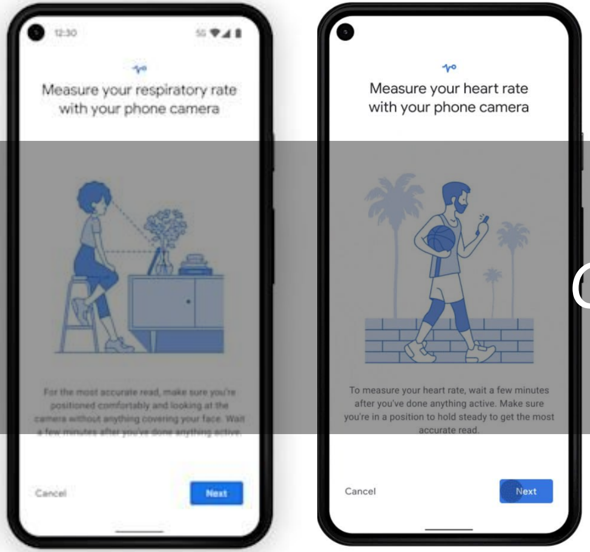
Reliable and Trustworthy Machine Learning for Health Using Dataset Shift Detection

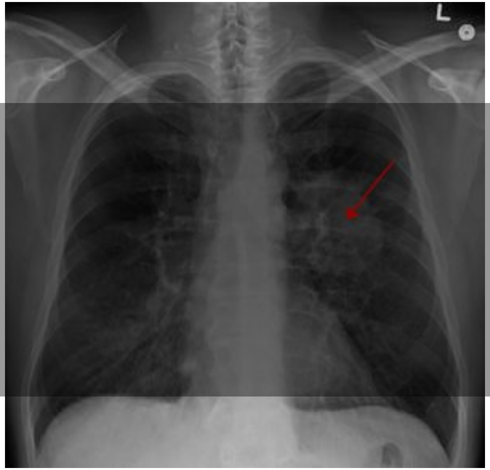
Chunjong Park, Anas Awadalla, Tadayoshi Kohno, Shwetak Patel

University of Washington

PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING
ubicomplab UNIVERSITY OF WASHINGTON
SECURITY & PRIVACY RESEARCH LAB UNIVERSITY of WASHINGTON

AI-Powered Health Applications





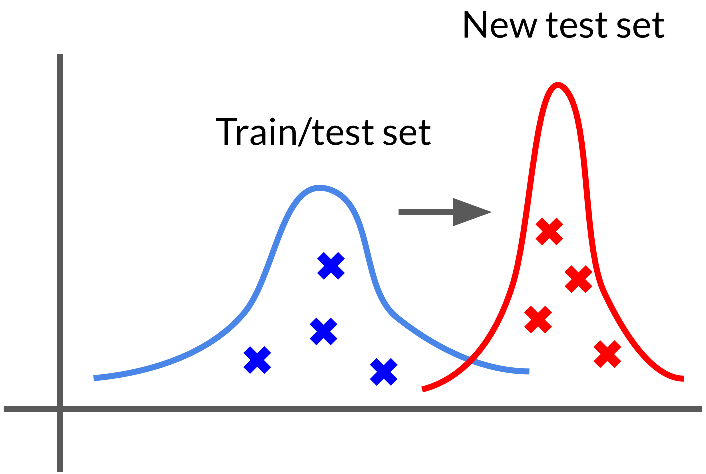
Early Screening
Continuous Monitoring
Access to Healthcare

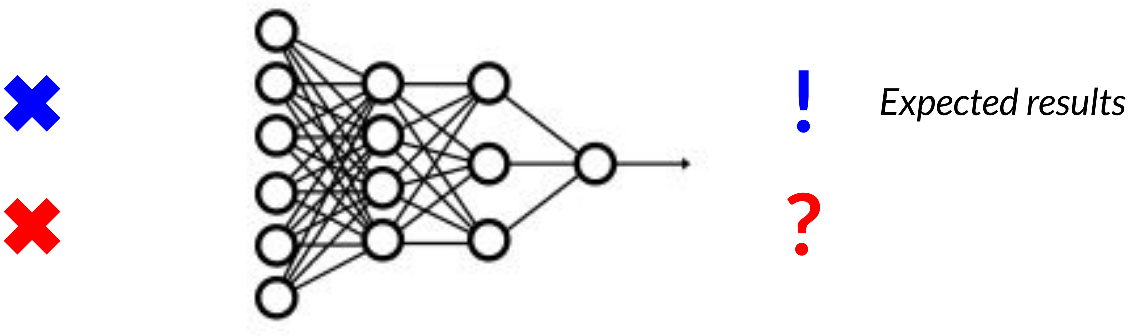
Google's heart/resp. rate

Apple's AFib detection

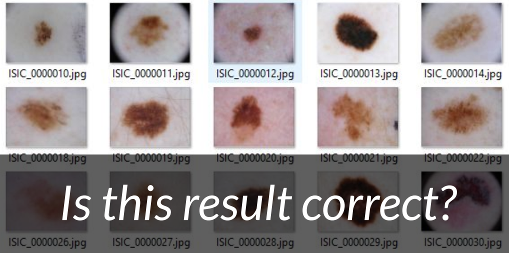
Cancer diagnosis

Dataset Shift





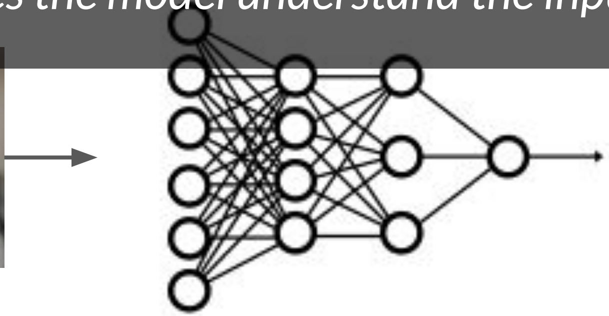
Dataset Shift in ML for Health



Is this result correct?

Can I trust this result?

Does the model understand the input image?

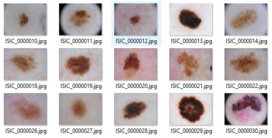


Malignant! ?

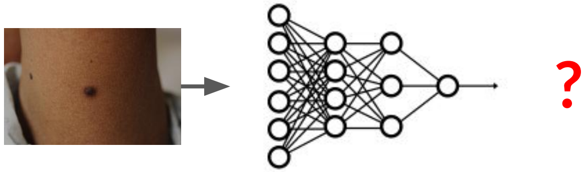
Dataset Shift in ML for Health

Difficult for non-experts to decide when to trust

- Medical decisions are high-stakes



Train dataset

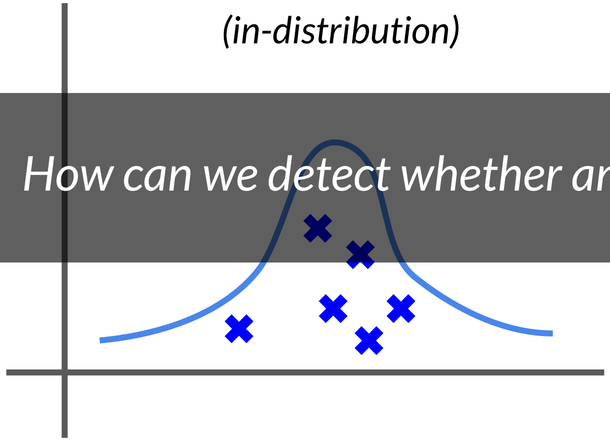


Difficult to get a complete coverage over a domain

- Emerging dataset for new diseases
- Device heterogeneity
- Potential bias within dataset

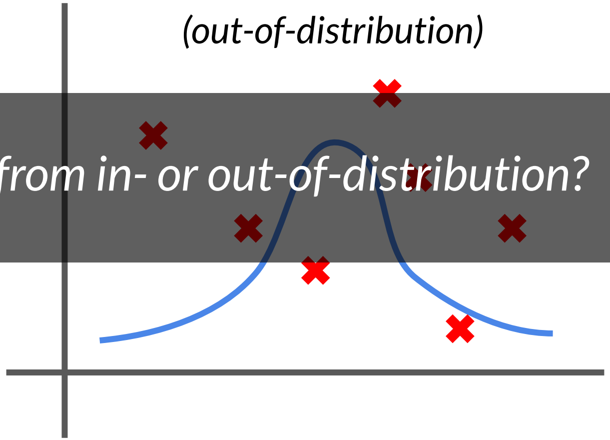
Expectation

Within train dataset
(in-distribution)



Reality

Outside of train dataset
(out-of-distribution)

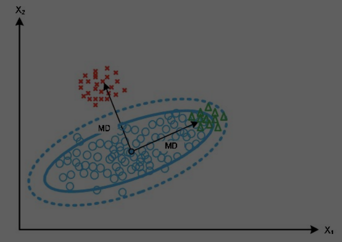


How can we detect whether an input example is from in- or out-of-distribution?

Out-of-Distribution Detection

Mahalanobis distance[1]

Distance from a distribution



Closest distance from class-conditional feature distributions at layer *l*

$$M_\ell = \arg\min_c \left(f_\ell(\mathbf{x}) - \mu_{\ell,c}\right)^\top \Sigma_\ell^{-1} \left(f_\ell(\mathbf{x}) - \mu_{\ell,c}\right)$$

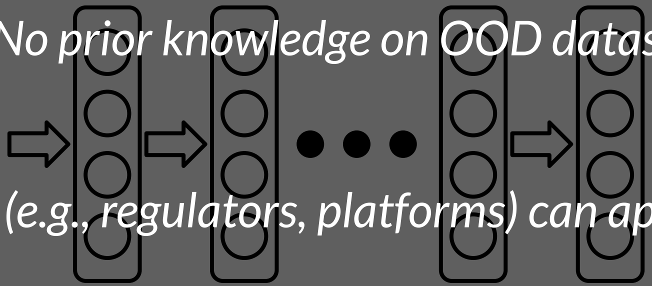
Weighted sum

$$\mathbf{M}(\mathbf{x}) = \textstyle\sum_\ell \alpha_\ell M_\ell$$

No re-training or network modifications

Works on any pre-trained models

No prior knowledge on OOD datasets



OOD score

3rd party stakeholders (e.g., regulators, platforms) can apply this to existing models

Gram matrices[2]

Pairwise feature correlation
Used for pattern and style encodings

p-th order Gram matrix at layer *l*

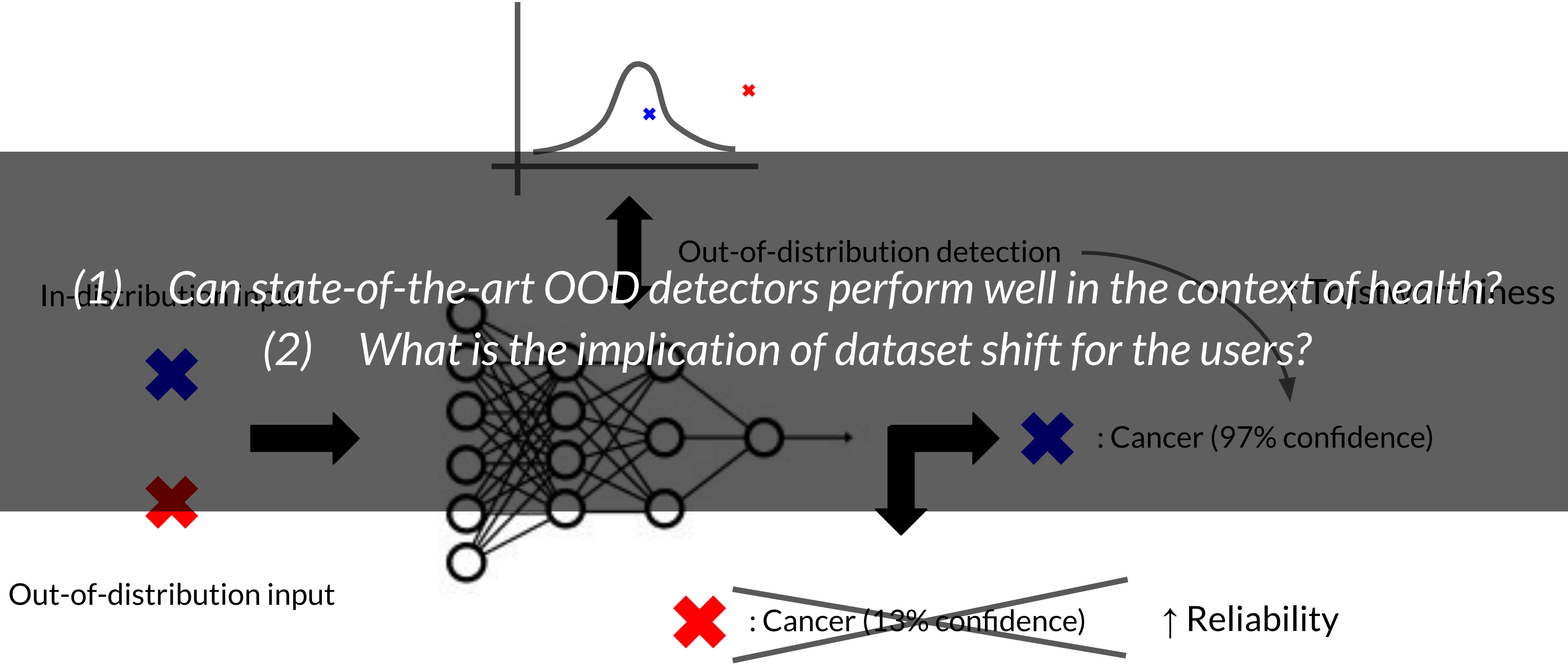
$$G_l^p = \left(F_l^p {F_l^p}^\top\right)^{\frac{1}{p}}$$

Normalized sum of layerwise deviation from in-distribution

$$\Delta(D) = \sum_{l=1}^{L} \frac{\delta_l(D)}{\mathbb{E}_{\mathrm{Va}}\left[\delta_l\right]}$$

1. Kimin Lee et al. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. NeurIPS 2018
2. Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with gram matrices. ICML 2020

Reliable and Trustworthy ML for Health



(1) Can state-of-the-art OOD detectors perform well in the context of health?

(2) What is the implication of dataset shift for the users?

Experiment Settings - OOD Detection

OOD methods: Mahalanobis distance, Gram matrices

	Skin Lesion Classifier	Lung Sound Classifier	Parkinson’s Classifier
Network	DenseNet-121	ResNet-34	5×1D-Conv
Train/test datasets	HAM10000 *skin lesion **images***	ICHBI 2017 *stethoscope lung **sound***	mPower ***acc. signal***
			Near-distribution
OOD datasets	ISIC2017	Digital Stethoscope	Kaggle Parkinson’s
	London Face	Audioset	MHEALTH
	CIFAR16		MotionSense

Far out-of-distribution

OOD Detection for Reliable ML for Health

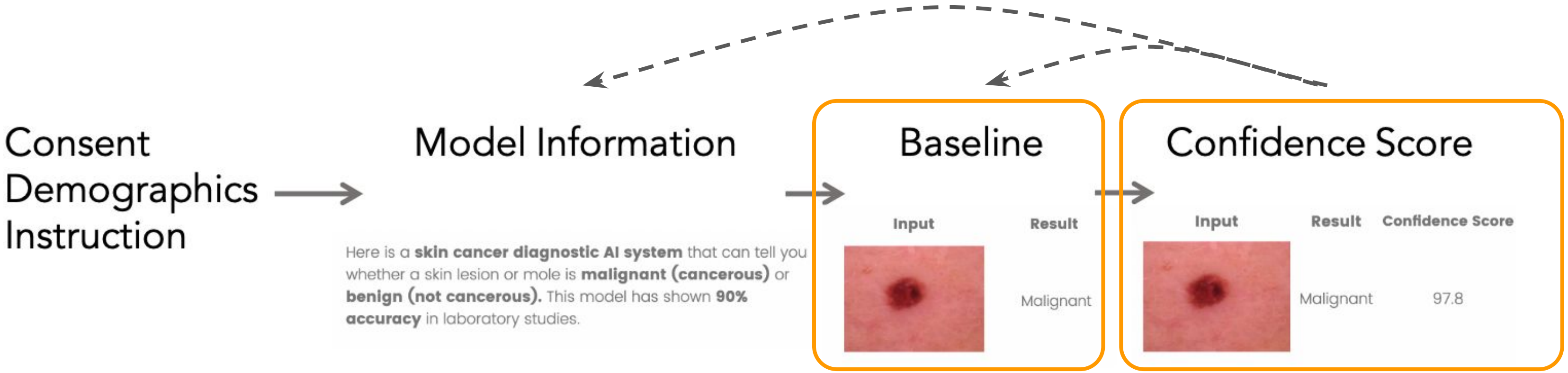
	OOD datasets	Detection Accuracy Mahalanobis distance	 Gram matrices
Skin lesion	ISIC2017	59.28	74.98
	London Face	**99.96**	**96.34**
	CIFAR16	**99.61**	**96.90**
Lung sound	Digital Stethoscope	80.57	76.05
	Audioset	**97.34**	**95.97**
Parkinson's	Kaggle Parkinson's	**99.47**	**99.67**
	MHEALTH	**100.00**	**99.99**
	MotionSense	**99.89**	**99.60**

Near-distribution

OOD Detection for Trustworthy Health ML Models

24 scenarios = 2 conditions (baseline vs. confidence score)
× 3 data types (image, audio, motion data)
× 2 *confidence score* (high vs. low)
× 2 results (positive vs. negative)

Online user study



Question for *baseline/confidence score*

1. User-perceived trustworthiness (5-point Likert scale)
2. Impact on making medical decisions (3-point Likert scale)

User Study Results

192 participants (155 male, 67 female, 42.7 ± 9.1 years old)

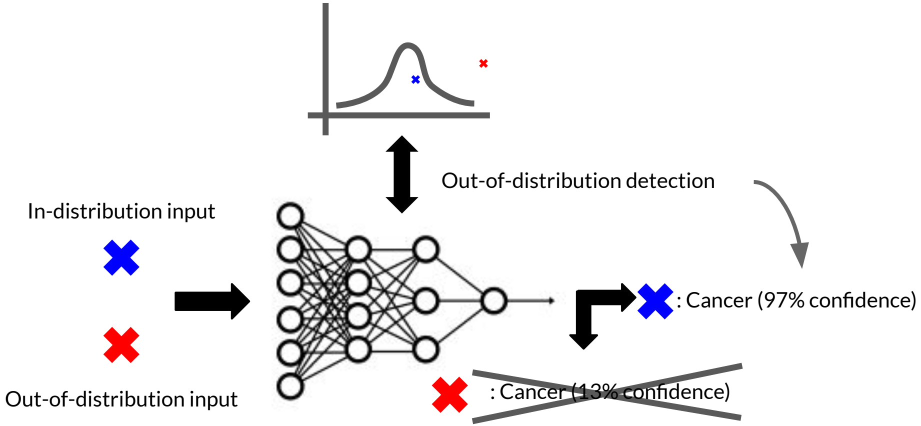
Higher trust for results with *confidence score* ($p<0.001$, $r=0.393$)

More willing make medical decisions with *confidence score* ($p<0.001$, $r=0.178$)

Larger effect in results with high *confidence score* ($r_{high}=0.475 > r_{low}=0.317$)

Effects **differ** by data types ($r_{image}=0.436 > r_{audio}=0.384 > r_{motion}=0.361$)

OOD Detection for ML for Healthcare



- Proposed a workflow for reliable/trustworthy ML for health
- OOD detectors can be applied to health ML using different data types
- OOD detection results improve user trustworthiness for health prediction results
- A step toward building trustworthy AI applications for high-stakes decision making

Chunjong "CJ" Park (cjparkuw@cs.washington.edu, www.cjpark.xyz)

PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING
ubicomplab UNIVERSITY OF WASHINGTON
SECURITY & PRIVACY RESEARCH LAB UNIVERSITY of WASHINGTON