



Distance metrics predict out-of-distribution performance of AI oncology models for clinical deployment

Syed Rakin Ahmed^{1,2,3,4}, Charles Lu³, Jayashree Kalpathy-Cramer^{1,5}

¹Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Boston, MA

²Harvard Graduate Program in Biophysics, Harvard Medical School, Harvard University, Cambridge, MA

³Massachusetts Institute of Technology, Cambridge, MA

⁴Geisel School of Medicine at Dartmouth, Dartmouth College, Hanover, NH

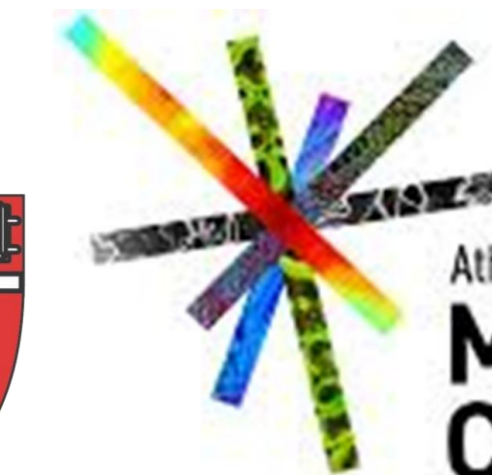
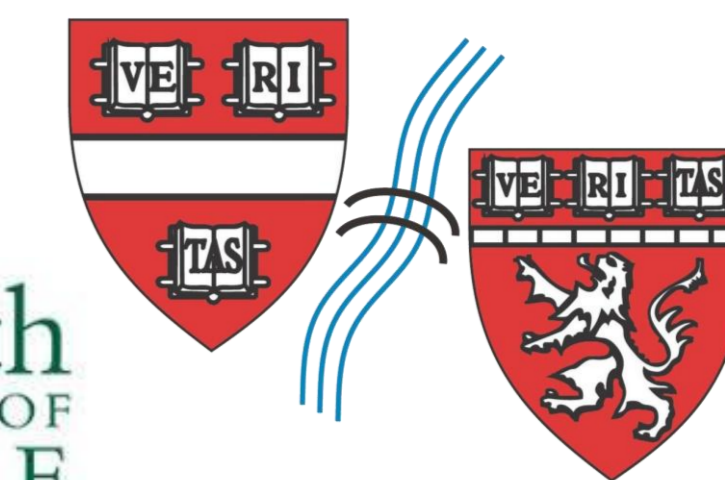
⁵Department of Ophthalmology, University of Colorado Anschutz, Denver, CO



THE HARVARD/MIT
MD-PHD PROGRAM

Dartmouth
Cancer Center

Dartmouth
GEISEL SCHOOL OF
MEDICINE



Athinoula A.
**Martinos
Center**
For Biomedical Imaging



BACKGROUND

While the development of artificial intelligence (AI) and deep learning pipelines for clinical tasks has surged in recent years¹⁻⁴, their translation into clinical practice remains sparse. A notable concern is the poor performance of models upon deployment due to differences in real-world clinical data from training data. Model robustness, defined as the ability to adapt well to out-of-distribution (OOD) data, is a particular challenge⁵, and there is a lack of explicit characterization and quantification of model robustness in the current literature.

MATERIALS AND METHODS

To assess model robustness, we designed and validated novel distance metrics and correlated them to retraining performance under dataset distribution shifts. We identified two axes of distance metric computation for a classification model:

- the n -dimensional feature (embeddings) vector space, $\mathbb{F} \in \mathbb{R}^n$; and
- the softmax prediction vector space, $\Delta^{|y|}$, where $|y|$ is the number of classes and Δ is the probability simplex.

Importantly, our approach treated the underlying model as a “black box”, relying solely on image-level model feature vectors and prediction vectors for metric computation.

- For each axis of distance metric computation, we calculated three metrics. In the features space, we calculated:
 - “Feature distance metrics”
 - cosine distance,
 - transport distance and,
 - KL divergence,
 with each metric computed between the centroids of the different data distributions.

- In the prediction vector space, we computed:
 - “Test estimation metrics”
 - average thresholded confidence with maximum confidence as the score function (ATC-MC)⁶,
 - average confidence and,
 - difference of confidences

We used the Digital Mammographic Imaging Screening Trial (DMIST) dataset⁷, a multi-device dataset for multi-class breast density classification consisting of 108,230 mammograms from 21,729 patients.

Classification metrics were evaluated on the test set (65% train : 10% validation : 25% test) for training runs encompassing distribution shifts between four mammography scanner devices (“ads”, “other”, “senograph”, “senoscan”), with UMAPs in Fig. 1.

- For each of the source-target pairs across each source training run, we calculated and subsequently correlated each “feature distance metric” and “test estimation metric” with relevant classification performance metrics, including:
 - average AUROC,
 - accuracy and,
 - linear kappa (LK).

RESULTS

- Among the “feature distance metrics”, transport distance (Eqn. 1)⁸ correlated best with classification performance for training runs on each source device (Fig. 2a).
- Intuitively, higher transport distance signifies a distribution farther from the source, and results in poorer model performance as reflected by the lower average AUROC, accuracy and LK.
- Among the “test estimation metrics”, ATC-MC (Eqn. 2)^{5,6} correlated best with classification performance (Fig. 2b).
- Since higher ATC-MC values indicate a stronger predicted classification performance, this is reflected in higher true classification performance.
- These correlations remained consistent across different model architectures (densenet121, Fig. 2i vs. resnet50, Fig. 2ii) and on repeat runs.
- Finally, the histograms in Fig. 3 highlight that the distribution of transport distance from the source centroid to each target point followed similar trends as the centroid-to-centroid “feature distance metric” in Fig. 2, thereby reinforcing true difference between the distributions.

$$\text{Transport Distance} = T(P, Q) = \inf_{\gamma \in \Pi(P, Q)} \mathbb{E}_{(x, y) \sim \gamma} [d(x, y)] \quad (\text{Eqn. 1})$$

where $\Pi(P, Q)$ is the set of all joint distributions whose marginals are P and Q .

Transport distance is also known as the Earth Mover's Distance or the Wasserstein Metric

$$\text{Average Thresholded Confidence} = \text{ATC}(X^T) = \mathbb{E}_{x \sim D^T} [\mathbb{1}[\max_{j \in y} s(x)_j > t]] \quad (\text{Eqn. 2})$$

where t is a threshold estimated on some score function, $s: \Delta^k \rightarrow \mathbb{R}$. In our case, the score function is taken to be the maximum confidence (MC)

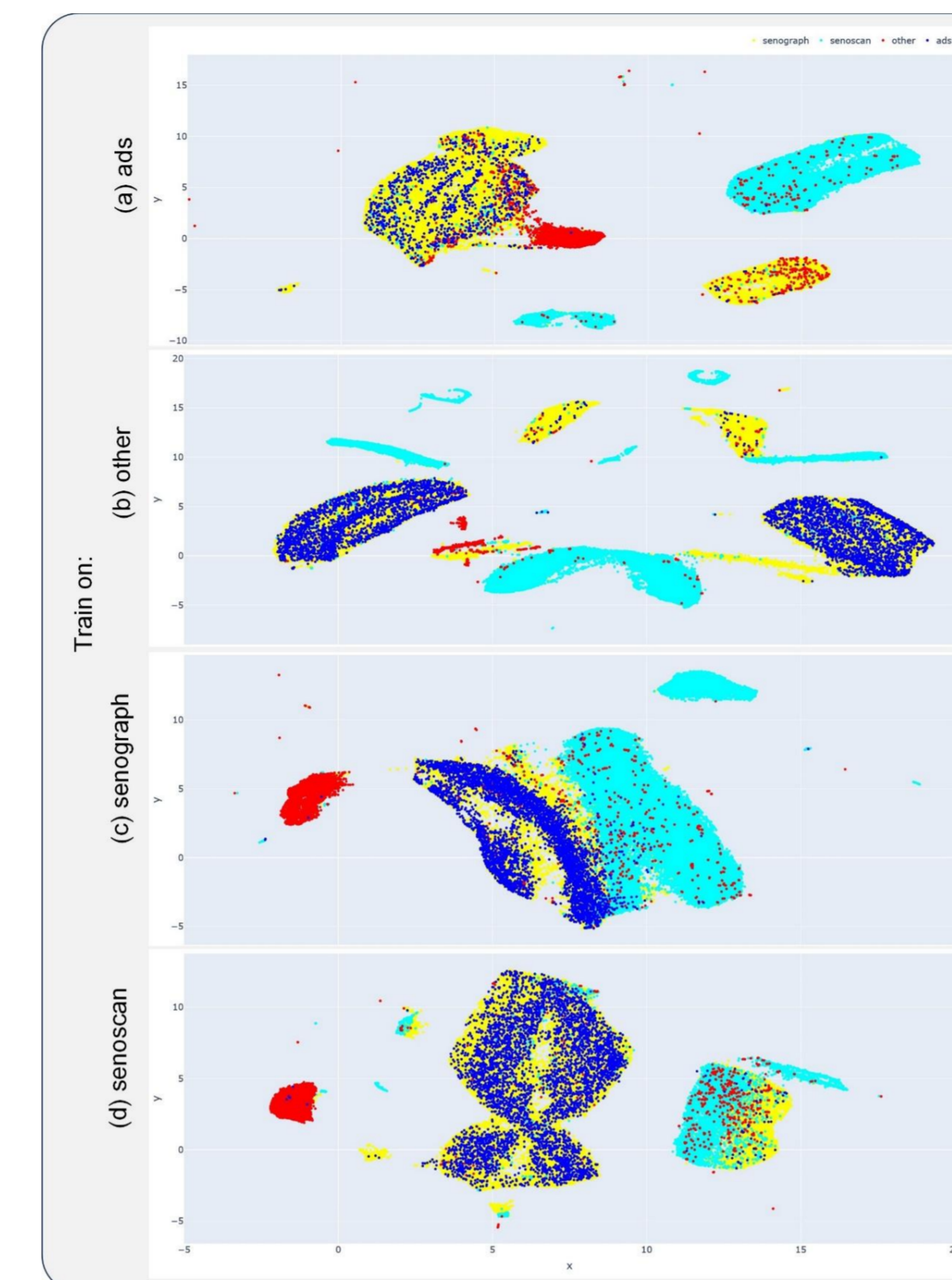


Figure 1: UMAPs generated on the test dataset for training runs on each source device for the breast density classification (DMIST) dataset. While a UMAP may be a useful dimension-reduced representation of a dataset, it is not necessarily indicative of model performance across axes of data heterogeneity. The higher dimensional feature space is a richer source of information likely to be more representative of dataset characteristics and is consequently used in this work for computation of “feature distance metrics”.

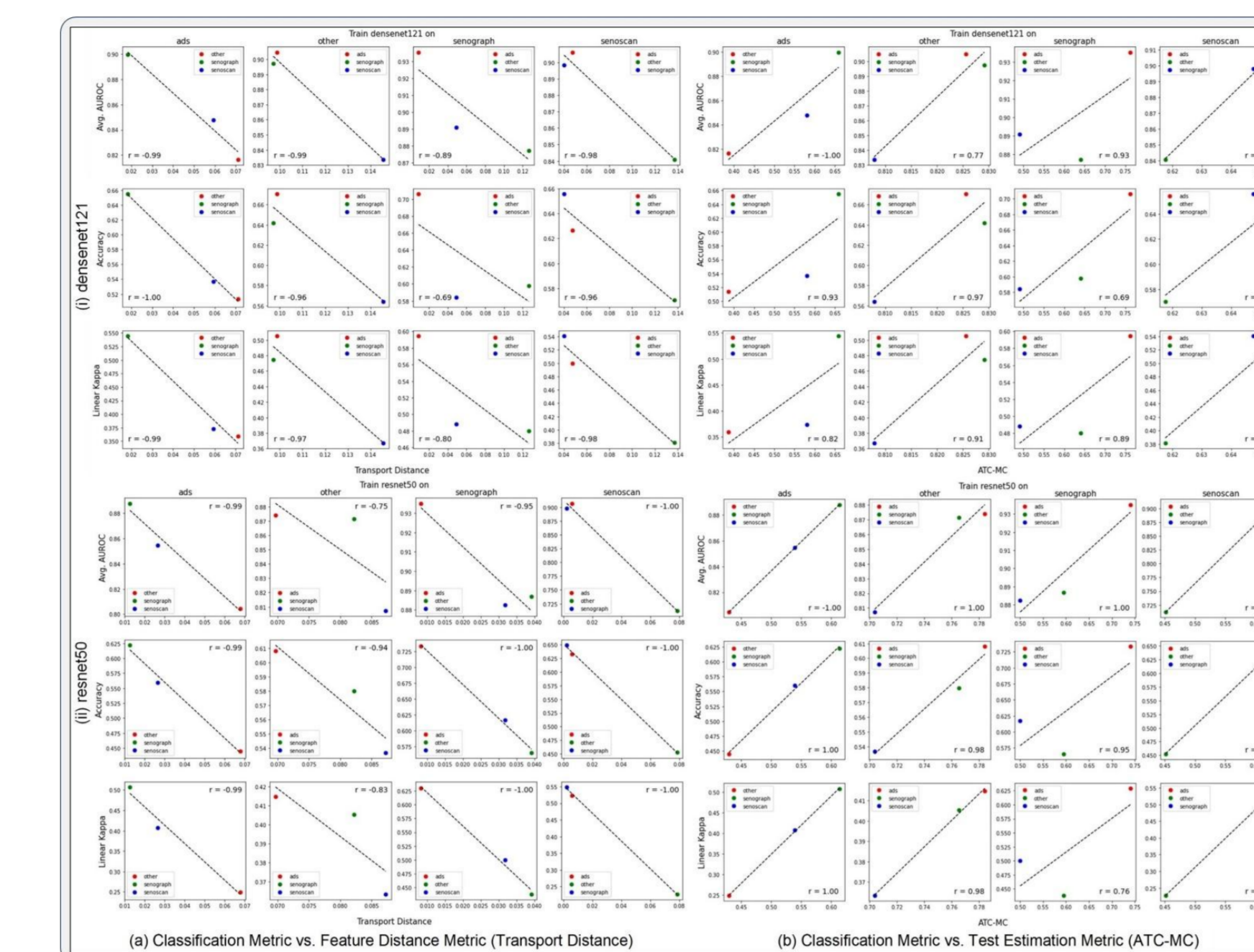


Figure 2: Plots highlighting the correlation between relevant classification performance metrics (avg AUROC, accuracy and linear kappa) with (a) transport distance (“feature distance metric”), and (b) ATC-MC (“test estimation metric”). All correlations are in the correct directions, with the corresponding Pearson's correlation coefficients, r , indicated on each plot. The correlations are robust to model architecture, as the trends remain consistent across the different model architectures, (i) densenet121 and (ii) resnet50, despite each architecture having a differently sized higher-dimensional embedding space.

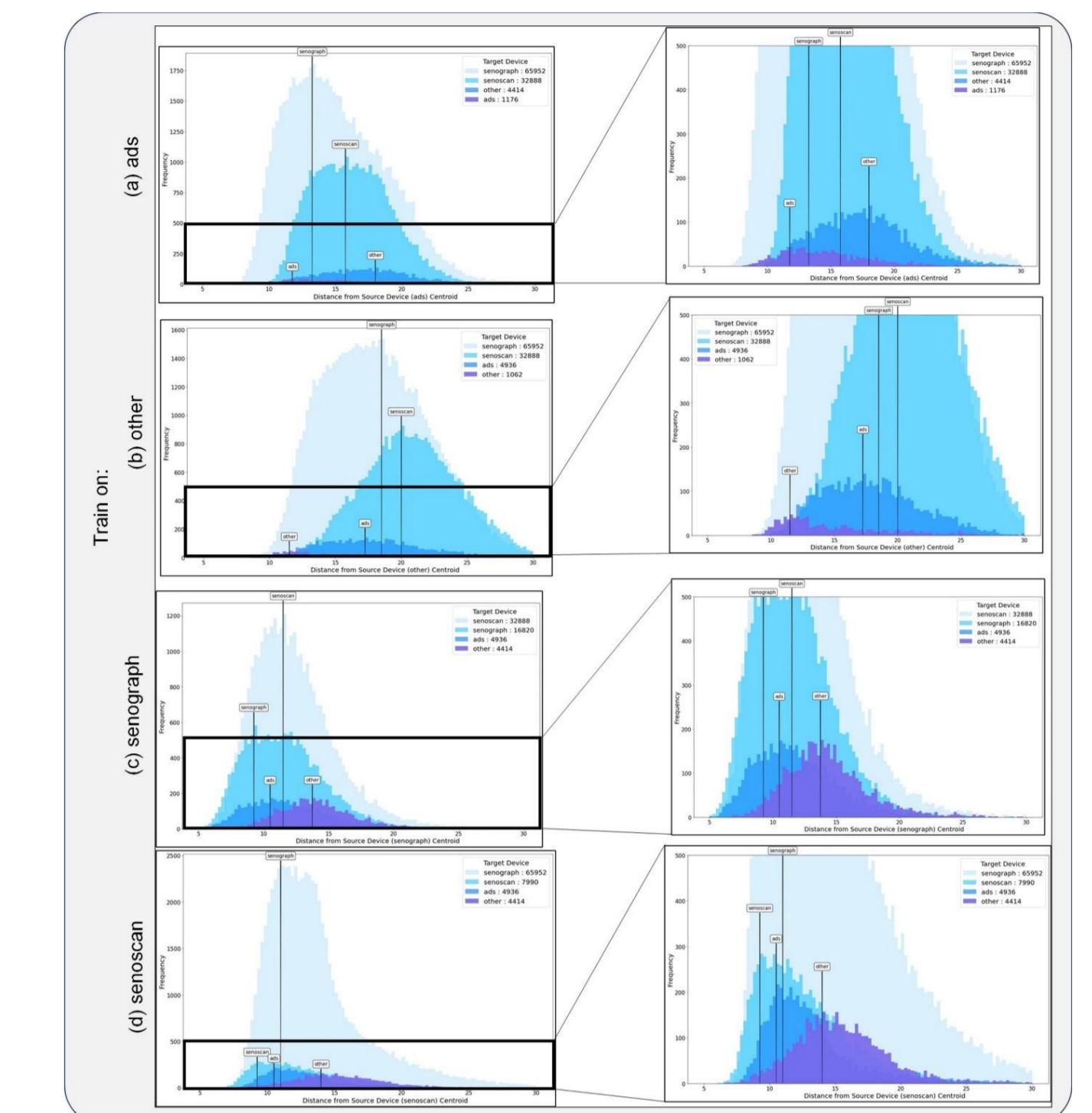


Figure 3: Histograms highlighting the distribution of transport distance (“feature distance metric”) between the feature vector of every image from the centroid of the source, for training runs on each source device. The right panel is a magnified view of the frequency range 0 – 500, and vertical lines highlight the mode of each device cluster. The distribution of transport distance from the source centroid to each target point follows similar trends as the centroid-to-centroid “feature distance metric” in Fig. 2, thereby reinforcing true difference/distance between the distributions. As an example, the (c) “senograph” training run produces a source-centroid-to-target-point transport distance distribution that follows the same trend as the source-centroid-to-target-centroid transport distance (i.e. distance from “ads” < “senoscan” < “other” throughout the dataset) and its corresponding correlation to model classification performance as highlighted in Fig. 2(a)(i); this is true for each of the source training runs in this figure.

CONCLUSIONS

- Prior to model deployment in the clinic, our approach can be used to generate calibration plots reflecting the relationship between distance metrics with model performance metrics across the different distributions in a well-characterized dataset.
- Subsequently, these calibration plots can be used to make predictions on the expected classification performance on newly acquired datasets of unknown distributions.
- In light of the FDA's October 2023 guidance on characterizing data distribution shifts to facilitate model deployment^{9,10}, our work shows strong correlations between well-validated distance metrics and model classification performance.
- Forthcoming work extends our analysis to additional datasets.

REFERENCES

- Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nat.* 2017 5427639 **542**, 115–118 (2017).
- Hannun, A. Y. *et al.* Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat. Med.* 2019 251 **25**, 65–69 (2019).
- Piccialli, F., Somma, V. Di, Giampaolo, F., Cuomo, S. & Fortino, G. A survey on deep learning in medicine: Why, how and when? *Inf. Fusion* **66**, 111–137 (2021).
- Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* 2019 251 **25**, 44–56 (2019).
- Lu, C., Ahmed, S. R., Singh, P. & Kalpathy-Cramer, J. Estimating Test Performance for AI Medical Devices under Distribution Shift with Conformal Prediction. (2022) doi:10.48550/arxiv.2207.05796.
- Garg, S., Balakrishnan, S., Lipton, Z. C., Neyshabur, B. & Sedghi, H. Leveraging Unlabeled Data to Predict Out-of-Distribution Performance. *ICLR 2022 - 10th Int. Conf. Learn.* (2022).
- Pisano, E. D. *et al.* Diagnostic Performance of Digital versus Film Mammography for Breast-Cancer Screening. *N. Engl. J. Med.* **353**, 1773–1783 (2005).
- Rubner, Y., Tomasi, C. & Guibas, L. J. Metric for distributions with applications to image databases. *Proc. IEEE Int. Conf. Comput. Vis.* 59–66 (1998) doi:10.1109/ICCV.1998.710701.
- Food and Drug Administration (FDA). Predetermined Change Control Plans for Machine Learning-Enabled Medical Devices: Guiding Principles. (2023).
- CDRH Issues Guiding Principles for Predetermined Change Control Plans for Machine Learning-Enabled Medical Devices | FDA. <https://www.fda.gov/medical-devices/medical-devices-news-and-events/cdrh-issues-guiding-principles-predetermined-change-control-plans-machine-learning-enabled-medical>.

CONTACT: rakin@mit.edu