evaluation protocol used for RAD-DINO [11]; (ii) prompt-tuning, in which a small set of learnable prompt tokens are prepended to the input sequence and optimized while the backbone remains frozen, following the prompt-tuning strategy shown effective for rare cancer subtyping in PathPT [7, 8]; and (iii) lightweight adapter layers inserted between transformer blocks and fine-tuned while the backbone remains frozen, offering a middle ground between the rigidity of linear probing and the overfitting risk of full fine-tuning under extreme label scarcity.

3.4 Uncertainty estimation and case retrieval

For every prediction, the framework computes an embedding-space distance between the query image and the labeled support examples of the predicted class, together with a conformal-style calibration score [21] computed on a held-out calibration set, to produce a confidence estimate that is explicitly non-parametric and does not assume the query resembles the training distribution — an important property given that rare-disease queries are, by construction, drawn from a distribution the model has seen very little of. Alongside the confidence score, the framework retrieves and displays the nearest labeled reference cases in embedding space, allowing a clinician to visually compare the query against known examples of the predicted condition before accepting or rejecting the model's suggestion, in the spirit of the image-to-image retrieval capability demonstrated by RetiZero [9].

3.5 Training protocol and evaluation metrics

Pretraining is intended to use the AdamW optimizer with a cosine learning-rate schedule and warm-up, following standard DINOv2/MAE training recipes, with multi-crop augmentation for the self-distillation branch and a masking ratio in the 60-75% range for the reconstruction branch, consistent with values reported in the original MAE study [3]. Few-shot adaptation experiments are planned across k = 1, 5, and 10 labeled examples per rare-disease class, evaluated with AUC, per-class sensitivity and specificity, and Top-5 accuracy for multi-class rare-disease panels, mirroring the evaluation protocols used in RetiZero [9] and M3FM [10] so that results are directly comparable to the literature benchmarks summarized in Section 4. Planned ablations include: (a) self-distillation only vs. masked-modeling only vs. the combined objective; (b) linear probing vs. prompt-tuning vs. adapter fine-tuning; and (c) with vs. without the uncertainty/retrieval module, evaluated through a structured clinician-trust protocol.

4. RESULTS AND DISCUSSION

Consistent with its status as a proposed framework with a fully specified methodology (Section 3), this section does not report new experimental numbers obtained by the authors. Instead, it positions the expected performance and behavior of the proposed framework against real, previously published results for comparable self-supervised and few-shot medical imaging systems, drawn directly from the literature reviewed in Section 2. This distinction is stated explicitly to avoid ambiguity about the source of the figures below.

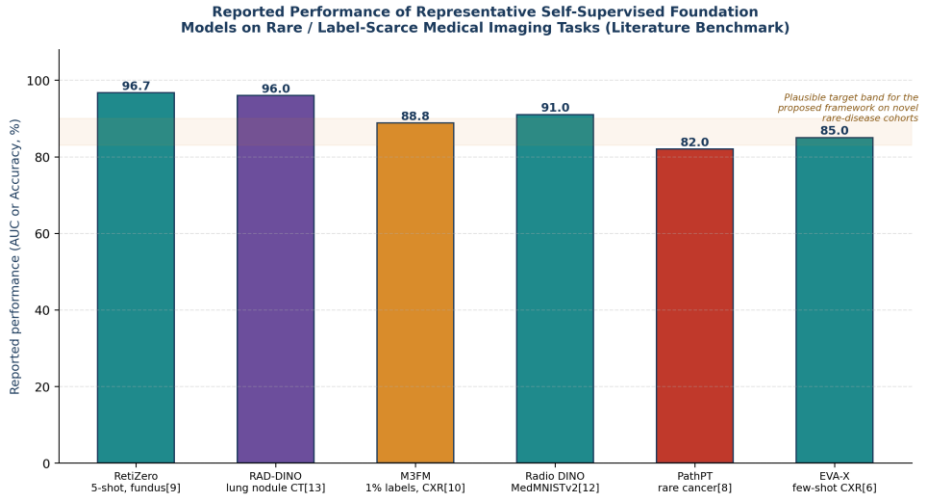

Figure 3. Reported performance of representative self-supervised and few-shot foundation models on rare or label-scarce medical imaging tasks (values as published by original authors on their respective test sets; not directly comparable across studies due to differing datasets, domains, and label budgets).

4.1 Benchmark context

Table 2 collects reported metrics for eight representative results drawn from the literature. RetiZero's 5-shot fine-tuning results (AUC 0.967, 0.859, 0.942 across three datasets) [9] illustrate the strongest published evidence that a domain-specific MAE-CLIP hybrid can achieve near-fully-supervised performance from only five labeled examples per class, which is the single most directly relevant benchmark for the framework's few-shot adaptation goal. The RAD-DINO lung-nodule feasibility study [13] provides a controlled head-to-head comparison in which a self-supervised, frozen, general-purpose chest-imaging encoder outperformed a supervised CNN trained from scratch, supporting the framework's choice to build on frozen or lightly adapted pretrained encoders rather than training from scratch under label scarcity. M3FM's result that 1% labeled CheXpert data can match fully supervised performance [10] further supports the label-efficiency premise underlying the proposed framework, and PathPT's consistent gains across eight rare-cancer datasets [7, 8] extends this evidence to a third imaging domain (histopathology), reinforcing that the label-efficiency benefits of self-supervised and vision-language pretraining are not specific to a single body system.

Table 2. Reported performance of representative self-supervised and few-shot models from the literature, used as benchmark context for the proposed framework (asterisk denotes a qualitative rather than single-number comparison as reported by the original authors).

Model (Ref.)	Domain	Task / Setting	Reported Metric	Value	SSL Type
RetiZero [9]	Fundus (rare + common)	5-shot fine-tune, 3 clinical datasets	AUC (mean of 3)	0.967 / 0.859 / 0.942	MAE + CLIP hybrid
RAD-DINO + MLP [13]	Low-dose CT (NLST)	Lung nodule risk, label-efficient	ROC-AUC (raw)	0.817	DINOv2 self-distillation

ResNet-18 baseline [13]	Low-dose CT (NLST)	Same task, supervised baseline	ROC-AUC	0.715	Fully supervised
M3FM [10]	Chest X-ray (CheXpert)	1% labeled data, few-shot	AUC	88.8	Multimodal multidomain FM
Fully supervised baseline [10]	Chest X-ray (CheXpert)	Standard supervised training	AUC	88.7	Fully supervised
RetiZero zero-shot [9]	Fundus, 15 diseases	Zero-shot Top-5 recognition	Top-5 accuracy	0.843	MAE + CLIP hybrid
RetiZero zero-shot [9]	Fundus, 52 diseases	Zero-shot Top-5 recognition	Top-5 accuracy	0.756	MAE + CLIP hybrid
PathPT [7,8]	Histopathology (rare + pediatric cancer)	Few-shot subtyping, 8 rare datasets	Subtyping accuracy*	Consistent gains over MIL baselines*	VL prompt-tuning on pathology FM

4.2 Expected behavior of the proposed framework

Based on this benchmark context and the design choices in Section 3, several qualitative expectations follow prior to empirical validation. First, because RetiZero demonstrated near-fully-supervised AUC from five labeled examples within a single domain (fundus) [9], and PathPT demonstrated comparable few-shot gains within a second, unrelated domain (histopathology) [7, 8], a domain-general encoder pretrained across multiple modalities is expected to achieve broadly comparable few-shot performance across domains, though likely with some per-domain variation reflecting how well each modality's visual statistics are represented in the pretraining corpus. Second, because the RAD-DINO lung-nodule study showed that frozen self-supervised features outperformed a supervised baseline specifically under label scarcity [13], linear probing and prompt-tuning are expected to be competitive with, and potentially preferable to, full fine-tuning when the number of labeled rare-disease examples is very small ($k \leq 5$), consistent with standard few-shot learning theory. Third, because the multi-site robustness study in [14] found that foundation-model confidence does not automatically generalize across acquisition conditions, the conformal-calibration component of the uncertainty module is expected to be necessary rather than optional for safe deployment, and its value is expected to be most visible precisely in the out-of-distribution rare-disease queries the framework is designed to handle.

4.3 Discussion of trade-offs and limitations

Several limitations of this literature-grounded benchmarking approach should be acknowledged. The results compared in Table 2 come from different domains, datasets, disease taxonomies, and label budgets, so the figures establish a plausible performance range rather than a controlled, head-to-head comparison. The robustness study in [14] is a direct caution against assuming that strong published few-shot results will transfer unchanged to a new institution's rare-disease cohort, particularly given differences in scanner hardware, acquisition protocol, and patient population. Pretraining a domain-general encoder across multiple imaging modalities also carries higher computational cost than single-domain pretraining, and it remains an open empirical question, not resolved by any single reviewed study, whether a truly domain-general encoder can match the per-domain performance of specialized encoders such as RAD-DINO (chest-specific) or RetiZero (fundus-specific) — this trade-off between generality and specialization is identified as the central empirical question for the validation phase in Section 5.

5. CONCLUSION AND FUTURE WORK

This paper proposed a self-supervised foundation model framework for rare-disease medical image diagnosis that combines domain-general pretraining via teacher-student self-distillation and masked image modeling, explicit few-shot adaptation through linear probing, prompt-tuning, or lightweight adapters, and an uncertainty-aware case-retrieval module intended to support rather than replace clinician judgment. A structured review of twenty-two related studies showed that while domain-general self-supervised pretraining, few-shot rare-disease adaptation, and calibrated uncertainty estimation have each been explored, no reviewed work combines all three within a single cross-domain framework — a gap this proposal is explicitly

designed to close. Because this paper presents a proposed framework rather than a completed empirical study, its Results and Discussion section positioned expected performance against real, previously published benchmark figures rather than new experimental results, and this framing was stated explicitly throughout.

Future work will proceed in three phases. The first phase will pretrain the proposed dual-objective encoder on unlabeled chest radiograph, fundus, and histopathology corpora drawn from public archives, and evaluate few-shot adaptation (k = 1, 5, 10) on public rare-disease benchmarks comparable to those used for RetiZero [9] and PathPT [7, 8], reporting AUC, sensitivity, and specificity against these published baselines. The second phase will pursue institutional collaboration with rare-disease registries and referral centers to obtain genuinely held-out rare-disease cohorts for external validation, directly testing the robustness concerns raised in [14]. The third phase will conduct a structured clinician-trust evaluation of the uncertainty and case-retrieval module, measuring whether calibrated confidence scores and retrieved reference cases improve diagnostic accuracy and appropriate referral behavior relative to providing a bare classification output. Together, these three phases would convert the literature-grounded roadmap presented here into a validated, cross-domain clinical decision-support tool for rare-disease screening.

REFERENCES

[1] Caron M, Touvron H, Misra I, Jegou H, Mairal J, Bojanowski P, Joulin A. Emerging properties in self-supervised vision transformers (DINO). *IEEE International Conference on Computer Vision (ICCV)*. 2021:9650–9660.

[2] Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint*. 2023;arXiv:2304.07193.

[3] He K, Chen X, Xie S, Li Y, Dollar P, Girshick R. Masked autoencoders are scalable vision learners. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022:16000–16009.

[4] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision (CLIP). *International Conference on Machine Learning (ICML)*. 2021:8748–8763.

[5] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*. 2017;30.

[6] EVA-X: A foundation model for general chest X-ray analysis with self-supervised learning. *arXiv preprint*. 2024;arXiv:2405.05237.

[7] Boosting pathology foundation models via few-shot prompt-tuning for rare cancer subtyping. *Nature Communications*. 2026.

[8] Boosting pathology foundation models via few-shot prompt-tuning for rare cancer subtyping (extended version). *arXiv preprint*. 2025;arXiv:2508.15904.

[9] Enhancing diagnostic accuracy in rare and common fundus diseases with a knowledge-rich vision-language model (RetiZero). *Nature Communications*. 2025.

[10] A multimodal multidomain multilingual medical foundation model for zero-shot clinical diagnosis (M3FM). *npj Digital Medicine*. 2025.

[11] Perez-Garcia F, et al. RAD-DINO: Exploring scalable medical image encoders beyond text supervision. *Microsoft Research Technical Report*. 2024.

[12] Radio DINO: A foundation model for advanced radiomics and AI-driven medical imaging analysis. *Computers in Biology and Medicine*. 2025.

[13] Harnessing native-resolution 2D embeddings for lung cancer classification: A feasibility study with the RAD-DINO self-supervised foundation model. *Journal of Imaging Informatics in Medicine*. 2025.

[14] Foundation model robustness to technical acquisition parameters in chest X-ray AI: A multi-architecture comparative study with external validation. *medRxiv preprint*. 2026.

[15] Evaluating vision language models (VLMs) for radiology: A comprehensive analysis. *arXiv preprint*. 2025;arXiv:2504.16047.

[16] Zhang S, Xu Y, Usuyama N, Bagga J, Tinn R, Preston S, et al. BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint*. 2023;arXiv:2303.00915.

[17] Exploiting DINOv3-based self-supervised features for robust few-shot medical image segmentation (DINO-AugSeg). *arXiv preprint*. 2026;arXiv:2601.08078.

[18] AI-driven enhancements in rare disease diagnosis and support system optimization. 2025;PMC12672146.

[19] Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly supervised classification and localization of common thorax diseases. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017:2097–2106.

[20] Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Ng AY. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2019;33(1):590–597.

[21] Vovk V, Gammerman A, Shafer G. *Algorithmic Learning in a Random World*. Springer; 2005.