Reliable and Trustworthy Machine Learning for Health Using Dataset Shift Detection

Chunjong Park, Anas Awadalla, Tadayoshi Kohno, Shwetak Patel
Paul G. Allen School of Computer Science & Engineering
University of Washington
{cjparkuw, anasa2, yoshi, shwetak}@cs.washington.edu

Abstract

Unpredictable ML model behavior on unseen data, especially in the health domain, raises serious concerns about its safety as repercussions for mistakes can be fatal. In this paper, we explore the feasibility of using state-of-the-art out-of-distribution detectors for reliable and trustworthy diagnostic predictions. We select publicly available deep learning models relating to various health conditions (e.g., skin cancer, lung sound, and Parkinson's disease) using various input data types (e.g., image, audio, and motion data). We demonstrate that these models show unreasonable predictions on out-of-distribution datasets. We show that Mahalanobis distance- and Gram matrices-based out-of-distribution detection methods are able to detect out-of-distribution data with high accuracy for the health models that operate on different modalities. We then translate the out-of-distribution score into a human interpretable CONFIDENCE SCORE to investigate its effect on the users' interaction with health ML applications. Our user study shows that the CONFIDENCE SCORE helped the participants only trust the results with a high score to make a medical decision and disregard results with a low score. Through this work, we demonstrate that dataset shift is a critical piece of information for high-stake ML applications, such as medical diagnosis and healthcare, to provide reliable and trustworthy predictions to the users.

1 Introduction

Advances in artificial intelligence and machine learning have made medical diagnostic and screening tools more accurate and accessible. AI-powered diagnostic tools [5, 14] are intended to assist medical personnel by making unbiased decision based on thousands of examples. In recent years, these models [15, 44, 48, 32] are even becoming available to consumers through the growth of mobile health with the intention of expediting diagnoses through increasingly frequent testing. Moreover, mobile health [4, 50] aims to improve access to medical expertise for those who are uninsured or live far away from hospitals.

Despite the potential benefits of health AI systems, there are concerns about their performance in real-world settings. Data-driven models learn from examples, making them heavily reliant on the data upon which they have been trained. However, datasets often fail to get complete coverage over a domain, particularly for emerging datasets; when new pulmonary diseases (e.g., MERS and COVID-19) emerge, a pulmonary classifier trained on the existing lung sounds would not be able to interpret sound of the new diseases. Previous work [42, 78] has found that machine learning models behave unpredictably on the unseen data. This problem [4, 69] is especially critical for medical diagnostic and screening tools since there are significant repercussions for mistakes.

Researchers have proposed methods to estimate the uncertainty of a machine learning models' predictions based on the input [29, 41, 40, 64, 43]. Out-of-distribution detection methods can distinguish whether the input lies within the distribution of the training dataset, with out-of-distribution

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

data leading to less reliable prediction results. However, such important information has not been widely explored in the context of health applications. When health applications are put into the hands of consumers with limited understanding of the underlying algorithms, they may upload poor quality data that lies outside the distribution of the data that was collected by experts. For example, consumers who are using a health application that involves image processing may take photographs in poor lighting conditions or framing of the target object. Even when the data is high-quality, it may be captured with a smartphone that has different hardware specifications than the devices that were used to collect the training dataset. Unless the models are explicitly designed or trained to detect invalid data, the models will incorrectly produce a diagnostically meaningless result.

In this work, we explore the utility of out-of-distribution detection for improving model performance and user-perceived trustworthiness of health-related models. We first benchmark our approach using publicly available deep learning models relating to various medical challenges and sensing domains — images for skin lesion classification, motion data for Parkinson's disease severity, and audio for lung sound classification. After demonstrating that these models are susceptible to dataset shift, we demonstrate that the state-of-the-art out-of-distribution detectors can effectively exclude such data with over 95% detection accuracy in most cases. We then explore the implications of this detector on user-perceived trustworthiness of the health models. After translating the out-of-distribution score into a human-interpretable metric, CONFIDENCE SCORE, we found that showing this information to end-users improved the user-perceived trustworthiness of the models. Furthermore, participants stated that they were more willing to make medical decisions based on models when they were shown the certainty metric. Our contributions in this work are as follows:

- We identify and quantify the limitations of current health deep learning models when encountered with unseen data,
- We evaluate the utility of out-of-distribution detection on various data types (e.g., image, audio, motion) for medical screening and diagnosis, and
- We evaluate the impact that dataset shift information has on user-perceived trustworthiness of health diagnostic results.

2 Related Work

Machine Learning-based Health Screening and Diagnosis In recent years, machine learning has been widely used for medical diagnosis and screening tool to help doctors diagnosis patients easier, faster, and more accurately. Machine learning models that learn from large-scale medical datasets are able to detect various symptoms and conditions, including mental health [26, 68], retinal disease [14], lung cancer [5]. With the increasing ubiquity of smartphone and advances in its computing power, machine learning-based health screening can be done on mobile devices. Various machine learning-based mobile health applications have been proposed to detect health conditions (e.g., traumatic brain injury [49], pancreatic cancer [48], jaundice [15]) and vital signals (e.g., heart rate [44], respiratory rate [44], heart rate variability [32], blood pressure [65], SpO2 [31]). Such mobile health applications can benefit nurses, health workers, and the general population for easier medical screening.

While the health machine learning models show high accuracy on their own test datasets, their performance is questionable in real-world settings where the input data can vary drastically, resulting in unreliable prediction results [70, 63]. Researchers have investigated the dataset shift problem for medical imaging (e.g., x-ray [11, 10], fundus eye images [11], CT scans [74], dermatology [63, 56]), focusing on developing and evaluating out-of-distribution detection methods for specific domains. However, as more consumer-facing health applications are available in the market, this issue can lead the users to make medical decision based on incorrect results. In this work, we aim to explore ways to leverage dataset shift information to make the health machine learning models more reliable and trustworthy to the users.

Dataset Shift Detection Recently researchers have proposed various methods to estimate the models' uncertainty due to dataset shift. The proposed methods leverage the output of the models to effectively detect *out-of-distribution* input that are different from the known distribution, *in-distribution*. Softmax confidence [29] has been the baseline for the out-of-distribution detection. Several work has been proposed for out-of-distribution detection using deep ensemble [38], Mahalanobis distance [40], Gram matrices [64], energy score [43], temperature scaling [41, 64], input perturbation [41, 40], mean and variance of channels activations [58]. Alternate training strategies [30, 39, 47, 52] have

been proposed to enable model to detect out-of-distribution. Generative models [54, 60, 53, 77] are proposed to detect out-of-distribution examples. However, many approaches require re-training and re-designing of the models and prior knowledge of out-of-distribution datasets; it is not realistic to apply these methods to the existing models. In this work, we explore Mahalanobis distance- [40], Gram matrices- [64], and energy-based [43] out-of-distribution detection methods for reliable and trustworthy machine learning for health since these methods show reasonable out-of-distribution detection performance, do not require retraining or prior knowledge of out-of-distribution datasets, and work on pre-trained discrimitive classifiers.

Trustworthy AI Machine learning systems are deployed in real-world settings to billions of users, making significant impacts on high-stake decision making such as healthcare, policy, economy, and transportation. Failures in machine learning systems can cause fatal consequences and building trustworthy AI is one of the most important problems in machine learning community. In recent years, there are active and ongoing efforts aimed at making machine learning systems causal [6, 55], explainable [2, 36, 45, 34, 35], fair [3, 59, 17], robust [18, 21, 28, 19, 72], and privacy-preserving [1, 57, 51, 71]. This work contributes to trustworthy AI by improving reliability and user-perceived trustworthiness of machine learning for health using estimated uncertainty. Bhatt et al. [8] proposed to leverage uncertainty for users making decision and placing trust in machine learning models. This work explores similar approach where we adopt out-of-distribution detection as a method to measure uncertainty. We took a step further to investigate and quantify its effect in improving reliability and trustworthiness in the context of machine learning for health.

3 Background: Dataset Shift Detection Methods

We aim to leverage state-of-the-art out-of-distribution detection methods [40, 64, 43] in the health domain for users to safely use health deep learning models. We selected three out-of-distribution methods that show high accuracy on different out-of-distribution datasets, do not require re-training and prior knowledge of out-of-distribution datasets, and work on pre-trained discrimitive classifiers. These characteristics are important to help developers or other stakeholders (e.g., regulators, auditors, platforms) easily adopt out-of-distribution detectors to any pre-trained models. In this section, we provide background on each out-of-distribution detection methods.

3.1 Mahalanobis Distance-Based Out-of-Distribution Detection

Mahalanobis distance is used to measure the proximity of a point to a certain Gaussian distribution. In the Mahalanobis distance-based out-of-distribution detector [40], this property is used to represent each class's samples at each layer of a network as a class conditional Gaussian distribution with mean $\hat{\mu}_{cl}$ and co-variance $\hat{\Sigma}_{cl}$, where c indicates the class and l indicates the layer in the model. Given a sample input x to a neural network, for each layer, it computes the minimum layer-wise class conditional Mahalanobis distances for x. That is, for each layer, it finds the Mahalanobis distance associated with the closest class to x. In other words, this is equivalent to $M(x) = \max_c -(f(x)-\mu_c)^T\Sigma^{-1}(f(x)-\mu_c)$. The authors have demonstrated that adding small noise to the input can help better distinguish between in-distribution and out-of-distribution data. As the authors suggested, for the real-world setting where the out-of-distribution datasets are generally not available, we obtain the input noise magnitude by generating adversarial samples generated by FGSM [27].

3.2 Gram Matrices-Based Out-of-Distribution Detection

Gram matrices are used to compute pairwise feature correlations and encode stylistic attributes. For out-of-distribution detection [64], higher order Gram matrices are used to compute class-conditional bounds of feature correlations at all hidden layers of the network as higher order shows better out-of-distribution detection performance. Higher order Gram matrices is expressed as $G_l^P = (F_l^P F_l^{P^T})^{\frac{1}{P}}$, where F_l is feature map at l-th layer and P is order. All elements of Gram matrices of an input at each layer are compared against the prepossessed minimum and maximum Gram matrices element values from in-distribution dataset to obtain deviation. If the input data is predicted as a certain class, the minimum and maximum values of the corresponding class will be used for comparison. The comparison is done for each layer to obtain layerwise deviations. Then, the deviations are used to get a total deviation, which is defined by the normalized sum of layerwise deviations. Whether the input data is from out-of-distribution is determined with a threshold which is defined as 95% percentile of the total deviations of in-distribution energy score distribution.

3.3 Energy-Based Out-of-Distribution Detection

The energy-based out-of-distribution detector [43] seeks to provide an alternative scoring function to the softmax function that is less susceptible to over-confidence and therefore can better distinguish between in and out-of-distribution inputs. It takes a discriminative classifier $f(c)$ that maps input $x \in \mathbb{R}^D$ to logits, which are traditionally used to derive a categorical confidence score using a softmax function. It defines the energy function on the classifier as $E(x; f) = -T \cdot log \sum_i^K e^{f_i(x)/T}$, where K is the number of classes in the model's output space and T is a temperature parameter that can be used to alter the shape of the energy score distribution. Energy score threshold that distinguishes between in- and out-of-distribution samples is defined at the 95% percentile of the in-distribution samples.

4 Performance Evaluation

In this section, we demonstrate the performance degradation of the existing health models when encountered with out-of-distribution datasets highlighting that the existing models are vulnerable to dataset shift. We then evaluate the performance of state-of-the-art out-of-distribution detectors for distinguishing between in- and out-of-distribution examples. We have selected out-of-distribution datasets that consist of both near- and far-from-distribution samples that represent realistic use cases in the real-world settings. For mobile health applications that use mobile sensors for health screening, non-expert users are expected to input data collected by themselves. Unlike clinicians, who may either receive training on how to operate these mobile apps or may already understand what must be done to generate high-quality, non-expert consumers may collect relevant but low-quality data due to environmental factors or totally irrelevant data by mistake or a lack of understanding. To reflect these scenarios, we include out-of-distribution datasets caused by covariate shift, label shift, and open-set recognition. The covariate and label shifts aim to evaluate a model's performance when tested on data pertaining to the same classification task but from different data sources and environment. Open-set recognition evaluates the model's performance on new classes not included in the training set. In Table 2, we indicate dataset shift type for each out-of-distribution dataset.

4.1 Models and Datasets

Skin lesion A DenseNet-121 based skin lesion classifier [56] was used in this work. The model aims to classify an image into seven different skin lesions: actinic keratoses, basal cell carcinoma, benign keratosis, dermatofibroma, melanoma, melanocytic nevi and vascular lesions. The following datasets are used for training and evaluation:

- (**In-distribution**) HAM10000 [73, 12]: (CC BY-NC 4.0) A dataset containing 10,000 samples of dermatoscopic skin tumor images taken using different devices and from different populations. These tumors are part of 7 classes: actinic keratoses, basal cell carcinoma, benign keratosis-like lesions, dermatofibroma, melanoma, melanocytic nevi, and vascular lesions.
- (**Out-of-distribution**) ISIC 2017 [13]: (CC BY-NC 4.0) A previous version of the HAM100000 dataset which contains 2000 dermatoscopic skin tumor images labelled for binary classification. A tumor is labelled malignant if it corresponds to melanoma to benign if it corresponds to nevus or seborrheic keratosis.
- (**Out-of-distribution**) Face [16]: (CC BY 4.0) A dataset containing frontal view face images of 102 adults without making a neutral facial expression. Face images are personally identifiable information. But, all individuals gave signed consent for their images to be "used in lab-based and web-based studies in their original or altered forms and to illustrate research (e.g., in scientific journals, news media or presentations)."
- (**Out-of-distribution**) CIFAR-10 [37]: (MIT License) A common image classification benchmark with 10 non-medical classes (airplane, car, cat, dog, horse, bird, deer, ship, frog, truck) which contains 6,000 images per class.

Lung Sound A lung sound classification model [23] classifies normal lung sound, wheeze, and crackle from an audio sample. This model is based on ResNet-34 and uses spectograms of audio samples as inputs and outputs 4 lung sound classes (normal, wheezing, crackle, and wheezing + crackle).

Health ML Models	In-Distribution	Out-of-Distribution		
Skin Lesion (DenseNet-121)	HAM10000 92.05%	ISIC 2017 74.00%	Face N/A	CIFAR N/A
Lung Sound (ResNet-34)	ICBHI 2017 78.50%	Stethoscope 2.10%	AudioSet N/A	
Parkinson's Disease (5×1D-Conv)	mPower 82.01%	Kaggle Parkinson's 26.67%	MotionSense 45.83%	MHEALTH 10.00%

Table 1: Accuracy of health deep learning models on in-distribution and out-of-distribution dataset. Accuracy is not available (N/A) for out-of-distribution datasets that do not have corresponding labels.

- (**In-distribution**) ICBHI 2017 Respiratory Challenge [61]: A dataset collected using multiple microphones and stethoscopes containing 6898 samples normal lung sound, wheeze, and crackle audio
- (**Out-of-distribution**) Stethoscope [22]: (CC BY 4.0) A dataset containing stethoscope respitory sounds with 336 samples of normal breathing, wheeze, and crackle audio sounds. The dataset was collected using a 3M Littmann Electronic Stethoscope.
- (**Out-of-distribution**) AudioSet [24]: (CC BY 4.0) A large dataset of millions of sound labelled YouTube audio of which a portion of the dataset contains breathing, cough, and wheezing samples which we use to create a suitable out-of-distribution dataset for this model.

Parkinson's Disease This is a binary classification model [76] that showed highest performance in Parkinson's disease digital biomarker DREAM challenge [66]. The model uses accelerometer signals to detect tremors in a person's movement and outputs whether a participant has Parkinson's. This model consists of 5 1D-convolutional layers and a single output.

- (**In-distribution**) mPower [9]: (CC BY 4.0) A dataset contains 30-second accelerometer readings from 3,100 participants at rest for both healthy and Parkinson's patients. The dataset was used in Parkinson's disease digital biomarker DREAM challenge [66].
- (**Out-of-distribution**) Kaggle Parkinson's disease [25]: (CC0 1.0) A dataset with accelerometer readings from healthy participants simulate movements of Parkinson's patients.
- (**Out-of-distribution**) MotionSense [46]: (MIT License) A dataset contains accelerometer readings from 24 participants performing various activities (e.g., walking, jogging, sitting, standing, etc).
- (**Out-of-distribution**) MHEALTH [7]: An activity classification dataset which contains accelerometer readings from 10 participants executing various activities (e.g., standing, sitting, walking, cycling, etc).

4.2 Performance Impact by Dataset Shift

In evaluating the model's performance on the out-of-distribution dataset, we used pre-trained models from the previous work[1] [23] when the authors make it available. Otherwise, we trained the model in the same way specified in their work[2] [56, 76]. We trained skin lesion model [56] for 150 epochs using Adam optimizer with a learning rate of 0.0001 and weight decay of 0.2. For Parkinson's disease model [76], we trained for 50 epochs using Adam optimizer with a learning rate of 0.0005. The pre-trained lung sound model [23] is trained for 200 epochs using SGD optimizer with a learning rate of 0.001 and momentum of 0.9. For all of these models, we used an 80/20 split and applied the same preprocessing for train and test sets. All training and testing is done in a server (Intel Xeon 2.1GHz, 64GB, GeForce RTX 2080 Ti) from an internal cluster. We then ran inference on each dataset and calculated the classification accuracy for the datasets that have corresponding labels. For the datasets that do not have the same labels from the in-distribution, the accuracy could not be computed. Table 1 summarizes the classification accuracy for the models on in- and out-of-distribution datasets.

[1] `https://github.com/microsoft/RespireNet`

[2] `https://github.com/GuanLab/PDDB`

We generally observed a significant performance drop for all health machine learning models that are tested with out-of-distribution datasets. The models output unreasonable and arbitrary predictions on datasets that are not related health conditions. For example, skin lesion classifier predicts all face images as vascular lesions and CIFAR10 images as various types skin lesions. Similarly, Parkinson's disease classifier predicts significant portion of physical activities by health participants as tremor caused Parkinson's disease. For lung sound classification, ordinary sound events (e.g., speech, walking, laughing) are classified as a certain type of lung sounds (e.g., crackle, wheezing). When the models are evaluated on out-of-distribution datasets that have similar data characteristics to the in-distribution data(i.e., near-distribution datasets), all health models exhibit a performance decrease that ranges from 18% to 76%. This implies that the models are also sensitive to small dataset shift, such as datasets collected with different devices and in different environments. All of these failure scenarios can occur in real-world deployment of health machine learning applications. Users can input a face image to skin lesion classifier, improperly record lung sound and input ambient sound to the lung sound classifier, or input motion data when they are not at rest to the Parkinson's disease classifier. Furthermore, diverse sensors and devices used in real-world deployment can cause significant performance drop. This evaluation demonstrates that users are exposed to the health machine learning applications that can provide unreliable diagnostic results.

4.3 Out-of-Distribution Detection Performance

The previous evaluation implies that it is crucial to determine whether the input data belongs to in- or out-of-distribution to avoid failures caused by dataset shift. In this section, we investigate the feasibility of using state-of-the-art out-of-distribution detection methods in the context of machine learning for health. We evaluate Mahalanobis distance- [40], Gram matrices- [64], and energy-based [43] methods, which work on any pre-trained discrimitive classifiers and do not need re-training and prior knowledge of out-of-distribution datasets, in detecting out-of-distribution data for different health models.

4.3.1 Experimental Setup

For Mahalanobis distance-based method[3], we extracted Mahalanobis distance-based scores from the output dense and residual block found in DenseNet and ResNet respectively. For the Parkinson's model which does not contain dense and residual blocks, we extracted the scores at the end of each convolutional layer. Then, we optimized the input noise magnitude using in-distribution samples and corresponding adversarial samples generated by FGSM [27]. The noise magnitude obtained is 0.0 for skin lesion classifier, 0.0005 for lung sound classifier, and 0.0 for Parkinson's disease classifier. For Gram matrices-based method[4], we extracted class-specific minimum and maximum correlation values for all orders of Gram matrices for all feature pairs. Total deviation values, which are used for out-of-distribution detection threshold, are computed with multiple sets of random samples from in-distribution datasets. For energy-based method[5], we use their method that does not require fine-tuning to avoid re-training of the network. We use the default temperature scaling ($T = 1$) as suggested in [43]. All evaluations are repeated for 5 trials and we report the mean (Table 2) and standard deviation (Table 4) of all metrics.

4.3.2 Evaluation Metrics

For out-of-distribution detection, it is common to use true negative rate (TNR) at 95% true positive rate (TPR), AUROC, and detection accuracy to evaluate the performance of a detector. Particularly, as the out-of-distribution problem is a binary classification problem, we consider out-of-distribution samples as negative and in-distribution samples as positive. TNR at TPR 95% is defined as the percentage of correctly detected out-of-distribution samples, when 95% of in-distribution samples are correctly detected. The AUROC metric measures the area under the TPR vs FPR curve. The detection accuracy measures the maximum possible classification accuracy over all possible thresholds in distinguishing between in-distribution and out-of-distribution examples. Detailed explanations on the metrics are available in Appendix B.

[3] https://github.com/pokaxpoka/deep_Mahalanobis_detector

[4] https://github.com/VectorInstitute/gram-ood-detection

[5] https://github.com/wetliu/energy_ood

Health ML Models	In-Distribution	Out-of-Distribution	Distribution Shift	Validation on OOD Samples (TNR @ TPR95/AUROC/Detection Accuracy)		
				Mahalanobis	Gram	Energy
Skin Lesion (DenseNet-121)	HAM10000	ISIC 2017	Covariate/label shift	10.13 / 58.21 / 59.28	25.90 / 81.14 / 74.98	14.28 / 76.20 / 70.76
		Face	Open-set recognition	100.00 / 99.98 / 99.96	95.01 / 98.20 / 96.34	0.00 / 80.45 / 84.81
		CIFAR10	Open-set recognition	99.83 / 99.90 / 99.61	95.14 / 98.66 / 96.90	5.06 / 58.33 / 57.89
Lung Sound (ResNet-34)	ICBHI	AudioSet	Open-set recognition	97.96 / 99.47 / 97.34	96.55 / 99.18 / 95.97	8.12 / 56.79 / 57.13
		Stethoscope	Covariate/label shift	45.60 / 86.27 / 80.57	41.77 / 83.75 / 76.05	7.29 / 60.98 / 58.94
Parkinson's Disease (5×1D-Conv)	mPower	MotionSense	Open-set recognition	100.00 / 99.86 / 99.89	100.00 / 99.94 / 99.60	0.00 / 58.71 / 64.96
		mHealth	Open-set recognition	100.00 / 100.00 / 100.00	100.0 / 99.99 / 99.99	0.00 / 41.41 / 59.44
		Kaggle Parkinson's	Covariate/label shift	98.00 / 99.89 / 99.47	98.96 / 99.96 / 99.67	70.00 / 95.91 / 93.34

Table 2: Out-of-distribution detection performance across different networks and datasets.

4.3.3 Results

Table 2 shows out-of-distribution detection performance for different methods across different health machine learning models and datasets. Overall, Mahalanobis distance- and Gram matrices-based out-of-distribution detection methods consistently show outstanding performance across different neural networks and different out-of-distribution datasets, showing TNR @ TPR95 of 95% or above. These methods show lower performance in distinguishing near-distribution datasets (e.g., ISIC 2017, Stethoscope, Kaggle Parkinson's), which aligns with the results from previous out-of-distribution work [64]. On the other hand, the energy-based method did not show reasonable performance in detecting out-of-distribution samples. We found that the energy scores of out-of-distribution samples were not able to effectively discriminated from in-distribution samples as shown in Appendix C. Note that we used the energy scores without fine-tuning; however, the authors of energy score-based method [43] have demonstrated that a classifier that is fine-tuned using the energy score in place of the softmax score shows significant improvement in out-of-distribution detection performance. This evaluation implies that state-of-the-art out-of-distribution detectors can be applied to health machine learning applications to provide reliable diagnostic results to the users.

5 User Study

According to the trustworthy AI literature [20], providing users with interpretable information can enhance the trustworthiness of the result and potentially impact users' decisions. We therefore conducted an online survey-based user study to validate this effect and the impact that our approach has on model trustworthiness. We first defined CONFIDENCE SCORE as how confident the model is in interpreting an input. In other words, in-distribution input would have a high CONFIDENCE SCORE, whereas out-of-distribution input would have a low CONFIDENCE SCORE. We compute CONFIDENCE SCORE by scaling raw out-of-distribution scores from out-of-distribution detectors [40, 64] to 0–100, where 0 is most likely to be an out-of-distribution example and 100 is most likely to be an in-distribution example. Scaling is done in a piecewise manner. When out-of-distribution scores are within an in-distribution threshold, which is set to include 95% of in-distribution examples, we compute min-max scaling that ranges from 90 to 100. In this way, we ensure that most of the in-distribution examples have confidence scores of 90 or above. When out-of-distribution scores outside of an in-distribution threshold, we compute min-max scaling from 0 to 90, where the same denominator is used as above since out-of-distribution examples might not be available in practice and any negative values are clipped to 0. We then investigate the effect of CONFIDENCE SCORE on user-perceived trustworthiness and its impact on medical decisions. Additionally, we also quantify potential learning that can be gained when it comes to distinguishing between in- and out-of-distribution input samples. Specifically, we aim to answer the following research questions:

- RQ. 1 How does the CONFIDENCE SCORE affect the perceived trustworthiness of diagnostic results?
- RQ. 2 How does the CONFIDENCE SCORE affect medical decisions based on diagnostic results?
- RQ. 3 Is there a potential learning effect from CONFIDENCE SCORE when it comes to distinguishing between input data with high and low CONFIDENCE SCORE?

5.1 Study Procedure and Participants

The overview of the online user study is illustrated in Figure 1 and a list of the user study interfaces is detailed in Appendix A. In short, the interface displays simulated results from the health screening models used in Section 4 (i.e., models for skin cancer, lung sound, and Parkinson's disease). We made

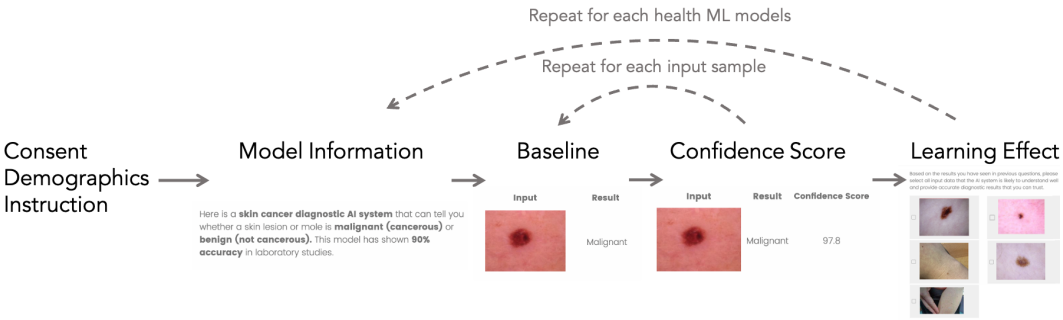


Figure 1: User study overview. The participants are first asked to give consent, read instruction, and provide demographics. Then, they report perceived trustworthiness and willingness to make a medical decision after seeing input samples that consist of different data types, diagnostic results, and CONFIDENCE SCORE for baseline and CONFIDENCE SCORE condition. Screenshots of the user study interface are demonstrated in Appendix A.

	User-perceived trustworthiness			**Impact on making medical decisions**		
	Wilcoxon Test (W)	p	Effect Size (r)	Wilcoxon Test (W)	p	Effect Size (r)
All	529,950.0	$< 0.001^{***}$	0.393	223,227.0	$< 0.001^{***}$	0.178
In-Distribution	138,778.5	$< 0.001^{***}$	0.475	52,056.0	$< 0.001^{***}$	0.200
Out-of-distribution	126,814.0	$< 0.001^{***}$	0.317	59,790.5	0.001^{***}	0.158
Negative result	131,910.0	$< 0.001^{***}$	0.393	51,197.5	0.026^{*}	0.100
Positive result	133,301.0	$< 0.001^{***}$	0.394	60,751.5	$< 0.001^{***}$	0.258
Image	55,890.0	$< 0.001^{***}$	0.436	20,884.0	$< 0.001^{***}$	0.225
Audio	64,440.0	$< 0.001^{***}$	0.384	25,848.0	0.002^{**}	0.173
Motion	56,767.5	$< 0.001^{***}$	0.361	28,084.5	0.019^{*}	0.133

Table 3: Results of Wilcoxon test in comparing baseline and CONFIDENCE SCORE conditions for the perceived trustworthiness and impact on decision making. All comparisons show statistically significant results. (***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$).

the input data as human-readable as possible to maximize interpretability. Images were shown as is, while audio was included in an audio player so that participants could play, pause, and stop the track. We presented the motion data as a time-series plot of accelerometer signals from x-, y-, and z-axis. Since time-series data can be particularly challenging for non-experts, we explain that high-amplitude signals are associated with fast motion while while low-amplitude signals are associated with slow motion. The interface explained the model's purpose and accuracy, which was fixed to 90% to remove potential bias. For each model, the interface presents prediction results in two different conditions: (baseline) input and result, and (confidence score) input, result, and CONFIDENCE SCORE. For each result, we asked participants how much they trust the model's prediction and whether they would be willing to make a medical decision based on that result. Participants saw a total of 24 scenarios (3 data types (image, audio, motion) $\times$ 2 conditions (baseline vs. CONFIDENCE SCORE $\times$ 2 CONFIDENCE SCORE (high vs. low) $\times$ 2 results (positive vs. negative). To provide realistic experience, we provide different skin tone images for skin lesion samples based on the reported skin tone. With the exception of the data type, the scenarios were shuffled across all other factors to avoid any ordering effects. After participants saw all of the scenarios for a given data type, we presented them with five data examples and asked them to pick the ones that the model would be confident in processing according to CONFIDENCE SCORE. We added these questions to assess whether people were able to learn about how the CONFIDENCE SCORE was being generated after seeing a series of examples.

We intend to target ordinary, non-expert consumers at random rather than expert clinicians for our study. Our research is primarily directed toward the boom in consumer-facing mobile health applications, where non-expert users are expected to collect input data themselves. We believe this is where models are most susceptible to out-of-distribution inputs, providing unreliable predictions to the users. For the safe use of AI-powered health applications, the users would need support via automated uncertainty measures. To this end, We recruited participants from Amazon Mechanical Turk and compensated with $3 USD for a 15-min online study. In total, 192 participants (155 male, 67 female) completed the online study with an average age of 42.7 $\pm$ 9.1 years. The study was approved by Institutional Review Boards at the University of Washington.

5.2 Results

We analyzed the responses using the Wilcoxon signed-rank test [75] to compute a pairwise comparison of the categorical responses between the baseline and CONFIDENCE SCORE conditions. Table 3 summarizes these statistical results along with the Rosenthal correlation coefficient [62] (r) for effect size.

RQ. 1: User-Perceived Trustworthiness In general, we found that user-perceived trustworthiness ($p < 0.001$) was higher in the CONFIDENCE SCORE condition with medium effect size ($r = 0.393$). In other words, the dataset shift information helped the participants decide when to trust or not to trust the output of the models. Higher CONFIDENCE SCORES led to increasing trustworthiness; high scores had a large effect size ($r = 0.475$), while low scores had a medium effect size ($r = 0.317$). The impact on trustworthiness was similar for positive and negative diagnostic results. The effect sizes varied for the different input data types, with the images having the largest effect size and motion having the smallest. We suspect that the effect size was correlated with the intuitiveness of the data types, with images being more intuitive than motion data.

RQ. 2: Impact on Making Medical Decisions When we examined the impact of CONFIDENCE SCORE on making medical decisions, we found that there was a statistically significant difference ($p < 0.001$) between the baseline and CONFIDENCE SCORE conditions. In other words, participants were more willing to make medical decisions when positive results were presented with high CONFIDENCE SCORE and vice versa. Similar to the results for user-perceived trustworthiness, the effect of CONFIDENCE SCORE was larger on input data with high scores than low scores and highest for images compared to audio and motion data.

RQ. 3: Learning Effect on Distinguishing In- and Out-of-Distribution Input Data We found that the participants were able to learn from their interaction with CONFIDENCE SCORE. The average Jaccard index when it came to selecting high CONFIDENCE SCORE input data was 0.75, 0.66, and 0.64 for image, audio, and motion data, respectively, which is a moderately high similarity. As with our other results, the Jaccard index was highest on images and lowest on the motion data, implying that ability to understand the input data also has impact on learning effect. This implies that the dataset shift information can make users better understand input data that the machine learning models can interpret for the future interaction.

6 Discussion

Dataset Shift Information for Health Application Users Based on the results from our performance evaluation and user study, we can imagine two potential use cases of the dataset shift information to improve safety and trust in mHealth applications. First, mHealth applications with machine learning models can exclude out-of-distribution samples to avoid making inferences and suggestions that are likely to be inaccurate and unreasonable. Second, our user study shows that the dataset shift information can enhance their interaction with the health machine learning applications. It was found to be particularly effective in improving trustworthiness for in-distribution data and leading the users to make the right medical decision. As the users interact with the health applications longer, they would have better understanding of importance of data quality for future interactions.

Dataset Shift Information for Health Application Developers Our dataset shift information not only improves the user experience, but also yields potential benefits for model developers. If a user correctly captures data but the model rejects it as being out-of-distribution, then there likely exists intrinsic problems or biases with the model. For example, if a skin lesion classifier is only trained on data from people with pale skin and a user with darker skin submits an image of their own, the out-of-distribution detector be triggered due to the incompleteness of the training dataset. The same issues may occur when training dataset is only collected from a specific set of sensors (e.g., camera, microphone, IMU) with particular specifications.

Limitations and Future Work Detecting near-distribution samples (e.g., ISIC 2017, Stethoscope, Kaggle Parkinson's) was a difficult problem for all the out-of-distribution detectors we evaluated. For the near-distribution datasets, we evaluated model's accuracy on the data that are distinguished as in-distribution. This issue is actively investigated by researchers and the improved near-distribution detection method would benefit this work.

Our user study was limited in the fact that it dealt with hypothetical scenarios. There were no repercussions for users decisions, so they may not have spent as much time making their decisions as

they would in real life. There are also many other factors that impact people's health-related decision making, such as the perceived severity of the medical condition and the perceived benefits of taking action [33, 67]. We tried to make some of the data more realistic by aligning data with the user's demographic information (e.g., we displayed skin lesion images based on their reported skin tone); nevertheless, participants were aware that the data was not their own. Additionally, increased trust might be affected by participants' own understanding and interpretation of the input data. Although we observed increased and decreased trust in examples with high and low CONFIDENCE SCORE, respectively, randomizing the CONFIDENCE SCORE of input data could further quantify impact of CONFIDENCE SCORE on user trust of health predictions.

In future work, we would like to (1) investigate the best way to present this information for the users, (2) leverage the dataset shift information for finding potential biases in the train dataset and inherent problems with the model, (3) investigate an out-of-distribution method for better near-distribution detection performance.

7 Conclusion

In this work, we investigated the utility of dataset shift information for improving reliability and trustworthiness of machine learning-based health applications. Using publicly available health deep learning models and datasets, we first demonstrated that the models fail when encountered with unseen data. We then evaluated the out-of-distribution detection performance of state-of-the-art methods, showing high accuracy in distinguishing between in- and out-of-distribution datasets for different input data types (e.g., image, audio, motion data). We conducted an online user study to investigate the effect of dataset shift information on potential users. We found that the participants trusted prediction results with high CONFIDENCE SCORE and are more willing to make a right medical decision, while they considered prediction results with low CONFIDENCE SCORE less trustworthy and are less willing to make medical decision. This work shows that the dataset shift is a meaningful piece of information for building consumer-facing trustworthy AI applications for high-stake decision making.

8 Acknowledgments and Disclosure of Funding

We would like to thank Alex Mariakakis for insightful feedback and discussions. This work was supported by NSF CNS-1565252. Shwetak Patel is on partial leave at Google.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016.

[2] Julius Adebayo, Justin Gilmer, Michael Christoph Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, pages 9505–9515, 2018.

[3] Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *36th International Conference on Machine Learning, ICML 2019*, pages 120–129, 2019.

[4] Saba Akbar, Enrico Coiera, and Farah Magrabi. Safety concerns with consumer-facing mobile health applications and their consequences: a scoping review. *Journal of the American Medical Informatics Association*, 27(2):330–340, 2020.

[5] Diego Ardila, Atilla P Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine*, 25(6):954–961, 2019.

[6] Susan Athey. Beyond prediction: Using big data for policy problems. *Science*, 355(6324):483–485, 2017.

[7] Oresti Banos, Rafael Garcia, Juan A Holgado-Terriza, Miguel Damas, Hector Pomares, Ignacio Rojas, Alejandro Saez, and Claudia Villalonga. mhealthdroid: a novel framework for agile development of mobile health applications. In *International workshop on ambient assisted living*, pages 91–98. Springer, 2014.

[8] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Gauthier Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, et al. Uncertainty as a

form of transparency: Measuring, communicating, and using uncertainty. *arXiv preprint arXiv:2011.07586*, 2020.

[9] Brian M Bot, Christine Suver, Elias Chaibub Neto, Michael Kellen, Arno Klein, Christopher Bare, Megan Doerr, Abhishek Pratap, John Wilbanks, E Ray Dorsey, et al. The mpower study, parkinson disease mobile data collected using researchkit. *Scientific data*, 3(1):1–9, 2016.

[10] Erdi Çallı, Keelin Murphy, Ecem Sogancioglu, and Bram Van Ginneken. Frodo: Free rejection of out-of-distribution samples: application to chest x-ray analysis. *arXiv preprint arXiv:1907.01253*, 2019.

[11] Tianshi Cao, Chinwei Huang, David Yu-Tung Hui, and Joseph Paul Cohen. A benchmark of medical out of distribution detection. *arXiv preprint arXiv:2007.04250*, 2020.

[12] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

[13] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

[14] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342–1350, 2018.

[15] Lilian De Greef, Mayank Goel, Min Joon Seo, Eric C Larson, James W Stout, James A Taylor, and Shwetak N Patel. Bilicam: using mobile phones to monitor newborn jaundice. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 331–342, 2014.

[16] Lisa DeBruine and Benedict Jones. Face Research Lab London Set. 4 2021.

[17] Miro Dudík, William Chen, Solon Barocas, Mario Inchiosa, Nick Lewins, Miruna Oprescu, Joy Qiao, Mehrnoosh Sameki, Mario Schlener, Jason Tuo, and Hanna Wallach. Assessing and mitigating unfairness in credit models with the fairlearn toolkit. 2020.

[18] Ivan Evtimov, Weidong Cui, Ece Kamar, Emre Kiciman, Tadayoshi Kohno, and Jerry Li. Security and machine learning in the real world. *arXiv preprint arXiv:2007.07205*, 2020.

[19] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018.

[20] Heike Felzmann, Eduard Fosch Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns:. *Big Data & Society*, 6(1):1–14, 2019.

[21] Nic Ford, Justin Gilmer, Nicolas Carlini, and Dogus Cubuk. Adversarial examples are a natural consequence of test error in noise. *arXiv preprint arXiv:1901.10513*, 2019.

[22] Mohammad Fraiwan, Luay Fraiwan, Basheer Khassawneh, and Ali Ibnian. A dataset of lung sounds recorded from the chest wall using an electronic stethoscope. *Data in Brief*, 35:106913, 2021.

[23] Siddhartha Gairola, Francis Tom, Nipun Kwatra, and Mohit Jain. Respirenet: A deep neural network for accurately detecting abnormal lung sounds in limited data setting, 2020.

[24] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[25] Giorgia. Simulation of parkinson movement disorders – kaggle, May 2018.

[26] George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim JP Hubbard, Richard JB Dobson, and Rina Dutta. Characterisation of mental health conditions in social media using informed deep learning. *Scientific reports*, 7(1):1–11, 2017.

[27] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[28] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

[29] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR (Poster)*, 2016.

[30] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

[31] Jason S Hoffman, Varun Viswanath, Xinyi Ding, Matthew J Thompson, Eric C Larson, Shwetak N Patel, and Edward Wang. Smartphone camera oximetry in an induced hypoxemia study. *arXiv preprint arXiv:2104.00038*, 2021.

[32] Sinh Huynh, Rajesh Krishna Balan, JeongGil Ko, and Youngki Lee. Vitamon: Measuring heart rate variability using smartphone front camera. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*, pages 1–14, 2019.

[33] Nancy K Janz and Marshall H Becker. The health belief model: A decade later. *Health education quarterly*, 11(1):1–47, 1984.

[34] Heinrich Jiang, Been Kim, Melody Y. Guan, and Maya R. Gupta. To trust or not to trust a classifier. In *Advances in Neural Information Processing Systems*, volume 31, pages 5541–5552, 2018.

[35] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[36] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[37] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[38] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, pages 6402–6413, 2017.

[39] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017.

[40] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[41] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

[42] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR, 2018.

[43] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21464–21475. Curran Associates, Inc., 2020.

[44] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *arXiv preprint arXiv:2006.03790*, 2020.

[45] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NIPS'17 Proceedings of the 31st International Conference on Neural Information Processing Systems*, volume 30, pages 4768–4777, 2017.

[46] Mohammad Malekzadeh, Richard G. Clegg, Andrea Cavallaro, and Hamed Haddadi. Mobile sensor data anonymization. In *Proceedings of the International Conference on Internet of Things Design and Implementation*, IoTDI '19, pages 49–58, New York, NY, USA, 2019. ACM.

[47] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *arXiv preprint arXiv:1802.10501*, 2018.

[48] Alex Mariakakis, Megan A Banks, Lauren Phillipi, Lei Yu, James Taylor, and Shwetak N Patel. Biliscreen: smartphone-based scleral jaundice monitoring for liver and pancreatic disorders. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(2):1–26, 2017.

[49] Alex Mariakakis, Jacob Baudin, Eric Whitmire, Vardhman Mehta, Megan A Banks, Anthony Law, Lynn Mcgrath, and Shwetak N Patel. Pupilscreen: using smartphones to assess traumatic brain injury. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–27, 2017.

[50] Alex Mariakakis, Edward Wang, Shwetak Patel, and Mayank Goel. Challenges in realizing smartphone-based health sensing. *IEEE Pervasive Computing*, 18(2):76–84, apr 2019.

[51] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 691–706, 2019.

[52] Sina Mohseni, Mandar Pitale, JBS Yadawa, and Zhangyang Wang. Self-supervised learning for generalizable out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5216–5223, 2020.

[53] Warren Morningstar, Cusuh Ham, Andrew Gallagher, Balaji Lakshminarayanan, Alex Alemi, and Joshua Dillon. Density of states estimation for out of distribution detection. In *International Conference on Artificial Intelligence and Statistics*, pages 3232–3240. PMLR, 2021.

[54] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using a test for typicality. *arXiv preprint arXiv:1906.02994*, 5:5, 2019.

[55] Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 2020.

[56] Andre GC Pacheco, Chandramouli S Sastry, Thomas Trappenberg, Sageev Oore, and Renato A Krohling. On out-of-distribution detection algorithms with deep neural skin cancer classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 732–733, 2020.

[57] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate. In *International Conference on Learning Representations*, 2018.

[58] Igor M. Quintanilha, Roberto de M. E. Filho, José Lezama, Mauricio Delbracio, and Leonardo O. Nunes. Detecting out-of-distribution samples using low-order deep features statistics. 2018.

[59] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring: evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 469–481, 2020.

[60] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A DePristo, Joshua V Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *arXiv preprint arXiv:1906.02845*, 2019.

[61] Bruno M Rocha, Dimitris Filos, Luís Mendes, Gorkem Serbes, Sezer Ulukaya, Yasemin P Kahya, Nikša Jakovljevic, Tatjana L Turukalo, Ioannis M Vogiatzis, Eleni Perantoni, et al. An open access database for the evaluation of respiratory sound classification algorithms. *Physiological measurement*, 40(3):035001, 2019.

[62] R. Rosenthal, H. Rosenthal, and inc Sage Publications. *Meta-Analytic Procedures for Social Research*. Applied Social Research Methods. SAGE Publications, 1991.

[63] Abhijit Guha Roy, Jie Ren, Shekoofeh Azizi, Aaron Loh, Vivek Natarajan, Basil Mustafa, Nick Pawlowski, Jan Freyberg, Yuan Liu, Zach Beaver, et al. Does your dermatology classifier know what it doesn't know? detecting the long-tail of unseen conditions. *arXiv preprint arXiv:2104.03829*, 2021.

[64] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with gram matrices. In *ICML 2020: 37th International Conference on Machine Learning*, volume 1, pages 8491–8501, 2020.

[65] Patrick Schoettker, Jean Degott, Gregory Hofmann, Martin Proença, Guillaume Bonnier, Alia Lemkaddem, Mathieu Lemay, Raoul Schorer, Urvan Christen, Jean-François Knebel, et al. Blood pressure measurements with the optibp smartphone app validated against reference auscultatory measurements. *Scientific Reports*, 10(1):1–12, 2020.

[66] Solveig K Sieberts, Jennifer Schaff, Marlena Duda, Bálint Ármin Pataki, Ming Sun, Phil Snyder, Jean-Francois Daneault, Federico Parisi, Gianluca Costante, Udi Rubin, et al. Crowdsourcing digital health measures to predict parkinson's disease severity: the parkinson's disease digital biomarker dream challenge. *NPJ digital medicine*, 4(1):1–12, 2021.

[67] Victor J Strecher and Irwin M Rosenstock. The health belief model. *Cambridge handbook of psychology, health and medicine*, 113:117, 1997.

[68] Chang Su, Zhenxing Xu, Jyotishman Pathak, and Fei Wang. Deep learning in mental health outcome research: a scoping review. *Translational Psychiatry*, 10(1):1–26, 2020.

[69] Adarsh Subbaswamy and Suchi Saria. From development to deployment: dataset shift, causality, and shift-stable models in health ai. *Biostatistics*, 21(2):345–352, 2020.

[70] Jayaraman J Thiagarajan, Prasanna Sattigeri, Deepta Rajan, and Bindya Venkatesh. Calibrating healthcare ai: Towards reliable and interpretable deep predictive models. *arXiv preprint arXiv:2004.14480*, 2020.

[71] Florian Tramer and Dan Boneh. Differentially private learning needs better features (or much more data). In *ICLR 2021: The Ninth International Conference on Learning Representations*, 2021.

[72] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel. Ensemble adversarial training: Attacks and defenses. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

[73] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

[74] Abinav Ravi Venkatakrishnan, Seong Tae Kim, Rami Eisawy, Franz Pfister, and Nassir Navab. Self-supervised out-of-distribution detection in brain ct scans. *arXiv preprint arXiv:2011.05428*, 2020.

[75] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992.

[76] Hanrui Zhang, Kaiwen Deng, Hongyang Li, Roger L Albin, and Yuanfang Guan. Deep learning identifies digital biomarkers for self-reported parkinson's disease. *Patterns*, 1(3):100042, 2020.

[77] Hongjie Zhang, Ang Li, Jie Guo, and Yanwen Guo. Hybrid models for open set recognition. In *European Conference on Computer Vision*, pages 102–117. Springer, 2020.

[78] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pages 819–827. PMLR, 2013.

A User Study Interface

In this section, we provide screenshots and list of examples that were used in the user study.

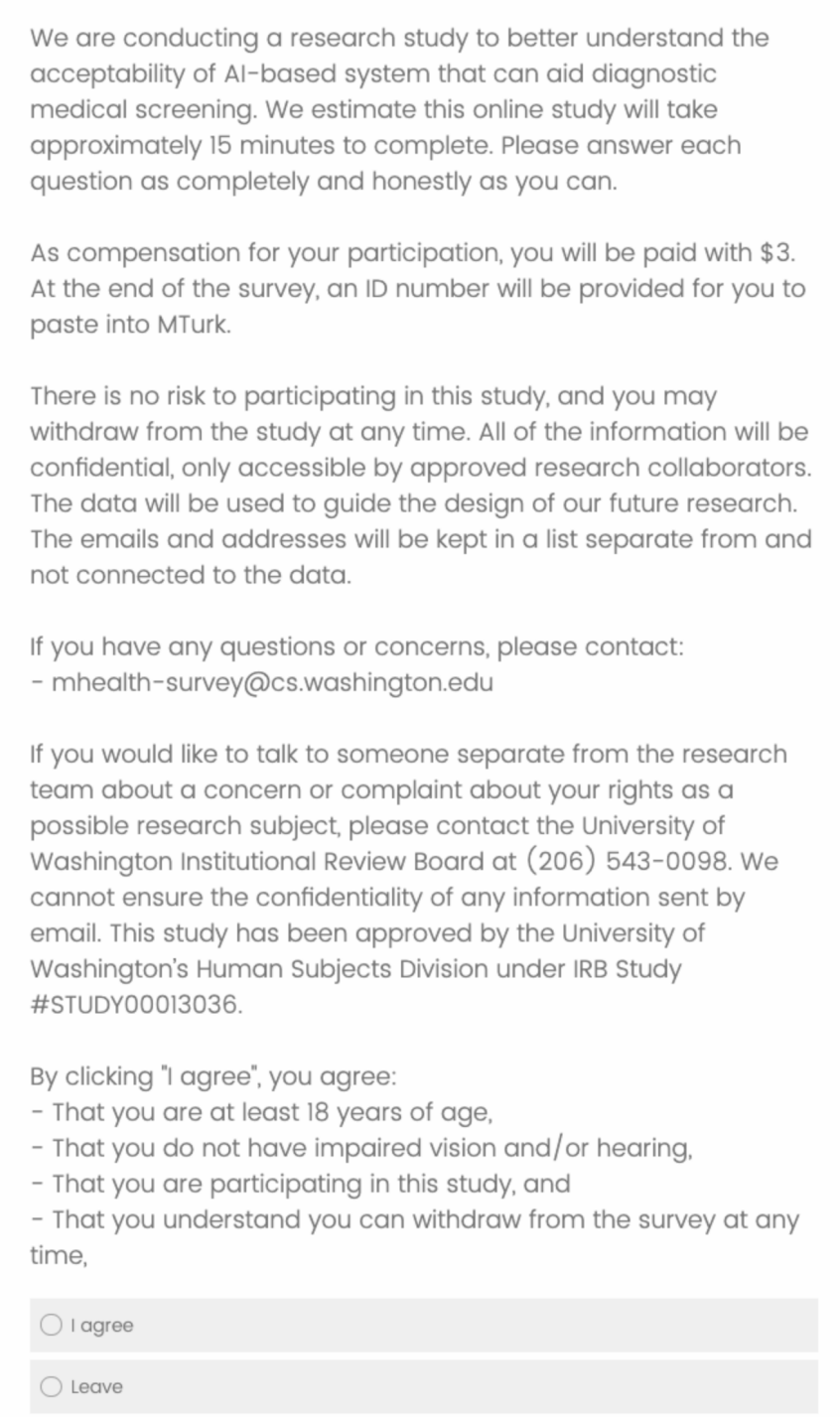
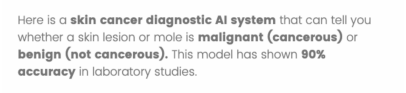
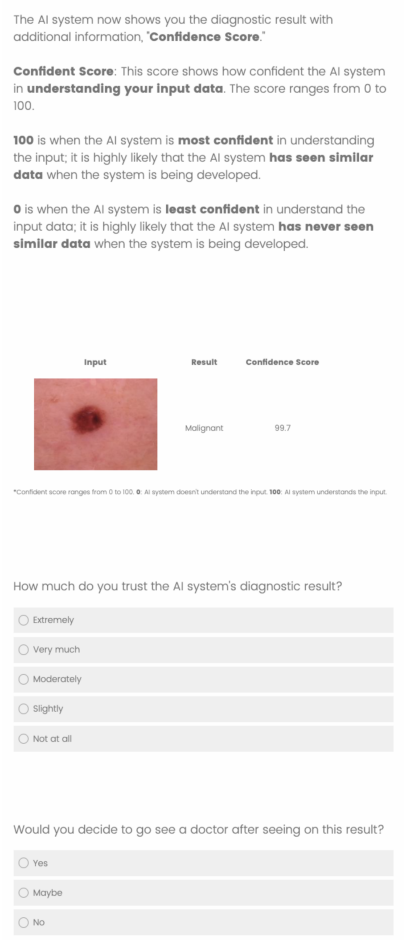


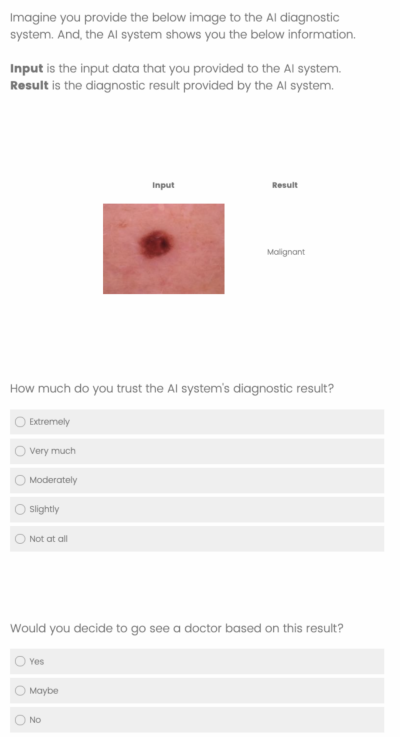
Figure 2: User study consent form. Note that the name of the institution is redacted for the review.



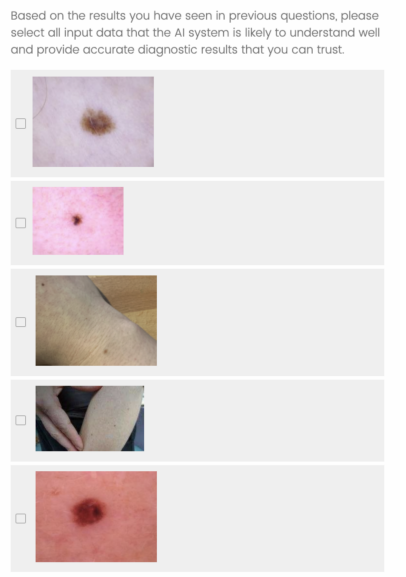
(a) Interface that shows information about a health machine learning model. It shows target health condition, possible prediction results, and its accuracy.



(c) Interface that shows CONFIDENCE SCORE condition. This condition only presents input data, prediction results, and CONFIDENCE SCORE.

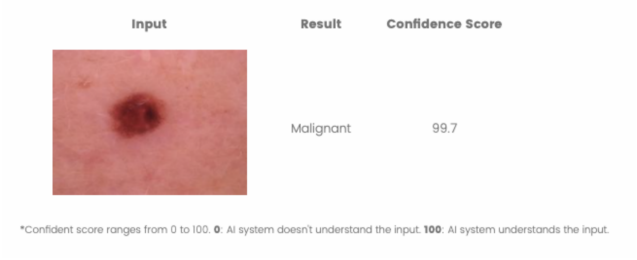


(b) Interface that shows baseline condition. This condition only presents input data and prediction results and asks questions on user-perceived trustworthiness and impact on making medical decisions.

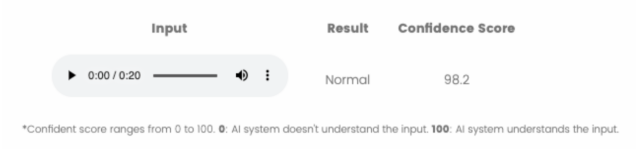


(d) Interface that asks users to select input data that would have high CONFIDENCE SCORE to explore potential learning effect through their interaction with CONFIDENCE SCORE.

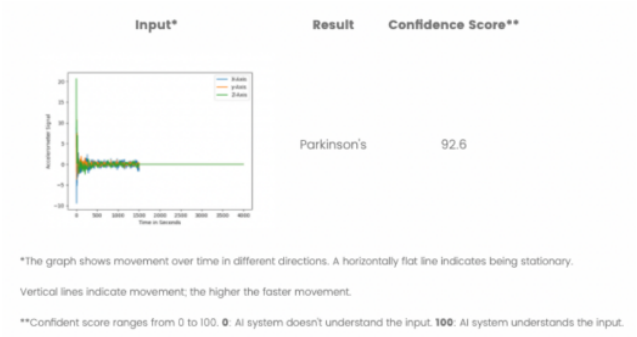
Figure 3: List of user study interface. This shows an example interface for skin cancer classifier.



(a) Image input is shown in a visible size.

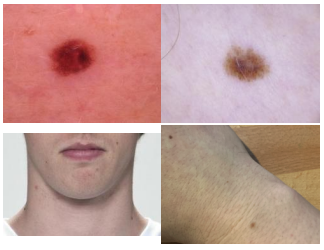


(b) Audio player is embedded for the participants to listen to the input data.

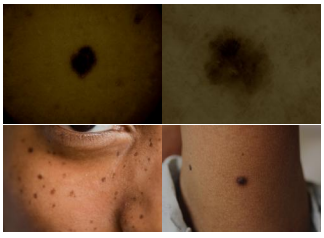


(c) Motion data is shown as a time-series plot of accelerometer signal. We provide additional explanation about how to interpret the signal.

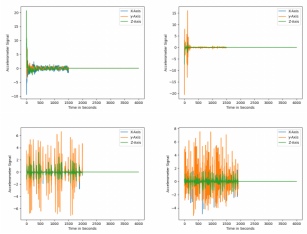
Figure 4: Interface to display different input data types.



(a) Input examples for skin cancer classifier for the participants who self-report to have light-colored skin tone.



(b) Input examples for skin cancer classifier for the participants who self-report to have dark-colored skin tone.



(c) Input examples for Parkinson's disease classifier.

Figure 5: List of input examples used in the user study. For each input type, top row shows in-distribution inputs and bottom row shows out-of-distribution inputs. Left column shows positive diagnostic results and right column shows negative diagnostic results. Note that audio samples are not included due to its difficulty to visualize.

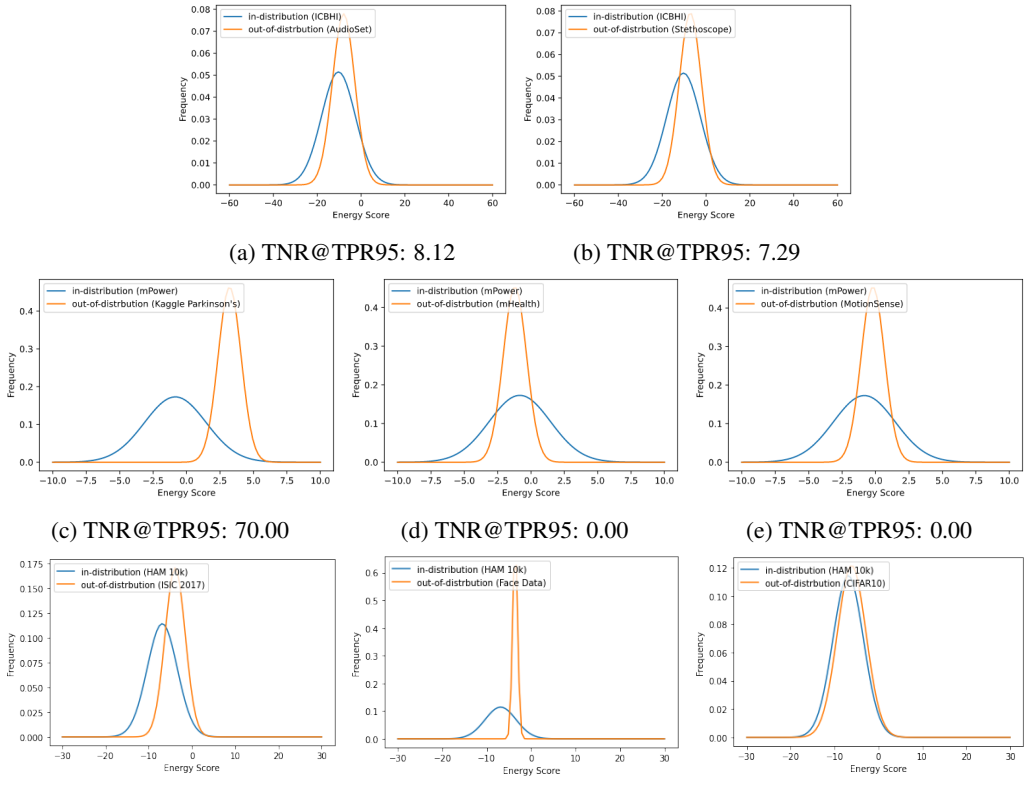
B Performance Metrics

In out-of-distribution performance evaluation in Section 4.3, we use the following metrics that has been used in previous out-of-distribution work [40, 64]:

- **True negative rate (TNR) at 95% true positive rate** (TPR) is defined as the percentage of correctly detected out-of-distribution samples, when 95% of in-distribution samples are correctly detected. TNR is calculated $TNR = TN/(FP + TN)$ and $TPR = TP/(TP + TN)$, where TP, TN, FP, and FN are true positive, true negative, false positive, and false negative, respectively.
- **Area under the receiver operating curve (AUROC)** is defined as the area under the plot of true positive rate (TPR) versue false positive rate (FPR), where $TPR = TP/(TP+FN)$ and $FPR = FP/(FP + TN)$.
- **Detection accuracy** is defined as the maximum classification accuracy over all possible thresholds in classifying in- and out-of-distribution data.

C Energy-Based OOD Detection Analysis

In Section 4.3, energy-based out-of-distribution detection method does not show comparable performance to methods using Mahalanobis distance and Gram matrices. We further analyze the method by comparing the distribution of energy score between in- and out-of-distribution as shown Figure 6. In most cases, the distribution of the energy scores are overlapped, making it difficulty to detect out-of-distribution samples using energy score. In this work, we use energy-based method without fine-tuning, which is suitable for adopting the method to any pre-trained models. However, as the authors have demonstrated in their paper [43], fine-tuned energy-based method that requires re-training of a classifier, shows significant improvement in detecting out-of-distribution samples.



(a) TNR@TPR95: 8.12 (b) TNR@TPR95: 7.29

(c) TNR@TPR95: 70.00 (d) TNR@TPR95: 0.00 (e) TNR@TPR95: 0.00

(f) TNR@TPR95: 14.28 (g) TNR@TPR95: 0.00 (h) TNR@TPR95: 5.06

Figure 6: Energy score distribution across different in- and out-of-distribution datasets.

C.1 Out-of-Distribution Performance with Confidence Interval

				Validation on OOD Samples (TNR @ TPR95/AUROC/Detection Accuracy)		
Health ML Models	**In-Distribution**	**Out-of-Distribution**	**Distribution Shift**	**Mahalanobis**	**Gram**	**Energy**
Skin Lesion (DenseNet-121)	HAM10000	ISIC 2017	Covariate/label shift	10.13 / 58.21 / 59.28 ±2.61/ ±3.30 / ±2.38	25.90 / 81.14 / 74.98 ±1.22 / ±1.89 / ±1.12	14.28 / 76.20 / 70.76 ±0.49 / ±0.18 / ±0.16
		Face	Open-set recognition	100.00 / 99.98 / 99.96 ±0.00 / ±0.02 / ±0.04	95.01 / 98.20 / 96.34 ±1.48 / ±0.41 / ±0.63	0.00 / 80.45 / 84.81 ±0.00 / ±0.14 / ±0.25
		CIFAR10	Open-set recognition	99.83 / 99.90 / 99.61 ±0.18 / ±0.10 / ±0.39	95.14 / 98.66 / 96.90 ±1.43 / ±1.37 / ±1.94	5.06 / 58.33 / 57.89 ±0.26 / ±0.92 / ±0.67
Lung Sound (ResNet-34)	ICBHI	AudioSet	Open-set recognition	97.96 / 99.47 / 97.34 ±0.73 / ±0.26 / ±0.45	96.55 / 99.18 / 95.97 ±1.67 / ±0.30 / ±0.62	8.12 / 56.79 / 57.13 ±0.24 / ±0.15 / ±0.14
		Stethoscope	Covariate/label shift	45.60 / 86.27 / 80.57 ±4.95 / ±1.42 / ±1.55	41.77 / 83.75 / 76.05 ±1.62 / ±0.63 / ±0.38	7.29 / 60.98 / 58.94 ±1.22 / ±0.74 / ±0.63
Parkinson's Disease (5×1D-Conv)	mPower	MotionSense	Open-set recognition	100.00 / 99.86 / 99.89 ±0.00 / ± 0.13 / ±0.10	100.00 / 99.94 / 99.60 ±0.00 / ±0.24 / ±0.14	0.00 / 58.71 / 64.96 ±0.00 / ±0.59 / ±0.32
		mHealth	Open-set recognition	100.00 / 100.00 / 100.00 ± 0.00 / ±0.00 / ±0.00	100.0 / 99.99 / 99.99 ±0.00 / ±0.02 / ±0.01	0.00 / 41.41 / 59.44 ±0.00 / ±1.09 / ±1.10
		Kaggle Parkinson's	Covariate/label shift	98.00 / 99.89 / 99.47 ±2.45 / ±0.14 / ±1.25	98.96 / 99.96 / 99.67 ±0.00 / ±0.02 / ±0.03	70.00 / 95.91 / 93.34 ±4.68 / ±0.30 / ±0.32

Table 4: Out-of-Distribution Detection Performance Across Multiple Tasks. Evaluation is repeated for 5 times. Mean and standard deviation of metrics are reported.