Title

ENRICHing medical imaging training sets enables more efficient machine learning.

Permalink

https://escholarship.org/uc/item/9vd24430

Journal

A Scholarly Journal of Informatics in Health and Biomedicine, 30(6)

Authors

Chinn, Erin
Arora, Rohit
Arnaout, Ramy
et al.

Publication Date

2023-05-19

DOI

10.1093/jamia/ocad055

Peer reviewed

Journal of the American Medical Informatics Association, 30(6), 2023, 1079–1090
https://doi.org/10.1093/jamia/ocad055
Advance Access Publication Date: 10 April 2023
Research and Applications

Research and Applications

ENRICHing medical imaging training sets enables more efficient machine learning

Erin Chinn[1], Rohit Arora[2], Ramy Arnaout[2,3], and Rima Arnaout[1]

[1]Department of Medicine, Division of Cardiology, Department of Radiology, Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, California, USA, [2]Division of Clinical Pathology, Department of Pathology, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA and [3]Division of Clinical Informatics, Department of Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA

Corresponding author: Rima Arnaout, MD, Department of Medicine, Division of Cardiology, Department of Radiology, Bakar Computational Health Sciences Institute, University of California, San Francisco, 521 Parnassus Avenue Rm 6222, San Francisco, CA 94143, USA; rima.arnaout@ucsf.edu

Ramy Arnaout and Rima Arnaout contributed equally to this work.

Received 13 December 2022; Revised 26 February 2023; Editorial Decision 28 February 2023; Accepted 15 March 2023

ABSTRACT

Objective: Deep learning (DL) has been applied in proofs of concept across biomedical imaging, including across modalities and medical specialties. Labeled data are critical to training and testing DL models, but human expert labelers are limited. In addition, DL traditionally requires copious training data, which is computationally expensive to process and iterate over. Consequently, it is useful to prioritize using those images that are most likely to improve a model's performance, a practice known as instance selection. The challenge is determining how best to prioritize. It is natural to prefer straightforward, robust, quantitative metrics as the basis for prioritization for instance selection. However, in current practice, such metrics are not tailored to, and almost never used for, image datasets.

Materials and Methods: To address this problem, we introduce ENRICH—Eliminate Noise and Redundancy for Imaging Challenges—a customizable method that prioritizes images based on how much diversity each image adds to the training set.

Results: First, we show that medical datasets are special in that in general each image adds less diversity than in nonmedical datasets. Next, we demonstrate that ENRICH achieves nearly maximal performance on classification and segmentation tasks on several medical image datasets using only a fraction of the available images and without up-front data labeling. ENRICH outperforms random image selection, the negative control. Finally, we show that ENRICH can also be used to identify errors and outliers in imaging datasets.

Conclusions: ENRICH is a simple, computationally efficient method for prioritizing images for expert labeling and use in DL.

Key words: deep learning, medical imaging, information theory, instance selection, data quality, data efficiency

INTRODUCTION

Deep learning (DL) models can classify images by disease or by the structure(s) they contain. They can also segment, track, and measure substructures within images.[1–17] DL thus has great promise for helping meet the overwhelming need for accurate, reliable, and scalable image interpretation that currently exists in medicine due to a near-universal shortage of trained human experts.[5,6,18–21] However, the data-hungry nature of DL model training threatens to hamper its effective use for medical imaging. First, the large amount of data required creates a costly and time-consuming labeling bottleneck for clinical experts.[3] (This is in contrast to labeling in nonmedical fields, which usually focuses on everyday objects and therefore can be performed more quickly and inexpensively by laypeople via crowdsourcing.[22]) Various unsupervised methods can help mitigate the labeling burden but cannot eliminate it, since experts are still needed to label test datasets in order to benchmark performance on high-stakes medical tasks. Second, large amounts of training data increase training times, slowing iteration during model development, requiring out-of-reach computational resources, or both. While early proofs of concept have used relatively small imaging datasets, the skyrocketing volume of medical image data promises to magnify these challenges.[23,24]

It has long been recognized that prioritizing the labeling of data that most benefit model performance, a practice known as *instance selection*, as opposed to random data, can be helpful for machine learning.[25,26] Instance selection methods generally balance some measure of the representativeness of a datapoint (instance) with some measure of how much that instance will add to the diversity of the resulting training set.[25] In contrast to *active learning*,[27,28] which requires iterative training of DL models to identify additional instances to include, instance selection algorithms are used once up front. Most work on instance selection has focused on nonimaging data and preceded recent developments in DL. In addition, instance selection methods described to date often require prior knowledge of the instance label and so would not reduce the data labeling burden. Therefore, how best to use instance selection for images, and how best to curate large image datasets in a label-free approach, are open questions.

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. Images from a particular medical domain often have similar subject matter (eg, the heart in cardiology, the retina in ophthalmology), pose (standard views), background (black), noise, lighting, and color. In the case of computed tomography, magnetic resonance imaging, ultrasound, and other common imaging modalities, image frames are often captured consecutively, resulting in similarity among images. For these reasons, we hypothesize that standardization in medical imaging creates greater redundancy in medical training data than in commonly used nonmedical datasets. We therefore propose that simply prioritizing nonredundant images is an efficient means of instance selection for DL in medical imaging.

As a test of this understanding, here we present a method called ENRICH: Eliminating Noise and Redundancy for Imaging Challenges. It consists 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 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 and size of the resulting training set without requiring up-front labeling, which can be laborious. We demonstrate proof of concept on classification and segmentation tasks on two large, well characterized medical datasets: ECHO-F,[3] which consists of fetal echocardiograms, and OCT,[29] which consists of adult retinal optical coherence tomography images. We also demonstrate the special nature of medical image datasets, showing differences in their pairwise similarities compared to STL10, a standard nonmedical image dataset used for various DL applications.[30]

MATERIALS AND METHODS

Datasets, tasks, and benchmarks

We searched for available datasets meeting the following criteria: minimum image size of 80 × 80 pixels; minimum image number 10 000 (classification) or 1000 (segmentation); not trivially simple (eg, MNIST); multiple labeled classes available; representing both nonmedical images as well as medical images of different anatomic structures, different imaging modalities, and different data structure (still images vs videos). Most cleaned, publicly available datasets have previously been cropped to square as part of their preparation. For consistency, where we encountered datasets where images were not already square, we removed any white edges from those images, found the center of the remining region, and cropped to a square region symmetric around it. Training and test sets are described below and in Table 1.

ECHO-F

ECHO-F consists of fetal echocardiogram images.[3] The binary classification task included the fetal axial four-chamber (A4C) view and the nontarget (NT) view. The multiclass task included the A4C and NT views as well as the fetal three-vessel view (3VV). In ultrasound, one or more clips are acquired per patient; each 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.

ECHO-F-SEG

ECHO-F-SEG consists of a subset of A4C images from ECHO-F. ECHO-F-SEG was used for multiclass segmentation with five classes: left ventricle, right ventricle, left atrium, right atrium, and background. Notably, ECHO-F-SEG had already been curated informally, in that only certain frames from each video clip were labeled.

OCT

OCT consists of adult retinal optical coherence tomography images.[29] Binary classification was between normal retina (NL) and choroidal neovascularization (CNV). The multiclass task included NL and CNV as well as drusen and diabetic macular edema. The train/test split of the dataset was adjusted from the authors' original description: instead of 250 images per lesion in the test set, we created disjoint train/test sets as we did for ECHO-F, split by patient ID, and increased the total size of the binary test set from 500 images to 17 638 images and the multiclass test set from 1000 images to 26 908 images to make the tests more difficult.

STL10

STL10 consists of images of animals and vehicles.[30] The binary classification task included images of airplanes (AIR) and trucks (TRUCK). The multiclass task included AIR, TRUCK, and ships/boats (SHIP).

Image processing

Grayscale conversion was done using Python3's OpenCV package. Image resizing was done using Python3's Scikit-Image package.

Table 1. Overall training and testing datasets

Dataset	Class	Training			Testing		
		Images	Clips	Patients	Images	Clips	Patients
ECHO-F	3VV	7159	837	500	890	96	51
	A4C	20 378	1495	652	3518	198	80
	NT	25 082	2849	281	4365	764	51
	Total[a]	52 619	4687	777	8773	976	89
ECHO-F-SEG	A4C	1248	299	186	173	48	20
OCT	NL	23 468	—	3193	3015	—	433
	CNV	22 696	—	653	14 623	—	267
	DME	6994	—	601	4550	—	166
	DRU	4144	—	574	4720	—	142
	Total[a]	57 302	—	4221	26 908	—	1254
STL10	AIR	6059	—	—	800	—	—
	TRUCK	4117	—	—	800	—	—
	SHIP	6600	—	—	800	—	—
	Total	16 776	—	—	2400	—	—

Abbreviations: 3VV: three-vessel view; A4C: apical four-chamber; AIR: airplane; CNV: choroidal neovascularization; DME: diabetic macular edema; DRU: drusen; NL: normal; NT: nontarget.

[a] Some clips and patients may contain more than one class of image.

ECHO-F

ECHO-F images were originally 300 × 400 pixel grayscale. For autoencoder input, the original images were cropped to square and resized to 64 × 64. For classification-model training and testing, images were cropped and resized to 80 × 80. For segmentation-model training and testing, original images were cropped to 272 × 272.

OCT

OCT images were originally grayscale and varied in pixel dimension. This dataset was standardized to correct region-of-interest misalignment and remove white-edge artifacts (discovered, in fact, by ENRICH). First, white sections at image edges were removed as above, then images underwent a cropping and resizing as above to allow images to be used with the autoencoder described.

STL10

STL10 images were originally 96 × 96 and converted to grayscale to for ease of comparison with the medical image datasets and resized as above.

Embeddings

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 except for the dimension of the embedding.[31] The β-VAE was trained on a subset of 5000 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, a matrix of pairwise image similarities was calculated. The similarity between two image embeddings was defined as the cosine similarity (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 DL task, an initial subset of images was chosen at random. For each image i in the remaining dataset, the maximum similarity to each image j in the training set, $\max_j(z_{ij})$, was read from the similarity matrix, and the image i with the smallest $\max_j(z_{ij})$ was added to the training set (ie, $\mathrm{argmin}_i(\max_j(z_{ij}))$). This step was iterated to grow the training set, and the quality of the training set was assessed at specific sizes by training a model and measuring its performance (below). The ranking algorithm was blind to class label. For statistical confidence, the ranking algorithm, including choice of the initial subset, was repeated three times for each task ("biological" replicates), and 10 models were trained on each resulting training set (technical replicates). Training subsets are summarized in Tables 2–4.

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 the classification tasks. Experiments for each dataset used the same model parameters throughout.

Human labeling time estimates

Human labeling time averaged across $n = 4$ labelers using several different labeling platforms was 3 seconds per image for classification in ECHO-F and 5 min per image for ECHO-F-SEG segmentation.

Evaluation metrics

For the binary classification tasks, model performance was assessed using the area under the receiver operator characteristic curve (AUCROC). For the multiclass tasks, a per class one-versus-rest AUCROC was averaged and compared between models. For the segmentation task, average Jaccard score of the four heart segments (left ventricle, right ventricle, left atrium, right atrium) was used, as previously described.[3] One-sided t tests were used to compare performance across experiments.

Several datasets contained hierarchical levels of organization, image < clip < patient. Representativeness of each training subset by level as applicable was calculated as a percentage of each of these in overall training set. Class balance was calculated as the effective number[32] form Shannon entropy, that is, exponentiating the Shannon entropy of the classes. At each level, class balance was calculated from the distribution of the number of unique images/clips/patients in each class.

RESULTS

Dataset diversity plots and the dataset diversity score

We developed a new way to visualize the diversity or redundancy of a dataset by plotting a cumulative histogram of the values in the similarity matrix, resulting in a *dataset diversity plot* (Figure 1). In such plots, high-diversity datasets—ones in which the images are very different from each other—will trace out curves to the upper left, while high-redundancy datasets will trace out curves to the lower right. The area under the diversity curve ranges from 0 to 1: 0 for datasets that are completely redundant and 1 for datasets that are maximally diverse. Thus, we used the area under the diversity curve as a natural a *dataset diversity score*, as we introduce and illustrate in Figure 1A. We created diversity plots and dataset diversity scores for each

Table 2. Select training subsets, ECHO-F

No. images	1000[a]	3000	5000	10 000	15 000	20 000	25 000	30 000	35 000	40 000	45 460
ECHO-F binary											
Pct. Images	2	7	11	22	33	44	55	66	77	88	100
No. images A4C (avg ± sd)	444 ± 7	1019 ± 7	1533 ± 11	2588 ± 8	4352 ± 22	6531 ± 13	9108 ± 160	11833 ± 18	15260 ± 8	20260 ± 8	20378 ± 0
No. images NT (avg ± sd)	556 ± 7	1981 ± 7	3467 ± 11	7412 ± 8	10648 ± 22	13469 ± 13	15891 ± 160	18167 ± 18	19739 ± 8	19739 ± 8	25082 ± 0
No. clips A4C (avg ± sd)	347 ± 16	882 ± 21	1319 ± 1	1480 ± 2	1493 ± 1	1495 ± 0	1495 ± 0	1495 ± 0	1495 ± 0	1495 ± 0	1495 ± 0
No. clips NT (avg ± sd)	424 ± 13	1346 ± 11	2139 ± 9	2741 ± 3	2794 ± 2	2810 ± 1	2817 ± 3	2826 ± 1	2833 ± 1	2833 ± 1	2849 ± 0
No. clips all classes (avg ± sd)	770 ± 9	2186 ± 12	3381 ± 7	4109 ± 3	4171 ± 2	4190 ± 1	4197 ± 3	4206 ± 1	4213 ± 1	4213 ± 1	4229 ± 0
No. patients A4C (avg ± sd)	266 ± 14	502 ± 8	630 ± 1	652 ± 0	652 ± 0	652 ± 0	652 ± 0	652 ± 0	652 ± 0	652 ± 0	652 ± 0
No. patients NT (avg ± sd)	104 ± 1	222 ± 6	269 ± 1	281 ± 1	281 ± 0	281 ± 0	281 ± 0	281 ± 0	281 ± 0	281 ± 0	281 ± 0
No. patients all classes (avg ± sd)	337 ± 10	592 ± 9	697 ± 2	713 ± 0	713 ± 0	713 ± 0	713 ± 0	713 ± 0	713 ± 0	713 ± 0	713 ± 0
No. images	1000[a]	3000	5000	10 000	15 000	20 000	25 000	30 000	35 000	40 000	52 619
ECHO-F multiclass											
Pct. images	2	6	10	19	29	38	48	57	67	76	100
No. images 3VV (avg ± sd)	139 ± 2	461 ± 17	742 ± 7	1174 ± 2	1728 ± 16	2391 ± 6	3899 ± 8	3899 ± 8	4788 ± 1	6937 ± 13	7159 ± 0
No. images A4C (avg ± sd)	381 ± 6	847 ± 17	1312 ± 4	2306 ± 8	3666 ± 13	5976 ± 28	7388 ± 5	10593 ± 33	12993 ± 25	13244 ± 28	20378 ± 0
No. images NT (avg ± sd)	480 ± 6	1692 ± 7	2946 ± 4	6520 ± 9	9606 ± 14	11633 ± 25	13713 ± 9	15508 ± 26	17219 ± 24	19819 ± 18	25082 ± 0
No. clips 3VV (avg ± sd)	118 ± 6	423 ± 9	676 ± 3	832 ± 1	836 ± 0	837 ± 1	837 ± 0	837 ± 0	837 ± 0	837 ± 0	837 ± 0
No. clips A4C (avg ± sd)	307 ± 3	741 ± 15	1155 ± 2	1476 ± 2	1489 ± 2	1494 ± 1	1495 ± 1	1495 ± 1	1495 ± 0	1495 ± 0	1495 ± 0
No. clips NT (avg ± sd)	378 ± 5	1181 ± 11	1907 ± 5	2697 ± 6	2783 ± 2	2798 ± 2	2809 ± 1	2817 ± 1	2820 ± 1	2830 ± 0	2849 ± 0
No. clips all classes (avg ± sd)	792 ± 8	2242 ± 10	3494 ± 5	4523 ± 6	4618 ± 1	4635 ± 0	4647 ± 1	4655 ± 1	4658 ± 0	4668 ± 0	4687 ± 0
No. patients 3VV (avg ± sd)	103 ± 8	313 ± 5	444 ± 6	500 ± 0	500 ± 0	500 ± 0	500 ± 0	500 ± 0	500 ± 0	500 ± 0	500 ± 0
No. patients A4C (avg ± sd)	246 ± 5	447 ± 10	588 ± 3	652 ± 1	652 ± 0	652 ± 0	652 ± 0	652 ± 0	652 ± 0	652 ± 0	652 ± 0
No. patients NT (avg ± sd)	98 ± 9	206 ± 2	251 ± 1	281 ± 0	281 ± 0	281 ± 0	281 ± 0	281 ± 0	281 ± 0	281 ± 0	281 ± 0
No. patients all classes (avg ± sd)	371 ± 7	629 ± 5	746 ± 3	777 ± 0	777 ± 0	777 ± 0	777 ± 0	777 ± 0	777 ± 0	777 ± 0	777 ± 0
No. images	200[a]	300	400	600	800	1000	1050	1100	1150	1200	1248
ECHO-F-SEG											
Pct. images	16	24	32	48	64	80	84	88	92	96	100
No. images A4C (avg ± sd)	200 ± 0	300 ± 0	400 ± 0	600 ± 0	800 ± 0	1000 ± 0	1050 ± 0	1100 ± 0	1150 ± 0	1200 ± 0	1248 ± 0
No. clips A4C (avg ± sd)	121 ± 4	156 ± 8	180 ± 7	224 ± 7	260 ± 5	281 ± 2	285 ± 3	289 ± 2	294 ± 1	298 ± 1	299 ± 0
No. patients A4C (avg ± sd)	100 ± 3	121 ± 7	130 ± 6	151 ± 5	167 ± 1	178 ± 2	179 ± 2	181 ± 2	184 ± 1	185 ± 0	186 ± 0

Abbreviations: 3VV: three-vessel view; A4C: axial four-chamber; NT: nontarget.

[a] Initial “seed” subset.

Table 3. Select training subsets, OCT

No. images	500[a]	1000	1500	2000	5000	7500	10000	15000	20000	25000	30000	46164			
OCT binary															
Pct. images	1	2.2	3.2	4.3	10.8	16.2	21.7	32.5	43.3	54.2	65.0	100			
No. images CNV (avg ± sd)	291 ± 7	750 ± 5	1199 ± 2	1649 ± 6	4189 ± 15	6109 ± 27	7870 ± 26	10730 ± 7	12879 ± 10	14399 ± 13	15538 ± 6	22696 ± 0			
No. images NORMAL (avg ± sd)	209 ± 7	250 ± 5	301 ± 2	351 ± 6	811 ± 15	1391 ± 27	2130 ± 26	4270 ± 7	7121 ± 10	10601 ± 13	14462 ± 6	23468 ± 0			
No. patients CNV (avg ± sd)	151 ± 4	244 ± 6	300 ± 10	336 ± 8	482 ± 3	540 ± 5	579 ± 3	621 ± 2	641 ± 1	648 ± 1	650 ± 1	653 ± 0			
No. patients NORMAL (avg ± sd)	198 ± 6	231 ± 1	266 ± 2	300 ± 4	571 ± 11	837 ± 5	1133 ± 17	1783 ± 6	2337 ± 5	2776 ± 17	3048 ± 2	3193 ± 0			
No. patients all classes (avg ± sd)	344 ± 9	465 ± 6	553 ± 11	621 ± 10	1011 ± 9	1309 ± 12	1625 ± 20	2268 ± 2	2808 ± 6	3232 ± 20	3494 ± 2	3635 ± 0			
OCT multiclass															
Pct. images	0.9	1.7	2.6	3.5	8.7	13.1	17.5	26.2	34.9	43.6	52.4	61.1	69.8	78.5	100
No. images	500[a]	1000	1500	2000	5000	7500	10 000	15 000	20 000	25 000	30 000	35 000	40 000	45 000	57 302
No. images CNV (avg ± sd)	195 ± 13	461 ± 21	758 ± 15	1067 ± 4	3025 ± 14	4592 ± 10	6096 ± 22	8785 ± 17	10969 ± 22	12664 ± 2	13981 ± 8	15005 ± 6	15927 ± 4	17003 ± 10	22696 ± 0
No. images DME (avg ± sd)	54 ± 13	269 ± 20	442 ± 10	601 ± 12	1361 ± 11	1901 ± 6	2410 ± 12	3256 ± 12	3930 ± 19	4499 ± 19	4937 ± 10	5290 ± 3	5605 ± 6	5907 ± 3	6994 ± 0
No. images DRUSEN (avg ± sd)	36 ± 4	36 ± 4	39 ± 3	47 ± 3	88 ± 3	139 ± 7	210 ± 3	452 ± 9	799 ± 20	1272 ± 25	1787 ± 5	2302 ± 18	2768 ± 13	3199 ± 5	4144 ± 0
No. images NORMAL (avg ± sd)	215 ± 4	234 ± 6	260 ± 9	286 ± 12	526 ± 12	868 ± 8	1284 ± 18	2507 ± 20	4302 ± 10	6565 ± 11	9295 ± 9	12402 ± 11	15700 ± 13	18891 ± 15	23468 ± 0
No. patients CNV (avg ± sd)	126 ± 8	191 ± 7	245 ± 4	284 ± 5	425 ± 5	493 ± 8	539 ± 6	593 ± 2	623 ± 1	639 ± 5	646 ± 3	649 ± 1	652 ± 1	652 ± 0	653 ± 0
No. patients DME (avg ± sd)	45 ± 10	121 ± 6	164 ± 8	200 ± 5	317 ± 4	376 ± 5	433 ± 2	489 ± 0	528 ± 4	561 ± 0	578 ± 2	592 ± 2	597 ± 1	600 ± 1	601 ± 0
No. patients DRUSEN (avg ± sd)	34 ± 3	34 ± 3	37 ± 1	40 ± 3	54 ± 1	73 ± 7	103 ± 5	194 ± 6	289 ± 8	387 ± 4	452 ± 4	501 ± 4	540 ± 6	567 ± 2	574 ± 0
No. patients NORMAL (avg ± sd)	202 ± 5	218 ± 6	238 ± 9	256 ± 13	416 ± 12	602 ± 7	804 ± 11	1273 ± 16	1791 ± 35	2255 ± 18	2635 ± 15	2909 ± 11	3088 ± 11	3181 ± 3	3193 ± 0
No. patients all classes (avg ± sd)	392 ± 8	538 ± 6	649 ± 3	737 ± 9	1114 ± 11	1401 ± 10	1675 ± 11	2223 ± 14	2783 ± 26	3277 ± 10	3665 ± 10	3938 ± 13	4118 ± 11	4210 ± 2	4221 ± 0

Abbreviations: CNV: choroidal neovascularization; DME: diabetic macular edema; DRU: drusen; NL: normal.

[a] Initial "seed" subset.

Table 4. Select training subsets, STL10

No. images	500[a]	1000	2000	3000	5000	7000	9000	10 176		
STL10 binary										
Pct. images	4.9	9.8	19.7	29.5	49.1	68.8	88.4	100		
No. images AIRPLANE (avg ± sd)	299 ± 6	407 ± 5	632 ± 7	946 ± 4	1887 ± 3	3457 ± 1	5362 ± 1	6059 ± 0		
No. images TRUCK (avg ± sd)	201 ± 6	593 ± 5	1368 ± 7	2054 ± 4	3113 ± 3	3543 ± 1	3638 ± 1	4117 ± 0		
No. images	500[a]	1000	2000	3000	5000	7000	9000	11 000	13 000	16 776
STL10 multiclass										
Pct. images	3	6	11.9	17.9	29.8	41.7	53.6	65.6	77.5	100
No. images AIRPLANE (avg ± sd)	183 ± 14	260 ± 15	438 ± 7	623 ± 11	1087 ± 16	1706 ± 8	2436 ± 33	3323 ± 10	4290 ± 10	6059 ± 0
No. images SHIP (avg ± sd)	194 ± 7	329 ± 8	566 ± 21	857 ± 8	1531 ± 18	2304 ± 14	3215 ± 32	4160 ± 10	5109 ± 10	4117 ± 0
No. images TRUCK (avg ± sd)	123 ± 9	411 ± 14	996 ± 16	1520 ± 3	2381 ± 3	2990 ± 7	3349 ± 6	3517 ± 4	3601 ± 1	6600 ± 0

[a] Initial "seed" subset.

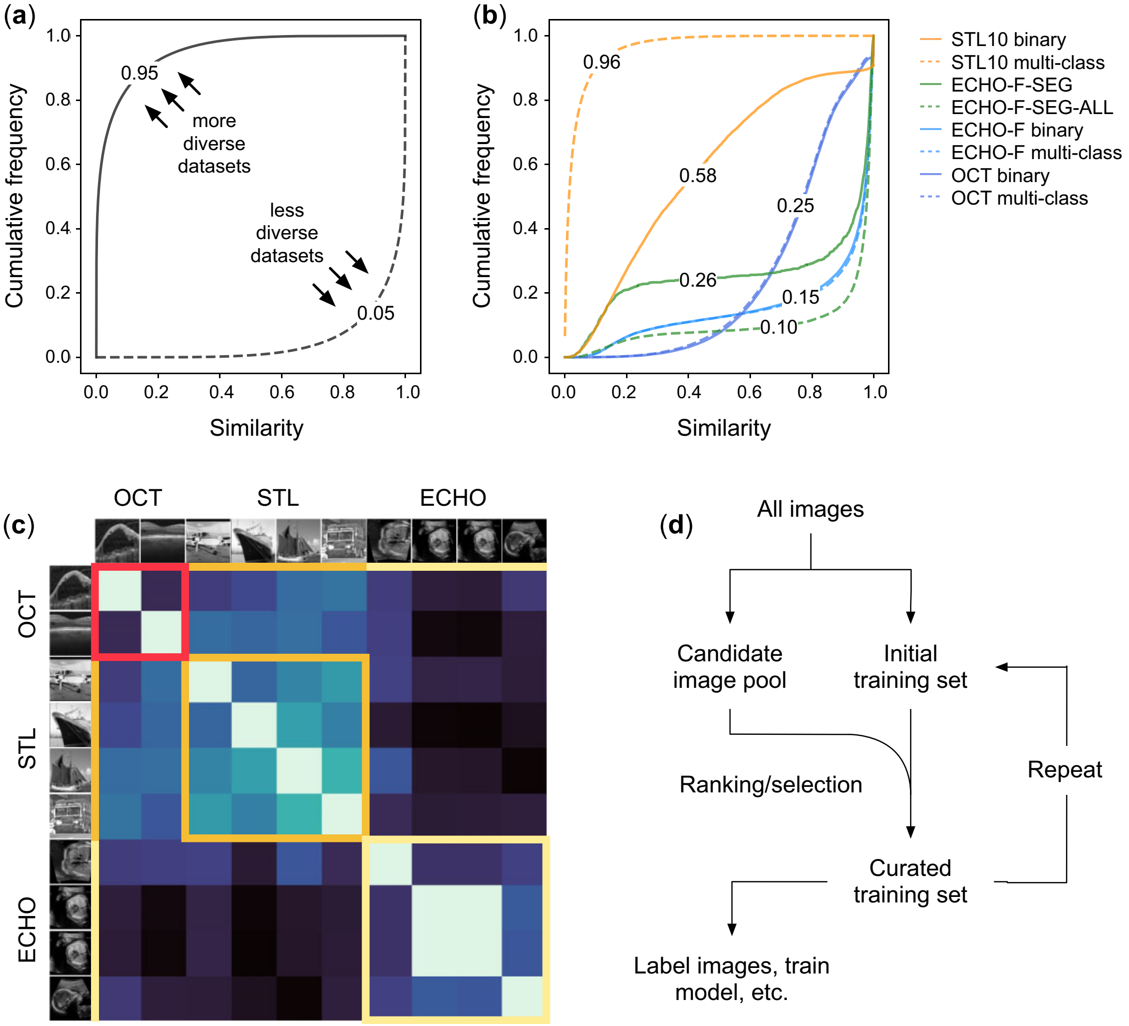

Figure 1. Similarity in imaging datasets and experimental approach. A, Schematic of the dataset diversity plot, a cumulative density plot of maximum pairwise similarities. Dataset diversity scores are indicated. B, Dataset diversity plots and scores for ECHO-F, OCT, STL10, and ECHO-F-SEG datasets. Also, included are the total images available for the ECHO-F segmentation task, ECHO-F-SEG-ALL. C, Pairwise image similarities in a handful of images drawn from OCT, STL10, and ECHO-F. Red-, orange-, and yellow-bordered squares indicate similarities within the OCT, STL, and ECHO datasets, respectively. D, Schematic summary of ENRICH. From all available images in a dataset, an initial training set is chosen at random. The remaining images comprise a candidate pool of images from which additional images can be selected. A matrix of pairwise image similarities (step 1 of ENRICH) is constructed. From this matrix, an algorithm is used to choose additional images to add to the initial training set; this is step 2 of ENRICH. This process is repeated, iteratively adding images to an initial subset.

dataset in our study (Figure 1B). These plots revealed significant redundancy in the medical image datasets ECHO-F and OCT (Figure 1B, blue and green), and higher redundancy in these medical datasets than in the nonmedical STL10 (Figure 1B, yellow).

Image redundancy in medical datasets

Our prior work suggested that medical image datasets are often quite redundant compared to nonmedical datasets[3] 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 calculated the maximum pairwise similarity for each image, with a similarity measure based on the cosine similarity between β-VAE embeddings (Materials and Methods). Supporting this hypothesis, we found that the majority of ECHO-F classification images had a maximum similarity greater than 0.9: most images had at least one other image in the dataset to which they were at least 90% similar (Figure 1B). The OCT dataset exhibited the same phenomenon, with roughly half of the images having a maximum pairwise similarity greater than 0.8. In contrast, most images in the nonmedical STL10 dataset had a maximum similarity less than 0.4 (Figure 1B).

Comparing ENRICH-curated training subsets versus control

ENRICH involves ranking images based on their similarity to each other and preferentially choosing the most unique—that is, lowest similarity—images for inclusion in the training set (Materials and Methods). Labeling is not required. For each dataset, we compared the performance of models trained on ENRICHed subsets of the dataset to those trained on control subsets, that is, subsets created by random sampling of the full dataset. We further compared the performance of both ENRICHed and control subsets of different sizes to the performance of models trained on the full training set, that is, all available training images, as the gold standard. For each dataset and task, we recorded whether and at what size ENRICHed subsets outperformed control subsets, as well as the minimum size at which ENRICHed subsets performed indistinguishably from the gold standard (full dataset). We tested binary and multiclass classification as well as segmentation, with replicates for statistical confidence (Figure 2).

ENRICH achieved gold-standard performance with substantially smaller training sets and outperformed unbiased selection

ENRICHed subsets achieved gold-standard performance using training subsets there were only a fraction of the size of the full training set (Figure 2). Specifically, ENRICH required only 55% and 48% of available images for ECHO-F binary and multiclass tasks, respectively (Figure 2). In contrast, random sampling failed to reach this benchmark at any training set size short of the full dataset. ENRICH outperformed control on even the smallest subsets, for example, 11% of ECHO-F in the binary classification task, 9.5% of ECHO-F in multiclassification (P-values $<2 \times 10^{-9}$; Figure 2A and B).

ENRICH discovers dataset structure without labels

Like many medical datasets, ECHO-F has hierarchical structure, with individual video frames stratified by clip and patient. We found that ENRICH selected images that represented almost all of the available patients and video clips even at small subset sizes, significantly more so than random sampling (all P-values $<1 \times 10^{-4}$) (Table 2 and Figure 2A and B). The largest magnitude in those differences were evident in the smallest training subsets that first showed performance gains over random selection.

ENRICH achieves class balance without labels

Class balance was measured as the effective number[32–35] of classes at the image, clip, and patient levels (Figure 2). Effective number is a mathematically rigorous standard that takes similarities and frequencies into account. (For example, if the frequencies of two classes in a toy dataset were highly imbalanced at 0.99 and 0.01, the effective number of classes will be close to one, since in effect, only the first class is represented.[32] In such a case, one would want to use some class balancing technique on the data to bring the effective number of classes closer to two.) Using Shannon entropy (see Materials and Methods), we calculated the effective number of classes represented by the images, clips, and patients in the full training set and in each ENRICHed and control subset.

The number of images in the full training set for ECHO-F were nearly equally balanced between classes, and indeed the effective number of classes in the binary classification task is approximately two (1.99). The effective number of classes at the clip and patient levels in the full training set are 1.88 and 1.57, respectively. While random sampling saw image-level effective size stay constant and clip- and patient-level effective sizes decrease approximately linearly, ENRICHed training sets were significantly less diverse at image, clip, and image levels, most notably at the 11% and 22% training subsets where ENRICH first pulled ahead in binary classification performance (Figure 2A; P-values all $<5 \times 10^{-4}$). Similar behavior was present in the 9.5% subset in the multiclassification task (Figure 2B; all P-values $<.001$). For both tasks, imbalance favored the NT class (Table 2), which clinically is felt to be more diverse, since it can contain any nonheart image, while the A4C class contains only A4C images of the heart. Thus, ENRICH selectively enriched training sets for the more diverse class, resulting in better performance, as opposed to blindly maintaining class balance.

Performance on additional datasets

Model test performance on OCT binary classification achieved a mean AUCROC of 0.99 ($\pm 2.24 \times 10^{-5}$) when trained on the full training dataset. ENRICH outperformed control training sets at just 2% of all OCT images (mean AUCROC 0.995 vs 0.993, P-value 9.98×10^{-6}). Only 32.5% of the training dataset was needed to achieve gold-standard performance when training images were chosen using ENRICH versus 41% for the control. The OCT dataset structure had images and patients, but not clips (Table 1). At a patient level, the effective number of classes is only 1.5 due to more patients in the NORMAL class than in the CNV class. For OCT, the effective number of classes patients represented in ENRICHed subsets was higher even as control subsets had a higher effective number of classes among images (Figure 2D). As Table 2 shows, this is because ENRICH selected fewer frames and patients from the overrepresented NORMAL class as it chose training subsets that outperformed control sets. When trained on the full training dataset, OCT multiclass performance achieved an AUCROC of $0.99 \pm 2 \times 10^{-4}$. Neither ENRICHed nor control subsets were able to achieve this benchmark. ENRICH outperformed control subsets at just 6% of all OCT images (P-value 2.49×10^{-9}), with similar findings with representativeness and effective number of classes as in the binary task (Figure 2E and Table 2).

ECHO-F-SEG multiclass segmentation

The ECHO-F-SEG dataset contains only two images per clip—an intuitive decision to economize on labeling. We evaluated training subsets chosen via ENRICH on multiclass segmentation. Using all available training data (Table 1), average Jaccard index was 0.68. With 80% of the training data, ENRICH achieved an average Jaccard of 0.66 (Figure 2C). ENRICH did not statistically significantly outperform random selection. Consistent with this, clip-level and patient-level representativeness was statistically indistinguishable between training subsets curated by ENRICH versus not. (Effective number of classes in this dataset is 1 by definition.)

Model test performance on binary classification in STL10 achieved a mean AUCROC of $0.99 \pm 2.04 \times 10^{-4}$ when trained on the full training dataset. In contrast to the medical datasets above, initially, control subsets narrowly but statistically outperformed ENRICH (at 20%; P-value 6×10^{-4}). At 50% of all STL10 images, the trend reversed, and ENRICH narrowly outperformed control subsets (P-value 5×10^{-6}) and continued to outperform it as sample size increased. Ninety percent of the total dataset was needed in order to achieve gold-standard performance (P-value .42; Figure 2F). We were not able to achieve the same benchmark without ENRICH. For multiclass classification with STL10, model test performance achieved a mean AUCROC of 0.99 ($\pm 5.57 \times 10^{-5}$) when trained on the full training dataset. Neither subset selection method, ENRICH nor control, was able to achieve the same benchmark. ENRICH was outperformed except at 30% and 40% of total, at which sizes performance was indistinguishable. For both binary and multiclass tasks, representativeness and class balance for STL10 were only present at the image level (Figure 2F and G).

Potential time savings in labeling

W estimated the time required to label all the images in ECHO-F for classification and ECHO-F-SEG for segmentation tasks, from

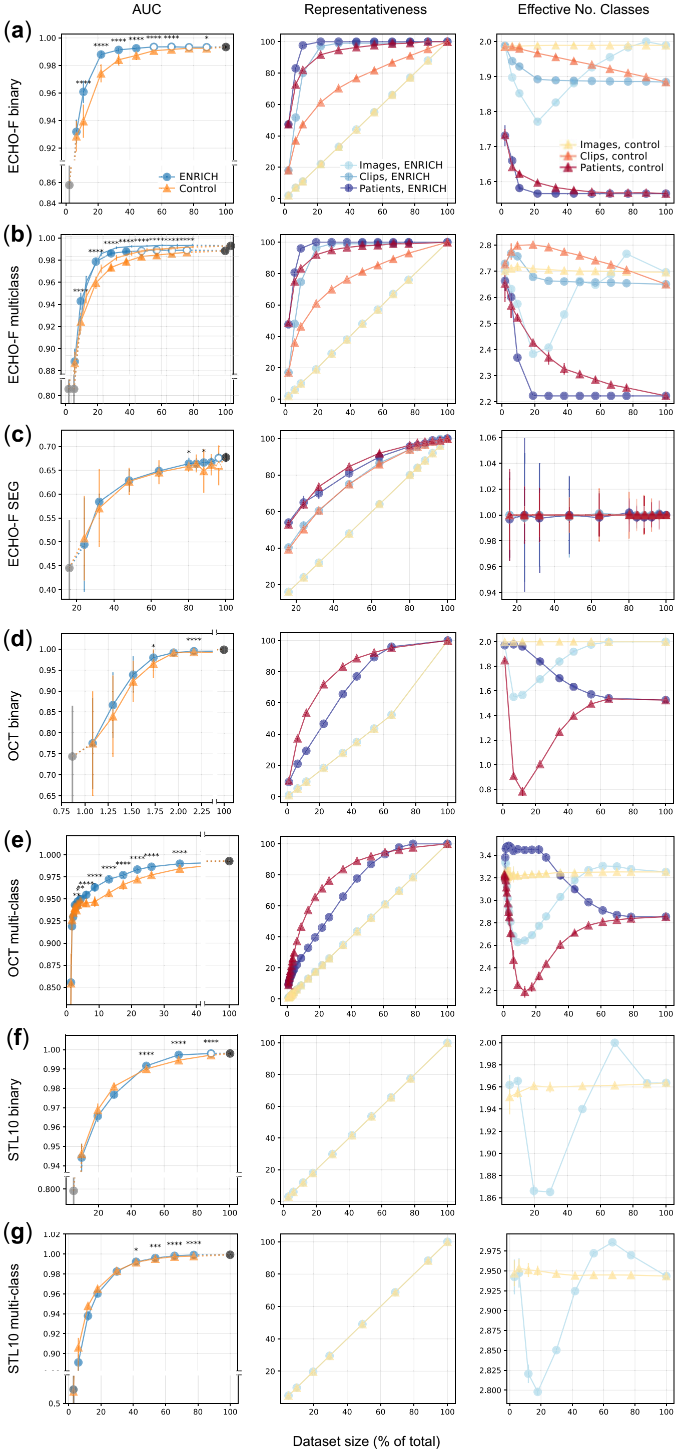

Figure 2. Performance of ENRICHed training datasets compared to randomly selected training datasets. (A) ECHO-F binary, (B) ECHO-F multiclass, (C) ECHO-F-SEG segmentation, (D) OCT binary, (E) OCT multiclass, (F) STL10 binary, and (G) STL10 multiclass. Each panel shows test performance on top, representativeness of images in the middle, and effective class size on the bottom. Performance testing, top: from a common initial random starting dataset (gray), additional images were added to grow increasingly larger training subsets using ENRICH (blue circle) versus random addition (yellow triangle). Each datapoint represents mean AUCROC on the test set from 30 replicates; error bars for each datapoint show one standard deviation around the mean. Asterisks for each training data subset represent statistical differences between ENRICH and random according to the standard convention (*ns* = $P > .05$; * = $P \leq .05$; ** = $P \leq .01$; *** = $P \leq .001$; **** = $P \leq .0001$). Empty symbols are statistically indistinguishable from model performance using the full training set (100% of training images; black dot). Representativeness, middle: for ENRICH (cool colors, circles) and random selection (warm colors, triangles), for each training subset, the percentage of the total training set is shown at an image (light blue circle, light yellow triangle), clip (medium blue circle, orange triangle), and patient (dark blue circle, red triangle) levels where applicable). Effective number of classes, bottom: for ENRICH (cool colors, circles) and random selection (warm colors, triangles), for each training subset, the effective number of classes is shown at an image (light blue circle, light yellow triangle), clip (medium blue circle, orange triangle), and patient (dark blue circle, red triangle) levels where applicable). For representativeness and effective size as well, error bars are shown but are small, and relevant *P*-values are summarized in the text.

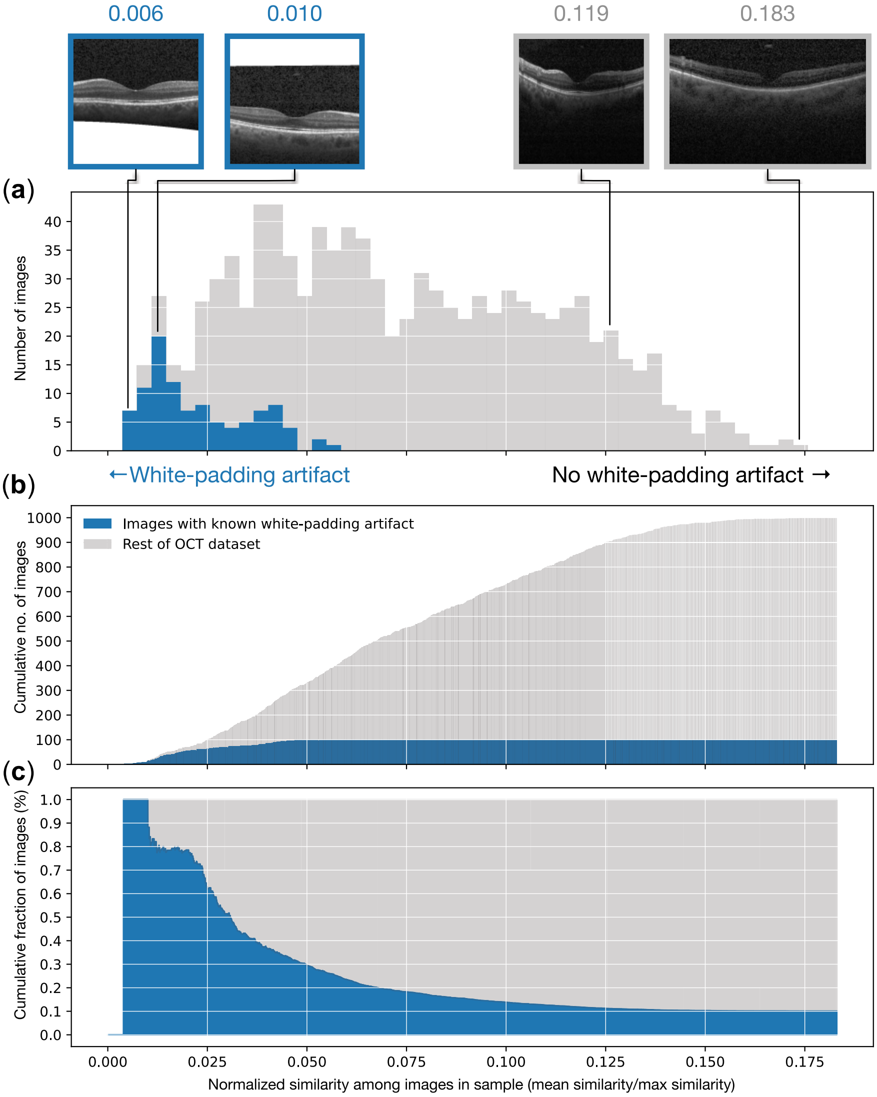

Figure 3. ENRICH aids in screening medical datasets for artifacts. A pairwise-similarity matrix was constructed from a sample of 1000 images in OCT. For each image in the matrix, a mean of the similarities to all other images (one row of the matrix) was calculated and normalized by the maximum similarity across the entire matrix. A, A stacked-bar histogram of these values, where images most different from the others are to the left, and most similar images are to the right. Blue (darker color) indicates images known to have a white-padding artifact; two examples are shown above, with their mean/max ratio as indicated. (B) Stacked cumulative distribution and (C) cumulative fraction of images in the sample, demonstrating how mean/max ratio of image similarities facilitates identification of images with artifacts. For example, in this thousand-image sample, about 10% of images have the white-padding artifact.

having measured the time it took to manually label these datasets for their original use. We compared this to the time that would have been required for the smallest ENRICHed subsets that achieved desired performance (55% for classification and 80% for segmentation). Using an ENRICHed training subset would have conferred a savings of 38 hours of full-time work, nearly an entire working week, for an expert labeler, on even this relatively small dataset. Labeling time on OCT and STL10 is unknown.

ENRICH identifies outliers in image datasets

ENRICH preferentially identifies outlier images, a property that can be exploited as a preprocessing step to screen for noise in large image datasets. In the OCT dataset, for example, we found about 10% of images to be similar to each other but different from the rest of images. Investigating this showed that these images had a white-space padding artifact (Figure 3) that is important to address (eg, by testing for segregation by class or by using preprocessing or data augmentation) in subsequent model training.

Taken together, these data demonstrate that ENRICH can outperform random selection, curating a high-performance training subset that is a fraction of the entire training dataset; and it does so by recognizing structure in the dataset (images, clips, patients) and optimizing both representativeness and class balance for this structure, even though this information is not an explicit part of the current implementation of ENRICH. When such structure does not exist in the dataset—either purposefully removed from medical image data, or in nonmedical datasets we encountered—the performance of the current implementation of ENRICH was more modest (Figure 2C, F, and G).

DISCUSSION

In DL for medical imaging, investigators generally rely on a crude metric for dataset quality and content: the number of images in the dataset. With ENRICH, we offer a mathematically rigorous and scalable way to look beyond size to dataset content, a practice that has largely been overlooked or only intuited in DL for medical imaging to date.

Instance selection provides a general strategy for labeling training datasets efficiently. ENRICH curates medical image datasets based on pairwise 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. While the canonical teaching is to split training and test sets by patient, in ECHO-F, most redundancy is at clip level, while in OCT, it appears to be above patient level (perhaps explaining why such a small portion of training data was needed to approximate full training set performance). In addition, there is a diversity to different classes which mere number of images does not fully describe.

Using ENRICH demonstrated that (1) redundant images do not aid significantly in DL model training, (2) this behavior is a property of the image dataset rather than the DL task (binary classification, multiclassification, and segmentation were tested), (3) image labels are not needed in order to curate image datasets according to redundancy nor to optimize representativeness and class balance, (4) images with artifacts can be systematically screened for using pairwise similarities, and (5) for some medical datasets, state-of-the-art performance can be achieved using only a fraction of the full training dataset.

For scientific rigor, we demonstrated these findings across several different image datasets, with several replicates per experiment totaling over two thousand model trainings. In some of the experiments presented, performance differences were small in magnitude (but still statistically significant); however, even these can translate to significant dataset savings. Furthermore, performance improvements of a few percentage points often redefine the state of the art when it comes to DL architectures.

There are several points in the above experiments worth mention. First, we note that classification-model training demonstrated here did not include standard data augmentations. This choice was made in order to remove data augmentation as a potential confounding factor in measuring the performance of ENRICH. However, in the future, data augmentation can be applied to ENRICH training subsets at the point when they first outperform random selection (eg, approximately 11% of ECHO-F and 2% of OCT), requiring even fewer images to meet optimum test performance.

Second, ENRICH was less helpful for the ECHO-F segmentation task studied here than for the classification tasks. However, for this task, the ECHO-F-SEG dataset had already been intuitively ENRICHed, as only a few image frames per clip were chosen. In this setting, the finding that an additional 20% of the already-intuitively reduced dataset was not needed to reach full dataset performance is still an additional gain in efficiency over informal curation and is another reason that quantitative methods for dataset curation represent an improvement over simple intuition. When considering that labeling each image for segmentation took several minutes, and 20% of the training dataset for segmentation comprised 249 image frames, the potential time savings in labeling *even on an already-intuitively-reduced dataset* is significant.

Third, the training/test split for the OCT dataset had to be adjusted in our study because the original test set (500 and 1000 images in the binary and multiclass datasets respectively) was too easy to classify. Experiments resulted in perfect test set separability (AUCROC = 1.0) despite very small training set sizes (<2% total images available). Even with this adjustment, the OCT test dataset was still very separable. In theory, the same methods used in this study to curate training data can be used to curate testing datasets to provide the most efficient and most representative benchmarks for generalizability.

Fourth, ENRICH curated datasets without a requirement for upfront image labeling. This means that one can first curate a small subset of images, and then invest in labeling only those. Furthermore, a study of the class balance curated by ENRICH in medical datasets yielded interesting results in that the most efficient data subsets did not have equal numbers of images among each class. This latter finding suggests that instead of simply balancing classes by raw number of instances per class, optimal class balance may make use of other measures that better account for diversity—this is an exciting avenue for future study.

The implications of our findings for economizing on data in DL tasks are clear. Perhaps future studies using medical imaging datasets might benefit from choosing a small, diverse, ENRICHed subset of images to label and use for model development. ENRICH may provide useful metrics on a dataset's quality and content. While the choice for image-similarity metric in the current implementation of ENRICH aided in demonstrating image redundancy as well as noisy images in the dataset, we anticipate that different choices for similarity metric and curation algorithm will yield additional quality metrics. Finally, while many data reduction methods are model-guided: for example in active learning, the model selects the "best" images for learning,[36–38] in an iterative process, ENRICH is data-guided: the data determine which images are best removed, and it can be used once. ENRICH may therefore be used in conjunction with or instead of active learning.

CONCLUSION

In the future, investigating alternative similarity measures and ranking algorithms offers opportunities to test and potentially optimize the ENRICH framework. For example, other pairwise image-similarity metrics may prove more informative or simpler to compute. 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. In addition, while we ran over two thousand experiments on several datasets, testing ENRICH on still more datasets can only improve its utility. We make our code available so that others can run ENRICH on their datasets, both for their own benefit and for advancement of the field.

Quantitative measures of similarity have been shown to add useful insights in other fields.[35,39] ENRICH is expected to be a useful new avenue for decreasing labeling burden and speeding iterative training and testing of DL models in development.

FUNDING

EC, RoA, RaA, and RiA were supported by the Department of Defense (W81XWH-19-1-0294) and the National Heart, Lung, and Blood Institute (NIH R01HL146398). RoA and RaA were supported by the National Institutes of Allergy and Infectious Diseases (NIH

R01AI148747-01). RaA and RiA were supported by the Gordon and Betty Moore Foundation. EC and RiA were supported by the American Heart Association (17IGMV33870001).

AUTHOR CONTRIBUTIONS

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

CONFLICT OF INTEREST STATEMENT

None declared.

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 cannot be made available to the public.

REFERENCES

1. Madani A, Arnaout R, Mofrad M, Arnaout R. Fast and accurate view classification of echocardiograms using deep learning. *NPJ Digit Med* 2018; 1.
2. Kornblith AE, Addo N, Dong R, *et al*. Development and validation of a deep learning strategy for automated view classification of pediatric focused assessment with sonography for trauma. *J Ultrasound Med* 2022; 41 (8): 1915–24.
3. Arnaout R, Curran L, Zhao Y, *et al*. An ensemble of neural networks provides expert-level prenatal detection of complex congenital heart disease. *Nat Med* 2021; 27 (5): 882–91.
4. Lee G, Fujita H, eds. *Deep Learning in Medical Image Analysis: Challenges and Applications*. Switzerland: Springer International Publishing; 2020. doi:10.1007/978-3-030-33128-3.
5. Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; 542 (7639): 115–8. 10.1038/nature21056.
6. Gulshan V, Peng L, Coram M, *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016; 316 (22): 2402–10.
7. Xu J, Zhang M, Turk EA, *et al*. Fetal pose estimation in volumetric MRI using a 3D convolution neural network. *Med Image Comput Comput Assist Interv* 2019; 11767: 403–10.
8. Rhee DJ, Jhingran A, Rigaud B, *et al*. Automatic contouring system for cervical cancer using convolutional neural networks. *Med Phys* 2020; 47 (11): 5648–58.
9. Gjesteby L, Shan H, Yang Q, *et al*. A dual-stream deep convolutional network for reducing metal streak artifacts in CT images. *Phys Med Biol* 2019; 64 (23): 235003.
10. Li H, He L, Dudley JA, *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* 2021; 51 (3): 392–402.
11. Anderson BM, Lin EY, Cardenas CE, *et al*. Automated contouring of contrast and noncontrast computed tomography liver images with fully convolutional networks. *Adv Radiat Oncol* 2021; 6 (1): 100464.
12. Shen Y, Wu N, Phang J, *et al*. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Med Image Anal* 2021; 68: 101908.
13. Shao M, Han S, Carass A, *et al*. Shortcomings of ventricle segmentation using deep convolutional networks. *Underst Interpret Mach Learn Med Image Comput Appl (2018)* 2018; 11038: 79–86.
14. Kaye EA, Aherne EA, Duzgol C, *et al*. Accelerating prostate diffusion-weighted MRI using a guided denoising convolutional neural network: retrospective feasibility study. *Radiol Artif Intell* 2020; 2 (5): e200007.
15. Vidyaratne L, Alam M, Shboul Z, Iftekharuddin KM. Deep learning and texture-based semantic label fusion for brain tumor segmentation. *Proc SPIE Int Soc Opt Eng* 2018; 2018: 105750D.
16. Zhang J, Gajjala S, Agrawal P, *et al*. Fully automated echocardiogram interpretation in clinical practice. *Circulation* 2018; 138 (16): 1623–35.
17. Fan L, *et al*. Rapid dealiasing of undersampled, non-Cartesian cardiac perfusion images using U-net. *NMR Biomed* 2020; 33: e4239.
18. Rosenkrantz AB, Hughes DR, Duszak R. The U.S. Radiologist Workforce: an analysis of temporal and geographic variation by using large national datasets. *Radiology* 2016; 279 (1): 175–84.
19. WHO. Global maps for diagnostic imaging. https://web.archive.org/web/20200422195643/https://www.who.int/diagnostic_imaging/collaboration/global_collab_maps/en/. Accessed January 24, 2021.
20. WHO. *Global Atlas of Medical Devices*. WHO; 2021. http://www.who.int/medical_devices/publications/global_atlas_meddev2017/en/. Accessed January 24, 2021.
21. *The Complexities of Physician Supply and Demand: Projections from 2019 to 2034*. AAMC. https://www.aamc.org/data-reports/workforce/data/complexities-physician-supply-and-demand-projections-2019-2034. Accessed January 24, 2021.
22. *Data Labeling Pricing—Amazon SageMaker Ground Truth—Amazon Web Services*. Amazon Web Services, Inc. https://aws.amazon.com/sagemaker/data-labeling/pricing/. Accessed January 24, 2021.
23. Culbertson N. Council post: the skyrocketing volume of healthcare data makes privacy imperative. *Forbes*. https://www.forbes.com/sites/forbestechcouncil/2021/08/06/the-skyrocketing-volume-of-healthcare-data-makes-privacy-imperative/.
24. Jercich K. The imaging AI field is exploding, but it carries unique challenges. *Healthcare IT News*. 2021. https://www.healthcareitnews.com/news/imaging-ai-field-exploding-it-carries-unique-challenges. Accessed January 24, 2021.
25. Olvera-López J, Carrasco-Ochoa J, Martínez-Trinidad JF, Kittler J. A review of instance selection methods. *Artif Intell Rev* 2010; 34: 133–43.
26. Joshi A, Porikli F, Papanikolopoulos N. Multi-class active learning for image classification. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009: 2372–9. doi:10.1109/CVPR.2009.5206627.
27. Hoyer L, Dai D, Wang Q, Chen Y, Van Gool L. Improving semi-supervised and domain-adaptive semantic segmentation with self-supervised depth estimation. *arXiv E-prints*, 2021. http://arxiv.org/abs/2108.12545.
28. Mehta R, Shui C, Nichyporuk B, Arbel T. Information gain sampling for active learning in medical image classification. *arXiv E-prints*, 2022. http://arxiv.org/abs/2208.00974.
29. Kermany DS, Goldbaum M, Cai W, *et al*. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* 2018; 172 (5): 1122–31.e9.
30. 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. JMLR workshop and conference proceedings; 2011: 215–223.

31. Burgess CP, *et al.* Understanding disentangling in β-VAE. *arXiv E-prints*, 2018. https://doi.org/10.48550/arXiv.1804.03599.
32. Leinster T. Entropy and diversity: the axiomatic approach. *arXiv E-prints*, 2020. https://doi.org/10.48550/arXiv.2012.02113
33. Jost L. What do we mean by diversity? The path towards quantification. *Mèt Sci Stud J Annu Rev* 2019; 9: 55–61.
34. Kaplinsky J, Arnaout R. Robust estimates of overall immune-repertoire diversity from high-throughput measurements on samples. *Nat Commun* 2016; 7: 11881.
35. Arora R, Arnaout R. Repertoire-scale measures of antigen binding. *Proc Natl Acad Sci U S A* 2022; 119 (34): e2203505119.
36. Cohn D, Atlas L, Ladner R. Improving generalization with active learning. *Mach Learn* 1994; 15: 201–21.
37. Wang K, Zhang D, Li Y, Zhang R, Lin L. Cost-effective active learning for deep image classification. *IEEE Trans Circuits Syst Video Technol* 2017; 27: 2591–600.
38. Fang M, Li Y, Cohn T. Learning how to active learn: a deep reinforcement learning approach. *arXiv E-prints*, 2017. https://doi.org/10.48550/arXiv.1708.02383.
39. Arora R, Arnaout R. Private antibody repertoires are public. bioRxiv 2020.06.18.159699, 2020. doi:10.1101/2020.06.18.159699.