Health Care Science

HEALTH CARE SCIENCE

ORIGINAL ARTICLE OPEN ACCESS

Rethinking Domain-Specific Pretraining by Supervised or Self-Supervised Learning for Chest Radiograph Classification: A Comparative Study Against ImageNet Counterparts in Cold-Start Active Learning

Han Yuan[1] | Mingcheng Zhu[1,2] | Rui Yang[1] | Han Liu[3] | Irene Li[4] | Chuan Hong[5]

[1]Duke-NUS Medical School, Centre for Quantitative Medicine, Singapore, Singapore | [2]Department of Engineering Science, University of Oxford, Oxford, UK | [3]Department of Computer Science, Vanderbilt University, Nashville, Tennessee, USA | [4]Information Technology Center, University of Tokyo, Bunkyo-ku, Japan | [5]Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina, USA

Correspondence: Chuan Hong (chuan.hong@duke.edu)

Received: 10 October 2024 | **Revised:** 5 January 2025 | **Accepted:** 26 January 2025

Funding: The authors received no specific funding for this work.

Keywords: chest radiograph analysis | cold-start active learning | COVID-19 | psychiatric pneumonia | radiology foundation model

ABSTRACT

Objective: Deep learning (DL) has become the prevailing method in chest radiograph analysis, yet its performance heavily depends on large quantities of annotated images. To mitigate the cost, cold-start active learning (AL), comprising an initialization followed by subsequent learning, selects a small subset of informative data points for labeling. Recent advancements in pretrained models by supervised or self-supervised learning tailored to chest radiograph have shown broad applicability to diverse downstream tasks. However, their potential in cold-start AL remains unexplored.

Methods: To validate the efficacy of domain-specific pretraining, we compared two foundation models: supervised TXRV and self-supervised REMEDIS with their general domain counterparts pretrained on ImageNet. Model performance was evaluated at both initialization and subsequent learning stages on two diagnostic tasks: psychiatric pneumonia and COVID-19. For initialization, we assessed their integration with three strategies: diversity, uncertainty, and hybrid sampling. For subsequent learning, we focused on uncertainty sampling powered by different pretrained models. We also conducted statistical tests to compare the foundation models with ImageNet counterparts, investigate the relationship between initialization and subsequent learning, examine the performance of one-shot initialization against the full AL process, and investigate the influence of class balance in initialization samples on initialization and subsequent learning.

Results: First, domain-specific foundation models failed to outperform ImageNet counterparts in six out of eight experiments on informative sample selection. Both domain-specific and general pretrained models were unable to generate representations that could substitute for the original images as model inputs in seven of the eight scenarios. However, pretrained model-based initialization surpassed random sampling, the default approach in cold-start AL. Second, initialization performance was positively correlated with subsequent learning performance, highlighting the importance of initialization strategies. Third, one-shot initialization performed comparably to the full AL process, demonstrating the potential of reducing experts' repeated waiting during AL iterations. Last, a U-shaped correlation was observed between the class balance of initialization samples and model

Abbreviations: AL, active learning; AUPRC, area under the precision-recall curve; AUROC, area under the receiver operating characteristic curve; DL, deep learning; MC, Monte Carlo; MLP-3, three-layer multilayer perceptron; REMEDIS, Robust and Efficient MEDical Imaging with Self-supervision; SGD, stochastic gradient descent; TXRV, TorchXRayVision.

performance, suggesting that the class balance is more strongly associated with performance at middle budget levels than at low or high budgets.

Conclusions: In this study, we highlighted the limitations of medical pretraining compared to general pretraining in the context of cold-start AL. We also identified promising outcomes related to cold-start AL, including initialization based on pretrained models, the positive influence of initialization on subsequent learning, the potential for one-shot initialization, and the influence of class balance on middle-budget AL. Researchers are encouraged to improve medical pretraining for versatile DL foundations and explore novel AL methods.

1 | Background

This section begins by motivating the use of active learning (AL) to reduce annotation costs in deep learning (DL) models, especially its orthogonal value compared to other strategies for DL on limited annotated samples, including data augmentation, transfer learning, and semisupervised learning. We then differentiate between warm-start and cold-start AL, highlighting that cold-start AL better addresses real-world scenarios. Next, we illustrate common strategies and related work for cold-start AL, focusing on both the initialization and subsequent learning stages. After that, we demonstrate the potential of domain-specific pretrained models, also known as foundation models, in enhancing cold-start AL. Finally, we summarize our key contributions.

1.1 | Motivation

DL has achieved remarkable success in chest radiograph analysis [1–3], but its performance heavily relies on large volumes of chest radiographs and high-quality diagnostic annotations [4, 5]. Unlike natural scene labeling, which primarily relies on common sense [6, 7] and can leverage crowdsourcing platforms [8], chest radiograph annotation requires specialized expertise [9], making it time-consuming and cost-intensive [10]. To reduce annotation costs, alleviate clinician workload, and optimize computational resources by avoiding redundant data [11–13], AL has been proposed to iteratively select a small subset of data points whose annotations are most beneficial for model convergence, querying these labels from experienced medical professionals as oracles [10, 14, 15].

Although various strategies have been proposed to address DL under the constraint of limited annotated samples, AL offers a distinct and irreplaceable advantage worthy of focused investigation. A commonly considered approach is data augmentation, which involves applying transformations to existing labeled data. However, this strategy potentially fails to introduce truly novel or representative information and can even degrade DL performance when augmented samples are physically implausible or semantically meaningless [16]. In contrast, AL selects genuine data samples, avoiding the reinforcement of existing biases, and the misrepresentation of real-world properties. Another popular technique is transfer learning, which leverages a model pretrained on a source data set for the target data set. However, transfer learning can suffer from distribution mismatches between the source and target data sets [17]. Even with fine-tuning on small annotated samples, biases inherent in the source data set are often challenging to mitigate [18]. Under the same annotation budget, AL directly optimizes the model for the target data set, ensuring more efficient use of resources. A third approach is semisupervised learning, which builds on limited annotated data by exploiting the structure of unlabeled data. However, if the initial labeled data set is poorly representative, semi-supervised learning risks propagating errors and failing to generalize effectively [19]. AL, however, dynamically adapts to the model's learning state and iteratively refines the model through active querying, thus mitigating the risks of error propagation and poor generalization. It is important to highlight that this comparison aims to underscore AL's complementary value rather than diminish the utility of other methods. Indeed, these techniques can be integrated with AL to enhance performance, as suggested in prior studies [20–22]. In this work, we focus exclusively on AL to systematically investigate the potential of AL in the context of recent advancements in foundation models.

1.2 | Related Work

1.2.1 | Warm-Start and Cold-Start AL

AL methods can be broadly classified into warm-start and cold-start AL, depending on the initialization stage [23]. Specifically, warm-start AL typically involves two stages: an initial phase where the model is trained on a small, preselected, annotated subset of images, and a subsequent learning phase where various query strategies are employed to select additional images for annotation and model fine-tuning based on the trained model [24]. Cold-start AL also comprises two stages, but unlike warm-start AL, it begins without any annotated samples. Instead, it autonomously selects initial samples, sends them to oracles for annotation, and proceeds with model training [25].

Although warm-start AL is commonly studied and has been applied to a spectrum of clinical tasks such as breast mass localization [9], white matter tract segmentation [10], optical coherence tomography segmentation [26], and so on [6, 7, 27–33], it requires preselection of sample annotation belonging to diverse classes in the initial stage. This reliance is often impractical in real-world AL scenarios, where none of the samples in a new data set are labeled, making it impossible to prepare representative instances for each category [34], especially in medical scenarios with a class-skewed distribution [35]. Therefore, cold-start AL is more suited to real-world applications and becomes the focus of our study.

1.2.2 | Cold-Start AL Strategies

The primary challenge in cold-start AL lies in the initialization phase: how to select annotation-worthy samples that cover diverse classes and significantly contribute to model convergence in the absence of label information. In other words, how can raw image pixels be utilized to identify samples that merit labeling? Upon completion of the initial sample selection and annotation, standard warm-start strategies can be employed because the model, following the initialization stage, has developed sufficient competency on the target data and task, thereby satisfying the prerequisites for warm-start AL.

For the initialization stage, random sampling is often the first method considered by researchers. Although this method works well on balanced data sets, it typically requires selecting a large number of instances to capture all potential classes in imbalanced scenarios [35], which is impractical for AL formulations and overlooks the informative sample features [36]. To address these limitations, diversity sampling, also known as representativeness sampling, has been proposed. This method selects samples that are representative of the underlying data distribution of diverse classes [31] based on the modeling of raw image pixels [37] in an unsupervised or self-supervised manner. For example, He et al. proposed a two-stage clustering approach to address the cold-start problem in AL initialization, which is adaptable to class imbalance [35]. In the first stage, the density peak clustering algorithm [38] was used to separate samples from majority and minority classes into distinct clusters. In the second stage, a cluster-adaptive method was employed to identify the most representative samples within each cluster. This approach effectively selects samples that improve both class coverage and model performance.

For the subsequent learning stage, previous initialization methods, such as random sampling [39] and diversity [40], can also be employed. These approaches do not rely on annotation information or models developed from the initialization phase, whereas other methods for subsequent learning typically do. Due to its simplicity [41] and outperformance [42], uncertainty sampling, also known as informativeness sampling, is the most widely used approach [31]. Specifically, this method selects data points where the current model exhibits the greatest uncertainty, often those near the decision boundary [41]. Common uncertainty metrics include margin of confidence, least confidence, and entropy [43, 44]. Uncertainty sampling can be further integrated with diversity or other strategies to form hybrid methods [23, 45–47]. For instance, Yang et al. proposed an annotation suggestion method that integrates uncertainty and diversity [48]. They first calculated the variance across a set of bootstrap-aggregated models [49], and then identified high-variance unlabeled samples [50]. Among these, the samples with the highest similarity sum to all other unlabeled samples were deemed representative and selected for annotation. Shen et al. introduced a three-step integrative strategy to gradually identify the most informative samples [51]. First, they selected a large subset with the highest uncertainty based on Monte Carlo (MC) dropout [52]. Next, they refined the subset by retaining samples that could represent the entire unlabeled set. Finally, they excluded samples already similar to annotated data. An alternative to multistep integration is the use of weighted combinations of different metrics in a single step [53]. For example, Mahapatra et al. computed sample informativeness as a weighted sum of entropy-based uncertainty and the mean squared distance between the feature vectors of candidate images and all other unlabeled samples [54].

1.2.3 | Domain-Specific Pretraining for Cold-Start AL

In chest radiograph analysis, pretraining plays a crucial role in reducing the need for large training data sets while improving model performance. Traditionally, pretraining involves the collection and annotation of large-scale data sets similar to the target data set, followed by supervised learning to develop DL models with optimal initial parameters for downstream tasks. However, the rapid growth of unlabeled data has outpaced the capacity of experts to provide annotations. To address this, researchers have introduced self-supervised learning, which exploits the inherent structure and relationships within the data to derive effective initial parameters. Self-supervised learning has been deployed on large-scale medical data sets that span different levels of specificity, from organ- or task-specific models such as those for abdominal organs [55] or sight-threatening eye diseases [56], to domain-specific models like those for chest radiographs [57, 58], and even general models capable of handling multiple domains, including dermatology photographs, fundus imaging, digital pathology, chest radiographs, and mammography [59]. Both supervised and self-supervised models can generate low-dimensional yet information-rich representation vectors for external data sets from the same target domains that they were not trained on. These numeric representations provide one of the overarching advantages of pretrained models, serving as feature inputs for downstream specialized models. Therefore, we refer to these domain-specific pretrained models, whether derived from supervised or self-supervised learning, as foundation models, and use this term interchangeably with domain-specific pretrained models in the following sections. By reducing the dimensionality relative to the original images, foundation models allow for more compact model parameters and lower the computational cost of model training [60].

Foundation model-based representations hold significant potential for use in the initialization and subsequent learning stages of cold-start AL. In the initialization phase, clustering is a common diversity sampling method, but it often faces convergence challenges due to the high dimensionality of original image pixel features [61]. These challenges can be mitigated by employing low-dimensional representation vectors [62]. Additionally, these representations can replace raw image pixels in model design, enabling more efficient parameterization during both the initialization and subsequent training stages. For instance, researchers applied the BERT foundation model [63] to address cold-start sentence classification [64]. They encoded samples into novel vectors that captured diversity through hidden representations and uncertainty via model confidence scores. Based on these vectors, they used K-means++ clustering [65] to select initial samples for annotation.

However, pretraining is not a novel concept in the DL domain. Before the advent of domain-specific foundation models,

various DL models pretrained on ImageNet [66], a general domain data set with human annotation, had already been applied to diverse external data sets [67, 68], demonstrating the ability to generate informative representations [69]. As such, ImageNet pretrained backbones should be considered valuable counterparts to foundation models, particularly because many foundation models, such as CXR foundation [57], leverage classic network architectures like ResNet [70], which also offer ImageNet pretrained versions. Therefore, a thorough evaluation of foundation models and their influence on cold-start AL is essential to understanding the capability of these models, inspiring both new application scenarios for foundation models and the development of novel AL methods in the era of self-supervised learning.

1.3 | Contributions

In this work, we provide the following contributions. First, we contribute a systematic, quantitative, and reproducible analysis to examine the effectiveness of domain-specific pretrained models against their ImageNet counterparts in both the initialization and subsequent learning stages of cold-start AL. Second, we propose a representation-based uncertainty sampling in the initialization stage of cold-start AL to address the difficulty that uncertainty sampling strategies have no access to sample labels in the initialization stage. Third, we conduct rigorous statistical tests to reveal the relationship between AL initialization and subsequent learning, the comparison between the lightweight representation-based model and raw image-based model, and the comparability of one-shot initialization and the complete AL under the same annotation budget. Last, we implement correlation tests to identify the impact of class balance of initialization samples on AL initialization and subsequent learning.

2 | Methods

In this section, we begin by introducing key notations and presenting a general formulation for cold-start AL, encompassing both the initialization and subsequent learning stages. We then detail specific strategies for each stage, leveraging domain-specific foundation models and their ImageNet counterparts. We focus primarily on binary image classification to align with the real-world experiments in the following section.

2.1 | Cold-Start AL Formulation

2.1.1 | Initialization

We denote the initial unlabeled and labeled image data sets as D_0^u and D_0^l, respectively. Before the initialization, $D_0^l = \varnothing$ is an empty set, and the unlabeled pool D_0^u contains N_0^u unlabeled two-dimensional images I_i with the width of W_0 and the height of H_0. Then a query strategy Q_0 is leveraged to select N_0^l images I_j to be annotated by oracles based on original image pixels. After that, D_0^u is updated into $D_0^u\langle(I_j), j = 1, 2, ..., N_0^l\rangle$ by removing the selected images $\langle(I_j), j = 1, 2, ..., N_0^l\rangle$ and professional medical experts such as radiologists give binary labels Y_j to each of the selected images I_j, updating D_0^l into $\langle(I_j, Y_j), j = 1, 2, ..., N_0^l\rangle$. Based on the D_0^l consisting of samples and corresponding annotations, a classifier M_c is trained to learn the projection from I_j to Y_j. When the training of M_c is completed, the initialization stage of cold-start AL is finished and proceeded into the next stage of subsequent learning.

2.1.2 | Subsequent Learning

Different from the initialization stage without any annotation information, subsequent learning has a classifier M_c with certain discriminability on the target task and therefore can use a classifier-based informative sampling strategy Q_1. Assuming a total annotation budget $B = N_0^l + k \times N_1^l$, in each query iteration $\tau = 1, 2, 3, ..., k$, the subsequent learning strategy Q_1 select N_1^l samples I_s from the unlabeled pool $D_{\tau-1}^u$ and send them to oracles for annotation, forming an annotated set $D_\tau^l = D_{\tau-1}^l \cup \langle(I_s, Y_s), s = 1, 2, ..., N_1^l\rangle$. Meanwhile, the unlabeled pool $D_\tau^u = D_{\tau-1}^u \backslash \langle(I_s), s = 1, 2, ..., N_1^l\rangle$is updated by removing the selected images I_s. Based on the updated annotated set D_τ^l, the classifier M_c can be further trained, and upon completion, the subsequent learning process can advance to the next iteration.

2.2 | Initialization Strategy

In the previous subsection, we introduced a general formulation for cold-start AL. Here, we present three different strategies based on foundation models: diversity sampling, uncertainty sampling, and hybrid sampling. Additionally, we discuss random sampling, a common approach that does not require a pretrained model. Figure 1a illustrates the cold-start AL initialization process for binary disease diagnoses. Unlike the general formulation, the three strategies require a foundation model M_f to process images I_i into embeddings E_i which has much lower dimensions than the image dimension of $W_0 \times H_0$. Also, previous literature [57, 71] has demonstrated that E_i can replace I_i as model inputs, enabling the development of a simplified model M_c' with comparable or superior performance, as depicted by the dashed line in Figure 1a.

2.2.1 | Diversity Sampling

The core of AL is to select informative samples, though the precise definition of informativeness remains an open question [9]. Some researchers suggest that an effective strategy, known as diversity sampling, is to select images that are representative of the overall data set while avoiding redundancy from visually similar images [51]. Among various diversity sampling strategies, clustering methods are considered classic approaches [72–75]. These methods have been validated as effective for partitioning chest radiograph data sets into distinct clusters based on image features [76]. The centroids of each cluster are deemed diverse, as they originate from different clusters, and representative, as they serve as the central points of these clusters [77, 78]. We select K-means [79] as the clustering method and follow previous studies that split the same amount of clusters as the annotation budget [27, 80–82]. Formally, the query strategy Q_0 divides $D_0^u = \langle(I_i), i = 1, 2, ..., N_0^u\rangle$ into N_0^l subgroups through E_i^u-based K-means, selects the centroid sample I_j from each

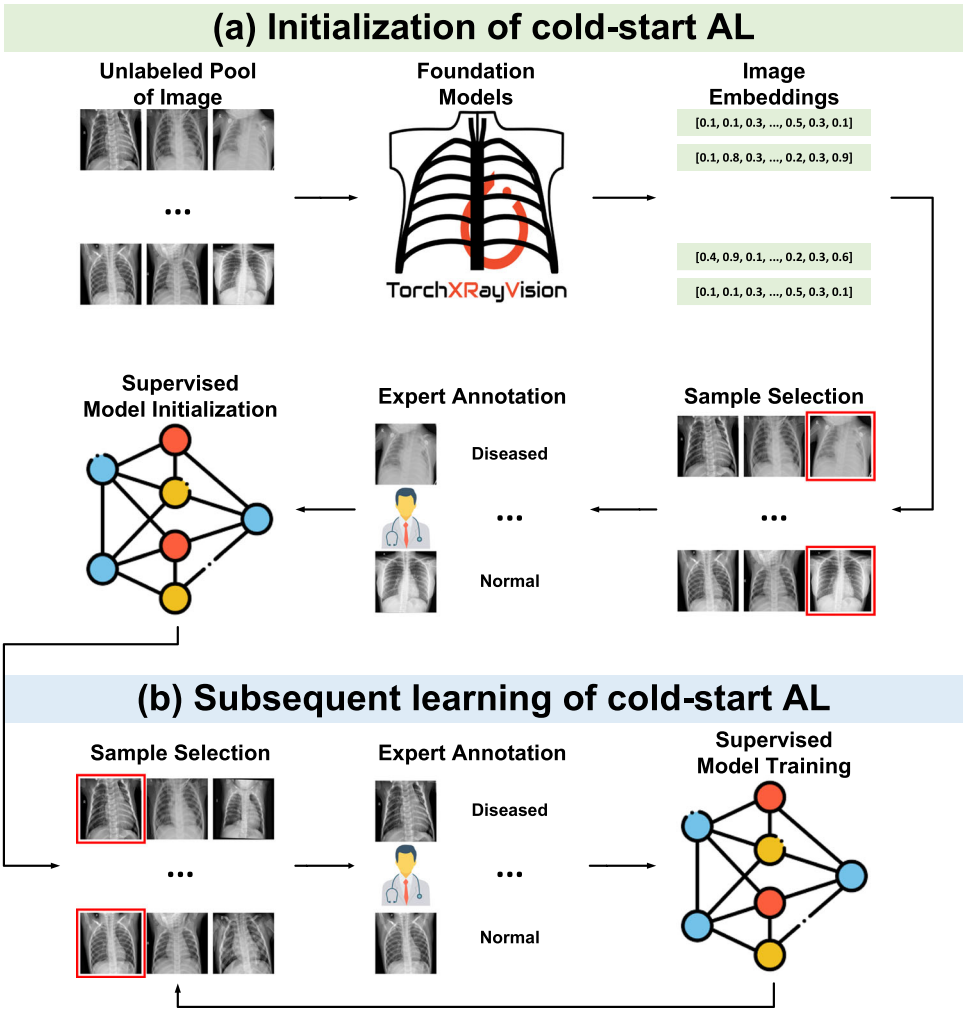

FIGURE 1 | Cold-start AL workflow based on domain-specific foundation models. (a) The initialization stage. (b) The subsequent learning stage. After sample selection and expert annotation, supervised model training can be conducted using either the original image pixels or embeddings derived from foundation models as classifier inputs.

cluster, and sends the selected samples to oracles for annotation to constitute $D_0^l = \langle (I_j, Y_j), j = 1, 2, ..., N_0^l \rangle$.

2.2.2 | Uncertainty Sampling

In contrast to diversity sampling, which aims to select representative samples, an alternative approach, called uncertainty sampling, emphasizes the selection of the most uncertain samples for the current model, positing that these samples contribute most to model convergence. In the initialization stage, sample uncertainty for the target task is unavailable. Nath et al. [83] introduced a proxy task for image segmentation using morphological operations and employed MC dropout to estimate sample uncertainty [52]. However, their approach was restricted to computed tomography. In contrast, we developed a more generalizable auxiliary task based on foundation models to enable the computation of sample uncertainty across a broader range of applications.

Given that domain-specific pretrained representation E_i encapsulates high-density information from the original image I_i [84–86], enabling a range of downstream tasks [87–89], we hypothesize that E_i^u can serve as an auxiliary prediction target. If a model exhibits uncertainty in predicting E_i, this suggests that the mapping between I_i and the lower dimensional E_i is challenging. Consequently, such samples are likely to be difficult for downstream tasks, including the target task. Formally, based on the image-representation pair $\langle (I_i, E_i), i = 1, 2, ..., N_0^u \rangle$, we adopt the same architecture of M_c and modify its final layer to match the dimension of E_i, generating the auxiliary model M_u for uncertainty estimation. When the training of M_u is completed, we follow the previous studies [51, 90–93] and use MC dropout to approximate sample uncertainty [52]. Specifically, the trained model M_u takes I_i as the input and feedforward it T times. In iteration φ, a random dropout pattern is activated with a probability of P, and the model output $M_{u,\varphi}(I_i)$ is recorded. Based on the aggregation set $\langle (M_{u,\varphi}(I_i)), \varphi = 1, 2, ..., T \rangle$, the inference variance is calculated via $(1/T)\sum_{\varphi=1}^{T}(M_{u,\varphi}(I_i) - (1/T)\sum_{\varphi=1}^{T} M_{u,\varphi}(I_i))^2$. A large variance demonstrates that the process of I_i by M_u is either highly sensitive to neuron connectivity altering or akin to random guessing, reflecting significant uncertainty [51, 93]. The I_i with the highest uncertainty, that is, inference variance, are selected and sent to oracles for annotation to constitute $D_0^l = \langle (I_j, Y_j), j = 1, 2, ..., N_0^l \rangle$.

2.2.3 | Hybrid Sampling of Diversity and Uncertainty

Diversity sampling selects representative samples while within a limited budget of AL, it might choose uninformative samples that are easy to distinguish and contribute marginally to model capability [94]. Uncertainty sampling suffers from selecting redundant samples to be labeled due to similar high uncertainty values [95] and a potential improvement could be a hybrid method to identify highly diverse and uncertain samples to convey more information with the same amount of annotated data [96, 97].

We employ a classic two-step hybrid method [81, 98, 99] to first partition $D_0^u = \langle (I_i), i = 1, 2, ..., N_0^u \rangle$ into N_0^l subgroups through E_i^u-based K-means and then from each cluster, selects the most uncertain I_i with the highest inference variance calculated by MC dropout as detailed in the previous subsection. This two-step approach ensures the selection of diverse and uncertain samples to construct D_0^l comprising N_0^l samples along with their annotations.

2.2.4 | Random Sampling

In addition to the three initialization strategies based on domain-specific foundation models, random sampling remains the most widely used and traditional method, as shown in previous studies [6, 7, 10, 27–33, 74]. Although commonly employed, random sampling is not without limitations. For instance, prior research has demonstrated that it does not ensure the informativeness of the initial samples selected for annotation, which may negatively impact downstream AL performance [10]. Moreover, random sampling is prone to issues related to data imbalance, especially during initialization, where selecting minority samples can require a substantial budget [35, 100].

2.3 | Subsequent Learning Strategy

Upon completion of the initial sample selection and annotation, a model with sufficient competency on the target data and task becomes available, and based on the model, we can step into the subsequent learning stage of cold-start AL, as depicted in Figure 1b. Following the initial sample selection and annotation, a discriminative model M_c or M_c' is developed, enabling the next stage of subsequent learning. Although a range of warm-start strategies could be applied at this stage, our study focuses on comparing foundation models with their ImageNet counterparts. We therefore employ classic uncertainty sampling strategies [100], which have demonstrated strong performance in prior studies [78, 101–103]. It is important to note that the uncertainty in this phase, given the availability of a discriminative model, differs from that discussed during the initialization stage, which will be further elaborated in the following paragraph.

In the context of subsequent learning, uncertainty sampling encompasses three primary methods: least confidence sampling, margin of confidence sampling, and entropy-based sampling [30]. Notably, these methods converge on the same conclusion: the most uncertain samples are those for which model predictions $M_c(I_s)$ or $M_c'(E_s)$ approach 0.5 in our experimental settings of binary classification [41]. Formally, we denote the predictive probability of I_s towards the positive class as $M_{c,1}(I_s)$ and the negative class as $M_{c,0}(I_s) = 1 - M_{c,1}(I_s)$.

Least confidence sampling selects the samples whose predictive probabilities $P(I_s)$ of the most probable class are low. In binary classification, the most probable class is either positive or negative. Given that $0 \leq M_{c,1}(I_s) \leq 1$, if $M_{c,1}(I_s) > 0.5$, the most probable class is the positive class, resulting in $P(I_s) = M_{c,1}(I_s) > 0.5$. Conversely, if $M_{c,1}(I_s) < 0.5$, the most probable class is the negative class, and $P(I_s) = 1 - M_{c,1}(I_s) > 0.5$. Therefore, the lowest probability occurs when $M_{c,1}(I_s) = 0.5$.

Margin of confidence sampling identifies the samples with small difference between the first and second most probable classes. In binary classification, the difference is expressed as $|M_{c,1}(I_s) - M_{c,0}(I_s)| = |M_{c,1}(I_s) - 1 + M_{c,1}(I_s)|$. Clearly, the lowest difference is 0, which is achieved when $M_{c,1}(I_s) = 0.5$.

Entropy sampling [28, 50] chooses the samples with the highest entropy sum of predictive probabilities across all classes. The sum is expressed as $-M_{c,1}(I_s)\log(M_{c,1}(I_s)) - M_{c,0}(I_s)\log(M_{c,0}(I_s)) = -M_{c,1}(I_s)\log(M_{c,1}(I_s)) - (1 - M_{c,1}(I_s))\log(1 - M_{c,1}(I_s))$ in binary classification. The first derivative of the sum is $-\log(M_{c,1}(I_s)/(1 - M_{c,1}(I_s)))$, with a stationary point occurring at $M_{c,1}(I_s) = 0.5$. Consequently, the maximum of the sum is attained when $M_{c,1}(I_s) = 0.5$.

In addition to uncertainty sampling strategies, we implement random sampling, a widely used and well-established approach [10, 32, 48, 76, 104, 105], to select samples in the subsequent learning stage.

3 | Experiments

This section begins by presenting a comprehensive overview of the experimental settings, including data sets, AL strategies, DL implementation details, evaluation metrics, and statistical tests. Next, we present the results of various AL strategies applied across previous experimental settings and foundation models developed by supervised and self-supervised learning. Finally, the statistical test results are analyzed to compare foundation models with their ImageNet counterparts, evaluate pixel-based classifiers against representation-based classifiers, examine the relationship between initialization and subsequent learning, and assess the effectiveness of different query strategies: one-shot initialization versus initialization followed by iterative subsequent learning.

3.1 | Experimental Settings

3.1.1 | Data Sets

To ensure the robustness of our experimental results [106], we employed two data sets featuring diverse population cohorts, sample sizes, and disease categories: the Guangzhou data set [107, 108] and the Pakistan data set [109, 110]. The Guangzhou

data set, collected by the Guangzhou Women and Children's Medical Center, comprised 5856 chest radiographs from retrospective cohorts of pediatric patients aged 1–5 years, with 4273 images diagnosed with pneumonia. In contrast, the Pakistan data set was considerably smaller and contained a total of 450 chest radiographs from a local hospital in Pakistan, among which 390 images were diagnosed with COVID-19. Both data sets were collected after the release of foundation models and were specifically chosen to simulate real-world scenarios, enabling an assessment of the benefits these models bring to AL. We resized all radiographs from the two data sets into the resolution of 224×224 to comply with DL classifiers [111], foundation models [59], and ImageNet counterparts [70].

The data sets were split using an 80/20 ratio for the Guangzhou data set, resulting in 4686 images for training (3419 diseased) and 1170 images for testing (854 diseased). For the Pakistan data set, a 50/50 split was applied, yielding 225 images for both training and testing sets, with 195 diseased images in each set. A larger proportion of samples was allocated to the testing set in the Pakistan data set due to its small sample size, ensuring greater stability in testing. We did not create separate validation sets with diagnostic labels, unlike prior cold-start AL studies [48, 112] to replicate real-world cold-start scenarios where no annotated samples were available at the outset [113]. Additionally, all labels were hidden during the initial sample selection and remained inaccessible until chosen by the query strategy in subsequent learning stages, simulating the cold-start AL process [51].

3.1.2 | Foundation Models and ImageNet Counterparts

TorchXRayVision (TXRV) [114] is an open-source library developed for chest radiograph analysis, offering a range of representation learning models trained on 950,778 chest radiographs from 13 data sets collected across diverse regions, including the United States, China, Spain, and Vietnam. These models served as feature extractors (representation providers). For input images with a resolution of 224×224, TXRV utilizes DenseNet-121 [115] as its backbone. Notably, TXRV was trained using fully supervised methods rather than self-supervised approaches. In this study, we used TXRV to compare a domain-specific supervised model with a general supervised model, specifically the ImageNet pretrained DenseNet-121 [116].

Robust and Efficient MEDical Imaging with Self-supervision (REMEDIS) strategy [59] combines supervised pretraining on natural images with contrastive self-supervised pretraining on chest radiographs. Specifically, it employs the ResNet-152 architecture [70] with pretrained weights from BiT-L [117], which were trained on a large-scale database of natural images (JFT-300M) [118]. REMEDIS was then trained using the self-supervised technique of SimCLR [119] on unlabeled medical data sets across five domains: chest radiographs, fundus imaging, digital pathology, mammography, and clinical dermatology. After that, REMEDIS learned generalizable representations that can be paired with a classifier head to map them to domain-specific labels for downstream tasks. REMEDIS has proven particularly effective for chest radiograph classification [120], and therefore we adopted it as a domain-specific self-supervised model for comparison with ImageNet pretrained ResNet-152. For both TXRV and REMEDIS, we utilized the embeddings from the final layer preceding the classification head as model representations for diverse AL strategies and simplified classifiers.

3.1.3 | AL Strategies

For cold-start initialization, we performed five experiments using budgets ranging from 10 to 50, with an incremental step of 10 samples. Diversity sampling employed classic K-means clustering, generating a number of clusters equal to the budget, and selecting the sample closest to each cluster centroid. In uncertainty sampling, estimators were trained using inputs of original images and outputs of representations from foundation models or ImageNet counterparts. Estimators then processed each sample 100 times with a dropout activation probability of 0.5 to stably compute the variance of model predictions [121], selecting the samples with the top variance per the allocated budget. The hybrid method integrated diversity and uncertainty strategies by generating a number of clusters equal to the budget and then selecting the sample with the highest variance from each cluster to form the initialization set. Random sampling was simulated 100 times for each initialization budget. For each simulation, the high-budget group included all samples from the low-budget group to ensure comparability in downstream analyses.

For subsequent learning, we allocated an initial budget of 10 and a subsequent learning budget of 40 to enable a direct comparison between one-shot initialization and the full AL strategy with both initialization and subsequent learning. The subsequent learning stage consisted of 4 iterations, each with a budget of 10. Due to the convergence of the three uncertainty strategies, 10 samples with predictive probabilities closest to 0.5 were selected based on the current classifier in each iteration. Random sampling, similar to the initialization, was benchmarked 100 times for comparison. In each iteration of the subsequent learning process, 10 samples were randomly selected from the unlabeled set.

For both initialization and subsequent learning stages, we established the same upper bound of model performance by training classifiers using all available training samples and their expert annotations. This budget was referred to as the "all samples."

3.1.4 | Implementation Details

Two main categories of DL models were developed in this study: one for binary classification tasks and another for uncertainty estimation during the initialization phase. We implemented binary classifiers based on VGG-11 [111] to distinguish either pneumonia or COVID-19 in the Guangzhou data set and the Pakistan data set, respectively. VGG-11 architecture was selected for its extensive use in AL studies [7] and its reliable convergence on small-sample data sets [122–124]. In addition to the full VGG-11 architecture, which used original images as inputs, we also developed simplified models based on previous studies [57, 125]. Specifically, we implemented three-layer multilayer perceptron

(MLP-3) models with intermediate layers of 512 and 256 neurons, using representations generated by foundation models as inputs [60]. The second category of DL models focused on uncertainty estimation, using the same VGG-11 architecture but with output layers modified to match the dimensionality of the target representations from foundation models.

For the training of binary classifiers, we utilized a Stochastic Gradient Descent (SGD) optimizer [126] with a learning rate of 1e−3 and a momentum of 0.9. The batch size was fixed at 10, matching the initialization budget. To mitigate the imbalance between major and minor samples [127, 128], a weighted cross-entropy loss function was applied. Training was conducted for 200 epochs, with a linear scheduler that reduced the learning rate by a factor of 0.5 if no improvement was observed over 10 consecutive epochs. An early stopping criterion was employed if no progress was made over 20 consecutive epochs. For the training of uncertainty estimators, all configurations were kept constant, except for the learning rate, which was adjusted to 1e−4, and the loss function, which was changed to mean squared error to align with the predictive targets of image representations in the continuous space.

Upon the completion of sample annotation queries and binary classifiers' training, the classification performance was assessed on the hold-out test sets from both the Guangzhou and Pakistan data sets. Following the previous literature [129, 130], the evaluation utilized the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC) due to their reliability in scenarios involving imbalanced data [131, 132]. Metrics such as accuracy, sensitivity, specificity, positive predictive value, and negative predictive value were excluded due to their vulnerability to instability in the presence of extreme data imbalances [133, 134]. Standard deviations for each metric were estimated using the nonparametric bootstrap method [135]. The study was conducted in PyTorch 1.12.1 and the code has been open access [136] for reproducibility. All experiments were implemented on a Dell Precision 7920 Tower Workstation with an Intel Xeon Silver 4210 CPU and an NVIDIA GeForce RTX 2080 Super GPU.

3.1.5 | Statistical Tests

We conducted statistical tests to assess whether significant differences exist in cold-start AL performance across various configurations [35]. Our first inquiry sought to determine whether domain-specific foundation models outperform their ImageNet pretrained counterparts. We also investigated whether simplified classifiers based on feature representations could surpass more complex classifiers relying on original image pixels. Additionally, we aimed to establish whether effective initialization contributes to enhanced subsequent learning. Finally, we examined whether a one-shot initialization can achieve performance comparable to the complete cold-start AL process, which includes initialization and multiple iterations of subsequent learning; this approach offers greater ease of implementation and user-friendliness [137, 138]. For the analysis of the relationship between initialization and subsequent learning, we employed the Pearson correlation coefficient [139]. The same correlation test was performed to evaluate the influence of class balance in the initialization samples on model performance in both initialization and subsequent learning. In addressing the other three questions, we utilized the paired *t*-test [140] to compare the performance of the two competing approaches.

3.2 | Results

3.2.1 | Cold-Start Initialization

Figures 2 and 3 illustrate the AUROC and AUPRC performance of various strategies and model backbones during the cold-start AL initialization on the Guangzhou and Pakistan data sets, respectively. In the odd-numbered columns, curve plots depict the mean values of AUROC and AUPRC, whereas the even-numbered columns present bar plots displaying their standard deviations, calculated via nonparametric bootstrap. The first row in both figures displays two baseline query strategies: all samples and random sampling. The horizontal dashed lines in Figures 2a,c and 3a,c represent the upper bound of classification performance, achieved by training on the full set of samples and annotations. Random sampling was the most common practice in cold-start initialization, and we compared this method with initialization strategies based on foundation models and ImageNet pretrained counterparts.

The subplots in the second, third, and fourth rows present model performance based on samples selected by diversity, uncertainty, and hybrid sampling, respectively, each applied to the four representation generation models. Using the samples queried by these diverse strategies, we developed both a full VGG-11 model and a simplified MLP-3. The MLP-3, based on representations, consistently underperformed the VGG-11 model trained on original pixels, suggesting that low-dimensional representations derived from foundation models and ImageNet counterparts may lose critical information embedded in the original images. For the VGG-11 classifiers, representation-based sampling outperformed random sampling in 47 out of 60 scenarios for the Guangzhou data set and 46 out of 60 for the Pakistan data set, indicating that representation-based strategies reduced annotation requirements while providing superior initializations. Among the three representation-based strategies, diversity and hybrid sampling achieved the best performance in 16 out of 20 and 14 out of 20 scenarios for the Guangzhou and Pakistan data sets, respectively. This suggested that for data sets with small sample sizes, such as the Pakistan data set, the hybrid method that incorporated both diversity and uncertainty may be preferred. In contrast, for larger data sets, such as the Guangzhou data set, diversity sampling remained competitive. Additionally, across all strategies, as the initialization budget increased, the standard deviation of model performance decreased, demonstrating that a larger sample size not only improved model accuracy but also enhanced prediction robustness. For detailed numeric results, see Tables A1 and A2.

3.2.2 | Cold-Start Subsequent Learning

Based on the classifiers trained on 10 samples selected by different initialization strategies, subsequent learning was performed using uncertainty-based iterations [23], with 10 samples

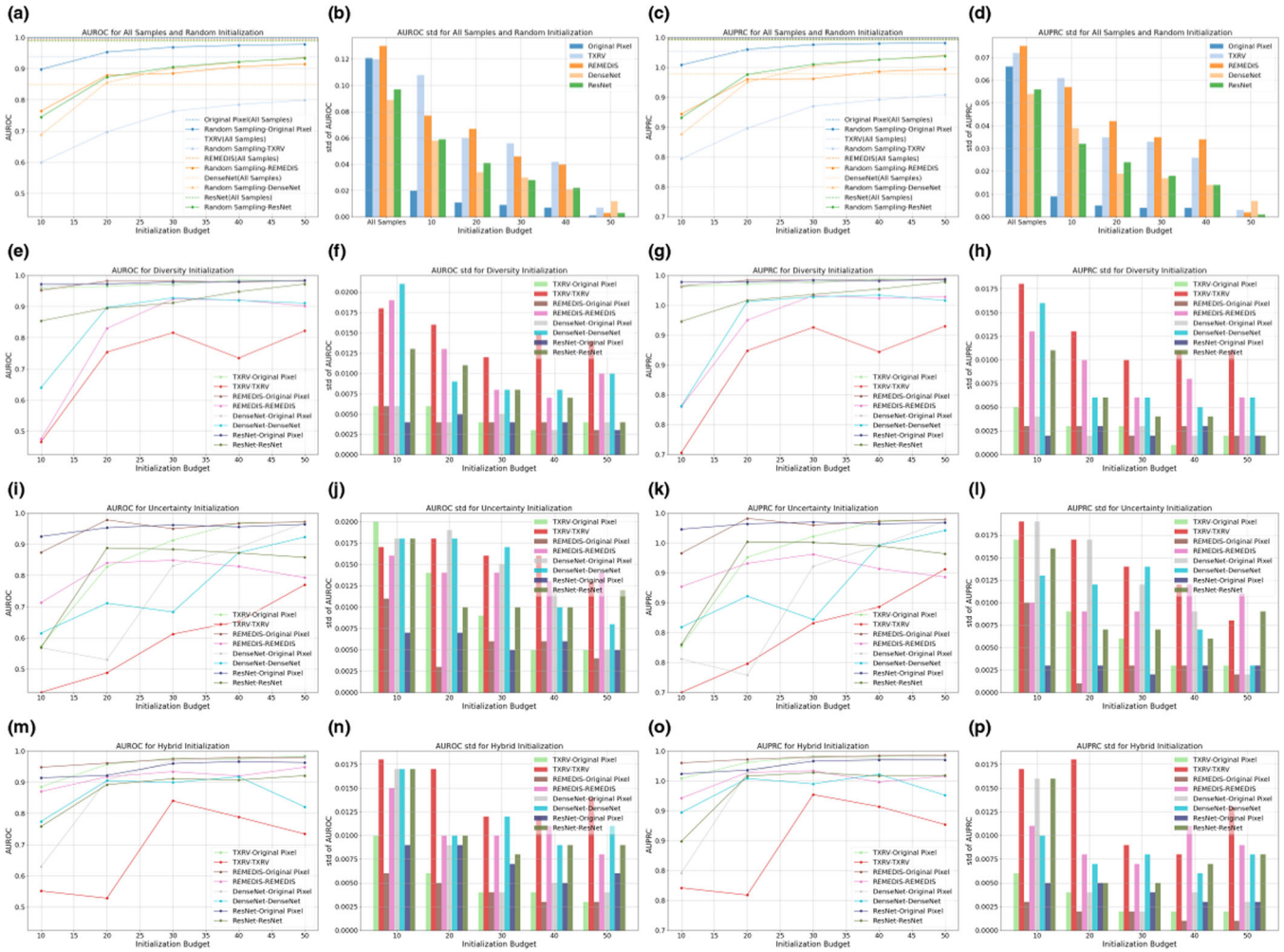

FIGURE 2 | One-shot initialization performance on the Guangzhou data set. Subgraphs (a–p) present specific information detailed in the following explanations. The first, second, third, and last columns present the mean values of AUROC, the standard deviation of AUROC, the mean values of AUPRC, and the standard deviation of AUPRC, respectively. The first row displays the results for all samples with annotations, the upper bound, and random sampling. The second, third, and final rows show the outcomes for diversity sampling, uncertainty sampling, and hybrid sampling, respectively.

queried per iteration. Figures 4 and 5 illustrate the classification performance on the Guangzhou and Pakistan data sets, respectively. The arrangement of subplots in Figures 4 and 5 mirrors that of Figures 2 and 3, with the following distinctions: (1) the learning strategy was based solely on classifier uncertainty, and the legend in each subplot indicates the initialization strategy; (2) samples selected under high-budget conditions included all samples from low budgets as they were consecutive procedures, which was not guaranteed in the initialization stage; and (3) the $\boldsymbol{X}$-axis representing the overall budget included both the initialization and subsequent learning phases: For example, a budget of $10 + 20$ denoted an initialization budget of 10 samples, followed by an additional budget of 20 samples for the subsequent learning.

As annotation budgets increased, VGG-11 performance initially improved over multiple iterations before quickly converging, with additional samples yielding only marginal gains. This phenomenon was attributed to the relative simplicity of the two binary classification tasks compared to more complex clinical tasks, such as low-contrast lesion segmentation [141–143], as demonstrated by the strong classification performance before subsequent learning. Although some performance improvement to the upper bound in Figures 4a,c and 5a,c remained possible, achieving this would require approximately 100 times more annotations for the Guangzhou data set and 10 times more for the Pakistan data set. This underscored the effectiveness of AL in balancing annotation costs with model performance. Compared with VGG-11 models, the representation-based MLP-3 classifiers were inferior, aligning with previous initialization results. Additionally, MLP-3 classifiers exhibited the risk of overfitting in Figures 4e,g and 5i,k,m,o: their predictive performance declined when more samples were added to the labeled training set. For detailed numeric results, see Tables A3 and A4.

Models initialized with diversity and hybrid sampling consistently outperformed uncertainty sampling, achieving the highest performance in 8 and 7 out of 16 learning scenarios for the Guangzhou data set, and 8 and 5 out of 16 scenarios for the Pakistan data set. This consistent outperformance of diversity and hybrid sampling highlighted the benefit of effective initialization for both the initialization and subsequent learning stages. In the next subsection, we will extend these observational findings with

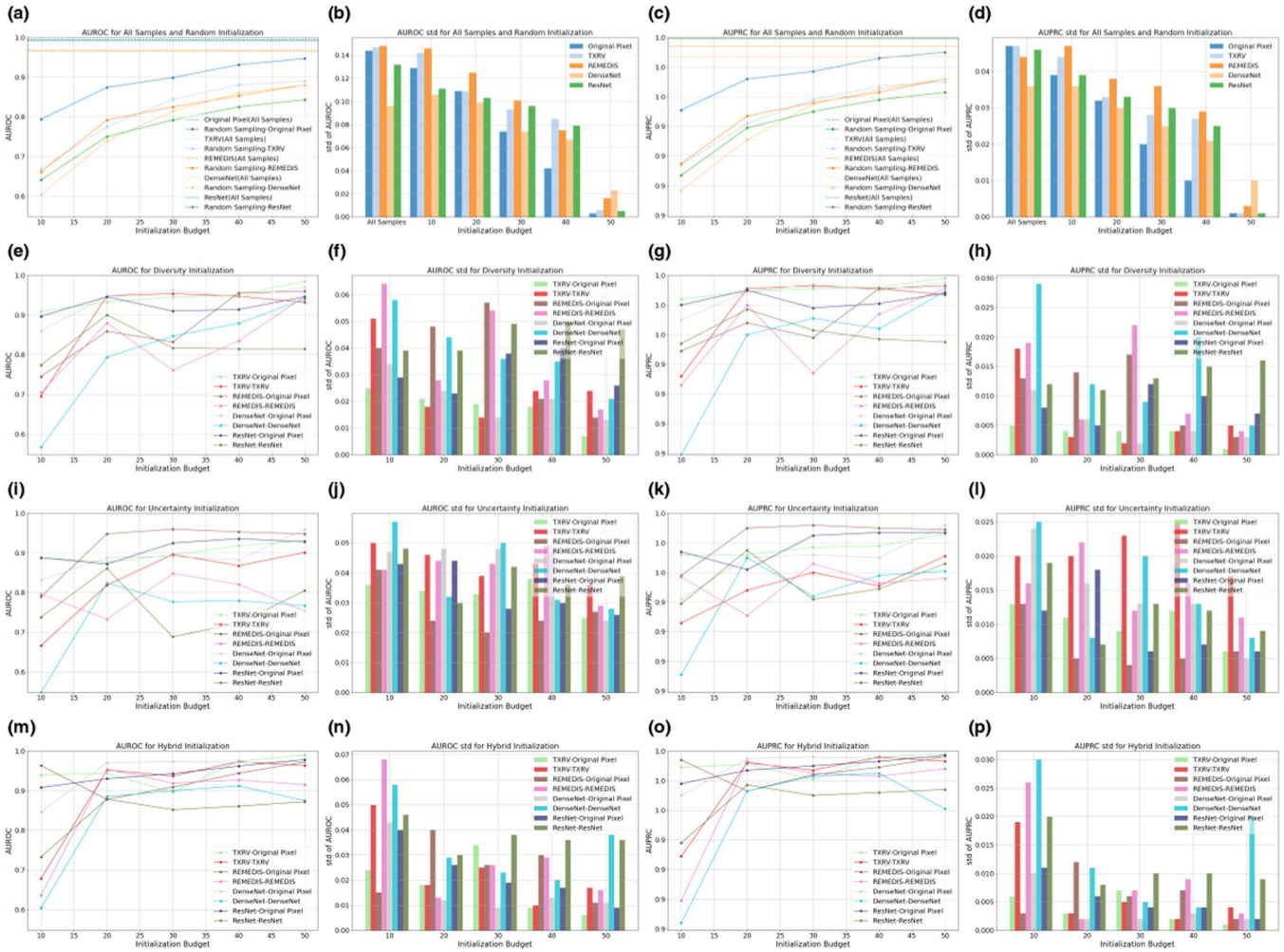

FIGURE 3 | One-shot initialization performance on the Pakistan data set. Subgraphs (a–p) present specific information detailed in the following explanations. The first, second, third, and last columns present the mean values of AUROC, the standard deviation of AUROC, the mean values of AUPRC, and the standard deviation of AUPRC, respectively. The first row displays the results for all samples with annotations, the upper bound, and random sampling. The second, third, and final rows show the outcomes for diversity sampling, uncertainty sampling, and hybrid sampling, respectively.

rigorous statistical tests. Specifically, we would compare representations from foundation models and ImageNet counterparts, evaluate different classification backbones, explore the relationship between initialization and subsequent learning, and contrast one-shot initialization with the complete AL process.

3.2.3 | Comparative Study on Classification Backbones

The primary question addressed in this study was whether foundation models designed for chest radiograph analysis outperformed their ImageNet counterparts pretrained on images from the natural domain. We conducted a paired *t*-test to compare their performance during both the initialization and subsequent learning stages using the Guangzhou and Pakistan data sets. The null hypothesis posited that the performance of foundation models and ImageNet counterparts would be statistically identical, whereas the alternative hypothesis suggested that the performance of foundation models is superior to that of ImageNet counterparts. As shown in Table 1, foundation models outperformed their ImageNet counterparts in only two out of eight experiments. Consequently, in the context of cold-start AL, foundation models failed to meet our expectations as generalist models.

Another objective of the foundation model was to generate representations that could be directly utilized as input features, thereby facilitating lightweight classification backbones, such as MLP, to achieve high-fidelity predictions with reduced computational costs [57, 59, 114]. To assess this, we compared the performance of VGG-11 with that of the lightweight MLP-3 [57, 125]. The null hypothesis posited that the performance of MLP-3 using generated representations was equivalent to that of VGG-11 using original pixel data, whereas the alternative hypothesis proposed that the performance of MLP-3 was inferior to that of VGG-11. Table 2 illustrates that MLP-3 statistically significantly underperformed VGG-11 in seven out of eight scenarios.

We identified that representation-based strategies outperformed default random sampling during the initialization stage. However, the extent to which these benefits extend to subsequent learning stages remained inadequately explored. To investigate this, we calculated the Pearson correlation coefficient of the AUROC and the AUPRC between model performance in

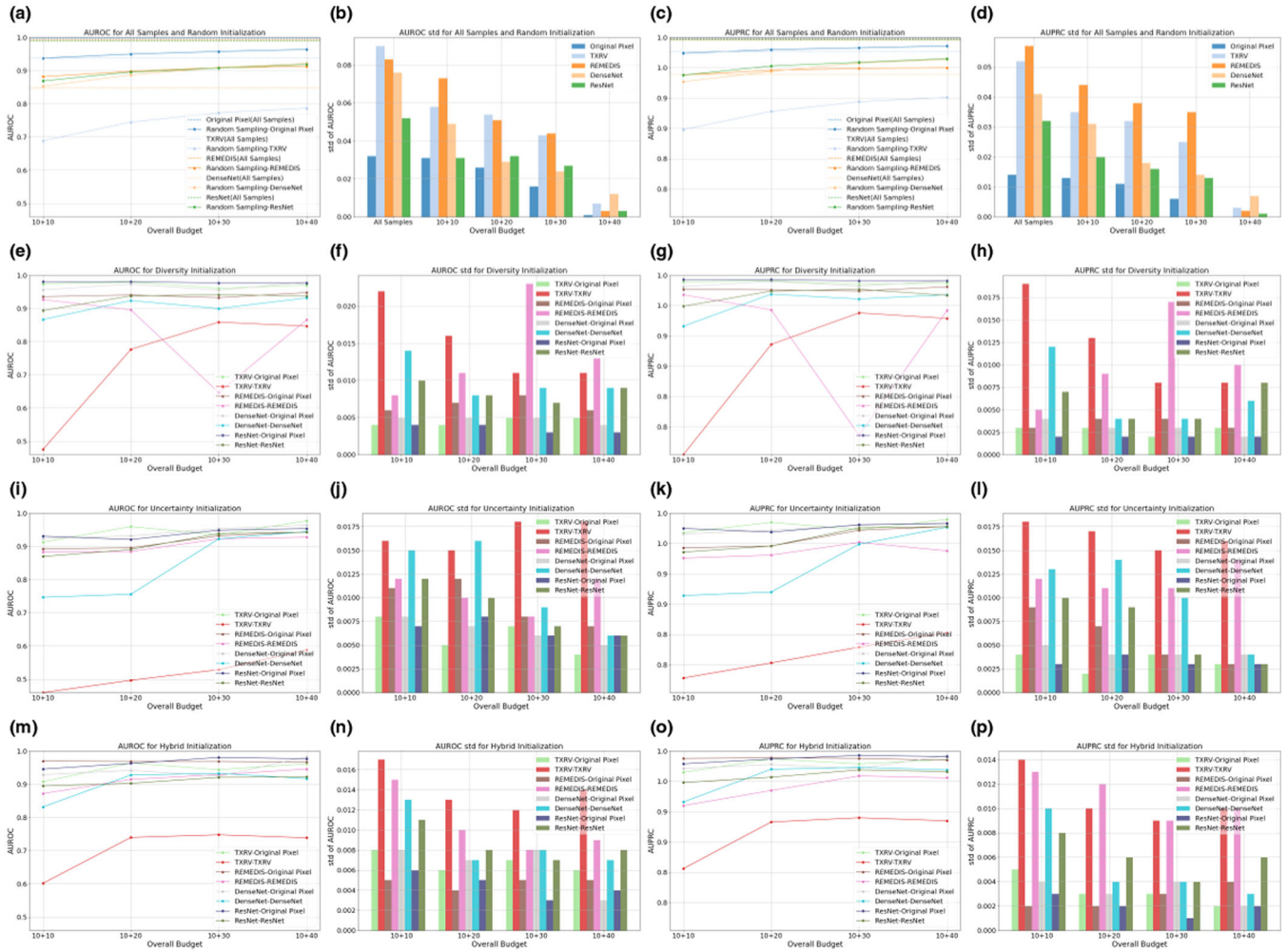

FIGURE 4 | Subsequent learning performance on the Guangzhou data set. Subgraphs (a–p) present specific information detailed in the following explanations. The first, second, third, and last columns present the mean values of AUROC, the standard deviation of AUROC, the mean values of AUPRC, and the standard deviation of AUPRC, respectively. The first row displays the results for all samples with annotations, the upper bound, and random sampling. The second, third, and final rows show the outcomes for diversity sampling, uncertainty sampling, and hybrid sampling, respectively.

the initialization stage and the subsequent learning stage. Our null hypothesis was that the correlation coefficient between the performance of cold-start initialization and subsequent learning did not significantly deviate from zero, whereas the alternative hypothesis asserted that this correlation was significantly greater than zero. As shown in Table 3, model performance during the initialization stage was positively correlated with performance in the subsequent learning stage, suggesting that researchers should pay more attention to effective initialization strategies instead of using random sampling as a default [144].

Another question we sought to address was whether one-shot initialization identified samples capable of training models with performance comparable to those selected through both initialization and iterative learning stages. Consistent with the first and second statistical tests, we conducted a paired *t*-test between the two approaches using an equivalent overall budget. The null hypothesis posited that the average performance of classifiers utilizing one-shot initialization was identical to that of classifiers employing a full AL cycle of both initialization and subsequent learning. Conversely, the alternative hypothesis asserted that the performance of classifiers using one-shot initialization was inferior to that of classifiers employing the integrated approach. As presented in Table 4, all *p*-values exceeded 0.05, indicating that one-shot initialization was comparable to the complete AL cycle in the medical task of chest radiograph classification.

Finally, we investigated whether model performance was influenced by the class balance of the initialization samples, specifically testing the hypothesis that a balanced class distribution could enhance performance. The null hypothesis posited no significant correlation deviating from zero between the minority class proportion in the initialization samples and the classifier's performance during cold-start initialization or subsequent learning. In contrast, the alternative hypothesis suggested that this correlation was significantly greater than zero. As shown in Table 5, no statistically significant correlation was observed in both stages. Interestingly, a U-shaped trend in *p*-values was observed during both stages of the Guangzhou data set and the cold-start initialization of the Pakistan data set, indicating that the class balance was more strongly correlated with performance at intermediate budget levels compared to low or high budgets.

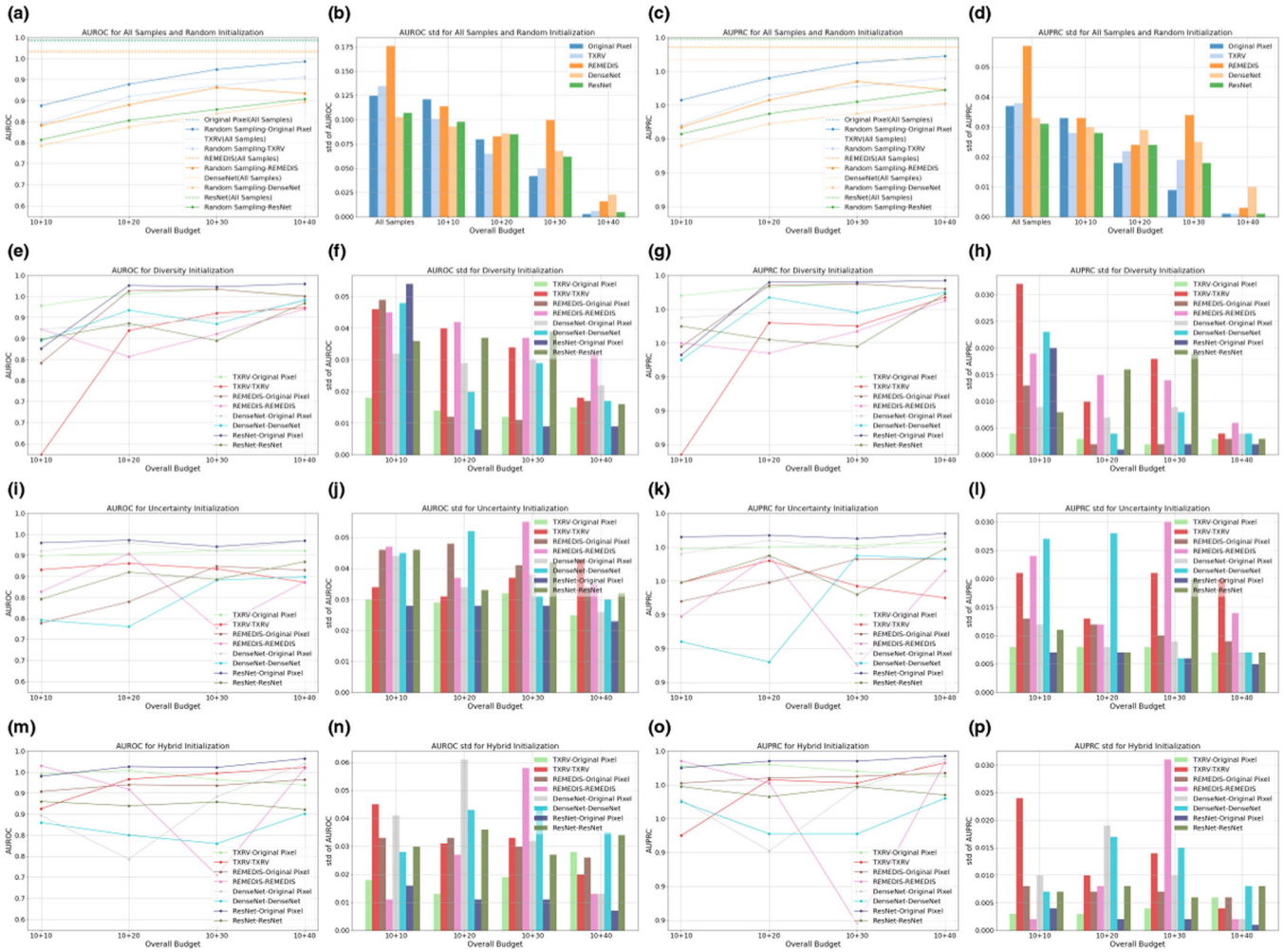

FIGURE 5 | Subsequent learning performance on the Pakistan data set. Subgraphs (a–p) present specific information detailed in the following explanations. The first, second, third, and last columns present the mean values of AUROC, the standard deviation of AUROC, the mean values of AUPRC, and the standard deviation of AUPRC, respectively. The first row displays the results for all samples with annotations, the upper bound, and random sampling. The second, third, and final rows show the outcomes for diversity sampling, uncertainty sampling, and hybrid sampling, respectively.

TABLE 1 | Paired *t*-test between classifiers initialized based on foundation models and ImageNet pretrained counterparts.

Data set	Foundation model	AL stage	*p*-value of AUROC	*p*-value of AUPRC
Guangzhou data set	TXRV	Cold-start initialization	2.74e−2*	2.89e−2*
		Subsequent learning	3.10e−1	1.57e−1
	REMEDIS	Cold-start initialization	1.59e−1	2.94e−1
		Subsequent learning	9.96e−1	9.97e−1
Pakistan data set	TXRV	Cold-start initialization	2.59e−1	9.34e−2
		Subsequent learning	4.80e−2*	2.99e−2*
	REMEDIS	Cold-start initialization	8.47e−1	7.54e−1
		Subsequent learning	9.98e−1	9.96e−1

*The *p*-value is less than 0.05, demonstrating statistical significance at a confidence level of 95%.

4 | Discussion

In this study, we conducted a quantitative analysis to evaluate the performance of domain-specific pretrained models compared to their ImageNet counterparts during both the initialization and learning stages of cold-start AL. Unlike foundation models in natural language processing [145–147], our findings reveal a notable disparity in the efficiency of pretrained models within the domain of medical imaging [148]. In most experiments, models pretrained on chest radiographs, whether through supervised or self-supervised learning, did not surpass those pretrained on natural images in selecting informative

TABLE 2 | Paired *t*-test between MLP-3 using representations and VGG-11 using original pixels in cold-start initialization.

Data set	Representation source	AL stage	*p*-value of AUROC	*p*-value of AUPRC
Guangzhou data set	TXRV	Cold-start initialization	3.62e−10*	3.52e−9*
		Subsequent learning	2.30e−7*	8.49e−7*
	REMEDIS	Cold-start initialization	3.35e−5*	1.81e−6*
		Subsequent learning	2.35e−3*	1.83e−3*
Pakistan data set	TXRV	Cold-start initialization	1.98e−3*	1.88e−4*
		Subsequent learning	9.99e−3*	1.33e−3*
	REMEDIS	Cold-start initialization	3.83e−4*	6.29e−4*
		Subsequent learning	7.91e−2	1.42e−2*

*The *p*-value is less than 0.05, demonstrating statistical significance at a confidence level of 95%.

TABLE 3 | Pearson correlation coefficient between cold-start initialization and subsequent learning. Classifiers were developed using the same architecture of VGG-11 and original pixels.

Data set	Overall budget	*p*-value of AUROC	*p*-value of AUPRC
Guangzhou data set	10 + 10	2.69e−5*	2.13e−5*
	10 + 20	4.03e−5*	8.33e−5*
	10 + 30	2.56e−5*	7.98e−4*
	10 + 40	2.05e−4*	5.52e−4*
Pakistan data set	10 + 10	3.66e−3*	4.92e−3*
	10 + 20	1.19e−3*	1.82e−3*
	10 + 30	1.23e−3*	6.84e−3*
	10 + 40	2.89e−3*	7.61e−3*

*The *p*-value is less than 0.05, demonstrating statistical significance at a confidence level of 95%.

samples for cold-start AL. Also, the representation did not improve the performance of a simplified model based on MLP architectures, contrary to researchers' expectations that it would surpass the performance of a more complex model using original images as inputs. Additionally, the class balance of initialization samples did not consistently exhibit a positive correlation with model performance across varying budgets in AL initialization and subsequent learning.

The relative inefficiency of domain-specific pretrained models compared to ImageNet-trained models can be attributed to several factors. In general, when domain-specific models are trained on a limited number of samples, their generalization capabilities are often inferior to those of ImageNet-trained models, primarily due to differences in sample and class diversity [149]. However, in our experiments, both domain-specific models were trained on data sets of comparable size to ImageNet. Beyond data scale, model architecture also influences the representation learning capacity of pretrained models [150, 151]. In this study, we standardized the architecture across domain-specific and ImageNet-trained models, ensuring that model architecture did not influence the comparison outcomes. We hypothesize that the inefficiency raised because the latent features in the two data sets may not be fundamentally complex, as evidenced by the rapid model convergence with only a few annotated samples. Thus, despite chest radiographs being visually distinct from general domain images, the low- to mid-level features learned from ImageNet appear sufficient for effectively discriminating between different images in this context [152–154].

A recent study by Huix et al. [145] also reports similar experimental results. They evaluated five vision transformer-based foundation models: SAM [155], SEEM [156], DINOv2 [157], CLIP [158], and BLIP [159], across four well-established medical imaging data sets. All five models employ transformer-based architectures, allowing for direct comparison with the baseline ImageNet pretrained vision transformer. The results revealed that only one model, DINOv2, consistently outperformed the ImageNet pretrained counterpart in four comparative experimental configurations, including whether a linear head or a complex DeiT [160] was used on top of the foundation models, and whether the foundation model parameters were frozen or not. Similar to our findings, the frozen foundation models with a linear head performed worse than those with DeiT, a more complex architecture. Interestingly, when the foundation model parameters were fine-tuned using target data, the linear head outperformed the transformer, a finding that merits further investigation. Although their work focused on comparing general domain foundation models with ImageNet-trained models in diverse medical imaging tasks, our study addresses a gap by further assessing whether models specifically designed for chest radiograph analysis can outperform ImageNet-trained counterparts in tasks within the target domains.

This study also uncovered inspiring findings. First, compared to the commonly used random sampling strategy, which has demonstrated decent performance in prior work [74, 161, 162], both chest radiograph pretrained models and their ImageNet counterparts led to improved performance. This suggests that representation-based initialization may be a superior alternative to random sampling for future AL applications, potentially achieving classifier performance comparable to models trained on fully annotated data sets [141, 142]. The advantages of a robust initialization were further supported by a statistically significant positive correlation between initial model performance and subsequent learning outcomes.

Second, we found that one-shot initialization performed on par with complete AL across both the Guangzhou and Pakistan

TABLE 4 | Paired *t*-test between classifiers developed using one-shot initialization and complete AL cycle.

Data set	Overall budget: Initialization-only	Overall budget: Initialization + subsequent learning	*p*-value of AUROC	*p*-value of AUPRC
Guangzhou data set	20	10 + 10	4.38e−1	5.48e−1
	30	10 + 20	8.04e−1	7.99e−1
	40	10 + 30	8.08e−1	7.81e−1
	50	10 + 40	5.41e−1	5.74e−1
Pakistan data set	20	10 + 10	9.74e−1	9.84e−1
	30	10 + 20	4.56e−1	5.15e−1
	40	10 + 30	8.94e−1	9.36e−1
	50	10 + 40	5.38e−1	4.58e−1

TABLE 5 | Pearson correlation coefficient between minority class proportion and model performance in different AL stages. Classifiers were developed using the same architecture of VGG-11 and original pixels.

Data set	AL stage	Overall budget	*p*-value of AUROC	*p*-value of AUPRC
Guangzhou data set	Cold-start initialization	10	0.283	0.234
		20	0.099	0.094
		30	0.056	0.052
		40	0.063	0.055
		50	0.453	0.530
	Subsequent learning	10 + 10	0.240	0.102
		10 + 20	0.089	0.057
		10 + 30	0.079	0.069
		10 + 40	0.300	0.316
Pakistan data set	Cold-start initialization	10	0.196	0.535
		20	0.297	0.506
		30	0.131	0.168
		40	0.816	0.813
		50	0.793	0.753
	Subsequent learning	10 + 10	0.080	0.098
		10 + 20	0.186	0.210
		10 + 30	0.667	0.658
		10 + 40	0.809	0.867

data sets. This approach alleviates the need for repeated experts' involvement during AL iterations, enabling continuous training of DL models without delays caused by awaiting new sample annotations [137, 138]. Similar one-shot initialization strategies, such as representative annotation, have also been explored in recent studies [72]. Specifically, it has two components to select samples: the first component uses autoencoder [163], variational autoencoder [164], or generative adversarial networks [165] to learn efficient data representation in an unsupervised manner. Based on these clustering-friendly representations, the second component uses agglomerative clustering and applies the greedy max-cover strategy to select images from each cluster. In 2D gland segmentation, the one-shot initialization method demonstrated performance comparable to state-of-the-art iterative approaches while remarkably reducing experts' waiting times. This time-saving advantage was even more pronounced in 3D segmentation of myocardium and great vessels. Jin et al. [138] proposed a one-shot AL method that integrates contrastive learning with diversity sampling. Their approach demonstrated superior performance compared to random sampling and two iterative AL strategies of Bayesian sample query [166] and core-set [167] in skin lesion segmentation, remote sensing image segmentation, and chest x-ray segmentation. The two preceding one-shot AL methods rely on informative representations, highlighting the potential of exploring domain-specific foundation models as representation providers.

Third, although no statistically significant correlation was observed between the sample balance ratio and model performance, a U-shaped trend in *p*-values suggests that the class balance

is more strongly associated with performance at intermediate budget levels than at low or high budgets. We propose that this phenomenon arises because, in low-budget scenarios, the data set's balance ratio exerts minimal influence on model performance, as the limited number of training samples constrains the achievable upper bound of model performance. Conversely, in high-budget scenarios, the abundant training samples ensure the lower bound of model performance, thereby limiting the observable impact of the balance ratio. In both cases, model performance is confined within a relatively narrow range, making it challenging to detect significant correlations.

Our study has limitations that warrant future investigation. First, this study exclusively examined the use of supervised learning based on labeled samples. Future researchers may explore the augmentation of labeled samples [168, 169] or training strategy of semisupervision [54, 112] or metalearning [170, 171] to further upgrade model performance without additional annotation burden [29]. From the data augmentation perspective, Shi et al. proposed to stitch four intraclass images together and resize them to the same size as the original image to unleash the potential value of limited annotated samples [31]. From both empirical improvements in AL performance and theoretical distribution similarity in high-level semantic space, they validated the positive impact of data augmentation towards AL [172]. Beyond aggregation of existing samples in the pixel space [173], Mahapatra et al. employed generative adversarial networks [165] to synthesize realistic chest radiographs from a limited set of anatomy annotations [174]. By incorporating these generated samples and associated annotations into the training set, they achieved a substantial improvement in model accuracy. From the semi-supervised view, Bai et al. proposed to combine expert-annotating labels with model-predicting pseudo labels to boost model performance [175]. To eliminate the training instability caused by pseudo labels, they designed a noise filter to filter pseudo labels with low fidelity, avoiding the improvement brought by informative pseudo labels being impaired by noisy ones [176]. Metalearning is another direction to improve DL performance using auxiliary tasks to generate a robust model that converges to the target task with minimal labeled samples [177]. Yuan et al. designed a training strategy that combines metalearning with AL, including two phases where the first phase aims to pretrain a metalearner possessing sensitive perceptron on the target data domain and the second phase is to select samples with the highest uncertainty on the target task [90]. The main difference between the foundation model and metamodel is that the generation mechanism is based on self-supervised learning or auxiliary tasks-based supervised learning and the integration of these two techniques has been investigated by recent studies [178, 179].

Second, our pipeline for cold-start AL was designed with modular components, and substituting the current techniques with alternative methods would enhance the credibility of our current findings. Foundation models can be replaced with momentum contrast for chest x-rays [180, 181] or in-house developed self-supervised models [72, 182]. Sampling strategies can also be extended to advanced techniques. For example, the current diversity sampling used K-means as the backbone and designated the sample closest to cluster centers as the representative one. Moving forward, we will include refined metrics of representativeness such as information density, which calculates the similarity between embeddings of a particular sample and others within the same cluster [23]. K-means can also be substituted with alternative methods such as BIRCH [183], which empirical evidence suggests is more robust against noisy data and imbalanced labels [23]. Similarly, the hybrid sampling strategy follows a static combination of representativeness and uncertainty while some dynamic reweighting combinations may achieve superior performance [27]. Furthermore, the static strategies can be enhanced with reinforcement learning policies such as multiarmed bandit [184] or actor-critic method [30] to actively switch different sampling strategies based on the state of classifiers and the current environment [185]. Additionally, we evaluated the performance of cold-start AL within a limited set of configurations, comprising one DL backbone, one imaging modality, two close-set binary classification targets [186], and no consideration of real labeling time. Future endeavors may encompass alternative DL backbones including vision transformer [187], additional imaging modalities such as positron emission tomography [188], different targets like open-set classification [28] or lesion segmentation, and comparison at both levels of sample numbers and overall annotation time [32] for a more thorough comparison [189, 190]. These comprehensive experiments would further substantiate the findings of this study regarding the application of foundation models in AL.

Last, our study represents an initial attempt to leverage domain-specific foundation models in AL and highlights promising avenues for future research in both foundation models and AL strategies. For foundation models, they did not exhibit superior generalization capabilities compared to general pretrained models, highlighting the need for further refinement to achieve their intended objective of versatile performance across diverse tasks [191]. Future research could explore the integration of medical knowledge or the adoption of a federated learning framework [192–194] to construct substantially larger training data sets, thereby enhancing model performance in accordance with scaling laws [87, 195]. For AL strategies, foundation models can serve as providers of representations. A low-hanging fruit is to integrate foundation models with strategies that require representations at specific stages of AL or to directly replace DL backbones in target tasks with foundation models, thereby exploring whether AL performance can be enhanced. A more ambitious direction is to exploit the potential of foundation models across different modalities for joint AL. On the one hand, joint AL can involve multiple modalities within medical imaging, such as magnetic resonance imaging, computed tomography, and positron emission tomography [196–198]. On the other hand, it can involve combining a medical imaging modality with another modality, such as chest radiographs and radiological reports [199–203].

5 | Conclusion

Pretraining has been a cornerstone of DL-based chest radiograph analysis, yet it remains unresolved whether domain-specific pretraining outperforms general domain pretraining in the context of cold-start AL. In this study, we demonstrated the inefficiency of domain-specific foundation models compared to general pretrained ImageNet models for two binary classification tasks. Despite this, initialization methods based on both models

significantly outperformed random sampling, the default method for cold-start AL initialization. Furthermore, we uncovered a positive correlation between different stages of cold-start AL and found comparable performance between one-shot initialization and full AL processes. In addition, the influence of class balance in the initialization samples on subsequent learning outcomes warrants careful consideration, particularly in middle-budget scenarios. We anticipate that this study will inspire researchers to enhance pretraining for generalist medical artificial intelligence and explore novel AL methods based on various pretrained models.

Author Contributions

Han Yuan: conceptualization (lead), data curation (lead), formal analysis (lead), investigation (lead), methodology (lead), software (lead), validation (lead), visualization (lead), writing – original draft (lead), writing – review and editing (lead). **Mingcheng Zhu:** formal analysis (supporting), visualization (lead), writing – review and editing (supporting). **Rui Yang:** formal analysis (supporting), validation (lead), writing – review and editing (supporting). **Han Liu:** investigation (supporting), methodology (supporting), writing – review and editing (supporting). **Irene Li:** formal analysis (supporting), resources (lead), writing – review and editing (supporting). **Chuan Hong:** formal analysis (supporting), investigation (supporting), methodology (supporting), project administration (lead), resources (lead), writing – review and editing (supporting).

Acknowledgments

We would like to acknowledge Prof. Nan Liu at Duke-NUS Medical School for his invaluable support.

Declaration of Generative AI in Scientific Writing

During the revision of the initial draft, Han Yuan used GPT-4o mini to check grammar. After using this tool, Han Yuan and other authors reviewed and edited the content as needed. Han Yuan takes full responsibility for the content of the publication.

Ethics Statement

Ethics approval was not required for this study as it utilized retrospective data sets that are publicly accessible. Researchers seeking access to the original data should request permission from the data owners and comply with their established protocols on data privacy and confidentiality.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data sets used in this study are available on https://data.mendeley.com/datasets/rscbjbr9sj/3 and https://www.kaggle.com/datasets/muhammadshahbazkhan/covid19-pakistani-patients-xray-image-dataset.

References

1. G. Litjens, T. Kooi, B. E. Bejnordi, et al., "A Survey on Deep Learning in Medical Image Analysis," *Medical Image Analysis* 42 (2017): 60–88.

2. E. Çallı, E. Sogancioglu, B. van Ginneken, K. G. van Leeuwen, and K. Murphy, "Deep Learning for Chest X-Ray Analysis: A Survey," *Medical Image Analysis* 72 (2021): 102125.

3. H. Yuan, K. Yu, F. Xie, M. Liu, and S. Sun, "Automated Machine Learning With Interpretation: A Systematic Review of Methodologies and Applications in Healthcare," *Medicine Advances* 2, no. 3 (2024): 205–237.

4. W. Li, J. Li, Z. Wang, et al., "Pathal: An Active Learning Framework for Histopathology Image Analysis," *IEEE Transactions on Medical Imaging* 41, no. 5 (2022): 1176–1187.

5. B. Sayin, E. Krivosheev, J. Yang, A. Passerini, and F. Casati, "A Review and Experimental Analysis of Active Learning Over Crowdsourced Data," *Artificial Intelligence Review* 54 (2021): 5283–5305.

6. J. Liu, L. Cao, and Y. Tian, "Deep Active Learning for Effective Pulmonary Nodule Detection," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (2020), 609–618.

7. A. Sadafi, N. Koehler, A. Makhro, et al., "Multiclass Deep Active Learning for Detecting Red Blood Cell Subtypes in Brightfield Microscopy," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (2019), 685–693.

8. D.-G. Hong, Y.-C. Lee, J. Lee, and S.-W. Kim, "CrowdStart: Warming Up Cold-Start Items Using Crowdsourcing," *Expert Systems With Applications* 138 (2019): 112813.

9. Y. Yan, P.-H. Conze, M. Lamard, et al., "Deep Active Learning for Dual-View Mammogram Analysis," in *Proceedings of the International Workshop on Machine Learning in Medical Imaging* (2021), 180–189.

10. R. Peretzke, K. H. Maier-Hein, J. Bohn, et al., "atTRACTive: Semi-Automatic White Matter Tract Segmentation Using Active Learning," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (2023), 237–246.

11. P. Qian, Y. Chen, J.-W. Kuo, et al., "mDixon-Based Synthetic CT Generation for PET Attenuation Correction on Abdomen and Pelvis Jointly Using Transfer Fuzzy Clustering and Active Learning-Based Classification," *IEEE Transactions on Medical Imaging* 39, no. 4 (2020): 819–832.

12. T. Wan, K. Xu, T. Yu, et al., "A Survey of Deep Active Learning for Foundation Models," *Intelligent Computing* 2 (2023): 0058.

13. H. Yuan, "Overcoming Computational Resource Limitations in Deep Learning for Healthcare: Strategies Targeting Data, Model, and Computing," *Medicine Advances* 3 (2025): 1–4.

14. Y. Tang, Y. Hu, J. Li, et al., "PLD-AL: Pseudo-Label Divergence-Based Active Learning in Carotid Intima-Media Segmentation for Ultrasound Images," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (2023), 57–67.

15. H. Yuan, L. Kang, Y. Li, and Z. Fan, "Human-in-the-Loop Machine Learning for Healthcare: Current Progress and Future Opportunities in Electronic Health Records," *Medicine Advances* 2, no. 3 (2024): 318–322.

16. A. Mumuni and F. Mumuni, "Data Augmentation: A Comprehensive Survey of Modern Approaches," *Array* 16 (2022): 100258.

17. S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering* 22, no. 10 (2010): 1345–1359.

18. K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A Survey of Transfer Learning," *Journal of Big data* 3 (2016): 9.

19. K. Wang, C. Zhang, Y. Geng, and H. Ma, "Evidential Pseudo-Label Ensemble for Semi-Supervised Classification," *Pattern Recognition Letters* 177 (2024): 135–141.

20. J. Fonseca and F. Bacao, "Improving Active Learning Performance Through the Use of Data Augmentation," *International Journal of Intelligent Systems* 2023, no. 1 (2023): 7941878.

21. L. Yang, S. Hanneke, and J. Carbonell, "A Theory of Transfer Learning With Applications to Active Learning," *Machine Learning* 90 (2013): 161–189.

22. Y. Leng, X. Xu, and G. Qi, "Combining Active Learning and Semi-Supervised Learning to Construct SVM Classifier," *Knowledge-Based Systems* 44 (2013): 121–131.

23. Q. Jin, M. Yuan, S. Li, H. Wang, M. Wang, and Z. Song, "Cold-Start Active Learning for Image Classification," *Information Sciences* 616 (2022): 16–36.

24. Y. Zhou, A. Renduchintala, X. Li, S. Wang, Y. Mehdad, and A. Ghoshal, "Towards Understanding the Behaviors of Optimal Deep Active Learning Algorithms," in *Proceedings of the International Conference on Artificial Intelligence and Statistics* (2021), 1486–1494.

25. N. Houlsby, J. M. Hernández-Lobato, and Z. Ghahramani, "Cold-Start Active Learning With Robust Ordinal Matrix Factorization," in *Proceedings of the International Conference on Machine Learning* (2014), 766–774.

26. M. A. Kadir, H. M. T. Alam, and D. Sonntag, "EdgeAL: An Edge Estimation Based Active Learning Approach for OCT Segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (2023), 79–89.

27. P. Donmez, J. G. Carbonell, and P. N. Bennett, "Dual Strategy Active Learning," in *Proceedings of the European Conference on Machine Learning* (2007), 116–127.

28. L. Qu, Y. Ma, Z. Yang, M. Wang, and Z. Song, "Openal: An Efficient Deep Active Learning Framework for Open-Set Pathology Image Classification," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (2023), 3–13.

29. S. Balaram, C. M. Nguyen, A. Kassim, and P. Krishnaswamy, "Consistency-Based Semi-Supervised Evidential Active Learning for Diagnostic Radiograph Classification," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (2022), 675–685.

30. J. Wang, Y. Yan, Y. Zhang, G. Cao, M. Yang, and M. K. Ng, "Deep Reinforcement Active Learning for Medical Image Classification," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (2020), 33–42.

31. X. Shi, Q. Dou, C. Xue, J. Qin, H. Chen, and P.-A. Heng, "An Active Learning Approach for Reducing Annotation Cost in Skin Lesion Analysis," in *Proceedings of the International Workshop on Machine Learning in Medical Imaging* (2019), 628–636.

32. W. Kuo, C. Häne, E. Yuh, P. Mukherjee, and J. Malik, "Cost-Sensitive Active Learning for Intracranial Hemorrhage Detection," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (2018), 715–723.

33. A. Mosinska, J. Tarnawski, and P. Fua, "Active Learning and Proofreading for Delineation of Curvilinear Structures," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (2017), 165–173.

34. Y. Zhu, J. Lin, S. He, et al., "Addressing the Item Cold-Start Problem by Attribute-Driven Active Learning," *IEEE Transactions on Knowledge and Data Engineering* 32, no. 4 (2020): 631–644.

35. D. He, H. Yu, G. Wang, and J. Li, "A Two-Stage Clustering-Based Cold-Start Method for Active Learning," *Intelligent Data Analysis* 25, no. 5 (2021): 1169–1185.

36. H. Yu, X. Yang, S. Zheng, and C. Sun, "Active Learning From Imbalanced Data: A Solution of Online Weighted Extreme Learning Machine," *IEEE Transactions on Neural Networks and Learning Systems* 30, no. 4 (2019): 1088–1103.

37. X. Li, M. Xia, J. Jiao, et al., "HAL-IA: A Hybrid Active Learning Framework Using Interactive Annotation for Medical Image Segmentation," *Medical Image Analysis* 88 (2023): 102862.

38. A. Rodriguez and A. Laio, "Clustering by Fast Search and Find of Density Peaks," *Science* 344, no. 6191 (2014): 1492–1496.

39. R. L. Figueroa, Q. Zeng-Treitler, L. H. Ngo, S. Goryachev, and E. P. Wiechmann, "Active Learning for Clinical Text Classification: Is It Better Than Random Sampling?," *Journal of the American Medical Informatics Association* 19, no. 5 (2012): 809–816.

40. M. Wang, F. Min, Z.-H. Zhang, and Y.-X. Wu, "Active Learning Through Density Clustering," *Expert Systems With Applications* 85 (2017): 305–317.

41. A. Raj and F. Bach, "Convergence of Uncertainty Sampling for Active Learning," in *Proceedings of the International Conference on Machine Learning* (2022), 18310–18331.

42. J. Bernard, M. Zeppelzauer, M. Lehmann, M. Müller, and M. Sedlmair, "Towards User-Centered Active Learning Algorithms," *Computer Graphics Forum* 37, no. 3 (2018): 121–132.

43. V.-L. Nguyen, M. H. Shaker, and E. Hüllermeier, "How to Measure Uncertainty in Uncertainty Sampling for Active Learning," *Machine Learning* 111, no. 1 (2022): 89–122.

44. J. Bernard, M. Hutter, M. Sedlmair, M. Zeppelzauer, and T. Munzner, "A Taxonomy of Property Measures to Unify Active Learning and Human-Centered Approaches to Data Labeling," *ACM Transactions on Interactive Intelligent Systems* 11, no. 3/4 (2021): 1–42.

45. M. Gaillochet, C. Desrosiers, and H. Lombaert, "Active Learning for Medical Image Segmentation With Stochastic Batches," *Medical Image Analysis* 90 (2023): 102958.

46. S. Kee, E. Del Castillo, and G. Runger, "Query-by-Committee Improvement With Diversity and Density in Batch Active Learning," *Information Sciences* 454–455 (2018): 401–418.

47. Z. Dong, S. Niu, X. Gao, J. Li, and X. Shao, "Uncertainty-Weighted Prototype Active Learning in Domain Adaptive Semantic Segmentation," *Expert Systems With Applications* 245 (2024): 123094.

48. L. Yang, Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen, "Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (2017), 399–407.

49. H. Du, E. Barut, and F. Jin, "Uncertainty Quantification in CNN Through the Bootstrap of Convex Neural Networks," in *Proceedings of the AAAI Conference on Artificial Intelligence* (2021), 12078–12085.

50. C. I. Sánchez, M. Niemeijer, M. D. Abràmoff, and B. van Ginneken, "Active Learning for an Efficient Training Strategy of Computer-Aided Diagnosis Systems: Application to Diabetic Retinopathy Screening," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (2010), 603–610.

51. H. Shen, K. Tian, P. Dong, et al., "Deep Active Learning for Breast Cancer Segmentation on Immunohistochemistry Images," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (2020), 509–518.

52. Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," in *Proceedings of the International Conference on Machine Learning* (2016), 1050–1059.

53. Z. Wang, X. Fang, X. Tang, and C. Wu, "Multi-Class Active Learning by Integrating Uncertainty and Diversity," *IEEE Access* 6 (2018): 22794–22803.

54. D. Mahapatra, P. J. Schüffler, J. A. Tielbeek, F. M. Vos, and J. M. Buhmann, "Semi-Supervised and Active Learning for Automatic Segmentation of Crohn's Disease," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (2013), 214–221.

55. X. Luo, W. Liao, J. Xiao, et al., "WORD: A Large Scale Dataset, Benchmark and Clinical Applicable Study for Abdominal Organ Segmentation From CT Image," *Medical Image Analysis* 82 (2022): 102642.

56. Y. Zhou, M. A. Chia, S. K. Wagner, et al., "A Foundation Model for Generalizable Disease Detection From Retinal Images," *Nature* 622, no. 7981 (2023): 156–163.

57. A. B. Sellergren, C. Chen, Z. Nabulsi, et al., "Simplified Transfer Learning for Chest Radiography Models Using Less Data," *Radiology* 305, no. 2 (2022): 454–465.

58. H. Yuan, "Clinical Decision Making: Evolving From the Hypothetico-Deductive Model to Knowledge-Enhanced Machine Learning," *Medicine Advances* 2, no. 4 (2024): 375–379.

59. S. Azizi, L. Culp, J. Freyberg, et al., "Robust and Data-Efficient Generalization of Self-Supervised Machine Learning for Diagnostic Imaging," *Nature Biomedical Engineering* 7, no. 6 (2023): 756–779.

60. M. Moor, O. Banerjee, Z. S. H. Abad, et al., "Foundation Models for Generalist Medical Artificial Intelligence," *Nature* 616, no. 7956 (2023): 259–265.

61. C. Ding, X. He, H. Zha, and H. D. Simon, "Adaptive Dimension Reduction for Clustering High Dimensional Data," in *Proceedings of the International Conference on Data Mining* (2002), 147–154.

62. H. Yuan and C. Hong, "Foundation Model Makes Clustering a Better Initialization for Cold-Start Active Learning," preprint, arXiv, 2024, arXiv:0240202561.

63. J. Devlin, "Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics* (2019), 1–16.

64. M. Yuan, H.-T. Lin, and J. Boyd-Graber, "Cold-Start Active Learning Through Self-Supervised Language Modeling," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (2020), 7935–7948.

65. D. Arthur and S. Vassilvitskii, "K-Means++: The Advantages of Careful Seeding," in *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms* (2007), 1027–1035.

66. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A Large-Scale Hierarchical Image Database," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2009), 248–255.

67. M. A. Morid, A. Borjali, and G. Del Fiol, "A Scoping Review of Transfer Learning Research on Medical Image Analysis Using Imagenet," *Computers in Biology and Medicine* 128 (2021): 104115.

68. F. Zhuang, Z. Qi, K. Duan, et al., "A Comprehensive Survey on Transfer Learning," *Proceedings of the IEEE* 109, no. 1 (2020): 43–76.

69. S. Zhang and D. Metaxas, "On the Challenges and Perspectives of Foundation Models for Medical Image Analysis," *Medical Image Analysis* 91 (2024): 102996.

70. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2016), 770–778.

71. K. Zhang, Y. Yang, J. Yu, et al., "Multi-Task Paired Masking With Alignment Modeling for Medical Vision-Language Pre-Training," *IEEE Transactions on Multimedia* 26 (2024): 4706–4721.

72. H. Zheng, L. Yang, J. Chen, et al., "Biomedical Image Segmentation via Representative Annotation," in *Proceedings of the AAAI Conference on Artificial Intelligence* (2019), 5901–5908.

73. R. S. Bressan, G. Camargo, P. H. Bugatti, and P. T. M. Saito, "Exploring Active Learning Based on Representativeness and Uncertainty for Biomedical Data Classification," *IEEE Journal of Biomedical and Health Informatics* 23, no. 6 (2019): 2238–2244.

74. H. Liu, H. Li, X. Yao, et al., "COLosSAL: A Benchmark for Cold-Start Active Learning for 3D Medical Image Segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (2023), 25–34.

75. Y. Xiao, Z. Chang, and B. Liu, "An Efficient Active Learning Method for Multi-Task Learning," *Knowledge-Based Systems* 190 (2020): 105137.

76. Y. Zhu, S. Zhang, W. Liu, and D. N. Metaxas, "Scalable Histopathological Image Analysis via Active Learning," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (2014), 369–376.

77. Z. Xu, K. Yu, V. Tresp, X. Xu, and J. Wang, "Representative Sampling for Text Classification Using Support Vector Machines," in *Proceedings of the European Conference on Information Retrieval* (2003), 393–407.

78. E. Lughofer, "Hybrid Active Learning for Reducing the Annotation Effort of Operators in Classification Systems," *Pattern Recognition* 45, no. 2 (2012): 884–896.

79. J. A. Hartigan and M. A. Wong, "A K-Means Clustering Algorithm," *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 28, no. 1 (1979): 100–108.

80. H. T. Nguyen and A. Smeulders, "Active Learning Using Pre-Clustering," in *Proceedings of the International Conference on Machine Learning* (2004), 79.

81. M. Tang, X. Luo, and S. Roukos, "Active Learning for Statistical Natural Language Parsing," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (2002), 120–127.

82. J. Zhu, H. Wang, T. Yao, and B. K. Tsou, "Active Learning With Sampling by Uncertainty and Density for Word Sense Disambiguation and Text Classification," in *Proceedings of the International Conference on Computational Linguistics* (2008), 1137–1144.

83. V. Nath, D. Yang, H. R. Roth, and D. Xu, "Warm Start Active Learning With Proxy Labels and Selection via Semi-Supervised Fine-Tuning," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (2022), 297–308.

84. K. Zhang, R. Zhou, E. Adhikarla, et al., "A Generalist Vision–Language Foundation Model for Diverse Biomedical Tasks," *Nature Medicine* 30 (2024): 3129–3141.

85. W. Huang, C. Li, H.-Y. Zhou, et al., "Enhancing Representation in Radiography-Reports Foundation Model: A Granular Alignment Algorithm Using Masked Contrastive Learning," *Nature Communications* 15, no. 1 (2024): 7620.

86. F. Haghighi, M. R. H. Taher, M. B. Gotway, and J. Liang, "Dira: Discriminative, Restorative, and Adversarial Learning for Self-Supervised Medical Image Analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), 20824–20834.

87. R. J. Chen, T. Ding, M. Y. Lu, et al., "Towards a General-Purpose Foundation Model for Computational Pathology," *Nature Medicine* 30, no. 3 (2024): 850–862.

88. M. Long, Y. Cao, Z. Cao, J. Wang, and M. I. Jordan, "Transferable Representation Learning With Deep Adaptation Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, no. 12 (2019): 3071–3085.

89. A. Tendle and M. R. Hasan, "A Study of the Generalizability of Self-Supervised Representations," *Machine Learning With Applications* 6 (2021): 100124.

90. P. Yuan, A. Mobiny, J. Jahanipour, et al., "Few Is Enough: Task-Augmented Active Meta-Learning for Brain Cell Classification," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (2020), 367–377.

91. T. Nair, D. Precup, D. L. Arnold, and T. Arbel, "Exploring Uncertainty Measures in Deep Networks for Multiple Sclerosis Lesion Detection and Segmentation," *Medical Image Analysis* 59 (2020): 101557.

92. T. Nair, D. Precup, D. L. Arnold, and T. Arbel, "Exploring Uncertainty Measures in Deep Networks for Multiple Sclerosis Lesion Detection and Segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (2018), 655–663.

93. C. Leibig, V. Allken, M. S. Ayhan, P. Berens, and S. Wahl, "Leveraging Uncertainty Information From Deep Neural Networks for Disease Detection," *Scientific Reports* 7, no. 1 (2017): 17816.

94. V. Prabhu, A. Chandrasekaran, K. Saenko, and J. Hoffman, "Active Domain Adaptation via Clustering Uncertainty-Weighted Embeddings," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), 8505–8514.

95. Y. Yang, Z. Ma, F. Nie, X. Chang, and A. G. Hauptmann, "Multi-Class Active Learning by Uncertainty Sampling With Diversity Maximization," *International Journal of Computer Vision* 113 (2015): 113–127.

96. B. Du, Z. Wang, L. Zhang, et al., "Exploring Representativeness and Informativeness for Active Learning," *IEEE Transactions on Cybernetics* 47, no. 1 (2017): 14–26.

97. P. T. M. Saito, C. T. N. Suzuki, J. F. Gomes, P. J. de Rezende, and A. X. Falcão, "Robust Active Learning for the Diagnosis of Parasites," *Pattern Recognition* 48, no. 11 (2015): 3572–3583.

98. B. Demir, C. Persello, and L. Bruzzone, "Batch-Mode Active-Learning Methods for the Interactive Classification of Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing* 49, no. 3 (2011): 1014–1031.

99. S. Patra and L. Bruzzone, "A Cluster-Assumption Based Batch Mode Active Learning Technique," *Pattern Recognition Letters* 33, no. 9 (2012): 1042–1048.

100. D. D. Lewis, "A Sequential Algorithm for Training Text Classifiers: Corrigendum and Additional Data," *ACM SIGIR Forum* 29, no. 2 (1995): 13–19.

101. V. Nath, D. Yang, B. A. Landman, D. Xu, and H. R. Roth, "Diminishing Uncertainty Within the Training Pool: Active Learning for Medical Image Segmentation," *IEEE Transactions on Medical Imaging* 40, no. 10 (2021): 2534–2547.

102. M. Kulichenko, K. Barros, N. Lubbers, et al., "Uncertainty-Driven Dynamics for Active Learning of Interatomic Potentials," *Nature Computational Science* 3, no. 3 (2023): 230–239.

103. O. Borkowski, M. Koch, A. Zettor, et al., "Large Scale Active-Learning-Guided Exploration for In Vitro Protein Production Optimization," *Nature Communications* 11, no. 1 (2020): 1872.

104. H. Li and Z. Yin, "Attention, Suggestion and Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (2020), 3–13.

105. A. Top, G. Hamarneh, and R. Abugharbieh, "Active Learning for Interactive 3D Image Segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (2011), 603–610.

106. H. Yuan, "Toward Real-World Deployment of Machine Learning for Healthcare: External Validation, Continual Monitoring, and Randomized Clinical Trials," *Health Care Science* 3, no. 5 (2024): 360–364.

107. D. S. Kermany, M. Goldbaum, W. Cai, et al., "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning," *Cell* 172, no. 5 (2018): 1122–1131.e9.

108. D. S. Kermany, M. Goldbaum, W. Cai, et al., "Guangzhou Dataset," https://data.mendeley.com/datasets/rscbjbr9sj/3.

109. M. Umair, M. S. Khan, F. Ahmed, et al., "Detection of COVID-19 Using Transfer Learning and Grad-Cam Visualization on Indigenously Collected X-Ray Dataset," *Sensors* 21, no. 17 (2021): 5813.

110. M. Umair, M. S. Khan, F. Ahmed, et al., "Pakistan Dataset 2021," https://www.kaggle.com/datasets/muhammadshahbazkhan/covid19-pakistani-patients-xray-image-dataset.

111. K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *Proceedings of the International Conference on Learning Representations* (2015), 1–14.

112. W. Wang, Y. Lu, B. Wu, T. Chen, D. Z. Chen, and J. Wu, "Deep Active Self-Paced Learning for Accurate Pulmonary Nodule Segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (2018), 723–731.

113. A. W. Olthof, P. Shouche, E. M. Fennema, et al., "Machine Learning Based Natural Language Processing of Radiology Reports in Orthopaedic Trauma," *Computer Methods and Programs in Biomedicine* 208 (2021): 106304.

114. J. P. Cohen, J. D. Viviano, P. Bertin, et al., "TorchXRayVision: A Library of Chest X-Ray Datasets and Models," in *Proceedings of the International Conference on Medical Imaging With Deep Learning* (2022), 231–249.

115. G. Huang, Z. Liu, G. Pleiss, L. Maaten, and K. Q. Weinberger, "Convolutional Networks With Dense Connectivity," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, no. 12 (2022): 8704–8716.

116. Y. Wen, L. Chen, Y. Deng, and C. Zhou, "Rethinking Pre-Training on Medical Imaging," *Journal of Visual Communication and Image Representation* 78 (2021): 103145.

117. A. Kolesnikov, L. Beyer, X. Zhai, et al., "Big Transfer (bit): General Visual Representation Learning," in *Proceedings of the European Conference on Computer Vision* (2020), 491–507.

118. C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting Unreasonable Effectiveness of Data in Deep Learning Era," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2017), 843–852.

119. T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," in *Proceedings of the International Conference on Machine Learning* (2020), 1597–1607.

120. M. M. Afzal, M. O. Khan, and Y. Fang, "A Comprehensive Benchmark of Supervised and Self-Supervised Pre-Training on Multi-View Chest X-Ray Classification," in *Proceedings of the Medical Imaging With Deep Learning* (2024), 1–16.

121. H. Yuan, M. Liu, L. Kang, C. Miao, and Y. Wu, "An Empirical Study of the Effect of Background Data Size on the Stability of SHapley Additive exPlanations (SHAP) for Deep Learning Models," in *Proceedings of the International Conference on Learning Representations* (2023), 1–8.

122. E. Agliari, F. Alemanno, A. Barra, and G. De Marzo, "The Emergence of a Concept in Shallow Neural Networks," *Neural Networks* 148 (2022): 232–253.

123. H. Yuan, C. Hong, P.-T. Jiang, et al., "Clinical Domain Knowledge-Derived Template Improves Post Hoc AI Explanations in Pneumothorax Classification," *Journal of Biomedical Informatics* 156 (2024): 104673.

124. H. Yuan, "Anatomic Boundary-Aware Explanation for Convolutional Neural Networks in Diagnostic Radiology," *iRADIOLOGY* 3 (2024): 47–60.

125. B. Glocker, C. Jones, M. Roschewitz, and S. Winzeck, "Risk of Bias in Chest Radiography Deep Learning Foundation Models," *Radiology: Artificial Intelligence* 5, no. 6 (2023): e230060.

126. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning Representations by Back-Propagating Errors," *Nature* 323, no. 6088 (1986): 533–536.

127. Y. X. Zhao, H. Yuan, and Y. Wu, "Prediction of Adverse Drug Reaction Using Machine Learning and Deep Learning Based on an Imbalanced Electronic Medical Records Dataset," in *Proceedings of the International Conference on Medical and Health Informatics* (2021), 17–21.

128. H. Yuan, F. Xie, M. E. H. Ong, et al., "Autoscore-Imbalance: An Interpretable Machine Learning Tool for Development of Clinical Scores With Rare Events Data," *Journal of Biomedical Informatics* 129 (2022): 104072.

129. Q. Sui and S. K. Ghosh, "Similarity-Based Active Learning Methods," *Expert Systems With Applications* 251 (2024): 123849.

130. D. Hu, H. Zhang, S. Li, H. Duan, N. Wu, and X. Lu, "An Ensemble Learning With Active Sampling to Predict the Prognosis of Post-operative Non-Small Cell Lung Cancer Patients," *BMC Medical Informatics and Decision Making* 22, no. 1 (2022): 245.

131. F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, "Data Imbalance in Classification: Experimental Evaluation," *Information Sciences* 513 (2020): 429–441.

132. J. T. Hancock, T. M. Khoshgoftaar, and J. M. Johnson, "Evaluating Classifier Performance With Highly Imbalanced Big Data," *Journal of Big Data* 10, no. 1 (2023): 42.

133. M. Owusu-Adjei, J. Ben Hayfron-Acquah, T. Frimpong, and G. Abdul-Salaam, "Imbalanced Class Distribution and Performance Evaluation Metrics: A Systematic Review of Prediction Accuracy for Determining Model Performance in Healthcare Systems," *PLOS Digital Health* 2, no. 11 (2023): e0000290.

134. E. Mortaz, "Imbalance Accuracy Metric for Model Selection in Multi-Class Imbalance Classification Problems," *Knowledge-Based Systems* 210 (2020): 106490.

135. B. Efron and R. Tibshirani, "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy," *Statistical Science* 1, no. 1 (1986): 54–75.

136. "Code Repository of Foundation Model-Based Cold-Start Active Learning," 2024, https://github.com/Han-Yuan-Med/Active-Learning-CXR-Classification.

137. Q. Jin, S. Li, X. Du, M. Yuan, M. Wang, and Z. Song, "Density-Based One-Shot Active Learning for Image Segmentation," *Engineering Applications of Artificial Intelligence* 126 (2023): 106805.

138. Q. Jin, M. Yuan, Q. Qiao, and Z. Song, "One-Shot Active Learning for Image Segmentation via Contrastive Learning and Diversity-Based Sampling," *Knowledge-Based Systems* 241 (2022): 108278.

139. M. J. Rovine and A. Von Eye, "A 14th Way to Look at a Correlation Coefficient: Correlation as the Proportion of Matches," *American Statistician* 51, no. 1 (1997): 42–46.

140. D. W. Zimmerman, "A Note on Interpretation of the Paired-Samples *t* Test," *Journal of Educational and Behavioral Statistics* 22, no. 3 (1997): 349–360.

141. S. Zhou, D. Nie, E. Adeli, J. Yin, J. Lian, and D. Shen, "High-Resolution Encoder–Decoder Networks for Low-Contrast Medical Image Segmentation," *IEEE Transactions on Image Processing* 29 (2020): 461–475.

142. H. Dong, Y. Zhang, H. Gu, N. Konz, Y. Zhang, and M. A. Mazurowski, "Swssl: Sliding Window-Based Self-Supervised Learning for Anomaly Detection in High-Resolution Images," *IEEE Transactions on Medical Imaging* 42, no. 12 (2023): 3860–3870.

143. Y. Ji, H. Bai, C. Ge, et al., "Amos: A Large-Scale Abdominal Multi-Organ Benchmark for Versatile Medical Image Segmentation," in *Proceedings of the Advances in Neural Information Processing Systems* (2022), 36722–36732.

144. Y. Yang and M. Loog, "To Actively Initialize Active Learning," *Pattern Recognition* 131 (2022): 108836.

145. J. P. Huix, A. R. Ganeshan, J. F. Haslum, M. Söderberg, C. Matsoukas, and K. Smith, "Are Natural Domain Foundation Models Useful for Medical Image Classification?," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2024), 7634–7643.

146. B. Bago and J.-F. Bonnefon, "Generative AI as a Tool for Truth," *Science* 385, no. 6714 (2024): 1164–1165.

147. K. Singhal, S. Azizi, T. Tu, et al., "Large Language Models Encode Clinical Knowledge," *Nature* 620, no. 7972 (2023): 172–180.

148. W. F. Wiggins and A. S. Tejani, "On the Opportunities and Risks of Foundation Models for Natural Language Processing in Radiology," *Radiology: Artificial Intelligence* 4, no. 4 (2022): e220119.

149. S. Jain, H. Salman, A. Khaddaj, E. Wong, S. M. Park, and A. Mądry, "A Data-Based Perspective on Transfer Learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), 3613–3622.

150. F. Maleki, K. Ovens, R. Gupta, C. Reinhold, A. Spatz, and R. Forghani, "Generalizability of Machine Learning Models: Quantitative Evaluation of Three Methodological Pitfalls," *Radiology: Artificial Intelligence* 5, no. 1 (2022): e220028.

151. Y. Duan, X. Chen, H. Xu, et al., "Transnas-Bench-101: Improving Transferability and Generalizability of Cross-Task Neural Architecture Search," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), 5251–5260.

152. H. Wang, H. Jia, L. Lu, and Y. Xia, "Thorax-Net: An Attention Regularized Deep Neural Network for Classification of Thoracic Diseases on Chest Radiography," *IEEE Journal of Biomedical and Health Informatics* 24, no. 2 (2020): 475–485.

153. J. T. Wu, K. C. L. Wong, Y. Gur, et al., "Comparison of Chest Radiograph Interpretations by Artificial Intelligence Algorithm vs Radiology Residents," *JAMA Network Open* 3, no. 10 (2020): e2022779.

154. Y.-X. Tang, Y.-B. Tang, Y. Peng, et al., "Automated Abnormality Classification of Chest Radiographs Using Deep Convolutional Neural Networks," *NPJ Digital Medicine* 3, no. 1 (2020): 70.

155. A. Kirillov, E. Mintun, N. Ravi, et al., "Segment Anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), 4015–4026.

156. X. Zou, J. Yang, H. Zhang, et al., "Segment Everything Everywhere All at Once," in *Proceedings of the Advances in Neural Information Processing Systems* (2024), 19769–19782.

157. M. Oquab, T. Darcet, T. Moutakanni, et al., "DINOv2: Learning Robust Visual Features Without Supervision," *Transactions on Machine Learning Research* 3 (2024): 1–32.

158. A. Radford, J. W. Kim, C. Hallacy, et al., "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the International Conference on Machine Learning* (2021), 8748–8763.

159. J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping Language-Image Pre-Training for Unified Vision-Language Understanding and Generation," in *Proceedings of the International Conference on Machine Learning* (2022), 12888–12900.

160. H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training Data-Efficient Image Transformers & Distillation Through Attention," in *Proceedings of the International Conference on Machine Learning* (2021), 10347–10357.

161. S.-J. Huang, R. Jin, and Z.-H. Zhou, "Active Learning by Querying Informative and Representative Examples," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, no. 10 (2014): 1936–1949.

162. Z. Zhou, J. Y. Shin, S. R. Gurudu, M. B. Gotway, and J. Liang, "Active, Continual Fine Tuning of Convolutional Neural Networks for Reducing Annotation Efforts," *Medical Image Analysis* 71 (2021): 101997.

163. G. E. Hinton and R. R. Salakhutdinov, "Reducing the Dimensionality of Data With Neural Networks," *Science* 313, no. 5786 (2006): 504–507.

164. D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *Proceedings of the International Conference on Learning Representations* (2014), 1–14.

165. I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., "Generative Adversarial Networks," *Communications of the ACM* 63, no. 11 (2020): 139–144.

166. F. Ozdemir, Z. Peng, P. Fuernstahl, C. Tanner, and O. Goksel, "Active Learning for Segmentation Based on Bayesian Sample Queries," *Knowledge-Based Systems* 214 (2021): 106531.

167. O. Sener and S. Savarese, "Active Learning for Convolutional Neural Networks: A Core-Set Approach," in *Proceedings of the International Conference on Learning Representations* (2018), 1–13.

168. D. Mahapatra, B. Bozorgtabar, Z. Ge, and M. Reyes, "GANDALF: Graph-Based Transformer and Data Augmentation Active Learning Framework With Interpretable Features for Multi-Label Chest X-Ray Classification," *Medical Image Analysis* 93 (2024): 103075.

169. J. Fonseca and F. Bacao, "Tabular and Latent Space Synthetic Data Generation: A Literature Review," *Journal of Big Data* 10, no. 1 (2023): 115.

170. D. Park, Y. Shin, J. Bang, Y. Lee, H. Song, and J.-G. Lee, "Meta-Query-Net: Resolving Purity-Informativeness Dilemma in Open-Set Active Learning," in *Proceedings of the Advances in Neural Information Processing Systems* (2022), 31416–31429.

171. S. Ho, M. Liu, S. Gao, and L. Gao, "Learning to Learn for Few-Shot Continual Active Learning," *Artificial Intelligence Review* 57, no. 10 (2024): 280.

172. V. Verma, A. Lamb, C. Beckham, et al., "Manifold Mixup: Better Representations by Interpolating Hidden States," in *Proceedings of the International Conference on Machine Learning* (2019), 6438–6447.

173. L. Carratino, M. Cissé, R. Jenatton, and J.-P. Vert, "On Mixup Regularization," *Journal of Machine Learning Research* 23, no. 325 (2022): 1–31.

174. D. Mahapatra, B. Bozorgtabar, J.-P. Thiran, and M. Reyes, "Efficient Active Learning for Image Classification and Segmentation Using a Sample Selection and Conditional Generative Adversarial Network," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (2018), 580–588.

175. F. Bai, X. Xing, Y. Shen, H. Ma, and M. Q.-H. Meng, "Discrepancy-Based Active Learning for Weakly Supervised Bleeding Segmentation in Wireless Capsule Endoscopy Images," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (2022), 24–34.

176. H. Yuan, C. Hong, N. T. A. Tran, X. Xu, and N. Liu, "Leveraging Anatomical Constraints With Uncertainty for Pneumothorax Segmentation," *Health Care Science* 3, no. 6 (2024): 456–474.

177. M. Huisman, J. N. Van Rijn, and A. Plaat, "A Survey of Deep Meta-Learning," *Artificial Intelligence Review* 54, no. 6 (2021): 4483–4541.

178. T. Gong, X. Zheng, and X. Lu, "Meta Self-Supervised Learning for Distribution Shifted Few-Shot Scene Classification," *IEEE Geoscience and Remote Sensing Letters* 19 (2022): 1–5.

179. S. Bucci, A. D'Innocente, Y. Liao, F. M. Carlucci, B. Caputo, and T. Tommasi, "Self-Supervised Learning Across Domains," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, no. 9 (2021): 1.

180. H. Sowrirajan, J. Yang, A. Y. Ng, and P. Rajpurkar, "Moco Pretraining Improves Representation and Transferability of Chest X-Ray Models," in *Proceedings of the International Conference on Medical Imaging With Deep Learning* (2021), 728–744.

181. C. Bluethgen, P. Chambon, J.-B. Delbrouck, et al., "A Vision–Language Foundation Model for the Generation of Realistic Chest X-Ray Images," *Nature Biomedical Engineering* 8, no. 9 (2024): 1–13.

182. X. Liu, F. Zhang, Z. Hou, et al., "Self-Supervised Learning: Generative or Contrastive," *IEEE Transactions on Knowledge and Data Engineering* 35, no. 1 (2021): 1.

183. T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases," *ACM Sigmod Record* 25, no. 2 (1996): 103–114.

184. N. Alon, N. Cesa-Bianchi, C. Gentile, S. Mannor, Y. Mansour, and O. Shamir, "Nonstochastic Multi-Armed Bandits With Graph-Structured Feedback," *SIAM Journal on Computing* 46, no. 6 (2017): 1785–1826.

185. G. Hacohen, A. Dekel, and D. Weinshall, "Active Learning on a Budget: Opposite Strategies Suit High and Low Budgets," in *Proceedings of the International Conference on Machine Learning* (2022), 8175–8195.

186. C. Geng, S. J. Huang, and S. Chen, "Recent Advances in Open Set Recognition: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, no. 10 (2021): 3614–3631.

187. K. Han, Y. Wang, H. Chen, et al., "A Survey on Vision Transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, no. 1 (2023): 87–110.

188. J. S. Lee, "A Review of Deep-Learning-Based Approaches for Attenuation Correction in Positron Emission Tomography," *IEEE Transactions on Radiation and Plasma Medical Sciences* 5, no. 2 (2021): 160–184.

189. D. Holzmüller, V. Zaverkin, J. Kästner, and I. Steinwart, "A Framework and Benchmark for Deep Batch Active Learning for Regression," *Journal of Machine Learning Research* 24, no. 164 (2023): 1–81.

190. Y. Yang and M. Loog, "A Benchmark and Comparison of Active Learning for Logistic Regression," *Pattern Recognition* 83 (2018): 401–415.

191. H. Yuan, "Toward the Comprehensive Evaluation of Medical Text Generation by Large Language Models: Programmatic Metrics, Human Assessment, and Large Language Models Judgment," *Medicine Advances* 3 (2025): 1–4.

192. S. Li, Y. Ning, M. E. H. Ong, et al., "FedScore: A Privacy-Preserving Framework for Federated Scoring System Development," *Journal of Biomedical Informatics* 146 (2023): 104485.

193. H. Li, C. Li, J. Wang, et al., "Review on Security of Federated Learning and Its Application in Healthcare," *Future Generation Computer Systems* 144 (2023): 271–290.

194. H. Guan, P.-T. Yap, A. Bozoki, and M. Liu, "Federated Learning for Medical Image Analysis: A Survey," *Pattern Recognition* 151 (2024): 110424.

195. M. Cherti, R. Beaumont, R. Wightman, et al., "Reproducible Scaling Laws for Contrastive Language-Image Learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), 2818–2829.

196. Z. Guo, X. Li, H. Huang, N. Guo, and Q. Li, "Medical Image Segmentation Based on Multi-Modal Convolutional Neural Network: Study on Image Fusion Schemes," in *Proceedings of the International Symposium on Biomedical Imaging* (2018), 903–907.

197. L. Fang, X. Wang, and L. Wang, "Multi-Modal Medical Image Segmentation Based on Vector-Valued Active Contour Models," *Information Sciences* 513 (2020): 504–518.

198. H. Liu, Y. Zhuang, E. Song, et al., "A Modality-Collaborative Convolution and Transformer Hybrid Network for Unpaired Multi-Modal Medical Image Segmentation With Limited Annotations," *Medical Physics* 50, no. 9 (2023): 5460–5478.

199. H. Yuan, "Natural Language Processing for Chest X-Ray Reports in the Transformer Era: BERT-Like Encoders for Comprehension and GPT-Like Decoders for Generation," *iRADIOLOGY* 3 (2025): 1–7.

200. H. Yuan, "Agentic Large Language Models for Healthcare: Current Progress and Future Opportunities," *Medicine Advances* 3 (2025): 1–5.

201. S. Yang, X. Wu, S. Ge, Z. Zheng, S. K. Zhou, and L. Xiao, "Radiology Report Generation With a Learned Knowledge Base and Multi-Modal Alignment," *Medical Image Analysis* 86 (2023): 102798.

202. C. Shang, S. Cui, T. Li, X. Wang, Y. Li, and J. Jiang, "MATNet: Exploiting Multi-Modal Features for Radiology Report Generation," *IEEE Signal Processing Letters* 29 (2022): 2692–2696.

203. X. Huang, J. Liu, Z. Zhang, et al., "Cross-Modal Recipe Retrieval With Fine-Grained Prompting Alignment and Evidential Semantic Consistency," *IEEE Transactions on Multimedia* 27 (2024): 1–12.

Appendix A

TABLE A1 | Initialization performance of random sampling, diversity, uncertainty, and hybrid methods on the Guangzhou data set.

Initialization budget	Foundation model	Initialization method	Model input	AUROC	AUPRC
All samples	/	/	Original pixel	0.998 (0.001)	0.999 (0.000)
			TXRV	0.938 (0.007)	0.977 (0.003)
			REMEDIS	0.993 (0.003)	0.997 (0.002)
			DenseNet	0.848 (0.012)	0.939 (0.007)
			ResNet	0.990 (0.003)	0.996 (0.001)
10	/	Random	Original pixel	0.898 (0.121)	0.954 (0.066)
			TXRV	0.599 (0.120)	0.797 (0.072)
			REMEDIS	0.764 (0.130)	0.872 (0.075)
			DenseNet	0.688 (0.089)	0.838 (0.054)
			ResNet	0.745 (0.097)	0.866 (0.056)
	TXRV	Diversity	Original pixel	0.958 (0.006)	0.981 (0.005)
			TXRV	0.466 (0.018)	0.703 (0.018)
		Uncertainty	Original pixel	0.573 (0.020)	0.777 (0.017)
			TXRV	0.425 (0.017)	0.700 (0.019)
		Hybrid	Original pixel	0.885 (0.010)	0.954 (0.006)
			TXRV	0.551 (0.018)	0.771 (0.017)
	REMEDIS	Diversity	Original pixel	0.952 (0.006)	0.982 (0.003)
			REMEDIS	0.475 (0.019)	0.780 (0.013)
		Uncertainty	Original pixel	0.874 (0.011)	0.933 (0.010)
			REMEDIS	0.713 (0.016)	0.877 (0.010)
		Hybrid	Original pixel	0.948 (0.006)	0.980 (0.003)
			REMEDIS	0.870 (0.015)	0.921 (0.011)
	DenseNet	Diversity	Original pixel	0.963 (0.006)	0.983 (0.004)
			DenseNet	0.640 (0.021)	0.781 (0.016)
		Uncertainty	Original pixel	0.568 (0.018)	0.756 (0.019)
			DenseNet	0.615 (0.018)	0.809 (0.013)
		Hybrid	Original pixel	0.629 (0.017)	0.796 (0.016)
			DenseNet	0.774 (0.017)	0.897 (0.010)
	ResNet	Diversity	Original pixel	0.972 (0.004)	0.989 (0.002)
			ResNet	0.854 (0.013)	0.923 (0.011)
		Uncertainty	Original pixel	0.925 (0.007)	0.973 (0.003)
			ResNet	0.569 (0.018)	0.780 (0.016)
		Hybrid	Original pixel	0.914 (0.009)	0.962 (0.005)
			ResNet	0.758 (0.017)	0.849 (0.016)
20	/	Random	Original pixel	0.953 (0.020)	0.980 (0.009)
			TXRV	0.697 (0.108)	0.848 (0.061)
			REMEDIS	0.879 (0.077)	0.930 (0.057)
			DenseNet	0.856 (0.058)	0.925 (0.039)
			ResNet	0.873 (0.059)	0.938 (0.032)
	TXRV	Diversity	Original pixel	0.964 (0.006)	0.985 (0.003)
			TXRV	0.754 (0.016)	0.874 (0.013)

(Continues)

TABLE A1 | (Continued)

Initialization budget	Foundation model	Initialization method	Model input	AUROC	AUPRC
		Uncertainty	Original pixel	0.828 (0.014)	0.926 (0.009)
			TXRV	0.488 (0.018)	0.748 (0.017)
		Hybrid	Original pixel	0.957 (0.006)	0.981 (0.004)
			TXRV	0.528 (0.017)	0.759 (0.018)
	REMEDIS	Diversity	Original pixel	0.982 (0.004)	0.992 (0.003)
			REMEDIS	0.830 (0.013)	0.925 (0.010)
		Uncertainty	Original pixel	0.978 (0.003)	0.991 (0.001)
			REMEDIS	0.840 (0.014)	0.916 (0.009)
		Hybrid	Original pixel	0.961 (0.005)	0.986 (0.002)
			REMEDIS	0.917 (0.010)	0.964 (0.008)
	DenseNet	Diversity	Original pixel	0.977 (0.004)	0.991 (0.002)
			DenseNet	0.897 (0.009)	0.956 (0.006)
		Uncertainty	Original pixel	0.530 (0.019)	0.729 (0.017)
			DenseNet	0.711 (0.018)	0.861 (0.012)
		Hybrid	Original pixel	0.906 (0.009)	0.964 (0.004)
			DenseNet	0.905 (0.010)	0.954 (0.007)
	ResNet	Diversity	Original pixel	0.972 (0.005)	0.989 (0.003)
			ResNet	0.895 (0.011)	0.958 (0.006)
		Uncertainty	Original pixel	0.953 (0.007)	0.982 (0.003)
			ResNet	0.888 (0.010)	0.952 (0.007)
		Hybrid	Original pixel	0.922 (0.009)	0.968 (0.005)
			ResNet	0.892 (0.010)	0.958 (0.005)
30	/	Random	Original pixel	0.969 (0.011)	0.988 (0.005)
			TXRV	0.763 (0.060)	0.885 (0.035)
			REMEDIS	0.885 (0.067)	0.931 (0.042)
			DenseNet	0.901 (0.034)	0.952 (0.019)
			ResNet	0.905 (0.041)	0.955 (0.024)
	TXRV	Diversity	Original pixel	0.973 (0.004)	0.989 (0.003)
			TXRV	0.816 (0.012)	0.913 (0.010)
		Uncertainty	Original pixel	0.913 (0.009)	0.961 (0.006)
			TXRV	0.612 (0.016)	0.816 (0.014)
		Hybrid	Original pixel	0.977 (0.004)	0.991 (0.002)
			TXRV	0.840 (0.012)	0.927 (0.009)
	REMEDIS	Diversity	Original pixel	0.982 (0.004)	0.992 (0.002)
			REMEDIS	0.924 (0.008)	0.966 (0.006)
		Uncertainty	Original pixel	0.950 (0.006)	0.980 (0.003)
			REMEDIS	0.849 (0.014)	0.931 (0.009)
		Hybrid	Original pixel	0.974 (0.004)	0.990 (0.002)
			REMEDIS	0.934 (0.010)	0.967 (0.007)
	DenseNet	Diversity	Original pixel	0.968 (0.005)	0.987 (0.003)
			DenseNet	0.928 (0.008)	0.964 (0.006)
		Uncertainty	Original pixel	0.831 (0.015)	0.911 (0.012)
			DenseNet	0.683 (0.017)	0.822 (0.014)

(Continues)

TABLE A1 | (Continued)

Initialization budget	Foundation model	Initialization method	Model input	AUROC	AUPRC
		Hybrid	Original pixel	0.969 (0.004)	0.989 (0.002)
			DenseNet	0.899 (0.012)	0.945 (0.008)
	ResNet	Diversity	Original pixel	0.979 (0.004)	0.992 (0.002)
			ResNet	0.912 (0.008)	0.968 (0.004)
		Uncertainty	Original pixel	0.962 (0.005)	0.985 (0.002)
			ResNet	0.884 (0.010)	0.951 (0.007)
		Hybrid	Original pixel	0.960 (0.007)	0.983 (0.004)
			ResNet	0.911 (0.008)	0.964 (0.005)
40	/	Random	Original pixel	0.975 (0.009)	0.990 (0.004)
			TXRV	0.785 (0.056)	0.896 (0.033)
			REMEDIS	0.906 (0.046)	0.943 (0.035)
			DenseNet	0.921 (0.030)	0.963 (0.017)
			ResNet	0.922 (0.028)	0.963 (0.018)
	TXRV	Diversity	Original pixel	0.985 (0.003)	0.994 (0.001)
			TXRV	0.735 (0.015)	0.872 (0.011)
		Uncertainty	Original pixel	0.969 (0.005)	0.987 (0.003)
			TXRV	0.653 (0.016)	0.843 (0.012)
		Hybrid	Original pixel	0.976 (0.004)	0.991 (0.002)
			TXRV	0.788 (0.012)	0.907 (0.008)
	REMEDIS	Diversity	Original pixel	0.979 (0.004)	0.992 (0.003)
			REMEDIS	0.921 (0.007)	0.962 (0.008)
		Uncertainty	Original pixel	0.967 (0.006)	0.986 (0.003)
			REMEDIS	0.829 (0.013)	0.907 (0.012)
		Hybrid	Original pixel	0.979 (0.003)	0.992 (0.001)
			REMEDIS	0.920 (0.011)	0.948 (0.011)
	DenseNet	Diversity	Original pixel	0.983 (0.003)	0.993 (0.002)
			DenseNet	0.920 (0.008)	0.967 (0.005)
		Uncertainty	Original pixel	0.890 (0.012)	0.946 (0.009)
			DenseNet	0.873 (0.010)	0.946 (0.007)
		Hybrid	Original pixel	0.973 (0.005)	0.987 (0.004)
			DenseNet	0.917 (0.009)	0.961 (0.006)
	ResNet	Diversity	Original pixel	0.980 (0.004)	0.991 (0.003)
			ResNet	0.948 (0.007)	0.977 (0.004)
		Uncertainty	Original pixel	0.956 (0.006)	0.982 (0.003)
			ResNet	0.872 (0.010)	0.945 (0.006)
		Hybrid	Original pixel	0.966 (0.005)	0.985 (0.003)
			ResNet	0.907 (0.009)	0.958 (0.007)
50	/	Random	Original pixel	0.978 (0.007)	0.991 (0.004)
			TXRV	0.799 (0.042)	0.904 (0.026)
			REMEDIS	0.915 (0.040)	0.947 (0.034)
			DenseNet	0.936 (0.021)	0.970 (0.014)
			ResNet	0.934 (0.022)	0.969 (0.014)
	TXRV	Diversity	Original pixel	0.982 (0.004)	0.993 (0.002)

(Continues)

TABLE A1 | (Continued)

Initialization budget	Foundation model	Initialization method	Model input	AUROC	AUPRC
			TXRV	0.822 (0.014)	0.915 (0.011)
		Uncertainty	Original pixel	0.971 (0.005)	0.989 (0.003)
			TXRV	0.770 (0.013)	0.906 (0.008)
		Hybrid	Original pixel	0.984 (0.003)	0.993 (0.002)
			TXRV	0.734 (0.014)	0.877 (0.013)
	REMEDIS	Diversity	Original pixel	0.983 (0.003)	0.994 (0.002)
			REMEDIS	0.902 (0.010)	0.964 (0.006)
		Uncertainty	Original pixel	0.972 (0.004)	0.989 (0.002)
			REMEDIS	0.793 (0.014)	0.893 (0.011)
		Hybrid	Original pixel	0.980 (0.003)	0.993 (0.001)
			REMEDIS	0.948 (0.008)	0.958 (0.009)
	DenseNet	Diversity	Original pixel	0.981 (0.004)	0.992 (0.002)
			DenseNet	0.911 (0.010)	0.958 (0.006)
		Uncertainty	Original pixel	0.968 (0.005)	0.987 (0.002)
			DenseNet	0.923 (0.008)	0.971 (0.003)
		Hybrid	Original pixel	0.978 (0.004)	0.990 (0.003)
			DenseNet	0.821 (0.011)	0.926 (0.008)
	ResNet	Diversity	Original pixel	0.984 (0.003)	0.993 (0.002)
			ResNet	0.972 (0.004)	0.989 (0.002)
		Uncertainty	Original pixel	0.964 (0.005)	0.984 (0.003)
			ResNet	0.858 (0.012)	0.932 (0.009)
		Hybrid	Original pixel	0.963 (0.006)	0.985 (0.003)
			ResNet	0.921 (0.009)	0.959 (0.008)

TABLE A2 | Initialization performance of random sampling, diversity, uncertainty, and hybrid methods on the Pakistan data set.

Initialization budget	Foundation model	Initialization method	Model input	AUROC	AUPRC
All samples	/	/	Original pixel	0.994 (0.003)	0.999 (0.001)
			TXRV	0.991 (0.006)	0.999 (0.001)
			REMEDIS	0.968 (0.016)	0.994 (0.003)
			DenseNet	0.964 (0.023)	0.987 (0.010)
			ResNet	0.993 (0.005)	0.999 (0.001)
10	/	Random	Original pixel	0.794 (0.144)	0.951 (0.047)
			TXRV	0.669 (0.147)	0.914 (0.047)
			REMEDIS	0.661 (0.148)	0.915 (0.044)
			DenseNet	0.604 (0.096)	0.897 (0.036)
			ResNet	0.641 (0.132)	0.907 (0.046)
	TXRV	Diversity	Original pixel	0.909 (0.025)	0.984 (0.005)
			TXRV	0.696 (0.051)	0.932 (0.018)
		Uncertainty	Original pixel	0.888 (0.036)	0.972 (0.013)
			TXRV	0.666 (0.050)	0.926 (0.020)
		Hybrid	Original pixel	0.940 (0.024)	0.989 (0.006)
			TXRV	0.678 (0.050)	0.929 (0.019)

(Continues)

TABLE A2 | (Continued)

Initialization budget	Foundation model	Initialization method	Model input	AUROC	AUPRC
	REMEDIS	Diversity	Original pixel	0.745 (0.040)	0.949 (0.013)
			REMEDIS	0.705 (0.064)	0.926 (0.019)
		Uncertainty	Original pixel	0.790 (0.041)	0.958 (0.013)
			REMEDIS	0.796 (0.041)	0.957 (0.016)
		Hybrid	Original pixel	0.963 (0.015)	0.994 (0.003)
			REMEDIS	0.636 (0.068)	0.899 (0.026)
	DenseNet	Diversity	Original pixel	0.860 (0.034)	0.970 (0.011)
			DenseNet	0.567 (0.058)	0.879 (0.029)
		Uncertainty	Original pixel	0.832 (0.047)	0.942 (0.024)
			DenseNet	0.548 (0.057)	0.891 (0.025)
		Hybrid	Original pixel	0.846 (0.043)	0.970 (0.010)
			DenseNet	0.604 (0.058)	0.884 (0.030)
	ResNet	Diversity	Original pixel	0.897 (0.029)	0.980 (0.008)
			ResNet	0.774 (0.039)	0.954 (0.012)
		Uncertainty	Original pixel	0.888 (0.043)	0.974 (0.012)
			ResNet	0.738 (0.048)	0.939 (0.019)
		Hybrid	Original pixel	0.908 (0.040)	0.978 (0.011)
			ResNet	0.733 (0.046)	0.938 (0.020)
20	/	Random	Original pixel	0.874 (0.129)	0.972 (0.039)
			TXRV	0.775 (0.142)	0.942 (0.044)
			REMEDIS	0.792 (0.146)	0.947 (0.047)
			DenseNet	0.740 (0.106)	0.931 (0.036)
			ResNet	0.750 (0.111)	0.939 (0.039)
	TXRV	Diversity	Original pixel	0.933 (0.021)	0.989 (0.004)
			TXRV	0.948 (0.018)	0.991 (0.003)
		Uncertainty	Original pixel	0.877 (0.034)	0.972 (0.011)
			TXRV	0.819 (0.046)	0.948 (0.020)
		Hybrid	Original pixel	0.944 (0.018)	0.991 (0.003)
			TXRV	0.952 (0.018)	0.992 (0.003)
	REMEDIS	Diversity	Original pixel	0.859 (0.048)	0.968 (0.014)
			REMEDIS	0.880 (0.028)	0.980 (0.006)
		Uncertainty	Original pixel	0.948 (0.024)	0.990 (0.005)
			REMEDIS	0.733 (0.044)	0.931 (0.022)
		Hybrid	Original pixel	0.878 (0.040)	0.973 (0.012)
			REMEDIS	0.954 (0.013)	0.993 (0.002)
	DenseNet	Diversity	Original pixel	0.948 (0.024)	0.989 (0.006)
			DenseNet	0.794 (0.044)	0.960 (0.012)
		Uncertainty	Original pixel	0.888 (0.048)	0.970 (0.016)
			DenseNet	0.824 (0.032)	0.970 (0.008)
		Hybrid	Original pixel	0.970 (0.012)	0.995 (0.002)
			DenseNet	0.885 (0.029)	0.973 (0.011)
	ResNet	Diversity	Original pixel	0.946 (0.023)	0.990 (0.005)
			ResNet	0.900 (0.039)	0.977 (0.011)

(Continues)

TABLE A2 | (Continued)

Initialization budget	Foundation model	Initialization method	Model input	AUROC	AUPRC
		Uncertainty	Original pixel	0.872 (0.044)	0.962 (0.018)
			ResNet	0.861 (0.030)	0.975 (0.007)
		Hybrid	Original pixel	0.930 (0.026)	0.987 (0.006)
			ResNet	0.879 (0.030)	0.977 (0.008)
30	/	Random	Original pixel	0.899 (0.109)	0.977 (0.032)
			TXRV	0.844 (0.109)	0.958 (0.033)
			REMEDIS	0.825 (0.125)	0.956 (0.038)
			DenseNet	0.814 (0.099)	0.955 (0.030)
			ResNet	0.792 (0.103)	0.950 (0.033)
	TXRV	Diversity	Original pixel	0.945 (0.019)	0.991 (0.004)
			TXRV	0.954 (0.014)	0.993 (0.002)
		Uncertainty	Original pixel	0.895 (0.033)	0.977 (0.009)
			TXRV	0.896 (0.039)	0.960 (0.023)
		Hybrid	Original pixel	0.896 (0.034)	0.981 (0.007)
			TXRV	0.937 (0.025)	0.987 (0.005)
	REMEDIS	Diversity	Original pixel	0.832 (0.057)	0.958 (0.017)
			REMEDIS	0.761 (0.054)	0.934 (0.022)
		Uncertainty	Original pixel	0.960 (0.020)	0.992 (0.004)
			REMEDIS	0.848 (0.043)	0.966 (0.012)
		Hybrid	Original pixel	0.909 (0.026)	0.984 (0.006)
			REMEDIS	0.919 (0.026)	0.985 (0.007)
	DenseNet	Diversity	Original pixel	0.963 (0.014)	0.994 (0.002)
			DenseNet	0.847 (0.036)	0.971 (0.009)
		Uncertainty	Original pixel	0.891 (0.048)	0.972 (0.013)
			DenseNet	0.777 (0.050)	0.944 (0.020)
		Hybrid	Original pixel	0.973 (0.009)	0.996 (0.002)
			DenseNet	0.900 (0.023)	0.983 (0.005)
	ResNet	Diversity	Original pixel	0.910 (0.038)	0.978 (0.012)
			ResNet	0.817 (0.049)	0.963 (0.013)
		Uncertainty	Original pixel	0.925 (0.028)	0.985 (0.006)
			ResNet	0.688 (0.042)	0.942 (0.013)
		Hybrid	Original pixel	0.943 (0.019)	0.990 (0.004)
			ResNet	0.852 (0.038)	0.970 (0.010)
40	/	Random	Original pixel	0.931 (0.074)	0.986 (0.020)
			TXRV	0.880 (0.093)	0.967 (0.028)
			REMEDIS	0.853 (0.101)	0.963 (0.036)
			DenseNet	0.860 (0.074)	0.965 (0.025)
			ResNet	0.825 (0.096)	0.958 (0.030)
	TXRV	Diversity	Original pixel	0.951 (0.018)	0.992 (0.004)
			TXRV	0.947 (0.024)	0.991 (0.004)
		Uncertainty	Original pixel	0.918 (0.038)	0.978 (0.012)
			TXRV	0.868 (0.043)	0.951 (0.025)
		Hybrid	Original pixel	0.974 (0.009)	0.996 (0.002)

(Continues)

TABLE A2 | (Continued)

Initialization budget	Foundation model	Initialization method	Model input	AUROC	AUPRC
			TXRV	0.973 (0.010)	0.996 (0.002)
	REMEDIS	Diversity	Original pixel	0.956 (0.021)	0.991 (0.005)
			REMEDIS	0.835 (0.028)	0.974 (0.007)
		Uncertainty	Original pixel	0.953 (0.024)	0.990 (0.005)
			REMEDIS	0.821 (0.049)	0.953 (0.016)
		Hybrid	Original pixel	0.944 (0.030)	0.989 (0.007)
			REMEDIS	0.927 (0.029)	0.983 (0.009)
	DenseNet	Diversity	Original pixel	0.945 (0.021)	0.990 (0.004)
			DenseNet	0.879 (0.035)	0.964 (0.020)
		Uncertainty	Original pixel	0.882 (0.047)	0.970 (0.013)
			DenseNet	0.780 (0.031)	0.958 (0.013)
		Hybrid	Original pixel	0.971 (0.013)	0.995 (0.003)
			DenseNet	0.912 (0.020)	0.985 (0.004)
	ResNet	Diversity	Original pixel	0.914 (0.040)	0.981 (0.010)
			ResNet	0.815 (0.050)	0.957 (0.015)
		Uncertainty	Original pixel	0.936 (0.030)	0.987 (0.007)
			ResNet	0.723 (0.036)	0.949 (0.012)
		Hybrid	Original pixel	0.962 (0.017)	0.993 (0.004)
			ResNet	0.861 (0.036)	0.972 (0.010)
50	/	Random	Original pixel	0.947 (0.042)	0.990 (0.010)
			TXRV	0.890 (0.085)	0.970 (0.027)
			REMEDIS	0.880 (0.075)	0.972 (0.029)
			DenseNet	0.880 (0.067)	0.972 (0.021)
			ResNet	0.843 (0.079)	0.963 (0.025)
	TXRV	Diversity	Original pixel	0.985 (0.007)	0.998 (0.001)
			TXRV	0.932 (0.024)	0.987 (0.005)
		Uncertainty	Original pixel	0.931 (0.025)	0.986 (0.006)
			TXRV	0.901 (0.036)	0.971 (0.017)
		Hybrid	Original pixel	0.990 (0.006)	0.998 (0.001)
			TXRV	0.963 (0.017)	0.993 (0.004)
	REMEDIS	Diversity	Original pixel	0.960 (0.014)	0.993 (0.003)
			REMEDIS	0.947 (0.017)	0.991 (0.004)
		Uncertainty	Original pixel	0.948 (0.027)	0.989 (0.006)
			REMEDIS	0.754 (0.029)	0.956 (0.011)
		Hybrid	Original pixel	0.972 (0.011)	0.996 (0.002)
			REMEDIS	0.915 (0.016)	0.988 (0.003)
	DenseNet	Diversity	Original pixel	0.970 (0.013)	0.995 (0.003)
			DenseNet	0.940 (0.021)	0.989 (0.005)
		Uncertainty	Original pixel	0.959 (0.024)	0.992 (0.005)
			DenseNet	0.767 (0.028)	0.961 (0.008)
		Hybrid	Original pixel	0.973 (0.011)	0.996 (0.002)
			DenseNet	0.875 (0.038)	0.961 (0.020)
	ResNet	Diversity	Original pixel	0.946 (0.026)	0.988 (0.007)

(Continues)

TABLE A2 | (Continued)

Initialization budget	Foundation model	Initialization method	Model input	AUROC	AUPRC
			ResNet	0.815 (0.047)	0.955 (0.016)
		Uncertainty	Original pixel	0.928 (0.026)	0.987 (0.006)
			ResNet	0.805 (0.039)	0.966 (0.009)
		Hybrid	Original pixel	0.978 (0.009)	0.997 (0.002)
			ResNet	0.872 (0.036)	0.974 (0.009)

TABLE A3 | Subsequent learning performance of random sampling, diversity, uncertainty, and hybrid methods on the Guangzhou data set.

Overall budget	Foundation model	Initialization method	Model input	AUROC	AUPRC
All samples	/	/	Original pixel	0.998 (0.001)	0.999 (0.000)
			TXRV	0.938 (0.007)	0.977 (0.003)
			REMEDIS	0.993 (0.003)	0.997 (0.002)
			DenseNet	0.848 (0.012)	0.939 (0.007)
			ResNet	0.990 (0.003)	0.996 (0.001)
10 + 10	/	Random	Original pixel	0.938 (0.032)	0.974 (0.014)
			TXRV	0.689 (0.090)	0.848 (0.052)
			REMEDIS	0.882 (0.083)	0.938 (0.057)
			DenseNet	0.853 (0.076)	0.927 (0.041)
			ResNet	0.869 (0.052)	0.938 (0.032)
	TXRV	Diversity	Original pixel	0.974 (0.004)	0.989 (0.003)
			TXRV	0.476 (0.022)	0.704 (0.019)
		Uncertainty	Original pixel	0.913 (0.008)	0.968 (0.004)
			TXRV	0.460 (0.016)	0.728 (0.018)
		Hybrid	Original pixel	0.908 (0.008)	0.965 (0.005)
			TXRV	0.602 (0.017)	0.806 (0.014)
	REMEDIS	Diversity	Original pixel	0.936 (0.006)	0.977 (0.003)
			REMEDIS	0.927 (0.008)	0.968 (0.005)
		Uncertainty	Original pixel	0.893 (0.011)	0.943 (0.009)
			REMEDIS	0.884 (0.012)	0.926 (0.012)
		Hybrid	Original pixel	0.970 (0.005)	0.988 (0.002)
			REMEDIS	0.872 (0.015)	0.910 (0.013)
	DenseNet	Diversity	Original pixel	0.956 (0.005)	0.981 (0.004)
			DenseNet	0.867 (0.014)	0.916 (0.012)
		Uncertainty	Original pixel	0.924 (0.008)	0.966 (0.005)
			DenseNet	0.747 (0.015)	0.864 (0.013)
		Hybrid	Original pixel	0.929 (0.008)	0.971 (0.004)
			DenseNet	0.832 (0.013)	0.916 (0.010)
	ResNet	Diversity	Original pixel	0.981 (0.004)	0.993 (0.002)
			ResNet	0.894 (0.010)	0.949 (0.007)
		Uncertainty	Original pixel	0.931 (0.007)	0.975 (0.003)
			ResNet	0.870 (0.012)	0.936 (0.010)
		Hybrid	Original pixel	0.946 (0.006)	0.979 (0.003)
			ResNet	0.895 (0.011)	0.948 (0.008)
10 + 20[a]	/	Random	Original pixel	0.950 (0.031)	0.980 (0.013)

(Continues)

TABLE A3 | (Continued)

Overall budget	Foundation model	Initialization method	Model input	AUROC	AUPRC
			TXRV	0.745 (0.058)	0.878 (0.035)
			REMEDIS	0.898 (0.073)	0.946 (0.044)
			DenseNet	0.887 (0.049)	0.944 (0.031)
			ResNet	0.896 (0.031)	0.953 (0.020)
	TXRV	Diversity	Original pixel	0.980 (0.004)	0.991 (0.003)
			TXRV	0.777 (0.016)	0.886 (0.013)
		Uncertainty	Original pixel	0.960 (0.005)	0.985 (0.002)
			TXRV	0.497 (0.015)	0.753 (0.017)
		Hybrid	Original pixel	0.964 (0.006)	0.987 (0.003)
			TXRV	0.740 (0.013)	0.883 (0.010)
	REMEDIS	Diversity	Original pixel	0.941 (0.007)	0.976 (0.004)
			REMEDIS	0.896 (0.011)	0.943 (0.009)
		Uncertainty	Original pixel	0.896 (0.012)	0.946 (0.007)
			REMEDIS	0.885 (0.010)	0.931 (0.011)
		Hybrid	Original pixel	0.969 (0.004)	0.989 (0.002)
			REMEDIS	0.915 (0.010)	0.935 (0.012)
	DenseNet	Diversity	Original pixel	0.973 (0.005)	0.990 (0.003)
			DenseNet	0.924 (0.008)	0.969 (0.004)
		Uncertainty	Original pixel	0.932 (0.007)	0.972 (0.004)
			DenseNet	0.756 (0.016)	0.870 (0.014)
		Hybrid	Original pixel	0.942 (0.007)	0.978 (0.003)
			DenseNet	0.928 (0.007)	0.970 (0.004)
	ResNet	Diversity	Original pixel	0.981 (0.004)	0.993 (0.002)
			ResNet	0.938 (0.008)	0.974 (0.004)
		Uncertainty	Original pixel	0.921 (0.008)	0.969 (0.004)
			ResNet	0.891 (0.010)	0.946 (0.009)
		Hybrid	Original pixel	0.963 (0.005)	0.987 (0.002)
			ResNet	0.903 (0.008)	0.957 (0.006)
10 + 30	/	Random	Original pixel	0.958 (0.026)	0.983 (0.011)
			TXRV	0.773 (0.054)	0.894 (0.032)
			REMEDIS	0.909 (0.051)	0.949 (0.038)
			DenseNet	0.910 (0.029)	0.958 (0.018)
			ResNet	0.908 (0.032)	0.959 (0.016)
	TXRV	Diversity	Original pixel	0.961 (0.005)	0.985 (0.002)
			TXRV	0.859 (0.011)	0.938 (0.008)
		Uncertainty	Original pixel	0.935 (0.007)	0.973 (0.004)
			TXRV	0.529 (0.018)	0.779 (0.015)
		Hybrid	Original pixel	0.944 (0.007)	0.979 (0.003)
			TXRV	0.748 (0.012)	0.890 (0.009)
	REMEDIS	Diversity	Original pixel	0.933 (0.008)	0.974 (0.004)
			REMEDIS	0.647 (0.023)	0.736 (0.017)
		Uncertainty	Original pixel	0.932 (0.008)	0.972 (0.004)
			REMEDIS	0.923 (0.008)	0.952 (0.011)

(Continues)

TABLE A3 | (Continued)

Overall budget	Foundation model	Initialization method	Model input	AUROC	AUPRC
		Hybrid	Original pixel	0.969 (0.005)	0.988 (0.003)
			REMEDIS	0.929 (0.008)	0.959 (0.009)
	DenseNet	Diversity	Original pixel	0.955 (0.005)	0.982 (0.003)
			DenseNet	0.900 (0.009)	0.961 (0.004)
		Uncertainty	Original pixel	0.954 (0.006)	0.980 (0.004)
			DenseNet	0.923 (0.009)	0.949 (0.010)
		Hybrid	Original pixel	0.924 (0.008)	0.971 (0.004)
			DenseNet	0.933 (0.008)	0.973 (0.004)
	ResNet	Diversity	Original pixel	0.977 (0.003)	0.991 (0.002)
			ResNet	0.942 (0.007)	0.977 (0.004)
		Uncertainty	Original pixel	0.949 (0.006)	0.981 (0.003)
			ResNet	0.939 (0.007)	0.976 (0.004)
		Hybrid	Original pixel	0.981 (0.003)	0.993 (0.001)
			ResNet	0.921 (0.007)	0.969 (0.004)
10 + 40	/	Random	Original pixel	0.964 (0.016)	0.986 (0.006)
			TXRV	0.787 (0.043)	0.901 (0.025)
			REMEDIS	0.913 (0.044)	0.950 (0.035)
			DenseNet	0.922 (0.024)	0.964 (0.014)
			ResNet	0.919 (0.027)	0.965 (0.013)
	TXRV	Diversity	Original pixel	0.970 (0.005)	0.987 (0.003)
			TXRV	0.847 (0.011)	0.929 (0.008)
		Uncertainty	Original pixel	0.977 (0.004)	0.990 (0.003)
			TXRV	0.588 (0.018)	0.803 (0.016)
		Hybrid	Original pixel	0.962 (0.006)	0.986 (0.002)
			TXRV	0.739 (0.014)	0.885 (0.010)
	REMEDIS	Diversity	Original pixel	0.948 (0.006)	0.981 (0.003)
			REMEDIS	0.866 (0.013)	0.942 (0.010)
		Uncertainty	Original pixel	0.944 (0.007)	0.978 (0.003)
			REMEDIS	0.928 (0.012)	0.938 (0.014)
		Hybrid	Original pixel	0.967 (0.005)	0.985 (0.004)
			REMEDIS	0.946 (0.009)	0.956 (0.010)
	DenseNet	Diversity	Original pixel	0.980 (0.004)	0.991 (0.002)
			DenseNet	0.932 (0.009)	0.968 (0.006)
		Uncertainty	Original pixel	0.964 (0.005)	0.985 (0.004)
			DenseNet	0.945 (0.006)	0.977 (0.004)
		Hybrid	Original pixel	0.983 (0.003)	0.993 (0.002)
			DenseNet	0.917 (0.007)	0.969 (0.003)
	ResNet	Diversity	Original pixel	0.976 (0.003)	0.991 (0.002)
			ResNet	0.938 (0.009)	0.967 (0.008)
		Uncertainty	Original pixel	0.954 (0.006)	0.983 (0.003)
			ResNet	0.943 (0.006)	0.978 (0.003)
		Hybrid	Original pixel	0.977 (0.004)	0.991 (0.002)
			ResNet	0.922 (0.008)	0.966 (0.006)

[a]The overall budget consists of both the initialization and subsequent learning components. For instance, a budget of 10 + 20 indicates an initialization budget comprising 10 samples and a subsequent learning budget also comprising 20 samples.

TABLE A4 | Subsequent learning performance of random sampling, diversity, uncertainty, and hybrid methods on the Pakistan data set.

Overall budget	Foundation model	Initialization method	Model input	AUROC	AUPRC
All samples	/	/	Original pixel	0.994 (0.003)	0.999 (0.001)
			TXRV	0.991 (0.006)	0.999 (0.001)
			REMEDIS	0.968 (0.016)	0.994 (0.003)
			DenseNet	0.964 (0.023)	0.987 (0.010)
			ResNet	0.993 (0.005)	0.999 (0.001)
10 + 10	/	Random	Original pixel	0.838 (0.125)	0.963 (0.037)
			TXRV	0.795 (0.135)	0.948 (0.038)
			REMEDIS	0.791 (0.176)	0.947 (0.057)
			DenseNet	0.743 (0.103)	0.936 (0.033)
			ResNet	0.757 (0.107)	0.943 (0.031)
	TXRV	Diversity	Original pixel	0.928 (0.018)	0.988 (0.004)
			TXRV	0.574 (0.046)	0.894 (0.032)
		Uncertainty	Original pixel	0.899 (0.030)	0.979 (0.008)
			TXRV	0.866 (0.034)	0.959 (0.021)
		Hybrid	Original pixel	0.945 (0.018)	0.991 (0.003)
			TXRV	0.863 (0.045)	0.950 (0.024)
	REMEDIS	Diversity	Original pixel	0.792 (0.049)	0.958 (0.013)
			REMEDIS	0.872 (0.045)	0.960 (0.019)
		Uncertainty	Original pixel	0.739 (0.046)	0.948 (0.013)
			REMEDIS	0.814 (0.047)	0.939 (0.024)
		Hybrid	Original pixel	0.904 (0.033)	0.981 (0.008)
			REMEDIS	0.965 (0.011)	0.994 (0.002)
	DenseNet	Diversity	Original pixel	0.873 (0.032)	0.975 (0.009)
			DenseNet	0.845 (0.048)	0.950 (0.023)
		Uncertainty	Original pixel	0.910 (0.044)	0.976 (0.012)
			DenseNet	0.746 (0.045)	0.924 (0.027)
		Hybrid	Original pixel	0.847 (0.041)	0.971 (0.010)
			DenseNet	0.830 (0.028)	0.970 (0.007)
	ResNet	Diversity	Original pixel	0.826 (0.054)	0.953 (0.020)
			ResNet	0.848 (0.036)	0.970 (0.008)
		Uncertainty	Original pixel	0.930 (0.028)	0.986 (0.007)
			ResNet	0.796 (0.046)	0.959 (0.011)
		Hybrid	Original pixel	0.940 (0.016)	0.990 (0.004)
			ResNet	0.880 (0.030)	0.979 (0.007)
10 + 20[a]	/	Random	Original pixel	0.889 (0.121)	0.976 (0.033)
			TXRV	0.860 (0.101)	0.966 (0.028)
			REMEDIS	0.840 (0.114)	0.963 (0.033)
			DenseNet	0.787 (0.093)	0.949 (0.030)
			ResNet	0.803 (0.098)	0.955 (0.028)
	TXRV	Diversity	Original pixel	0.956 (0.014)	0.993 (0.003)
			TXRV	0.869 (0.040)	0.972 (0.010)
		Uncertainty	Original pixel	0.904 (0.029)	0.980 (0.008)
			TXRV	0.881 (0.031)	0.972 (0.013)

(Continues)

TABLE A4 | (Continued)

Overall budget	Foundation model	Initialization method	Model input	AUROC	AUPRC
		Hybrid	Original pixel	0.954 (0.013)	0.992 (0.003)
			TXRV	0.933 (0.031)	0.983 (0.010)
	REMEDIS	Diversity	Original pixel	0.963 (0.012)	0.994 (0.002)
			REMEDIS	0.807 (0.042)	0.954 (0.015)
		Uncertainty	Original pixel	0.790 (0.048)	0.959 (0.012)
			REMEDIS	0.904 (0.037)	0.975 (0.012)
		Hybrid	Original pixel	0.920 (0.033)	0.984 (0.007)
			REMEDIS	0.909 (0.027)	0.981 (0.008)
	DenseNet	Diversity	Original pixel	0.880 (0.029)	0.978 (0.007)
			DenseNet	0.917 (0.020)	0.987 (0.004)
		Uncertainty	Original pixel	0.929 (0.034)	0.984 (0.008)
			DenseNet	0.731 (0.052)	0.912 (0.028)
		Hybrid	Original pixel	0.743 (0.061)	0.941 (0.019)
			DenseNet	0.800 (0.043)	0.951 (0.017)
	ResNet	Diversity	Original pixel	0.976 (0.008)	0.996 (0.001)
			ResNet	0.886 (0.037)	0.962 (0.016)
		Uncertainty	Original pixel	0.936 (0.028)	0.987 (0.007)
			ResNet	0.860 (0.033)	0.975 (0.007)
		Hybrid	Original pixel	0.963 (0.011)	0.994 (0.002)
			ResNet	0.870 (0.036)	0.973 (0.008)
10 + 30	/	Random	Original pixel	0.924 (0.080)	0.985 (0.018)
			TXRV	0.885 (0.065)	0.971 (0.022)
			REMEDIS	0.881 (0.083)	0.974 (0.024)
			DenseNet	0.819 (0.086)	0.955 (0.029)
			ResNet	0.829 (0.085)	0.962 (0.024)
	TXRV	Diversity	Original pixel	0.967 (0.012)	0.995 (0.002)
			TXRV	0.910 (0.034)	0.970 (0.018)
		Uncertainty	Original pixel	0.911 (0.032)	0.981 (0.008)
			TXRV	0.868 (0.037)	0.957 (0.021)
		Hybrid	Original pixel	0.932 (0.019)	0.988 (0.004)
			TXRV	0.947 (0.033)	0.981 (0.014)
	REMEDIS	Diversity	Original pixel	0.967 (0.011)	0.995 (0.002)
			REMEDIS	0.861 (0.037)	0.967 (0.014)
		Uncertainty	Original pixel	0.874 (0.041)	0.973 (0.010)
			REMEDIS	0.728 (0.055)	0.910 (0.030)
		Hybrid	Original pixel	0.918 (0.030)	0.985 (0.007)
			REMEDIS	0.705 (0.058)	0.898 (0.031)
	DenseNet	Diversity	Original pixel	0.899 (0.030)	0.978 (0.009)
			DenseNet	0.885 (0.029)	0.978 (0.008)
		Uncertainty	Original pixel	0.910 (0.038)	0.979 (0.009)
			DenseNet	0.841 (0.031)	0.975 (0.006)
		Hybrid	Original pixel	0.892 (0.032)	0.978 (0.010)
			DenseNet	0.780 (0.044)	0.951 (0.015)

(Continues)

TABLE A4 | (Continued)

Overall budget	Foundation model	Initialization method	Model input	AUROC	AUPRC
	ResNet	Diversity	Original pixel	0.973 (0.009)	0.996 (0.002)
			ResNet	0.845 (0.039)	0.958 (0.019)
		Uncertainty	Original pixel	0.921 (0.028)	0.985 (0.006)
			ResNet	0.843 (0.042)	0.952 (0.020)
		Hybrid	Original pixel	0.961 (0.011)	0.994 (0.002)
			ResNet	0.879 (0.027)	0.979 (0.006)
10 + 40	/	Random	Original pixel	0.943 (0.042)	0.989 (0.009)
			TXRV	0.906 (0.050)	0.976 (0.019)
			REMEDIS	0.867 (0.100)	0.969 (0.034)
			DenseNet	0.846 (0.068)	0.961 (0.025)
			ResNet	0.854 (0.062)	0.969 (0.018)
	TXRV	Diversity	Original pixel	0.951 (0.015)	0.992 (0.003)
			TXRV	0.923 (0.018)	0.987 (0.004)
		Uncertainty	Original pixel	0.911 (0.025)	0.983 (0.007)
			TXRV	0.836 (0.043)	0.950 (0.020)
		Hybrid	Original pixel	0.919 (0.028)	0.985 (0.006)
			TXRV	0.961 (0.020)	0.993 (0.004)
	REMEDIS	Diversity	Original pixel	0.950 (0.017)	0.992 (0.003)
			REMEDIS	0.920 (0.032)	0.985 (0.006)
		Uncertainty	Original pixel	0.865 (0.034)	0.973 (0.009)
			REMEDIS	0.837 (0.035)	0.966 (0.014)
		Hybrid	Original pixel	0.932 (0.026)	0.987 (0.006)
			REMEDIS	0.959 (0.013)	0.993 (0.002)
	DenseNet	Diversity	Original pixel	0.938 (0.022)	0.990 (0.004)
			DenseNet	0.942 (0.017)	0.990 (0.004)
		Uncertainty	Original pixel	0.935 (0.026)	0.986 (0.007)
			DenseNet	0.849 (0.030)	0.973 (0.007)
		Hybrid	Original pixel	0.970 (0.013)	0.995 (0.002)
			DenseNet	0.851 (0.035)	0.972 (0.008)
	ResNet	Diversity	Original pixel	0.980 (0.009)	0.997 (0.002)
			ResNet	0.934 (0.016)	0.989 (0.003)
		Uncertainty	Original pixel	0.935 (0.023)	0.988 (0.005)
			ResNet	0.885 (0.032)	0.979 (0.007)
		Hybrid	Original pixel	0.982 (0.007)	0.997 (0.001)
			ResNet	0.861 (0.034)	0.974 (0.008)

[a]The overall budget consists of both the initialization and subsequent learning components. For instance, a budget of 10 + 20 indicates an initialization budget comprising 10 samples and a subsequent learning budget comprising 20 samples.