WILDS: A Survey and Benchmark of in-the-Wild Distribution Shifts

Anonymous Author(s)
Affiliation
Address
email

Abstract

Distribution shifts can cause significant failures in ML systems deployed in the wild. However, many widely-used datasets in the ML community today were designed for evaluating algorithms that make standard i.i.d. assumptions, and attempts to retrofit distribution shifts onto these datasets have generally not been representative of the kinds of shifts that arise in real-world applications. In this paper, we present WILDS, a benchmark of in-the-wild distribution shifts that builds on top of recent data-collection efforts by domain experts in a broad array of applications, from tumor identification to wildlife monitoring and poverty mapping. These datasets reflect distribution shifts arising from testing on different hospitals, camera locations, countries, time periods, demographics, etc., all of which cause substantial performance drops in the baseline models that we train. Finally, we discuss other datasets and shifts for which we did not see an appreciable performance drop, and survey other application areas where distribution shifts naturally arise. By unifying datasets from a variety of application areas and making them accessible to the ML community, we hope to encourage the development of general-purpose methods that are anchored to real-world distribution shifts and that can work well across different application areas and problem settings.

1 Introduction

Distribution shifts—when the training distribution does not match the test distribution—pose significant challenges for ML systems deployed in the wild, with many examples in prior work of state-of-the-art neural networks that achieve high accuracies on independent and identically distributed (i.i.d.) test sets but fail when the test distributions shift . Such shifts arise naturally in many real-world scenarios. In some settings, the training and test distributions might comprise data from related but distinct domains: e.g., model performance can be substantially worse on patients from different hospitals (Zech et al., 2018; Beede et al., 2020; DeGrave et al., 2020), wildlife photos taken at different camera trap locations (Beery et al., 2018), biological assays from different cell types (Li et al., 2019a), or satellite images from different countries (Jean et al., 2016). In other settings, we might be interested in a test distribution that is a subpopulation of the training distribution: e.g., models can fail to make accurate predictions on people from minority groups (Buolamwini & Gebru, 2018; Borkan et al., 2019; Koenecke et al., 2020), raising issues of equity and generalization.

However, many datasets that are widely used in the ML community today were designed for evaluating algorithms that make standard i.i.d. assumptions, and attempts to retrofit distribution shifts onto these datasets have generally not been representative of the kinds of shifts encountered in the wild. Instead, recent work tends to prioritize distribution shifts that are controllable and targeted, which are often easier to study and work with than their real-world counterparts. For example, a substantial amount of recent work on distribution shifts in the ML community has been on synthetic transformations for object recognition, such as changing the colors of MNIST digits (Arjovsky et al., 2019), corrupting

Submitted to the ML Retrospectives Workshop at NeurIPS 2020. Do not distribute.

Dataset	Data (x)	Prediction target (y)	Size	Domains	Domain count	Train/test domain overlap	Train domain labels	Test domain labels
CIVILCOMMENTS (Borkan et al., 2019)	online comments	toxicity	?	demographics	16	✓	partial	-
AMAZON (Ni et al., 2019)	product reviews	sentiment	1.4M	users	7,642	-	✓	✓
CAMELYON17 (Bandi et al., 2018)	tissue slides	tumor	?	hospitals	5	-	✓	✓
IWILDCAM2020 (Beery et al., 2020a)	camera trap photos	species	?	trap locations	291	-	✓	✓
POVERTYMAP (Yeh et al., 2020)	satellite images	asset wealth	?	countries	23	-	✓	✓
FMOW (Christie et al., 2018)	satellite images	land use	?	time continents	?	- ✓	✓	✓

Table 1: Summary of datasets in the WILDS benchmark.

images with noise (Hendrycks & Dietterich, 2019), or swapping out image backgrounds (Xiao et al., 2020). Datasets that do not rely on synthetic perturbations might instead split the data to induce a shift that is more extreme than what is likely to occur in the wild, e.g., learning to classify photos solely from stylized representations (Li et al., 2017a), or learning to classify objects at different scales solely from objects at a single scale (Hendrycks et al., 2020).

Datasets like the ones above play an important role in developing methods for handling distribution shifts, as they tend to contain shifts that are controllable, targeted, and convenient to work with. However, it is also crucial to ensure that these methods can translate to real-world settings. Datasets from earlier work in the ML community on real-world distribution shifts are not as widely used today as they tend to be much smaller than modern datasets, e.g., in object recognition (Saenko et al., 2010; Gong et al., 2012), sentiment analysis (Blitzer et al., 2007; Pan et al., 2010), part-of-speech tagging (Marcus et al., 1993; Daumé III, 2007), or land cover classification (Bruzzone & Marconcini, 2009). On the other hand, domain experts applying ML in their specific areas are often intimately familiar with the distribution shifts that arise in their applications, e.g., in medicine (Chen et al., 2020), computational biology (Leek et al., 2010), wildlife conservation (Beery et al., 2018), satellite imagery (Jean et al., 2016), and so on. However, these applications and their corresponding datasets can be less accessible and convenient for ML researchers who are not domain experts, and existing methods for mitigating these shifts can be highly domain-specific.

In this paper, we present the WILDS benchmark, a curated collection of datasets that we believe are representative of the kinds of distribution shift challenges that ML algorithms face in the wild (Table 1). WILDS builds on top of extensive data-collection efforts by domain experts in a broad array of applications with natural distribution shifts: predicting text toxicity (Borkan et al., 2019), sentiment analysis (Ni et al., 2019), poverty mapping (Yeh et al., 2020), building and land use classification (Christie et al., 2018), animal species categorization Beery et al. (2020a), and tumor identification (Bandi et al., 2018). By unifying datasets from a variety of application areas and making them accessible to the ML community through careful preprocessing and standardized benchmarking, we hope to encourage the development of *general-purpose* methods that are anchored to real-world distribution shifts and that can work well across different application areas and problem settings.

To design WILDS, we consulted domain experts and researchers working on distribution shifts in different application areas, selecting and adapting datasets to fulfill the following criteria:

- *Distribution shifts with corresponding performance drops*. The train/test splits reflect distribution shifts, with significant gaps between the in-distribution versus out-of-distribution performances of models trained on them.
- *Real-world relevance*. These distribution shifts arise naturally in ML applications, with training/test splits and evaluation metrics that are motivated by real-world scenarios.
- *Potential leverage*. It is unrealistic to expect models to generalize well to arbitrary distribution shifts. For each dataset, we articulate why we might still expect methods to be able to do well on the specific distribution shift in the dataset. For example, each training set

contains data from multiple domains, which in theory could allow a model to learn not to rely on domain-specific features.

At present, there are 6 datasets in WILDS (Table 1), reflecting distribution shifts arising from different hospitals, cell types, users, demographics, camera locations, countries, and time periods. Formally, we cast these as examples of *domain shifts*, wherein we assume that each point belongs to a domain z (corresponding to hospitals, cell types, etc.), each domain z has an associated distribution P_z over data points, and the training and test distributions comprise a (different) mixture of domains. Each dataset is packaged in a consistent and accessible manner, with pre-processed versions of the data together with standard models and evaluation code that replicate the baseline results presented in this paper. For each dataset, we also discuss its broader context and relation to other tasks and distribution shifts in the same application area.

We aim for WILDS to be extensible and community-driven. To this end, we have also developed a template for describing and evaluating a distribution shift dataset that researchers can use to contribute datasets in their area of work that fit the criteria above. In Section 5, we survey potential application areas—algorithmic fairness, medicine and healthcare, natural language and speech processing, education, and robotics—that could be promising sources of appropriate datasets, as well as the challenges associated with each of them. Finally, in Section 6, we discuss datasets where distribution shifts did not seem to affect model performance, as well as promising avenues of future work in method development.

2 Comparison with existing benchmarks

A plethora of distribution shift tasks have been studied in the ML community, including domain adaptation (Ben-David et al., 2006; Daumé III, 2007), domain generalization (Blanchard et al., 2011; Muandet et al., 2013), and test-time adaptation (Sun et al., 2020; Zhang et al., 2020). These prior works have largely focused on distribution shifts that are induced by synthetic transformations, data splits, and dataset combinations, as such shifts tend to be more controllable, targeted, and accessible. Moreover, they predominantly consider object recognition tasks. The WILDS benchmark complements these prior works by focusing instead on naturally-occurring distribution shifts across a diverse set of applications, which we believe are currently underrepresented in the ML literature.

Distribution shifts from transformations. Some of the most widely-adopted benchmarks induce distribution shifts by synthetically transforming the data. Examples include rotated and translated versions of MNIST and CIFAR (Worrall et al., 2017; Gulrajani & Lopez-Paz, 2020); surface variations such as texture, color, and corruptions like blur in Colored MNIST (Gulrajani & Lopez-Paz, 2020), Stylized ImageNet (Geirhos et al., 2018), and ImageNet-C (Hendrycks & Dietterich, 2019); and datasets that crop out objects and replace their backgrounds, as in the Backgrounds Challenge (Xiao et al., 2020) and other similar datasets (Sagawa et al., 2020). Fully synthetic datasets such as SYNTHIA (Ros et al., 2016) have also been adopted for domain adaptation and generalization, e.g., by testing robustness to transformations in the seasons, weather, time, or architectural style (Hoffman et al., 2018; Volpi et al., 2018). Benchmarks for adversarial robustness also fall in this category (Goodfellow et al., 2015; Croce et al., 2020), though it is not a focus of this work.

Distribution shifts from constrained splits. Other benchmarks do not rely on transformations but instead collect and split the data in a way that induces particular distribution shifts, yielding distribution shift datasets that are more controlled and targeted but not reflective of the real-world challenges. For example, BREEDS (Santurkar et al., 2020) tests generalization to unseen subclasses by holding out subclasses as specified by several controllable parameters; DeepFashion-Remixed (Hendrycks et al., 2020) constrains the training set to include only photos from a single camera viewpoint and tests generalization to unseen camera viewpoints; and ObjectNet (Barbu et al., 2019) comprises images taken from a few pre-specified viewpoints, allowing for systematic evaluation for robustness to camera angle changes but deviating from natural camera angles.

Distribution shifts across datasets. Finally, standard benchmarks for domain adaptation and domain generalization sometimes combine several disparate datasets (Torralba & Efros, 2011). However, many of these well-studied distribution shifts are more drastic than might naturally arise. For example, standard domain adaptation benchmarks include transfers across digit classification datasets such as MNIST and SVHN (LeCun et al., 1998; Yuval et al., 2011; Tzeng et al., 2017; Hoffman et al., 2018), transfers across different renditions (e.g., photos, clipart, sketches) in DomainNet (Peng et al.,

2019), as well as transfers from synthetic to real data (Ganin & Lempitsky, 2015; Richter et al., 2016; Peng et al., 2018), the last of which is not a focus of our work. Standard domain generalization benchmarks, many of which are in DomainBed (Gulrajani & Lopez-Paz, 2020), include generalization across different renditions in PACS (Li et al., 2017a), DomainNet (Peng et al., 2019), Office-Home (Venkateswara et al., 2017), and ImageNet-R (Hendrycks et al., 2020) as well as generalization across different object recognition datasets in VLCS (Fang et al., 2013). While good test beds, these cross-dataset transfers are not anchored to a real-world challenge and it is unclear if there is sufficient leverage for successful adaptation.

Motivation for WILDS. While the above benchmarks are useful test beds, it is also important to assess robustness to distribution shifts in the wild to ensure methodological progress applicable for real-world challenges. In fact, some prior work suggests that robustness does not necessarily transfer from one shift to another, underscoring the importance of evaluating directly on real-world shifts; for example, robustness to synthetic shifts such as corruption and textural changes fails to transfer to robustness to shifts across datasets (Taori et al., 2020), and a method can improve robustness on a standard vision dataset, but consistently harm robustness on real-world satellite imagery datasets (Xie et al., 2020). To this end, several datasets with realistic shifts have been proposed recently (Atwood et al., 2020; Shankar et al., 2019), but they still largely focus on object recognition tasks, similarly to the vast majority of the aforementioned benchmarks. We seek to complement the above prior work by presenting a standardized suite of datasets with distribution shifts in the wild spanning diverse applications.

3 Overview

We provide some context for the discussion on datasets in Section 4. We describe our approach for dataset selection, formalize domain shifts and problem settings, and describe baseline algorithms.

3.1 Approach

The WILDS benchmark is currently a collection of six datasets (Table 1), assembled according to the following criteria.

Real-world relevance and diversity in applications. We curate and modify existing datasets from various application areas, from natural language processing to pathology to satellite imagery. In our curation, we adopt a bottom-up approach, identifying distribution shifts that are real-world hurdles in various applications. To find such shifts and datasets, we consult with domain experts and survey these application areas, which has a large body of prior work on real-world distribution shifts and corresponding datasets.

Potential leverage. Rather than simply providing a training set and a shifted test set, we provide additional data and information that can be helpful for achieving distributional robustness, to the extent that it is realistic for a given application. Examples include multi-domain training data, annotations on domains and their subdomains, metadata, and unlabeled data at test time. Furthermore, we describe why these additional resources can reasonably provide sufficient leverage for achieving distributional robustness for each dataset.

Performance drops. To demonstrate that that the presented datasets are valid test beds for out-of-distribution generalization, we report the performance of various baseline models. First, we show that there are substantial performance drops due to distribution shifts, comparing the in-distribution and out-of-distribution performance of standard models. In addition, we show that existing methods for improving robustness to distribution shifts do not close the above gap, demosntrating room for substantial improvement. We discuss the above baselines in more detail in Section 3.3.

3.2 Problem settings

In the WILDS benchmark, we consider several problem settings on *domain shifts*, a broad class of distribution shifts in which the distribution over domains changes during train and test time. We now

formally define domain shifts and introduce specific problem settings: subpopulation shifts, domain generalization, and test-time adaptation.

Domain shifts. Each point (x, y, z) is a tuple of input $x \in \mathcal{X}$, target $y \in \mathcal{Y}$, and domain $z \in \mathcal{Z}$, and each domain z corresponds to a fixed data distribution P_z over (x, y, z) given z. The training data distribution is composed of a mixture of domains in $\mathcal{Z}^{\text{train}} \subseteq \mathcal{Z}$ with domain weights q^{train},

$$P^{\text{train}} = \sum_{z \in \mathcal{Z}^{\text{train}}} q_z^{\text{train}} P_z. \tag{1}$$

At test time, we have a different mixture of domains with weights q^{test} and composed of a potentially different subset of domains $\mathcal{Z}^{\text{test}} \subseteq \mathcal{Z}$,

$$P^{\text{test}} = \sum_{z \in \mathcal{Z}^{\text{test}}} q_z^{\text{test}} P_z. \tag{2}$$

Overview of problem settings. Within the general framework of domain shifts, problem settings can differ along the following axes of variation:

1. *Train/test domain overlap.* Train and test domains can have different degrees of overlap, from fully overlapping ($\mathcal{Z}^{\text{train}} = \mathcal{Z}^{\text{test}}$) to fully disjoint ($\mathcal{Z}^{\text{train}} \cap \mathcal{Z}^{\text{test}} = \emptyset$). Robustness in the former problem setting requires good performance on subsets of the training distribution, whereas the latter requires generalization to unseen domains.
2. *Train-time domain annotations.* The domain identity z may be observed for none, some, or all of the training examples. Train-time domain annotations allow training algorithms to take the domain information into account.
3. *Test-time domain annotations.* The domain identity z may be observed for none, some, or all of the test examples. Test-time domain annotations allow models to be domain-specific, for example by treating domain identity as a feature or by adapting to each domain.
4. *Test-time unlabeled data*. Varying amounts of test-time unlabeled data—samples of x drawn from the test distribution P^{test}—may be available, from none to a small batch to a large pool. This affects the degree to which models can adapt to test distributions.

Each combination of the above four factors corresponds to a problem setting, each with a specific set of applicable methodologies. In particular, we focus on three problem settings in the current version of the WILDS benchmark (Table 2): (i) *subpopulation shift*, in which proportions of known domains shift between train and test time, (ii) *domain generalization*, in which we seek to generalize to unseen and apriori unknown domains, and (iii) *test-time adaptation*, in which we seek to adapt to unseen domains by leveraging a small batch of unlabeled data.

Problem Setting	**Train/Test Domain Overlap**	**Train-Time Domain Annotations**	**Test-Time Domain Annotations**	**Test-Time Unlabeled Data**
Subpopulation shift	✓	✓	-	-
Domain generalization	-	✓	✓	-
Test-time adaptation	-	✓	✓	batch

Table 2: Problem settings.

3.3 Baselines

Demonstrating performance drops. To demonstrate performance drops upon distribution shifts, we compare the in-distribution and out-of-distribution performance of models trained via empirical risk minimization (ERM), which minimizes the average training loss,

$$\mathcal{R}_{\text{ERM}}(\theta) := \hat{\mathbb{E}}_{P^{\text{train}}} \left[\ell(\theta; (x, y)) \right]. \tag{3}$$

Subpopulation shift baselines. To train models robust to subpopulation shifts, a simple but a strong baseline is to train models using the reweighted objective that upweights minority domains (Shimodaira, 2000),

$$\mathcal{R}_{\mathsf{REW}}(\theta) := \hat{\mathbb{E}}_{P^{\mathsf{train}}}\left[\frac{1}{q_z^{\mathsf{train}}}\ell(\theta;(x,y))\right]. \tag{4}$$

The above reweighted objective is a heuristic for achieving low loss not only on common domains, but also for minority domains. Group DRO (Sagawa et al., 2020) explicitly minimizes the loss for the worst-case domain,

$$\mathcal{R}_{\mathsf{DRO}}(\theta) := \max_{z\in\mathcal{Z}^{\mathsf{train}}} \hat{\mathbb{E}}_{P_z}\left[\ell(\theta;(x,y))\right]. \tag{5}$$

While we evaluate the above representative baselines, additional methodologies for subpopulation shifts include more sophisticated reweighting methods (Cui et al., 2019), combining the above objectives with unsupervised clustering (Oren et al., 2019; Sohoni et al., 2020), adaptive Lipschitz regularization (Cao et al., 2020), slice-based learning (Chen et al., 2019b; Ré et al., 2019), and style transfer across domains (Goel et al., 2020).

Domain generalization baselines. Many methodologies have been proposed for domain generalization, including group DRO (Sagawa et al., 2020) , domain-invariant learning (Ganin et al., 2016; Sun & Saenko, 2016), invariant risk minimization (Gulrajani & Lopez-Paz, 2020), and meta-learning-based methods (Li et al., 2017b; Dou et al., 2019), Recently, Gulrajani & Lopez-Paz (2020) benchmarked many of the above methodologies on standard domain generalization datasets and found that they all perform comparably to and no better than ERM. We thus report the performance of models trainved via ERM and group DRO as representative baselines, even though we are currently working to test more baseline methodologies.

Test-time adaptation baselines. In the presence of distribution shift, one source of information that can be leveraged even without access to labels is to simply observe multiple test points, either sequentially or in a batch. A number of recent methods that leverage batches of unlabeled test points can be classified as test-time adaptation methods (Lipton et al., 2018; Li et al., 2017c; Sun et al., 2020; Wang et al., 2020a). Recently, Zhang et al. (2020) proposed adaptive risk minimization (ARM), a problem formulation that allows for meta-learning models that can better leverage these test-time adaptation procedures for improved performance. We report performance of models trained via test-time training (TTT) (Sun et al., 2020) and ARM as baselines.

4 Datasets

We now discuss the six datasets in the WILDS benchmark, summarized in Table 1. For each dataset, we describe the setup and the evaluation criteria, present baseline results, and discuss the real-world relevance and connection to other shifts in the literature. Because we modify the dataset from the original version in terms of the evaluation, splits, and data through substantial pre-processing, we use a -WILDS suffix to avoid confusion between our modified datasets and their original sources.

4.1 CIVILCOMMENTS-WILDS: Toxicity classification across demographic identities

Automatic review of user-generated text—e.g., detecting if an online comment is toxic—is an important tool for moderating the sheer volume of text being written every day on the Internet. Unfortunately, prior work has documented biases in automatic moderation tools; for example, toxicity classifiers have been shown to spuriously associate the mention of certain demographic groups with toxicity (Park et al., 2018; Dixon et al., 2018). We study this issue through the CIVILCOMMENTS-WILDS dataset (Borkan et al., 2019), a large collection of 450,000 comments each annotated for toxicity and demographic mentions by multiple crowdworkers. Dataset and model details are in Appendix A.1.

Task. Binary classification, where the input is a comment (comprising one or more sentences of text) on an online article, and the output is whether it is toxic or not, i.e., if it had been rated as toxic by a majority of the crowdworkers who saw that comment. Each comment was reviewed by at least 10 crowdworkers.

Toxic	Comment Text	Male	Female	LGBTQ	White	Black	…	Christian
0	I applaud your father. He was a good man! We need more like him.	1	0	0	0	0	…	0
0	As a Christian, I will not be patronizing any of those businesses.	0	0	0	0	0	…	1
0	What do Black and LGBT people have to do with bicycle licensing?	0	0	1	0	1	…	0
0	Government agencies track down foreign baddies and protect law-abiding white citizens. How many shows does that describe?	0	0	0	1	0	…	0
1	Maybe you should learn to write a coherent sentence so we can understand WTF your point is.	0	0	0	0	0	…	0

Figure 1: Example comments from CIVILCOMMENTS-WILDS.

Model	Average accuracy	Worst-group accuracy
Standard	**92.7**	54.3
Group DRO (class balance)	85.5	51.5
Group DRO (class $\times$ any identity)	88.1	**70.9**
Group DRO (all combinations of class/identities)	86.4	59.8

Table 3: Baseline results on CIVILCOMMENTS-WILDS.

Distribution shift and evaluation. We focus on *subpopulation shift* with respect to different demographic identities. Each comment is annotated with 8 binary indicators corresponding to whether a majority of crowdworkers believe it mentions each of the 8 demographic identities *male*, *female*, *LGBTQ*, *Christian*, *Muslim*, *other religion*, *Black*, and *White*. For each identity (e.g., "male"), we form 2 groups based on the toxicity label (e.g., one group of comments that mention the male gender and are toxic, and another group that mentions the male gender and are not toxic), for a total of 16 groups. These groups overlap (a comment might mention multiple identities) and are not a complete partition (a comment might not mention any identity). We measure a model's performance by its worst-group accuracy, i.e., its lowest accuracy over these 16 groups, and its average accuracy. Equivalently, the training distribution is the empirical data distribution, while the test distributions correspond to different subpopulations of the data distribution, divided along demographic lines. A high worst-group accuracy implies that the model is not spuriously associating a demographic identity with toxicity.

Baseline out-of-distribution results (Table 3). We fine-tuned a standard BERT-base-uncased model (Devlin et al., 2019) by minimizing the average training loss, and found that it does well on average but poorly on the worst group. Group DRO models (Sagawa et al., 2020), which seek to do well on all specified groups in the data, have higher worst-group accuracy; however, this comes at the cost of lower average accuracy, and there still remains a large gap between average and worst-group accuracies.

In-distribution results. The small sizes of the groups with low accuracies make it challenging to measure "in-distribution" performance. For example, the best-performing DRO model above does most poorly on the group of non-toxic Black comments (i.e., it often mistakes them for toxic comments). Ideally, we would train a model only on Black comments and measure its in-distribution accuracy on distinguishing toxic from non-toxic Black comments; but Black comments comprise only <4% of the training data, which is insufficient to obtain high in-distribution accuracy. As a proxy, we train a reweighted model by equally sampling toxic and non-toxic Black comments (and likewise, toxic and non-toxic non-Black comments) in each minibatch. This obtains an accuracy of 75.3% on non-toxic Black comments and 76.6% on toxic Black comments.[1] This is an imperfect proxy, in that it is likely an underestimate of true in-distribution accuracy, but nevertheless already higher than the baseline worst-group accuracies.

Discussion. We believe there is room for substantial progress on CIVILCOMMENTS-WILDS: since identity annotations are provided at training time, we have an IID dataset available at training time for each of the test distributions of interest (corresponding to each group), and it is also reasonable to expect that a model should be able to accurately predict which group a test example belongs to.

[1] Unfortunately, this reweighting scheme does not solve the original task: the accuracy on a different group (non-toxic LGBTQ comments) drops to 46.3%, resulting in an even lower overall worst-group accuracy.

Potential approaches on this task include adapting methods like reweighting (Shimodaira, 2000) and group DRO (Sagawa et al., 2020) to handle multiple overlapping groups, which were not studied in their original settings, or by using baselining to account for different groups having different intrinsic levels of difficulty (Oren et al., 2019).

The original CivilComments dataset (Borkan et al., 2019) also contains ≈1.5M training examples that have toxicity (label) annotations but not identity (group) annotations. For simplicity, we omitted these from the baselines above, but have included these with the CIVILCOMMENTS-WILDS dataset release. These additional data points substantially enlarge the training set and could be used, e.g., by first inferring which group each additional point belongs to, and then running group DRO or a similar algorithm that uses group annotations at training time.

Broader context. Demographic disparities in natural language and speech processing have been widely documented (Hovy & Spruit, 2016). For example, NLP models have been shown to obtain worse performance on African-American Vernacular English compared to Standard American English on part-of-speech tagging (Jørgensen et al., 2015), dependency parsing (Blodgett et al., 2016), language identification (Blodgett & O'Connor, 2017), and auto-correct systems (Hashimoto et al., 2018). Similar disparities exist in speech, with state-of-the-art commerical systems obtaining higher word error rates on particular races (Koenecke et al., 2020) and genders and dialects (Tatman, 2017). Note that the CIVILCOMMENTS-WILDS dataset does not assume that user demographics are available. Instead, it uses mentions of different demographic identities in the actual comment text (e.g., we want a model that does not associate comments that mention being Black with being toxic, regardless of whether a Black or non-Black person wrote the comment).

These disparities are present not just in academic models, but in large-scale commercial systems that are already widely deployed, e.g., in speech-to-text systems from Amazon, Apple, Google, IBM, and Microsoft (Tatman, 2017; Koenecke et al., 2020) or language identification systems from IBM, Microsoft, and Twitter (Blodgett & O'Connor, 2017). Indeed, the original CivilComments dataset was developed by Google's Conversation AI team, which is also behind a public toxicity classifier (Perspective API) that was developed in partnership with The New York Times (NYTimes, 2016).

4.2 AMAZON-WILDS: Sentiment classification across different users

Models are often trained on data collected on a set of users and then deployed as a general-purpose model across a wide range of users, and yet they can exhibit large performance disparities across individuals (Li et al., 2019b; Caldas et al., 2018; Geva et al., 2019; Tatman, 2017; Koenecke et al., 2020). These performance gaps are practical limitations in applications that call for good performance across a wide range of users (e.g., user-facing models). In addition, they can be indicative of unfairness of models (Li et al., 2019b; Dwork et al., 2012) as well as their failure to learn the actual task in a generalizable fashion, with models learning the biases specific to individuals instead (Geva et al., 2019). We study this issue of inter-individual performance disparities in the sentiment classification task on the AMAZON-WILDS dataset (Ni et al., 2019), in which our goal is to train models with consistently high performance across reviewers. Dataset details and supplemental results are in Appendix A.2.

Setup. We consider a multi-class classification task, in which the input is a review text and the label is a corresponding star rating out of 5. Each training and test example is annotated with the reviewer ID. In addition, at test time, we provide a batch of 75 unlabeled reviews for each user for test-time adaptation. We consider disjoint reviewers at training versus test time.

Distribution shift and evaluation. The goal is to train a model that performs well not only on an average reviewer seen at training time, but also on a wide range of unseen reviewers. Concretely, we evaluate models by their accuracy on the reviewer at the 10th percentile, following federated learning literature.

Out-of-distribution baseline results. First, we show that there are significant performance disparities across reviewers by presenting results on a standard model. A standard BERT-base-uncased model trained with the standard ERM objective performs well on average, but their performance vary widely across users (Figure 3, Table 4). Despite the high average accuracy of 74.1%, accuracies on reviewers vary widely between 100.0% and 9.3% (worse than random), with accuracy at the

	Reviewer ID	Review Text	Stars
Train	1	They are decent shoes. Material quality is good but the color fades very quickly. Not as black in person as shown.	5
		Super easy to put together. Very well built.	5
	2	This works well and was easy to install. The only thing I don't like is that it tilts forward a little bit and I can't figure out how to stop it.	4
		Perfect for the trail camera	5
		…	
	10,000	I am disappointed in the quality of these. They have significantly deteriorated in just a few uses. I am going to stick with using foil.	1
		Very sturdy especially at this price point. I have a memory foam mattress on it with nothing underneath and the slats perform well.	5
Test	10,001	Solidly built plug in. I have had 4 devices plugged in and all charge just fine.	5
		Works perfectly on the wall to hang our wreath without having to do any permanent damage.	5
		…	

Figure 2: AMAZON-WILDS dataset.

Model	Average accuracy	10th percentile accuracy
Standard BERT	0.76	0.57
Reweighted BERT (class balance)	0.70	0.53
Group DRO BERT (reviewer)	0.71	0.55

Table 4: Baseline results on AMAZON-WILDS.

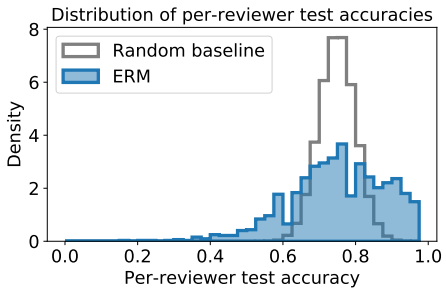


Figure 3: Distribution of per-reviewer accuracy on the test set for the ERM model (blue). The corresponding random baseline would have per-reviewer accuracy distribution in grey.

tenth percentile of 57.3%. The above variation is larger than expected from randomness; a random binomial baseline with equal average accuracy would have a tenth percentile accuracy of 68.5%.

In addition, we show that a few existing robust training procedures fail to solve the issue (Table 4). We observe that resampling to achieve uniform class balance fails to improve the 10th percentile accuracy, showing that variation across users cannot be solved simply by accounting for label imbalance. In addition, we show that group DRO fails to improve performance on unseen users as expected.

In-distribution results. We now show that it is possible to achieve high performance even on reviewers that the standard models perform poorly on, suggesting that there is substantial room for improvement. Concretely, we train oracle baseline models specific to each of those reviewers by fine-tuning on their own reviews, and show that they achieve high accuracy. We consider reviewers at the tenth percentile or below in terms of accuracy of the standard ERM model, and in particular those with the highest number of reviews. Despite being trained on data that are orders of magnitude smaller (thousands of reviews per user, compared to the full training set of 1 million reviews), the oracle baseline models achieve 74.1% accuracy on average, outperforming the standard models by 18.7% on those reviewers.

Discussion. The above baseline results suggest that even though current training procedures yield models that fail to consistently perform well across reviewers, it is possible to have a reviewer-specific model that performs well on each user given labeled examples from each user. Our set-up seeks to train reviewer-specific models through *test-time adaptation*, leveraging a batch of unlabeled examples available for each user at test time as well as reviewer ID annotations at both training and test time.

Broader context. Performance disparities across individuals have been observed in a wide range of tasks and applications, including in natural language processing (Geva et al., 2019), automatic speech recognition (Koenecke et al., 2020; Tatman, 2017), federated learning (Li et al., 2019b; Caldas et al., 2018), and medical imaging (Badgeley et al., 2019). These performance gaps are practical limitations in applications that call for good performance across a wide range of users, including many user-facing applications such as speech recognition (Koenecke et al., 2020; Tatman, 2017) and personalized recommender systems (Patro et al., 2020), tools used for analysis of individuals such as sentiment classification in computational social science (West et al., 2014) and user analytics (Lau et al., 2014), and applications in federated learning. These performance disparities have also been studied in the context of algorithmic fairness, including in the federated learning literature, in which uniform performance across individuals is cast as a goal toward fairness (Li et al., 2019b; Dwork et al., 2012). Lastly, these performance disparities can also highlight models' failures to learn the actual task in a generalizable manner; instead, some models have been shown learn the biases specific to individuals. Prior work has shown that individuals—technicians for medical imaging in this case—can not only be identified from data, but also are predictive of the diagnosis, highlighting the risk of learning to classify technicians rather than the medical condition (Badgeley et al., 2019). More directly, across a few natural language processing tasks where examples are annotated by crowdworkers, models have been observed to perform well on annotators that are commonly seen at training time, but fail to generalize to unseen annotators, suggesting that models are merely learning annotator-specific patterns and not the task (Geva et al., 2019).

4.3 CAMELYON17-WILDS: Tumor identification across different hospitals

Variation in data collection and processing across different hospitals can degrade model accuracy on data from hospitals not included in the training set (e.g., Zech et al. (2018); AlBadawy et al. (2018)). In histopathology applications—studying tissue slides under a microscope—this variation can arise from sources like differences in the patient population or in slide staining and image acquisition (Veta et al., 2016; Komura & Ishikawa, 2018; Tellez et al., 2019). We study this distribution shift through the CAMELYON17-WILDS dataset (Figure 4), which comprises 450,000 patches extracted from 50 whole-slide images of breast cancer metastases in lymph node sections, with 10 slides from each of 5 hospitals in the Netherlands (Bandi et al., 2018). Dataset and model details are in Appendix A.3.

Task. Binary classification, where the input is a 96x96 patch, and the output is whether the central 32x32 region is purely normal tissue or instead contains any tumor tissue. The tumor regions were manually annotated by pathologists.

Distribution shift and evaluation. Our goal is to evaluate how well a model generalizes to data from a hospital that it was not trained on. To this end, the training data comprises patches extracted from slides from 4 hospitals; the validation data comprises patches extracted from a disjoint set of slides from the same 4 hospitals; and the test data comprises patches from the 5th hospital. Each split has a 50/50 class balance, and models are evaluated on average accuracy.

Baseline out-of-distribution results (Table 5). We trained a standard DenseNet-121 (Huang et al., 2017) that was pretrained on ImageNet to minimize average training loss, following prior work (Veeling et al., 2018). Across different hyperparameters (learning rate, L_2 regularization, and random seeds), this model was consistently accurate on the validation set—which was from the same hospitals as the training set—but wildly inconsistent on the test set. In Table 5, we show the ranges corresponding to all runs that were within 2 $\times$ the estimated standard deviation due to random seeds. Test accuracy is extremely variable, implying that it is difficult to reliably select a model with high test accuracy: the run with the highest validation accuracy (96.5%) had a test accuracy of 68.3%, and another run with a validation accuracy of 95.1% had a test accuracy of 55.7%. Group DRO models (treating each hospital as a group) perform similarly.

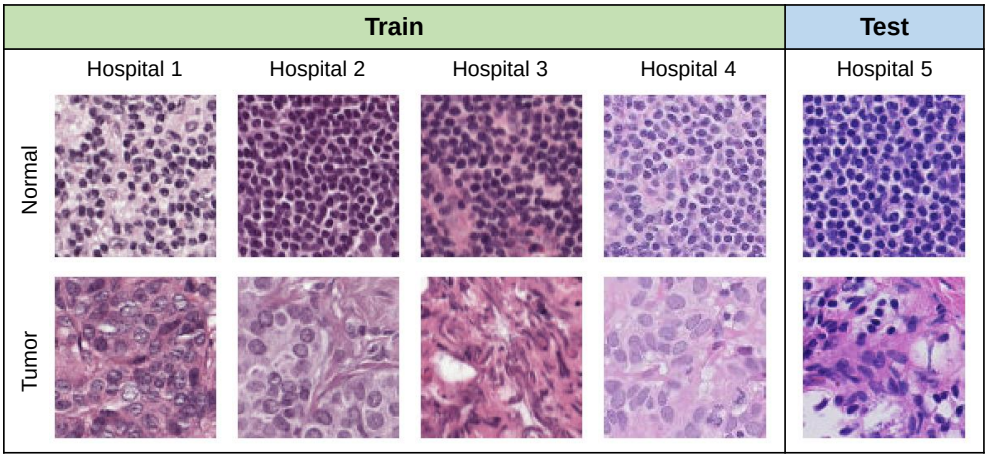


Figure 4: Sample patches from each hospital in CAMELYON17-WILDS. Each column contains two patches, one of normal tissue and the other of tumor tissue, from the same slide.

Model	Validation (in-distribution) accuracy	Test (out-of-distribution) accuracy
Standard	95.5–96.5	68.3–90.3
Group DRO	95.0–96.5	62.6–92.5

Table 5: Baseline results on CAMELYON17-WILDS.

In-distribution results. We evaluated in-distribution accuracy by moving 1 of the 10 slides from the test hospital to the training set, corresponding to about 15% of the test patches,[2] and testing on (the patches from) the remaining 9. The resulting model gets similar accuracy (94.8%–96.7%) on the same validation set but also achieves consistently high test accuracy (92.7%–95.6%), indicating that in-distribution test accuracy is stable.[3]

Discussion. These results reveal a subtle failure mode in out-of-distribution accuracy: there are models (i.e., choices of hyperparameters) that do well both in- and out-of-distribution, but we cannot reliably choose these models from just the training/validation set. This is consistent with previous reports of generalization instability in NLP models (McCoy et al., 2019a; Kim & Linzen, 2020). As pathology datasets tend to have small numbers of slides (though a potentially large number of patches extracted from these slides), predictive accuracy can be correlated across patches from the same slides, exacerbating instability (Zhou et al., 2020).

Prior work has shown that differences in staining between hospitals are the primary source of variation in this dataset, and that specialized stain augmentation methods can close the in- and out-of-distribution accuracy gap on a variant of the dataset based on the same underlying slides (Tellez et al., 2019). In this way, the CAMELYON17-WILDS dataset is a controlled testbed for general-purpose methods that can learn to be robust to stain variation between hospitals, given a training set from multiple hospitals.

Histopathology datasets can be unwieldy for ML models, as individual images can be several gigabytes large; extracting patches involves many design choices; and evaluation often relies on more complex slide-level measures such as the free-response receiver operating characteristic (FROC) (Gurcan et al., 2009). To improve accessibility, we pre-process the slides into patches and balance the dataset so that average accuracy is a reasonable measure of performance (Veeling et al., 2018; Tellez et al., 2019).

[2]Some slides contribute more patches than othersbecause the tumor regions are larger on those slides.

[3]Removing 1 slide from the test set does not change the original test accuracy much. In the original training split, the test accuracy on the remaining 9 test slides is similarly unstable (70.7%–91.2%).

Broader context. Beyond histopathology applications, variation between different hospitals/deployment sites has also been shown to degrade model accuracy in other medical applications such as diabetic retinopathy (Beede et al., 2020) and chest radiographs (Zech et al., 2018; Phillips et al., 2020), including recent work on COVID-19 detection (DeGrave et al., 2020). Even within the same hospital, process variables like which scanner/technician took the image can significantly change model predictions (Badgeley et al., 2019).

In these medical applications, the gold standard is to evaluate models on an independent test set collected from a different hospital (e.g.,Beck et al. (2011); Liu et al. (2017); McKinney et al. (2020)) or at least with a different scanner within the same hospital (e.g., Campanella et al. (2019)). However, this practice has not been ubiquitous due to the difficulty of obtaining data spanning multiple hospitals (Esteva et al., 2017; Bejnordi et al., 2017; Codella et al., 2019; Veta et al., 2019). As the CAMELYON17-WILDS dataset features multiple hospitals in the training set and an independent hospital in the test set, we hope that it will be useful for developing models that can generalize to new hospitals and contexts (Chen et al., 2020).

4.4 iWildCam2020-wilds: Species classification across different camera traps

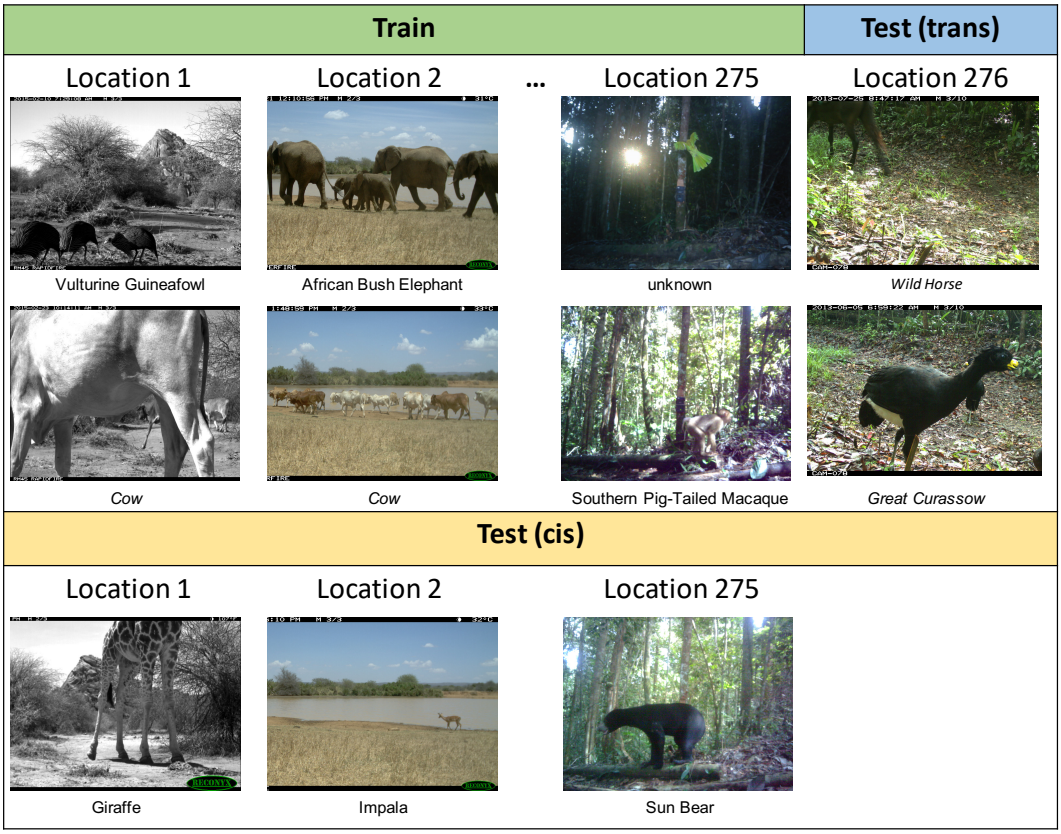


Figure 5: iWildCam2020-wilds dataset.

In the 2020 Living Planet Report (Grooten et al., 2020) , the WWF found that animal populations have declined 68% on average since 1970. In the current climate crisis, understanding the connection between climate change, human impact, and wildlife biodiversity loss is more important than ever. Camera traps, heat or motion activated static cameras placed in the wild, are one of the primary methods for monitoring wildlife in the ecology community Wearn & Glover-Kapfer (2017). Camera traps collect data much faster than experts can process it, and so the ecologists have turned to computer vision as an efficient solution (Ahumada et al., 2020; Weinstein, 2018; Norouzzadeh et al., 2019; Tabak et al., 2019; Beery et al., 2019). However, camera traps, and in fact all static sensors, capture signals that are correlated in time and space. This correlation results in overfitting and poor generalization to new sensor deployments, reducing the scalability of computer vision solutions

Model	Average accuracy	Macro F1
ERM	0.6573	0.218
Upsampled ERM (class balance)	0.558	0.1726
ARM-BN (location)	0.613	0.15

Table 6: Baseline results on iWildCam2020-wilds.

(Beery et al., 2018). We study this problem using the iWildCam 2020 Competition Dataset (Beery et al., 2020a), where the goal is to classify animal species in static camera traps and generalize to images from new camera deployment locations. Dataset and model details are in Appendix A.4.

Setup. The task is multi-class classification where the input is an image, and the output is one of 196 classes. Each image is annotated with one of 292 location IDs, and at test time we can assume that all images in each batch are coming from the same location, enabling test-time adaptation.

Distribution shift and evaluation. We wish to evaluate how well the model generalizes to images from camera trap locations that were not in the train set (trans setting), as well to locations that were in the train set (cis setting). Therefore, data was split into five groups: train, validation-cis, validation-trans, test-cis and test-trans. We first formed the validation-trans and test-trans splits by randomly splitting off locations. Then, using the remaining locations, we formed the train, validation-cis, and test-cis splits, by randomly splitting by date (see Appendix A.4 for details). In the natural world, protected and endangered species are rare by definition, and are often the most important to accurately monitor. However, common species are much more likely to be captured in camera trap images, leading to a vast imbalance in data representation from rare species to common ones. To capture both per-class and overall performance across the imbalanced class set, we report Macro F1 as well as average accuracy.

Out-of-distribution baseline results. We trained a Densenet-121 that was pretrained on Imagenet. Average accuracy on out-of-distribution locations was around 65%. We also ran another baseline where we upsample classes based on inverse class frequency. This performed significantly worse than ERM, possibly due to upweighting many classes in the train set that are not in the test set. We also investigated the ARM-Batchnorm (Zhang et al., 2020) which adapts to the location at test time. This also performed worse than ERM.

In-distribution results. Average accuracy for ERM on in-distribution locations (test-cis) was consistently around .95, and .91 for ARM-BN and .80 for ERM upsampling. Note that we are using the same trained models as when we are evaluating on out-of-distribution locations.

Discussion. The large discrepancy between the in-distribution location and out-of-distribution locations suggests that there is large room for improvement. Even though there is significant label imbalance, the label distribution approximately is the same in the cis and the trans split, suggesting that it is not primarily label shift that accounts for the performance drop. Across locations, there is drastic variation in illumination, camera angle, and background, vegetation and color. This variation coupled with a considerable shift in the distribution of animals, likely encourages the model to pick overfit to specific locations which may account for the performance drop.

Though this setting seems suitable for test-time adaptation, this specific version of adaptive risk minimization, namely adapting using batchnorm, does not do well on this task. Since there is considerable label shift, between locations, we expect adaptive methods to be able to adapt to location specific species distribution and specific backgrounds. In line with this, Beery et al. (2020b) show that using images from the full sequence can significantly improve performance over a models that utilize single frames only.

Broader context. Differences across data distributions at different sensor locations is a common challenge in automated wildlife monitoring applications, including using audio sensors to monitor animals that are easier heard than seen such as primates, birds, and marine mammals (Crunchant et al., 2020; Stowell et al., 2019; Shiu et al., 2020), and using static sonar to count fish underwater to help

maintain sustainable fishing industries (Pipal et al., 2012; Vatnehol et al., 2018; Schneider & Zhuang, 2020). As with camera traps, each static audio sensor has a specific species distribution as well as a sensor specific background noise signature, making generalization to new sensors challenging. Similarly, static sonar that is used to measure fish escapement have sensor specific background reflectance based on the shape of the river bottom. Moreover, since species are distributed in a non-uniform and long-tailed fashion across the globe, it is incredibly challenging to collect enough samples for each species. Implicitly representing camera-specific distributions and background features in per-camera memory banks and extracting relevant information from these via attention has been shown to help overcome some of these challenges for static cameras Beery et al. (2020b).

More broadly, shifts in background, image illumination and viewpoint have been studied in computer vision research. First, several works have shown that object classifiers often rely on the background rather than the object to make its classification (Rosenfeld et al., 2018; Shetty et al., 2019; Xiao et al., 2020). Second, common perturbations such as blurriness or shifts in illumination, tend to reduce performance (Dodge & Karam, 2017; Temel et al., 2018; Hendrycks & Dietterich, 2019). Finally, shifts in rotation and viewpoint of the object has been shown to degrade performance (Barbu et al., 2019).

4.5 POVERTYMAP-WILDS: Mapping poverty across different countries

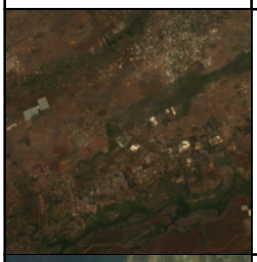
Input	Auxiliary Information	Target
	*Survey year:* `2014` *Location:* `(-1.05313, 37.087)` *Nighttime Light Intensity:* `0.84`	*Asset Wealth Index:* `0.369`
	*Survey year:* `2010` *Location:* `(-1.68038, 29.2651)` *Nighttime Light Intensity:* `0.23`	*Asset Wealth Index:* `1.048`

Figure 6: Examples from the POVERTYMAP-WILDS dataset. The top example is from an urban location in Kenya with a low asset wealth score, while the bottom example is from a rural location in Rwanda with a higher asset wealth score. There may be significant economic and cultural differences across country borders that contribute to the spatial distribution shift.

High-resolution predictions of poverty measures are essential for targeted humanitarian efforts and directing policy decisions in developing countries (Espey et al., 2015; Abelson et al., 2014). However, ground truth measurements of poverty are lacking for much of the developing world, since gathering them requires conducting expensive surveys in the field (Blumenstock et al., 2015; Xie et al., 2016; Jean et al., 2016; Yeh et al., 2020). At least 4 years pass between nationally representative consumption or asset wealth surveys in the majority of African countries, with seven countries that had either never conducted a survey or had gaps of over a decade between surveys (Yeh et al., 2020). The lack of labels in certain countries creates a natural scenario where we desire model generalization to unseen countries. We study this cross-border distribution shift in the POVERTYMAP-WILDS dataset (Figure 6), which assembles satellite imagery and survey data at 19,669 villages from 23 African countries between 2009 and 2016 (Yeh et al., 2020).

Task. Regression, where the input is a $224 \times 224 \times 7$ multispectral image from the LandSat 7 satellite and the regression target is a real-valued asset wealth index computed from survey data.

	Validation (in-distribution) MSE	Test (out-of-distribution) MSE	Validation r^2	Test r^2
Baseline	0.051	0.156	0.611	0.484
Oracle	0.134	0.138	0.540	0.497

Table 7: Mean squared error (MSE) and r^2 on in-distribution and out-of-distribution (unseen countries) held-out sets in POVERTYMAP-WILDS. All results are averaged over 5 folds taken from Yeh et al. (2020).

Survey data comes from the Demographic and Health Surveys (DHS). Each example comes with location coordinates, the survey year, and an additional eighth image channel for nighttime light intensity from a separate satellite, which correlates with poverty measures (Noor et al., 2008; Elvidge et al., 2009). Additional unlabeled satellite imagery with corresponding nighttime light intensities, sampled around DHS survey locations, are also available.

Distribution shift and evaluation. We aim to generalize to countries not in the training set. Yeh et al. (2020) presents two sets of data splits meant for testing in-country and out-of-country generalization. The out-of-country scheme splits the 23 countries into 5 folds such that each fold has roughly the same number of examples. The in-country scheme splits the 19669 villages into 5 folds such that there is no overlap in the spatial extent of the satellite images between any fold.

We combine these two schemes to test in-country and out-of-country generalization simultaneously. For the i-th fold, we first take the i-th in-country training/validation fold and remove all countries from the i-th out-of-country fold from these splits. We use the the out-of-country data from the validation set as the test set. We provide the folds we use for ease of use. The models are evaluated on mean squared error (MSE) and squared Pearson correlation (r^2), as is standard in the literature (Jean et al., 2016; Yeh et al., 2020).

Out-of-distribution baseline results. Following Yeh et al. (2020), we trained a ResNet-18 taking only the 7 multispectral channels from LandSat to minimize squared error. Table 7 shows that the baseline model suffers about a $3\times$ increase in MSE and a 0.12 drop in r^2 on test examples from unseen countries.

In-distribution baseline results. The baseline model achieves low MSE (0.051 to 0.06) and r^2 over 0.6 on the validation set. We estimate results without spatial shift with an oracle model trained on data sampled from all the countries uniformly so that there is no unseen country. We use the same number of training points for this oracle training set as for the baseline training data. As a result, the oracle training set has fewer samples in each country (but has data from more countries) and the validation MSE increases, but the gap between the validation and test MSE becomes small.

Discussion. These results corroborate performance drops seen in previous out-of-country generalization tests for poverty prediction from satellite imagery (Jean et al., 2016). In general, differences in infrastructure, economic development, agricultural practices, and even cultural differences can cause large shifts across country borders. Differences between urban and rural subpopulations have also been well-documented (Jean et al., 2016; Yeh et al., 2020). Models based on nighttime light information could suffer more in rural areas where nighttime light intensity is uniformly low or even zero. Given large shifts, is it possible to generalize across borders? We note that some indicators of wealth are known to be robust and are able to be seen from space. For example, roof type (e.g. thatched or metal roofing) has been shown to be a reliable proxy for wealth (Abelson et al., 2014), and context factors such as nearby cropland health, presence of paved roads, and connections to urban areas are plausibly reliable signals for measuring poverty.

Since survey years are also available, we could also investigate the robustness of the model over time. This would enable the models to be used for a longer time before needing more updated survey data, and we leave this to future work. Yeh et al. (2020) investigated predicting the change in asset wealth for individual villages in the World Bank Living Standards Measurement Surveys (LSMS), which is a longitudinal study containing multiple samples from the same village, finding that the village level task is relatively difficult ($r^2 = 0.35$) while aggregate predictions at the district level are more promising ($r^2 = 0.51$). Instead of predicting a time-series at each village, we can also consider shifts

across years for cross-sectional samples such as in DHS using POVERTYMAP-WILDS, which we leave to future work.

Since satellite imagery is available globally independent of survey data, some works also leverage unlabeled data in a semi-supervised method (Xie et al., 2016; Jean et al., 2018; Xie et al., 2020). POVERTYMAP-WILDS is a potential testbed for robust models that leverage unlabeled data, along with auxiliary sources of satellite information such as climate or nighttime light data, to improve generalization to domains with low labeled data.

Broader context. Computational sustainability applications in the developing world includes not only poverty mapping but also tracking child mortality (Osgood-Zimmerman et al., 2018; Reiner et al., 2018; Burke et al., 2016), educational attainment (Graetz et al., 2018), food security and crop yield prediction (Wang et al., 2020b; You et al., 2017; Xie et al., 2020). Remote sensing data and satellite imagery has the potential to enable high-resolution maps of many of these sustainability challenges, but similarly to poverty, ground truth labels in these applications expensive surveys or observations from human workers in the field. We hope that POVERTYMAP-WILDS can be used to improve the robustness of machine learning techniques on satellite data, providing an avenue for cheaper and faster measurements that can be used to make progress on a general set of computational sustainability challenges.

4.6 FMOW-WILDS: Classifying building and land use across different regions and years

Input	Auxiliary Information	Target
	*Timestamp:* `2009-08-13T08:09:51Z` *Location:* `(-1.10265, 37.0211)` *Country Code:* `KEN` *Region:* `Africa`	*Building/Land Class (of 62):* `"multi-unit residential"`
	*Timestamp:* `2017-02-12T08:05:31Z` *Location:* `(-1.10265, 37.0211)` *Country Code:* `KEN` *Region:* `Africa`	*Building/Land Class (of 62):* `"multi-unit residential"`

Figure 7: Examples from the FMOW-WILDS dataset. The two examples are from the same location in Kenya, taken about 5.5 years apart.

As human activity and environmental processes change the natural environment and human-made infrastructure, ML models on satellite imagery must be robust to distribution shifts over time. Disparities in data available between regions can also lead to performance disparities over spatial locations. We study these shifts with the Functional Map of the World (FMOW-WILDS) dataset (Christie et al., 2018), which collects and categorizes over 1 million high-resolution satellite images from over 200 countries based on the functional purpose of the buildings or land in the image, over the years 2002-2018 (see Figure 7).

Task. We use the RGB version of the original dataset, which contains 523846 total examples, excluding the multispectral version of the images. The simplest formulation is as a classification problem, where the input is a $224 \times 224 \times 3$ satellite image and the classification target is one of 62 building or land use categories. Each example is paired with corresponding auxiliary information, such as the timestamp and location coordinate. Methods that can utilize a sequence of images can

	Validation (<2013) accuracy	Test (≥ 2013) accuracy	2017 accuracy
Baseline	59.22	55.51	48.04
Oracle	59.71	59.75	54.72

Table 8: Time shift results for models trained on data before 2013 and tested on held-out locations from validation or test in FMoW-WILDS. The baseline accuracy drops significantly in the last year of the dataset.

	Americas	Africa	Asia	Europe	Oceania	Worst region
Baseline (test)	55.87	35.80	56.42	59.20	58.99	35.80
Oracle (test)	60.02	51.74	58.61	61.76	64.77	51.74
Baseline (validation)	61.05	65.07	58.66	57.61	72.48	57.61
Oracle (validation)	61.69	70.33	57.06	58.95	71.10	57.06

Table 9: Region shift results (accuracy, %) for models trained on data before 2013 and tested on held-out locations from the validation set (<2013) or test set (≥ 2013) in FMoW-WILDS.

group the images from the same location across multiple years together as input, but we consider the simple formulation here for our baseline evaluation.

Distribution shift and evaluation. We aim to generalize to shifts across time. We first hold out all images from 2013-2018 (≥ 2013) from the original validation set(Christie et al., 2018) as the out-of-distribution test set. We split the remaining portion (<2013) of the given training and validation sets as our training and validation set. The training, validation, and test splits contain images from disjoint location coordinates. We do not artificially constrain the training data to be from different regions. We use the full distribution over regions in the training data, but evaluate the sub-population performance on different regions.

Out-of-distribution baseline results. Following Christie et al. (2018), we train a Densenet-121 model pretrained on ImageNet to minimize the cross entropy loss on data from the <2013 training split. Table 8 shows that accuracy drops almost 4% when evaluated on test set (≥ 2013), and that the accuracy drop is especially large (11%) on images from the last year of the dataset (2017), furthest in the future from the training set.

We also study the baseline performance on different regions of the world. Figure 8 shows that there is a disparity in the number of examples in each region, where Africa and Oceania have the least examples (this could be due to bias in sampling, lack of infrastructure/land data in certain regions). Table 9 shows that on the test set, the baseline model performs much worse in Africa (35.80%) than other regions.

In-distribution baseline results. Overall, the baseline model does better on the validation set than the time-shifted test set. We estimate results without time shift with an oracle model trained on data sampled 50-50 between <2013 and ≥ 2013. For time shifts (Table 8, the oracle does not have a shift in performance between validation and test, and the shift on images from 2017 is much smaller. For

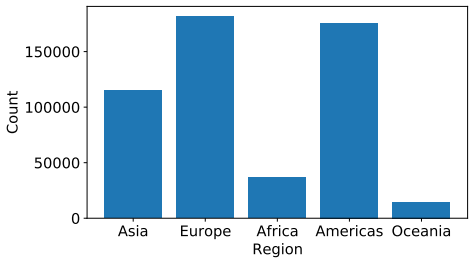


Figure 8: Number of examples from each region of the world in FMoW-WILDS. There is much less data from Africa and Oceania than other regions.

region shifts (Table 9), the oracle does not show as much of a shift across regions, although Africa is still the worst region on the test set.

Discussion. Intriguingly, region shift results on the validation set show that the combination of time and region shift is necessary to see a large drop in performance across regions. This is corroborated by the oracle region shift results (Table 9), which do not have time shift between training and test and do not display a large drop in performance across regions. As we show in Appendix A.6, we find that there is a large label distribution shift between non-African regions and Africa, suggesting that the drop in performance may be in some part due to label shift.

Despite having the smallest number of training examples (Figure 8), the baseline models do not suffer a drop in performance in Oceania on validation or test sets (Table 9). We hypothesize that infrastructure in Oceania is more similar to regions with a large amount of data than Africa. In contrast, Africa may be more distinct and may have changed more drastically over 2002-2018, the time extent of the dataset. This suggests that the subpopulation shift is not merely a function of the number of training examples.

As seen by the large drop in performance in Africa over time, developing countries may present a bigger challenge for robustness over time. One source of leverage for mitigating the effect of time shift is the presence of other countries around the world in varying stages of economic development, from which we can transfer some knowledge. Another source of leverage is globalization — trade of materials and expertise across the world help to create more commonalities across nations. Finally, changes over time are often gradual, and utilizing this gradual shift structure may enable adaptation across longer time periods (Kumar et al., 2020).

Compared to POVERTYMAP-WILDS, FMOW-WILDS contains much higher resolution images (sub-meter resolution vs. 30m resolution) and contains a larger variety of viewpoints/tilts, both of which could present computational or algorithmic challenges. For computational purposes, we resized all images to 224×224 (following Christie et al. (2018)), but raw images can be thousands of pixels wide. Some recent works have tried to balance this tradeoff between viewing overall context and the fine-grained detail (Uzkent & Ermon, 2020; Kim et al., 2016), but how best to do this is an open question. FMOW-WILDS also contains additional information on azimuth and cloud cover which could be used to correct for the variety in viewpoints and image quality.

Broader context. Recognizing infrastructure and land features is crucial to remote sensing applications such as crop yield prediction (Wang et al., 2020b), tracking deforestation (Hansen et al., 2013), population density mapping Tiecke et al. (2017), poverty mapping (Abelson et al., 2014; Jean et al., 2016), and other economic tracking applications (Katona et al., 2018). It is natural to have labeled data with limited temporal or spatial extent since ground truth generally must be verified on the ground or requires manual annotations from domain experts (i.e. often hard to be crowdsourced). A number of existing remote sensing datasets have limited spatial or temporal scope, including the UC Merced Land Use Dataset (Yang & Newsam, 2010), TorontoCity (Wang et al., 2017), and SpaceNet DigitalGlobe & Works (2016).

Although the data is typically limited, we desire generalization on a global scale without requiring frequent large-scale efforts to gather more ground-truth data. FMOW-WILDS has a larger spatial extent, has multiple timestamped views of a location over time, and encompasses more functional categories than the previous datasets. We hope FMOW-WILDS can serve as a benchmark for ML models on high-resolution satellite imagery that are robust to time and region shifts.

5 Potential extensions to other application areas

Beyond the datasets and application areas currently included in WILDS, there are many other applications where it is critical to handle distribution shifts. Here, we discuss some of these promising applications and the challenges of finding appropriate benchmark datasets in those areas. These all represent important avenues of future work, and we would highly welcome community contributions of benchmark datasets in these areas.

Algorithmic fairness. Distribution shifts which worsen algorithmic performance in historically disadvantaged or minority populations have been frequently discussed in the algorithmic fairness literature and represent an important area for research. Commercial gender classifiers are more likely

to misclassify the gender of darker-skinned women, likely in part because training datasets overrepresent lighter-skinned subjects (Buolamwini & Gebru, 2018). Algorithmic pedestrian detection systems achieve poorer performance when recognizing darker-skinned pedestrians (Wilson et al., 2019). As discussed in Section 4.1, NLP models also show racial bias.

Publicly available algorithmic fairness benchmarks (Mehrabi et al., 2019)—e.g., the COMPAS recidivism dataset (Larson et al., 2016)—often suffer from several limitations; ameliorating these represents a promising direction for future work. First, the datasets are often quite small by the standards of modern machine learning: the COMPAS dataset has only a few thousand rows (Larson et al., 2016). Second, the datasets sometimes have relatively few features, allowing even simple algorithms to achieve state-of-the-art performance and limiting the benefit of more sophisticated approaches. On the COMPAS dataset, logistic regression on a small number of features performs comparably to a black-box commercial algorithm (Jung et al., 2020; Dressel & Farid, 2018). Relatedly, disparities in performance across subgroups are not always large, again limiting opportunity for improvement from more sophisticated algorithms (Larrazabal et al., 2020). Third, the datasets sometimes represent "toy" problems which, while useful for illustrating the theoretical properties of an approach, do not represent real-world problems of interest. For example, the UCI Adult Income dataset (Asuncion & Newman, 2007) is widely used as a fairness benchmark, but the classification task—predicting whether a person will have an income above $50,000—does not represent a real-world application. Finally, because many of the domains in which algorithmic fairness is of most concern—for example, criminal justice and healthcare—are high-stakes and disparities are politically sensitive, it can be difficult to make datasets publicly available.

Creating algorithmic fairness benchmarks which do not suffer from these limitations represents a promising direction for future work. In particular, such datasets would ideally have: 1) information about a sensitive attribute like race or gender; 2) a prediction task which is of immediate real-world interest; 3) enough samples, a rich enough feature set, and large enough disparities in group performance that more sophisticated machine learning approaches would plausibly produce improvement over naive approaches.

Medicine and healthcare. Substantial evidence indicates the potential for distribution shifts in medical settings. One concern is *demographic* distribution shifts (eg, across race, gender, or socioeconomic status) (Chen et al., 2020), similar to the algorithmic fairness concerns discussed above. Historically disadvantaged populations are underrepresented in many medical datasets, potentially producing inferior algorithmic performance on these groups. A second source of distribution shifts is heterogeneity *across hospitals*, as discussed in Section 4.3. Finally, changes *over time* in care settings can also produce distribution shifts and drops in algorithmic performance: Nestor et al. (2019) shows that switching between two electronic health record (EHR) systems produced a drop in algorithmic performance. Similarly, temporal shifts in conditions affecting patient populations could cause distribution shifts: for example, the COVID-19 epidemic has affected the distribution of chest radiographs (Wong et al., 2020).

Creating medical distribution shift benchmarks thus represents a promising direction for future work, if several challenges can be overcome. First, while there are large demographic disparities in healthcare outcomes (eg, by race, or socioeconomic status), many of them are not due to distribution shifts, but to disparities in non-algorithmic factors (eg, access to care or prevalence of comorbidities (Chen et al., 2020)) or to algorithmic problems unrelated to distribution shift (eg, choice of a biased outcome variable (Obermeyer et al., 2019)). Several previous investigations have found relatively small disparities in algorithmic performance across demographic groups (Chen et al., 2019a; Larrazabal et al., 2020); Seyyed-Kalantari et al. (2020) finds larger disparities in TPR across demographic groups, and future work should investigate whether there are disparities in other performance metrics.

A second challenge to overcome is data availability, in part because stringent medical privacy laws often preclude data sharing (Price & Cohen, 2019). For example, EHR datasets are fundamental to medical decision-making, but there are few widely adopted EHR benchmarks (the MIMIC database representing one example (Johnson et al., 2016)) and relatively little progress in predictive performance has been made on them (Bellamy et al., 2020).

Natural language and speech processing. As mentioned in Section 4.1, recent work found that automated speech recognition (ASR) systems have higher error rates for black speakers than for white speakers (Koenecke et al., 2020) and for women and speakers of some dialects (Tatman, 2017). This is a natural setting for developing methods that are robust to subpopulation shifts, like in

CIVILCOMMENTS-WILDS. The aforementioned papers use commercial ASR systems to demonstrate these disparities. However, there are many public speech datasets with speaker metadata that could potentially be used to construct a benchmark, e.g., LibriSpeech (Panayotov et al., 2015), the Speech Accent Archive (Weinberger, 2015), VoxCeleb2 (Chung et al., 2018), the Spoken Wikipedia Corpus (Baumann et al., 2019), and Common Voice (Ardila et al., 2020).

In natural language processing (NLP), a current focus is on challenge datasets that are carefully crafted to test particular aspects of models, e.g., HANS (McCoy et al., 2019b), PAWS (Zhang et al., 2019), counterfactually-augmented datasets (Kaushik et al., 2019), or CheckList (Ribeiro et al., 2020). These challenge datasets represent a form of distribution shift, as their test distributions are often (deliberately) quite different from the data distributions that the models were originally trained on. However, one issue is that these datasets represent just a few of the many potential shifts in NLP applications, so hill-climbing on the specific types of shifts that these datasets represent might not translate to progress on general robust NLP models. Similarly, there are several synthetic datasets designed to test compositional generalization, such as CLEVR (Johnson et al., 2017), SCAN (Lake & Baroni, 2018), and COGS (Kim & Linzen, 2020). The test sets in these datasets are chosen such that models need to generalize to novel combinations of, e.g., familiar primitives and grammatical roles (Kim & Linzen, 2020).

All of the NLP examples we have discussed so far deal with English-language models; other languages typically have fewer and smaller datasets available for training and benchmarking models. Multi-lingual models and benchmarks (e.g., Conneau et al. (2018); Conneau & Lample (2019); Hu et al. (2020); Clark et al. (2020)) represent another source of subpopulation shifts with corresponding disparities in model performance: training sets might contain fewer examples in low-resource languages (Nekoto et al., 2020), but we would still hope for high model performance on these minority groups. A challenge here is that multi-lingual models are often more complex, requiring larger datasets, than their mono-lingual counterparts.

Education. ML models can help in educational settings in a variety of ways: e.g., assisting in grading (Piech et al., 2013; Shermis, 2014; Kulkarni et al., 2014; Taghipour & Ng, 2016), estimating student knowledge and ability (Desmarais & Baker, 2012; Wu et al., 2020), identifying students who need assistance (Ahadi et al., 2015), or automatically generating explanations for student submissions (Williams et al., 2016; Wu et al., 2019a). However, there are substantial distribution shift issues to deal with in these settings as well. For example, automatic essay scoring has been found to be affected by rater bias (Amorim et al., 2018) and spurious correlations like essay length (Perelman, 2014), leading to problems with subpopulation shift; and these systems would also ideally generalize across different educational contexts, e.g., a model for scoring grammar should work well across multiple different essay prompts. Unfortunately, finding a suitable education benchmark is difficult due to a general lack of standardized datasets, in part due to student privacy concerns and the proprietary nature of large-scale standardized tests. Nevertheless, dataset construction for ML in education is an active area—e.g., the NeurIPS 2020 workshop on Machine Learning for Education[4] has a segment devoted to finding "ImageNets for education"—and we hope to be able to include one in the future.

Robotics. Robot learning has emerged as a strong paradigm for automatically acquiring complex and skilled behaviors such as locomotion (Yang et al., 2019; Peng et al., 2020), navigation (Mirowski et al., 2017; Kahn et al., 2020), and manipulation (Gu et al., 2017; et al, 2019). However, the advent of learning based techniques for robotics has not convincingly addressed, and has perhaps even exasperated, problems stemming from distribution shift. These problems have manifested in many ways, including shifts induced by weather and lighting changes (Wulfmeier et al., 2018), location changes (Gupta et al., 2018), and the simulation-to-real-world gap (Sadeghi & Levine, 2017; Tobin et al., 2017). Dealing with these challenging scenarios is critical to deploying robots in the real world, especially in high-stakes decision making scenarios.

Consider, for example, the rollout of autonomous driving systems. Autonomous cars are already deployed in many places around the world, but as a society, we do not yet place full faith in these systems. This is not only because driving is a complex task; in this setting, we also require that these systems work reliably and robustly across the huge variety of conditions that exist in the real world, such as locations, lighting and weather conditions, and sensor intrinsics. This is a challenging requirement, as many of these conditions may be underrepresented, or not represented at all, by the available training data. As a result, some prior work has shown that naïvely trained models can suffer

[4] `https://www.ml4ed.org/`

at segmenting nighttime driving scenes (Dai & Van Gool, 2018), recognizing traffic signs in adverse weather conditions (Temel et al., 2017), and, as discussed earlier, detecting pedestrians with darker skin tones (Wilson et al., 2019).

Thus, we hope to add a dataset centered around autonomous driving that can highlight some of the key challenges in this area regarding distribution shift. Concrete examples of relevant problems include shifts arising from image compression, shifts caused by differing camera properties, and unexpected shifts such as defective cameras. One promising approach is to explore using datasets such as BDD100K (Yu et al., 2020) and nuImages (Caesar et al., 2019). Both datasets are publicly available, enabling adoption by the larger community, and they include meta data such as camera IDs and GPS locations. This facilitates the construction of distribution shift problems that are both realistic and challenging, and one concrete direction to explore is to split the data according to these particular meta data values to assess a model's ability to generalize to new cameras and locations.

6 Discussion

6.1 Trends.

Dataset specificity. We find that the performance drops due to distribution shifts are highly specific to the task and the dataset in line with prior work. We consistently observe that a type of shift leads to a substantial performance drop in one dataset, but not in others. For example, while we observed substantial performance gaps due to time shifts in FMoW-WILDS, we observe no performance gap on AMAZON-WILDS. Moreover, when we consider the same sentiment classification task on a similar review dataset (Yelp), we see modest but significant performance drops due to time shifts, unlike on AMAZON-WILDS. Similarly, we see large variation in model performance across users on AMAZON-WILDS, not on the Yelp dataset. Finally, location shifts affects model performance differently across the two satellite imagery datasets, FMoW-WILDS and POVERTYMAP-WILDS; while we see substantial performance gaps due to location shifts in POVERTYMAP-WILDS when we simply split by location, we only see such gaps in FMoW-WILDS after splitting by time.

Diversity of training domains. We observe that domain diversity in the training data is helpful for training a model that generalizes to unseen domains. On category shifts in AMAZON-WILDS, increasing the number of training categories from one to four (Books versus Books, Electronics, Movies, and Home) yields significant improvement in generalization to unseen categories. In fact, with respect to in-distribution baseline models that are trained on each target category, we observe substantial performance gaps for the model trained on a single source category, but the gap is largely eliminated when the model is trained on four categories. Similarly, in CAMELYON17-WILDS, increasing the number of training hospitals from one to four improves performance on the unseen hospital, even though both models perform poorly. The above observations underscore the importance of including as many domains as possible in the training data to the extent it is realistic, not only in practice, but also for robustness benchmarks.

6.2 Related problem settings

As our baseline results on the WILDS datasets suggest, building models that are robust to distribution shifts is a challenging open problem. We end by discussing two related problem settings that are potentially more tractable and that can also be tested on the WILDS benchmark.

Unsupervised domain adaptation. In the unsupervised domain adaptation setting, one has access to a large amount of unlabeled data from each test distribution of interest, as well as the resources to train a separate model for each test distribution. For example, in a satellite imagery setting like FMoW, it might be appropriate to assume that we have access to a large set of unlabeled recent satellite images from each continent and can train a separate model for each continent. Many of the methods for domain generalization discussed in Section 3 were originally methods for domain adaptation, since methods for both settings share the common goal of learning models that can transfer between domains; for example, methods that learn features that have similar distributions across domains are equally applicable to both settings (e.g., Ben-David et al. (2006); Long et al. (2015); Sun et al. (2016); Ganin et al. (2016); Tzeng et al. (2017); Shen et al. (2018); Wu et al. (2019b)). Other methods rely on knowing the test distribution and are thus specific to domain adaptation, e.g., learning to map data points from source to target domains (Hoffman et al., 2018).

Selective prediction. In the selective prediction setting, models are allowed to abstain on points where their confidence is below a certain threshold. This is appropriate in settings where, for example, abstentions can be handled by backing off to human experts, such as pathologists for CAMELYON17, content moderators for CIVILCOMMENTS, wildlife experts for iWILDCAM2020, etc. Many methods for selective prediction have been developed, from simply using softmax probabilities as a proxy for confidence (Cordella et al., 1995; Geifman & El-Yaniv, 2017), to methods involving ensembles of models (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017; Geifman et al., 2018) or jointly learning to abstain and classify (Bartlett & Wegkamp, 2008; Geifman & El-Yaniv, 2019; Feng et al., 2019).

Intuitively, even if a model is not robust to a distribution shift, it might at least be able to maintain high accuracies on some subset of points that are close to the training distribution, while abstaining on the other points. Indeed, prior work has shown that selective prediction can improve model accuracy under distribution shifts (Pimentel et al., 2014; Hendrycks & Gimpel, 2017; Liang et al., 2018; Ovadia et al., 2019; Feng et al., 2019; Kamath et al., 2020). However, distribution shifts still pose a problem to selective prediction methods; for instance, it is difficult to maintain desired abstention rates under distribution shifts (Kompa et al., 2020), and confidence estimates have been found to drift over time (e.g., Davis et al. (2017)).

References

B. Abelson, K. R. Varshney, and J. Sun. Targeting direct cash transfers to the extremely poor. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, 2014.

A. Ahadi, R. Lister, H. Haapala, and A. Vihavainen. Exploring machine learning methods to automatically identify students in need of assistance. In *Proceedings of the eleventh annual International Conference on International Computing Education Research*, pp. 121–130, 2015.

Jorge A Ahumada, Eric Fegraus, Tanya Birch, Nicole Flores, Roland Kays, Timothy G O'Brien, Jonathan Palmer, Stephanie Schuttler, Jennifer Y Zhao, Walter Jetz, et al. Wildlife insights: A platform to maximize the potential of camera trap and other passive sensor wildlife data for the planet. *Environmental Conservation*, 47(1):1–6, 2020.

E. AlBadawy, A. Saha, and M. Mazurowski. Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. *Med Phys.*, 45, 2018.

E. Amorim, M. Cançado, and A. Veloso. Automated essay scoring in the presence of biased ratings. In *Association for Computational Linguistics (ACL)*, pp. 229–237, 2018.

R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber. Common voice: A massively-multilingual speech corpus. In *Language Resources and Evaluation Conference (LREC)*, pp. 4218–4222, 2020.

M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Arthur Asuncion and David Newman. UCI machine learning repository, 2007.

J. Atwood, Y. Halpern, P. Baljekar, E. Breck, D. Sculley, P. Ostyakov, S. I. Nikolenko, I. Ivanov, R. Solovyev, W. Wang, et al. The inclusive images competition. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 155–186, 2020.

M. A. Badgeley, J. R. Zech, L. Oakden-Rayner, B. S. Glicksberg, M. Liu, W. Gale, M. V. McConnell, B. Percha, T. M. Snyder, and J. T. Dudley. Deep learning predicts hip fracture using confounding patient and healthcare variables. *npj Digital Medicine*, 2, 2019.

P. Bandi, O. Geessink, Q. Manson, M. V. Dijk, M. Balkenhol, M. Hermsen, B. E. Bejnordi, B. Lee, K. Paeng, A. Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2018.

A. Barbu, D. Mayo, J. Alverio, W. Luo, C. Wang, D. Gutfreund, J. Tenenbaum, and B. Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 9453–9463, 2019.

P. L. Bartlett and M. H. Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research (JMLR)*, 9(0):1823–1840, 2008.

T. Baumann, A. Köhn, and F. Hennig. The Spoken Wikipedia Corpus collection: Harvesting, alignment and an application to hyperlistening. *Language Resources and Evaluation*, 53(2): 303–329, 2019.

A. H. Beck, A. R. Sangoi, S. Leung, R. J. Marinelli, T. O. Nielsen, M. J. V. D. Vijver, R. B. West, M. V. D. Rijn, and D. Koller. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Science*, 3(108), 2011.

E. Beede, E. Baylor, F. Hersch, A. Iurchenko, L. Wilcox, P. Ruamviboonsuk, and L. M. Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Conference on Human Factors in Computing Systems (CHI)*, pp. 1–12, 2020.

S. Beery, G. V. Horn, and P. Perona. Recognition in terra incognita. In *European Conference on Computer Vision (ECCV)*, pp. 456–473, 2018.

S. Beery, E. Cole, and A. Gjoka. The iwildcam 2020 competition dataset. *arXiv preprint arXiv:2004.10340*, 2020a.

Sara Beery, Dan Morris, and Siyu Yang. Efficient pipeline for camera trap image review. *arXiv preprint arXiv:1907.06772*, 2019.

Sara Beery, Guanhang Wu, Vivek Rathod, Ronny Votel, and Jonathan Huang. Context r-cnn: Long term temporal context for per-camera object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13075–13085, 2020b.

B. E. Bejnordi, M. Veta, P. J. V. Diest, B. V. Ginneken, N. Karssemeijer, G. Litjens, J. A. V. D. Laak, M. Hermsen, Q. F. Manson, M. Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22): 2199–2210, 2017.

D. Bellamy, L. Celi, and A. L. Beam. Evaluating progress on machine learning for longitudinal electronic healthcare data. *arXiv preprint arXiv:2010.01149*, 2020.

S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 137–144, 2006.

A. BenTaieb and G. Hamarneh. Adversarial stain transfer for histopathology image analysis. *IEEE transactions on medical imaging*, 37(3):792–802, 2017.

G. Blanchard, G. Lee, and C. Scott. Generalizing from several related classification tasks to a new unlabeled sample. In *Advances in neural information processing systems*, pp. 2178–2186, 2011.

J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pp. 440–447, 2007.

S. L. Blodgett and B. O'Connor. Racial disparity in natural language processing: A case study of social media African-American English. *arXiv preprint arXiv:1707.00061*, 2017.

S. L. Blodgett, L. Green, and B. O'Connor. Demographic dialectal variation in social media: A case study of African-American English. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1119–1130, 2016.

J. Blumenstock, G. Cadamuro, and R. On. Predicting poverty and wealth from mobile phone metadata. *Science*, 350, 2015.

D. Borkan, L. Dixon, J. Sorensen, N. Thain, and L. Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *WWW*, pp. 491–500, 2019.

L. Bruzzone and M. Marconcini. Domain adaptation problems: A DASVM classification technique and a circular validation strategy. *IEEE transactions on pattern analysis and machine intelligence*, 32(5):770–787, 2009.

D. Bug, S. Schneider, A. Grote, E. Oswald, F. Feuerhake, J. Schüler, and D. Merhof. Context-based normalization of histological stains using deep convolutional features. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 135–142, 2017.

J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pp. 77–91, 2018.

M. Burke, S. Heft-Neal, and E. Bendavid. Sources of variation in under-5 mortality across sub-saharan africa: a spatial analysis. *Lancet Global Health*, 4, 2016.

H. Caesar, V. Bankiti, A. Lang, S. Vora, V. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. nuScenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019.

S. Caldas, P. Wu, T. Li, J. Konečnỳ, H. B. McMahan, V. Smith, and A. Talwalkar. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018.

G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. W. K. Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019.

K. Cao, Y. Chen, J. Lu, N. Arechiga, A. Gaidon, and T. Ma. Heteroskedastic and imbalanced deep learning with adaptive regularization. *arXiv preprint arXiv:2006.15766*, 2020.

I. Y. Chen, P. Szolovits, and M. Ghassemi. Can AI help reduce disparities in general medical and mental health care? *AMA Journal of Ethics*, 21(2):167–179, 2019a.

I. Y. Chen, E. Pierson, S. Rose, S. Joshi, K. Ferryman, and M. Ghassemi. Ethical machine learning in health care. *arXiv preprint arXiv:2009.10576*, 2020.

V. Chen, S. Wu, A. J. Ratner, J. Weng, and C. Ré. Slice-based learning: A programming model for residual learning in critical data slices. In *Advances in neural information processing systems*, pp. 9397–9407, 2019b.

G. Christie, N. Fendley, J. Wilson, and R. Mukherjee. Functional map of the world. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. *Proc. Interspeech*, pp. 1086–1090, 2018.

J. H. Clark, E. Choi, M. Collins, D. Garrette, T. Kwiatkowski, V. Nikolaev, and J. Palomaki. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *arXiv preprint arXiv:2003.05002*, 2020.

N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. 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.

A. Conneau and G. Lample. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 7059–7069, 2019.

A. Conneau, R. Rinott, G. Lample, A. Williams, S. Bowman, H. Schwenk, and V. Stoyanov. Xnli: Evaluating cross-lingual sentence representations. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2475–2485, 2018.

L. P. Cordella, C. D. Stefano, F. Tortorella, and M. Vento. A method for improving classification reliability of multilayer perceptrons. *IEEE Transactions on Neural Networks*, 6(5):1140–1147, 1995.

F. Croce, M. Andriushchenko, V. Sehwag, N. Flammarion, M. Chiang, P. Mittal, and M. Hein. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

Anne-Sophie Crunchant, David Borchers, Hjalmar Kühl, and Alex Piel. Listening and watching: Do camera traps or acoustic sensors more efficiently detect wild chimpanzees in an open habitat? *Methods in Ecology and Evolution*, 11(4):542–552, 2020.

Y. Cui, M. Jia, T. Lin, Y. Song, and S. Belongie. Class-balanced loss based on effective number of samples. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 9268–9277, 2019.

D. Dai and L. Van Gool. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *International Conference on Intelligent Transportation Systems (ITSC)*, 2018.

H. Daumé III. Frustratingly easy domain adaptation. In *Association for Computational Linguistics (ACL)*, 2007.

S. E. Davis, T. A. Lasko, G. Chen, E. D. Siew, and M. E. Matheny. Calibration drift in regression and machine learning models for acute kidney injury. *Journal of the American Medical Informatics Association*, 24(6):1052–1061, 2017.

A. J. DeGrave, J. D. Janizek, and S. Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *medRxiv*, 2020.

M. C. Desmarais and R. Baker. A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, 22(1):9–38, 2012.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Association for Computational Linguistics (ACL)*, pp. 4171–4186, 2019.

N. DigitalGlobe and C. Works. Spacenet. *https://aws.amazon.com/publicdatasets/spacenet/*, 2016.

L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman. Measuring and mitigating unintended bias in text classification. In *Association for the Advancement of Artificial Intelligence (AAAI)*, pp. 67–73, 2018.

Samuel Dodge and Lina Karam. A study and comparison of human and deep learning recognition performance under visual distortions. In *2017 26th international conference on computer communication and networks (ICCCN)*, pp. 1–7. IEEE, 2017.

Q. Dou, D. Castro, K. Kamnitsas, and B. Glocker. Domain generalization via model-agnostic learning of semantic features. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science (ITCS)*, pp. 214–226, 2012.

C. D. Elvidge, P. C. Sutton, T. Ghosh, B. T. Tuttle, K. E. Baugh, B. Bhaduri, and E. Bright. A global poverty map derived from satellite data. *Computers and Geosciences*, 35, 2009.

J. Espey, E. Swanson, S. Badiee, Z. Chistensen, A. Fischer, M. Levy, G. Yetman, A. de Sherbinin, R. Chen, Y. Qiu, G. Greenwell, T. Klein, , J. Jutting, M. Jerven, G. Cameron, A. M. A. Rivera, V. C. Arias, , S. L. Mills, and A. Motivans. Data for development: A needs assessment for SDG monitoring and statistical capacity development. *Sustainable Development Solutions Network*, 2015.

A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

OpenAI et al. Solving Rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

C. Fang, Y. Xu, and D. N. Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *International Conference on Computer Vision (ICCV)*, pp. 1657–1664, 2013.

J. Feng, A. Sondhi, J. Perry, and N. Simon. Selective prediction-set models with coverage guarantees. *arXiv preprint arXiv:1906.05473*, 2019.

D. Filmer and K. Scott. Assessing asset indices. *Demography*, 49, 2011.

Y. Gal and Z. Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, 2016.

Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning (ICML)*, pp. 1180–1189, 2015.

Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research (JMLR)*, 17, 2016.

Y. Geifman and R. El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Y. Geifman and R. El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning (ICML)*, 2019.

Y. Geifman, G. Uziel, and R. El-Yaniv. Bias-reduced uncertainty estimation for deep neural classifiers. In *International Conference on Learning Representations (ICLR)*, 2018.

R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

M. Geva, Y. Goldberg, and J. Berant. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

K. Goel, A. Gu, Y. Li, and C. Ré. Model patching: Closing the subgroup performance gap with data augmentation. *arXiv preprint arXiv:2008.06775*, 2020.

B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 2066–2073, 2012.

I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

N. Graetz, J. Friedman, A. Osgood-Zimmerman, R. Burstein, M. H. Biehl, C. Shields, J. F. Mosser, D. C. Casey, A. Deshpande, L. Earl, R. C. Reiner, S. E. Ray, N. Fullman, A. J. Levine, R. W. Stubbs, B. K. Mayala, J. Longbottom, A. J. Browne, S. Bhatt, D. J. Weiss, P. W. Gething, A. H. Mokdad, S. S. Lim, C. J. L. Murray, E. Gakidou, and S. I. Hay. Mapping local variation in educational attainment across africa. *Nature*, 555, 2018.

M Grooten, T Peterson, and R.E.A Almond. *Living Planet Report 2020 - Bending the curve of biodiversity loss*. WWF, Gland, Switzerland, 2020.

S. Gu, E. Holly, T. Lillicrap, and S. Levine. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *International Conference on Robotics and Automation (ICRA)*, 2017.

I. Gulrajani and D. Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

A. Gupta, A. Murali, D. Gandhi, and L. Pinto. Robot learning in homes: Improving generalization and reducing dataset bias. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.

M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener. Histopathological image analysis: A review. *IEEE reviews in biomedical engineering*, 2:147–171, 2009.

M. C. Hansen, P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice, and J. R. G. Townshend. High-resolution global maps of 21st-century forest cover change. *Science*, 342, 2013.

T. B. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning (ICML)*, 2018.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

B. E. Henderson, N. H. Lee, V. Seewaldt, and H. Shen. The influence of race and ethnicity on the biology of cancer. *Nature Reviews Cancer*, 12(9):648–653, 2012.

D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017.

D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020.

J. Hoffman, E. Tzeng, T. Park, J. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell. Cycada: Cycle consistent adversarial domain adaptation. In *International Conference on Machine Learning (ICML)*, 2018.

D. Hovy and S. L. Spruit. The social impact of natural language processing. In *Association for Computational Linguistics (ACL)*, pp. 591–598, 2016.

J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *arXiv preprint arXiv:2003.11080*, 2020.

G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353, 2016.

N. Jean, S. M. Xie, and S. Ermon. Semi-supervised deep kernel learning: Regression with unlabeled data by minimizing predictive variance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

Jongbin Jung, Sharad Goel, Jennifer Skeem, et al. The limits of human predictions of recidivism. *Science advances*, 6(7):eaaz0652, 2020.

A. K. Jørgensen, D. Hovy, and A. Søgaard. Challenges of studying and processing dialects in social media. In *ACL Workshop on Noisy User-generated Text*, pp. 9–18, 2015.

G. Kahn, P. Abbeel, and S. Levine. BADGR: An autonomous self-supervised learning-based navigation system. *arXiv preprint arXiv:2002.05700*, 2020.

A. Kamath, R. Jia, and P. Liang. Selective question answering under domain shift. In *Association for Computational Linguistics (ACL)*, 2020.

Z. Katona, M. Painter, P. N. Patatoukas, and J. Zeng. On the capital market consequences of alternative data: Evidence from outer space. *Miami Behavioral Finance Conference*, 2018.

D. Kaushik, E. Hovy, and Z. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations (ICLR)*, 2019.

M. Kearns, S. Neel, A. Roth, and Z. S. Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning (ICML)*, pp. 2564–2572, 2018.

J. H. Kim, M. Xie, N. Jean, and S. Ermon. Incorporating spatial context and fine-grained detail from satellite imagery to predict poverty. *Stanford University*, 2016.

N. Kim and T. Linzen. Cogs: A compositional generalization challenge based on semantic interpretation. *arXiv preprint arXiv:2010.05465*, 2020.

A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel. Racial disparities in automated speech recognition. *Science*, 117(14): 7684–7689, 2020.

B. Kompa, J. Snoek, and A. Beam. Empirical frequentist coverage of deep learning uncertainty quantification procedures. *arXiv preprint arXiv:2010.03039*, 2020.

D. Komura and S. Ishikawa. Machine learning methods for histopathological image analysis. *Computational and structural biotechnology journal*, 16:34–42, 2018.

C. E. Kulkarni, R. Socher, M. S. Bernstein, and S. R. Klemmer. Scaling short-answer grading by combining peer assessment with algorithmic scoring. In *Proceedings of the first ACM conference on Learning@Scale conference*, pp. 99–108, 2014.

A. Kumar, T. Ma, and P. Liang. Understanding self-training for gradual domain adaptation. In *International Conference on Machine Learning (ICML)*, 2020.

B. Lake and M. Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning (ICML)*, 2018.

B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020.

Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm. *ProPublica*, 9(1), 2016.

R. Y. Lau, C. Li, and S. S. Liao. Social analytics: Learning fuzzy product ontologies for aspect-oriented sentiment analysis. *Decision Support Systems*, 65:80–94, 2014.

Y. LeCun, C. Cortes, and C. J. Burges. The MNIST database of handwritten digits. *http://yann.lecun.com/exdb/mnist/*, 1998.

J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10), 2010.

D. Li, Y. Yang, Y. Song, and T. M. Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017a.

H. Li, D. Quang, and Y. Guan. Anchor: trans-cell type prediction of transcription factor binding sites. *Genome research*, 29(2):281–292, 2019a.

J. Li, A. H. Miller, S. Chopra, M. Ranzato, and J. Weston. Learning through dialogue interactions by asking questions. In *International Conference on Learning Representations (ICLR)*, 2017b.

T. Li, M. Sanjabi, A. Beirami, and V. Smith. Fair resource allocation in federated learning. *arXiv preprint arXiv:1905.10497*, 2019b.

Y. Li, N. Wang, J. Shi, J. Liu, and X. Hou. Revisiting batch normalization for practical domain adaptation. In *International Conference on Learning Representations Workshop (ICLRW)*, 2017c.

S. Liang, Y. Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Z. Lipton, Y. Wang, and A. Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning (ICML)*, 2018.

Y. Liu, K. Gadepalli, M. Norouzi, G. E. Dahl, T. Kohlberger, A. Boyko, S. Venugopalan, A. Timofeev, P. Q. Nelson, G. S. Corrado, et al. Detecting cancer metastases on gigapixel pathology images. *arXiv preprint arXiv:1703.02442*, 2017.

M. Long, Y. Cao, J. Wang, and M. Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pp. 97–105, 2015.

M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt, and N. E. Thomas. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 1107–1110, 2009.

M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19:313–330, 1993.

R. T. McCoy, J. Min, and T. Linzen. Berts of a feather do not generalize together: Large variability in generalization across models with similar test set performance. *arXiv preprint arXiv:1911.02969*, 2019a.

R. T. McCoy, E. Pavlick, and T. Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Association for Computational Linguistics (ACL)*, 2019b.

S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. C. Corrado, A. Darzi, et al. International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788):89–94, 2020.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

P. Mirowski, R. Pascanu, F. Viola, H. Soyer, A. Ballard, A. Banino, M. Denil, R. Goroshin, L. Sifre, C. Kavukcuoglu, D. Kumaran, and R. Hadsell. Learning to navigate in complex environments. In *International Conference on Learning Representations (ICLR)*, 2017.

K. Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning (ICML)*, pp. 10–18, 2013.

W. Nekoto, V. Marivate, T. Matsila, T. Fasubaa, T. Kolawole, T. Fagbohungbe, S. O. Akinola, S. H. Muhammad, S. Kabongo, S. Osei, S. Freshia, R. A. Niyongabo, R. Macharm, P. Ogayo, O. Ahia, M. Meressa, M. Adeyemi, M. Mokgesi-Selinga, L. Okegbemi, L. J. Martinus, K. Tajudeen, K. Degila, K. Ogueji, K. Siminyu, J. Kreutzer, J. Webster, J. T. Ali, J. Abbott, I. Orife, I. Ezeani, I. A. Dangana, H. Kamper, H. Elsahar, G. Duru, G. Kioko, E. Murhabazi, E. van Biljon, D. Whitenack, C. Onyefuluchi, C. Emezue, B. Dossou, B. Sibanda, B. I. Bassey, A. Olabiyi, A. Ramkilowan, A. Öktem, A. Akinfaderin, and A. Bashir. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of Empirical Methods in Natural Language Processing (Findings of EMNLP)*, 2020.

B. Nestor, M. McDermott, W. Boag, G. Berner, T. Naumann, M. C. Hughes, A. Goldenberg, and M. Ghassemi. Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. *arXiv preprint arXiv:1908.00690*, 2019.

J. Ni, J. Li, and J. McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 188–197, 2019.

A. Noor, V. Alegana, P. Gething, A. Tatem, and R. Snow. Using remotely sensed night-time light as a proxy for poverty in africa. *Population Health Metrics*, 6, 2008.

Mohammad Sadegh Norouzzadeh, Dan Morris, Sara Beery, Neel Joshi, Nebojsa Jojic, and Jeff Clune. A deep active learning system for species identification and counting in camera trap images. *arXiv preprint arXiv:1910.09716*, 2019.

NYTimes. The Times is partnering with Jigsaw to expand comment capabilities. *The New York Times*, 2016. URL `https://www.nytco.com/press/the-times-is-partnering-with-jigsaw-to-expand-comment-capabilities/`.

Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Y. Oren, S. Sagawa, T. Hashimoto, and P. Liang. Distributionally robust language modeling. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

A. Osgood-Zimmerman, A. I. Millear, R. W. Stubbs, C. Shields, B. V. Pickering, L. Earl, N. Graetz, D. K. Kinyoki, S. E. Ray, S. Bhatt, A. J. Browne, R. Burstein, E. Cameron, D. C. Casey, A. Deshpande, N. Fullman, P. W. Gething, H. S. Gibson, N. J. Henry, M. Herrero, L. K. Krause, I. D. Letourneau, A. J. Levine, P. Y. Liu, J. Longbottom, B. K. Mayala, J. F. Mosser, A. M. Noor, D. M. Pigott, E. G. Piwoz, P. Rao, R. Rawat, R. C. Reiner, D. L. Smith, D. J. Weiss, K. E. Wiens, A. H. Mokdad, S. S. Lim, C. J. L. Murray, N. J. Kassebaum, and S. I. Hay. Mapping child growth failure in africa between 2000 and 2015. *Nature*, 555, 2018.

Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

S. J. Pan, X. Ni, J. Sun, Q. Yang, and Z. Chen. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th international conference on World wide web*, pp. 751–760, 2010.

V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5206–5210, 2015.

J. H. Park, J. Shin, and P. Fung. Reducing gender bias in abusive language detection. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2799–2804, 2018.

G. K. Patro, A. Biswas, N. Ganguly, K. P. Gummadi, and A. Chakraborty. Fairrec: Two-sided fairness for personalized recommendations in two-sided platforms. In *Proceedings of The Web Conference 2020*, pp. 1194–1204, 2020.

X. Peng, B. Usman, N. Kaushik, D. Wang, J. Hoffman, and K. Saenko. Visda: A synthetic-to-real benchmark for visual domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 2021–2026, 2018.

X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang. Moment matching for multi-source domain adaptation. In *International Conference on Computer Vision (ICCV)*, 2019.

X. Peng, E. Coumans, T. Zhang, T. Lee, J. Tan, and S. Levine. Learning agile robotic locomotion skills by imitating animals. In *Robotics: Science and Systems (RSS)*, 2020.

L. Perelman. When "the state of the art" is counting words. *Assessing Writing*, 21:104–111, 2014.

N. A. Phillips, P. Rajpurkar, M. Sabini, R. Krishnan, S. Zhou, A. Pareek, N. M. Phu, C. Wang, A. Y. Ng, and M. P. Lungren. Chexphoto: 10,000+ smartphone photos and synthetic photographic transformations of chest x-rays for benchmarking deep learning robustness. *arXiv preprint arXiv:2007.06199*, 2020.

C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller. Tuned models of peer assessment in moocs. *Educational Data Mining*, 2013.

M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014.

Kerrie A Pipal, Jeremy J Notch, Sean A Hayes, and Peter B Adams. Estimating escapement for a low-abundance steelhead population using dual-frequency identification sonar (didson). *North American Journal of Fisheries Management*, 32(5):880–893, 2012.

W Nicholson Price and I Glenn Cohen. Privacy in the age of medical big data. *Nature medicine*, 25 (1):37–43, 2019.

C. Ré, F. Niu, P. Gudipati, and C. Srisuwananukorn. Overton: A data system for monitoring and improving machine-learned products. *arXiv preprint arXiv:1909.05372*, 2019.

R. C. Reiner, N. Graetz, D. C. Casey, C. Troeger, G. M. Garcia, J. F. Mosser, A. Deshpande, S. J. Swartz, S. E. Ray, B. F. Blacker, P. C. Rao, A. Osgood-Zimmerman, R. Burstein, D. M. Pigott, I. M. Davis, I. D. Letourneau, L. Earl, J. M. Ross, I. A. Khalil, T. H. Farag, O. J. Brady, M. U. Kraemer, D. L. Smith, S. Bhatt, D. J. Weiss, P. W. Gething, N. J. Kassebaum, A. H. Mokdad, C. J. Murray, and S. I. Hay. Variation in childhood diarrheal morbidity and mortality in africa, 2000–2015. *New England Journal of Medicine*, 379, 2018.

M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Association for Computational Linguistics (ACL)*, pp. 4902–4912, 2020.

S. R. Richter, V. Vineet, S. Roth, and V. Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pp. 102–118, 2016.

G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3234–3243, 2016.

Amir Rosenfeld, Richard Zemel, and John K Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018.

F. Sadeghi and S. Levine. CAD2RL: Real single-image flight without a single real image. In *Robotics: Science and Systems (RSS)*, 2017.

K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pp. 213–226, 2010.

S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

D. E. Sahn and D. Stifel. Exploring alternative measures of welfare in the absence of expenditure data. *The Review of Income and Wealth*, 49, 2003.

S. Santurkar, D. Tsipras, and A. Madry. Breeds: Benchmarks for subpopulation shift. *arXiv*, 2020.

Stefan Schneider and Alex Zhuang. Counting fish and dolphins in sonar images using deep learning. *arXiv preprint arXiv:2007.12808*, 2020.

L. Seyyed-Kalantari, G. Liu, M. McDermott, and M. Ghassemi. Chexclusion: Fairness gaps in deep chest X-ray classifiers. *arXiv preprint arXiv:2003.00827*, 2020.

V. Shankar, A. Dave, R. Roelofs, D. Ramanan, B. Recht, and L. Schmidt. Do image classifiers generalize across time? *arXiv preprint arXiv:1906.02168*, 2019.

J. Shen, Y. Qu, W. Zhang, and Y. Yu. Wasserstein distance guided representation learning for domain adaptation. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018.

M. D. Shermis. State-of-the-art automated essay scoring: Competition, results, and future directions from a united states demonstration. *Assessing Writing*, 20:53–76, 2014.

Rakshith Shetty, Bernt Schiele, and Mario Fritz. Not using the car to see the sidewalk–quantifying and controlling the effects of context in classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8218–8226, 2019.

H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90:227–244, 2000.

Yu Shiu, KJ Palmer, Marie A Roch, Erica Fleishman, Xiaobai Liu, Eva-Marie Nosal, Tyler Helble, Danielle Cholewiak, Douglas Gillespie, and Holger Klinck. Deep neural networks for automated detection of marine mammal species. *Scientific Reports*, 10(1):1–12, 2020.

N. Sohoni, J. Dunnmon, G. Angus, A. Gu, and C. Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Dan Stowell, Michael D Wood, Hanna Pamuła, Yannis Stylianou, and Hervé Glotin. Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge. *Methods in Ecology and Evolution*, 10(3):368–380, 2019.

B. Sun and K. Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pp. 443–450, 2016.

B. Sun, J. Feng, and K. Saenko. Return of frustratingly easy domain adaptation. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2016.

Y. Sun, X. Wang, Z. Liu, J. Miller, A. A. Efros, and M. Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning (ICML)*, 2020.

Michael A Tabak, Mohammad S Norouzzadeh, David W Wolfson, Steven J Sweeney, Kurt C VerCauteren, Nathan P Snow, Joseph M Halseth, Paul A Di Salvo, Jesse S Lewis, Michael D White, et al. Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology and Evolution*, 10(4):585–590, 2019.

K. Taghipour and H. T. Ng. A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 1882–1891, 2016.

R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt. Measuring robustness to natural distribution shifts in image classification. *arXiv preprint arXiv:2007.00644*, 2020.

R. Tatman. Gender and dialect bias in YouTube's automatic captions. In *Workshop on Ethics in Natural Langauge Processing*, volume 1, pp. 53–59, 2017.

D. Tellez, M. Balkenhol, I. Otte-Höller, R. van de Loo, R. Vogels, P. Bult, C. Wauters, W. Vreuls, S. Mol, N. Karssemeijer, et al. Whole-slide mitosis detection in h&e breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks. *IEEE transactions on medical imaging*, 37(9):2126–2136, 2018.

D. Tellez, G. Litjens, P. Bándi, W. Bulten, J. Bokhorst, F. Ciompi, and J. van der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical image analysis*, 58, 2019.

D. Temel, G. Kwon, M. Prabhushankar, and G. AlRegib. CURE-TSR: Challenging unreal and real environments for traffic sign recognition. *arXiv preprint arXiv:1712.02463*, 2017.

Dogancan Temel, Jinsol Lee, and Ghassan AlRegib. Cure-or: Challenging unreal and real environments for object recognition. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 137–144. IEEE, 2018.

T. G. Tiecke, X. Liu, A. Zhang, A. Gros, N. Li, G. Yetman, T. Kilic, S. Murray, B. Blankespoor, E. B. Prydz, and H. H. Dang. Mapping the world population one building at a time. *arXiv*, 2017.

J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *International Conference on Intelligent Robots and Systems (IROS)*, 2017.

A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 1521–1528, 2011.

E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

B. Uzkent and S. Ermon. Learning when and where to zoom with deep reinforcement learning. In *Computer Vision and Pattern Recognition (CVPR)*, 2020.

Sindre Vatnehol, Hector Peña, and Nils Olav Handegard. A method to automatically detect fish aggregations using horizontally scanning sonar. *ICES Journal of Marine Science*, 75(5):1803–1812, 2018.

B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling. Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 210–218, 2018.

H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 5018–5027, 2017.

M. Veta, P. J. V. Diest, M. Jiwa, S. Al-Janabi, and J. P. Pluim. Mitosis counting in breast cancer: Object-level interobserver agreement and comparison to an automatic method. *PloS one*, 11(8), 2016.

M. Veta, Y. J. Heng, N. Stathonikos, B. E. Bejnordi, F. Beca, T. Wollmann, K. Rohr, M. A. Shah, D. Wang, M. Rousson, et al. Predicting breast tumor proliferation from whole-slide images: the tupac16 challenge. *Medical image analysis*, 54:111–121, 2019.

R. Volpi, H. Namkoong, O. Sener, J. Duchi, V. Murino, and S. Savarese. Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell. Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020a.

S. Wang, M. Bai, G. Mattyus, H. Chu, W. Luo, B. Yang, J. Liang, J. Cheverie, S. Fidler, and R. Urtasun. Torontocity: Seeing the world with a million eyes. In *International Conference on Computer Vision (ICCV)*, 2017.

S. Wang, W. Chen, S. M. Xie, G. Azzari, and D. B. Lobell. Weakly supervised deep learning for segmentation of remote sensing imagery. *Remote Sensing*, 12, 2020b.

OR Wearn and P Glover-Kapfer. Camera-trapping for conservation: a guide to best-practices. *WWF conservation technology series*, 1(1):2019–04, 2017.

S. Weinberger. Speech accent archive. *George Mason University*, 2015.

Ben G Weinstein. A computer vision for animal ecology. *Journal of Animal Ecology*, 87(3):533–545, 2018.

J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, C. G. A. R. Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10), 2013.

R. West, H. S. Paskov, J. Leskovec, and C. Potts. Exploiting social network structure for person-to-person sentiment analysis. *Transactions of the Association for Computational Linguistics (TACL)*, 2:297–310, 2014.

J. J. Williams, J. Kim, A. Rafferty, S. Maldonado, K. Z. Gajos, W. S. Lasecki, and N. Heffernan. Axis: Generating explanations at scale with learnersourcing and machine learning. In *Proceedings of the Third (2016) ACM Conference on Learning@Scale*, pp. 379–388, 2016.

Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. Predictive inequity in object detection. *arXiv preprint arXiv:1902.11097*, 2019.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew. HuggingFace's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Ho Yuen Frank Wong, Hiu Yin Sonia Lam, Ambrose Ho-Tung Fong, Siu Ting Leung, Thomas Wing-Yan Chin, Christine Shing Yen Lo, Macy Mei-Sze Lui, Jonan Chun Yin Lee, Keith Wan-Hang Chiu, Tom Chung, et al. Frequency and distribution of chest radiographic findings in covid-19 positive patients. *Radiology*, pp. 201160, 2020.

D. E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow. Harmonic networks: Deep translation and rotation equivariance. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 5028–5037, 2017.

M. Wu, M. Mosse, N. Goodman, and C. Piech. Zero shot learning for code education: Rubric sampling with deep learning inference. In *Association for the Advancement of Artificial Intelligence (AAAI)*, volume 33, pp. 782–790, 2019a.

M. Wu, R. L. Davis, B. W. Domingue, C. Piech, and N. Goodman. Variational item response theory: Fast, accurate, and expressive. *International Conference on Educational Data Mining*, 2020.

Y. Wu, E. Winston, D. Kaushik, and Z. Lipton. Domain adaptation with asymmetrically-relaxed distribution alignment. In *International Conference on Machine Learning (ICML)*, pp. 6872–6881, 2019b.

M. Wulfmeier, A. Bewley, and I. Posner. Incremental adversarial domain adaptation for continually changing environments. In *International Conference on Robotics and Automation (ICRA)*, 2018.

K. Xiao, L. Engstrom, A. Ilyas, and A. Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020.

M. Xie, N. Jean, M. Burke, D. Lobell, and S. Ermon. Transfer learning from deep features for remote sensing and poverty mapping. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2016.

S. M. Xie, A. Kumar, R. Jones, F. Khani, T. Ma, and P. Liang. In-N-out: Pre-training and self-training using auxiliary information for out-of-distribution robustness. *arXiv*, 2020.

Y. Yang and S. Newsam. Bag-of-visual-words and spatial extensions for land-use classification. *Geographic Information Systems*, 2010.

Y. Yang, K. Caluwaerts, A. Iscen, T. Zhang, J. Tan, and V. Sindhwani. Data efficient reinforcement learning for legged robots. In *Conference on Robot Learning (CoRL)*, 2019.

C. Yeh, A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, and M. Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature Communications*, 11, 2020.

J. You, X. Li, M. Low, D. Lobell, and S. Ermon. Deep gaussian process for crop yield prediction based on remote sensing data. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2017.

F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

N. Yuval, W. Tao, C. Adam, B. Alessandro, W. Bo, and N. A. Y. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. In *PLOS Medicine*, 2018.

M. Zhang, H. Marklund, N. Dhawan, A. Gupta, S. Levine, and C. Finn. Adaptive risk minimization: A meta-learning approach for tackling group shift. *arXiv preprint arXiv:2007.02931*, 2020.

Y. Zhang, J. Baldridge, and L. He. Paws: Paraphrase adversaries from word scrambling. In *North American Association for Computational Linguistics (NAACL)*, 2019.

X. Zhou, Y. Nie, H. Tan, and M. Bansal. The curse of performance instability in analysis datasets: Consequences, source, and suggestions. *arXiv preprint arXiv:2004.13606*, 2020.

A Datasets

A.1 CivilComments-wilds

Additional dataset details. The CivilComments-wilds dataset comprises comments from a large set of articles from the Civil Comments platform, annotated for toxicity and demographic identities (Borkan et al., 2019). We partitioned the articles into disjoint training, validation, and test splits, and then formed the corresponding datasets by taking all comments on the articles in those splits. In total, the training set comprised 269,038 comments (60% of the data); the validation set comprised 45,180 comments (10%); and the test set comprised 133,782 (30%).

Differences from the original dataset. The original dataset[5] also had a training and test split with disjoint articles. These splits are related to ours in the following way. Let the number of articles in the original test split be m. To form our validation split, we took m articles (sampled uniformly at random) from the original training split, and to form our test split, we took $2m$ articles (also sampled uniformly at random) from the original training split and added it to the existing test split. We added a fixed validation set to allow other researchers to be able to compare methods more consistently, and we tripled the size of the test set to allow for more accurate worst-group accuracy measurement.

Similarly, we combined some of the demographic identities in the original dataset to obtain larger groups (for which we could more accurately estimate accuracy). Specifically, we created an aggregate *LGBTQ* identity that combines the original *homosexual_gay_or_lesbian*, *bisexual*, *other_sexual_orientation*, *transgender*, and *other_gender* identities (e.g., it is 1 if any of those identities are 1), and an aggregate *other_religions* identity that combines the original *jewish*, *hindu*, *buddhist*, *atheist*, and *other_religion* identities. We also omitted the *psychiatric_or_mental_illness* identity, which was evaluated in the original Kaggle competition, because of a lack of sufficient data for accurate estimation; but we note that baseline group accuracies for that identity seemed higher than for the other groups, so it is unlikely to factor into worst-group accuracy. In our new split, each identity we evaluate on (*male*, *female*, *LGBTQ*, *Christian*, *Muslim*, *other_religion*, *Black*, and *White*) has at least 500 positive and 500 negative examples. We also have an *identity_any* identity that we use in one of the baseline models; this identity combines all of the identities in the original dataset, including *psychiatric_or_mental_illness* and related identities.

The evaluation metric used in the original competition was a complex weighted combination of various metrics, including subgroup AUCs for each demographic identity, and a new pinned AUC metric introduced by the original authors (Borkan et al., 2019); conceptually, these metrics also measure the degree to which model accuracy is uniform across the different identities. After discussion with the original authors, we replace the composite metric with worst-group accuracy for simplicity and consistency with other datasets.

Baseline model details. Our baseline models are all fine-tuned BERT-base-uncased models, using the implementation from Wolf et al. (2019), and with the following hyperparameter settings: batch size 16; learning rate 10^{-5} using the AdamW optimizer; L_2-regularization strength 10^{-2}; and a maximum number of tokens of 300 (99.95% of the data had fewer than or equal to 300 tokens). We chose these hyperparameters through a brief grid search (with batch sizes of 16, 24, 32, with a correspondingly smaller maximum number of tokens, so as to fit in memory; learning rates 10^{-6},

[5] `www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/`

Model	Average accuracy	Worst-group accuracy
Standard	**0.93**	0.54
Reweighted (class balance)	0.87	0.56
Group DRO (class balance)	0.86	0.52
Reweighted (class $\times$ any identity)	0.87	0.60
Group DRO (class $\times$ any identity)	0.88	**0.71**
Reweighted (all combinations of class/identities)	0.87	0.63
Group DRO (all combinations of class/identities)	0.86	0.60

Table 10: Baseline results on CIVILCOMMENTS-WILDS.

10^{-5}, 2×10^{-5}; and L_2-regularization strength 0, 10^{-3}, 10^{-2}), picking the hyperparameters with the best worst-group validation accuracy, and using a standard model fine-tuned through empirical risk minimization. However, in general, we found little difference between these hyperparameter settings in terms of average and worst-group accuracies on the training or validation set. We used the same hyperparameter settings for all baselines. For all baselines, we also do early stopping on worst-group validation accuracy, which in practice means stopping after the first epoch; this is consistent with previous findings (Sagawa et al., 2020).

We also run reweighted (Shimodaira, 2000) and group DRO baselines (Sagawa et al., 2020). These involve partitioning the training data into disjoint subsets. We study several different choices of subsets, corresponding to different rows in Table 10:

1. *Class balance*: 2 subsets, 1 for each class.
2. *Class $\times$ any identity*: 4 subsets, 1 for each combination of class and *identity_any*.
3. *All combinations of class and identities*: $2^9 = 512$ subsets, 1 for each combination of class and the 8 identities.

Full results are in Table 10.

Additional data sources. All of the data, including the data with identity annotations that we use and the data with just label annotations, are also annotated for additional toxicity subtype attributes, specifically *severe_toxicity*, *obscene*, *threat*, *insult*, *identity_attack*, and *sexual_explicit*. These annotations can be used to train models that are more aware of the different ways that a comment can be toxic; in particular, learning from the *identity_attack* attribute which comments are toxic because of the use of identities might help the model learn how to avoid spurious associations between toxicity and identity.

Additional discussion. Measuring worst-group accuracy treats false positives and false negatives equally; in deployment systems, one might want to weight these differently, e.g., using cost-sensitive learning or by simply raising or lowering the classification threshold. One could also binarize the labels and identities differently: in this benchmark, we simply use majority voting from the annotators.

In practice, models might do poorly on intersections of groups (Kearns et al., 2018), e.g., on comments that mention multiple identities. Given the size of the dataset and comparative rarity of some identities and of toxic comments in general, accuracies on these intersections are difficult to estimate from this dataset. A potential avenue of future work is to develop methods for evaluating models on such subgroups, e.g., by generating data in particular groups through templates (Park et al., 2018; Ribeiro et al., 2020).

A.2 AMAZON-WILDS

Additional setup The input is a review text with a maximum token length of 512, and the label is the star rating out of 5 with $\mathcal{Y} = \{1, 2, 3, 4, 5\}$. For each example, the following additional metadata is available at both training and evaluation time: reviewer ID, product ID, product category, review time, and summary. At test time, we provide unlabeled examples for each reviewer for test-time adaptation.

Additional dataset details. We consider a subset of the Amazon reviews dataset (Ni et al., 2019). We consider disjoint reviewers between the training set and the official validation and test sets, but we

also provide separate validation and test sets for reviewers seen during training for additional baseline experiments. These reviewers are selected uniformly at random from the reviewer pool, with the constraint that they have at least 150 reviews in the pre-processed dataset. Statistics for each split are described in Table 13. Notably, each reviewer has at least 75 reviews in the training set and exactly 75 reviews in the validation and test sets. For more details on pre-processing and subset selection, see the below subsection of pre-processing.

Split	# Reviews	# Reviewers	# Reviews per reviewer (mean / minimum / maximum)
Train	1,000,182	4,974	201 / 75 / 3,198
Validation (unseen)	100,050	1,334	75 / 75 / 75
Test (unseen)	100,050	1,334	75 / 75 / 75
Validation (seen)	100,050	1,334	75 / 75 / 75
Test (seen)	100,050	1,334	75 / 75 / 75

Table 11: Dataset details for AMAZON-WILDS.

Modifications from the original dataset. The original dataset does not consider a task nor a split. We consider a standard task of sentiment classification, but we depart from a standard split, considering disjoint users between training and evaluation time as described above. In addition, we pre-process the data as detailed below.

Pre-processing and subset selection. We first eliminate reviews that are longer than 512 tokens, reviews without any text, and any duplicate reviews with identical star rating, reviewer ID, product ID, and time. We then obtain the 30-core subset of the reviews, which contains the maximal set of reviewers and products each of which have at least 30 reviews; this is a standard pre-processing procedure for the original dataset (Ni et al., 2019). To construct the dataset for reviewer shifts in particular, we further eliminate the following reviews: (i) reviews that contain HTML, (ii) reviews with identical text within a user in order to ensure sufficiently high effective sample size per reviewer, and (iii) reviews with identical text across users to eliminate generic reviews. Once we have the filtered set of reviews, we consider reviewers with at least 150 reviews and sample uniformly at random until the training set contains 1 million reviews and each evaluation set contains at least 100,000 reviews. As we construct the training set, we reserve a random sample of 75 reviews for each user for evaluation and put all other reviews to the training set. For evaluation set, we put a random sample of 75 reviews for each user.

Additional baseline results. We report additional results on the baseline models in Table 12, reporting the worst-group accuracy on unseen users as well as accuracies on seen users. We observe performance disparities across seen reviewers as well, and the performance gaps between seen and unseen reviewers are small across various metrics for our baseline models. We note that while group DRO successfully improves worst-group accuracy on seen users as advertised, it fails to significantly improve the wort-group accuracy for unseen users as well as the 10th percentile accuracies on seen and unseen users.

Model	Worst-group, unseen	Average, seen	10th percentile, seen	Worst-group, seen
Standard BERT	9.3	75.8	58.7	17.3
Reweighted BERT (class balance)	12.0	72.4	56.0	17.3
Group DRO BERT (reviewer)	14.7	73.5	58.7	29.3

Table 12: Additional test accuracies of baseline models on AMAZON-WILDS.

Baseline model details. For all baseline experiments, we fine-tune BERT-base-uncased models, using the implementation from Wolf et al. (2019), and with the following hyperparameter settings: batch size 8; learning rate 5×10^{-6}; L_2-regularization strength 0; 3 epochs; and a maximum number of tokens of 512. We select the above hyperparameters based on a grid search, considering L_2-regularization strengths 0, 10^{-2} and learning rates $5 \times 10^{-6}, 1 \times 10^{-5}, 2 \times 10^{-5}, 5 \times 10^{-5}$, and epochs 1-3. We select the hyperparameters that yield best average accuracy on the validation set.

A.3 CAMELYON17-WILDS

Additional dataset details. The CAMELYON17-WILDS dataset is adapted from whole-slide images (WSIs) of breast cancer metastases in lymph nodes sections, obtained from the CAMELYON17 challenge (Bandi et al., 2018). The training set comprises 303,821 patches taken from 36 WSIs, with 9 WSIs from each of the 4 hospitals in the training set. The validation set comprises 67,078 patches taken from 4 WSIs, with 1 WSI from each of the same 4 hospitals. These WSIs were selected by taking, for each hospital, the WSI whose number of patches was closest to 20% of the total number of patches from the 10 WSIs in that hospital. The test set comprises 85,054 patches taken from 10 WSIs from a 5th hospital that is not represented in the training or validation sets. We selected the test set hospital as the one that led to the largest out-of-distribution accuracy drop; the other hospitals were more similar to each other and did not display large out-of-distribution drops when used as the test set.

From these WSIs, we extracted patches in a standard manner. The WSIs were scanned at a resolution of 0.23μm–0.25μm in the original dataset, and each WSI contains multiple resolution levels, with approximately 10,000×20,000 pixels at the highest resolution level (Bandi et al., 2018). We used the third-highest resolution level, corresponding to reducing the size of each dimension by a factor of 4. We then tiled each slide with overlapping 96×96 pixel patches with a step size of 32 pixels in each direction (such that none of the central 32×32 regions overlap), labeling them as the following:

- *Tumor* patches have at least one pixel of tumor tissue in the central 32×32 region. We used the pathologist-annotated tumor annotations provided with the WSIs.
- *Normal* patches have no tumor and have at least 20% normal tissue in the central 32×32 region. We used Otsu thresholding to distinguish normal tissue from background.

We discarded all patches that had no tumor and <20% normal tissue in the central 32×32 region.

To maintain an equal class balance, we then subsampled the extracted patches in the following way. First, for each WSI, we kept all tumor patches unless the WSI had fewer normal than tumor patches, which was the case for a single WSI; in that case, we randomly discarded tumor patches from that WSI until the numbers of tumor and normal patches were equal. Then, for each training hospital, we picked the WSI whose number of tumor patches was closest to 20% of the total number of tumor patches from that hospital to be part of the validation set. Finally, we randomly selected normal patches for inclusion such that for each hospital and split, there was an equal number of tumor and normal patches.

Differences from the original dataset. The task in the original CAMELYON17 challenge was the patient-level classification task of determining the pathologic lymph node stage of the tumor present in all slides from a patient. In contrast, our task is a lesion-level classification task. Patient-level, slide-level, and lesion-level tasks are all common in histopathology applications. As mentioned above, the original dataset provided WSIs and tumor annotations, but not a standardized set of patches or data splits, which we provide here.

The CAMELYON17-WILDS patch-based dataset is similar to one of the datasets used in Tellez et al. (2019), which was also derived from the CAMELYON17 challenge; there, only one hospital is used as the training set, and the other hospitals are all part of the test set. CAMELYON17-WILDS is also similar to PCam (Veeling et al., 2018), which is a patch-based dataset based on an earlier CAMELYON16 challenge; the data there is derived from only two hospitals.

Baseline model details. Our baseline models are DenseNet-121 (Huang et al., 2017) models pre-trained on ImageNet with empirical risk minimization (ERM). We ran a hyperparameter grid search with the learning rates 10^{-4}, 10^{-3}, 10^{-2}, and L_2-regularization strengths 0, 10^{-3}, 10^{-2}, with 3 random seeds per hyperparameter setting. Optimization was done through stochastic gradient descent with momentum (set to 0.9). We estimated the variance in accuracy due to random seeds by computing the sample variance across random seeds for each hyperparameter setting, and then taking its average; to avoid inflating the variance, we discarded the hyperparameter setting with learning rate 10^{-2} and L_2-regularization strength 10^{-2} as it had substantially worse validation accuracy (around 80-85%) and larger variance. We observed that models with the lowest learning rate (10^{-4}) generally had high test accuracy, although they were hard to distinguish from models with higher learning rates based on just the validation accuracy.

For the group DRO models, we used the same hyperparameter grid search, treating the patches from each hospital as one group (i.e., 4 groups in the training set and 4 groups in the validation set). In Table 5 in the main text, we report average validation accuracy for consistency with the standard ERM models. However, using worst-group validation accuracy for model selection does similarly, with an equally-large spread of test accuracies.

Additional data sources. The full, original CAMELYON17 dataset contains 1000 WSIs from the same 5 hospitals, although only 50 of them (which we use here) have tumor annotations. The other 950 WSIs may be used as unlabeled data. Beyond the CAMELYON17 dataset, the largest source of unlabeled WSI data is the Cancer Genome Atlas (Weinstein et al., 2013), which typically has patient-level annotations (e.g., patient demographics and clinical outcomes).

Additional discussion. Many specialized methods have been developed to handle stain variation in the context of digital histopathology. These typically fall into one of two categories: data augmentation methods that perturb the colors in the training images (e.g., Liu et al. (2017); Bug et al. (2017); Tellez et al. (2018)) or stain normalization methods that seek to standardize colors across training images (e.g., Macenko et al. (2009); BenTaieb & Hamarneh (2017)). These methods are reasonably effective at mitigating stain variation, at least in some contexts (Tellez et al., 2019), though the general problem of learning digital histopathology models that can be effectively deployed across multiple hospitals/sites is still open.

Beyond stain variation, there are many other distribution shifts that might occur in histopathology applications. For example, as with all medical applications, patient demographics might differ from hospital to hospital, e.g., some hospitals might tend to see patients who are older or more sick, and patients from different backgrounds and countries vary in terms of cancer susceptibility (Henderson et al., 2012). Some cancer subtypes and tissues of origin are also more common than others, leading to potential subpopulation shift issues, e.g,. a rare cancer subtype in one context might be more common in another; or even if it remains rare, we would seek to leverage the greater quantity of data from other subtypes to improve model accuracy on the rare subtype (Weinstein et al., 2013).

A.4 iWildCam2020-wilds

Additional dataset details. We generate the splits in three steps. First, to generate the trans-splits, we randomly split all locations into three groups, remaining, val-trans and test-trans. Then, to generate the cis-split, we split "remaining" by date into three groups: train, val-cis, and test-cis. When doing the cis split, according to date, some locations only ended up in some of but not all of train, val-cis, and test-cis. For instance, if there were very few dates for a specific location, it may be that no examples from that location ended up in the train split. This defeats the purpose of the cis split, which is to test performance on locations that were seen during training. Thus, these locations were removed. Finally, any images in the test set with classes not present in the train set were also removed.

Split	# Examples	# Locations
Train	151906	211
Validation-cis	8442	211
Test-cis	8201	211
Validation-trans	20230	32
Test-trans	29133	48

Table 13: Dataset details for iWildCam2020-wilds.

Modifications from the original dataset. In the competition on Kaggle there is a held-out test set that we are not utilizing since it is not public. Instead we constructed our own test set, by splitting the kaggle competition "train data" into our split: train, validation-cis, validation-trans, test-cis, test-trans. Moreover, as we mentioned before, we leave out 49 locations that did not span cis splits. Images are organized into sequences, but we treat each image separately. In the iWildCam 2020 competition, the top participants utilize the sequence data and also use a pretrained animal detection model that outputs bounding boxes over the animals. The images are cropped and fed into a classification network.

Baseline model details. We train a Densenet-121 with batch size 32 for 10 epochs. The optimizer is Adam and the learning rate is 5e-4. For the upsampling baseline, we weight the samples by the inverse frequency of the class. In ARM, each minibatch of size 16, contains only examples from one location, to which the model adapts. In this case, we use ARM Batch Norm, where the batch statistics are used to adapt to the location. Support size is 16, and meta batch size is 2.

Additional discussion Beery et al. (2018) studied shifts in camera traps with the Caltech Camera Traps-20 dataset (CCT-20) and showed that there were considerable performance drops when evaluating the model on held-out test locations. They identified several classification challenges arising from camera trap images, including poor illumination, motion blur, and size of the animal in the image, which could account for the large drop in performance. They show that making use of an animal detection model that first locates the animal, followed by classification can significantly reduce the generalization gap. However, this requires the collection of bounding box annotations, either using an already trained animal detection model or having humans annotate the images, which is costly, and non-trivial. Since there is already much data with image-level species labels, whilst not nearly as much data for training animal detection models, it would be useful to be able to train models that directly classify images without relying on the intermediate step of first detecting the animal. It has also been shown that utilizing the temporal signal, for instance, taking the median prediction across a burst of images captured for a single motion trigger can reduce the gap.

A.5 POVERTYMAP-WILDS

The POVERTYMAP-WILDS dataset is derived from Yeh et al. (2020), which gathers LandSat imagery and Demographic and Health Surveys (DHS) data from 19669 villages in Africa across 23 countries. The images are 224×224 pixels large over 7 multispectral channels, with an optional eighth nighttime light intensity channel. The LandSat satellite has a 30m resolution, meaning that each pixel of the image covers a $30m^2$ spatial area. The location metadata is perturbed by the DHS as a privacy protection scheme; urban locations are randomly displaced by up to 2km and rural locations are perturbed by up to 10km. This adds some noise to the data, but having a large enough image can guarantee that the location is in the image most of the time. The target is a composite asset wealth index computed as the first principal component of survey responses about household assets, which is thought to be a less noisy measure of households' longer-run economic well-being than other welfare measurements such as consumption expenditure (Sahn & Stifel, 2003; Filmer & Scott, 2011). Asset wealth also has the advantage of not requiring adjustments for inflation or for purchasing power parity (PPP), as it is not based on a currency.

Model. Following (Yeh et al., 2020), we use a ResNet-18 model (He et al., 2016) trained with the Adam optimizer and mean squared-error loss function. We use almost all the same hyperparameters, including a batch size of 64 and decaying the learning rate by a factor of 0.96 after each epoch (initial learning rate 1e-3). We train for 200 epochs with early stopping on an in-distribution development set.

We also experimented with naively including the nighttime light imagery (from a separate DMSP or VIIRS satellite) as an 8th image channel. We did not find that this improved the baseline results, although Yeh et al. (2020) found gains when using nighttime light by incorporating the information at a later layer in the model.

Data processing and augmentation. We normalize each channel by the pixel-wise mean and standard deviation for each channel, following (Yeh et al., 2020). We also do a similar data augmentation scheme, adding random horizontal and vertical flips as well as color jitter (brightness factor 0.8, contrast factor 0.8, saturation factor 0.8, hue factor 0.1).

The data download process provided by Yeh et al. (2020) involves downloading and processing imagery from Google Earth Engine. We process all the imagery into a single NumPy array. We also provide all the metadata in a CSV format. We will provide a PyTorch implementation of a dataset loader and a baseline model training pipeline.

Comparison to other results. We see a much larger drop due to spatial shift than in Yeh et al. (2020). To explain this, we note that our data splitting method is slightly different to theirs. Since they

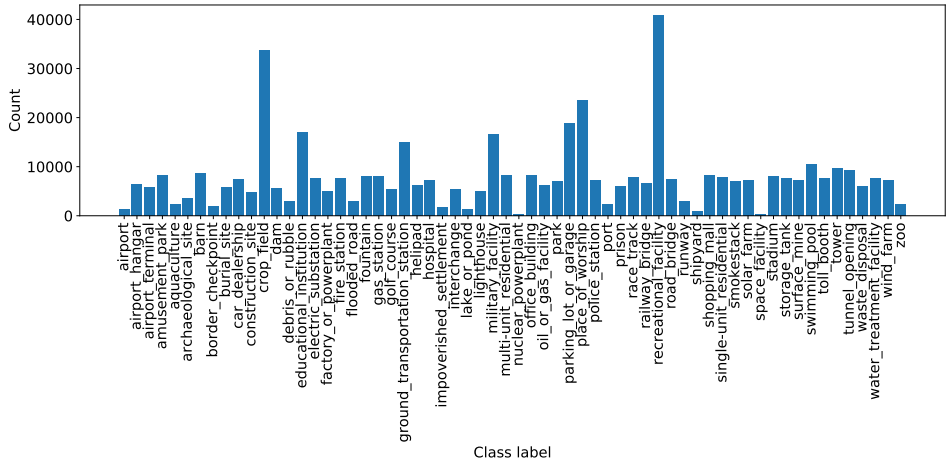


Figure 9: Number of examples from each category in FMoW-WILDS in non-African regions.

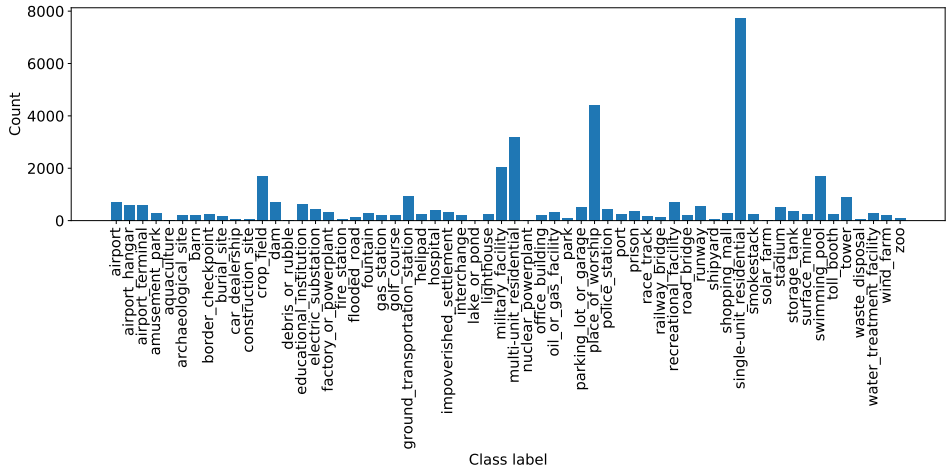


Figure 10: Number of examples from each category in FMoW-WILDS in Africa. There is a large label shift between non-African regions and Africa.

have two separate experiments (with different data splits) to test in-distribution vs. out-of-distribution generalization, we compare both metrics simultaneously on one model as a more direct comparison.

A.6 FMoW-WILDS

The FMoW-WILDS dataset is derived from Christie et al. (2018), which collects over 1 million satellite images from over 200 countries over 2002-2018.

Model. We use a Densenet-121 model (Huang et al., 2017) pretrained on ImageNet for 62-way classification. We train to minimize cross entropy loss using the Adam optimizer, decaying the learning rate by 0.96 per epoch (initial learning rate 1e-4). We use a batch size of 64 and train for 50 epochs, using early stopping with an in-distribution development set.

Data splits. We use the training, development, and validation splits from Christie et al. (2018) and remove all the examples with timestamp on or after 2013-01-01 as out-of-distribution. We use the validation out-of-distribution examples ($\geq$ 2013) as the test set. All of the data splits have examples from distinct locations.

Data processing. The original dataset from Christie et al. (2018) is provided as a set of hierarchical directories with JPEG images of varying sizes. We process the images as a collection of 100 NumPy arrays, where can we use the fast memory mapped reading mode to load our training set of over 150,000 images (< 2013) in under a second. We also collect all the metadata into CSV format for easy processing. We will provide a PyTorch implementation of a dataset loader and baseline model training pipeline, including how to split along time and region axes.

Label shift between non-African regions and Africa. Figures 9 and 10 plot the category frequencies of examples restricted to non-African regions or Africa-only. We find that there is a large label shift when moving to Africa, especially with a drop in recreational facilities and a large increase in single residential units. We do not find a similarly large label shift between < 2013 and ≥ 2013 splits of the dataset.