medRxiv preprint doi: https://doi.org/10.1101/2021.05.22.21257645; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

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

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.05.22.21257645; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Abstract

Deep learning (DL) has been applied with success in proofs of concept across biomedical imaging, including across modalities and medical specialties [1–17]. Labeled data is critical to training and testing DL models, and such models traditionally require large amounts of training data, straining the limited (human) resources available for expert labeling/annotation. It would be ideal to prioritize labeling those images that are most likely to improve model performance and skip images that are redundant. However, straightforward, robust, and quantitative metrics for measuring and eliminating redundancy in datasets have not yet been described. Here, we introduce a new method, ENRICH (Eliminate Needless Redundancy for Imaging Challenges), for assessing image dataset redundancy and test it on a well-benchmarked medical imaging dataset[3]. First, we compute pairwise similarity metrics for images in a given dataset, resulting in a matrix of pairwise-similarity values. We then rank images based on this matrix and use these rankings to curate the dataset, to minimize dataset redundancy. Using this method, we achieve similar AUC scores in a binary classification task with just a fraction of our original dataset (AUC of 0.99 ± 1.35e-05 on 44 percent of available images vs. AUC of 0.99 ± 9.32e-06 on all available images, p-value 0.0002) and better scores than same-sized training subsets chosen at random. We also demonstrate similar Jaccard sores in a multi-class segmentation task while eliminating redundant images. (average Jaccard index of 0.58 on 80 percent of available images vs 0.60 on all available images). Thus, algorithms that reduce dataset redundancy based on image similarity can significantly reduce the number of training images required, while preserving performance, in medical imaging datasets.

medRxiv preprint doi: https://doi.org/10.1101/2021.05.22.21257645; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Introduction

Labeling and annotation can be costly bottlenecks for supervised learning tasks in medical imaging. These tasks often require experts with years of training, and very limited bandwidth, to label the images in a dataset[5,6,18–21]. Sometimes these tasks 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, expert labeling is still needed for labeling images in test datasets, in order to benchmark performance on high-stakes medical tasks. This is in contrast to labeling for DL in non-medical fields, which can often rely on a layperson to identify everyday objects or animals and can therefore be done quickly and inexpensively using crowdsourcing techniques[22]. For medical imaging, labeling represents a significant bottleneck for deep learning research in terms of financial, computational, and human resources, as well as in terms of time.

To reduce the labeling burden, it would be ideal then to limit medical imaging datasets to the minimum set of images that still maximizes model performance. How can one find such a set? We propose this can be done by examining image similarity and redundancy within medical imaging datasets. We demonstrate proof of concept on a large, well characterized/well benchmarked dataset of fetal-ultrasound echocardiograms: ECHO-F[3].

Unlike standard images, medical images often have many aspects in common due to standardization of imaging protocols for patient care. Images from a particular medical domain often have similar subject matter (e.g. the heart), 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, images may be captured consecutively, resulting in similarity among consecutive images. In contrast, standard non-medical images are typically captured in a wide range of environments with much greater pairwise differences. For example, the CIFAR10 class of car images consists of images of a wide variety of car makes and models with different backgrounds, photographed from different angles, at different distances, at different times of day[23]. We hypothesize that similarity between common aspects found in medical datasets can create far greater redundancy in medical-image training data and propose that redundant images do not add significant value to the task of training neural networks.

medRxiv preprint doi: https://doi.org/10.1101/2021.05.22.21257645; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

To economize on expert labeling, we require ways to measure and eliminate redundancy in imaging datasets. Here, we present a method, Eliminating Needless Redundancy for Imaging Challenges (ENRICH) with two main steps. First, we develop a similarity metric for pairs of images and compute pairwise similarity for all images in a given dataset, resulting in a matrix of similarity values. Second, we develop an algorithm that operates on the matrix to identify for removal those images with the highest redundancy, thereby minimizing the redundancy of the resulting smaller dataset.

We demonstrate proof of concept using a similarity metric calculated from image embeddings from variational autoencoders. Autoencoders have been widely used for image dimensionality reduction[24]. In particular, disentangled variational autoencoders (d-VAE) have been shown to create embeddings which encode distinct image characteristics [25,26,27]. While other standard whole-image similarity metrics (SSIM) may be used for this task as well, our demonstration here focuses on measuring and comparing redundancy using embeddings.

Methods

Datasets and benchmarks. Datasets consisted of labeled fetal echo ultrasound images, ECHO-F, as previously described[3]. These data were used to explore dataset efficiency-boosting methods on both a classification task and a segmentation task. The (binary) classification task was to classify the fetal axial 4-chamber (A4C) view vs. the non-target (NT) view. The (multi-class) segmentation task was performed on the A4C view, with 5 class labels for image pixels (left ventricle, right ventricle, left atrium, right atrium, background). Training and test datasets are described in Table 1. In ultrasound, one or more video clips are acquired per patient; each video clip consists of as few as one, or as many as hundreds, of consecutive image frames. Training and test sets were divided by patient identifier (ID) and were disjoint from each other (mutually exclusive). Notably, the segmentation dataset had already been curated informally, in that only certain frames from each video clip were labeled.

Image processing. All images were originally 300x400 and converted to grayscale using Python3's OpenCV package. For similarity calculations and creating embeddings, images were cropped and resized to 64x64. For classification model training and testing, the original images were cropped and resized to 80x80. All resizing was done using Python3's Scikit-Image

medRxiv preprint doi: https://doi.org/10.1101/2021.05.22.21257645; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

package. For segmentation model-training and testing, original images were cropped to 272x272, and no resizing was performed.

Embeddings from autoencoders. The bottleneck layer of a dissentagled variational autoencoder (d-VAE) was used to compress each image into a 128-element vector embedding. The d-VAE used was based on the architecture as described previously with the exception of having a 128-element embedding[27]. The d-VAE was trained on a subset of 5,000 images from the entire training dataset, as previously described[3], using combined loss, reconstruction loss and KL divergence, and standard stopping conditions.

Pairwise image similarity metrics. For both classification and segmentation datasets (Table 1), a matrix of pairwise image similarity measures was calculated. The similarity measure between two image embeddings was defined as the complement of the cosine distance between each embedding (resulting in pairwise similarity measures 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 from the training dataset (Table 1) was chosen uniformly at random: 1,000 starting images for the classification task (438 A4C, 562 NT; 2 percent of full training set) and 200 starting images for the segmentation task. Additional images from the remaining dataset determined by the ranking algorithm to have the lowest similarity to the initial subset were iteratively added to grow the dataset (by approximately 5,000 images at a time for classification and 200 images for segmentation). For classification, this resulted in training subsets of 3,000, 5,000, 10,000, 15,000, 20,000, 25,000, 30,000, 35,000, and 40,000 images (corresponding to 7, 11, 22, 33, 44, 55, 66, 77 and 88 percent of the full training set, respectively). For segmentation, this resulted in training subsets of 200, 300, 400, 600, 800, and 1,000 images (corresponding to 16, 24, 32, 48, 64, and 80 percent of the full training set, respectively). This process was repeated, 10 times for the classification task and 3 times for the segmentation task, to create replicates at each subset size.

For classification, the ranking algorithm was blind to class label during this iterative selection; the label was only revealed/assigned after an image was chosen. We then trained a new classification model with the resulting training set, at specific subset sizes, and predicted on the test set.

medRxiv preprint doi: https://doi.org/10.1101/2021.05.22.21257645; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

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, but not for the classification task. For each task the same model parameters were used throughout the experiments.

Labeling time estimates. Labeling time was estimated at 1 image per second for classification tasks and 5 minutes per image for segmentation tasks.

Evaluation metrics. For both deep-learning tasks, we calculated overall redundancy at a threshold *k* as the percentage of the entire training set that had pairwise similarity values above certain values of *k*: 0.5, 0.7, 0.9. This excluded images compared with themselves (which would have had similarity=1.0).

Model performance was compared using AUC for the classification task 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 between AUC performance on test data for different model trainings were performed using the Mann-Whitney U test.

Results

ENRICH involves two steps:(1) computing a matrix of pairwise similarity measures among images in a given dataset, and (2) ranking images by similarity measure for inclusion in the curated dataset. Here, we used ENRICH with a pairwise similarity measure based on the distance between d-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 a medical image dataset (ECHO-F), to our knowledge the first quantitative demonstration of this property. We also demonstrated that using ENRICH to curate the training dataset, the same test performance on well benchmarked binary classification (AUC) and multi-class segmentation tasks was achieved using only 44 percent of available training 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

medRxiv preprint doi: https://doi.org/10.1101/2021.05.22.21257645; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

distributed across the dataset. Using a pairwise image-similarity measure (Methods), we measured the proportion of pairwise image comparisons for ECHO-F classification and segmentation datasets that were highly similar, moderately similar, and largely dissimilar (Fig. 1). The majority of pairwise image comparisons (67% classification and 63% segmentation) are greater than 0.9, where a value of 0 indicates a highly dissimilar image pair and 1 indicates an identical image pair.

Using ENRICH to find smaller training datasets that can achieve benchmark performance on medical-imaging tasks. We trained a binary classification task with different subsets of the ECHO-F training dataset, all tested on the same test images (Table 1). As in the Methods, training image subsets of increasing size were curated using ENRICH vs random selection (Fig. 2a). Area under the curve (AUC) on classifying test data served as the primary performance metric. When trained on the full training dataset, model test performance achieved an AUC of 0.99 ($\pm$ 9×10^{-6}). We then compared model performance using training data subsets when images in those subsets were chosen via ENRICH vs. when images were chosen randomly (Figure 2b). Even with the smallest image subsets, almost all of the patients and video clips were represented in the training set (Table 2).

With as little as 22 percent of images in the full training dataset, ENRICH outperformed random selection of training images (mean AUC 0.99 vs 0.97, MWU p-value 1×10^{-4}; 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 44 percent of the training dataset was needed to achieve benchmark performance (AUC 0.99 $\pm$ 1×10^{-5}, MWU p-value 2×10^{-4}). When images were chosen at random, 66 percent of the training data were needed (Figure 2b)—50 percent more images than ENRICH required to achieve the same benchmark.

Similar to the above, we next 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 0.60. With 64 percent of the training data, ENRICH achieved an average Jaccard of 0.56. In contrast, 80 percent of training data was needed to reach the same performance when image subsets were chosen randomly (Table 3, Fig. 2c).

Potential time savings in labeling. As an example, we estimated the time required to label all the images in ECHO-F for both classification and segmentation tasks, and the time that would have

medRxiv preprint doi: https://doi.org/10.1101/2021.05.22.21257645; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

been required for the smallest ENRICHed subsets that achieved desired performance (44 percent for classification and 64 percent for segmentation), suggesting a savings of a week of full-time work for an expert labeler, on even this relatively small dataset (Fig. 3).

Discussion

Deep learning on medical-imaging datasets requires labeling that can burden clinical experts. To address this bottleneck, we developed ENRICH, a method for curating medical image datasets based on quantitative measures of image similarity. Our results show that quantitative methods can be used to identify redundancy in image-training datasets, showing that medical datasets such as ECHO-F contain significant redundancy. Quantitative methods such as ENRICH can therefore be used to curate training images in a label-free approach. For both classification and segmentation tasks, using ENRICH demonstrated that *(i)* redundant images do not aid significantly in DL model training, *(ii)* image labels are not needed for curating image datasets for redundancy, and *(iii)* for some medical datasets, state-of-the-art performance can be achieved using only a fraction of the full training dataset.

ENRICH eliminates the need to label large portions of available image data—over half, in the case of our binary classification task (Fig. 2b)—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 of only 20 percent of the training dataset still had an AUC of 0.98 (Fig. 2b). The implications of these findings for economizing on expert clinical image labeling are clear. 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.

For the classification task presented here, model trainings did not include standard data augmentations (such as rotating or flipping images; see Methods) in order to *(i)* remove data augmentation as a potential confounder 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.

medRxiv preprint doi: https://doi.org/10.1101/2021.05.22.21257645; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

For the segmentation task presented here, a larger subset of training data was needed to approximate full training set performance, suggesting that ENRICH was less helpful for segmentation. However, it is important to note that for the segmentation task, labeling was so time-consuming that we had *already* chosen not to label every frame in each video clip (Methods; Table 1). Therefore, the finding that an additional 36 percent of the already-intuitively-reduced dataset was not needed to reach full dataset performance is 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 36 percent of the training dataset for segmentation comprised 449 image frames, the potential time savings in labeling *even on an already-intuitively-reduced dataset* is significant.

ENRICH can accommodate different choices of similarity measure (step 1) and ranking algorithm (step 2). Here, we used a d-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. Additionally, 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 [28,29]. 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).

medRxiv preprint doi: https://doi.org/10.1101/2021.05.22.21257645; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Figure Legends

Figure 1. ECHO-F dataset has significant redundancy among images. Pairwise similarity measures among images in the ECHO-F classification and segmentation datasets are shown (excluding images compared with themselves). Shading indicates the percent of image similarities with values of 0.7-0.9 (dark gray), 0.5-0.7 (medium gray), and 0-0.5 (light gray), where a similarity measure of 0 indicates a highly dissimilar image pair and a measure of 1 indicates an identical image pair. As shown, over 60 percent of image similarity measures have a score of 0.7-0.9 for both classification and segmentation datasets, indicating a high number of similar images in these datasets.

Figure 2. ENRICH allows for similar test performance using a fraction of available training data, outperforming training on randomly selected image subsets. (a) The overall schematic for iteratively adding images to an initial, small group of images is shown. The algorithm used can be ENRICH or one that adds images at random. **(b)** From a common 1000-image starting dataset (grey), additional images were added to create increasingly larger training subsets using ENRICH (blue) vs random addition (yellow). Dots represent mean AUC on the test set (Table 1) from 10 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 (*ns* = $p>0.05$; * = $p\leq0.05$; ** = $p\leq0.01$; *** = $p\leq0.001$; **** = $p\leq0.0001$). Datapoints circled in red are statistically indistinguishable from model performance using the full training set (100% of training images; black dot). **(c)** For a multi-class segmentation task, average Jaccard index for segmented structures is shown as a function of increasing training set size using ENRICH (blue) vs. randomly added images (yellow).

Figure 3. Labeling time savings using ENRICH. Time estimates for classification (left) and segmentation (right) for ECHO-F.

medRxiv preprint doi: https://doi.org/10.1101/2021.05.22.21257645; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

References

1. Madani, A., Arnaout, R., Mofrad, M. & Arnaout, R. Fast and accurate view classification of echocardiograms using deep learning. *NPJ Digit. Med.* **1**, (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). *medRxiv* 2020.10.14.20206607 (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.* (2020) doi:10.1007/s00247-020-04854-3.

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

medRxiv preprint doi: https://doi.org/10.1101/2021.05.22.21257645; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

100464 (2021).

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. First Int. Workshop MLCN 2018 DLF 2018 IMIMIC 2018 Held Conjunction MICCAI 2018 Granada Spain Sept. 16-20 2018* **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**, (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.

medRxiv preprint doi: https://doi.org/10.1101/2021.05.22.21257645; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

https://aws.amazon.com/sagemaker/groundtruth/pricing/.

23. CIFAR-10 and CIFAR-100 datasets. https://www.cs.toronto.edu/~kriz/cifar.html.

24. Petscharnig, S., Lux, M. & Chatzichristofis, S. Dimensionality Reduction for Image Features using Deep Learning and Autoencoders. in *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing* 1–6 (ACM, 2017). doi:10.1145/3095713.3095737.

25. Li, Y. *et al.* Disentangled Variational Auto-Encoder for semi-supervised learning. *Inf. Sci.* **482**, 73–85 (2019).

26. Bepler, T., Zhong, E. D., Kelley, K., Brignole, E. & Berger, B. Explicitly disentangling image content from translation and rotation with spatial-VAE. *ArXiv190911663 Cs Q-Bio* (2019).

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

28. Arora, R. & Arnaout, R. Private Antibody Repertoires Are Public. *bioRxiv* 2020.06.18.159699 (2020) doi:10.1101/2020.06.18.159699.

29. Arora, R., Burke, H. M. & Arnaout, R. Immunological Diversity with Similarity. *bioRxiv* 483131 (2018) doi:10.1101/483131.

medRxiv preprint doi: https://doi.org/10.1101/2021.05.22.21257645; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Tables and Figures

Table 1. Overall training and test datasets

	Classification: A4C		**Classification: NT**		**Segmentation: A4C**	
	Training	Test	Training	Test	Training	Test
Number of images	20,378	4,365	25,082	3,618	1,248	173
Number of patients	652	80	281	51	186	20
Number of video clips	1,495	198	2849	764	299	48

medRxiv preprint doi: https://doi.org/10.1101/2021.05.22.21257645; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Table 2. Training data subsets for classification tasks

Dataset	**NT**										
% of full training set	2	7	11	22	33	44	55	66	77	88	100
Number of images	562	1,979	3,470	7,421	10,671	13,476	15,955	18,188	19,731	19,731	25,082
Number of patients	105	225	270	281	281	281	281	281	281	281	281
Number of video clips	412	1,341	2,131	2,740	2,792	2,810	2,818	2,827	2,834	2,834	2,849
Dataset	**A4C**										
% of full training set	2	7	11	22	33	44	55	66	77	88	100
Number of images	438	1,021	1,530	2,579	4,329	6,524	9,045	11,812	15,269	20,269	20,378
Number of patients	268	505	630	652	652	652	652	652	652	652	652
Number of video clips	355	898	1,320	1,479	1,493	1,495	1,495	1,495	1,495	1,495	1,495

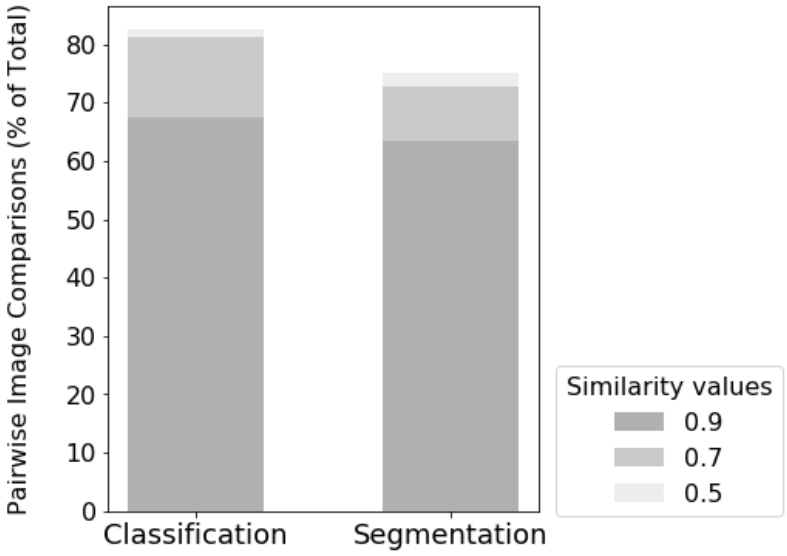
medRxiv preprint doi: https://doi.org/10.1101/2021.05.22.21257645; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Table 3. Training-data subsets for segmentation task

	Segmentation: A4C						
% of full training dataset	16	24	32	48	64	80	100
Number of images	200	300	400	600	800	1,000	1,248
Number of patients	99	119	130	152	168	179	186
Number of video clips	121	154	182	231	264	283	299

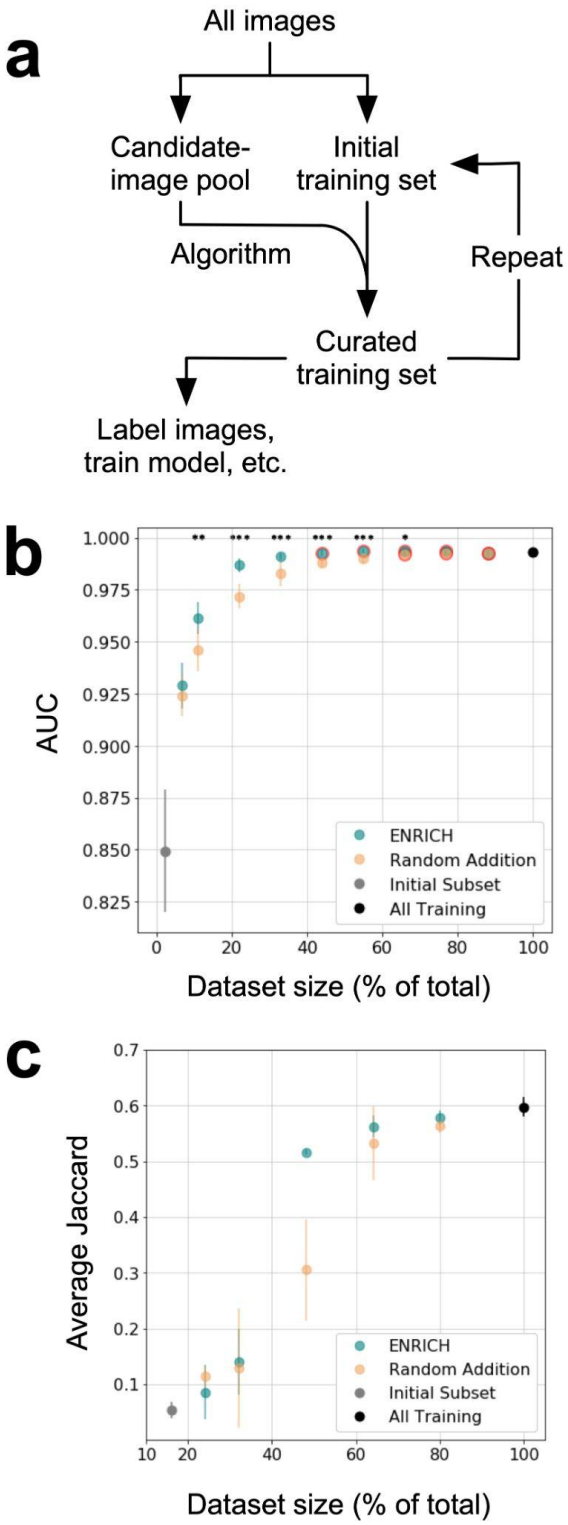
medRxiv preprint doi: https://doi.org/10.1101/2021.05.22.21257645; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Figure 1. ECHO-F dataset has significant redundancy among images



medRxiv preprint doi: https://doi.org/10.1101/2021.05.22.21257645; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Figure 2. ENRICH allows for similar test performance using a fraction of available training data, outperforming training on randomly selected image subsets



medRxiv preprint doi: https://doi.org/10.1101/2021.05.22.21257645; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

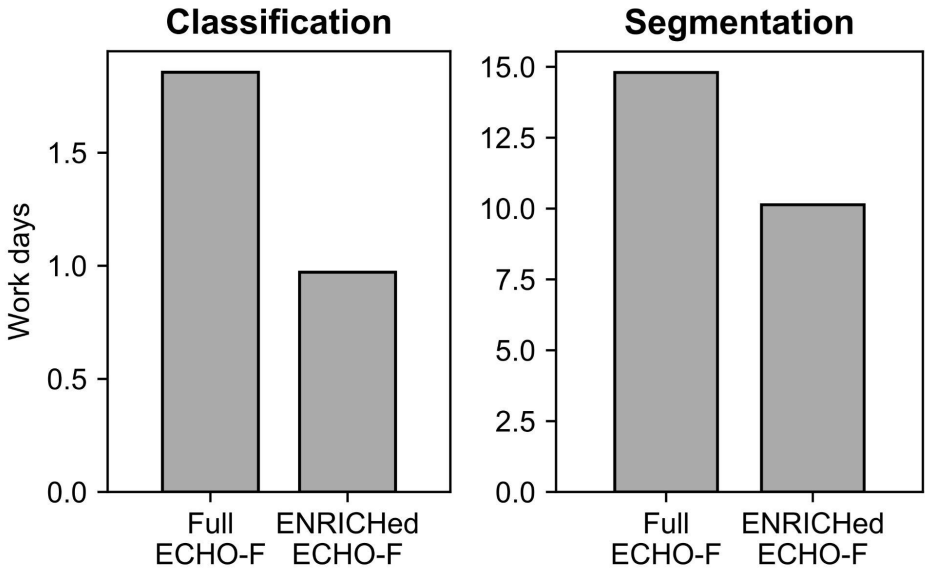



Figure 3. Labeling time savings using ENRICH