A LEARNING BASED HYPOTHESIS TEST FOR HARMFUL COVARIATE SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

The ability to quickly and accurately identify covariate shift at test time is a critical and often overlooked component of safe machine learning systems deployed in high-risk domains. While methods exist for detecting when predictions should not be made on out-of-distribution test examples, identifying distributional level differences between training and test time can help determine when a model should be removed from the deployment setting and retrained. In this work, we define *harmful covariate shift* (HCS) as a change in distribution that may weaken the generalization of a predictive model. To detect HCS, we use the discordance between an ensemble of classifiers trained to agree on training data and disagree on test data. We derive a loss function for training this ensemble and show that the disagreement rate and entropy represent powerful discriminative statistics for HCS. Empirically, we demonstrate the ability of our method to detect harmful covariate shift with statistical certainty on a variety of high-dimensional datasets. Across numerous domains and modalities, we show state-of-the-art performance compared to existing methods, particularly when the number of observed test samples is small[1].

1 INTRODUCTION

Machine learning models operate on the assumption, albeit incorrectly that they will be deployed on data distributed identically to what they were trained on. The violation of this assumption is known as distribution shift and can often result in significant degradation of performance (Bickel et al., 2009; Rabanser et al., 2019; Otles et al., 2021; Ovadia et al., 2019). There are several cases where a mismatch between training and deployment data results in very real consequences on human beings. In healthcare, machine learning models have been deployed for predicting the likelihood of sepsis. Yet, as (Habib et al., 2021) show, such models can be miscalibrated for large groups of individuals, directly affecting the quality of care they experience. The deployment of classifiers in the criminal justice system (Hao, 2019), hiring and recruitment pipelines (Dastin, 2018) and self-driving cars (Smiley, 2022) have all seen humans affected by the failures of learning models. The need for methods that quickly detect, characterize and respond to covariate shift is, therefore, a fundamental problem in trustworthy machine learning. We study a special case of distribution shift, commonly known as *covariate shift*, which considers shifts only in the distribution of input data $\mathbb{P}(X)$ while the relation between the inputs and outputs $\mathbb{P}(Y|X)$ remains fixed. In a standard deployment setting where ground truth labels are not available, covariate shift is the only type of distribution shift that can be identified.

For practitioners, regulatory agencies and individuals to have faith in deployed predictive models without the need for laborious manual audits, we need methods for the identification of covariate shift that are *sample-efficient* (identifying shifts from a small number of samples), *informed* (identifying shifts relevant to the domain and learning algorithm), *model-agnostic* (identifying shifts regardless of the functional class of the predictive model) and *statistically sound* (identifying true shifts while avoiding false positives with high-confidence).

We build off recent progress in understanding model performance under covariate shift using the PQ-learning framework (Goldwasser et al., 2020), a framework for selective classifiers that may

[1]Code available at `https://anonymous.4open.science/r/deeptst-4483`

either predict on or reject a given sample, that provides strong performance guarantees on arbitrary test distributions. Our work uses and extends PQ-learning to develop a practical, model-based hypothesis test, named *the Detectron*, to identify potentially harmful covariate shifts given any existing classification model already in deployment.

Our work makes the following key contributions:

- We show how to construct an ensemble of classifiers that maximize out-of-domain disagreement while behaving consistently in the training domain. We propose the *disagreement cross entropy* for models learned via continuous gradient-based methods (e.g., neural networks), as well as a generalization for those learned via discrete optimization (e.g., random forest).
- We show that the rejection rate and the entropy of the learning ensemble can be used to define a model-aware hypothesis test for covariate shift, the Detectron, that in idealized settings can provably detect covariate shift.
- On high-dimensional image and tabular data, using both neural networks and gradient boosted decision trees, our method outperforms state-of-the-art techniques for detecting covariate shift, particularly when given access to as few as ten test examples.

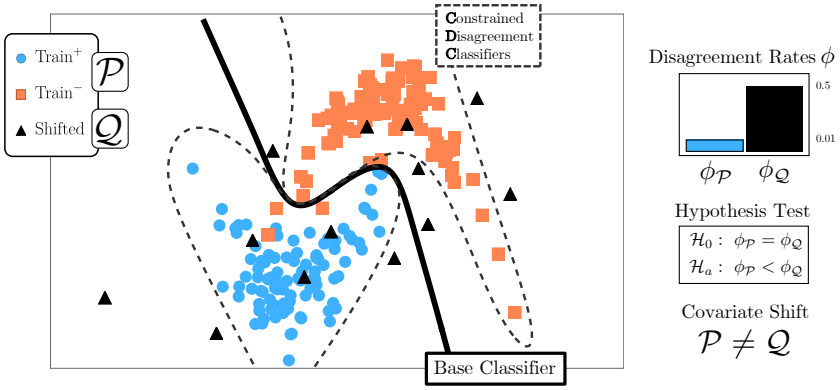

Figure 1: **Detectron (PQ Learning for Covariate Shift Detection):** Starting with a base classifier trained on labeled samples from distribution $\mathcal{P}$ we train new ***C**onstrained **D**isagreement **C**lassifiers* (CDCs) on a small set of observed unlabeled samples from a new distribution $\mathcal{Q}$. CDCs aim to maximize classification disagreement on $\mathcal{Q}$ while constrained to agree with the base classifier on $\mathcal{P}$. The rate ϕ that CDCs disagree, as well as their entropy, on test data is a powerful and sample efficient statistic for identifying covariate shift $\mathcal{P} \neq \mathcal{Q}$.

2 Background and Related Work

Covariate Shift Detection. Covariate shift is the tendency for a distribution at test time $p_{\text{test}}(x)$ to differ from that seen during training $p_{\text{train}}(x)$ while the underlying prediction concept y remains fixed e.g. $p_{\text{train}}(y|x) = p_{\text{test}}(y|x)$. Many methods for detecting shift apply dimensionality reduction followed by statistical hypothesis tests for distributional differences in the outputs (from a reference and target) (Rabanser et al., 2019). However, applying statistical tests to non-invertible representations of data can never guarantee to capture arbitrary covariate shifts, as there may always exist multiple distributions that collapse to the same test statistic (Zhang et al., 2021). Kifer et al. (2004); Ben-David et al. (2006) introduce some of the earliest learning theoretic approaches for identifying and correcting for covariate shift based on discriminative learning with finite samples. More recent approaches for covariate shift detection including classifier two sample tests (Lopez-Paz & Oquab, 2017), deep kernel MMD (Liu et al., 2020) and H-Divergence (Zhao et al., 2022) rely on analyzing the outputs of unsupervised learning models (see Appendix subsection E.3 for more details). In our work we take a transductive learning approach and construct a method to directly use the structure of a supervised classification problem to improve the statistical power for detecting shifts.

Out of Distribution Detection. Out of distribution (OOD) detection focuses on identifying when a specific data point x' admits low likelihood under the original training distributions ($p_{\text{train}}(x') \approx 0$)—a useful tool to have at inference time. Ren et al. (2019); Morningstar et al. (2021) represent a broad class of work that uses density estimation to pose the identification of covariate shift as anomaly detection.

However, in finite samples, density estimation for high-dimensional data can be difficult which in turn affects the accuracy of anomaly detection (Zhang et al., 2021). Others, including ODIN (Liang et al., 2018), Deep Mahalanobis Detectors (Lee et al., 2018) and, Gram Matrices (Sastry & Oore, 2020) directly use the predictive model (e.g. information from the intermediate representations of neural networks). Such methods are largely based on heuristics on the manifold of neural networks offering little to no theoretical guarantees on detecting subtle types of covariate shifts encountered in real-world settings. Furthermore, the majority of methods in this space have been designed exclusively for deep neural networks, an uncommon modelling choice particularly for tabular data (Borisov et al., 2021). Related to OOD, uncertainty estimation concerns developing models that identify sources of uncertainty in their predictions (Lakshminarayanan et al., 2017; Ovadia et al., 2019). Naturally, uncertainty should be large when samples are OOD, however (Ovadia et al., 2019) perform a large-scale empirical comparison of uncertainty estimation methods and find that while deep ensembles generally provide the best results, the quality of uncertainty estimations, regardless of method, consistently degrades with increasing covariate shift.

Selective Classification and PQ Learning. Selective classification concerns building classifiers that may either predict on or reject on test samples (Geifman & El-Yaniv, 2019). Recent work by (Goldwasser et al., 2020) develops a formal framework known as PQ learning which extends probably approximately correct (PAC) learning (Haussler, 1990) to arbitrary test distributions by allowing for selective classification. While PAC learning concerns the development of a classifier with a bounded finite-sample error rate on its training distribution, PQ learning seeks a selective classifier with jointly bounded finite-sample error and rejection rates on arbitrary test distributions. The Rejectron algorithms proposed therein builds an ensemble of models that produce different outputs relative to a perfect baseline on a set of unlabeled test samples. We provide a summary of the original Rejectron algorithm in the supplementary material (see Appendix A). PQ-learning represents a major theoretical leap for learning guarantees under covariate shift; however, the majority of the underlying ideas have not been implemented/tested experimentally using real-world data. We show how to build a PQ learner by generalizing the Rejectron algorithm, overcoming several limitations and assumptions made by the original work including extending beyond simple binary classification to general multiclass/multilabel tasks and reducing the number of samples required for learning at each iteration. We go on to show how a PQ learner can be used to characterize covariate shifts in real-world data.

3 DETECTRON

Problem Setup. Let $f : X \to Y$ be classification model from a function class F that maps from space of covariates X to a discrete set of classes $Y = \{1, \ldots, N\}$. We assume f was trained on a dataset $\mathbf{P}$ consisting of labeled samples $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ where each x_i is drawn identically from a distribution $\mathcal{P}$ over X. In deployment, f is then made to predict on new unlabeled samples $\mathbf{Q}$ from a distribution $\mathcal{Q}$ over X. Our goal is to determine whether f may be trusted to do so accurately. The problem we address is how to automatically detect, from only a set of finite samples $\mathbf{Q}$, if the new covariate distribution $\mathcal{Q}$ has shifted from $\mathcal{P}$ in such a way that f can no longer be assumed to generalize — we refer to this type of dataset shift as *harmful covariate shift*.

Harmful Covariate Shift. A shift in the data distribution is not always harmful. In many practical problems, a practitioner may use domain knowledge to embed invariances with the explicit goal of ensuring the predictive performance of a classifier does not, by construction, change under certain shifts. This may be done directly via translation invariance in convolutional neural networks, permutation invariance in DeepSets (Zaheer et al., 2017) or indirectly via data augmentation or domain adaptation. Such practical heuristics can lead to models generalizing to a more broad range of distributions than can be characterized by just the training set. We refer to such an induced generalization set as $\mathcal{R}$. Although $\mathcal{R}$ is difficult to characterize and will in general depend on the model architecture, learning algorithm and training dataset, we seek a practical method for detecting shift that is explicitly tied to $\mathcal{R}$[2]. We present a more formal definition of *harmful covariate shift* as well as a connection to domain identification and $\mathcal{A}$ distance (Kifer et al., 2004) in Appendix H.

Our approach is based both on PQ learning and intuition from learning theory. If there exists a set of classifiers with the same generalization set $\mathcal{R}$ but behave inconsistently on samples from a

[2]See Appendix G for further intuition and a concrete example of the generalization set.

distribution $\mathcal{Q}$, then $\mathcal{Q}$ must not be a member $\mathcal{R}$. Our strategy will be to create an ensemble of *constrained disagreement classifiers*, classifiers constrained to predict consistently (i.e., predict the same as f) on $\mathcal{R}$ but as differently as possible on $\mathcal{Q}$. If $\mathcal{Q}$ is within $\mathcal{R}$ then such an ensemble will fail to predict differently. When we can find an ensemble that exhibits inconsistent behaviour on $\mathcal{Q}$, there must be covariate shift that explicitly lies outside $\mathcal{R}$. To make the idea of constrained disagreement classifiers tangible we propose a simple definition which we will translate into a learning algorithm in the following sections.

Constrained Disagreement Classifier (CDC). A constrained disagreement classifier $g_{(f,\mathbf{P},\mathbf{Q})}$, or simply g if f, $\mathbf{P}$ and $\mathbf{Q}$ are clear with context, is a classifier with the following properties:

1. g belongs to the same model class as $f \in F$ and is trained with the same learning algorithm on $\mathbf{P}$
2. g achieves similar held out performance on $\mathcal{P}$
3. g disagrees maximally with f on elements of dataset $\mathbf{Q}$ while not violating 1 and 2

Our definition of a CDC aims to explicitly capture the concept of a classifier that learns the same generalization region as f while behaving as inconsistently as possible on $\mathbf{Q}$[3].

Limitations of PQ Learning.

As our work builds on PQ learning, we provide a summary of the original framework and clearly state the distinctions in our methodology. In PQ learning we seek a selective classifier h that achieves a bounded tradeoff between its in distribution rejection rate $\text{rej}_h(\mathbf{x})$ and its out of distribution error $\text{err}_h(\tilde{\mathbf{x}})$ with respect to a ground truth decision function d. Formally, this tradeoff is defined using the following learning theoretic bound.

PQ learning (Goldwasser et al., 2020) Learner L (ϵ, δ, n)-PQ-learns F if for any distributions $\mathcal{P}, \mathcal{Q}$ over X and any ground truth function $d \in F$, its output $h := L(\mathbf{P}, d(\mathbf{P}), \mathbf{Q})$ satisfies

$$\mathbb{P}_{\mathbf{x}\sim\mathcal{P}^n,\ \tilde{\mathbf{x}}\sim\mathcal{Q}^n}\left[\text{rej}_h(\mathbf{x}) + \text{err}_h(\tilde{\mathbf{x}}) \leq \epsilon\right] \geq 1 - \delta \tag{1}$$

Goldwasser et al. propose the Rejectron algorithm for PQ learning in a noiseless binary classification setting with zero training error and access to a perfect empirical risk minimization (ERM) oracle. Rejectron sequentially searches for new classifiers that achieve zero training error while attempting to predict the opposite label on subsets of the unlabeled set $\mathbf{Q}$. We highlight several pitfalls that prevent a practical realization of Rejectron (1) the classification problem must be binary (2) the optimization objective at each step requires an ERM query with $\Omega(|\mathbf{P}|^2)$ samples and (3) most significantly there is no process to control overfitting. By tackling the above issues we derive a practical PQ learner and propose a powerful hypothesis test to identify harmful covariate shift.

Learning to Disagree with the Disagreement Cross Entropy. To train a classifier to disagree in the binary setting, it suffices to flip the labels. However, in the multi-class classification, it is unclear what a good objective function is. We formulate an explicit loss function that can be minimized via gradient descent to learn a CDC. For classification problems, letting $\hat{y} := g(x_i)$ be the predictive distribution over N classes, $f(x_i) \in \{1, \ldots N\}$ the label predicted by f and $[\cdot]$ a binary indicator, we define the *disagreement-cross-entropy* $\tilde{\ell}$ as:

$$\tilde{\ell}(\hat{y}, f(x_i)) = \frac{1}{1-N}\sum_{c=1}^{N}[f(x_i) \neq c]\log(\hat{y}_c) \tag{2}$$

$\tilde{\ell}$ corresponds to taking the cross entropy of $\hat{y}$ with the uniform distribution over all classes except $f(x_i)$. Since the primary criteria is that $g(x_i)$ disagrees with $f(x_i)$, $\tilde{\ell}$ is designed to minimize the probability of g's prediction for the output of f's while maximizing its overall entropy. Our definition of $\tilde{\ell}$ is significantly more stable to optimize compared to simply maximizing the regular cross entropy as it has a bounded global minimum.

Our goal is to agree on $\mathbf{P}$ and disagree on $\mathbf{Q}$. Consequently, we learn with the loss in Equation 3. ℓ denotes the standard cross entropy loss and $\tilde{\ell}$ is the disagreement cross entropy. λ is a scalar parameter that controls the trade off between agreement and disagreement.

$$\mathcal{L}_g(\{x_1, \ldots x_n\}) = \frac{1}{n}\left(\sum_{i=1}^{n}\ell(g(x_i), y_i)[x_i \in \mathbf{P}] + \lambda\sum_{i=1}^{n}\tilde{\ell}(g(x_i), f(x_i))[x_i \in \mathbf{Q}]\right) \tag{3}$$

[3]A discussion on the relation between model complexity of F and the behaviour of CDCs and associated limitations of our defintion is given in Appendix I.

When learning CDCs in practice, $\mathcal{L}_g$ should be combined with any additional regularization and data augmentation used in the original training process of f to ensure that we retain the true generalization region of f. Furthermore, training and validation metrics must be closely monitored on unseen samples from $\mathcal{P}$ to ensure that g achieves similar generalization performance on $\mathcal{P}$.

Choosing λ. In the original formulation of Rejectron, selective classifiers are trained on a dataset consisting of $\mathbf{P}$ replicated $|\mathbf{P}|$ times and $\mathbf{Q}$. Calling an ERM oracle on this data ensures that a misclassification on $\mathbf{P}$ is significantly more costly than one on $\mathbf{Q}$ but requires $\Omega(|\mathbf{P}|^2)$ samples, an impractical number for large datasets. We show that instead we can choose the scalar parameter λ in Equation 3 to set learning $\mathbf{P}$ as the primary learning objective and only when it cannot be improved, we allow g to learn how to disagree on $\mathbf{Q}$. The reasoning is a simple counting argument. Suppose agreeing on each sample in $\mathbf{P}$ incurs a reward of 1 and disagreeing with each sample in $\mathbf{Q}$ a reward of λ. To encourage agreement on $\mathbf{P}$ as the primary objective, we set λ such that the extra reward obtained by going from *zero* to *all* disagreements on $\mathbf{Q}$ is less than that achieved with only one extra agreement on $\mathbf{P}$, this gives $\lambda|\mathbf{Q}| < 1$. Practically, we chose $\lambda = 1/(|\mathbf{Q}| + 1)$ and find that no tuning is required.

Some predictive models require learning with non-differentiable loss functions and/or discrete optimization (e.g., random forest). In such cases we fall back and formulate a general disagreement loss by duplicating each sample in $\mathbf{Q}$ $(N-1)$ times, giving each a unique label that is not the target and weight of $1/(N-1)$. In the continuous case, this corresponds exactly to Equation 2.

To learn richer disagreement rules, we create an ensemble of CDCs where the kth model is trained only to disagree on the subset of $\mathbf{Q}$ that has yet to be disagreed on by models 1 through $k-1$. The final disagreement rate $\phi_{\mathbf{Q}}$ is the fraction of unlabelled samples where any CDC provides an alternate decision from f. In what follows we use this rate to characterize shift.

From Constrained Disagreement to Detecting Shift with Hypothesis Tests. A natural way to apply the concept of constrained disagreement to the identification of covariate shift is to partition $\mathbf{Q}$ into two sets, using the first to train a CDC ensemble and the second to compute an unbiased estimate of its held out disagreement rate $\phi_{\mathcal{Q}}$. We would statistically compare this disagreement rate using a 2×2 exact hypothesis test against a baseline estimate for the disagreement rate on $\mathcal{P}$. The following shows that this results in a provably correct method to detect shift.

Theorem 1 (Disagreement implies covariate shift)**.** Let f be a classifier trained on dataset $\mathbf{P}$ consisting of N samples drawn identically from $\mathcal{P}$ and their corresponding labels. Let g be a classifier that is observed to agree (classify identically) with f on $\mathbf{P}$ and disagree on a dataset $\mathbf{Q}$ drawn from $\mathcal{Q}$. If the rate which g disagrees with f on n unseen samples from $\mathcal{Q}$ is greater then that from n unseen samples from $\mathcal{P}$ w.p. greater than $p^\star := \frac{1}{2}\left(1 - 4^{-n}\binom{2n}{n}\right)$ there must be covariate shift.

Sketch of Proof. We show that under the null hypothesis where $\mathcal{P} = \mathcal{Q}$ the tightest upper bound on the probability that g is more likely to disagree on $\mathcal{Q}$ compared to $\mathcal{P}$ is $p^\star$. The contrapositive argument then states if we deem the probability to be greater than $p^\star$ there must be a covariate shift. This result motivates a hypothesis testing approach to determine how probable it is that g is truly more likely to disagree on $\mathcal{Q}$ given only a set of finite observations. The full proof can be found in Appendix C.

Remark. While Theorem 1 takes a frequentist's approach to identifying covariate shift, we show that there is a Bayesian formulation which can provide a closed-form and informative lower bound on the probability of covariate shift given an observation of finite sample disagreement rates on $\mathbf{P}$ and $\mathbf{Q}$.

Our theory, while simple, has a limitation that prevents its direct application. Any approach that requires unseen samples from $\mathbf{Q}$ is ill-suited for the low data regime, as it requires splitting $\mathbf{Q}$ leaving an even smaller set for computing the disagreement rate. Estimators from small samples result in high variance and ultimately low statistical power. Since our objective is to detect covariate shift from as few test samples as possible, splitting $\mathbf{Q}$ is not a good option. To tackle this issue practically, we take a transductive approach based on intuition from learning theory: creating learning models to disagree on samples from $\mathbf{Q}$ while generalizing to $\mathcal{R}$ is a far easier task when $\mathcal{Q}$ is not in $\mathcal{R}$. We can therefore use the *relative increase* in disagreement between CDCs on $\mathbf{Q}$ and $\mathbf{P}$ to capture a quantity that is nearly as informative as the unbiased statistic without reducing samples from $\mathbf{Q}$ that we can use.

The Detectron Test. Our proposed method is to train two CDC ensembles $g_{\mathbf{Q}}$ and $g_{\mathbf{P}}$. $g_{\mathbf{Q}}$ is trained to disagree on **all** of $\mathbf{Q}$ and a $g_{\mathbf{P}}$ is a baseline trained to disagree an unseen set $\mathbf{P}^\star$ from $\mathcal{P}$. Once trained,

we compute the ratio of samples $\phi_{\mathbf{Q}}$ and $\phi_{\mathbf{P}}$ that $g_{\mathbf{Q}}$ and $g_{\mathbf{P}}$ learn to disagree on their sets respectively. Under the null hypothesis where $\mathcal{Q}$ belongs to the generalization region of f and $\mathcal{P}$, $\mathbb{E}[\phi_{\mathbf{Q}}] \leq \mathbb{E}[\phi_{\mathbf{P}}]$, while harmful shift is expressed as a one sided alternative $H_a : \mathbb{E}[\phi_{\mathbf{Q}}] > \mathbb{E}[\phi_{\mathbf{P}}]$ (i.e., it is easier to learn how to reject on $\mathcal{Q}$ compared to $\mathcal{P}$). We refer to this test as *Detectron (Disagreement)*. To compute the test result at a significance level α we first estimate the null distribution of $\phi_{\mathbf{P}}$ for a fixed sample size N by training the Detectron for K calibration rounds with different random seeds and sets $\mathbf{P}^\star$. The test result is significant if the observed disagreement rate $\phi_{\mathbf{Q}}$ is greater than the $(1-\alpha)$ quantile of the null distribution. For more information on the testing procedure see algorithm 1 below and a detailed description in Appendix B.

We consider an additional variant, *Detectron (Entropy)*, which computes the prediction entropy of each sample under the CDC ensemble instead of relying solely on disagreement rates. The CDC entropy is computed from the mean probabilities over each N classes of the base classifier f and set of k CDCs $g_1, \ldots, g_k$.

$$\text{CDC}_{\text{entropy}}(x) = -\sum_{c=1}^{N} \hat{p}_c \log(\hat{p}_c) \;\text{ where }\; \hat{p} := \frac{1}{k+1}\left(f(x) + \sum_{i=1}^{k} g_i(x)\right) \tag{4}$$

We use a KS test to compute a p-value for covariate shift directly on the entropy distributions computed for $\mathbf{Q}$ and $\mathbf{P}^\star$ and guarantee significance using the same strategy as above. The intuition for *Detectron (Entropy)* draws from the fact that when CDCs satisfy their objective (i.e., in the case of harmful shift) they learn to predict with high entropy on $\mathbf{Q}$ and low entropy on $\mathbf{P}^\star$, resulting in a natural way to distinguish between distributions.

Algorithm 1: The Detectron algorithm for detecting harmful covariate shift

Input: $\mathbf{P}$: labeled dataset, $\mathbf{Q}$: unlabeled dataset, L: learning algorithm,
K: calibration rounds $= 100$, $\aleph$: ensemble size $= 5$, α: significance level $= 0.05$
Output: test result for harmful covariate shift at significance level α

```
P_train, P_val, P* ← Partition(P)
N ← |Q|; Φ_P ← []
f ← L(P_train, P_val)  # Load or train a base classifier on P
repeat
    p* ← RandomSample(P*, N)
    while n > 0 and iterations ≤ ℵ do  # Train an ensemble of CDCs on P*
        # See Appendix algorithm 3 for ConstrainedDisagreement
        g ← ConstrainedDisagreement(L, P_train, P_val, p*, f)
        p* ← {x | x ∈ p* and f(x) = g(x)}  # Filter out disagreed on data
        φ_P ← 1 − |p*|/N  # Update disagreement rate
    end
    Append φ_P to Φ_P
until K iterations elapse;
while n > 0 and iterations ≤ ℵ do  # Train an ensemble of CDCs on Q
    g ← ConstrainedDisagreement(L, P_train, P_val, Q, f)
    Q ← {x | x ∈ Q and f(x) = g(x)}
    φ_Q ← 1 − |Q|/N
end
return φ_Q > [(1 − α) quantile of Φ_P]
```

4 Empirical Evaluation

Our experiments are carried out on natural distribution shifts across multiple domains, modalities, and model types. We use the *CIFAR-10.1* dataset (Recht et al., 2019) where shift comes from subtle changes in the dataset creation processes, the *Camelyon17 dataset* (Veeling et al., 2018) for metastases detection in histopathological slides from multiple hospitals, as well as the *UCI heart disease* dataset (Janosi et al., 1988) which contains tabular features collected across international health systems and indicators of heart disease. We present unseen source an target domain performance of base models trained on source data in Appendix Table 2 which shows significant performance drops as an indicator for the hamrfulness of these shifts. See Appendix F for more details on datasets.

Learning Constrained Disagreement. We begin by training ensembles of 10 CDCs using the *disagreement cross entropy* (DCE) with CIFAR-10 as $\mathbf{P}$ and CIFAR-10.1 as $\mathbf{Q}$ for 100 random runs

at a sample size of 50 (see Appendix E for training details). The results in Figure 2 empirically validates minimizng the DCE as a CDC learning objective. The first observation is that when an unseen set is drawn from a shifted distribution $\mathcal{Q}$, the empirical disagreement rate $\phi_{\mathbf{Q}}$ grows significantly larger than the baseline disagreement rate $\phi_{\mathbf{P}}$. Next, we see that CDCs preserve accuracy on data from the training distribution. Finally, as the ensemble size increases (and disagreed upon points are removed) we see that the accuracy of the classifier increases. This indicates that the points that are disagreed upon early on in the algorithm are those that would have been misclassified.

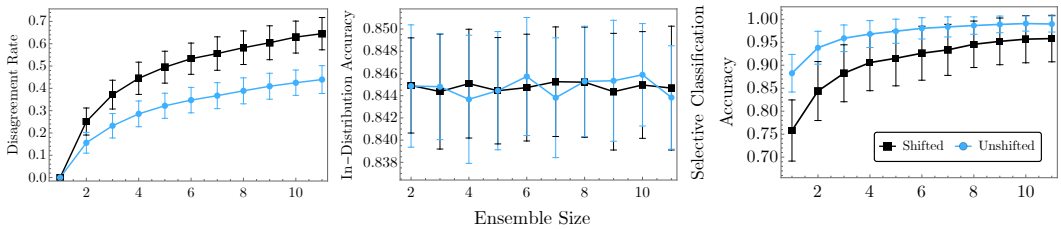

Figure 2: **Ensemble Size vs Properties of Constrained disagreement classifiers on CIFAR-10/10.1:** (Left) We see that for all ensemble sizes, there is lower disagreement on unshifted data (CIFAR-10) compared to disagreement on shifted data (CIFAR-10.1). (Center) Constrained disagreement does not compromise in-distribution performance. (Right) As the ensemble grows the selective classification accuracy computed on the set of test examples that all models agree on increases both on in-distribution and out-of-distribution data. Confidence intervals are computed as $\pm$ one standard deviation across experiments.

Shift Detection Setup. We evaluate the Detectron in a standard two-sample testing scenario similar to prior work (Zhao et al., 2022). Given two datasets $\mathbf{P} = \{x_1, \ldots, x_N\}$ (drawn from $\mathcal{P}$) and $\mathbf{Q} = \{\hat{x}_1, \ldots, \hat{x}_n\}$ (drawn from $\mathcal{Q}$) and classifier f, we seek to rule out the null hypothesis ($\mathcal{P} = \mathcal{Q}$) at the 5% significance level. To guarantee fixed significance we employ a permutation test by first sampling from the distribution of p-values derived by the Detectron where the null hypothesis $\mathcal{P} = \mathcal{Q}$ holds (i.e., $\mathbf{Q}$ is drawn $\mathcal{P}$). We then compute a threshold over the observed set of p-values that sets the false positive rate to 5%. This step can be performed in advance of deployment as it only requires access to $\mathbf{P}$. To mimic deployment settings where we wish to identify covariate shift quickly, we assume access to far more samples from $\mathcal{P}$ compared to $\mathcal{Q}$. For each dataset, we begin by training a base classifier on the unshifted dataset. We evaluate the detection of covariate shift on 100 randomly selected test sets of $n = 10$, 20 and 50 samples from $\mathcal{Q}$. In all cases we train a maximum ensemble size of 5 (parameter $\aleph$ in algorithm 1). To prevent CDCs from overfitting in the case of small test set sizes, we perform early stopping if in-distribution validation performance drops by over 5% from the measured performance of the base classifier. Hyperparameters and training details for all models can be found in Appendix E.

Evaluation. We report the *True Positive Rate at 5% Significance Level (TPR@5)* aggregated over 100 randomly selected sets $\mathbf{Q}$. This signifies how often our method correctly identifies covariate shift ($\mathcal{P} \neq \mathcal{Q}$) while only incorrectly identifying shift 5% of the time. This is also referred to as the statistical power of a test where the significance level or probability of making a type I error is 5%.

Baselines. We compare the Detectron against several methods for OOD detection, uncertainty estimation and covariate shift detection. *Deep Ensembles* Ovadia et al. (2019) using both (1) *disagreement* and (2) *entropy* scoring methods as a direct ablation to the CDC approach (3) *Black Box Shift Detection (BBSD)* (Lipton et al., 2018). (4) *Relative Mahalanobis Distance (RMD)* (Ren et al., 2021). (5) *Classifier Two Sample Test (CTST)* (Lopez-Paz & Oquab, 2017). (6) *Deep Kernel MMD (MMD-D)* (Liu et al., 2020). (7) *H Divergence (H-Div)* (Zhao et al., 2022). For more information on baselines see Appendix subsection E.3.

Shift Detection Experiments. We begin with an analysis of the performance of the Detectron on the UCI Heart Disease dataset. Using a sample size ranging from 10 to 100 we compute the TPR@5 averaged over 100 trials and plot the results in Figure 3. As the Detectron is model agnostic we use gradient boosted trees (XGBoost (Chen & Guestrin, 2016)) for Detectron and CTST methods while the remaining baselines use a 2 layer MLP that achieves similar in-distribution performance. Further results for sample sizes of 10, 20 and 50 on all datasets are shown in Table 1. We report the mean and standard error of TPR@5 computed on 100 samples for each trial.

Discussion. *Overall Performance*. We observe in the bottom rows of Table 1 that Detectron methods outperform all baselines across all three tasks. This confirms our intuition that designing distribution

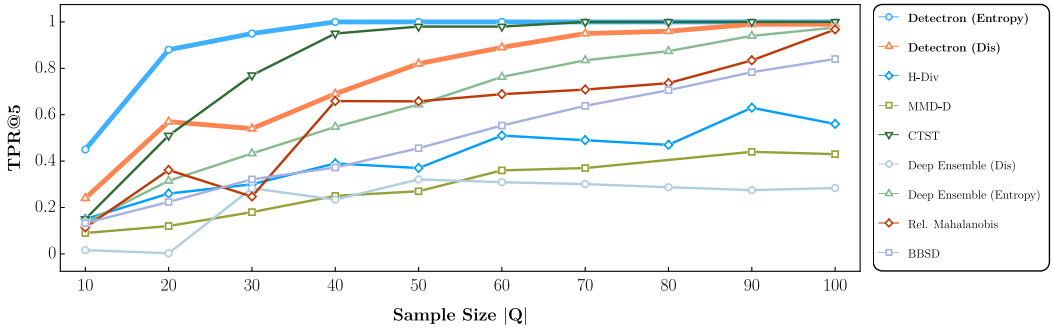

Figure 3: **True positive rate at the 5% significance level** for the Detectron and baseline methods for detection of covariate shift on the UCI heart disease dataset. The Detectron (Entropy) is shown to uniformly outperform baselines. Confidence intervals are excluded for visual clarity but are found in Table 1.

Table 1: **Results (true positive rate at the 5% significance level) for detection of harmful covariate shift** on CIFAR-10.1, Camelyon 17 and UCI Heart Disease benchmarks. The **best** result for each column is bolded, results that are within 2% of the best are underlined and the *best baseline* method is italicized.

	CIFAR 10.1			**Camelyon 17**			**UCI Heart Disease**		
$\lvert\mathbf{Q}\rvert$	10	20	50	10	20	50	10	20	50
BBSD	$.07 \pm .03$	$.05 \pm .02$	$.12 \pm .03$	$.16 \pm .04$	$.38 \pm .05$	$.87 \pm .03$	$.13 \pm .03$	$.22 \pm .04$	$.46 \pm .05$
Rel. Mahalanobis	$.05 \pm .02$	$.03 \pm .03$	$.04 \pm .02$	$.16 \pm .04$	$.40 \pm .05$	$\mathit{.89 \pm .03}$	$.11 \pm .03$	$.36 \pm .05$	$.66 \pm .05$
Deep Ensemble (Dis)	$.23 \pm .04$	$.40 \pm .05$	$\mathit{.74 \pm .04}$	$.10 \pm .03$	$.11 \pm .03$	$.13 \pm .03$	$.02 \pm .01$	$.00 \pm .00$	$.32 \pm .05$
Deep Ensemble (Entropy)	$\mathit{.33 \pm .05}$	$\mathit{.52 \pm .05}$	$.68 \pm .05$	$.14 \pm .03$	$.26 \pm .04$	$.82 \pm .04$	$.13 \pm .03$	$.32 \pm .05$	$.64 \pm .05$
CTST	$.03 \pm .02$	$.04 \pm .02$	$.04 \pm .02$	$.11 \pm .03$	$.59 \pm .05$	$.59 \pm .05$	$\mathit{.15 \pm .04}$	$\mathit{.51 \pm .05}$	$\underline{\mathit{.98 \pm .01}}$
MMD-D	$.24 \pm .04$	$.10 \pm .03$	$.05 \pm .02$	$\mathit{.42 \pm .05}$	$\mathit{.62 \pm .05}$	$.69 \pm .05$	$.09 \pm .03$	$.12 \pm .03$	$.27 \pm .04$
H-Div	$.02 \pm .01$	$.05 \pm .02$	$.04 \pm .02$	$.03 \pm .02$	$.07 \pm .03$	$.23 \pm .04$	$\mathit{.15 \pm .04}$	$.26 \pm .04$	$.37 \pm .05$
Detectron (Dis)	$\mathbf{.37 \pm .05}$	$\underline{.54 \pm .05}$	$.83 \pm .04$	$\mathbf{.97 \pm .02}$	$\mathbf{1.0 \pm .00}$	$.96 \pm .02$	$.24 \pm 0.04$	$.57 \pm 0.05$	$.82 \pm 0.04$
Detectron (Entropy)	$\underline{.35 \pm .05}$	$\mathbf{.56 \pm .05}$	$\mathbf{.92 \pm .03}$	$\mathbf{.97 \pm .02}$	$\mathbf{1.0 \pm .00}$	$\mathbf{1.0 \pm .00}$	$\mathbf{.45 \pm .05}$	$\mathbf{.88 \pm 0.03}$	$\mathbf{1.0 \pm .00}$

tests based specifically on available data and the outputs of learning algorithms is a promising avenue for improving existing methods in the high dimensional/low data regime.

Sample Efficiency. For more significant shifts (Camelyon and UCI), we see in Table 1 the most significant improvements over baselines in the lowest sample regime (10 data points). The fine-grained result in Figure 3 shows that CTST catches up to Detectron at 40 samples while deep ensemble, BBSD, and Mahalanobis catch up at 100.

Disagreement vs Entropy. For the experiments on imaging datasets with deep neural networks Detectron (Disagreement) often performs nearly as well as Detectron (Entropy), while Detectron (Entropy) is strictly superior for the UCI dataset. While we recommend entropy as the method to maximize test power, disagreement is a more interpretable statistic as it is correlates well with the portion of misclassified samples (see (right) Figure 2).

Comparison to baselines. Amongst the baselines, there is no clear best method. Although on average, ensemble entropy is superior on CIFAR, MMD-D on Camelyon, and CTST on UCI. Our method may be thought of as a combination of ensembles, CTST, and H-Divergence. As ensembles, we leverage the variation in outputs between a set of classifiers; as CTST, we learn in a domain adversarial setting; and as H-Divergence, we compute a test statistic based on data that a model was trained on. Lastly, while MMD-D and H-Divergence were shown to be the previous state-of-the-art, their performance was validated only on larger sample sizes (≥ 200).

On Tabular Data. The Detectron shows promise for deployment on tabular datasets (bottom right of Table 1 and Figure 3), where (1) the computational cost of training models is low, (2) the model

agnostic nature of the Detectron is beneficial as random forests often outperform neural networks in tabular data (Borisov et al., 2021), and (3) based on our discussions with medical professionals, the ability to detect covariate shift from small test sizes is of particular interest in the healthcare domain where population shift is a constant problem burden for maintaining the reliability of deployed models.

On computational cost: Our method is more computationally expensive than some existing methods for detecting shifts such as BBSD and Mahalanobis Scores, but is similar complexity to other approaches such as Ensembles, MMD-D and H-Divergence which may require training multiple deep models. However, as the Detectron leverages a pretrained model already in deployment, we find in practice that only a small number of training rounds are required to create each CDC. For instance, on CIFAR 10/10.1 a CDC ensemble of 5 models using a ResNet 18 architecture can train in under $\approx$ 2m using an unoptimized PyTorch implementation on 1 GPU. Furthermore, looking at the runtime behavior in Figure 4 we see that while allowing for more computation time increases the fidelity of the Detectron, only a small number of training batches may be required to achieve a desirable level of statistical significance.

In scenarios where the deployed classifier is deemed high-risk (e.g. healthcare, justice system, education) where each data point is a decision that affects a human being, we believe the additional computational expense is justified for an accurate and sensitive assessment of whether the classifier needs updating. Having established the utility, accelerating the Detectron as well as building a deeper understanding of the runtime performance tradeoffs, is fertile ground for future work.

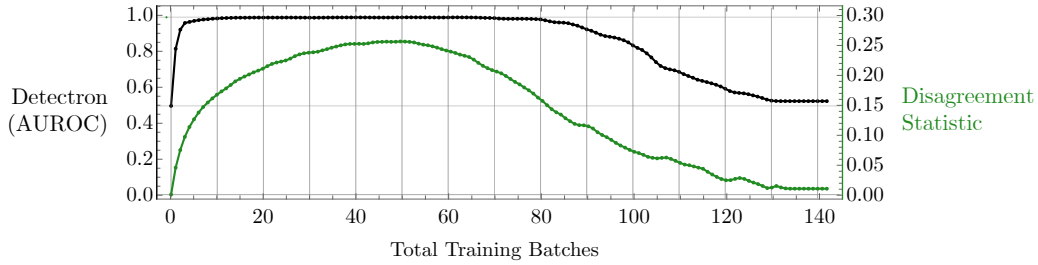

Figure 4: **Runtime Characteristics:** We train 100 random runs of CDCs on 100 samples from CIFAR 10 and 10.1 and compute the *disagreement satistic* as the difference $\psi := \mathbb{E}[\phi_{\mathbf{Q}} - \phi_{\mathbf{P}}]$. While we see that while ψ peaks near 50 training batches, only 10 batches are required for the Detectron disagreement test to reach an area under the TPR vs FPR curve (AUROC) of nearly 1 (i.e., perfect discrimination). Training CDCs for too long eventually lowers ψ as $\mathbb{E}[\phi_{\mathbf{Q}}] \approx \mathbb{E}[\phi_{\mathbf{P}}] \approx 1$ meaning CDCs eventually overfit to disagreeing on all of their data.

5 Conclusion, Limitations and Future Work

Our work presents a practical application of PQ learning towards creating a method capable of detecting covariate shifts given a pre-existing classifier. On both neural networks and random forests, we showcase the efficacy of our method in being sensitive enough to detect covariate shift using a small number of unlabelled examples across several real-world datasets. We remark on several characteristics of our algorithm that represent potential directions for future work:

Beyond Classification: Our work here focuses on the case of classification (since a large number of pre-existing benchmarks in the literature focus on the same). However, we believe there is a viable extension of our work to regression models where constrained predictors are explicitly learned to maximize test error according to the existing metric, such as mean squared error. We leave this exploration for future work.

Beyond Covariate Shifts: While covariate shift is the only type of shift that can be discovered from unlabeled data without additional assumptions, we acknowledge that other types of shift, such as label and concept shift, are prevalent in the real world. Building learning-based methods to identify these types of shifts is another direction for future work.

Finally, we wish to highlight that while auditing systems such as the Detectron show promise to ease concerns when using learning systems in high-risk domains, practitioners interfacing with these systems should not place blind trust in their outputs.

ETHICS STATEMENT

The speed of the adoption of ML into risk scenarios raises a critical need for methods that ensure trustworthiness and reliability. However, over-reliance on such methods brings about critical ethical considerations. As we have seen, Detectron is highly sensitive to picking out discriminative features of data distributions, and, as such, its usage may prevent practitioners from deploying models in new environments. As a result, the individuals in those environments may become subject to unfair treatment. For instance, if Detectron determines a model trained on hospital A safe to deploy in hospitals B and C, but not in D, the individuals in population D may experience a lower level of care. In a real example, Detectron, when tested on a model trained on a subset of light-skinned celebrities (CelebA dataset–Liu et al. (2015)), quickly raises the alarm when given images of those that are not light-skinned. While Detectron can help mitigate potential disasters encountered by deploying models in hazardous domains, it should not be an excuse for practitioners to avoid collecting richer and more diverse datasets as a primary strategy to ensure model reliability.

REFERENCES

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In B. Schölkopf, J. Platt, and T. Hoffman (eds.), *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. URL https://proceedings.neurips.cc/paper/2006/file/b1b0432ceafb0ce714426e9114852ac7-Paper.pdf.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(75):2137–2155, 2009. URL http://jmlr.org/papers/v10/bickel09a.html.

Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *CoRR*, abs/2110.01889, 2021. URL https://arxiv.org/abs/2110.01889.

Dipankar Chaki, Arkadeep Das, and Moinul Islam Zaber. A comparison of three discrete methods for classification of heart disease data. *Bangladesh Journal of Scientific and Industrial Research*, 50:293–296, 2015.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL http://doi.acm.org/10.1145/2939672.2939785.

Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G, 2018. (Accessed on 05/11/2022).

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Shafi Goldwasser, Adam Tauman Kalai, Yael Kalai, and Omar Montasser. Beyond Perturbations: Learning Guarantees with Arbitrary Adversarial Test Examples. In H Larochelle, M Ranzato, R Hadsell, M F Balcan, and H Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 15859–15870. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/b6c8cf4c587f2ead0c08955ee6e2502b-Paper.pdf.

Anand R. Habib, Anthony L. Lin, and Richard W. Grant. The Epic Sepsis Model Falls Short—The Importance of External Validation. *JAMA Internal Medicine*, 181(8):1040–1041, 08 2021.

Karen Hao. Ai is sending people to jail—and getting it wrong. https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/, 2019. (Accessed on 05/11/2022).

David Haussler. Probably approximately correct learning. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, pp. 1101–1108. AAAI Press, 1990.

J. L. Hodges. The significance probability of the smirnov two-sample test. *Arkiv för Matematik*, 3(5):469–486, 1958. doi: 10.1007/BF02589501. URL https://doi.org/10.1007/BF02589501.

Wolfram Research, Inc. Mathematica, Version 13.0.0. URL https://www.wolfram.com/mathematica. Champaign, IL, 2021.

Andras Janosi, William Steinbrunn, Matthias Pfisterer, and Robert Detrano. Heart Disease. UCI Machine Learning Repository, 1988.

Daniel Kifer, Shai Ben-David, and Johannes Gehrke. Detecting change in data streams. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*, VLDB '04, pp. 180–191. VLDB Endowment, 2004. ISBN 0120884690.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. The cifar-10 dataset. *online: http://www. cs. toronto. edu/kriz/cifar. html*, 55(5), 2014.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6405–6416, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4): 541–551, 12 1989. ISSN 0899-7667. doi: 10.1162/neco.1989.1.4.541. URL https://doi.org/10.1162/neco.1989.1.4.541.

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

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1VGkIxRZ.

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

Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Danica J Sutherland. Learning deep kernels for non-parametric two-sample tests. In *International conference on machine learning*, pp. 6316–6326. PMLR, 2020.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJkXfE5xx.

Sébastien Marcel and Yann Rodriguez. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pp. 1485–1488, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605589336. doi: 10.1145/1873951.1874254. URL https://doi.org/10.1145/1873951.1874254.

Warren Morningstar, Cusuh Ham, Andrew Gallagher, Balaji Lakshminarayanan, Alex Alemi, and Joshua Dillon. Density of states estimation for out of distribution detection. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3232–3240. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/morningstar21a.html.

Tristan Needham. *Complex Integration: Cauchy's Theorem*, pp. 377–417. Oxford University Press, 2000.

Erkin Otles, Jeeheh Oh, Benjamin Li, Michelle Bochinski, Hyeon Joo, Justin Ortwine, Erica Shenoy, Laraine Washer, Vincent B Young, Krishna Rao, et al. Mind the performance gap: examining dataset shift during prospective validation. In *Machine Learning for Healthcare Conference*, pp. 506–534. PMLR, 2021.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Stephan Rabanser, Stephan Günnemann, and Zachary C. Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 1394–1406, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/846c260d715e5b854ffad5f70a516c88-Abstract.html.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do cifar-10 classifiers generalize to cifar-10?, 2018. URL https://arxiv.org/abs/1806.00451.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/recht19a.html.

Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *Advances in neural information processing systems*, 32, 2019.

Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *ICML workshop on Uncertainty and Robustness in Deep Learning*, 2021. URL http://www.gatsby.ucl.ac.uk/~balaji/udl2021/accepted-papers/UDL2021-paper-007.pdf.

Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with Gram matrices. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8491–8501. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/sastry20a.html.

Lauren Smiley. 'i'm the operator': The aftermath of a self-driving tragedy. https://www.wired.com/story/uber-self-driving-car-fatal-crash/, 2022. (Accessed on 05/11/2022).

Nikolai G. Ushakov. *Continuity theorems and inversion formulas*. De Gruyter, 2011.

Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995. ISBN 0-387-94559-8.

Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant CNNs for digital pathology. June 2018.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

Lily Zhang, Mark Goldstein, and Rajesh Ranganath. Understanding failures in out-of-distribution detection with deep generative models. In *International Conference on Machine Learning*, pp. 12427–12436. PMLR, 2021.

Shengjia Zhao, Abhishek Sinha, Yutong He, Aidan Perreault, Jiaming Song, and Stefano Ermon. Comparing distributions by measuring differences that affect decision making. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=KB5onONJIAU.