TMED 2: A Dataset for Semi-Supervised Classification of Echocardiograms

Zhe Huang [1] **Gary Long** [2 3] **Benjamin S. Wessler** [2] **Michael C. Hughes** [1]

Abstract

Deep learning can automate the interpretation of medical images and potentially improve the reliability of current diagnostic practice. However, a common roadblock is a lack of labeled data. Recent developments in semi-supervised learning (SSL) promise high accuracy from the combination of a small labeled set and a large unlabeled set. But validation of SSL on real medical data is sorely needed, as common benchmarks represent *artificial* "best-case" scenarios for SSL (unlabeled sets are obtained by dropping labels) and are too *curated* (unlabeled images all show task-relevant classes). To address this need, we release an upgraded open-access dataset – The Tufts Medical Echocardiogram Dataset (TMED 2) – with higher resolution, more labels, and an *authentic, uncurated* unlabeled set double the size of our original release. In a view classification task, state-of-the-art SSL training via FixMatch improves accuracy over a labeled-set-only baseline. In a diagnosis task using many images from a single scan, the benefits of SSL are less clear; we find multi-task methods that do not use unlabeled data work better. We hope this dataset catalyzes a new wave of methods development that might improve patient care despite limited labeled data.

1. Introduction

When developing medical image classifiers, a primary barrier to success is the need to assemble a large-enough labeled dataset for the intended task. Labeling often requires expensive, time-consuming work from human experts. If only a tiny labeled set is available but access to a larger *unlabeled* set of images is possible, recent advances in *semi-supervised learning* (SSL) are promising (Miyato et al., 2019; Berthelot et al., 2019). Table 1 shows SSL's progress on the SVHN benchmark, recognizing digits in photographs of address numbers on houses. With only 100 labeled examples per class, supervised neural nets have an error rate over 12%.

Num. Labeled	Num. Unlabeled	SSL Method	Error rate
100 per class	0	labeled-set-only	12.8% (†)
100 per class	64932	VAT	5.6% (†)
100 per class	64932	FixMatch	2.4% (‡)

Table 1: SSL can improve image classifiers given limited labeled data but a huge unlabeled set, using a standard architecture (WideResNet-28-2) on the non-medical SVHN benchmark. We ask, *Are similar gains possible for ultrasound images of the heart?*. †: Tab. 5 of Oliver et al. (2018), ‡: Tab. 2 of Sohn et al. (2020)

Using SSL and a large unlabeled set, a recent method called FixMatch drops error below 2.5% (Sohn et al., 2020).

Applications of SSL to medical imaging (Madani et al., 2018; Calderon-Ramirez et al., 2021) are exciting but relatively rare. Most SSL methods development continues to focus on benchmark datasets originally intended for supervised classification, such as SVHN, CIFAR-10, or ImageNet. This is a problem for two reasons. First, these data represent *artificial* "best-case" scenarios for SSL, because the unlabeled set is created by dropping known labels. Second, these data are *too curated*, because the unlabeled set contains only "known" classes with balanced distribution. In a real medical application, the unlabeled set will have truly unknown labels and may contain examples not belonging to any category of task-specific interest. Even the examples of known classes may have frequencies that differ from the labeled train set. Several recent efforts have pointed out potential limitations of SSL when unlabeled sets differ from the labeled set (Oliver et al., 2018; Ganev & Aitchison, 2021), but most methods-focused papers continue to evaluate SSL on artificial, curated unlabeled sets.

Recently, we introduced a dataset, the Tufts Echocardiogram Dataset (TMED), to stress-test modern SSL methods on real medical tasks (Huang et al., 2021). TMED's focus is on a particular medical imaging problem: developing classifiers to diagnose *aortic stenosis* (AS) from echocardiograms (ultrasound images of the heart). Using a large unlabeled set and a small labeled set, we found modern SSL methods like MixMatch could improve two classification tasks relevant to AS. Our original release of the TMED dataset now has more than 70 approved users in 22 countries across 5 continents.

In this present work, we release a significantly upgraded dataset, which we call TMED 2, which will be available for non-commercial research use at TMED.cs.tufts.edu starting in July 2022. This new release *doubles* the size of both

[1] Dept. of Computer Science, Tufts University [2]Division of Cardiology, Tufts Medical Center [3]CVAI Solutions.

Accepted paper at the DataPerf workshop at ICML 2022.

the labeled set and the unlabeled set and expands the set of possible labels. These changes allow external validation on other open data. In the rest of this paper, we will make the case that this new dataset highlights the promise of SSL for medicine (view task in Fig. 1), while also showing where current SSL methods fall short (diagnosis task in Fig. 2). A forthcoming journal article describes our data and investigations for a medical research audience.

2. Clinically-relevant Classification Tasks

We consider two classification tasks: image-level view classification and study-level aortic stenosis (AS) severity classification. Both advance our goal of improving early detection of AS and potentially reducing its high mortality rate.

2.1. Per-image View Classification Task

In a typical trans-thoracic echocardiography (TTE) scan or *study*, a human sonographer holds a handheld transducer over the patient's chest, manually choosing acquisition angles in order to capture the heart's complex anatomy. We focus on *2-dimensional* (2D) view types, leaving other modalities such as Doppler profiles or m-mode imaging for future work. 2D imaging results in *multiple* short videos of the heart, each depicting a potentially different anatomical view throughout the cardiac cycle. From each short video, we can extract a representative image. A typical scan thus consists of 68 images (median=68, 10-90th percentile range=27-97).

Among 2D TTE views, at least 9 canonical view types are possible (Mitchell et al., 2019). Each one displays distinct aspects of the heart's anatomy. For our goal of supporting diagnosis of AS, two specific views are *relevant*: parasternal long axis (PLAX) and parasternal short axis (PSAX), because the aortic valve's structure and function is visible.

While the sonographer intentionally captures multiple views, when data is stored to the electronic medical record a view type annotation is not usually applied. Given many raw images alone it is difficult to automatically find a specific view. This practical difficulty motivates the need for a *view classifier*. Can we reliably find the views relevant to AS from the dozens of images in a typical study?

For the concrete view-classification task supported by TMED 2 data, the problem is to take one input image (112x112 pixels) and determine its view type. There are 5 possible types: PLAX, PSAX, apical 2-chamber (A2C), apical 4-chamber (A4C), or Other. A4C and A2C views are less relevant to AS diagnosis but allow connections to other datasets, plus unlock tasks we may study in future. The "Other" category is a super-category that contains other possible types distinct from PLAX, PSAX, A4C, and A2C.

2.2. Per-study AS Severity Classification Task

Toward our ultimate goal of automated early screening for AS, we formulate a *diagnosis* task. As input, we are given the dozens of images from one echocardiogram study (the exact number may vary). The predicted output for the whole study is one of 3 severity levels: no AS, early AS, or significant AS. This task mimics how real AS diagnoses work in practice: Cardiologists have access to many images captured by the sonographer, varying in view type and other details. They need to identify which images show relevant anatomical structure and then look for signs of disease in these images to determine a severity level.

3. TMED 2 Dataset

The TMED 2 dataset represents a significant upgrade from the original TMED 1 release in size, resolution, and available labels. TMED 2 contains 3 kinds of studies: a fully-labeled set (with labels for both tasks: view and AS severity), a view-only-labeled set, and a huge, uncurated unlabeled set. These are summarized in Table 3. Example images can be found in App. D.

Image Acquisition All images originate from trans-thoracic echocardiograms performed during routine clinical care between 2011-2020 at Tufts Medical Center, a high-volume tertiary care center in Boston, MA. The source devices span several major vendors (Philips®, Toshiba®, Siemens®), so that derived classifiers are not specific to one vendor.

Access and Ethical Oversight. We post-processed all images to remove any protected health information. The TMED 2 dataset contains only fully de-identified images and labels. Non-commercial research use of these de-identified images was approved by the Tufts Medical Center IRB, enabling sharing of our dataset under a apply-for-access model whose license [1] balances the benefits of sharing with the best interests of the patients this data comes from.

Diagnostic Label Acquisition. Diagnostic labels were assigned by a cardiologist with specialty training in echocardiography during a routine clinical interpretation of the entire study, following current guidelines for AS severity (Baumgartner et al., 2017). We simplified fine-grained severity levels into 3 classes: "no AS", "early AS" (including mild and mild-to-moderate), and "significant AS" (including moderate and severe). Although these labels are technically recorded for most studies where an expert prepares a summary report, under our current records system extracting this label from the report into a form amenable to machine learning requires substantial manual effort.

View Label Acquisition. Board certified echocardiographers and American Registry for Diagnostic Medical Sonog-

[1] TMED.cs.tufts.edu/data_license.html

Dataset	Clinical Goal	Unlabeled Set	Labeled Set			
			Num. Patients	Num. Images	Label Type	View Types Included
TMED 2[a]	detect aortic valve disease	353500 images	1284	24964 images	image-level view study-level severity	all available (complete studies)
Stanford EchoNet Dynamic[b]	ventricle measurements	none	10030	10030 videos	pixel-level	A4C only
Stanford EchoNet LVH[c]	ventricle measurements	none	12000	12000 videos	pixel-level	PLAX only
Unity Imaging Collaborative[d]	ventricle measurements	none	2056	7523 images	pixel-level	PLAX only
CAMUS[e] (Univ. Lyon)	structure segmentation	none	500	2000 images	pixel-level	A2C and A4C

a: TMED.cs.tufts.edu
b: Ouyang et al. (2020) https://echonet.github.io/dynamic/
c: Duffy et al. (2022) https://echonet.github.io/lvh/
d: Howard et al. (2021) https://data.unityimaging.net/
e: Leclerc et al. (2019) https://www.creatis.insa-lyon.fr/Challenge/camus/

Table 2: Comparison of ML-ready open-access datasets of echocardiograms. TMED 2 differs in its focus on aortic valve disease, its large unlabeled set for SSL, and the release of complete studies containing *many diverse images* (not just manually-selected specific views). Previous datasets do provide useful *pixel-level* labels, meaning that specific points, regions, or contours are annotated within the image.

Set	Num. Studies	Num. Images	
		Labeled	Unlabeled
labeled train	360	10066	16468
labeled valid	119	3602	5046
labeled test	120	3602	5082
view-only train	722	7694	37576
unlabeled train	5486	0	353500

Table 3: TMED 2 dataset contents. Showing the image count for split 0 in our experiment. Image count for different splits differ by less than 10 % for different parts of the dataset. We further ensure the ratio of studies with different diagnosis class are roughly the same for labeled train, labeled valid and labeled test ($\sim$ 21% no AS, $\sim$ 29% early AS, and $\sim$ 50% significant AS).

raphy certified sonographers used a custom annotation tool to provide view type labels to specific images.

Our dataset represents the union of two rounds of labeling. In the first round (181 studies), 3 possible labels were assigned to the majority of images in a study: PLAX, PSAX, or a label representing the union of A2C, A4C, and Other. Round 1 labeled a median of 60 images per study (10-90th percentile range=47-78). In the second round (1140 studies), to expedite diverse labels from more studies, labelers were asked to label a few examples from each study for each of 4 types (PLAX, PSAX, A2C, or A4C). Round 2 labeled a median of 10 images per study (10-90th percentile range=2-24). Images without labels may belong to any category.

Image preprocessing. To prepare raw echocardiogram DICOM files for classification, we keep only 2D images. From each cineloop, we extract the first frame as a representative image (clinicians verified they could perform all tasks from such frame). We convert to gray-scale, pad the shorter axis to a square aspect ratio and resize to 112x112 pixels (increasing resolution from TMED 1's 64x64).

4. Related Work

In the last few years, several laudable efforts around the world (Leclerc et al., 2019; Ouyang et al., 2020; Howard et al., 2021) have released echocardiogram images as well as annotations for training predictive models. These are summarized in Table 2. We emphasize that ours is the only dataset that represents all available views in a study instead of a prefiltered subset. Similarly, ours is the only one providing severity level labels for AS.

5. Classification Methods

Overall, our methods are drawn from our previous work on TMED 1 (Huang et al., 2021). For all experiments, we use the same "Wide ResNet-28" architecture (Zagoruyko & Komodakis, 2017). One such network f with weights θ_V produces *view type* probabilities (a vector of size 5). Another network g with weights θ_D produces probabilities for AS diagnosis (a vector of size 3).

Training CNNs to recognise views. To train weights θ_V, our baseline labeled-set-only approach minimizes cross entropy on all view-labeled training data (rows 1 and 4 of Tab. 3). As an exemplar of state-of-the-art SSL, we use FixMatch (Sohn et al., 2020), which we train using the same labeled set, as well as all images in the unlabeled set.

Our view task is unusual in our wish to predict an "Other" class despite only having super-class labels for "A4C-or-A2C-or-Other". To handle this super-class, we simply add together the 3 probabilities predicted by network f_{θ_V} to form the super-class probability, and compute the corresponding cross-entropy loss for the super-class.

Training CNNs to diagnose individual images. To train weights θ_D, we tried FixMatch SSL (using data from row 1 and 5), but it generally did not outperform the labeled-set-only *multi-task learning* method. Let α be *shared* weights for all but the output layer of the networks, and let ω_V, ω_D be the output layers of each network, we train these weights to minimize the sum of a diagnostic loss and a view loss:

$$\min_{\alpha,\omega_D,\omega_V} \sum_{x,y,v} \ell(y, g_{\alpha,\omega_D}(x)) + \gamma \ell(v, f_{\alpha,\omega_V}(x)), \quad (1)$$

where x is an image, y is its one-hot diagnosis label, v is its one-hot view label, and ℓ is a weighted cross-entropy loss. Hyperparameter γ controls the strength of the view loss. We select γ to optimize diagnosis performance on validation.

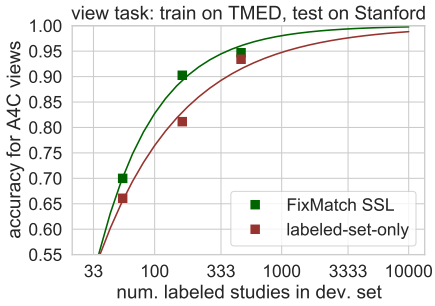

Figure 1: The *promise* of SSL for medical imaging using our TMED 2 data. View classifier accuracy vs. size of labeled data for models trained on TMED 2 data then tested on images from Stanford Echonet. SSL methods use all 5486 studies in the TMED 2 unlabeled set. Lines show least-squared-error fit of the power-law projection $\text{acc}(n) = 1 - \alpha n^{-\beta}$ to aid extrapolation.

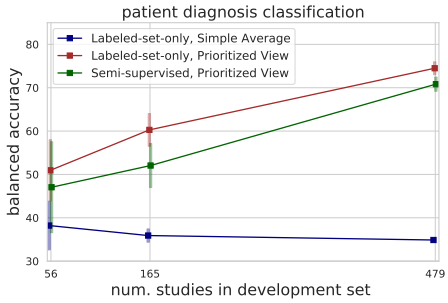

Figure 2: The *perils* of current SSL for medical imaging using our TMED 2 data. Balanced accuracy on the diagnosis task as size of labeled data increases. Square marker gives mean accuracy over 3 splits; bars show standard deviation. Prioritized view averaging is described around Eq. (2). The SSL method is FixMatch.

Aggregating across images for study-level diagnosis. Given image-level networks with weights θ_V, θ_D, we turn image-level diagnoses into study-level ones via the weighted averaging procedure in Huang et al. (2021). Let i index the I_n total images of study n. Let weight $w(x_{ni})$ be the view classifier's probabilistic confidence that image i shows a *clinically-relevant* view for our task: $w(x_{ni}) = p(v_{ni} \in \{\text{PLAX,PSAX}\}|x_{ni}, \theta_V)$. We predict the probability of diagnostic label c for study n via a weighted average that prioritizes relevant views:

$$p(y_n = c|x_{n,1:I_n}) \propto \sum_{i=1}^{I_n} w(x_{ni}) g_{\theta_D}(x_{ni})[c], \quad (2)$$

Here, $g_{\theta_D}(x_{ni})[c]$ denotes the c-th class probability produced by the image-level diagnosis network on image x_{ni}. For TMED 2, we found a further *thresholding* operation that forces $w(x_{ni})$ to zero if it falls below a chosen value helped improve performance further, with threshold value selected on the validation set. We see significant gains from this choice in practice (see App. C).

6. Experimental Results

Our TMED 2 release defines 3 separate, independently sampled splits into train, validation, and test sets, all obeying the sizes in Tab. 3. Within each split, each patient's data belongs to only train, only valid, or only test. For robustness, we recommend reporting the mean test score across splits.

Per-Image View Classification Results. We assess generalization by training models on TMED2, then evaluating on 10,030 external images from Stanford's Echonet Dynamic dataset. All such images are A4C views, so this can only test generalization to new A4C images, not other view types. We report accuracy (fraction of all A4Cs classified correctly, higher is better). App. B provides further internal validation.

Fig. 1 shows that for all methods, external A4C accuracy improves as the size of the development set (training plus validation) increases. At each dataset size, the semi-supervised approach consistently outperforms the labeled-set-only baseline. This suggests the value of unlabeled data and modern SSL methods on real medical data.

We visualize saliency maps (Selvaraju et al., 2020) for the view classifier in App. E. Our clinical co-author confirmed the highlighted regions are medically relevant.

Per-study AS Diagnosis Results. In Fig. 2, we evaluate our severity-level classifiers on the internal TMED test set. First, we see that our prioritized view weighted averaging in Eq. 2 is far more effective than a simpler averaging that treats all images equally regardless of view. Second, we see that at each development set size, using only the labeled set outperforms the corresponding SSL method that learns from the additional large unlabeled set. We suspect the semi-supervised diagnosis model tends to predict overly confident relevance weights $w(x_{ni})$ in Eq. 2 for unlabeled images that should be irrelevant. We plan to investigate this issue further in the future.

7. Conclusions

Labeled medical data are notoriously difficult and expensive to collect. Some very recent methods have tried to ensure SSL delivers real-world value by effectively making use of *uncurated* unlabeled sets that may look different from the labeled set (Chen et al., 2020; Guo et al., 2020; Saito et al., 2021). However, this line of work paradoxically makes this point by repurposing too-curated datasets like CIFAR-10. Using TMED 2 data, we can demonstrate both the promise (Fig. 1) and the challenges (Fig. 2) that current SSL methods face given a large, uncurated unlabeled set for a real medical problem. We hope our open-access release encourages new methods to address the challenge of medical image classification with limited available labels. Performance on our diagnosis task in particular is far from saturated. We hope improvements could help patients with AS.

References

Baumgartner, H., Hung, J., Bermejo, J., Chambers, J. B., Edvardsen, T., Goldstein, S., Lancellotti, P., LeFevre, M., Miller, F., et al. Recommendations on the echocardiographic assessment of aortic valve stenosis: A focused update from the European Association of Cardiovascular Imaging and the American Society of Echocardiography. *European Heart Journal Cardiovascular Imaging*, 18(3): 254–275, 2017.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. MixMatch: A Holistic Approach to Semi-Supervised Learning. In *Advances in Neural Information Processing Systems*, 2019. URL http://arxiv.org/abs/1905.02249.

Calderon-Ramirez, S., Giri, R., Yang, S., Moemeni, A., Umana, M., Elizondo, D., Torrents-Barrena, J., and Molina-Cabello, M. A. Dealing with scarce labelled data: Semi-supervised deep learning with mix match for covid-19 detection using chest x-ray images. In *International Conference on Pattern Recognition (ICPR)*, 2021.

Chen, Y., Zhu, X., Li, W., and Gong, S. Semi-Supervised Learning under Class Distribution Mismatch. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34 (04):3569–3576, 2020.

Duffy, G., Cheng, P. P., Yuan, N., He, B., Kwan, A. C., Shun-Shin, M. J., Alexander, K. M., Ebinger, J., Lungren, M. P., Rader, F., Liang, D. H., Schnittger, I., Ashley, E. A., Zou, J. Y., Patel, J., Witteles, R., Cheng, S., and Ouyang, D. High-Throughput Precision Phenotyping of Left Ventricular Hypertrophy With Cardiovascular Deep Learning. *JAMA Cardiology*, 7(4):386–395, 2022. ISSN 2380-6583. doi: 10.1001/jamacardio.2021.6059. URL https://doi.org/10.1001/jamacardio.2021.6059.

Ganev, S. and Aitchison, L. Semi-supervised learning objectives as log-likelihoods in a generative model of data curation. *arXiv:2008.05913 [cs, stat]*, 2021. URL http://arxiv.org/abs/2008.05913.

Guo, L.-Z., Zhang, Z.-Y., Jiang, Y., Li, Y.-F., and Zhou, Z.-H. Safe Deep Semi-Supervised Learning for Unseen-Class Unlabeled Data. In *International Conference on Machine Learning*, pp. 10, 2020. URL http://proceedings.mlr.press/v119/guo20i/guo20i.pdf.

Howard, J. P., Stowell, C. C., Cole, G. D., Ananthan, K., Demetrescu, C. D., Pearce, K., Rajani, R., Sehmi, J., Vimalesvaran, K., et al. Automated Left Ventricular Dimension Assessment Using Artificial Intelligence Developed and Validated by a UK-Wide Collaborative. *Circulation: Cardiovascular Imaging*, 14(5):e011951, 2021.

Huang, Z., Long, G., Wessler, B., and Hughes, M. C. A New Semi-supervised Learning Benchmark for Classifying View and Diagnosing Aortic Stenosis from Echocardiograms. In *Proceedings of the 6th Machine Learning for Healthcare Conference*. PMLR, 2021. URL https://proceedings.mlr.press/v149/huang21a.html.

Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E. A. R., Jodoin, P.-M., et al. Deep Learning for Segmentation Using an Open Large-Scale Dataset in 2D Echocardiography. *IEEE Transactions on Medical Imaging*, 38(9): 2198–2210, 2019.

Madani, A., Ong, J. R., Tibrewal, A., and Mofrad, M. R. Deep echocardiography: Data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease. *NPJ digital medicine*, 1(1):1–11, 2018.

Mitchell, C., Rahko, P. S., Blauwet, L. A., Canaday, B., Finstuen, J. A., Foster, M. C., Horton, K., Ogunyankin, K. O., Palma, R. A., et al. Guidelines for Performing a Comprehensive Transthoracic Echocardiographic Examination in Adults: Recommendations from the American Society of Echocardiography. *Journal of the American Society of Echocardiography*, 32(1):1–64, 2019.

Miyato, T., Maeda, S.-I., Koyama, M., and Ishii, S. Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41 (8):1979–1993, 2019. URL https://ieeexplore.ieee.org/document/8417973/.

Oliver, A., Odena, A., Raffel, C. A., Cubuk, E. D., and Goodfellow, I. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems*, 2018. URL https://papers.nips.cc/paper/2018/file/c1fea270c48e8079d8ddf7d06d26ab52-Paper.pdf.

Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C. P., Heidenreich, P. A., Harrington, R. A., Liang, D. H., et al. Video-based AI for beat-to-beat assessment of cardiac function. *Nature*, 580(7802):252–256, 2020.

Saito, K., Kim, D., and Saenko, K. OpenMatch: Open-set Consistency Regularization for Semi-supervised Learning with Outliers. In *Advances in Neural Information Processing Systems*, pp. 12, 2021. URL https://proceedings.neurips.cc/paper/2021/file/

da11e8cd1811acb79ccf0fd62cd58f86-Paper.pdf.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, 128(2): 336–359, 2020.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A., Li, C.-L., et al. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper/2020/file/06964dce9addb1c5cb5d6e3d9838f733-Paper.pdf.

Zagoruyko, S. and Komodakis, N. Wide Residual Networks. *arXiv:1605.07146 [cs]*, 2017. URL http://arxiv.org/abs/1605.07146.

A. Data Release

We plan the public release of TMED 2 in July 2022, in time for the DataPerf workshop at ICML.

Meanwhile, interested readers can access and download TMED 1 via our website: TMED.cs.tufts.edu

Deidentified images have been approved for release by our institutional review board (Tufts IRB #MODCR-03-12678)

B. Further Results

Fig. B.1 shows our internal experiments on view, training on TMED2 train set and evaluating on the TMED2 test set.

C. Method Details

C.1. How View and Diagnosis Classifiers are Combined

For image-level view classifiers, we trained two versions of weights: one θ_V via FixMatch SSL (Sohn et al., 2020), and another via minimizing cross entropy on the labeled set.

For image-level diagnosis classifiers, we always train θ_D via the labeled-set-only multi-task objective described in the main text, setting θ_D to the concatenation of learned shared weights α and output layer weights ω_D. We found this was generally superior to any SSL method to learn θ_D.

We then perform patient-level diagnosis with 3 strategies that vary the averaging method (Eq. (2)) used to make one diagnosis from many images (these are the 3 lines in Fig. 2.):

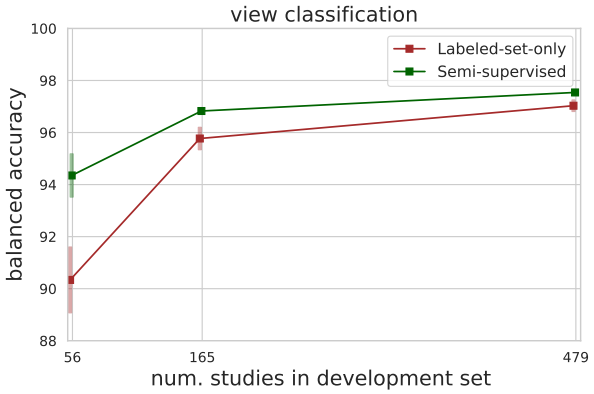

Figure B.1: View classification performance of semi-supervised learning model vs baseline models. This figure describes balanced accuracy (y-axis) as the number of labeled studies increase (x-axis). Square represent the average over 3 splits, and the color bar represent standard deviation. We used FixMatch as the SSL algorithm used for these experiments.

- Multi-task θ_D and simple averaging (each image has uniform weight; no need for any θ_V)
- Multi-task θ_D plus prioritized-view averaging (using the SSL θ_V to compute $w(x_{ni})$)
- Multi-task θ_D plus prioritized-view averaging (using the baseline labeled-set-only θ_V to compute $w(x_{ni})$)

C.2. Thresholding the Weights

In Eq. (2), we described a weighted averaging process that used the view classifier to obtain per-image weights $w(x_{ni})$ that determined the relevancy of each image to the diagnostic task. Given weights $w(x_{ni})$ for each image i in study n, we found an additional thresholding post-processing step was useful:

$$w'(x_{ni}) = \begin{cases} 0 & \text{if } w(x_{ni}) < \tau_1 \\ 0 & \text{if } H[f(x_{ni})] > \tau_2 \\ w(x_{ni}) & \text{otherwise} \end{cases} \quad (3)$$

Here, each image's final weight in Eq. (2) is set to zero if either of two conditions suggesting that the view classifier is not confident occur. The first condition is that the original relevant-view probability $w(x_{ni})$ is too low. The second condition is that the entropy of the predicted probability vector $f(x_{ni})$ is too high (too close to uniform over all 5 possible view types).

On the TMED 2 test set for split 0, we find this thresholding improves the balanced accuracy of all methods:

- 72.5 (without) to 74.6 (with thresholding) for labeled-set-only and prioritized view averaging
- 52.3 (without) to 73.0 (with thresholding) for SSL-enabled prioritized view averaging

These numbers are for one split; others show similar gains.

D. Example Echocardiogram Images of Different View Types

Figure D.1: **Example echocardiograms of each view type.** Each row shows 3 randomly-chosen examples from our TMED2 dataset for each type of view: PLAX (1st row), PSAX (2nd row), A4C (3rd row) and A2C (4th row) view and the super category A4CorA2CorOther (5th row). *Note the examples here are in their original resolution. Our public release contains images resized to a standard resolution of 112x112.*

E. Sailency Visualizations

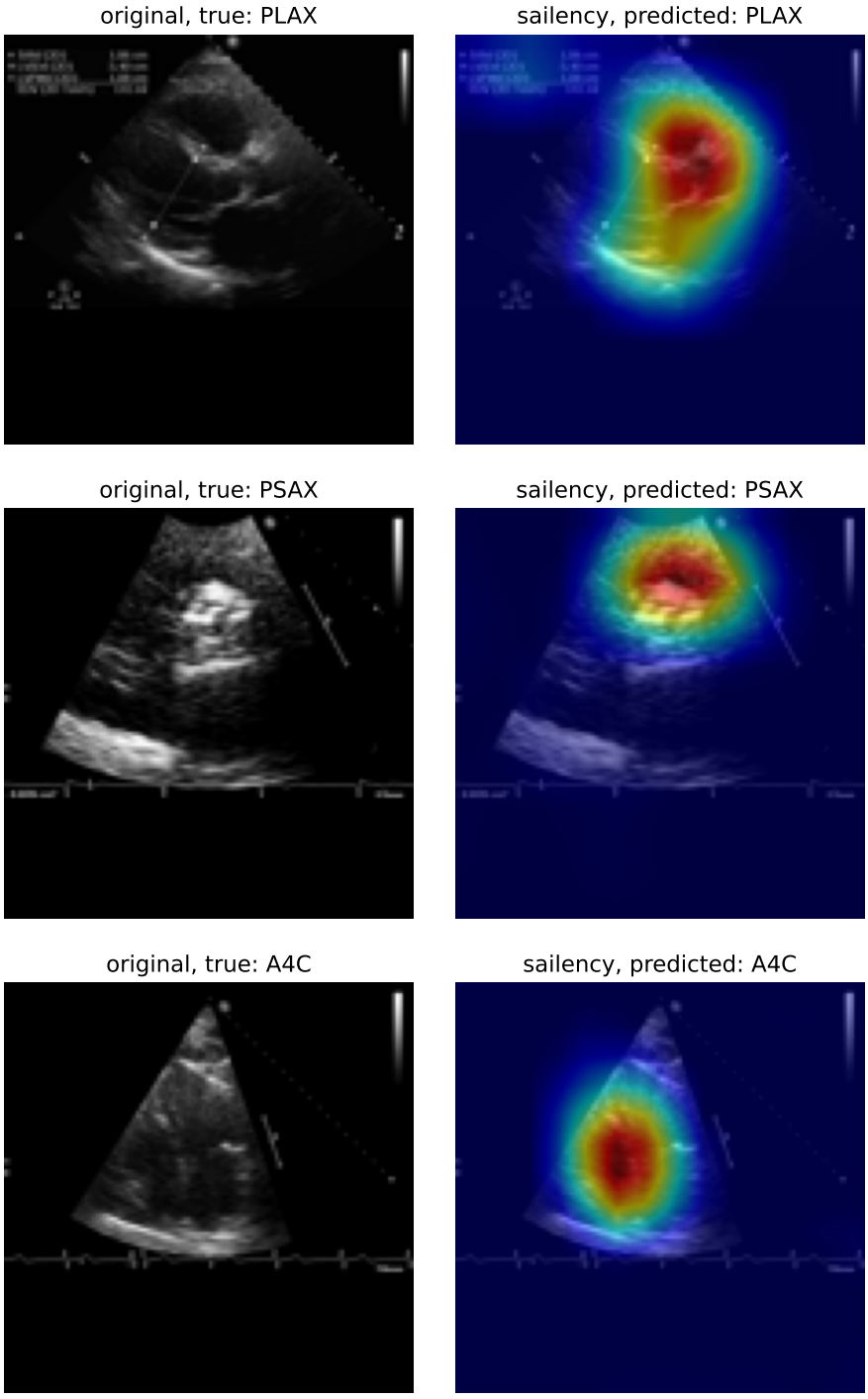

Figure E.1: **Example saliency maps for our view classifiers (cont'd on next page).** Each row shows an example image from our test set (left) and the corresponding Grad-CAM saliency map for our view classifier's prediction (right). The model correctly predicted the correct label of the image based on relevant region of the heart. The hotter the color, the more important the pixel in making the class discriminative decisions.

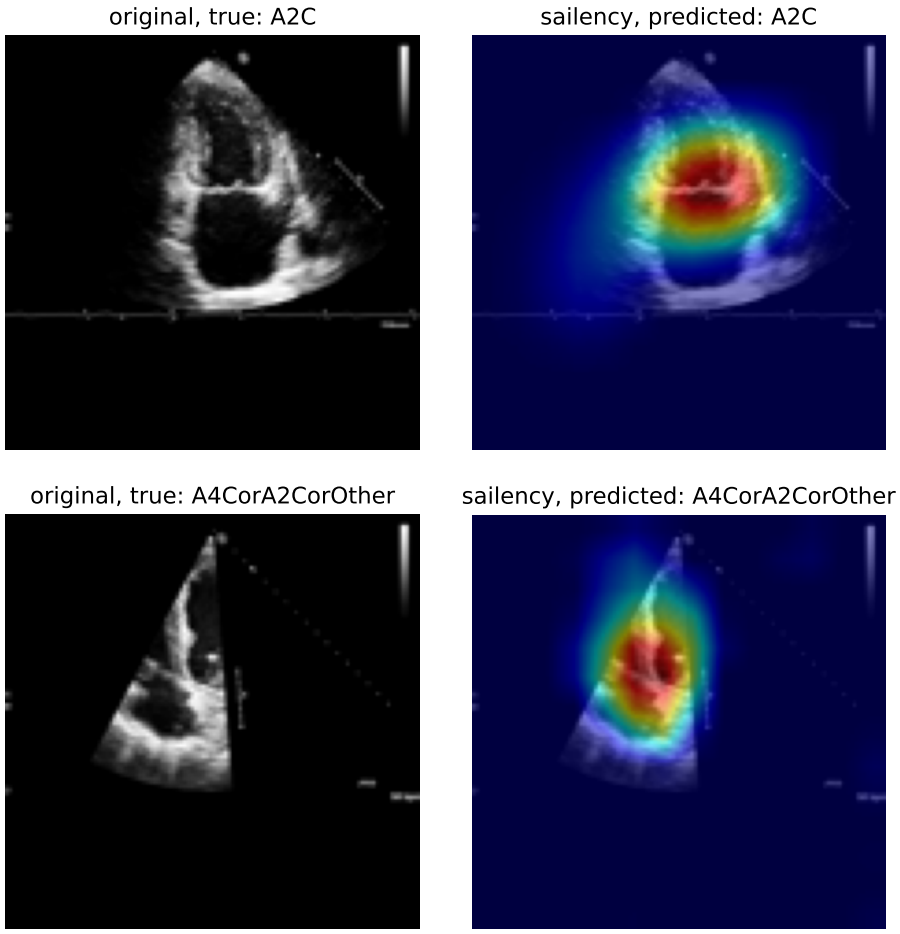

Figure E.1: **Example saliency maps for our view classifiers (cont'd from prev. page).** Each row shows an example image from our test set (left) and the corresponding Grad-CAM saliency map for our view classifier's prediction (right). The model correctly predicted the correct label of the image based on relevant region of the heart. The hotter the color, the more important the pixel in making the class discriminative decisions.