ENRICH: Exploiting Image Similarity to Maximize Efficient Machine Learning in Medical Imaging

Erin Chinn MS[1], Rohit Arora PhD [2†], Ramy Arnaout MD DPhil[2-3*], Rima Arnaout MD[1*]

1 Department of Medicine, Division of Cardiology, Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, CA

Email: rima.arnaout@ucsf.edu

2 Division of Clinical Pathology, Department of Pathology, Beth Israel Deaconess Medical Center, Boston, MA 02215

Email: rarnaout@bidmc.harvard.edu

3 Division of Clinical Informatics, Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA 02215.

† Present address: Iktos Inc. 50 Milk Street, Floor 16, Boston, MA, 02109

* indicates co-corresponding authors

Abstract

Deep learning (DL) requires labeled data. Labeling medical images requires medical expertise, which is often a bottleneck. It is therefore useful to prioritize labeling those images that are most likely to improve a model's performance, a practice known as instance selection. Here we introduce ENRICH, a method that selects images for labeling based on how much novelty each image adds to the growing training set. In our implementation, we use cosine similarity between autoencoder embeddings to measure that novelty. We show that ENRICH achieves nearly maximal performance on classification and segmentation tasks using only a fraction of available images, and outperforms the default practice of selecting images at random. We also present evidence that instance selection may perform categorically better on medical vs. non-medical imaging tasks. In conclusion, ENRICH is a simple, computationally efficient method for prioritizing images for expert labeling for DL.

Introduction

Deep learning (DL) has been applied with success in proofs of concept across biomedical imaging modalities and medical specialties [1–17]. DL models can classify images by disease or structure and can segment, track, and measure structures within images. DL thus has great promise for helping meet the overwhelming need for accurate, reliable, and scalable image interpretation that currently exists due to a near-universal shortage of trained experts [5,6,18–21]. However, DL requires labeled data, and labeling and annotation require those same experts. Labeling and annotation may even require agreement from multiple experts before assigning a gold-standard label [3]. Even when semi-supervised or unsupervised methods are used to train a DL model, or when weak labels are used, experts can still be needed to label images in test datasets, in order to benchmark performance on high-stakes medical tasks. As a result, labeling can be a costly and time-consuming bottleneck for DL in medical imaging. This is in contrast to labeling for DL in non-medical fields, which usually focuses on everyday objects and therefore can be performed more quickly and inexpensively by laypeople via crowdsourcing [22].

It has long been recognized that prioritizing training data that most benefits model performance, *instance selection*, as opposed to choosing data at random, should reduce the labeling burden for DL [23,24]. The challenge is determining which training data to prioritize. Another well-established form of data selection is *active learning* [25,26]. In active learning, a model (e.g. an image classifier) is trained on a labeled subset of data from a larger unlabeled pool (e.g. unlabeled images). The remaining unlabeled data are evaluated according to the model, and some selection criterion is applied. The best-performing instance is then added to the training set, a new model is trained on the now-larger training set, and the cycle is repeated. In image classification many selection criteria have been evaluated, including measures of the uncertainty of the model's classification of a candidate image, the image's contribution to the training set's entropy, and the image's representativeness of the pool [27,28]. These investigations have been fruitful but have not identified a universal best performer across datasets and modeling approaches. Iterative model retraining can make active learning computationally intensive for DL, a drawback somewhat alleviated by selecting images in batches instead of individually (at the cost of making learning less "active") [29].

In contrast, instance selection involves choosing images based only on their relationships to the rest of the images in the pool and growing the training set, avoiding the computational cost of iterative retraining. Image-selection criteria generally balance some measure of the informativeness of a candidate image with some measure of how much that image will add to the diversity of the resulting training set [23]. Instance selection has been shown to reduce training-set sizes in many settings, but again which criteria perform best seems to depend on the type of model and dataset. Also, the majority of work on instance selection involves non-imaging datasets and precedes recent developments in DL.

Medical images differ from images of everyday objects in ways that we hypothesized could be leveraged for a new instance selection approach. Unlike images of everyday objects, which typically exhibit multiple lighting conditions and are captured at a range of distances, angles, and contexts, medical images are often more uniform in these respects, a result of standardization of imaging protocols for patient care (more true for images of macroscopic and/or intact structures than for histology images). Images from a particular medical domain often have similar subject matter (e.g., the heart in cardiology, the retina in ophthalmology), pose (standard views), background (black), noise, lighting, and color (monochrome). In the case of computed tomography (CT), magnetic resonance imaging (MRI), ultrasound (US), and other common modalities, image frames may be captured consecutively, resulting in similarity among consecutive images. We hypothesized that standardization in medical imaging creates greater redundancy in medical training data than in commonly used non-medical datasets, and propose that simply prioritizing non-redundant images is an efficient means of instance selection for DL in medical imaging.

Here, we present a method called ENRICH—Eliminating Needless Redundancy for Imaging Challenges—consisting of two main steps. First, a similarity metric is calculated for all pairs of images in a given dataset, forming a matrix of pairwise-similarity values. Second, an algorithm operates on the matrix to identify those unlabeled images that are least similar to images in an existing seed training set and thereby hypothetically most informative. The result is a meaningful decrease in the redundancy, size, and labeling burden of the resulting dataset. We demonstrate proof of concept on classification and segmentation tasks on two large, well characterized/well benchmarked medical datasets: ECHO-F [3], which consists of fetal echocardiograms, and OCT [30], which consists of adult retinal optical coherence tomography images. We also demonstrate the special nature of medical image datasets, demonstrating differences in their pairwise

similarities compared to STL10, a standard non-medical image dataset used for unsupervised, self-taught learning [31].

Methods

Datasets and benchmarks. Training and test sets are described in **Table 1**.

ECHO-F consists of labeled fetal echocardiogram images [3]. The classification task was to predict the fetal axial 4-chamber (A4C) view vs. the non-target (NT) view. In ultrasound, one or more video clips are acquired per patient; each video clip consists of one to several hundred consecutive image frames. Training and test sets were divided by patient identifier (ID) and were disjoint from each other (mutually exclusive).

ECHO-F-SEG consists of a subset of A4C images from the ECHO-F dataset above. ECHO-F-SEG was used for a (multi-class) segmentation task, with 5 class labels for image pixels: left ventricle, right ventricle, left atrium, right atrium, background. Notably, the segmentation dataset had already been curated informally, in that only certain frames from each video clip were labeled.

OCT consists of labeled adult retinal optical coherence tomography images [30]. This dataset was used to classify between a normal retina (NL) vs. choroidal neovascularization (CNV). The train/test split of the dataset was adjusted from the original authors' description: instead of 250 images per lesion in the test set, we created disjoint train/test sets, similar to ECHO-F, split by patient ID. This increased the total size of the test set from 500 images to 17,638 images.

STL10 consists of labeled and unlabeled images of animals and vehicles for unsupervised learning tasks (self-taught learning) [31]. Our classification task for this dataset was predicting images of airplanes (AIR) vs. images of trucks (TRUCK). In addition to the initial 500 labeled training images per class, a larger set of labeled training data was curated from the 100K unlabeled training images using a progression of classifiers and human sorting. The train/test split of the dataset remained unchanged, with 800 images per class.

Image processing. All grayscale conversion was done using Python3's OpenCV package; all image resizing was done using Python3's Scikit-Image package.

ECHO-F images were originally 300x400 and converted to grayscale. For autoencoder input, images were cropped and resized to 64x64. For classification model training and testing, the

original images were cropped and resized to 80x80. For segmentation model training and testing, original images were cropped to 272x272, and no resizing was performed.

OCT images were originally grayscale and varied in size. This dataset was put through an additional preprocessing step to correct region-of-interest misalignment and remove white-edge artifacts. First, white sections at image edges were removed, then images underwent a similar cropping and resizing process performed on the ECHO-F classification dataset.

STL10 images were originally 96x96 and converted to grayscale. For autoencoder input, images were resized to 64x64. For model training and testing, the original images were resized to 80x80.

Embeddings from autoencoders. The bottleneck layer of a disentangled variational autoencoder (β-VAE) was used to compress each image into a 128-element vector embedding. The β-VAE used was based on the architecture as described previously with the exception of having a 128-element embedding [32]. The β-VAE was trained on a subset of 5,000 images from the entire ECHO-F training dataset as previously described [3], using combined loss (reconstruction loss and Kullback-Leibler divergence) and standard stopping conditions.

Pairwise image similarities. For each dataset (**Table 1**), a matrix of pairwise image similarities was calculated. The similarity between two image embeddings was defined as the complement of the cosine distance between each embedding (resulting in pairwise similarities ranging from 0 for highly dissimilar images to 1 for identical images).

Ranking algorithm. For each deep-learning task (classification and segmentation), an initial subset of images was chosen uniformly at random. Additional images from the remaining dataset, determined by the ranking algorithm to have the lowest similarity to the initial subset, were added iteratively to grow the subset. For statistical confidence, this process of random subset initialization and subset growth was repeated 3 times for each task to provide replicates. We then trained 10 new models with the resulting training sets grown from each initialization, at specific subset sizes (**Table 2**), and predicted on the test set. For classification, the ranking algorithm was blind to class label during iterative selection; the label was revealed/assigned only after an image was chosen. For segmentation, the ranking algorithm was blind to segmentation label.

ECHO-F classification experiments started with 1,000 images, roughly 2 percent of the full training set (45,460 images). Approximately 5,000 images at a time were added with each iteration of subset growth.

ECHO-F-SEG segmentation experiments started with 200 images, 16 percent of the full training set (1,248 images). Approximately 100 images at a time were added with each iteration of subset growth.

OCT experiments started with 400 images, roughly 1 percent of the full training set (46,164 images). Approximately 500 images at a time were added with each iteration of subset growth.

STL10 experiments started with 500 images, roughly 5 percent of the full training set (10,176 images). Approximately 1,000 images at a time were added with each iteration of subset growth.

Model training. Resnet and U-net architectures were used to train classification and segmentation models, respectively, as previously described [3]. Data augmentation was used for the segmentation task as previously described [3] but not for any of the classification tasks. Experiments for each dataset used the same model parameters throughout.

Human labeling time estimates. Human labeling time was estimated at 3 seconds per image for classification tasks and 5 minutes per image for segmentation tasks, based on the average time it took for these tasks for ECHO-F and ECHO-F-SEG (n=4 labelers across several different labeling platforms)

Evaluation metrics. For each dataset, we calculated the highest pairwise similarity for each image. This excluded images compared with themselves (which would have had similarity=1.0). We then plotted the cumulative distribution of these maximum similarity values for each dataset (**Fig. 1**).

Model performance was compared using the area under the receiver operator characteristic curve (AUC) for the classification tasks and average Jaccard score of the four heart segments (left ventricle, right ventricle, left atrium, right atrium) for the segmentation task, as previously described [3].

Statistical Comparisons. Model performance was compared using a 2-tailed Mann-Whitney U test. For statistical confidence, at each training subset a total of 60 models were compared; 30 models trained using ENRICH to add images to the starting seed and 30 models trained using images added at random. Performance was also compared using standardization at each subset size (by subtracting the mean performance and scaling the resulting values to a standard deviation of 1), pooling the resulting standardized values, and using Mann-Whitney U to compare the results for ENRICH to those for random selection (to compare performance across sample sizes) (**Fig. 5**).

Results

ENRICH involves two steps: (1) computing a matrix of pairwise-similarity values for all pairs of images in a given dataset and (2) ranking images by similarity for inclusion in the curated dataset. Here, we used ENRICH with a pairwise similarity measure based on the distance between β-VAE embeddings, and a ranking algorithm designed to minimize image redundancy. Note that these choices do not require image labels to be assigned prior to training. We thereby demonstrated significant redundancy in the medical image datasets ECHO-F and OCT, and higher redundancy in these medical datasets than in the non-medical STL10. To our knowledge this is the first quantitative demonstration of this property and the first such comparison. We also demonstrated that by using ENRICH to curate these training datasets, the same test performance on well benchmarked binary classification was achieved using only a fraction of the available training images (ECHO-F 55 percent of available images, OCT 32.5 percent of available images).

Image redundancy in medical datasets. Based on prior experience with medical data [3], we hypothesized that medical image datasets have significant redundancy among images, and that such redundancy is not confined to images from a given patient or video clip but instead is distributed across the dataset. To test this hypothesis, for each dataset we identified the maximum pairwise similarity for each image. The majority of ECHO-F classification images had a maximum similarity greater than 0.9, i.e. the majority of images had at least one other image in the dataset that they are at least 90% similar to. The OCT dataset also demonstrates considerable redundancy, with roughly half of the images having a maximum pairwise similarity greater than 0.8. In stark contrast, close to 60 percent of the images in the STL10 dataset have a maximum similarity less than 0.4 (**Fig. 1**).

Using ENRICH to find smaller training datasets that can achieve benchmark performance on medical imaging tasks. We trained binary classification and multi-class segmentation models with different subsets of each of the training datasets. Each trained model was tested on the same set of test images (**Table 1**). As in the Methods, training image subsets of increasing size were curated using ENRICH vs random selection (**Fig. 3**).

ECHO-F Classification. When trained on the full training dataset, model test performance achieved a mean AUC of 0.99 (± 3.43×10^{-4}). Even with the smallest training subsets, ENRICH selected images that represented almost all of the available patients and video clips (**Table 2**).

With just 11 percent of images from the full ECHO-F training dataset, ENRICH outperformed random selection of training images (mean AUC 0.96 vs 0.94, MWU p-value 3.88×10^{-9}; **Fig. 2b**). The size of the training subset required to achieve statistically similar results to the full training dataset was also investigated. When training images were chosen using ENRICH, only 55 percent of the training dataset (25,000/45,460 images) was needed to achieve benchmark performance (AUC 0.99 ± 4.53×10^{-4}). We were not able to achieve the same benchmark from random sampling.

OCT Classification. Model test performance achieved a mean AUC of 0.99 (± 2.24×10^{-5}) when trained on the full training dataset. ENRICH outperformed random selection at just 2 percent of all OCT images (mean AUC 0.995 vs 0.993, MWU p-value 9.98×10^{-6}). Only 32.5 percent of the training dataset (15,000/46,164 images) was needed to achieve benchmark performance (AUC 0.99 ± 2.84×10^{-5}) when training images were chosen using ENRICH. When chosen at random, 41 percent of the training dataset (19,000/46,164 images) was needed to achieve the same benchmark (AUC 0.99 ± 2.70×10^{-5}).

STL10 Classification. Model test performance achieved a mean AUC of 0.99 (± 2.04×10^{-4}) when trained on the full training dataset. Initially, random image selection outperformed ENRICH (20 percent: mean AUC 0.969 vs 0.966, MWU p-value 5.87×10^{-4}). At 50 percent of all STL10 images, ENRICH outperformed random selection (mean AUC 0.992 vs 0.990, MWU p-value 5.45×10^{-6}) and continued to outperform random sampling. In order to achieve benchmark performance 90 percent of the total dataset (9,000/10,176 images) was needed (mean AUC 0.99 vs 0.99, MWU p-value 0.42). We were not able to achieve the same benchmark with random sampling.

ECHO-F-SEG Segmentation. We also compared training subsets chosen by ENRICH vs randomly chosen image subsets for a multi-class segmentation task. Using all available training data (**Table 1**), average Jaccard index was 0.68. With 80 percent of the training data, ENRICH achieved an average Jaccard of 0.66 on 80 percent of available images (1,000/1,248 images).

Potential time savings in labeling. As an example, we estimated the time required to label all the images in ECHO-F for classification and ECHO-F-SEG for segmentation tasks. We compared this to the time that would have been required for the smallest ENRICHed subsets that achieved desired performance (55 percent for classification and 80 percent for segmentation). This suggests a savings of 38 hours of full-time work for an expert labeler, on even this relatively small dataset (**Fig. 4**).

Discussion

In DL for medical imaging, investigators generally rely on a crude metric for dataset quality and content: the number of images in a dataset. However, “more is better” is untenable as the field progresses and datasets grow: requiring more storage and compute, overburdening human labelers, hamstringing agile development of DL models, and excluding smaller research groups from the field. Therefore, there is a critical and urgent need for better metrics for image dataset content.

Instance selection provides a general strategy for addressing these shortcomings. Standardization in medical imaging suggested to us that images that are the least similar to each other, when preprocessed appropriately, would be among the most valuable to label. Our method, ENRICH, curates medical image datasets based on quantitative measures of image similarity. Our results show that ENRICH can be used to identify redundancy in image training datasets. We further demonstrate that medical datasets such as ECHO-F and OCT contain significant redundancy. Using ENRICH demonstrated that *(i)* redundant images do not aid significantly in DL model training, *(ii)* image labels are not needed in order to curate image datasets according to redundancy, and *(iii)* for some medical datasets, state-of-the-art performance can be achieved using only a fraction—sometimes a small fraction—of the full training dataset.

ENRICH eliminates the need to label large portions of available medical imaging data (**Fig. 3**) while still achieving the same performance as when all images are used. Furthermore, with only a minor hit to performance, even fewer images can be used. For example, while not statistically the same as full dataset performance, the performance on just 22 percent of the ECHO-F classification training dataset had an AUC of 0.98, and a dataset of only 2 percent of the size of the full OCT training dataset still had an AUC of 0.99 (**Fig. 3**). It is reasonable to conclude that some medical classification problems may be more straightforward, in the sense of being less data-hungry, than they have traditionally appeared, given the appropriate (i.e. ENRICHed or otherwise low-redundancy) dataset.

Furthermore, it is important to note that classification model trainings demonstrated here did not include standard data augmentations (such as rotating or flipping images; see Methods). This choice was made in order to *(i)* remove data augmentation as a potential confounding factor in measuring ENRICH performance *(ii)* mimic clinical DL model-training situations where data

augmentation may not be desired. However, in the future, data augmentation can be applied to enrich training subsets, hopefully requiring even fewer images to meet optimum test performance.

The implications of these findings for economizing on expert clinical image labeling are clear. Perhaps future studies using medical imaging datasets might benefit from choosing a small, diverse, ENRICHed, subset of images to human label. Focusing resources on a select few images, rather than relying on weak labeling large batches of potentially redundant images, which may result in mislabeled, noisy data [33]. Also, ENRICH may be used either in conjunction with or instead of transfer learning from models trained on enormous, unrelated datasets.

It is notable that ENRICH was less helpful for the segmentation task studied here than for the classification tasks. However, for this task, labeling was so time-consuming that we had *already* chosen not to label every frame in each video clip, thus ECHO-F-SEG was already a subset of ECHO-F (Methods). Therefore, the finding that an additional 20 percent of the already-intuitively-reduced dataset was not needed to reach full dataset performance is still an additional gain in efficiency over informal curation. The segmentation task therefore demonstrates that a quantitative approach to image dataset curation has advantages over intuitive approaches. When considering that labeling each image for segmentation took several minutes, and 20 percent of the training dataset for segmentation comprised 249 image frames, the potential time savings in labeling *even on an already-manually-reduced dataset* is significant.

The OCT dataset split was adjusted due to the test set, originally 500 images, being too easy to classify. Experiments resulted in perfect test set separability (AUC=1.0) despite very small training set size (<2 percent total images available). Even with this adjustment, the OCT test dataset was still very separable (**Fig. 3**). In theory, the same methods used here to curate training data can be used to curate testing datasets, in order to provide the most efficient and most representative benchmarks for generalizability.

Results from the STL10 dataset provided an intriguing counterexample to the medical imaging datasets. Although ENRICH experiments resulted in a subset (90 percent) achieving benchmark performance of the full training dataset, STL10 is simply less redundant—more diverse— than the other datasets studied here (**Fig. 1**). We suspect this will prove generally true for

non-medical datasets, because of the variety in how they capture real-world objects. If so, it would point to the value of designing methods and model architectures explicitly for medical data, as opposed to porting them from other domains, for efficient machine learning in medicine.

Our main goal with ENRICH is to alleviate the human-labeling burden that medical imaging datasets often present by removing redundant images. Smaller datasets can also help data science researchers economize on storage and compute, especially as they iterate in model development. We have shown that model performance does not suffer by removing these images, and in some cases a significant proportion of the dataset can be removed without detriment to performance. Data reduction methods such as active learning are model-guided: the model selects the "best" images for learning [34–36], while ours is data-guided: the data determine which images are redundant and are best removed. Potential combinations of ENRICH with active learning are interesting directions for future exploration.

ENRICH can accommodate arbitrary choices of similarity measure (step 1) and ranking algorithm (step 2) (**Fig. 2**). As proof of concept, here we used embeddings from a β-VAE to provide a pairwise image similarity measure based on imaging data of the same general type used in our experiments, and we used a ranking algorithm that did not require *a priori* labeling even of the starting training images. In the future, investigating alternative similarity measures and ranking algorithms offers opportunities to test and potentially optimize ENRICH for specific image datasets or imaging tasks. For example, other pairwise image-similarity metrics may prove more informative or simpler to compute; other ranking algorithms may account for class balance, which is important in classification tasks. In addition, different algorithm choices as well as code optimizations can be explored to maximize the utility of ENRICH while minimizing time and computational load. Quantitative measures of similarity have been shown to add useful insights in other fields [37,38]. ENRICHment, in various forms, is expected to be a useful new avenue for decreasing labeling burden and speeding iterative training and testing of DL models in development.

Acknowledgements

EC, RA, RA, and RA were supported by the Department of Defense (W81XWH-19-1-0294) and the National Heart, Lung, and Blood Institute (NIH R01HL146398). RA and RA were supported by the National Institutes of Allergy and Infectious Diseases (NIH R01AI148747-01). EC and RA were supported by the American Heart Association (17IGMV33870001).

Author Contributions

RA and RA conceived of the study. Similarity metric, algorithm design, image preprocessing, and neural-network design and testing were implemented by EC with input from RA and RA. RA, RA, and EC all contributed to the writing of the manuscript.

Competing Interests

The authors have no potential competing interests to report.

Code and Data Availability

Code will be made available at https://github.com/ArnaoutLabUCSF/cardioML upon publication. The datasets OCT and STL10 are publicly available at the Mendeley Data repository and the Stanford University Computer Science Department's webpage, https://data.mendeley.com/datasets/rscbjbr9sj/2 and https://cs.stanford.edu/~acoates/stl10/ respectively. Due to patient privacy constraints the ECHO-F and ECHO-F-SEG datasets will not be made available to the public.

Figure Legends

Figure 1. Cumulative density of maximum pairwise similarities. ECHO-F, OCT, STL10, and ECHO-F-SEG datasets are shown. Also included are the total images available for the ECHO-F segmentation task, ECHO-F-SEG-ALL.

Figure 2. Experimental schematic. From all available images in a dataset, an initial subset (a) is chosen at random. The remaining images comprise a candidate pool of images (b) from which additional images can be selected. A matrix of pairwise image similarities (Step 1 of ENRICH) is constructed (c). From this matrix, (d) an algorithm is used to choose additional images to add to the initial training set; this is Step 2 of ENRICH. A redundancy-reducing ranking algorithm was used in this work, compared to random choice of images as a negative control. (e) this process was repeated, iteratively adding images to an initial subset.

Figure 3. Performance of ENRICHed training datasets compared to randomly selected training datasets (a) ECHO-F classification, (b) ECHO-F-SEG segmentation (c) OCT classification, and (d) STL10 classification. From a common initial random starting dataset (grey), additional images were added to grow increasingly larger training subsets using ENRICH (blue) vs random addition (yellow). Dots represent mean AUC on the test set from 30 replicates for each datapoint; error bars for each datapoint show 1 standard deviation around the mean. Asterisks for each training data subset represent statistical differences between ENRICH and random according to the standard convention (*n*s = $p>0.05$; * = $p\leq0.05$; ** = $p\leq0.01$; *** = $p\leq0.001$; **** = $p\leq0.0001$). Datapoints circled in red are statistically indistinguishable from model performance using the full training set (100 percent of training images; black dot).

Figure 4. Labeling time savings using ENRICH. Time estimates for labeling ECHO-F and ECHO-F-SEG datasets.

Figure 5. Standardized combined subsets demonstrate overall differences between ENRICH and random selection (a) ECHO-F classification, (b) ECHO-F-SEG segmentation (c) OCT classification, and (d) STL10 classification. Comparison of ENRICH to random performance on grouped subsets.

References

1. Madani, A., Arnaout, R., Mofrad, M. & Arnaout, R. Fast and accurate view classification of echocardiograms using deep learning. *NPJ Digit. Med.* **1**, 6 (2018).

2. Kornblith, A. E. *et al. Development and Validation of a Deep Learning Model for Automated View Classification of Pediatric Focused Assessment with Sonography for Trauma (FAST)*. 2020.10.14.20206607 https://www.medrxiv.org/content/10.1101/2020.10.14.20206607v1 (2020) doi:10.1101/2020.10.14.20206607.

3. Arnaout, R. *et al.* An ensemble of neural networks provides expert-level prenatal detection of complex congenital heart disease. *Nat. Med.* **27**, 882–891 (2021).

4. *Deep Learning in Medical Image Analysis : Challenges and Applications*. (Springer International Publishing, 2020). doi:10.1007/978-3-030-33128-3.

5. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).

6. Gulshan, V. *et al.* Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **316**, 2402–2410 (2016).

7. Xu, J. *et al.* Fetal Pose Estimation in Volumetric MRI using a 3D Convolution Neural Network. *Med. Image Comput. Comput.-Assist. Interv. MICCAI Int. Conf. Med. Image Comput. Comput.-Assist. Interv.* **11767**, 403–410 (2019).

8. Rhee, D. J. *et al.* Automatic contouring system for cervical cancer using convolutional neural networks. *Med. Phys.* **47**, 5648–5658 (2020).

9. Gjesteby, L. *et al.* A dual-stream deep convolutional network for reducing metal streak artifacts in CT images. *Phys. Med. Biol.* **64**, 235003 (2019).

10. Li, H. *et al.* DeepLiverNet: a deep transfer learning model for classifying liver stiffness using clinical and T2-weighted magnetic resonance imaging data in children and young adults. *Pediatr. Radiol.* **51**, 392–402 (2021).

11. Anderson, B. M. *et al.* Automated Contouring of Contrast and Noncontrast Computed Tomography Liver Images With Fully Convolutional Networks. *Adv. Radiat. Oncol.* **6**, 100464 (2020).

12. Shen, Y. *et al.* An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Med. Image Anal.* **68**, 101908 (2021).

13. Shao, M. *et al.* Shortcomings of Ventricle Segmentation Using Deep Convolutional Networks. *Underst. Interpret. Mach. Learn. Med. Image Comput. Appl.* **11038**, 79–86 (2018).

14. Kaye, E. A. *et al.* Accelerating Prostate Diffusion-weighted MRI Using a Guided Denoising Convolutional Neural Network: Retrospective Feasibility Study. *Radiol. Artif. Intell.* **2**, e200007 (2020).

15. Vidyaratne, L., Alam, M., Shboul, Z. & Iftekharuddin, K. M. Deep Learning and Texture-Based Semantic Label Fusion for Brain Tumor Segmentation. *Proc. SPIE-- Int. Soc. Opt. Eng.* **2018**, 105750D (2018).

16. Zhang, J. *et al.* Fully Automated Echocardiogram Interpretation in Clinical Practice. *Circulation* **138**, 1623–1635 (2018).

17. Fan, L. *et al.* Rapid dealiasing of undersampled, non-Cartesian cardiac perfusion images using U-net. *NMR Biomed.* **33**, e4239 (2020).

18. Rosenkrantz, A. B., Hughes, D. R. & Duszak, R. The U.S. Radiologist Workforce: An Analysis of Temporal and Geographic Variation by Using Large National Datasets. *Radiology* **279**, 175–184 (2016).

19. WHO | Global Maps for Diagnostic Imaging. WHO

https://www.who.int/diagnostic_imaging/collaboration/global_collab_maps/en/.

20. WHO | Global Atlas of medical devices. WHO http://www.who.int/medical_devices/publications/global_atlas_meddev2017/en/.

21. The Complexities of Physician Supply and Demand: Projections From 2018 to 2033. 92 (2018).

22. Amazon SageMaker Ground Truth Pricing | AWS. https://aws.amazon.com/sagemaker/groundtruth/pricing/.

23. Olvera-López, J., Carrasco-Ochoa, J., Martínez-Trinidad, J. F. & Kittler, J. A review of instance selection methods. *Artif Intell Rev* **34**, 133–143 (2010).

24. Joshi, A. J., Porikli, F. & Papanikolopoulos, N. Multi-class active learning for image classification. *2009 IEEE Conf. Comput. Vis. Pattern Recognit.* (2009) doi:10.1109/cvprw.2009.5206627.

25. Settles, B. Active Learning Literature Survey. 47.

26. Fu, Y., Zhu, X. & Li, B. A survey on instance selection for active learning. *Knowl. Inf. Syst.* **35**, 249–283 (2013).

27. Kumar, P. & Gupta, A. Active Learning Query Strategies for Classification, Regression, and Clustering: A Survey. *J. Comput. Sci. Technol.* **35**, 913–945 (2020).

28. Sahiner, B. *et al.* Deep learning in medical imaging and radiation therapy. *Med. Phys.* **46**, e1–e36 (2019).

29. Hoi, S., Jin, R., Zhu, J. & Lyu, M. Batch Mode Active Learning and Its Application to Medical Image Classiflcation. in vol. 2006 417–424 (2006).

30. Kermany, D. S. *et al.* Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell* **172**, 1122-1131.e9 (2018).

31. Coates, A., Ng, A. & Lee, H. An Analysis of Single-Layer Networks in Unsupervised

Feature Learning. in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* 215–223 (JMLR Workshop and Conference Proceedings, 2011).

32. Burgess, C. P. *et al.* Understanding disentangling in β-VAE. *ArXiv180403599 Cs Stat* (2018).

33. Jang, R. *et al.* Assessment of the Robustness of Convolutional Neural Networks in Labeling Noise by Using Chest X-Ray Images From Multiple Centers. *JMIR Med. Inform.* **8**, e18089 (2020).

34. Cohn, D., Atlas, L. & Ladner, R. Improving generalization with active learning. *Mach. Learn.* **15**, 201–221 (1994).

35. Wang, K., Zhang, D., Li, Y., Zhang, R. & Lin, L. Cost-Effective Active Learning for Deep Image Classification. *IEEE Trans. Circuits Syst. Video Technol.* **27**, 2591–2600 (2017).

36. Fang, M., Li, Y. & Cohn, T. Learning how to Active Learn: A Deep Reinforcement Learning Approach. *ArXiv170802383 Cs* (2017).

37. Arora, R. & Arnaout, R. *Private Antibody Repertoires Are Public*. 2020.06.18.159699 https://www.biorxiv.org/content/10.1101/2020.06.18.159699v1 (2020) doi:10.1101/2020.06.18.159699.

38. Arora, R., Burke, H. M. & Arnaout, R. *Immunological Diversity with Similarity*. 483131 https://www.biorxiv.org/content/10.1101/483131v1 (2018) doi:10.1101/483131.

Tables and Figures

Table 1. Overall training and test datasets.

	ECHO-F: A4C		ECHO-F: NT		ECHO-F-SEG		OCT: NL		OCT: CNV		STL10: AIR		STL10: TRUCK	
	Training	Test	Training	Test	Training	Test	Training	Test	Training	Test	Training	Test	Training	Test
No. images	20378	4365	25082	3618	1248	173	23468	3015	22696	14623	6059	800	4117	800
No. patients	652	80	281	51	186	20	3193	433	653	267				
No. video clips	1495	198	2849	764	299	48								

A4C, axial 4-chamber; NT, non-target; NL, normal; CNV, choroidal neovascularization; AIR, airplane

Table 2. Average training subsets.

Dataset	ECHO-F: NT										
% of full training set	2	7	11	22	33	44	55	66	77	88	100
No. images	556	1981	3467	7412	10648	13469	15891	18167	19739	19739	25082
No. patients	104	222	270	281	281	281	281	281	281	281	281
No. video clips	424	1346	2139	2741	2794	2817	2826	2,827	2833	2833	2849
Dataset	**ECHO-F: A4C**										
% of full training set	2	7	11	22	33	44	55	66	77	88	100
No. images	444	1018	1533	2588	4352	6531	9108	11833	15260	20260	20378
No. patients	266	502	629	652	652	652	652	652	652	652	652
No. video clips	347	882	1319	1480	1492	1495	1495	1495	1495	1495	1495

A4C, axial 4-chamber; NT, non-target

Dataset	ECHO-F-SEG								
% of full training set	16	24	32	48	64	80	88	96	100
No. images	200	300	400	600	800	1000	1100	1200	1248
No. patients	100	121	130	151	167	178	181	185	186
No. video clips	121	156	180	224	260	281	289	298	299

Dataset	OCT: NL										
% of full training set	1.3	3.2	5.4	15	26	32.5	34.6	41	54	65	100
No. images	209	301	410	1264	2888	4270	4766	6490	10601	14462	23468
No. patients	204	266	338	777	1403	1783	1898	2241	2776	3048	3193
Dataset	**OCT: CNV**										
% of full training set	1.3	3.2	5.4	15	26	32.5	34.6	41	54	65	100
No. images	384	1199	2090	4189	9112	10730	11234	12510	14399	15538	22696
No. patients	176	300	373	529	600	621	627	638	648	650	653

NL, normal; CNV, choroidal neovascularization

Dataset	STL10: AIRPLANE							
% of full training set	5	10	20	30	50	70	88	100
No. images	299	407	632	946	1886	3457	5361	6059
Dataset	**STL10: TRUCK**							
% of full training set	5	10	20	30	50	70	88	100
No. images	201	593	1368	2054	3113	3543	3638	4117

Figure 1. Cumulative density of maximum pairwise similarities.

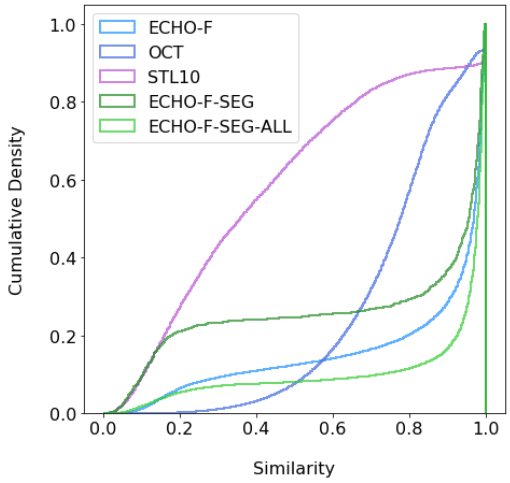

Figure 2. Experimental schematic.

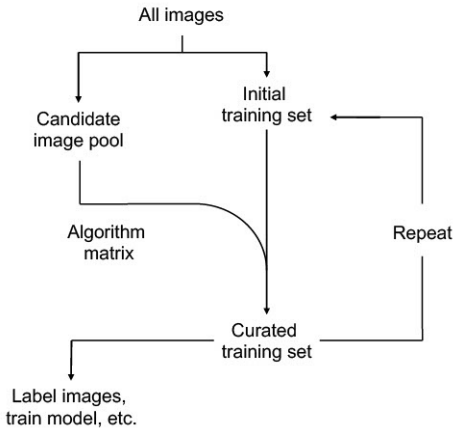

Figure 3. Performance of ENRICHed training datasets compared to randomly selected training datasets (a) ECHO-F classification, (b) ECHO-F-SEG segmentation (c) OCT classification, and (d) STL10 classification.

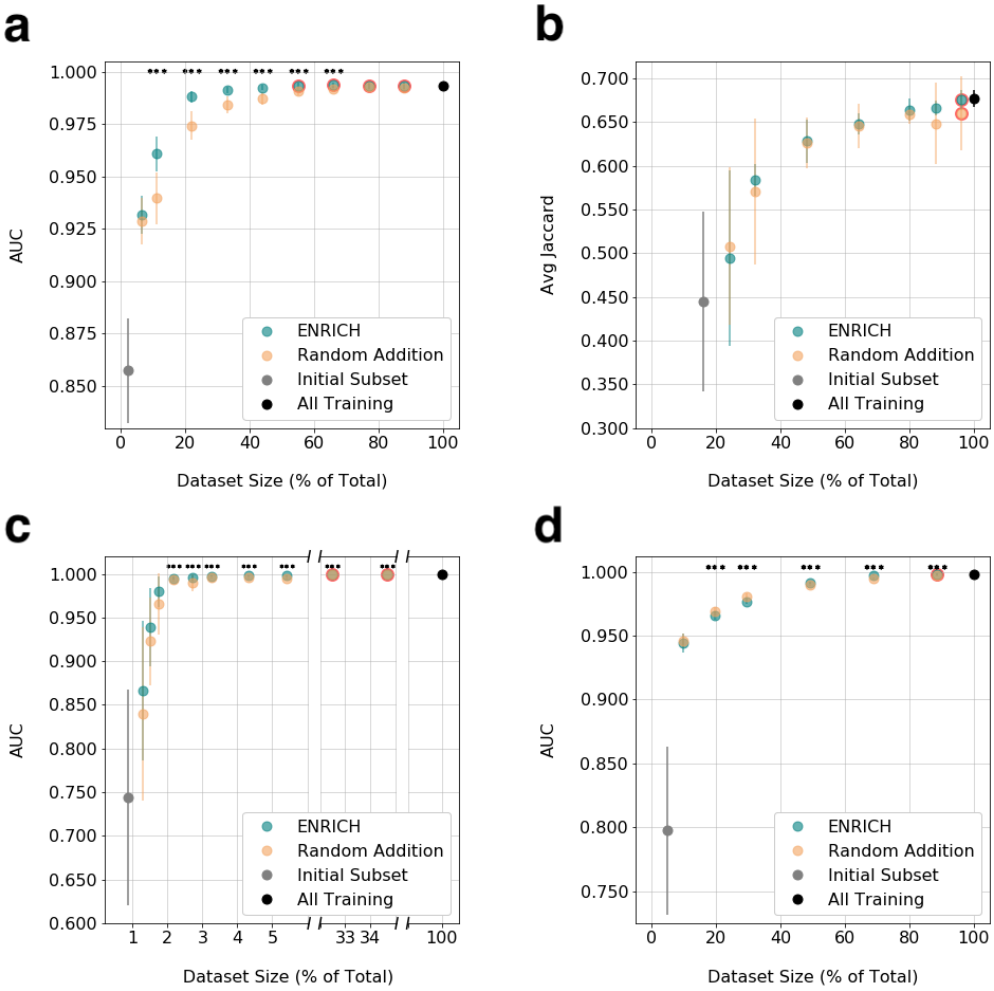

Figure 4. Labeling time savings using ENRICH.

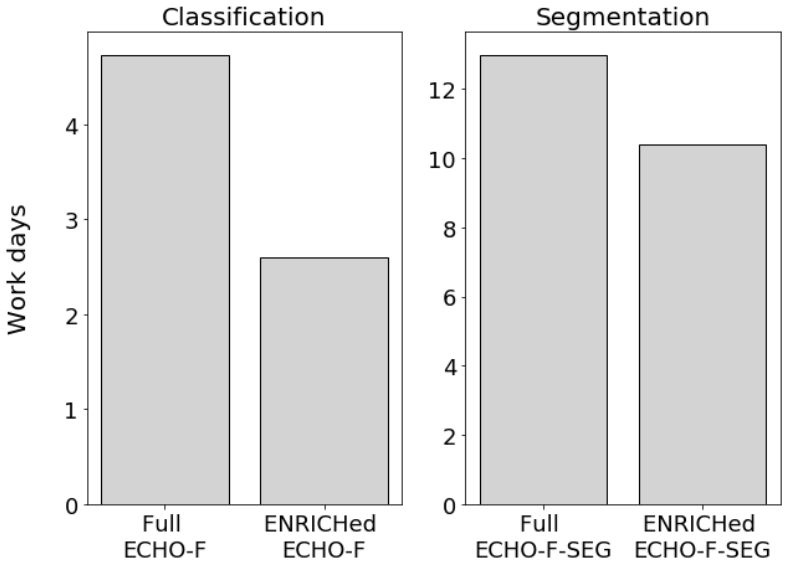

Figure 5. Standardized combined subsets demonstrate overall differences between ENRICH and random selection (a) ECHO-F classification, (b) ECHO-F-SEG segmentation (c) OCT classification, and (d) STL10 classification.