

# BenchMD: A Benchmark for Modality-Agnostic Learning on Medical Images and Sensors

Kathryn Wantlin<sup>\*</sup>  
Princeton University

Chenwei Wu<sup>†</sup>  
Harvard University

Shih-Cheng Huang<sup>†</sup>  
Stanford University

Oishi Banerjee  
Harvard Medical School

Farah Dadabhoy  
Massachusetts General Hospital

Veeral Vipin Mehta  
Stony Brook University Hospital

Ryan Wonhee Han  
Stanford University

Fang Cao  
Stanford University

Raja R. Narayan  
Harvard Medical School

Errol Colak  
University of Toronto

Adewole Adamson  
University of Texas at Austin

Laura Heacock  
NYU Langone

Geoffrey H. Tison  
University of California, San Francisco

Alex Tamkin<sup>†</sup>  
Stanford University

Pranav Rajpurkar<sup>†</sup>  
Harvard Medical School

## Abstract

Recent advances in transformers and self-supervised learning (SSL) offer remarkable versatility and can be applied flexibly across modalities, while also improving generalization performance across distributions within each modality. However, the benchmarks used for evaluating such advances have been criticized for detaching tasks from real-world context. Furthermore, the medical domain poses a particular challenge for current modality-agnostic methods, due to the heterogeneity of modalities produced by dozens of different technologies, the frequent distribution shifts, and the scarcity of data and labels. To address these problems, we present BenchMD, a benchmark that tests how modality-agnostic methods, including both architectures and training techniques (e.g. SSL, ImageNet pretraining), perform on medical tasks. BenchMD combines 19 publicly available datasets for 7 diverse medical modalities, including 1D sensor data, 2D images, and 3D volumetric scans. Reflecting real-world label constraints, our benchmark tests model performance across multiple settings where different amounts of labels are available, including challenging few-

shot settings that incentivize the use of SSL. We also evaluate performance on out-of-distribution data collected at different hospitals than the training data, representing naturally occurring distribution shifts that frequently degrade the performance of medical AI models. BenchMD offers an easy-to-use benchmark for identifying high-performing, modality-agnostic methods with the potential to transform medicine. Additionally, our initial baselines demonstrate mixed results, with no technique achieving strong performance across all modalities. Thus, there is ample room for improvement on this benchmark.

## 1. Introduction

Recent advances in transformers and self-supervised learning (SSL) have enabled state-of-the-art performance across many modalities, including images and videos [19]. These methods offer remarkable versatility, reduce the need for labeled data, and can be applied flexibly across modalities, while also improving generalization performance across distributions within each modality. Benchmarks that go beyond 2D images and include tasks across multiple modalities are needed to measure progress in this

<sup>\*</sup>kw2960@princeton.edu

<sup>†</sup>Equal Authorship

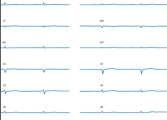
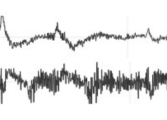

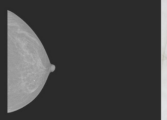
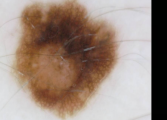
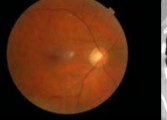
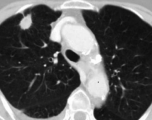
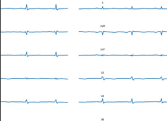
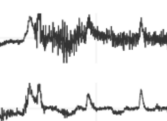

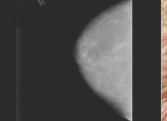

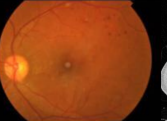
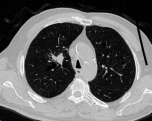
	1D		2D				3D
Modality	ECG	EEG	CXR	MAMMO	DERM	FUNDUS	LDCT
Task	Arrhythmia classification (multiclass)	Sleep stage classification (multiclass)	Abnormality Detection (multilabel)	Lesion Grading (multiclass)	Lesion Classification (multiclass)	Diabetic Retinopathy Grading (multiclass)	Lung Nodule Detection
Label Support	NORM, CD, HYP, MI, STTC, AF, Other	REM, N1, N2, N3, WAKE	Atelectasis, Cardiomegaly, Consolidation, Edema, Pleural Effusion	BIRADS Score (1-5)	AKIEC, BCC, MEL, NEV, OTHER	ICDR Grade (0-4)	Positive, Negative
Distribution Shift	Technology, Demographics, Annotations						
Train Data (# examples)	PTB-XL (21.8k)	SHHS (5.8k)	MIMIC-CXR (10k)	VinDR-Mammo (20k)	BCN-20000 (19k)	Messidor 2 (1.4k)	LIDC-IDRI (1.3k)
Train Example							
Out of Distribution Test Data (# examples)	Chapman (10k) Georgia (10.3k) CPSC 2020 (6.9k)	ISRUC (126)	VINDR-CXR (18k) CheXpert (224k)	CBIS-DDM (595)	HAM10000 (10k) PAD-UFES-20 (2.3k)	APTOS 2019 (733) Jinchi Medical University (2k)	LNDb (229)
Test Example							

Figure 1. The BenchMD benchmark consists of 19 real-world medical datasets across 7 medical modalities. Successful methods will achieve high performance when evaluated on out-of-distribution data.

space. However, previous benchmarks have been criticized for detaching tasks from real-world context, with problem formulations that are not grounded in the understanding of domain experts [12, 40, 44]. This problem has been specifically noted in the medical domain, where existing benchmarks often address questions that are synthetic or otherwise clinically irrelevant [10].

To address this gap, we propose BenchMD (A **Benchmark** across Medical **M**odalities and **D**istributions), a new modality-agnostic learning benchmark grounded in real-world interpretation tasks and distribution shifts. BenchMD tests different modality-agnostic methods, including both architectures and training techniques (e.g. SSL, ImageNet pretraining), on 19 datasets for 7 medical modalities. The wide variety of modalities reflects the heterogeneity of medical image and sensor data, which can be produced by dozens of different technologies [55]. Specifically, we evaluate methods using 1D data from electrocardiogram (ECG) and electroencephalogram (EEG) sensors, 2D image data from chest X-rays (CXR), mammograms, dermoscopic images, and fundus images, and 3D volumetric data from low-dose computed tomography (LDCT) scans (see Fig. 1). Current methods are often specialized for these different modalities; for example, contrastive learning techniques typically require modality-specific data augmentations [33]. In contrast, we encourage the development of flexible, **modality-agnostic** methods that can be applied out-of-the-box without customization.

We construct our benchmark to enable advances on two

additional fronts. First, label shortages have historically posed a serious obstacle to model development, so our benchmark tests performance under severe **data scarcity**, incentivizing the use of SSL techniques that exploit unlabeled data. In order to explore the label efficiency of different methods, we assess performance across multiple settings where increasingly small subsets of the source dataset have labels. Second, we investigate how models perform under **naturally-occurring distribution shifts**, such as when they are trained on data from one hospital and deployed in another. To this end, we train models on one in-distribution (ID) source dataset and then test zero-shot transfer performance on out-of-distribution (OOD) data from unseen target datasets collected at different hospitals.

We present BenchMD as an easy-to-use benchmark for assessing performance widely across medical modalities and distributions. To make using BenchMD **simple**, we standardize preprocessing steps and validation metrics (details in the Appendix), so users simply need to plug in new architectures and training tasks. Additionally, we use only **publicly available datasets**, allowing users to easily access BenchMD and replicate results. Using our benchmark, we provide initial baselines that demonstrate significant variations in performance, with no technique achieving strong results across all modalities. These results offer significant room for improvement on this benchmark, and we discuss possible directions for future work. We expect our work will accelerate the development of versatile methods for medicine and provide a valuable tool for measuring ad-

vances in modality-agnostic methods.

## 2. Related Work

### Modality-Agnostic Techniques for Natural Images:

Recent advances in deep learning have produced methods that enable high performance and can be flexibly applied across modalities. Self-supervised learning (SSL) techniques such as masked data modeling [21] and contrastive learning [34,51] can be used to learn from unlabeled datasets across many modalities. Architectures are also increasingly able to take in different modalities as input, producing models that can interchangeably process 2D images and 3D videos [5,36]. Benchmarks like ours can rigorously evaluate these new methods, assessing their performance on real-world tasks across multiple modalities.

**Modality-Agnostic Medical AI:** There have been limited efforts to unify architectures and training techniques across medical image modalities in particular [20,57,60]. For example, Zhou *et al.* recently found that self-supervised MAE pretraining on medical images offered better performance than ImageNet pretraining when interpreting chest X-rays, MRI scans, and CT scans [59]. Similarly, Azizi *et al.* found that a training procedure combining supervised learning on natural images with SSL pretraining on medical images offered high performance on 6 2D medical image modalities [8]. BenchMD tracks progress in this area and includes an unprecedented range of medical modalities, addressing a range of 1D, 2D, and 3D medical images and sensors.

**Existing Benchmarks for Multiple Modalities:** Our work extends the line of thinking in the DABS benchmarks, which evaluate the performance of modality-agnostic SSL techniques across modalities ranging from text and genomics to X-rays and wearable sensor data [50,51]. We also take inspiration from the WILDS benchmarks, which evaluate performance on out-of-distribution data within several modalities [32,46,56]. BenchMD combines the strengths of both approaches, creating a benchmark that is rooted in real-world modalities with direct clinical applicability: evaluating whether techniques generalize well across modalities as well as to new distributions within each modality. Furthermore, while some of DABS’s training datasets are unlabeled, we, similar to WILDS, consistently provide labeled source data to facilitate comparison against non-SSL techniques such as supervised learning. Unlike both DABS and WILDS, BenchMD tests model performance across settings with few-shot learning, exploring how label availability affects performance. Our work also differs from both DABS and WILDS because we focus on real-world medical tasks and cover a broader range of medical modalities, including 3D volumetric scans.

## 3. Modalities and Datasets

We have curated a list of high-impact modalities and selected source and target datasets for evaluating out-of-distribution (OOD) performance. Each modality we present in this benchmark is used to test for prevalent diseases and significantly contributes to clinician workloads in current practice [4,6,14,25,27,37,47,61]. For each modality, we select a highly-cited, large dataset as the source dataset, which we use for training. We also choose labeled target datasets, which we use to test performance on OOD data. We design a set of tasks that are both clinically relevant and unified across source and target datasets.

### 3.1. Electrocardiograms

12-lead electrocardiogram (ECG) measures the three-dimensional electrical activity of the heart over time using electrodes placed on the skin. Classifying cardiovascular abnormalities from ECGs is challenging because there are 12 1D channels, each corresponding to a different spatial axis, and because diagnosis requires distinguishing irregular cardiovascular signals from noisy data.

**Task** We perform a single-label classification task on 5 second recordings of 12-channel ECG signals with a sampling rate of 500Hz. We use a set of 7 labels unified across datasets derived at the discretion of medical experts: Normal, Conduction Disturbance, Hypertrophy, Myocardial Infarction, Ischemic ST-T Changes, Atrial fibrillation/atrial flutter, and Other. We consider four publicly available 12-lead ECG-waveform datasets: PTB-XL (source) [22,53,54], Chapman-Shaoxing (target) [58], Georgia 12-Lead ECG Challenge (target) [23], and China Physiological Signal Challenge (CPSC, target) [13].

**Distribution Shifts** *Demographics:* PTB-XL’s data was collected between 1989 and 1996, Chapman-Shaoxing’s in 2020, CPSC’s in 2018, and Georgia’s in 2020. The Chapman-Shaoxing and CPSC datasets’ patients are based in China, the Georgia datasets’ in the southeastern United States, and the PTB-XL datasets’ in Germany.

*Collection Technology:* The PTB-XL dataset used devices provided by Schiller AG, while the Chapman-Shaoxing dataset was collected using devices from Zhejiang Cachet Jetboom Medical Devices.

*Annotation Details:* Although we group abnormalities into 7 categories that are consistent across datasets, different datasets provide varying levels of additional granularity in their labels, with different label distributions across datasets.

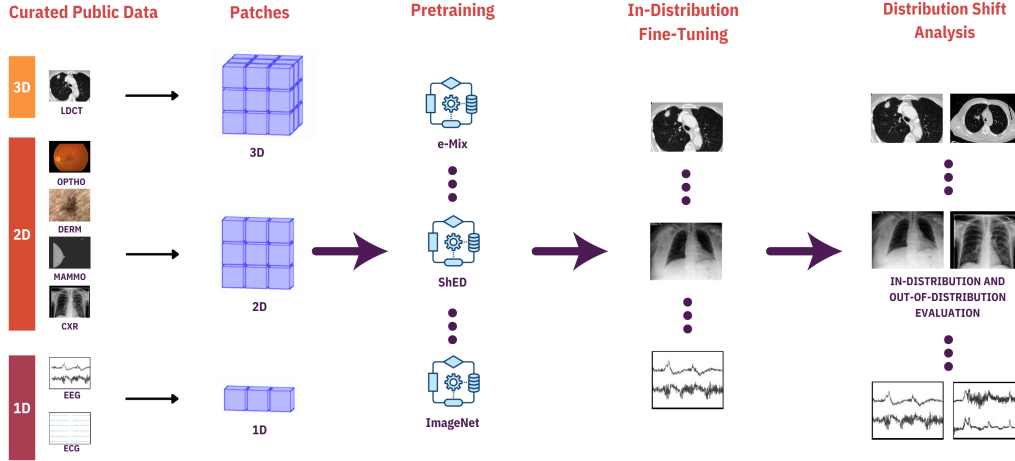


Figure 2. Models for each modality are first trained on a source dataset, using modality-agnostic methods. They are then evaluated on out-of-distribution data from one or more target datasets.

### 3.2. Electroencephalograms

Electroencephalograms (EEG) measure multi-channel 1D signals of electrical activity in the brain and are used to diagnose sleep and seizure disorders [9]. Noise and intra-class variability in sampling rates, signal quality, the number of leads used, and the length of the captured rhythm makes distinguishing sleep stages difficult in EEGs.

**Task** We perform a single-label sleep stage classification task on 2 traditional central derivations channels (C3 and C4), 125 Hz, 30 second EEG signal inputs. We use the American Academy of Sleep Medicine’s standard 5 labels: Wake, Rapid Eye Movement, Non-REM Stage 1, Non-REM stage 2, and Non-REM stage 3 [18]. We consider two publicly available datasets: the Sleep Heart Health Study (SHHS) dataset (source) [43] and the ISRUC-Sleep dataset (target) [31].

**Distribution Shifts Demographics:** The SHHS dataset was collected between 1995-1998, while the ISRUC dataset was collected between 2009–2013. The SHHS dataset includes 5,804 adults aged 40 and older, while the ISRUC dataset was collected from subjects whose ages range from 20 years old to 85 years old, with an average age of 51. Moreover, the ISRUC dataset was collected from a hospital in Coimbra, Portugal, while SHHS was collected by the National Heart Lung & Blood Institute in the US.

**Collection Technology:** The SHHS dataset was collected at a sampling rate of 125 Hz while ISRUC was collected at 150 Hz. SHHS was also collected from studies conducted in patient homes, while ISRUC was collected in a hospital setting.

### 3.3. Chest X-Rays

Chest X-rays are 2D grayscale projection radiographs of a patient’s heart, lungs, blood vessels, airways, chest bones, and spine and are crucial for the diagnosis of cardiovascular diseases such as atelectasis and edema. Chest X-ray classification is uniquely challenging compared to natural image classification since radiographs are grayscale and always have similar frontal or lateral spatial structures, with relevant abnormalities only occurring in a small region of the image [30].

**Task** For this modality, we are performing a single-label classification task on 2D grayscale chest x-rays using 5 prevalent labels: Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion. We utilize three publicly available datasets: MIMIC-CXR (source) [23, 28, 29], CheXpert (target) [27], and VinDr-CXR (target) [38].

**Distribution Shifts Demographics:** MIMIC-CXR’s images were collected between 2011 and 2016, CheXpert’s between 2002 and 2017, and VinDr-CXR’s between 2018 and 2020. CheXpert’s data was collected by Stanford University School of Medicine, MIMIC-CXR’s from the Beth Israel Deaconess Medical Center in Boston, and VinDr-CXR’s from the Hanoi Medical University Hospital and Hospital 108 in Vietnam.

**Annotation Details:** MIMIC-CXR and CheXpert depends on automated natural language labelers, while VinDr-CXR uses radiologist-generated annotations.



### 3.4. Mammograms

Mammograms consist of 2D grayscale images of the cranio-caudal (CC) view and the mediolateral-oblique (MLO) view of the left and right breast of a patient (4 images possible per patient), and are the main imaging tool for the screening and diagnosis of breast cancer [11]. Mammograms are high-resolution images with millions of pixels, and while the breast views are highly standardized, disease classification depends on abnormalities in small regions of interest, making diagnosis challenging for AI models.

**Task** We perform a single-label task of predicting the Breast Imaging Reporting and Data System (BI-RADS) assessment category (from 1 to 5) for each breast image. We consider two datasets: VinDr-Mammo (source) [39] and CBIS-DDSM (target) [15, 35, 48].

**Distribution Shifts** *Demographics:* VinDr-Mammo was compiled from a pool of mammography examinations taken between 2018 and 2020, while CBIS-DDSM was compiled from exams conducted between 1988 and 1999. VinDr-Mammo exams were collected by hospitals in Vietnam, while CBIS-DDSM’s were collected by United States hospitals.

*Collection Technology:* VinDr-Mammo contains full-field digital mammogram images while CBIS-DDSM contains scanned film mammogram images. Several different scanners from multiple manufacturers were used to collect the CBIS-DDSM mammograms.

*Annotation Details:* VinDr-Mammo contains some lesionless images, which still have a breast-level BI-RADS score. The CBIS-DDSM dataset, however, exclusively contains images with lesions, and does not annotate breast-level BI-RADS scores. To test our model on this target dataset, for each breast we use the maximum of lesion-level BI-RADS scores as the breast-level BI-RADS score.

### 3.5. Dermoscopic Images

Dermoscopy produces 2D RGB images showing subsurface skin structures in the epidermis, at the dermoepidermal junction, and in the papillary dermis, and is used to assess cancer in skin lesions. Performing tasks on dermoscopic images is complicated by intraclass variability in lesion texture, scale, and color due to presence of different skin colors, hair, veins, and irregular lesion borders [25].

**Task** We perform a single-label classification of 2D RGB dermoscopy images across 5 unified labels extracted by clinicians: “AKIEC” (includes actinic keratoses, intraepithelial carcinoma, and squamous cell carcinoma as all of these are with the continuum of squamous cell carcinoma), “BCC” (basal cell carcinoma), “MEL” (melanoma),

“NEV” (nevus), and “Other diseases” ( dermatofibroma, etc). We utilize three publicly available datasets: BCN 20000 (source) [16], HAM 10000 (target) [52], PAD-UFES-20 Smartphone image-set (target) [41].

**Distribution Shifts** *Demographics:* BCN20000’s images were collected from 2010 to 2016, PAD-UEFS-20’s from 2020, and HAM10000’s from the past 20 years. PAD-UEFS-20’s images were collected by hospitals in Brazil, HAM10000’s in Austria and Australia, and BCN20000’s in Spain.

*Collection Technology:* BCN20000 and HAM10000 images were collected using dermatoscopes, while PAD-UFES-20 images were collected by smartphone cameras.

*Annotation Details:* Although we grouped the abnormality annotations across datasets into 5 general categories, the granularity within each label varies depending on the dataset. For example, the “Other diseases” category for HAM10000 includes benign keratosis-like lesions while BCN20000’s doesn’t.

### 3.6. Fundus Images

Eye fundus images are 2D RGB images showing the interior surface of a single eye, including the retina, fovea, optic disc, macula, and posterior pole, and are crucial for the diagnosis of diabetic retinopathy (DR). The detection of DR is complicated by spurious correlations with other undetected conditions such as diabetic macular edema [4].

**Task** For each 2D RGB fundus image, we perform the single-label task of predicting the severity of diabetic retinopathy (DR) in an image of each eye. We use the International Clinic Diabetic Retinopathy (ICDR) classification scale, which classifies DR on a five-stage severity scale from 0-4 [45]. We consider three datasets: Messidor-2 (source) [3, 17], APTOS 2019 (target) [1, 2], and the Jinchi Medical University dataset (target) [49].

**Distribution Shifts** *Demographics:* Messidor-2 images were collected from 2004 to 2010, while Jinchi Medical University images were collected between May 2011 and June 2015. The total collection period for APTOS 2019 is unknown. The Messidor2 data was collected from French institutions, APTOS 2019 data from the Aravind Eye Care System in India, and the Jinchi Medical University data from Japan. The Messidor-2 and Jinchi Medical University datasets consist of high quality retinal images, while APTOS 2019 exhibits more variation in data quality, including images with artifacts.

*Collection Technology:* Messidor-2 training images were taken with a Topcon TRC NW6 non-mydratic camera. The Jinchi Medical University dataset also uses a non-mydratic

camera, but a different model (AFC-230). The APTOS dataset contains images taken from both mydriatic (lower-quality imaging requiring pupil dilation) and non-mydriatic cameras, with the full range of camera models unknown. The Jinchi Medical University dataset was collected in a single-site, exploratory study performed in an institutional setting, whereas the other datasets contain images taken in clinical settings for diagnostic purposes.

*Annotation Details:* Jinchi Medical University, unlike Messidor-2 and APTOS, consolidates the similar ICDR classes 1 and 2 into a single superclass, termed a modified Davis grading. This creates an easier classification task than what we train our models for on the source dataset.

### 3.7. Low Dose Computed Tomography Scans

Low dose computed tomography (LDCT) is a procedure that uses an x-ray machine linked with a computer to create 3D images of a patient’s tissues and organs. LDCT is typically used to detect early-stage nodules of lung cancer in high-risk patients. The LDCT nodule classification task is challenging since LDCT scans are 3D images originally recorded in single-channel Hounsfield units with varying numbers of slices between patients. In addition, while scans have a large field of view with hundreds of slices, nodules only occupy a small volume of the scan, especially in early cancer stages [26].

**Task** Inputs are partitioned by sliding windows, representing 24 CT slices in single channel Hounsfield units. We perform two binary classification tasks, determining 1) whether a small nodule (diameter  $\leq 3\text{mm}$ ) exists in the current CT scan window and 2) whether a large nodule (diameter  $\leq 3\text{mm}$ ) exists in the current CT scan window. A sliding window is labeled positive for a nodule of either type if it contains more than 4 consecutive slices with positive labels. We utilize two public datasets: LIDC-IDRI (source) [7] and LNDb (target) [42].

**Distribution Shifts** *Demographics:* LIDC scans were collected in 2010, while LNDb scans were collected from 2016-2018. LIDC was collected from academic centers and medical imaging companies in the United States, while LNDb was collected at the Centro Hospitalar e Universitário de São João (CHUSJ) in Porto, Portugal.

*Collection Technology:* LIDC dataset collection involved a variety of scanner manufacturers and models, while the LNDb dataset was primarily collected by Siemens scanners. LIDC’s data was collected using a mean tube current of 222.1mA, while LNDb use a mean tube current of 161.9mA. The LIDC dataset includes slice thicknesses ranging from 0.6mm to 5mm, while the LNDb dataset has excluded CT scans where intravenous contrast had been used and those with a slice thickness greater than 1mm.

## 4. Experiments

We evaluate the performance of five baseline techniques: three SSL algorithms, ImageNet pretraining, and training from scratch. We then test performance on OOD target datasets using multiple transfer learning schemes.

### 4.1. Architecture

Following [51], we utilize a modality-agnostic transformer architecture across all experiments. We use separate 1D, 2D, and 3D embedding modules, which make minimal assumptions about the data and map all inputs to the same 256-dimensional embedding space, allowing users to mix inputs with different input dimensions. The transformer encoder is based on a standard ViT architecture, and we choose patch sizes to keep the number of embeddings similar across all datasets.

### 4.2. Pretraining

We evaluated the performance of three different SSL algorithms in this benchmark. The first two, Contrastive Embedding-Mixup (e-Mix) and Shuffled Embedding Prediction (ShED), follow [51]. e-Mix is a contrastive objective that additively mixes a batch of original input embeddings, weighting them with different coefficients. It then trains an encoder to produce a vector for a mixed embedding that is close to the original inputs’ embeddings in proportion to their mixing coefficients. ShED shuffles a fraction (0.85 in our experiments) of embeddings and trains the encoder with a classifier to predict which embeddings were perturbed. We also use a third Masked Autoencoding (MAE) objective, which masks a given fraction (0.75 in our experiments) of input embeddings and trains models to reconstruct them [24]. We standardize the pretraining process, running it for 100k steps with the Adam optimizer, learning rate  $1\text{e-}4$ , weight decay  $1\text{e-}4$ , and momentum 0.9.

Beyond SSL, we evaluate two other techniques. First, we train a modality-agnostic model from **scratch**, performing no pretraining. In addition, for 2D modalities, we evaluate models pretrained on ImageNet.

### 4.3. Transfer Learning and Out-of-Distribution Evaluation

We train models for particular tasks through both linear evaluation and finetuning, using labeled data from our in-distribution source datasets. We then evaluate zero-shot performance on OOD target datasets.

For linear evaluation, we freeze the model backbone and train a linear classifier head for the modality task using 100% of the source data labels. For finetuning, we run one set of experiments using 100% of the source labels but also test performance while varying label availability. For single-label tasks, we run experiments using 8, 64, or 256

labels per class, which we refer to as small, medium, and large label fractions, respectively. If the source dataset contains a class for which we have insufficient labels, we simply use all available labels for that class. For multi-label tasks, we create small/medium/large label sets by iterating through each class label and sampling labeled examples until we have 8, 64, or 256 labels for that class or have exhausted the available examples for that class. During both linear evaluation and finetuning, we train for 100 epochs with the same hyperparameters as in pretraining.

We then evaluate zero-shot transfer performance on OOD target datasets. After every epoch of linear evaluation or finetuning, we check the current model checkpoint’s performance on the source dataset’s validation set in order to perform model selection. We identify the top-performing checkpoint that achieves the highest average AUROC across tasks on the source validation set and report this AUROC in Fig. 3 under “In Distribution”. Next, we perform zero-shot transfer with this top-performing checkpoint by directly evaluating it on the target dataset(s), without any further training. We report AUROC, averaged across tasks and target datasets, in Fig. 3, which shows how techniques perform on OOD data for each modality.

#### 4.4. Results

Fig. 3 shows the performance of different techniques on the ID validation set and on our OOD test data. Fig. 4 shows how well different techniques perform and generalize, compared to training from scratch. We calculate the difference in performance between our pretraining techniques and training from scratch and compare how distribution shifts affect this difference. Our results are generally mixed.

**Does any SSL technique offer high performance across modalities?** No. e-Mix typically offers middling performance on OOD data; it outperforms other techniques on only on a few scattered experiments across ECG data, dermoscopic images and retinal fundus images. ShED is more promising, with particularly strong performance on ECG data and LDCT scans. However, it fails to maintain consistent performance across other modalities such as fundus images and mammograms, where ImageNet pretraining outperforms it across nearly all settings. Similarly, MAE achieves strong results on many experiments, especially on EEG data and on dermoscopic images from the PAD-UFES-20 dataset, it also performs poorly elsewhere. Despite being a top performer on EEG data, MAE achieves poor AUROC scores on ECG data, particularly the Georgia 12-Lead ECG Challenge and Chapman-Shaoxing datasets. This discrepancy indicates the difficulty of developing a single, high-performing technique even across different 1D sensor modalities. We see similar inconsistencies across different 2D image modalities, where no one SSL technique main-

tains high performance across OOD experiments. Future work may explore whether other SSL techniques or architectures offer more consistent performance.

**Does any other technique offer high performance across modalities?** No. We investigate the use of ImageNet pretraining on 2D modalities and find it performs well across several modalities, typically outperforming other techniques on CXRs, mammograms, and fundus images from the APTOS dataset. However, SSL methods sometimes outperform ImageNet pretraining, with a particularly large gap on OOD dermoscopic images. Additionally, while models trained from scratch are rarely top-performers in any experiment, they still remain competitive, frequently outperforming at least one other model. The fact that ImageNet pretraining and training from scratch can sometimes match SSL performance demonstrates the difficulty of using SSL techniques out-of-the-box, without customization for particular medical modalities. Future studies may explore two-stage approaches that pretrain models on other datasets, such as ImageNet, before performing SSL on medical images. Furthermore, while we standardized training times by performing the same number of iterations across all modalities, future work may explore other ways to set hyperparameters, such as by adjusting the number of iterations based on the shape of the loss curve.

**Does label availability affect performance?** Yes. Across all techniques and modalities, OOD performance typically stays the same or improves when more labels are available, though we see rare exceptions. For instance, MAE performance on mammograms drops when finetuning on 100% of the data, suggesting that the model may have overfit. Once again, future work may benefit from dynamically adjusting the finetuning process to prevent overfitting, perhaps by altering the learning rate.

**How does performance change across distributions?** Though we sometimes see promising generalization performance, there are also cases where performance drops on OOD datasets. On the 100% fine-tuning settings for ECG data, EEG data and mammograms, the top-performing technique on the in-distribution validation set also achieves the best performance across all OOD datasets, suggesting that generalization is successful. However, we see other cases of performance degradation due to distribution shift. For example, training from scratch achieves near-perfect performance on in-distribution dermoscopic images when finetuned with 64 or more data points. However, these models generalize poorly to OOD dermoscopic data, where training from scratch is never the top-performing technique on any experiment. Similarly, e-Mix is the top performer on most in-distribution LDCT experiments, yet it performs worst on

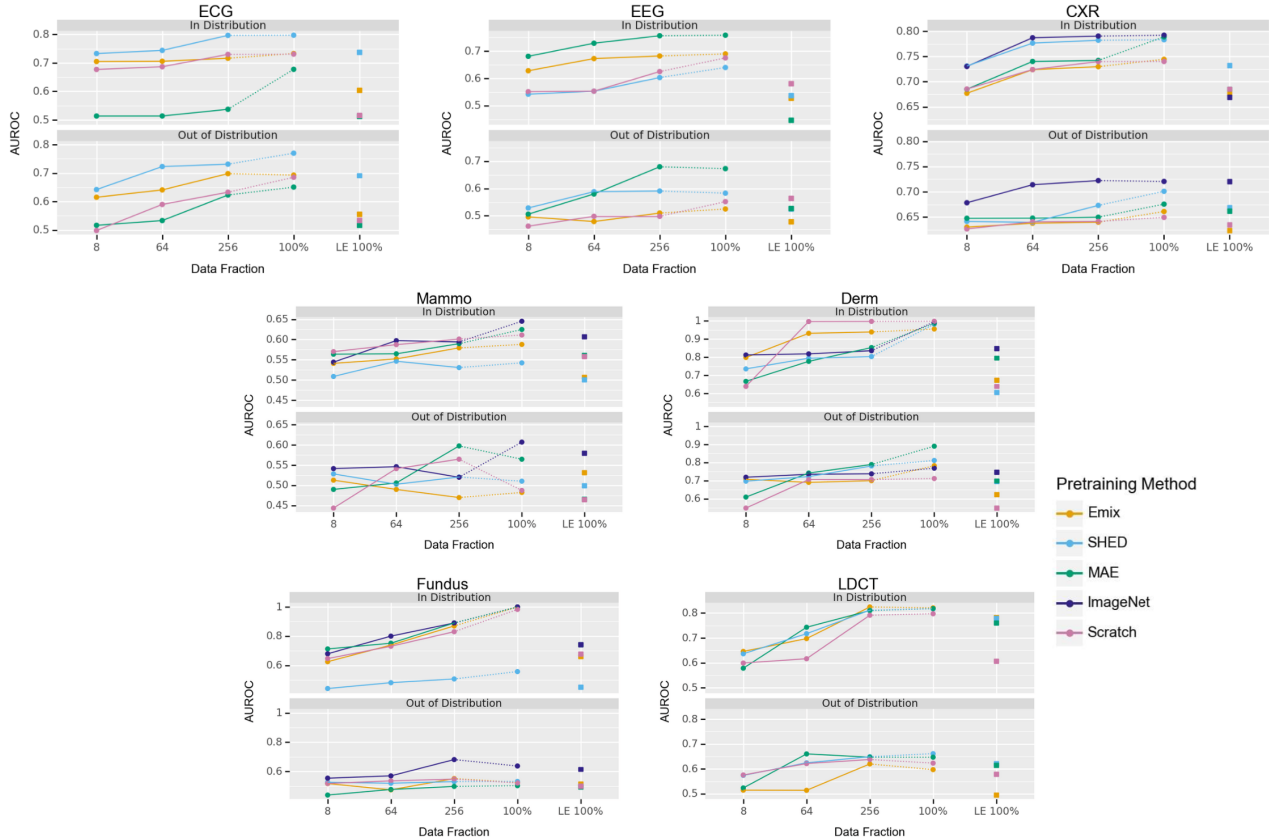


Figure 3. The in-distribution and out-of-distribution performance of models across modalities. OOD performance is averaged across target dataset(s).

all experiments using OOD data. Future work may use regularization techniques to improve generalization performance.

## 5. Limitations

While BenchMD generally aims to treat all modalities the same, we follow DABS’s example in providing different embedding modules for 1D, 2D, and 3D data [51]. Users can replace these modules with their own, and we are hopeful that future approaches will identify ways to unify even this step across modalities. Additionally, while SSL has demonstrated promise in the medical domain, our SSL baselines achieve modest performance and fail to provide consistent benefits over ImageNet pretraining on 2D modalities. Our strict approach to standardizing training procedures likely explains low performance, as we consistently use the same set of hyperparameters across all experiments. Future work may be able to improve performance by setting hyperparameters specifically for certain methods or by implementing modality-agnostic ways to dynamically adjust hyperparameters during training. The ImageNet training technique we present is also inherently limited, as it is

only appropriate for 2D images. As pretraining on natural images appears to offer benefits, future work may extend this approach to 1D and 3D data, such as by incorporating natural video pretraining as well. Finally, while we endeavor to cover a diverse range of medical modalities and datasets, it is impossible to fully represent the breadth of data in the medical domain. To protect patient safety, medical AI models should undergo further validation, such as through site-specific testing, before being deployed.

## 6. Conclusion

We present BenchMD, a benchmark for evaluating modality-agnostic methods on medical image and sensor modalities. While our initial baselines show some potential, there are ample opportunities for future work to improve both versatility and performance.

*Broader Impacts:* Methods that succeed on BenchMD may also be applicable to many other modalities and distributions and can have real-world impact on clinical practice. We hope BenchMD will help promote the development of high-performing, generalizable and label-efficient methods for modality-agnostic learning.



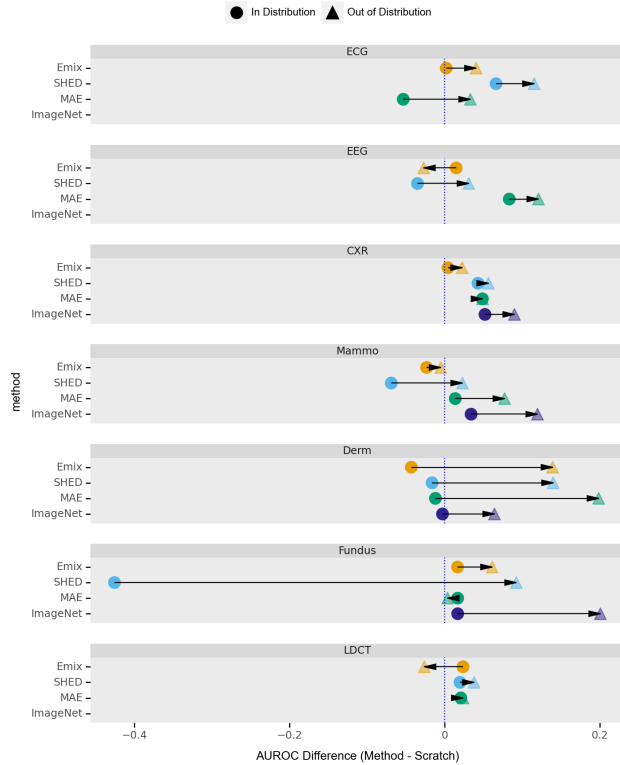


Figure 4. Performance relative to scratch on the ID validation set vs. OOD test datasets. Points that fall to the left of the line marking where the AUROC difference is 0 indicate that a technique performs worse than training from scratch. Arrows pointing from left to right show that a technique generalizes better than training from scratch. The length of such an arrow shows how much better a technique generalizes, compared to training from scratch. Typically, these techniques generalize better than training from scratch. Results displayed are for the Fine-Tune 100% setting and averaged across target datasets and tasks.

## References

- [1] The 4th asia pacific Tele-Ophthalmology society symposium. <https://2019.asiateleophth.org/>. Accessed: 2022-11-11. 5
- [2] APTOS 2019 blindness detection. <https://www.kaggle.com/competitions/aptos2019-blindness-detection/data>. Accessed: 2022-11-11. 5
- [3] Michael D Abramoff, James C Folk, Dennis P Han, Jonathan D Walker, David F Williams, Stephen R Russell, Pascale Massin, Beatrice Cochener, Philippe Gain, Li Tang, et al. Automated analysis of retinal images for detection of referable diabetic retinopathy. *JAMA ophthalmology*, 131(3):351–357, 2013. 5
- [4] Michael David Abramoff, Yiyue Lou, Ali Erginay, Warren Clarida, Ryan Amelon, James C Folk, and Meindert Niemeijer. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Invest. Ophthalmol. Vis. Sci.*, 57(13):5200–5206, Oct. 2016. 3, 5
- [5] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for Few-Shot learning. Apr. 2022. 3
- [6] Bruce M Altevogt, Harvey R Colten, et al. Sleep disorders and sleep deprivation: an unmet public health problem. 2006. 3
- [7] Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011. 6
- [8] Shekoofeh Azizi, Laura Culp, Jan Freyberg, Basil Mustafa, Sebastien Baur, Simon Kornblith, Ting Chen, Patricia MacWilliams, S Sara Mahdavi, Ellery Wulczyn, Boris Babenko, Megan Wilson, Aaron Loh, Po-Hsuan Cameron Chen, Yuan Liu, Pinal Bavishi, Scott Mayer McKinney, Jim Winkens, Abhijit Guha Roy, Zach Beaver, Fiona Ryan, Justin Krogue, Mozziyar Etemadi, Umesh Telang, Yun Liu, Lily Peng, Greg S Corrado, Dale R Webster, David Fleet, Geoffrey Hinton, Neil Houlsby, Alan Karthikesalingam, Mohammad Norouzi, and Vivek Natarajan. Robust and efficient medical imaging with Self-Supervision. May 2022. 3
- [9] Anuja Bandyopadhyay and Cathy Goldstein. Clinical applications of artificial intelligence in sleep medicine: a sleep clinician’s perspective. *Sleep and Breathing*, Mar. 2022. 4
- [10] Kathrin Blagec, Jakob Kraiger, Wolfgang Frühwirth, and Matthias Samwald. Benchmark datasets driving artificial intelligence development fail to capture the needs of medical professionals. Jan. 2022. 2
- [11] Said Boumaraf, Xiabi Liu, Chokri Ferkous, and Xiaohong Ma. A new Computer-Aided diagnosis system with modified genetic feature selection for BI-RADS classification of breast masses in mammograms. *Biomed Res. Int.*, 2020:7695207, May 2020. 5
- [12] Samuel R Bowman and George E Dahl. What will it take to fix benchmarking in natural language understanding? Apr. 2021. 2
- [13] Zhipeng Cai, Chengyu Liu, Hongxiang Gao, Xingyao Wang, Lina Zhao, Qin Shen, EYK Ng, and Jianqing Li. An open-access long-term wearable ecg database for premature ventricular contractions and supraventricular premature beat detection. *Journal of Medical Imaging and Health Informatics*, 10(11):2663–2667, 2020. 3
- [14] Whitney Chiao and Megan L Durr. Trends in sleep studies performed for medicare beneficiaries. *The Laryngoscope*, 127(12):2891–2896, 2017. 3

- [15] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. The cancer imaging archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging*, 26(6):1045–1057, Dec. 2013. 5
- [16] Marc Combalia, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, et al. Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*, 2019. 5
- [17] Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, Richard Ordonez, Pascale Massin, Ali Erginay, et al. Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology*, 33(3):231–234, 2014. 5
- [18] Brett Duce, Conchita Rego, Jasmina Milosavljevic, and Craig Hukins. The AASM recommended and acceptable EEG montages are comparable for the staging of sleep and scoring of EEG arousals. *J. Clin. Sleep Med.*, 10(7):803–809, July 2014. 4
- [19] Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M Hospedales. Self-Supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Process. Mag.*, 39(3):42–62, May 2022. 1
- [20] Florin C Ghesu, Bogdan Georgescu, Awais Mansoor, Youngjin Yoo, Dominik Neumann, Pragneshkumar Patel, R S Vishwanath, James M Balter, Yue Cao, Sasa Grbic, and Dorin Comaniciu. Self-supervised learning from 100 million medical images. Jan. 2022. 3
- [21] Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. OmniMAE: Single model masked pretraining on images and videos. June 2022. 3
- [22] A L Goldberger, L A Amaral, L Glass, J M Hausdorff, P C Ivanov, R G Mark, J E Mietus, G B Moody, C K Peng, and H E Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):E215–20, June 2000. 3
- [23] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000. 3, 4
- [24] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. Nov. 2021. 6
- [25] Isabelle Hoorens, Katrien Vossaert, Sven Lanssens, Laurence Dierckxsens, Giuseppe Argenziano, and Lieve Brochez. Value of dermoscopy in a Population-Based screening sample by dermatologists. *Dermatol Pract Concept*, 9(3):200–206, July 2019. 3, 5
- [26] E Immonen, J Wong, M Nieminen, L Kekkonen, S Roine, S Törnroos, L Lanca, F Guan, and E Metsälä. The use of deep learning towards dose optimization in low-dose computed tomography: A scoping review. *Radiography*, 2021. 6
- [27] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghighi, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019. 3, 4
- [28] Alistair Johnson, Tom Pollard, Roger Mark, Seth Berkowitz, and Steven Horng. MIMIC-CXR database, Sept. 2019. 4
- [29] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):1–8, 2019. 4
- [30] Alexander Ke, William Ellsworth, Oishi Banerjee, Andrew Y Ng, and Pranav Rajpurkar. CheXtransfer: performance and parameter efficiency of ImageNet models for chest X-Ray interpretation. In *Proceedings of the Conference on Health, Inference, and Learning, CHIL ’21*, pages 116–124, New York, NY, USA, Apr. 2021. Association for Computing Machinery. 4
- [31] Sirvan Khalighi, Teresa Sousa, José Moutinho Santos, and Urbano Nunes. Isruc-sleep: A comprehensive public dataset for sleep researchers. *Computer methods and programs in biomedicine*, 124:180–192, 2016. 4
- [32] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanus Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR, 2021. 3
- [33] Rayan Krishnan, Pranav Rajpurkar, and Eric J Topol. Self-supervised learning in medicine and healthcare. *Nat Biomed Eng*, Aug. 2022. 2
- [34] Kibok Lee, Yian Zhu, Kihyuk Sohn, Chun-Liang Li, Jinwoo Shin, and Honglak Lee. i-mix: A Domain-Agnostic strategy for contrastive representation learning. Oct. 2020. 3
- [35] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L Rubin. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific data*, 4(1):1–9, 2017. 5
- [36] Valerii Likhoshesterov, Anurag Arnab, Krzysztof Choromanski, Mario Lucic, Yi Tay, Adrian Weller, and Mostafa Dehghani. PolyViT: Co-training vision transformers on images, videos and audio. Nov. 2021. 3
- [37] Debra L Monticciolo, Sharp F Malak, Sarah M Friedewald, Peter R Eby, Mary S Newell, Linda Moy, Stamatia

- Destounis, Jessica W T Leung, R Edward Hendrick, and Dana Smetherman. Breast cancer screening recommendations inclusive of all women at average risk: Update from the ACR and society of breast imaging. *J. Am. Coll. Radiol.*, 18(9):1280–1288, Sept. 2021. 3
- [38] Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist’s annotations. *Scientific Data*, 9(1):1–7, 2022. 4
- [39] Hieu Trung Nguyen, Ha Quy Nguyen, Hieu Huy Pham, Khanh Lam, Linh Tuan Le, Minh Dao, and Van Vu. Vindr-mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *medRxiv*, 2022. 5
- [40] Simon Ott, Adriano Barbosa-Silva, Kathrin Blagec, Jan Brauner, and Matthias Samwald. Mapping global dynamics of benchmark creation and saturation in artificial intelligence. Mar. 2022. 2
- [41] Andre GC Pacheco, Gustavo R Lima, Amanda S Salomão, Breno Krohling, Igor P Biral, Gabriel G de Angelo, Fábio CR Alves Jr, José GM Esgario, Alana C Simora, Pedro BC Castro, et al. Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in brief*, 32:106221, 2020. 5
- [42] João Pedrosa, Guilherme Aresta, Carlos Ferreira, Márcio Rodrigues, Patrícia Leitão, André Silva Carvalho, João Rebelo, Eduardo Negrão, Isabel Ramos, António Cunha, et al. Lndb: a lung nodule database on computed tomography. *arXiv preprint arXiv:1911.08434*, 2019. 6
- [43] Stuart F Quan, Barbara V Howard, Conrad Iber, James P Kiley, F Javier Nieto, George T O’Connor, David M Rapoport, Susan Redline, John Robbins, Jonathan M Samet, et al. The sleep heart health study: design, rationale, and methods. *Sleep*, 20(12):1077–1085, 1997. 4
- [44] Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. AI and the everything in the whole wide world benchmark. Nov. 2021. 2
- [45] Satwik Ramchandre, Bhargav Patil, Shardul Pharande, Karan Javali, and Himangi Pande. A deep learning approach for diabetic retinopathy detection using transfer learning. In *2020 IEEE International Conference for Innovation in Technology (INOCON)*, pages 1–5, Nov. 2020. 5
- [46] Shiori Sagawa, Pang Wei Koh, Tony Lee, Irena Gao, Sang Michael Xie, Kendrick Shen, Ananya Kumar, Weihua Hu, Michihiro Yasunaga, Henrik Marklund, Sara Beery, Etienne David, Ian Stavness, Wei Guo, Jure Leskovec, Kate Saenko, Tatsunori Hashimoto, Sergey Levine, Chelsea Finn, and Percy Liang. Extending the WILDS benchmark for unsupervised adaptation. Dec. 2021. 3
- [47] Anton Schreuder, Ernst T Scholten, Bram van Ginneken, and Colin Jacobs. Artificial intelligence for detection and characterization of pulmonary nodules in lung cancer CT screening: ready for practice? *Transl Lung Cancer Res*, 10(5):2378–2388, May 2021. 3
- [48] Kirk Smith. Curated breast imaging subset of digital database for screening mammography (CBIS-DDSM) - the cancer imaging archive (TCIA) public access - cancer imaging archive wiki. <https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=22516629>. Accessed: 2022-11-11. 5
- [49] Hidenori Takahashi, Hironobu Tampo, Yusuke Arai, Yuji Inoue, and Hidetoshi Kawashima. Applying artificial intelligence to disease staging: Deep learning for improved staging of diabetic retinopathy. *PLoS One*, 12(6):e0179790, June 2017. 5
- [50] Alex Tamkin, Gaurab Banerjee, Mohamed Owda, Vincent Liu, Shashank Rammoorthy, and Noah Goodman. DABS 2.0: Improved datasets and algorithms for universal Self-Supervision. Oct. 2022. 3
- [51] Alex Tamkin, Vincent Liu, Rongfei Lu, Daniel Fein, Colin Schultz, and Noah Goodman. DABS: A Domain-Agnostic benchmark for Self-Supervised learning. Nov. 2021. 3, 6, 8
- [52] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018. 5
- [53] Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):1–15, 2020. 3
- [54] Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset, Nov. 2022. 3
- [55] Anthony B Wolbarst, Patrizio Capasso, and Andrew R Wyant. *Medical Imaging: Essentials for Physicians*. John Wiley & Sons, Apr. 2013. 2
- [56] Huaxiu Yao, Caroline Choi, Bochuan Cao, Yoonho Lee, Pang Wei Koh, and Chelsea Finn. Wild-Time: A benchmark of in-the-wild distribution shift over time. Oct. 2022. 3
- [57] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruysen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. A large-scale study of representation learning with the visual task adaptation benchmark. Oct. 2019. 3
- [58] Jianwei Zheng, Jianming Zhang, Sidy Danioko, Hai Yao, Hangyuan Guo, and Cyril Rakovski. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific data*, 7(1):1–8, 2020. 3
- [59] Lei Zhou, Huidong Liu, Joseph Bae, Junjun He, Dimitris Samaras, and Prateek Prasanna. Self pre-training with masked autoencoders for medical image analysis. Mar. 2022. 3
- [60] Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. Models genesis. *Med. Image Anal.*, 67:101840, Jan. 2021. 3
- [61] Hongling Zhu, Cheng Cheng, Hang Yin, Xingyi Li, Ping Zuo, Jia Ding, Fan Lin, Jingyi Wang, Beitong Zhou, Yonge Li, Shouxing Hu, Yulong Xiong, Binran Wang, Guohua Wan, Xiaoyun Yang, and Ye Yuan. Automatic multilabel electrocardiogram diagnosis of heart rhythm or conduction abnormalities with deep learning: a cohort study. *Lancet Digit Health*, 2(7):e348–e357, July 2020. 3