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

Pang Wei Koh [*1] **Shiori Sagawa** [*1] **Henrik Marklund** [1] **Sang Michael Xie** [1] **Marvin Zhang** [2]
Akshay Balsubramani [1] **Weihua Hu** [1] **Michihiro Yasunaga** [1] **Richard Lanas Phillips** [3] **Irena Gao** [1] **Tony Lee** [1]
Etienne David [4] **Ian Stavness** [5] **Wei Guo** [6] **Berton A. Earnshaw** [7] **Imran S. Haque** [7] **Sara Beery** [8]
Jure Leskovec [1] **Anshul Kundaje** [1] **Emma Pierson** [3 9] **Sergey Levine** [2] **Chelsea Finn** [1] **Percy Liang** [1]

Abstract

Distribution shifts—where the training distribution differs from the test distribution—can substantially degrade the accuracy of machine learning (ML) systems deployed in the wild. Despite their ubiquity in the real-world deployments, these distribution shifts are under-represented in the datasets widely used in the ML community today. To address this gap, we present WILDS, a curated benchmark of 10 datasets reflecting a diverse range of distribution shifts that naturally arise in real-world applications, such as shifts across hospitals for tumor identification; across camera traps for wildlife monitoring; and across time and location in satellite imaging and poverty mapping. On each dataset, we show that standard training yields substantially lower out-of-distribution than in-distribution performance. This gap remains even with models trained by existing methods for tackling distribution shifts, underscoring the need for new methods for training models that are more robust to the types of distribution shifts that arise in practice. To facilitate method development, we provide an open-source package that automates dataset loading, contains default model architectures and hyperparameters, and standardizes evaluations. **The full paper, code, and leaderboards are available at `https://wilds.stanford.edu`.**

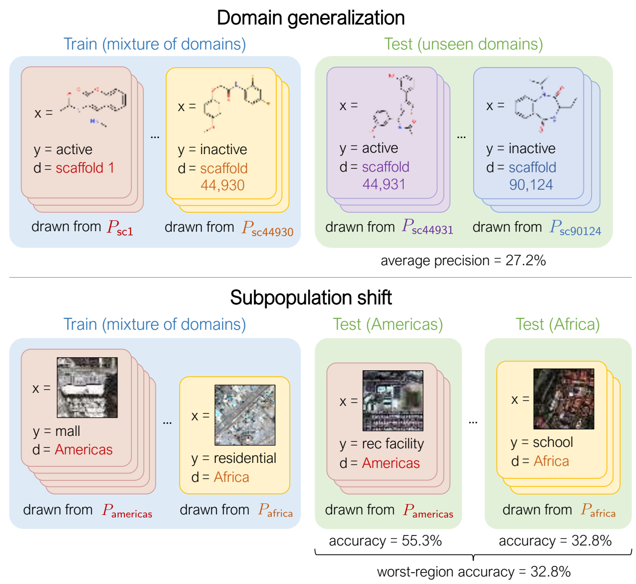

Figure 1: In each WILDS dataset, each data point (x, y, d) is associated with a domain d. Each domain corresponds to a distribution P_d over data points which are similar in some way, e.g., molecules with the same scaffold, or satellite images from the same region. We study two types of distribution shifts. **Top:** In *domain generalization*, we train and test on disjoint sets of domains. The goal is to generalize to domains unseen during training, e.g., molecules with a new scaffold in OGB-MOLPCBA (Hu et al., 2020b). **Bottom:** In *subpopulation shift*, the training and test domains overlap, but their relative proportions differ. We typically assess models by their worst performance over test domains, each of which correspond to a subpopulation of interest, e.g., different geographical regions in FMOW-WILDS (Christie et al., 2018).

1. Introduction

Distribution shifts—where the training distribution differs from the test distribution—pose significant challenges for machine learning (ML) systems deployed in the wild. In this work, we consider two common types of distribution shifts: *domain generalization* and *subpopulation shift* (Figure 1). Both of these shifts arise naturally in many real-world scenarios, and prior work has shown that they can substantially degrade model performance. In domain generalization, the training and test distributions comprise data from related but distinct domains, such as patients from different hospitals (Zech et al., 2018), images taken by different cameras (Beery et al., 2018), bioassays from different cell types (Li et al., 2019a), or satellite images from different countries and time periods (Jean et al., 2016). In subpopulation shift,

[*]Equal contribution [1]Stanford [2]UC Berkeley [3]Cornell [4]INRAE [5]USask [6]UTokyo [7]Recursion [8]Caltech [9]Microsoft Research. Correspondence to: Shiori Sagawa <ssagawa@cs.stanford.edu>, Pang Wei Koh <pangwei@cs.stanford.edu>, Percy Liang <pliang@cs.stanford.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

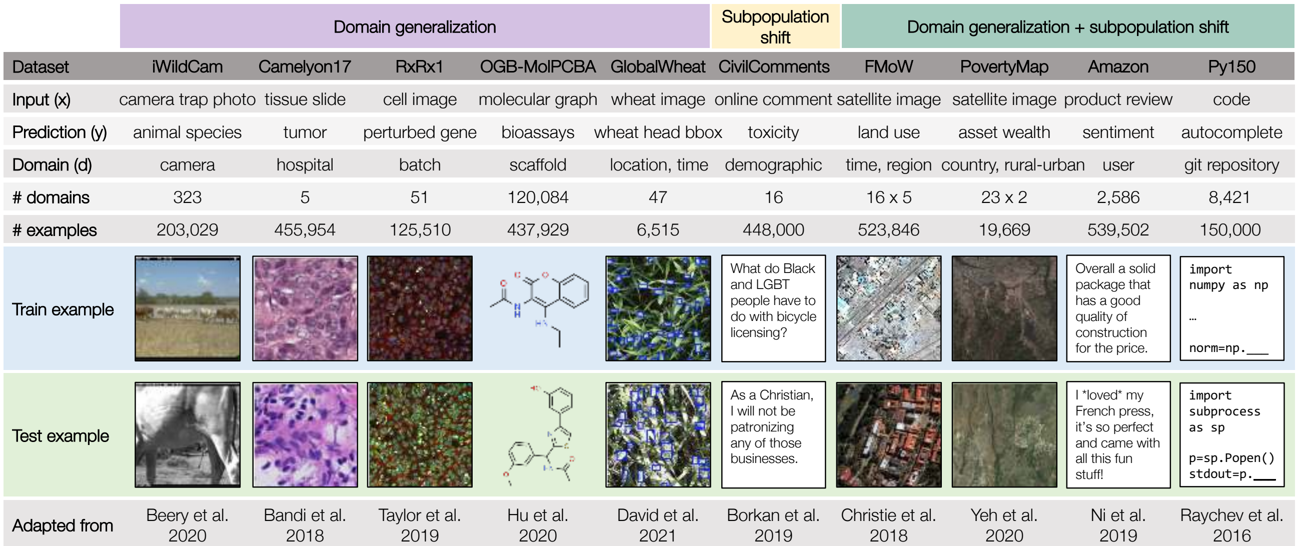

Figure 2: The WILDS benchmark contains 10 datasets across a diverse set of application areas, data modalities, and dataset sizes. Each dataset comprises data from different domains, and the benchmark is set up to evaluate models on distribution shifts across these domains.

we consider test distributions that are subpopulations of the training distribution, with the goal of doing well even on the worst-case subpopulation; e.g., we might seek models that perform well on all demographic subpopulations, including minority individuals (Buolamwini & Gebru, 2018).

Despite their ubiquity in real-world deployments, these types of distribution shifts are under-represented in the datasets widely used in the ML community today (Geirhos et al., 2020). Most of these datasets were designed for the standard i.i.d. setting, with training and test sets from the same distribution, and prior work on retrofitting them with distribution shifts has focused on shifts that are cleanly characterized but not always likely to arise in real-world deployments. For instance, many recent papers have studied datasets with shifts induced by synthetic transformations, such as changing the color of MNIST digits (Arjovsky et al., 2019), or by disparate data splits, such as generalizing from cartoons to photos (Li et al., 2017a). Datasets like these are important testbeds for systematic studies; but to develop and evaluate methods for real-world shifts, we need to complement them with datasets that capture shifts in the wild.

In this paper, we present WILDS, a curated benchmark of 10 datasets with evaluation metrics and train/test splits representing a broad array of distribution shifts that ML models face in the wild (Figure 2). WILDS datasets span many important applications: animal species categorization (Beery et al., 2020a), tumor identification (Bandi et al., 2018), bioassay prediction (Wu et al., 2018; Hu et al., 2020b), genetic perturbation classification (Taylor et al., 2019), wheat head detection (David et al., 2020), text toxicity classification (Borkan et al., 2019b), land use classification (Christie et al., 2018), poverty mapping (Yeh et al., 2020), sentiment analysis (Ni et al., 2019), and code completion (Raychev et al., 2016; Lu et al., 2021). These datasets reflect natural distribution shifts arising from different cameras, hospitals, molecular scaffolds, experiments, demographics, countries, time periods, users, and codebases.

WILDS builds on extensive data-collection efforts by domain experts, who are often forced to grapple with distribution shifts to make progress in their applications. To design WILDS, we worked with them to identify, select, and adapt datasets that fulfilled the following criteria:

1. **Distribution shifts with performance drops.** The train/test splits reflect shifts that substantially degrade model performance, i.e., with a large gap between in-distribution and out-of-distribution performance.

2. **Real-world relevance.** The training/test splits and evaluation metrics are motivated by real-world scenarios and chosen in conjunction with domain experts. In Appendix A, we further discuss the framework we use to assess the realism of a dataset.

3. **Potential leverage.** Distribution shift benchmarks must be non-trivial but also possible to solve, as models cannot be expected to generalize to arbitrary distribution shifts. We constructed each WILDS dataset to have training data from multiple domains, with domain annotations and other metadata available at training time. We hope that these can be used to learn robust models: e.g., for domain generalization, one could use these annotations to learn models that are invariant to domain-specific features, while for subpopulation shift, one could learn models that perform uniformly well across each subpopulation.

We chose the WILDS datasets to collectively encompass a diverse set of tasks, data modalities, dataset sizes, and numbers of domains, so as to enable evaluation across a broad range of real-world distribution shifts. In Appendix C, we further survey the distribution shifts that occur in other application areas—algorithmic fairness and policing, medicine and healthcare, genomics, natural language and speech processing, education, and robotics—and discuss examples of datasets from these areas that we considered but did not include in WILDS, as their distribution shifts did not cause an appreciable performance drop.

To make the WILDS datasets more accessible, we have substantially modified most of them, e.g., to clarify the distribution shift, standardize the data splits, or preprocess the data for use in standard ML frameworks. In Appendix F, we introduce our accompanying open-source Python package that fully automates data loading and evaluation. The package also includes default models appropriate for each dataset, allowing all of the baseline results reported in this paper to be easily replicated. To track the state-of-the-art in training algorithms and model architectures that are robust to these distribution shifts, we are also hosting a public leaderboard; we discuss guidelines for developers in Section 7. Code, leaderboards, and updates are available at `https://wilds.stanford.edu`.

Datasets are significant catalysts for ML research. Likewise, benchmarks that curate and standardize datasets—e.g., the GLUE and SuperGLUE benchmarks for language understanding (Wang et al., 2019a;b) and the Open Graph Benchmark for graph ML (Hu et al., 2020b)—can accelerate research by focusing community attention, easing development on multiple datasets, and enabling systematic comparisons between approaches. In this spirit, we hope that WILDS will facilitate the development of ML methods and models that are robust to real-world distribution shifts and can therefore be deployed reliably in the wild.

2. Comparison with existing ML benchmarks

Distribution shifts have been a longstanding problem in the ML research community (Hand, 2006; Quiñonero-Candela et al., 2009). Earlier work studied shifts in datasets for tasks including part-of-speech tagging (Marcus et al., 1993), sentiment analysis (Blitzer et al., 2007), land cover classification (Bruzzone & Marconcini, 2009), object recognition (Saenko et al., 2010), and flow cytometry (Blanchard et al., 2011). However, these datasets are not as widely used today, in part because they tend to be much smaller than modern datasets.

Instead, recent papers have focused on object recognition datasets with shifts induced by synthetic transformations, such as ImageNet-C (Hendrycks & Dietterich, 2019), which corrupts images with noise; the Backgrounds Challenge (Xiao et al., 2020) and Waterbirds (Sagawa et al., 2020a), which alter image backgrounds; or Colored MNIST (Arjovsky et al., 2019), which changes the colors of MNIST digits. It is also common to use data splits or combinations of disparate datasets to induce shifts, such as generalizing to photos solely from cartoons and other stylized images in PACS (Li et al., 2017a); generalizing to objects at different scales solely from a single scale in DeepFashion Remixed (Hendrycks et al., 2020b); or using training and test sets with disjoint subclasses in BREEDS (Santurkar et al., 2020) and similar datasets (Hendrycks & Dietterich, 2019). While our treatment here is necessarily brief, we discuss other similar datasets in Appendix B.

These existing benchmarks are useful and important testbeds for method development. As they typically target well-defined and isolated shifts, they facilitate clean analysis and controlled experimentation, e.g., studying the effect of backgrounds on image classification (Xiao et al., 2020), or showing that training with added Gaussian blur improves performance on real-world blurry images (Hendrycks et al., 2020b). Moreover, by studying how off-the-shelf models trained on standard datasets like ImageNet perform on different test datasets, we can better understand the robustness of these widely-used models (Geirhos et al., 2018b; Recht et al., 2019; Hendrycks & Dietterich, 2019; Taori et al., 2020; Djolonga et al., 2020; Hendrycks et al., 2020b).

However, existing benchmarks do not generally represent realistic distribution shifts, i.e., train/test splits that are likely to arise in real-world deployments. As model robustness need not transfer across shifts , it is important to develop and evaluate methods on real-world shifts. For example, models can be robust to image corruptions but not to shifts across datasets (Taori et al., 2020; Djolonga et al., 2020), and a method that improves robustness on a standard vision dataset can actually consistently harm robustness on real-world satellite imagery datasets (Xie et al., 2020). With WILDS, we seek to complement these existing benchmarks by focusing on datasets with realistic distribution shifts across a diverse set of data modalities and applications.

3. Problem settings

Each WILDS dataset is associated with a type of domain shift: domain generalization, subpopulation shift, or a hybrid of both (Figure 2). In each setting, we can view the overall data distribution as a mixture of D domains $\mathcal{D} = \{1, \dots, D\}$. Each domain $d \in \mathcal{D}$ corresponds to a fixed data distribution P_d over (x, y, d), where x is the input, y is the prediction target, and all points sampled from P_d have domain d. We encode the domain shift by assuming that the training distribution $P^{\text{train}} = \sum_{d \in \mathcal{D}} q_d^{\text{train}} P_d$ has mixture weights q_d^{train} for each domain d, while the test distribution $P^{\text{test}} = \sum_{d \in \mathcal{D}} q_d^{\text{test}} P_d$ is a different mixture of domains with weights q_d^{test}. For convenience, we define the set of training domains as $\mathcal{D}^{\text{train}} = \{d \in \mathcal{D} \mid q_d^{\text{train}} > 0\}$, and likewise, the set of test domains as $\mathcal{D}^{\text{test}} = \{d \in \mathcal{D} \mid q_d^{\text{test}} > 0\}$.

At training time, the learning algorithm gets to see the domain annotations d, i.e., the training set comprises points $(x, y, d) \sim P^{\text{train}}$. At test time, the model gets either x or (x, d) drawn from P^{test}, depending on the application.

Domain generalization (Figure 1-Top). In domain generalization, we aim to generalize to test domains $\mathcal{D}^{\text{test}}$ that are disjoint from the training domains $\mathcal{D}^{\text{train}}$, i.e., $\mathcal{D}^{\text{train}} \cap \mathcal{D}^{\text{test}} = \emptyset$. To make this problem tractable, the training and test domains are typically similar to each other: e.g., in CAMELYON17-WILDS, we train on data from some hospitals and test on a different hospital, and in IWILDCAM2020-WILDS, we train on data from some camera traps and test on different camera traps. We typically seek to minimize the average error on the test distribution.

Subpopulation shift (Figure 1-Bottom). In subpopulation shift, we aim to perform well across a wide range of domains seen during training time. Concretely, all test domains are seen at training, with $\mathcal{D}^{\text{test}} \subseteq \mathcal{D}^{\text{train}}$, but the proportions of the domains can change, with $q^{\text{test}} \neq q^{\text{train}}$. We typically seek to minimize the maximum error over all test domains. For example, in CIVILCOMMENTS-WILDS, the domains d represent particular demographics, some of which are a minority in the training set, and we seek high accuracy on each of these subpopulations without observing their demographic identity d at test time.

Hybrid settings. It is not always possible to cleanly define a problem as domain generalization or subpopulation shift; for example, a test domain might be present in the training set but at a very low frequency. In WILDS, we also consider some hybrid settings that combine both problem settings. For example, in FMOW-WILDS, the inputs are satellite images and the domains correspond to the year and geographical region in which they were taken. We simultaneously consider domain generalization across time (the training/test sets comprise images taken before/after a certain year) and subpopulation shift across regions (there are images from the same regions in the training and test sets, and we seek high performance across all regions).

4. WILDS datasets

We now briefly describe each WILDS dataset (Figure 2). For each dataset, we consider a problem setting—domain generalization, subpopulation shift, or a hybrid—that we believe best reflects the real-world challenges in the corresponding application area; see Appendix A for more discussion of these considerations. To avoid confusion between our modified datasets and their original sources, we append -WILDS to the dataset names. We provide more details and context on related distribution shifts for each dataset in Appendix H.

4.1. Domain generalization datasets

IWILDCAM2020-WILDS (Appendix H.1). Animal populations have declined 68% on average since 1970 (Grooten et al., 2020). To better understand and monitor wildlife biodiversity loss, ecologists commonly deploy camera traps—heat or motion-activated static cameras placed in the wild (Wearn & Glover-Kapfer, 2017)—and then use ML models to process the data collected (Weinstein, 2018; Norouzzadeh et al., 2019; Tabak et al., 2019; Beery et al., 2019; Ahumada et al., 2020). Typically, these models would be trained on photos from existing camera traps and then used across new camera trap deployments. However, across different camera traps, there is drastic variation in illumination, color, camera angle, background, vegetation, and relative animal frequencies, which results in models generalizing poorly to new camera trap deployments (Beery et al., 2018).

We study this shift on a variant of the iWildCam 2020 dataset (Beery et al., 2020a), where the input x is a photo from a camera trap, the label y is one of 182 animal species, and the domain d specifies the identity of the camera trap. The training and test sets comprise photos from disjoint sets of camera traps. As leverage, we include over 200 camera traps in the training set, capturing a wide range of variation. We evaluate models by their macro F1 scores, which emphasizes performance on rare species, as rare and endangered species are the most important to accurately monitor.

CAMELYON17-WILDS (Appendix H.2). Models for medical applications are often trained on data from a small number of hospitals, but with the goal of being deployed more generally across other hospitals. However, variations in data collection and processing can degrade model accuracy on data from new hospital deployments (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 shift on a patch-based variant of the Camelyon17 dataset (Bandi et al., 2018), where the input x is a 96x96 patch of a whole-slide image of a lymph node section from a patient with potentially metastatic breast cancer, the label y is whether the patch contains tumor, and the domain d specifies which of 5 hospitals the patch was from. The training and test sets comprise class-balanced patches from separate hospitals, and we evaluate models by their average accuracy. Prior work suggests that staining differences are the main source of variation between hospitals in similar datasets (Tellez et al., 2019). As we have training data from multiple hospitals, a model could use that as leverage to learn to be robust to stain variation.

RxRx1-wilds (Appendix H.3). High-throughput screening techniques that can generate large amounts of data are now common in many fields of biology, including transcriptomics (Harrill et al., 2019), genomics (Echeverri & Perrimon, 2006; Zhou et al., 2014), proteomics and metabolomics (Taylor et al., 2021), and drug discovery (Broach et al., 1996; Macarron et al., 2011; Swinney & Anthony, 2011; Boutros et al., 2015). Such large volumes of data, however, need to be created in experimental batches, or groups of experiments executed at similar times under similar conditions. Despite attempts to carefully control experimental variables such as temperature, humidity, and reagent concentration, measurements from these screens are confounded by technical artifacts that arise from differences in the execution of each batch. These *batch effects* make it difficult to draw conclusions from data across experimental batches (Leek et al., 2010; Parker & Leek, 2012; Soneson et al., 2014; Nygaard et al., 2016; Caicedo et al., 2017).

We study the shift induced by batch effects on a variant of the RxRx1 dataset (Taylor et al., 2019), where the input x is a 3-channel image of cells obtained by fluorescent microscopy (Bray et al., 2016), the label y indicates which of the 1,139 genetic treatments (including no treatment) the cells received, and the domain d specifies the batch in which the imaging experiment was run. The training and test sets consist of disjoint experimental batches; as leverage, the training set has images from 33 different batches, with each batch containing one sample for every class. We assess a model's ability to normalize batch effects while preserving biological signal by evaluating how well it can classify images of treated cells in the out-of-distribution test set.

OGB-MolPCBA (Appendix H.4). Accurate prediction of the biochemical properties of small molecules can significantly accelerate drug discovery by reducing the need for expensive lab experiments (Shoichet, 2004; Hughes et al., 2011). However, the experimental data available for training such models is limited compared to the extremely diverse and combinatorially large universe of candidate molecules that we would want to make predictions on (Bohacek et al., 1996; Sterling & Irwin, 2015; Lyu et al., 2019; McCloskey et al., 2020). This means that models need to generalize to out-of-distribution molecules that are structurally different from those seen in the training set.

We study this shift on the OGB-MolPCBA dataset, which is directly adopted from the Open Graph Benchmark (Hu et al., 2020b) and originally from MoleculeNet (Wu et al., 2018). It is a multi-label classification dataset, where the input x is a molecular graph, the label y is a 128-dimensional binary vector where each component corresponds to a biochemical assay result, and the domain d specifies the scaffold (i.e., a cluster of molecules with similar structure). The training and test sets comprise molecules with disjoint scaffolds; for leverage, the training set has molecules from over 40,000 scaffolds. We evaluate models by averaging the Average Precision (AP) across each of the 128 assays.

GlobalWheat-wilds (Appendix H.5). Models for automated, high-throughput plant phenotyping—measuring the physical characteristics of plants and crops, such as wheat head density and counts—are important tools for crop breeding (Thorp et al., 2018; Reynolds et al., 2020) and agricultural field management (Shi et al., 2016). These models are typically trained on data collected in a limited number of regions, even for crops grown worldwide such as wheat (Madec et al., 2019; Xiong et al., 2019; Ubbens et al., 2020; Ayalew et al., 2020). However, there can be substantial variation between regions, due to differences in crop varieties, growing conditions, and data collection protocols. Prior work on wheat head detection has shown that this variation can significantly degrade model performance on regions unseen during training (David et al., 2020).

We study this shift in an expanded version of the Global Wheat Head Dataset (David et al., 2020; 2021), a large set of wheat images collected from 12 countries around the world. It is a detection dataset, where the input x is a cropped overhead image of a wheat field, the label y is the set of bounding boxes for each wheat head visible in the image, and the domain d specifies an image acquisition session (i.e., a specific location, time, and sensor with which a set of images was collected). The data split captures a shift in location, with training and test sets comprising images from disjoint countries. As leverage, we include images from 18 acquisition sessions over 5 countries in the training set. We evaluate model performance on unseen countries by measuring accuracy at a fixed Intersection over Union (IoU) threshold, and averaging across acquisition sessions to account for imbalances in the numbers of images in them.

4.2. Subpopulation shift datasets

CivilComments-wilds (Appendix H.6). Automatic review of user-generated text is an important tool for moderating the sheer volume of text written on the Internet. We focus here on the task of detecting toxic comments. Prior work has shown that toxicity classifiers can pick up on biases in the training data and spuriously associate toxicity with the mention of certain demographics (Park et al., 2018; Dixon et al., 2018). These types of spurious correlations can significantly degrade model performance on particular subpopulations (Sagawa et al., 2020a).

We study this problem on a variant of the CivilComments dataset (Borkan et al., 2019b), a large collection of comments on online articles taken from the Civil Comments platform. The input x is a text comment, the label y is whether the comment was rated as toxic, and the domain d is a 8-dimensional binary vector where each component corresponds to whether the comment mentions one of the 8 demographic identities *male*, *female*, *LGBTQ*, *Christian*, *Muslim*, *other religions*, *Black*, and *White*. The training and test sets comprise comments on disjoint articles, and we evaluate models by the lowest true positive/negative rate over each of these 8 demographic groups; these groups overlap with each other, deviating slightly from the standard subpopulation shift framework in Section 3. Models can use the provided domain annotations as leverage to learn to perform well over each demographic group.

4.3. Hybrid datasets

FMoW-wilds (Appendix H.7). ML models for satellite imagery can enable global-scale monitoring of sustainability and economic challenges, aiding policy and humanitarian efforts in applications such as deforestation tracking (Hansen et al., 2013), population density mapping (Tiecke et al., 2017), crop yield prediction (Wang et al., 2020b), and other economic tracking applications (Katona et al., 2018). As satellite data constantly changes due to human activity and environmental processes, these models must be robust to distribution shifts over time. Moreover, as there can be disparities in the data available between regions, these models should ideally have uniformly high accuracies instead of only doing well on data-rich regions and countries.

We study this problem on a variant of the Functional Map of the World dataset (Christie et al., 2018), where the input x is an RGB satellite image, the label y is one of 62 building or land use categories, and the domain d represents the year the image was taken and its geographical region (Africa, the Americas, Oceania, Asia, or Europe). The different regions have different numbers of examples, e.g., there are far fewer images from Africa than the Americas. The training set comprises data from before 2013, while the test set comprises data from 2016 and after; years 2013 to 2015 are reserved for the validation set. We evaluate models by their test accuracy on the worst geographical region, which combines both a domain generalization problem over time and a subpopulation shift problem over regions. As we provide both time and region annotations, models can leverage the structure across both space and time to improve robustness.

PovertyMap-wilds (Appendix H.8). Global-scale poverty estimation is a specific remote sensing application which is essential for targeted humanitarian efforts in poor regions (Abelson et al., 2014; Espey et al., 2015). However, ground truth measurements of poverty are lacking for much of the developing world, as field surveys for collecting the ground truth are expensive (Blumenstock et al., 2015). This motivates the approach of training ML models on countries with ground truth labels and then deploying them on different countries where we have satellite data but no labels (Xie et al., 2016; Jean et al., 2016; Yeh et al., 2020).

We study this shift through a variant of the poverty mapping dataset collected by Yeh et al. (2020), where the input x is a multispectral satellite image, the output y is a real-valued asset wealth index from surveys, and the domain d represents the country the image was taken in and whether the image is of an urban or rural area. The training and test set comprise data from disjoint sets of countries, and we evaluate models by the correlation of their predictions with the ground truth. Specifically, we take the lower of the correlations over the urban and rural subpopulations, as prior work has shown that accurately predicting poverty within these subpopulations is especially challenging. As poverty measures are highly correlated across space (Jean et al., 2018; Rolf et al., 2020), methods can utilize the provided location coordinates, and the country and urban/rural annotations, to improve robustness.

Amazon-wilds (Appendix H.9). In many consumer-facing ML applications, models are trained on data collected on one set of users and then deployed across a wide range of potentially new users. These models can perform well on average but poorly on some users (Tatman, 2017; Caldas et al., 2018; Li et al., 2019b; Koenecke et al., 2020). These large performance disparities across users are practical concerns in consumer-facing applications, and they can also indicate that models are exploiting biases or spurious correlations in the data (Badgeley et al., 2019; Geva et al., 2019).

We study a variant of the Amazon review dataset (Ni et al., 2019), where the input x is the review text, the label y is the corresponding 1-to-5 star rating, and the domain d identifies the user who wrote the review. The training and test sets comprise reviews from disjoint sets of users; for leverage, the training set has reviews from 5,008 different users. As our goal is to train models with consistently high performance across users, we evaluate models by the 10th percentile of per-user accuracies. We discuss other distribution shifts on this dataset (e.g., by category) in Appendix I.3.

Py150-wilds (Appendix H.10). Code completion models—autocomplete tools used by programmers to suggest subsequent source code tokens, such as the names of API calls—are commonly used to reduce the effort of software development (Robbes & Lanza, 2008; Bruch et al., 2009; Nguyen & Nguyen, 2015; Proksch et al., 2015; Franks et al., 2015). These models are typically trained on data collected from existing codebases but then deployed more generally across other codebases, which may have different distributions of API usages (Nita & Notkin, 2010; Proksch et al., 2016; Allamanis & Brockschmidt, 2017). This shift across codebases can cause substantial performance drops in code completion models. Moreover, prior studies of real-world usage of code completion models have noted that they can generalize poorly on some important subpopulations of tokens such as method names (Hellendoorn et al., 2019).

Table 1: The in-distribution (ID) vs. out-of-distribution (OOD) performance of models trained with empirical risk minimization. The OOD test sets are drawn from the shifted test distributions described in Section 4, while the ID comparisons vary per dataset and are described in the main text. For each dataset, higher numbers are better. In all tables in this paper, we report in parentheses the standard deviation across 3+ replicates, which measures the variability between replicates; note that this is higher than the standard error of the mean, which measures the variability in the estimate of the mean across replicates.

Dataset	Metric	In-distribution type	In-distribution	Out-of-distribution
IWILDCAM2020-WILDS	Macro F1	Fixed-train	47.0 (1.4)	31.0 (1.3)
CAMELYON17-WILDS	Average accuracy	Fixed-train	93.2 (5.2)	70.3 (6.4)
RXRX1-WILDS	Average accuracy	Fixed-test	39.8 (0.2)	29.9 (0.4)
OGB-MOLPCBA	Average AP	Randomized	34.4 (0.9)	27.2 (0.3)
GLOBALWHEAT-WILDS	Average domain accuracy	Fixed-test	64.8 (0.4)	48.4 (1.8)
CIVILCOMMENTS-WILDS	Worst-group accuracy	Average	92.2 (0.1)	56.0 (3.6)
FMOW-WILDS	Worst-region accuracy	Fixed-test	48.6 (0.9)	32.3 (1.3)
POVERTYMAP-WILDS	Worst-U/R Pearson R	Fixed-test	0.60 (0.06)	0.45 (0.06)
AMAZON-WILDS	10th percentile accuracy	Average	71.9 (0.1)	53.8 (0.8)
PY150-WILDS	Method/class accuracy	Fixed-train	75.4 (0.4)	67.9 (0.1)

We study a variant of the Py150 Dataset (Raychev et al., 2016; Lu et al., 2021), where the goal is to predict the next token given the context of previous tokens. The input x is a sequence of source code tokens, the label y is the next token, and the domain d specifies the repository that the source code belongs to. The training and test sets comprise code from disjoint GitHub repositories. As leverage, we include over 5,300 repositories in the training set, capturing a wide range of source code variation. We evaluate models by their accuracy on the subpopulation of class and method tokens.

5. Performance drops from distribution shifts

For a dataset to be included in WILDS, the shift reflected in its train/test split should cause significant performance drops in standard models. We ascertained this for each dataset by training standard models using empirical risk minimization (ERM), i.e., minimizing the average training loss, and then comparing their out-of-distribution (OOD) vs. in-distribution (ID) performance. The OOD setting is captured by the default train/test split and the evaluation criteria described in Section 4: for domain generalization, we report performance on unseen domains, and for subpopulation shift, we report performance on the worst-case subpopulation. ID comparisons vary by dataset. Each dataset has at least one of the following types of ID comparisons:

1. **Fixed-train.** We hold the training set constant and evaluate on a separate ID test set of data from the same domains (e.g., camera traps) as the training set. This comparison is convenient because it does not require retraining the model, and we use it when we expect the training and test domains to be interchangeable in the sense of being randomly drawn from the same distribution, e.g., in IWILDCAM2020-WILDS, where the camera traps are randomly split across training and test sets.

2. **Fixed-test.** We hold the OOD test set approximately constant and modify the training set to mix in data from the (OOD) test distribution, while keeping the size of training set similar or smaller. We use this comparison when the training and test distributions are qualitatively different: e.g., in FMOW-WILDS, where the test distribution comes from a later time period, we replace half of the data in the training set with otherwise unused data from the test distribution.

3. **Randomized.** We shuffle all of the data into i.i.d. training, validation, and test splits. We use this for OGB-MOLPCBA, where the small size of the domains preclude the other options.

4. **Average.** For subpopulation shift datasets, where models are evaluated on a subpopulation of the data, we report the average performance across the entire OOD test set.

More details on the ID and OOD test sets, and additional results for datasets that admit multiple ID comparisons, are described in Appendix H. We further describe model selection and the general experimental protocol in Appendix G.

Results. Table 1 shows that for each dataset, OOD performance is consistently and substantially lower than ID performance. Moreover, on the datasets that allow for fixed-test ID comparisons, we show that oracle models trained on a mix of the ID and OOD distributions can simultaneously achieve high ID and OOD performance, indicating that lower OOD performance is not due to the OOD test sets being intrinsically more difficult than the ID test sets (Appendix H). Overall, these results demonstrate that the real-world distribution shifts reflected in the WILDS datasets meaningfully degrade standard model performance.

Table 2: The out-of-distribution test performance of models trained with different baseline algorithms: CORAL, originally designed for unsupervised domain adaptation; IRM, for domain generalization; and Group DRO, for subpopulation shifts. Evaluation metrics for each dataset are the same as in Table 1; higher is better. Overall, these algorithms failed to improve over ERM, except on CIVILCOMMENTS-WILDS where they perform better but still do not close the in-distribution gap in Table 1. For GLOBALWHEAT-WILDS, we omit CORAL and IRM as those methods do not port straightforwardly to detection settings; its ERM number also differs from Table 1 as its ID comparison required a slight change to the OOD test set. Parentheses show standard deviation across 3+ replicates.

Dataset	Setting	ERM	CORAL	IRM	Group DRO
IWILDCAM2020-WILDS	Domain gen.	31.0 (1.3)	**32.8 (0.1)**	15.1 (4.9)	23.9 (2.1)
CAMELYON17-WILDS	Domain gen.	**70.3 (6.4)**	59.5 (7.7)	64.2 (8.1)	68.4 (7.3)
RXRX1-WILDS	Domain gen.	**29.9 (0.4)**	28.4 (0.3)	8.2 (1.1)	23.0 (0.3)
OGB-MOLPCBA	Domain gen.	**27.2 (0.3)**	17.9 (0.5)	15.6 (0.3)	22.4 (0.6)
GLOBALWHEAT-WILDS	Domain gen.	**49.2 (1.5)**	—	—	46.1 (1.6)
CIVILCOMMENTS-WILDS	Subpop. shift	56.0 (3.6)	65.6 (1.3)	66.3 (2.1)	**70.0 (2.0)**
FMOW-WILDS	Hybrid	**32.3 (1.3)**	31.7 (1.2)	30.0 (1.4)	30.8 (0.8)
POVERTYMAP-WILDS	Hybrid	**0.45 (0.06)**	0.44 (0.06)	0.43 (0.07)	0.39 (0.06)
AMAZON-WILDS	Hybrid	**53.8 (0.8)**	52.9 (0.8)	52.4 (0.8)	53.3 (0.0)
PY150-WILDS	Hybrid	**67.9 (0.1)**	65.9 (0.1)	64.3 (0.2)	65.9 (0.1)

6. Baseline algorithms for distribution shifts

Many algorithms have been proposed for training models that are more robust to particular distribution shifts than standard ERM models. Unlike ERM, these algorithms tend to utilize domain annotations during training, with the goal of learning a model that can generalize across domains. In this section, we evaluate several representative algorithms from prior work and show that the out-of-distribution performance drops shown in Section 5 still remain.

Domain generalization baselines. Methods for domain generalization typically involve adding a penalty to the ERM objective that encourages some form of invariance across domains. We include two such methods as representatives:

- **CORAL** (Sun & Saenko, 2016), which penalizes differences in the means and covariances of the feature distributions (i.e., the distribution of last layer activations in a neural network) for each domain. Conceptually, CORAL is similar to other methods that encourage feature representations to have the same distribution across domains (Tzeng et al., 2014; Long et al., 2015; Ganin et al., 2016; Li et al., 2018c;b).
- **IRM** (Arjovsky et al., 2019), which penalizes feature distributions that have different optimal linear classifiers for each domain. This builds on earlier work on invariant predictors (Peters et al., 2016).

Other techniques for domain generalization include conditional variance regularization (Heinze-Deml & Meinshausen, 2017); self-supervision (Carlucci et al., 2019); and meta-learning-based approaches (Li et al., 2018a; Balaji et al., 2018; Dou et al., 2019).

Subpopulation shift baselines. In subpopulation shift settings, our aim is to train models that perform well on all relevant subpopulations. We test the following approach:

- **Group DRO** (Hu et al., 2018; Sagawa et al., 2020a), which uses distributionally robust optimization to explicitly minimize the loss on the worst-case domain during training. Group DRO builds on the maximin approach developed in Meinshausen & Bühlmann (2015).

Other methods for subpopulation shifts include reweighting methods based on class/domain frequencies (Shimodaira, 2000; Cui et al., 2019); label-distribution-aware margin losses (Cao et al., 2019); adaptive Lipschitz regularization (Cao et al., 2020); slice-based learning (Chen et al., 2019b; Ré et al., 2019); style transfer across domains (Goel et al., 2020); or other DRO algorithms that do not make use of explicit domain information and rely on, for example, unsupervised clustering (Oren et al., 2019; Sohoni et al., 2020).

Subpopulation shifts are also connected to the well-studied notions of tail performance and risk-averse optimization (Chapter 6 in Shapiro et al. (2014)). For example, optimizing for the worst case over all subpopulations of a certain size, regardless of domain, can guarantee a certain level of performance over the smaller set of subpopulations defined by domains (Duchi et al., 2020; Duchi & Namkoong, 2021).

Setup. We trained CORAL, IRM, and Group DRO models on each dataset. While Group DRO was originally developed for subpopulation shifts, for completeness, we also experiment with using it for domain generalization. In that setting, Group DRO models aim to achieve similar performance across domains: e.g., in CAMELYON17-WILDS, where the domains are hospitals, Group DRO optimizes for the training hospital with the highest loss. Similarly, we also test CORAL and IRM on subpopulation shifts, where they encourage models to learn invariant representations across subpopulations. As in Section 5, we used the same OOD validation set for early stopping and to tune the penalty weights for the CORAL and IRM algorithms. More experimental details are in Appendix G, and dataset-specific hyperparameters and domain choices are discussed in Appendix H.

Results. Table 2 shows that models trained with CORAL, IRM, and Group DRO generally fail to improve over models trained with ERM. The exception is the CIVILCOMMENTS-WILDS subpopulation shift dataset, where the worst-performing subpopulation is a minority domain. By upweighting the minority domain, Group DRO obtains an OOD accuracy of 70.0% on the worst-performing subpopulation compared to 56.0% for ERM, though this is comparable to simple class balancing (Appendix H.6) and is still substantially below the ERM model's average accuracy of 92.2% over the entire test set. CORAL and IRM also perform well on CIVILCOMMENTS-WILDS, though their gains stem largely from how our implementation heuristically upsamples the minority domain. All other datasets involve domain generalization; the failures here are consistent with other recent findings on standard domain generalization datasets (Gulrajani & Lopez-Paz, 2020).

These results indicate that training models to be robust to distribution shifts in the wild remains a significant open challenge. However, we are optimistic about future progress for two reasons. First, current methods were mostly designed for other problem settings besides domain generalization, e.g., CORAL for unsupervised domain adaptation and Group DRO for subpopulation shifts. Second, compared to existing distribution shift datasets, the WILDS datasets generally contain diverse training data from many more domains as well as metadata on these domains, which future algorithms might be able to leverage.

7. Discussion

We end by discussing extensions to WILDS and community guidelines for method development using WILDS.

Other applications and datasets. Distribution shifts are a challenge in many application areas beyond those covered in WILDS. In Appendix C, we survey other application areas—algorithmic fairness and policing, medicine and healthcare, natural language and speech processing, code, education, and robotics—and discuss relevant distribution shifts as well as the challenges associated with finding appropriate datasets in these areas. In Appendix I, we also present results on datasets from these areas that we had considered including in WILDS, but for which we did not see an appreciable performance drop under distribution shift. These include location and time shifts in the BDD100K autonomous driving dataset (Yu et al., 2020), location and race shifts in the New York stop-question-and-frisk dataset (Goel et al., 2016), and category and time shifts in the Amazon and Yelp review datasets (Ni et al., 2019). Understanding when distribution shifts result in large performance drops is an important question for future work to resolve.

Other problem settings. In this paper, we focused on the domain generalization and subpopulation shift problem settings. In Appendix D, we discuss how WILDS can be used to develop and evaluate models in other problem settings that allow training algorithms to leverage additional information, such as unlabeled test data in unsupervised domain adaptation (Ben-David et al., 2006).

Guidelines for algorithm development. WILDS is a benchmark for developing and evaluating algorithms for training models that are robust to distribution shifts. To facilitate systematic comparisons between these algorithms, we encourage algorithm developers to use the standardized datasets (i.e., with no external data), evaluation criteria, and default model architectures provided in WILDS. Moreover, we encourage developers to test their algorithms on all applicable WILDS datasets. We emphasize that it is still an open question if a single general-purpose training algorithm can produce models that do well on all of the datasets without accounting for the particular structure of the distribution shift in each dataset. As such, it would still be a substantial advance if an algorithm significantly improves performance on one type of shift but not others.

Methods beyond training algorithms. Beyond new training algorithms, there are many other promising directions for improving distributional robustness, including new model architectures and pre-training on additional external data beyond what is used in our default models. We encourage developers to test these approaches on WILDS as well, and we will track all such submissions on a separate leaderboard from the training algorithm leaderboard.

Avoiding overfitting to the test distribution. While each WILDS dataset aims to benchmark robustness to a type of distribution shift (e.g., shifts to unseen hospitals), practical limitations mean that for some datasets, we have data from only a limited number of domains (e.g., one OOD test hospital in CAMELYON17-WILDS). As there can be substantial variability in performance across domains, developers should be careful to avoid overfitting to the specific test sets in WILDS, especially on datasets like CAMELYON17-WILDS with limited test domains. We strongly encourage all model developers to use the provided OOD validation sets for development and model selection, and to only use the OOD test sets for their final evaluations.

Reproducibility

An executable version of our paper, hosted on CodaLab, can be found at `https://wilds.stanford.edu/codalab`. This contains the exact commands, code, environment, and data used for the experiments reported in our paper, as well as all trained model weights. The WILDS package is open-source and can be found at `https://github.com/p-lambda/wilds`.

Acknowledgements

Many people generously volunteered their time and expertise to advise us on WILDS. We are grateful for all of the helpful suggestions and constructive feedback from: Aditya Khosla, Andreas Schlueter, Annie Chen, Aleksander Madry, Alexander D'Amour, Allison Koenecke, Alyssa Lees, Ananya Kumar, Andrew Beck, Behzad Haghgoo, Charles Sutton, Christopher Yeh, Cody Coleman, Dan Hendrycks, Dan Jurafsky, Daniel Levy, Daphne Koller, David Tellez, Erik Jones, Evan Liu, Fisher Yu, Georgi Marinov, Hongseok Namkoong, Irene Chen, Jacky Kang, Jacob Schreiber, Jacob Steinhardt, Jared Dunnmon, Jean Feng, Jeffrey Sorensen, Jianmo Ni, John Hewitt, John Miller, Kate Saenko, Kelly Cochran, Kensen Shi, Kyle Loh, Li Jiang, Lucy Vasserman, Ludwig Schmidt, Luke Oakden-Rayner, Marco Tulio Ribeiro, Matthew Lungren, Megha Srivastava, Nelson Liu, Nimit Sohoni, Pranav Rajpurkar, Robin Jia, Rohan Taori, Sarah Bird, Sharad Goel, Sherrie Wang, Shyamal Buch, Stefano Ermon, Steve Yadlowsky, Tatsunori Hashimoto, Tengyu Ma, Vincent Hellendoorn, Yair Carmon, Zachary Lipton, and Zhenghao Chen.

The design of the WILDS benchmark was inspired by the Open Graph Benchmark (Hu et al., 2020b), and we are grateful to the Open Graph Benchmark team for their advice and help in setting up our benchmark.

This project was funded by an Open Philanthropy Project Award and NSF Award Grant No. 1805310. Shiori Sagawa was supported by the Herbert Kunzel Stanford Graduate Fellowship. Henrik Marklund was supported by the Dr. Tech. Marcus Wallenberg Foundation for Education in International Industrial Entrepreneurship, CIFAR, and Google. Sang Michael Xie and Marvin Zhang were supported by NDSEG Graduate Fellowships. Weihua Hu was supported by the Funai Overseas Scholarship and the Masason Foundation Fellowship. Sara Beery was supported by an NSF Graduate Research Fellowship and is a PIMCO Fellow in Data Science. Jure Leskovec is a Chan Zuckerberg Biohub investigator. Chelsea Finn is a CIFAR Fellow in the Learning in Machines and Brains Program.

We also gratefully acknowledge the support of DARPA under Nos. N660011924033 (MCS); ARO under Nos. W911NF-16-1-0342 (MURI), W911NF-16-1-0171 (DURIP); NSF under Nos. OAC-1835598 (CINES), OAC-1934578 (HDR), CCF-1918940 (Expeditions), IIS-2030477 (RAPID); Stanford Data Science Initiative, Wu Tsai Neurosciences Institute, Chan Zuckerberg Biohub, Amazon, JPMorgan Chase, Docomo, Hitachi, JD.com, KDDI, NVIDIA, Dell, Toshiba, and UnitedHealth Group.

References

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

Adragna, R., Creager, E., Madras, D., and Zemel, R. Fairness and robustness in invariant learning: A case study in toxicity classification. *arXiv preprint arXiv:2011.06485*, 2020.

Agrawal, A., Batra, D., Parikh, D., and Kembhavi, A. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 4971–4980, 2018.

Ahadi, A., Lister, R., Haapala, H., and Vihavainen, A. 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.

Ahumada, J. A., Fegraus, E., Birch, T., Flores, N., Kays, R., O'Brien, T. G., Palmer, J., Schuttler, S., Zhao, J. Y., Jetz, W., Kinnaird, M., Kulkarni, S., Lyet, A., Thau, D., Duong, M., Oliver, R., and Dancer, A. 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.

Aich, S., Josuttes, A., Ovsyannikov, I., Strueby, K., Ahmed, I., Duddu, H. S., Pozniak, C., Shirtliffe, S., and Stavness, I. Deepwheat: Estimating phenotypic traits from crop images with deep learning. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 323–332. IEEE, 2018.

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

Alexandari, A., Kundaje, A., and Shrikumar, A. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *International Conference on Machine Learning (ICML)*, pp. 222–232, 2020.

Allamanis, M. and Brockschmidt, M. Smartpaste: Learning to adapt source code. *arXiv preprint arXiv:1705.07867*, 2017.

Allamanis, M., Barr, E. T., Bird, C., and Sutton, C. Suggesting accurate method and class names. In *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering*, pp. 38–49, 2015.

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

Ando, D. M., McLean, C. Y., and Berndl, M. Improving phenotypic measurements in high-content imaging screens. *BioRxiv*, pp. 161422, 2017.

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

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

Asuncion, A. and Newman, D. UCI Machine Learning Repository, 2007.

Attene-Ramos, M. S., Miller, N., Huang, R., Michael, S., Itkin, M., Kavlock, R. J., Austin, C. P., Shinn, P., Simeonov, A., Tice, R. R., et al. The tox21 robotic platform for the assessment of environmental chemicals–from vision to reality. *Drug Discovery Today*, 18(15):716–723, 2013.

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

Aviv, R., Teichmann, S. A., Lander, E. S., Ido, A., Christophe, B., Ewan, B., Bernd, B., Campbell, P., Piero, C., Menna, C., et al. The human cell atlas. *eLife*, 6, 2017.

Avsec, Ž., Weilert, M., Shrikumar, A., Alexandari, A., Krueger, S., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., and Zeitlinger, J. Deep learning at base-resolution reveals motif syntax of the cis-regulatory code. *bioRxiv*, 2019.

Ayalew, T. W., Ubbens, J. R., and Stavness, I. Unsupervised domain adaptation for plant organ counting. In *European Conference on Computer Vision*, pp. 330–346. Springer, 2020.

Azizzadenesheli, K., Liu, A., Yang, F., and Anandkumar, A. Regularized learning for domain adaptation under label shifts. In *International Conference on Learning Representations (ICLR)*, 2019.

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

Balaji, Y., Sankaranarayanan, S., and Chellappa, R. Metareg: Towards domain generalization using meta-regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 998–1008, 2018.

Bandi, P., Geessink, O., Manson, Q., Dijk, M. V., Balkenhol, M., Hermsen, M., Bejnordi, B. E., Lee, B., Paeng, K., Zhong, A., 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.

Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., and Katz, B. 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.

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

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

BBC. A-levels and GCSEs: How did the exam algorithm work? *The British Broadcasting Corporation*, 2020. URL https://www.bbc.com/news/explainers-53807730.

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

Becke, A. D. Perspective: Fifty years of density-functional theory in chemical physics. *The Journal of Chemical Physics*, 140(18):18A301, 2014.

Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., and Vardoulakis, L. M. 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.

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

Beery, S., Morris, D., and Yang, S. Efficient pipeline for camera trap image review. *arXiv preprint arXiv:1907.06772*, 2019.

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

Beery, S., Wu, G., Rathod, V., Votel, R., and Huang, J. 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.

Bejnordi, B. E., Veta, M., Diest, P. J. V., Ginneken, B. V., Karssemeijer, N., Litjens, G., Laak, J. A. V. D., Hermsen, M., Manson, Q. F., Balkenhol, M., 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.

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

Bellemare, M. G., Candido, S., Castro, P. S., Gong, J., Machado, M. C., Moitra, S., Ponda, S. S., and Wang, Z. Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature*, 588, 2020.

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

Bender, E. M. and Friedman, B. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics (TACL)*, 6:587–604, 2018.

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

Berman, G., de la Rosa, S., and Accone, T. Ethical considerations when using geospatial technologies for evidence generation. *Innocenti Discussion Paper, UNICEF Office of Research*, 2018.

Beyene, A. A., Welemariam, T., Persson, M., and Lavesson, N. Improved concept drift handling in surgery prediction and other applications. *Knowledge and Information Systems*, 44(1):177–196, 2015.

Blanchard, G., Lee, G., and Scott, C. Generalizing from several related classification tasks to a new unlabeled sample. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2178–2186, 2011.

Blitzer, J., Dredze, M., and Pereira, F. 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.

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

Blodgett, S. L., Green, L., and O'Connor, B. 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.

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

Bohacek, R. S., McMartin, C., and Guida, W. C. The art and practice of structure-based drug design: a molecular modeling perspective. *Medicinal Research Reviews*, 16 (1):3–50, 1996.

Borkan, D., Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. Limitations of pinned AUC for measuring unintended bias. *arXiv preprint arXiv:1903.02088*, 2019a.

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

Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., and Snelson, E. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research (JMLR)*, 14:3207–3260, 2013.

Boutros, M., Heigwer, F., and Laufer, C. Microscopy-based high-content screening. *Cell*, 163(6):1314–1325, 2015.

Bray, M.-A., Singh, S., Han, H., Davis, C. T., Borgeson, B., Hartland, C., Kost-Alimova, M., Gustafsdottir, S. M., Gibson, C. C., and Carpenter, A. E. Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature protocols*, 11 (9):1757, 2016.

Broach, J. R., Thorner, J., et al. High-throughput screening for drug discovery. *Nature*, 384(6604):14–16, 1996.

Broussard, M. When algorithms give real students imaginary grades. *The New York Times*, 2020. URL https://www.nytimes.com/2020/09/08/opinion/international-baccalaureate-algorithm-grades.html.

Bruch, M., Monperrus, M., and Mezini, M. Learning from examples to improve code completion systems. In *European software engineering conference and the ACM SIGSOFT symposium on the foundations of software engineering*, 2009.

Bruzzone, L. and Marconcini, M. 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.

Bug, D., Schneider, S., Grote, A., Oswald, E., Feuerhake, F., Schüler, J., and Merhof, D. 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.

Bunel, R., Hausknecht, M., Devlin, J., Singh, R., and Kohli, P. Leveraging grammar and reinforcement learning for neural program synthesis. In *International Conference on Learning Representations (ICLR)*, 2018.

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

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

Byrd, J. and Lipton, Z. What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning (ICML)*, pp. 872–881, 2019.

Caicedo, J. C., Cooper, S., Heigwer, F., Warchal, S., Qiu, P., Molnar, C., Vasilevich, A. S., Barry, J. D., Bansal, H. S., Kraus, O., et al. Data-analysis strategies for image-based cell profiling. *Nature methods*, 14(9):849–863, 2017.

Caicedo, J. C., McQuin, C., Goodman, A., Singh, S., and Carpenter, A. E. Weakly supervised learning of single-cell feature embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9309–9318, 2018.

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

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

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

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

Carlucci, F. M., D'Innocente, A., Bucci, S., Caputo, B., and Tommasi, T. Domain generalization by solving jigsaw puzzles. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 2229–2238, 2019.

Chanussot, L., Das, A., Goyal, S., Lavril, T., Shuaibi, M., Riviere, M., Tran, K., Heras-Domingo, J., Ho, C., Hu, W., Palizhati, A., Sriram, A., Wood, B., Yoon, J., Parikh, D., Zitnick, C. L., and Ulissi, Z. The Open Catalyst 2020 (oc20) dataset and community challenges. *arXiv preprint arXiv:2010.09990*, 2020.

Chen, I., Johansson, F. D., and Sontag, D. Why is my classifier discriminatory? In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3539–3550, 2018.

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

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

Chen, V., Wu, S., Ratner, A. J., Weng, J., and Ré, C. Slice-based learning: A programming model for residual learning in critical data slices. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 9397–9407, 2019b.

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P., Zietz, M., Hoffman, M. M., et al. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15(141), 2018.

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

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

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

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

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

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

Consortium, E. P. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.

Consortium, G. et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, 369 (6509):1318–1330, 2020.

Consortium, H. et al. The human body at cellular resolution: the NIH human biomolecular atlas program. *Nature*, 574 (7777), 2019.

Corbett-Davies, S. and Goel, S. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

Corbett-Davies, S., Pierson, E., Feller, A., and Goel, S. A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. *Washington Post*, 2016. ISSN 0190-8286. URL https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797–806, 2017.

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

Courtiol, P., Maussion, C., Moarii, M., Pronier, E., Pilcer, S., Sefta, M., Manceron, P., Toldo, S., Zaslavskiy, M., Stang, N. L., et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nature Medicine*, 25(10):1519–1525, 2019.

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

Crunchant, A.-S., Borchers, D., Kühl, H., and Piel, A. 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.

Cuccarese, M. F., Earnshaw, B. A., Heiser, K., Fogelson, B., Davis, C. T., McLean, P. F., Gordon, H. B., Skelly, K., Weathersby, F. L., Rodic, V., et al. Functional immune mapping with deep-learning enabled phenomics applied to immunomodulatory and COVID-19 drug discovery. *bioRxiv*, 2020.

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

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

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020a.

D'Amour, A., Srinivasan, H., Atwood, J., Baljekar, P., Sculley, D., and Halpern, Y. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 525–534, 2020b.

David, E., Madec, S., Sadeghi-Tehran, P., Aasen, H., Zheng, B., Liu, S., Kirchgessner, N., Ishikawa, G., Nagasawa, K., Badhon, M. A., Pozniak, C., de Solan, B., Hund, A., Chapman, S. C., Baret, F., Stavness, I., and Guo, W. Global wheat head detection (gwhd) dataset: a large and diverse dataset of high-resolution rgb-labelled images to develop and benchmark wheat head detection methods. *Plant Phenomics*, 2020, 2020.

David, E., Serouart, M., Smith, D., Madec, S., Velumani, K., Liu, S., Wang, X., Espinosa, F. P., Shafiee, S., Tahir, I.

S. A., Tsujimoto, H., Nasuda, S., Zheng, B., Kichgessner, N., Aasen, H., Hund, A., Sadhegi-Tehran, P., Nagasawa, K., Ishikawa, G., Dandrifosse, S., Carlier, A., Mercatoris, B., Kuroki, K., Wang, H., Ishii, M., Badhon, M. A., Pozniak, C., LeBauer, D. S., Lilimo, M., Poland, J., Chapman, S., de Solan, B., Baret, F., Stavness, I., and Guo, W. Global wheat head dataset 2021: an update to improve the benchmarking wheat head localization with more diversity, 2021.

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

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

Desmarais, M. C. and Baker, R. 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.

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

Dill, K. A. and MacCallum, J. L. The protein-folding problem, 50 years on. *Science*, 338(6110):1042–1046, 2012.

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

Djolonga, J., Yung, J., Tschannen, M., Romijnders, R., Beyer, L., Kolesnikov, A., Puigcerver, J., Minderer, M., D'Amour, A., Moldovan, D., et al. On robustness and transferability of convolutional neural networks. *arXiv preprint arXiv:2007.08558*, 2020.

Dodge, S. and Karam, L. A study and comparison of human and deep learning recognition performance under visual distortions. In *26th International Conference on Computer Communication and Networks (ICCCN)*, pp. 1–7. IEEE, 2017.

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

Dreccer, M. F., Molero, G., Rivera-Amado, C., John-Bejai, C., and Wilson, Z. Yielding to the image: how phenotyping reproductive growth can assist crop improvement and production. *Plant science*, 282:73–82, 2019.

Dressel, J. and Farid, H. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1), 2018.

Duchi, J. and Namkoong, H. Learning models with uniform performance via distributionally robust optimization. *Annals of Statistics*, 2021.

Duchi, J., Hashimoto, T., and Namkoong, H. Distributionally robust losses for latent covariate mixtures. *arXiv preprint arXiv:2007.13982*, 2020.

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

Echeverri, C. J. and Perrimon, N. High-throughput rnai screening in cultured cells: a user's guide. *Nature Reviews Genetics*, 7(5):373, 2006.

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

Eraslan, G., Avsec, Ž., Gagneur, J., and Theis, F. J. Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, 20(7):389–403, 2019.

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

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

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

Fan, Z., Lu, J., Gong, M., Xie, H., and Goodman, E. D. Automatic tobacco plant detection in uav images via deep neural networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(3): 876–887, 2018.

Fang, C., Xu, Y., and Rockmore, D. N. 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.

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

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

Franks, C., Tu, Z., Devanbu, P., and Hellendoorn, V. Cacheca: A cache language model based code suggestion tool. In *International Conference on Software Engineering (ICSE)*, 2015.

Fuentes, A., Yoon, S., Kim, S. C., and Park, D. S. A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. *Sensors*, 17(9):2022, 2017.

Futoma, J., Simons, M., Panch, T., Doshi-Velez, F., and Celi, L. A. The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health*, 2(9):e489–e492, 2020.

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

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

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

Garg, S., Wu, Y., Balakrishnan, S., and Lipton, Z. C. A unified view of label shift estimation. *arXiv preprint arXiv:2003.07554*, 2020.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Ill, H. D., and Crawford, K. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.

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

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

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

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

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., and Wichmann, F. A. Generalisation in humans and deep neural networks. *Advances in Neural Information Processing Systems*, 31:7538–7550, 2018b.

Geirhos, R., Jacobsen, J., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *arXiv preprint arXiv:2004.07780*, 2020.

Gelman, A., Fagan, J., and Kiss, A. An Analysis of the New York City Police Department's "Stop-and-Frisk" Policy in the Context of Claims of Racial Bias. *Journal of the American Statistical Association*, 102(479):813–823, Sep 2007. ISSN 0162-1459. doi: 10.1198/016214506000001040. URL https://amstat.tandfonline.com/doi/abs/10.1198/016214506000001040. Publisher: Taylor & Francis.

Geva, M., Goldberg, Y., and Berant, J. 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.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In *International Conference on Machine Learning (ICML)*, pp. 1273–1272, 2017.

Godinez, W. J., Hossain, I., and Zhang, X. Unsupervised phenotypic analysis of cellular images with multi-scale convolutional neural networks. *BioRxiv*, pp. 361410, 2018.

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

Goel, S., Rao, J. M., and Shroff, R. Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy. *The Annals of Applied Statistics*, 10(1): 365–394, March 2016. ISSN 1932-6157. doi: 10.1214/15-AOAS897. URL http://projecteuclid.org/euclid.aoas/1458909920.

Gogoll, D., Lottes, P., Weyler, J., Petrinic, N., and Stachniss, C. Unsupervised domain adaptation for transferring plant classification systems to new field environments, crops, and robots. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2636–2642. IEEE, 2020.

Goh, W. W. B., Wang, W., and Wong, L. Why batch effects matter in omics data, and how to avoid them. *Trends in biotechnology*, 35(6):498–507, 2017.

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

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

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

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

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

Guo, J., Shah, D. J., and Barzilay, R. Multi-source domain adaptation with mixture of experts. *arXiv preprint arXiv:1809.02256*, 2018.

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

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

Han, X. and Tsvetkov, Y. Fortifying toxic speech detectors against veiled toxicity. *arXiv preprint arXiv:2010.03154*, 2020.

Hand, D. J. Classifier technology and the illusion of progress. *Statistical science*, pp. 1–14, 2006.

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

Harrill, J., Shah, I., Setzer, R. W., Haggard, D., Auerbach, S., Judson, R., and Thomas, R. S. Considerations for strategic use of high-throughput transcriptomics chemical screening data in regulatory decisions. *Current Opinion in Toxicology*, 15:64–75, 2019. ISSN 2468-2020. doi: https://doi.org/10.1016/j.cotox.2019.05.004. URL https://www.sciencedirect.com/science/article/pii/S2468202019300129. Risk Assessment in Toxicology.

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

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

He, Y., Shen, Z., and Cui, P. Towards non-IID image classification: A dataset and baselines. *Pattern Recognition*, 110, 2020.

Heinze-Deml, C. and Meinshausen, N. Conditional variance penalties and domain shift robustness. *arXiv preprint arXiv:1710.11469*, 2017.

Hellendoorn, V. J., Proksch, S., Gall, H. C., and Bacchelli, A. When code completion fails: A case study on real-world completions. In *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*, pp. 960–970. IEEE, 2019.

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

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

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

Hendrycks, D., Basart, S., Mazeika, M., Mostajabi, M., Steinhardt, J., and Song, D. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*, 2020a.

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

Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R., and Song, D. Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100*, 2020c.

Ho, J. W., Jung, Y. L., Liu, T., Alver, B. H., Lee, S., Ikegami, K., Sohn, K., Minoda, A., Tolstorukov, M. Y., Appert, A., et al. Comparative analysis of metazoan chromatin organization. *Nature*, 512(7515):449–452, 2014.

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

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

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

Hu, W., Niu, G., Sato, I., and Sugiyama, M. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning (ICML)*, 2018.

Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., Catasta, M., and Leskovec, J. Open graph benchmark: Datasets for machine learning on graphs. *arXiv preprint arXiv:2005.00687*, 2020b.

Huang, G., Liu, Z., Maaten, L. V. D., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017.

Hughes, J. P., Rees, S., Kalindjian, S. B., and Philpott, K. L. Principles of early drug discovery. *British journal of pharmacology*, 162(6):1239–1249, 2011.

Husain, H., Wu, H.-H., Gazit, T., Allamanis, M., and Brockschmidt, M. Codesearchnet challenge: Evaluating the state of semantic code search. *arXiv preprint arXiv:1909.09436*, 2019.

Jaganathan, K., Panagiotopoulou, S. K., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., Kosmicki, J. A., Arbelaez, J., Cui, W., Schwartz, G. B., et al. Predicting splicing from primary sequence with deep learning. *Cell*, 176(3):535–548, 2019.

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

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

Jin, W., Barzilay, R., and Jaakkola, T. Enforcing predictive invariance across structured biomedical domains. *arXiv preprint arXiv:2006.03908*, 2020.

Johnson, A. E., Pollard, T. J., Shen, L., Li-Wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):1–9, 2016.

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

Jones, E., Sagawa, S., Koh, P. W., Kumar, A., and Liang, P. Selective classification can magnify disparities across groups. In *International Conference on Learning Representations (ICLR)*, 2021.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Tunyasuvunakool, K., Ronneberger, O., Bates, R., Žídek, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Potapenko, A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Steinegger, M., Pacholska, M., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. High accuracy protein structure prediction using deep learning. *Fourteenth Critical Assessment of Techniques for Protein Structure Prediction*, 2020.

Jung, J., Goel, S., Skeem, J., et al. The limits of human predictions of recidivism. *Science Advances*, 6(7), 2020.

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

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

Kallus, N. and Zhou, A. Residual Unfairness in Fair Machine Learning from Prejudiced Data. *arXiv:1806.02887 [cs, stat]*, June 2018. URL http://arxiv.org/abs/1806.02887. arXiv: 1806.02887.

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

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

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

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

Keilwagen, J., Posch, S., and Grau, J. Accurate prediction of cell type-specific transcription factor binding. *Genome Biology*, 20(1), 2019.

Kelley, D. R., Snoek, J., and Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, 26(7): 990–999, 2016.

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

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

Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B. A., Wang, J., Yu, B., Zhang, J., and Bryant, S. H. Pubchem substance and compound databases. *Nucleic Acids Research*, 44(D1):D1202–D1213, 2016b.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

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

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. Concept bottleneck models. In *International Conference on Machine Learning (ICML)*, 2020.

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

Komura, D. and Ishikawa, S. Machine learning methods for histopathological image analysis. *Computational and Structural Biotechnology Journal*, 16:34–42, 2018.

Kulal, S., Pasupat, P., Chandra, K., Lee, M., Padon, O., Aiken, A., and Liang, P. S. Spoc: Search-based pseudocode to code. In *Advances in Neural Information Processing Systems*, pp. 11906–11917, 2019.

Kulkarni, C., Koh, P. W., Huy, H., Chia, D., Papadopoulos, K., Cheng, J., Koller, D., and Klemmer, S. R. Peer and self assessment in massive online classes. *Design Thinking Research*, pp. 131–168, 2015.

Kulkarni, C. E., Socher, R., Bernstein, M. S., and Klemmer, S. R. 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.

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

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015.

Kuznichov, D., Zvirin, A., Honen, Y., and Kimmel, R. Data augmentation for leaf segmentation and counting tasks in rosette plants. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

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

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

Landrum, G. et al. Rdkit: Open-source cheminformatics, 2006.

Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., and Ferrante, E. 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.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. How we analyzed the compas recidivism algorithm. *ProPublica*, 9(1), 2016.

Latessa, E. J., Lemke, R., Makarios, M., and Smith, P. The Creation and Validation of the Ohio Risk Assessment System (ORAS). *Federal Probation*, 74:16, 2010. URL https://heinonline.org/HOL/Page?handle=hein.journals/fedpro74&id=16&div=&collection=.

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

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

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

Li, D., Yang, Y., Song, Y., and Hospedales, T. M. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5542–5550, 2017a.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. Learning to generalize: Meta-learning for domain generalization. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018a.

Li, H. and Guan, Y. Leopard: fast decoding cell type-specific transcription factor binding landscape at single-nucleotide resolution. *bioRxiv*, 2019.

Li, H., Pan, S. J., Wang, S., and Kot, A. C. Domain generalization with adversarial feature learning. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 5400–5409, 2018b.

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

Li, J., Miller, A. H., Chopra, S., Ranzato, M., and Weston, J. Dialogue learning with human-in-the-loop. In *International Conference on Learning Representations (ICLR)*, 2017b.

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

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

Li, Y., Tian, X., Gong, M., Liu, Y., Liu, T., Zhang, K., and Tao, D. Deep domain generalization via conditional invariant adversarial networks. In *European Conference on Computer Vision (ECCV)*, pp. 624–639, 2018c.

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

Libbrecht, M. W. and Noble, W. S. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6):321–332, 2015.

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

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. Delayed impact of fair machine learning. In *International Conference on Machine Learning (ICML)*, 2018.

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

Ljosa, V., Sokolnicki, K. L., and Carpenter, A. E. Annotated high-throughput microscopy image sets for validation. *Nature methods*, 9(7):637–637, 2012.

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

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Lu, S., Guo, D., Ren, S., Huang, J., Svyatkovskiy, A., Blanco, A., Clement, C., Drain, D., Jiang, D., Tang, D., Li, G., Zhou, L., Shou, L., Zhou, L., Tufano, M., Gong, M., Zhou, M., Duan, N., Sundaresan, N., Deng, S. K., Fu, S., and Liu, S. Codexglue: A machine learning benchmark dataset for code understanding and generation. *arXiv preprint arXiv:2102.04664*, 2021.

Lum, K. and Isaac, W. To predict and serve? *Significance*, 13(5):14–19, 2016.

Lum, K. and Shah, T. Measures of fairness for New York City's Supervised Release Risk Assessment Tool. *Human Rights Data Analytics Group*, pp. 21, 2019.

Lyu, J., Wang, S., Balius, T. E., Singh, I., Levit, A., Moroz, Y. S., O'Meara, M. J., Che, T., Algaa, E., Tolmachova, K., et al. Ultra-large library docking for discovering new chemotypes. *Nature*, 566(7743):224–229, 2019.

Macarron, R., Banks, M. N., Bojanic, D., Burns, D. J., Cirovic, D. A., Garyantes, T., Green, D. V., Hertzberg, R. P., Janzen, W. P., Paslay, J. W., et al. Impact of high-throughput screening in biomedical research. *Nature reviews Drug discovery*, 10(3):188, 2011.

Macenko, M., Niethammer, M., Marron, J. S., Borland, D., Woosley, J. T., Guan, X., Schmitt, C., and Thomas, N. E. 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.

Madec, S., Jin, X., Lu, H., De Solan, B., Liu, S., Duyme, F., Heritier, E., and Baret, F. Ear density estimation from high resolution rgb imagery using deep learning technique. *Agricultural and forest meteorology*, 264:225–234, 2019.

Malloy, B. A. and Power, J. F. Quantifying the transition from python 2 to 3: an empirical study of python applications. In *2017 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pp. 314–323. IEEE, 2017.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain adaptation with multiple sources. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1041–1048, 2009.

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

McCloskey, K., Sigel, E. A., Kearnes, S., Xue, L., Tian, X., Moccia, D., Gikunju, D., Bazzaz, S., Chan, B., Clark, M. A., et al. Machine learning on DNA-encoded libraries: A new paradigm for hit finding. *Journal of Medicinal Chemistry*, 2020.

McCoy, R. T., Min, J., and Linzen, T. 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.

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

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

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

Meinshausen, N. and Bühlmann, P. Maximin effects in inhomogeneous large-scale data. *Annals of Statistics*, 43, 2015.

Miller, J., Krauth, K., Recht, B., and Schmidt, L. The effect of natural distribution shift on question answering models. *arXiv preprint arXiv:2004.14444*, 2020.

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

Moore, J. E., Purcaro, M. J., Pratt, H. E., Epstein, C. B., Shoresh, N., Adrian, J., Kawli, T., Davis, C. A., Dobin, A., Kaul, R., et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, 583 (7818):699–710, 2020.

Moult, J., Pedersen, J. T., Judson, R., and Fidelis, K. A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Bioinformatics*, 23(3):ii–iv, 1995.

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Kolawole, T., Fagbohungbe, T., Akinola, S. O., Muhammad, S. H., Kabongo, S., Osei, S., Freshia, S., Niyongabo, R. A., Macharm, R., Ogayo, P., Ahia, O., Meressa, M., Adeyemi, M., Mokgesi-Selinga, M., Okegbemi, L., Martinus, L. J., Tajudeen, K., Degila, K., Ogueji, K., Siminyu, K., Kreutzer, J., Webster, J., Ali, J. T., Abbott, J., Orife, I., Ezeani, I., Dangana, I. A., Kamper, H., Elsahar, H., Duru, G., Kioko, G., Murhabazi, E., van Biljon, E., Whitenack, D., Onyefuluchi, C., Emezue, C., Dossou, B., Sibanda, B., Bassey, B. I., Olabiyi, A., Ramkilowan, A., Öktem, A., Akinfaderin, A., and Bashir, A. 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.

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

Nguyen, A. T. and Nguyen, T. N. Graph-based statistical language model for code. In *International Conference on Software Engineering (ICSE)*, 2015.

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

Nita, M. and Notkin, D. Using twinning to adapt programs to alternative apis. In *2010 ACM/IEEE 32nd International Conference on Software Engineering*, volume 1, pp. 205–214. IEEE, 2010.

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

Norouzzadeh, M. S., Morris, D., Beery, S., Joshi, N., Jojic, N., and Clune, J. A deep active learning system for species identification and counting in camera trap images. *arXiv preprint arXiv:1910.09716*, 2019.

Nygaard, V., Rødland, E. A., and Hovig, E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics*, 17(1):29–39, 2016.

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/.

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

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

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

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

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

Parham, J., Crall, J., Stewart, C., Berger-Wolf, T., and Rubenstein, D. I. Animal population censusing at scale with citizen science and photographic identification. In *AAAI Spring Symposium-Technical Report*, 2017.

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

Parker, H. S. and Leek, J. T. The practical effect of batch on genomic prediction. *Statistical applications in genetics and molecular biology*, 11(3), 2012.

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

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

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

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

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

Peters, J., Bühlmann, P., and Meinshausen, N. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Methodological)*, 78, 2016.

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

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

Pierson, E., Corbett-Davies, S., and Goel, S. Fast Threshold Tests for Detecting Discrimination. *arXiv:1702.08536 [cs, stat]*, March 2018. URL http://arxiv.org/abs/1702.08536. arXiv: 1702.08536.

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

Pipal, K. A., Notch, J. J., Hayes, S. A., and Adams, P. B. 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.

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

Proksch, S., Lerch, J., and Mezini, M. Intelligent code completion with bayesian networks. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 2015.

Proksch, S., Amann, S., Nadi, S., and Mezini, M. Evaluating the evaluations of code recommender systems: A reality check. In *2016 31st IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 2016.

Quang, D. and Xie, X. Factornet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. *Methods*, 166:40–47, 2019.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset shift in machine learning*. The MIT Press, 2009.

Raychev, V., Vechev, M., and Yahav, E. Code completion with statistical language models. In *Proceedings of the 35th ACM SIGPLAN Conference on Programming Language Design and Implementation*, pp. 419–428, 2014.

Raychev, V., Bielik, P., and Vechev, M. Probabilistic model for code with decision trees. *ACM SIGPLAN Notices*, 2016.

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

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning (ICML)*, 2019.

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

Reker, D. Practical considerations for active machine learning in drug discovery. *Drug Discovery Today: Technologies*, 2020.

Ren, S., He, K., Girshick, R., and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

Reynolds, M., Chapman, S., Crespo-Herrera, L., Molero, G., Mondal, S., Pequeno, D. N., Pinto, F., Pinera-Chavez, F. J., Poland, J., Rivera-Amado, C., et al. Breeder friendly phenotyping. *Plant Science*, pp. 110396, 2020.

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

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

Rigaki, M. and Garcia, S. Bringing a GAN to a knife-fight: Adapting malware communication to avoid detection. In *2018 IEEE Security and Privacy Workshops (SPW)*, pp. 70–75, 2018.

Robbes, R. and Lanza, M. How program history can improve code completion. In *International Conference on Automated Software Engineering*, 2008.

Rolf, E., Jordan, M. I., and Recht, B. Post-estimation smoothing: A simple baseline for learning with side information. In *Artificial Intelligence and Statistics (AISTATS)*, 2020.

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., and Lopez, A. M. 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.

Rosenfeld, A., Zemel, R., and Tsotsos, J. K. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018.

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

Sadeghi-Tehran, P., Virlet, N., Sabermanesh, K., and Hawkesford, M. J. Multi-feature machine learning model for automatic segmentation of green fractional vegetation cover for high-throughput field phenotyping. *Plant methods*, 13(1):1–16, 2017.

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

Saerens, M., Latinne, P., and Decaestecker, C. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural Computation*, 14(1):21–41, 2002.

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

Sagawa, S., Raghunathan, A., Koh, P. W., and Liang, P. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning (ICML)*, 2020b.

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

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

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

Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N. A. The risk of racial bias in hate speech detection. In *Association for Computational Linguistics (ACL)*, 2019.

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

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

Shakoor, N., Lee, S., and Mockler, T. C. High throughput phenotyping to accelerate crop breeding and monitoring of diseases in the field. *Current opinion in plant biology*, 38:184–192, 2017.

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., and Sculley, D. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *Advances in Neural Information Processing Systems (NeurIPS) Workshop on Machine Learning for the Developing World*, 2017.

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

Shapiro, A., Dentcheva, D., and Ruszczyński, A. *Lectures on stochastic programming: modeling and theory*. SIAM, 2014.

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

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

Shetty, R., Schiele, B., and Fritz, M. 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.

Shi, Y., Thomasson, J. A., Murray, S. C., Pugh, N. A., Rooney, W. L., Shafian, S., Rajan, N., Rouze, G., Morgan, C. L. S., Neely, H. L., Rana, A., Bagavathiannan, M. V., Henrickson, J., Bowden, E., Valasek, J., Olsenholler, J., Bishop, M. P., Sheridan, R., Putman, E. B., Popescu, S., Burks, T., Cope, D., Ibrahim, A., McCutchen, B. F., Baltensperger, D. D., Avant, Jr, R. V., Vidrine, M., and Yang, C. Unmanned aerial vehicles for high-throughput phenotyping and agronomic research. *PLOS ONE*, 11(7):1–26, 07 2016. doi: 10.1371/journal.pone.0159781. URL https://doi.org/10.1371/journal.pone.0159781.

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

Shin, R., Kant, N., Gupta, K., Bender, C., Trabucco, B., Singh, R., and Song, D. Synthetic datasets for neural program synthesis. In *International Conference on Learning Representations (ICLR)*, 2019.

Shiu, Y., Palmer, K., Roch, M. A., Fleishman, E., Liu, X., Nosal, E.-M., Helble, T., Cholewiak, D., Gillespie, D., and Klinck, H. Deep neural networks for automated detection of marine mammal species. *Scientific Reports*, 10(1):1–12, 2020.

Shoichet, B. K. Virtual screening of chemical libraries. *Nature*, 432(7019):862–865, 2004.

Slack, D., Friedler, S., and Givental, E. Fairness Warnings and Fair-MAML: Learning Fairly with Minimal Data. *arXiv:1908.09092 [cs, stat]*, December 2019. URL http://arxiv.org/abs/1908.09092. arXiv: 1908.09092.

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

Soneson, C., Gerster, S., and Delorenzi, M. Batch effect confounding leads to strong bias in performance estimates obtained by cross-validation. *PloS one*, 9(6):e100335, 2014.

Srivastava, D. and Mahony, S. Sequence and chromatin determinants of transcription factor binding and the establishment of cell type-specific binding patterns. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1863(6), 2020.

Srivastava, M., Hashimoto, T., and Liang, P. Robustness to Spurious Correlations via Human Annotations. In *International Conference on Machine Learning*, pp. 9109–9119. PMLR, November 2020. URL http://proceedings.mlr.press/v119/srivastava20a.html. ISSN: 2640-3498.

Sterling, T. and Irwin, J. J. Zinc 15 – ligand discovery for everyone. *Journal of Chemical Information and Modeling*, 55(11):2324–2337, 2015. doi: 10.1021/acs.jcim.5b00559. PMID: 26479676.

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

Subbaswamy, A., Adams, R., and Saria, S. Evaluating model robustness to dataset shift. *arXiv preprint arXiv:2010.15100*, 2020.

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

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

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhao, S., Cheng, S., Zhang, Y., Shlens, J., Chen, Z., and Anguelov, D. Scalability in perception for autonomous driving: Waymo open dataset. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020a.

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

Svyatkovskiy, A., Zhao, Y., Fu, S., and Sundaresan, N. Pythia: ai-assisted code completion system. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2727–2735, 2019.

Swinney, D. C. and Anthony, J. How were new medicines discovered? *Nature reviews Drug discovery*, 10(7):507, 2011.

Tabak, G., Fan, M., Yang, S., Hoyer, S., and Davis, G. Correcting nuisance variation using wasserstein distance. *PeerJ*, 8:e8594, 2020.

Tabak, M. A., Norouzzadeh, M. S., Wolfson, D. W., Sweeney, S. J., VerCauteren, K. C., Snow, N. P., Halseth, J. M., Di Salvo, P. A., Lewis, J. S., White, M. D., 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.

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

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

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

Taylor, J., Earnshaw, B., Mabey, B., Victors, M., and Yosinski, J. Rxrx1: An image set for cellular morphological variation across many experimental batches. In *International Conference on Learning Representations (ICLR)*, 2019.

Taylor, M. J., Lukowski, J. K., and Anderton, C. R. Spatially resolved mass spectrometry at the single cell: Recent innovations in proteomics and metabolomics. *Journal of the American Society for Mass Spectrometry*, 32(4): 872–894, 2021.

Tellez, D., Balkenhol, M., Otte-Höller, I., van de Loo, R., Vogels, R., Bult, P., Wauters, C., Vreuls, W., Mol, S., Karssemeijer, N., 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.

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

Temel, D., Lee, J., and AlRegib, G. 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.

Thorp, K. R., Thompson, A. L., Harders, S. J., French, A. N., and Ward, R. W. High-throughput phenotyping of crop water use efficiency via multispectral drone imagery and a daily soil water balance model. *Remote Sensing*, 10 (11):1682, 2018.

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

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

Toda, Y. and Okura, F. How convolutional neural networks diagnose plant disease. *Plant Phenomics*, 2019, 2019.

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

Tuschl, T. Rna interference and small interfering rnas. *Chembiochem*, 2(4):239–245, 2001.

Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., and Darrell, T. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

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

Ubbens, J. R., Ayalew, T. W., Shirtliffe, S., Josuttes, A., Pozniak, C., and Stavness, I. Autocount: Unsupervised segmentation and counting of organs in field images. In *European Conference on Computer Vision*, pp. 391–399. Springer, 2020.

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

Vasic, M., Kanade, A., Maniatis, P., Bieber, D., and Singh, R. Neural program repair by jointly learning to localize and repair. In *International Conference on Learning Representations (ICLR)*, 2019.

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

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

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

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

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

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

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019a.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations (ICLR)*, 2019b.

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

Wang, H., Ge, S., Lipton, Z., and Xing, E. P. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019c.

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

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

Ward, D. and Moghadam, P. Scalable learning for bridging the species gap in image-based plant phenotyping. *Computer Vision and Image Understanding*, 197:103009, 2020.

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

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

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

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

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

Widmer, G. and Kubat, M. Learning in the presence of concept drift and hidden contexts. *Machine learning*, 23 (1):69–101, 1996.

Williams, J. J., Kim, J., Rafferty, A., Maldonado, S., Gajos, K. Z., Lasecki, W. S., and Heffernan, N. 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.

Wilson, B., Hoffman, J., and Morgenstern, J. Predictive inequity in object detection. *arXiv preprint arXiv:1902.11097*, 2019.

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

Wong, H. Y. F., Lam, H. Y. S., Fong, A. H.-T., Leung, S. T., Chin, T. W.-Y., Lo, C. S. Y., Lui, M. M.-S., Lee, J. C. Y., Chiu, K. W.-H., Chung, T., et al. Frequency and distribution of chest radiographic findings in covid-19 positive patients. *Radiology*, pp. 201160, 2020.

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

Wu, M., Mosse, M., Goodman, N., and Piech, C. 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.

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

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

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. Moleculenet: a benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018.

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

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

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

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

Xiong, H., Cao, Z., Lu, H., Madec, S., Liu, L., and Shen, C. Tasselnetv2: in-field counting of wheat spikes with context-augmented local regression networks. *Plant Methods*, 15(1):1–14, 2019.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? In *International Conference on Learning Representations (ICLR)*, 2018.

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

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

Yasunaga, M. and Liang, P. Graph-based, self-supervised program repair from diagnostic feedback. In *International Conference on Machine Learning (ICML)*, 2020.

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

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

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

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

Zafar, M. B., Valera, I., Rodriguez, M. G., Gummadi, K. P., and Weller, A. From Parity to Preference-based Notions of Fairness in Classification. *arXiv:1707.00010 [cs, stat]*, Nov 2017. URL http://arxiv.org/abs/1707.00010. arXiv: 1707.00010.

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

Zhang, K., Schölkopf, B., Muandet, K., and Wang, Z. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning (ICML)*, pp. 819–827, 2013.

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

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

Zhao, J., Wang, T., Yatskar, M., Ordoñez, V., and Chang, K. Gender bias in coreference resolution: Evaluation and debiasing methods. In *North American Association for Computational Linguistics (NAACL)*, 2018.

Zhou, J. and Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10):931–934, 2015.

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

Zhou, Y., Zhu, S., Cai, C., Yuan, P., Li, C., Huang, Y., and Wei, W. High-throughput screening of a crispr/cas9 library for functional genomics in human cells. *Nature*, 509(7501):487, 2014.

Zitnick, C. L., Chanussot, L., Das, A., Goyal, S., Heras-Domingo, J., Ho, C., Hu, W., Lavril, T., Palizhati, A., Riviere, M., Shuaibi, M., Sriram, A., Tran, K., Wood, B., Yoon, J., Parikh, D., and Ulissi, Z. An introduction to electrocatalyst design using machine learning for renewable energy storage. *arXiv preprint arXiv:2010.09435*, 2020.