Change is Hard: A Closer Look at Subpopulation Shift

Yuzhe Yang [*1] **Haoran Zhang** [*1] **Dina Katabi** [1] **Marzyeh Ghassemi** [1]

Abstract

Machine learning models often perform poorly on *subgroups* that are underrepresented in the training data. Yet, little is understood on the variation in mechanisms that cause subpopulation shifts, and how algorithms generalize across such diverse shifts at scale. In this work, we provide a fine-grained analysis of subpopulation shift. We first propose a unified framework that dissects and explains common shifts in subgroups. We then establish a comprehensive benchmark of 20 state-of-the-art algorithms evaluated on 12 real-world datasets in vision, language, and healthcare domains. With results obtained from training over 10,000 models, we reveal intriguing observations for future progress in this space. First, existing algorithms only improve subgroup robustness over certain types of shifts but not others. Moreover, while current algorithms rely on group-annotated validation data for model selection, we find that a simple selection criterion based on worst-class accuracy is surprisingly effective even without any group information. Finally, unlike existing works that solely aim to improve worst-group accuracy (WGA), we demonstrate the fundamental trade-off between WGA and other important metrics, highlighting the need to carefully choose testing metrics. Code and data are available at: `https://github.com/YyzHarry/SubpopBench`.

1. Introduction

Machine learning models frequently exhibit drops in performance under the presence of distribution shifts (Quinonero-Candela et al., 2008). Constructing machine learning models that are robust to these shifts is critical to the safe deployment of such models in the real-world (Amodei et al., 2016). One ubiquitous type of distribution shift is *subpopulation shift*, which is characterized by changes in the proportion of some subpopulations between training and deployment (Koh et al., 2021). In such settings, models may have high overall performance but still perform poorly in rare subgroups (Hashimoto et al., 2018; Zhang et al., 2020).

A well-studied type of subpopulation shift occurs when data contains *spurious correlations* (Geirhos et al., 2020) – non-causal relationships between the input and the label which may shift in deployment (Simon, 1954). For example, image classifiers frequently make use of non-robust features such as image backgrounds (Xiao et al., 2016), textures (Geirhos et al., 2018), and erroneous markings (DeGrave et al., 2021). However, there has been little work in defining subpopulation shift in a holistic way, understanding *when* these shifts happen, and *how* state-of-the-art (SOTA) algorithms generalize under diverse and realistic shifts. Subpopulation shift can encompass a much wider array of underlying mechanisms. First, different attributes in data often exhibit skewed distributions, inevitably causing *attribute imbalance* (Martinez et al., 2021). Moreover, certain labels can have significantly fewer observations, where such long-tailed label distribution induces severe *class imbalance* (Liu et al., 2019b). Finally, certain attributes may have no training data at all, which motivates the need for *attribute generalization* to unseen subpopulations (Santurkar et al., 2020).

In this work, we systematically investigate subpopulation shift in realistic evaluation settings. We first formalize a generic framework of subpopulation shift, which decomposes *attribute* and *class* to enable fine-grained analyses. We demonstrate that this modeling covers and explains the aforementioned common subgroup shifts, which are basic units of building more complex shifts that arise in real data. Using this framework, we can quantify the type and degree of different shift components in each given dataset.

We establish a realistic and comprehensive benchmark of subpopulation shift, consisting of **20** SOTA algorithms that span different learning strategies and **12** real-world datasets in vision, language, and healthcare domains. While existing analysis on subpopulation shift either focus on a single shift type, or have limited severity, our benchmark provides a much larger set of datasets that cover different types of realistic subgroup shifts. Our experimental framework can be easily extended to include new methods, shifts, and datasets.

Our work also evaluates current methods across different settings including attribute availability in training and/or

[*]Equal contribution [1]MIT CSAIL. Correspondence to: Yuzhe Yang <yuzhe@mit.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

validation set, model selection strategies, and a wide range of metrics for understanding subpopulation shift in-depth. With the established framework and over 10K trained models, we reveal intriguing observations for future research.

Concretely, we make the following contributions:

- We formalize a unified framework for subpopulation shift which defines basic types of shift, explains when and why shifts happen, and quantifies their degrees.
- We set up a comprehensive and realistic benchmark for systematic subpopulation shift evaluation, with 20 SOTA methods and 12 diverse datasets across various domains.
- Based on over 10K trained models, we verify that current algorithms only advance subgroup robustness over certain types of shift identified by our framework, but not others.
- We confirm that while successful algorithms rely on the access to group information for model selection, a simple criterion based on worst-class accuracy is surprisingly effective even without group-annotated validation data.
- We establish the fundamental tradeoff between worst-group accuracy (WGA) and important metrics such as worst-case precision, highlighting the need to rethink evaluation metrics in subpopulation shift beyond WGA.

2. Related Work

Subpopulation Shift. Machine learning models frequently experience performance degradation under *subpopulation shift*, where the proportion of some subpopulations differ between the training and test (Cai et al., 2021; Koh et al., 2021). Depending on the definition of such subpopulations, this could lead to vastly different problem settings. Prior works largely focus on the case of shortcut learning (Geirhos et al., 2020), where subpopulations are defined as the product of attributes and labels. In such settings, models trained to minimize overall loss tend to learn spurious correlations, resulting in poor performance in the minority subpopulation (DeGrave et al., 2021; Joshi et al., 2022). There have been a large set of methods developed to address this scenario, both when the attribute is known (Gowda et al., 2021; Izmailov et al., 2022; Menon et al., 2020; Nam et al., 2022; Sagawa et al., 2019; Yao et al., 2022), and unknown (Creager et al., 2021; Han et al., 2022; Idrissi et al., 2022; Liu et al., 2021).

However, subpopulations may also be defined using only the label. This setting corresponds to class-imbalanced learning, which has also been well studied with extensive proposed methods (Cao et al., 2019; Cui et al., 2019; Li et al., 2021; Yang & Xu, 2020; Yang et al., 2021; 2022).

Finally, when subpopulations are defined based on a particular attribute (e.g., demographic group) (Pfohl et al., 2022; Zong et al., 2022), the objective of maximizing performance for the worst-case group then becomes identical to minimax fairness (Lahoti et al., 2020; Martinez et al., 2020).

In this work, we present a unified framework of subpopulation shift across these aforementioned scenarios.

Distribution Shift Benchmarks. There have been few prior works which benchmark the performance of subpopulation shift methods. Koh et al. (2021) proposed the WILDS benchmark for domain generalization and subpopulation shift, though they only evaluated four methods over five datasets. Zhang et al. (2022) and Gulrajani & Lopez-Paz (2020) proposed the NICO++ and DomainBed benchmarks respectively for domain generalization, and we adapt elements of their benchmark into our subpopulation shift evaluation. Santurkar et al. (2020) proposed the BREEDS benchmark, which consists of multiple datasets constructed from ImageNet (Deng et al., 2009) using the WordNet hierarchy (Miller, 1995), aiming to evaluate generalization across unseen attributes. Finally, Wiles et al. (2021) conducted a similar analysis in the general distribution shift setting on four synthetic and two real-world datasets.

Our work differs from these prior works by evaluating a much larger set of algorithms that span different categories on many more real-world datasets. We further define, dissect and quantify the type and degree of shift components in each dataset, and relate it to the performance of each method. In addition, we analyze important yet overlooked factors such as model selection criteria and metrics to evaluate against, and reveal intriguing properties in subpopulation shift.

3. Unified Framework of Subpopulation Shift

Problem Setup. In the general subpopulation shift setting, given input $\mathbf{x} \in \mathcal{X}$ and label $y \in \mathcal{Y}$, the goal is to learn $f : \mathcal{X} \to \mathcal{Y}$. In addition, there exist attributes $a_1, ..., a_i, ..., a_m$, $a_i \in \mathcal{A}_i$, which may or may not be available when learning f. Then, discrete subpopulations can be defined based on the attribute and label, by some function $h : \mathcal{A} \times \mathcal{Y} \to \mathcal{G}$.

Let $\ell(y, f(\mathbf{x})) \to \mathbb{R}$ be a loss function. Consider the source distribution where $(\mathbf{x}, y)$ are drawn as a mixture of group-wise distributions: $P_{src} = \sum_{g \in \mathcal{G}} \alpha_g P_g$, where $\alpha \in \Delta_{|\mathcal{G}|}$. Further, consider some target distribution which is not observed: $P_{tar} = \sum_{g \in \mathcal{G}} \beta_g P_g$, where $\beta \in \Delta_{|\mathcal{G}|}$. The objective of subpopulation shift is to find (Sagawa et al., 2020):

$$f^* = \arg\min_f \sup_{\beta \in \Delta_{|\mathcal{G}|}} \mathbb{E}_{(\mathbf{x},y)\sim P_{tar}}[\ell(y, f(\mathbf{x}))].$$

This objective is equivalent to minimizing risk for the worst-case group (Sagawa et al., 2020), i.e.,

$$f^* = \arg\min_f \max_{g \in \mathcal{G}} \mathbb{E}_{(\mathbf{x},y)\sim P_g}[\ell(y, f(\mathbf{x}))].$$

3.1. A Generic Framework for Subpopulation Shift

As motivated earlier, both attribute a and label y can have specific skewed distributions, resulting in distinct types of

Table 1. Formulation summary of basic types of subpopulation shift under our framework.

Subpopulation Shift Type	Attribute Bias	Class Bias	Impact on Classification Model
Spurious Correlations (SC)	$p_{\text{train}}(a\|y, \mathbf{x}_{\text{core}}) \gg p_{\text{train}}(a\|\mathbf{x}_{\text{core}})$ $p_{\text{test}}(a\|y, \mathbf{x}_{\text{core}}) = p_{\text{test}}(a\|\mathbf{x}_{\text{core}})$	–	$\frac{\mathbb{P}(a\|y,\mathbf{x}_{\text{core}})}{\mathbb{P}(a\|\mathbf{x}_{\text{core}})} \gg 1 \Rightarrow \mathbb{P}(y\|\mathbf{x}) \uparrow$
Attribute Imbalance (AI)	$p_{\text{train}}(a\|y, \mathbf{x}_{\text{core}}) \gg p_{\text{train}}(a'\|y, \mathbf{x}_{\text{core}})$ $p_{\text{test}}(a\|y, \mathbf{x}_{\text{core}}) = p_{\text{test}}(a'\|y, \mathbf{x}_{\text{core}})$	–	$\frac{\mathbb{P}(a\|y,\mathbf{x}_{\text{core}})}{\mathbb{P}(a\|\mathbf{x}_{\text{core}})} \gg \frac{\mathbb{P}(a'\|y,\mathbf{x}_{\text{core}})}{\mathbb{P}(a'\|\mathbf{x}_{\text{core}})} \Rightarrow \mathbb{P}(y\|\mathbf{x}_{\text{core}}, a) \gg \mathbb{P}(y\|\mathbf{x}_{\text{core}}, a')$
Class Imbalance (CI)	–	$p_{\text{train}}(\mathbf{Y} = y) \gg p_{\text{train}}(\mathbf{Y} = y')$ $p_{\text{test}}(\mathbf{Y} = y) = p_{\text{test}}(\mathbf{Y} = y')$	$\mathbb{P}(y) \gg \mathbb{P}(y') \Rightarrow \mathbb{P}(y\|\mathbf{x}) \gg \mathbb{P}(y'\|\mathbf{x})$
Attribute Generalization (AG)	$p_{\text{train}}(a\|y, \mathbf{x}_{\text{core}}) = 0, \forall a \in \mathbb{A}^{\text{unseen}}$ $p_{\text{test}}(a\|y, \mathbf{x}_{\text{core}}) > 0, \forall a \in \mathbb{A}$	Unconstrained	Generalize to $\mathbb{A}^{\text{unseen}}$

subpopulation shift. To this end, we propose to decompose the effect of a and y given a multi-group dataset, and characterize general subpopulation shift into several ***basic shift*** components for fine-grained interpretation.

Specifically, we view each input $\mathbf{x}$ as being fully described or generated from a set of underlying core features $\mathbf{x}_{\text{core}}$ (representing the label) and a list of attributes $\mathbf{a}$ (Tang et al., 2022; Wang et al., 2021). Here, $\mathbf{x}_{\text{core}}$ denotes the underlying invariant components that are label-specific and support robust classification, whereas attributes $\mathbf{a}$ may have inconsistent distributions and are not label-specific. Such modeling helps us disentangle the attributes and examine how they affect the classification results $\mathbb{P}(y|\mathbf{x})$. Following Bayes' theorem, we can rewrite the classification model as:

$$\begin{aligned}\mathbb{P}(y|\mathbf{x}) &= \frac{\mathbb{P}(\mathbf{x}|y)}{\mathbb{P}(\mathbf{x})} \cdot \mathbb{P}(y) \\ &= \frac{\mathbb{P}(\mathbf{x}_{\text{core}}, \mathbf{a}|y)}{\mathbb{P}(\mathbf{x}_{\text{core}}, \mathbf{a})} \cdot \mathbb{P}(y) \\ &= \underbrace{\frac{\mathbb{P}(\mathbf{x}_{\text{core}}|y)}{\mathbb{P}(\mathbf{x}_{\text{core}})}}_{\text{PMI}} \cdot \underbrace{\frac{\mathbb{P}(\mathbf{a}|y, \mathbf{x}_{\text{core}})}{\mathbb{P}(\mathbf{a}|\mathbf{x}_{\text{core}})}}_{\text{attribute}} \cdot \underbrace{\mathbb{P}(y)}_{\text{class}}, \end{aligned} \tag{1}$$

where the first term in Eqn. (1) represents the pointwise mutual information (PMI) between $\mathbf{x}_{\text{core}}$ and y, the second term corresponds to the potential bias arising in the **attribute** distribution, and the third term explains the potential bias arising in the **class** (label) distribution. Given invariant $\mathbf{x}_{\text{core}}$ between training and testing distributions, we can ignore changes in first term (which is a robust indicator), and focus on how the second and third term, i.e., the *attribute* and *class*, influence the outcomes under subpopulation shift.

More formally, assuming the mutual independence and conditional independence across different attributes a_i (Wiles et al., 2021), we can further decompose the attribute term into a fine-grained version:

$$\frac{\mathbb{P}(\mathbf{a}|y, \mathbf{x}_{\text{core}})}{\mathbb{P}(\mathbf{a}|\mathbf{x}_{\text{core}})} \triangleq \prod_{a_i \in \mathbf{a}} \frac{\mathbb{P}(a_i|y, \mathbf{x}_{\text{core}})}{\mathbb{P}(a_i|\mathbf{x}_{\text{core}})}, \tag{2}$$

where each a_i corresponds to an attribute. Note that for benign attributes that are independent of y (i.e., $a_i \perp\!\!\!\perp y, \forall a_i \in \mathbf{a}_{\text{benign}}$), we have $\mathbb{P}(a_i|y, \mathbf{x}_{\text{core}}) = \mathbb{P}(a_i|\mathbf{x}_{\text{core}})$, indicating that the attribute term in Eqn. (2) is only driven by *biased* attributes that are label-dependent.

Using the formulation of "attribute-class" decomposition, we can intuitively explain *when* do common subpopulation shifts happen, and *how* they affect the classification results.

3.2. Characterizing Basic Types of Subpopulation Shift

We formally define and characterize four basic types of subpopulation shift using our framework: *spurious correlations*, *attribute imbalance*, *class imbalance*, and *attribute generalization* (see Table 1). In practice, we note that dataset often consists of multiple types of shift instead of one. The four cases constitute the *basic* shift units, and are important elements to explain complex subgroup shifts in real data.

Spurious Correlations (SC). Spurious correlations happen when certain a is spuriously correlated with y in training but not in test data. Under our framework, it implies that $p_{\text{train}}(a|y, \mathbf{x}_{\text{core}}) \gg p_{\text{train}}(a|\mathbf{x}_{\text{core}})$, which is not true of p_{test}. As a result, it introduces bias to the *attribute term*, which induces higher prediction confidence for certain label once given its spuriously correlated attribute (details in Table 1).

Attribute Imbalance (AI). Attributes often incur biased distributions in the wild. In our framework, it happens when certain attributes are sampled with a much smaller probability than others in p_{train}, but not in p_{test}. To disentangle the effect of labels, we assume no class bias under this basic shift. As such, it again affects the *attribute term* in Eqn. (1) where $p_{\text{train}}(a|y, \mathbf{x}_{\text{core}}) \gg p_{\text{train}}(a'|y, \mathbf{x}_{\text{core}})$, causing lower prediction confidence for underrepresented attributes.

Class Imbalance (CI). Similarly, class labels can exhibit imbalanced distributions, causing lower preference for minority labels. Within our framework, CI can be explained by biasing the *class term* in p_{train}, leading to higher prediction confidence for majority classes.

Attribute Generalization (AG). Certain attributes can be totally missing in p_{train}, but present in p_{test}, which motivates the need for attribute generalization. In our framework, this translates to $p_{\text{train}}(a|y, \mathbf{x}_{\text{core}}) = 0, a \in \mathbb{A}^{\text{unseen}}$, yet we have $p_{\text{test}}(a|y, \mathbf{x}_{\text{core}}) > 0$. AG requires learning robust $\mathbf{x}_{\text{core}}$ in order to generalize across unseen attributes, which is harder but more ubiquitous in real data (Santurkar et al., 2020).

Table 2. Overview of the datasets for evaluating subpopulation shift. Detailed statistics and example data are provided in Appendix B.

Dataset	Data type	# Attr.	# Classes	# Train set	# Val. set	# Test set	Max group	Min group	Shift type			
									SC	AI	CI	AG
Waterbirds	Image	2	2	4795	1199	5794	3498 (73.0%)	56 (1.2%)	✓	✓	✓	
CelebA	Image	2	2	162770	19867	19962	71629 (44.0%)	1387 (0.9%)	✓		✓	
MetaShift	Image	2	2	2276	349	874	789 (34.7%)	196 (8.6%)	✓			
ImageNetBG	Image	N/A	9	183006	7200	4050	N/A	N/A				✓
NICO++	Image	6	60	62657	8726	17483	811 (1.3%)	0 (0.0%)		✓	✓	✓
Living17	Image	N/A	17	39780	4420	1700	N/A	N/A				✓
MultiNLI	Text	2	3	206175	82462	123712	67376 (32.7%)	1521 (0.7%)		✓		
CivilComments	Text	8	2	148304	24278	71854	31282 (21.1%)	1003 (0.7%)		✓	✓	
MIMICNotes	Clinical text	2	2	16149	3229	6460	8359 (51.8%)	676 (4.2%)			✓	
MIMIC-CXR	Chest X-rays	6	2	303591	17859	35717	68575 (22.6%)	7846 (2.6%)		✓		
CheXpert	Chest X-rays	6	2	167093	22280	33419	51606 (30.9%)	506 (0.3%)		✓	✓	
CXRMultisite	Chest X-rays	2	2	338134	19891	39781	299089 (88.5%)	574 (0.2%)	✓	✓	✓	

4. Benchmarking Subpopulation Shift

Datasets. We explore subpopulation shift using **12** real-world datasets from a variety of modalities and tasks. First, for **vision** datasets, we use Waterbirds (Wah et al., 2011) and CelebA (Liu et al., 2015), which are commonly used in the spurious correlation literature (Liu et al., 2021). Similarly, we use the MetaShift cats *vs.* dogs dataset (Liang & Zou, 2022). We further convert the ImageNet backgrounds challenge (ImageNetBG) (Xiao et al., 2020), the NICO++ (Zhang et al., 2022) benchmark, and the Living17 dataset from the BREEDS benchmark (Santurkar et al., 2020) for subpopulation shift. Further, for **language** understanding datasets, we leverage CivilComments (Borkan et al., 2019) and MultiNLI (Williams et al., 2017), which are commonly used text datasets in subpopulation shift. Finally, we curate 4 datasets in the **medical** domain. We construct MIMIC-CXR (Johnson et al., 2019) and CheXpert (Irvin et al., 2019) to predict the presence of any pathology from a chest X-ray. We also construct MIMICNotes for mortality classification from clinical notes (Chen et al., 2019). Finally, we follow a recent work in evaluating subgroup shift and construct the CXRMultisite dataset (Puli et al., 2021). Table 2 reports the details of each dataset. We leave full information and descriptions for each of the datasets in Appendix B.1.

Algorithms. We evaluate **20** algorithms that span a broad range of learning strategies and categories, and relate their performance to different shifts defined in our framework. We believe this is the first work to comprehensively evaluate a large set of diverse algorithms in subpopulation shift. Concretely, these algorithms cover the following areas: (1) *vanilla:* **ERM** (Vapnik, 1999), (2) *subgroup robust methods:* **GroupDRO** (Sagawa et al., 2020), **CVaRDRO** (Duchi & Namkoong, 2018), **LfF** (Nam et al., 2020), **JTT** (Liu et al., 2021), **LISA** (Yao et al., 2022), **DFR** (Izmailov et al., 2022), (3) *data augmentation:* **Mixup** (Zhang et al., 2018), (4) *domain-invariant feature learning:* **IRM** (Arjovsky et al., 2019), **CORAL** (Sun & Saenko, 2016), **MMD** (Li et al., 2018), (5) *imbalanced learning:* **ReSample** (Japkowicz, 2000), **ReWeight** (Japkowicz, 2000), **Focal** (Lin et al., 2017), **CBLoss** (Cui et al., 2019), **LDAM** (Cao et al., 2019), **BSoftmax** (Ren et al., 2020), **CRT** (Kang et al., 2020), **ReWeightCRT** (Kang et al., 2020). Our framework can be easily extended to include new algorithms. We provide detailed descriptions for each algorithm in Appendix B.2.

Evaluation Metrics. Existing works on subpopulation shift mainly report *worst-group accuracy* (WGA) as the gold-standard. While WGA faithfully assesses worst-group performance, other important metrics (e.g., worst-case precision, calibration error, etc.) are also essential especially when involving subpopulation shift. Therefore, in our benchmark we include a variety of metrics aiming for a thorough evaluation from different aspects. In particular, besides **Avg Accuracy** and **Worst Accuracy**, we further include **Avg Precision**, **Worst Precision**, **Avg F1-score**, **Worst F1-score**, (Class-)**Balanced Accuracy**, **Adjusted Accuracy** (accuracy on a *group*-balanced dataset), and expected calibration error (**ECE**) (Guo et al., 2017). Detailed summaries of all metrics are in Appendix B.3.

Attribute Availability. Whether attribute is known in both (1) *training set* and (2) *validation set* has long been a vital factor for almost all subgroup algorithms (Izmailov et al., 2022). Specifically, classic methods (e.g., GroupDRO) assume access to attributes during training to define meaningful groups. Recently, a number of methods (e.g., JTT, LfF, DFR) try to improve worst-group accuracy without knowing the training attributes. Nevertheless, current approaches still require access to group-annotated validation set for model selection and hyperparameter tuning (Idrissi et al., 2022).

We systematically investigate this phenomenon by considering three settings in our benchmark: (1) *attributes are known in both training & validation*, (2) *attributes are unknown*

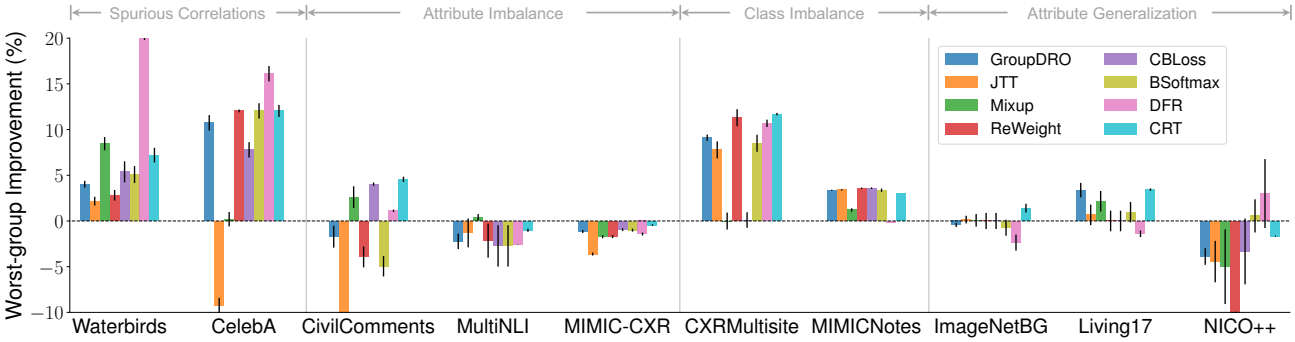


Figure 1. Worst-group improvements over ERM across different datasets when attributes are *unknown* in both training and validation set. SOTA algorithms only enhance subgroup robustness on certain types of shift (i.e., **SC** and **CI**). Complete results are in Appendix D.2.

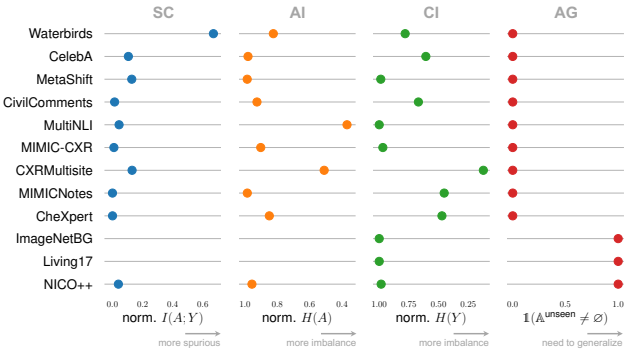


Figure 2. Quantification of the degree of different shifts over all datasets. Additional metrics are provided in Appendix D.1.

in training, but known in validation, and (3) *attributes are unknown in both training & validation*. Note that when training attributes are unknown, methods that operate over *subgroups* degenerate to operate over *classes*. Without further specification, we report results under the third setting, which is the hardest but the most realistic one. We include full results across all settings in Appendix E.

Model Selection. As mentioned earlier, model selection becomes essential when attributes are completely unknown. Significant drop (over 20%) in worst-group test accuracy has been observed if using the highest *average* validation accuracy as the model selection criterion without any group annotations (Idrissi et al., 2022). To this end, we provide a rigorous analysis on different model selection strategies, especially when attributes are fully unknown. Further details are provided in Appendix B.4.

Implementation. For a fair evaluation, following (Gulrajani & Lopez-Paz, 2021), for each algorithm we conduct a random search of 16 trials over a joint distribution of all hyperparameters (details are provided in Appendix C). We then use the validation set to select the best hyperparameters for each algorithm, fix them and rerun the experiments under three different random seeds to report the final average results with standard deviation. Such process ensures the comparison is best-versus-best, and the hyperparameters are optimized for all algorithms.

5. A Fine-Grained Analysis

5.1. Quantifying Subpopulation Shift

In order to quantify the degree of each shift for each dataset relative to others, we use several simple metrics. For *spurious correlations*, we use the normalized mutual information between A and Y, where norm $I(A;Y) = 1$ means that the two are perfectly correlated: norm $I(A;Y) = \frac{2I(A;Y)}{H(Y)+H(A)}$.

For *attribute* and *class imbalance*, we use the normalized entropy, where norm $H(Y) = 1$ indicates that the distribution is uniform (i.e., no imbalance): norm $H(Y) = \frac{H(Y)}{\log |\text{supp}(Y)|}$.

For *attribute generalization*, we simply examine whether there exist any subpopulations in the test set which do not appear during training via an indicator function (see Fig. 2). We provide several additional metrics in Appendix D.1.

We find that different datasets exhibit very different types of shift, and the degrees also greatly vary (Fig. 2). To further study how algorithms perform across various types of shift, we categorize each dataset into its most dominant shift type.

5.2. Performance across Different Types of Shift

As described earlier, we run experiments for all algorithms, datasets, and attribute availability settings. We use *worst-group accuracy* as the model selection criterion, and provide analysis for other metrics in Appendix D.3. When attributes are unknown in the validation set, this criterion degenerates to *worst-class accuracy*. Interestingly, we discover that this simple method is surprisingly effective (related results in Sec. 5.4). In total, we trained over 10,000 models.

We study model performance over different shifts. Specifically, we report results when attributes are *unknown* in both training and validation. Results for other settings are in Appendix D.2. We present main results in Fig. 1 and Table 3, where we make intriguing observations as follows.

SOTA algorithms only improve subgroup robustness on certain types of shift, but not others. As Fig. 1 illustrates, for *spurious correlations* and *class imbalance*, existing algo-

Table 3. Results on all tested subpopulation benchmarks, when attributes are *unknown* in both training and validation set. Full results for each dataset and other settings are in Appendix E. Methods that re-train classifier using a two-stage strategy are marked in gray.

Algorithm	Waterbirds	CelebA	CivilComments	MultiNLI	MetaShift	ImageNetBG	NICO++	MIMIC-CXR	MIMICNotes	CXRMultisite	CheXpert	Living17	Avg
ERM	69.1 ±4.7	57.6 ±0.8	63.2 ±1.2	66.4 ±2.3	82.1 ±0.8	76.8 ±0.9	35.0 ±4.1	68.6 ±0.2	80.4 ±0.2	50.1 ±0.9	41.7 ±3.4	27.7 ±1.1	59.9
Mixup	77.5 ±0.7	57.8 ±0.8	65.8 ±1.5	66.8 ±0.3	79.0 ±0.8	76.9 ±0.7	30.0 ±4.1	66.8 ±0.6	81.6 ±0.6	50.1 ±0.9	37.4 ±3.5	29.8 ±1.8	60.0
GroupDRO	73.1 ±0.4	68.3 ±0.9	61.5 ±1.8	64.1 ±0.8	83.1 ±0.7	76.4 ±0.2	31.1 ±0.9	67.4 ±0.5	83.7 ±0.1	59.2 ±0.3	74.7 ±0.3	31.1 ±1.0	64.5
CVaRDRO	75.5 ±2.2	60.2 ±3.0	62.9 ±3.8	48.2 ±3.4	83.5 ±0.5	74.8 ±0.8	27.8 ±2.3	68.0 ±0.2	65.6 ±1.5	50.2 ±0.9	50.2 ±1.8	27.3 ±1.6	57.8
JTT	71.2 ±0.5	48.3 ±1.5	51.0 ±4.2	65.1 ±1.6	82.6 ±0.4	77.0 ±0.4	30.6 ±2.3	64.9 ±0.3	83.8 ±0.1	57.9 ±2.1	60.4 ±4.8	28.3 ±1.1	60.1
LfF	75.0 ±0.7	53.0 ±4.3	42.2 ±7.2	57.3 ±5.7	72.3 ±1.3	70.1 ±1.4	28.8 ±2.0	62.2 ±2.4	84.0 ±0.1	50.1 ±0.9	13.7 ±9.8	26.4 ±1.3	52.9
LISA	77.5 ±0.7	57.8 ±0.8	65.8 ±1.5	66.8 ±0.3	79.0 ±0.8	76.9 ±0.7	30.0 ±4.1	66.8 ±0.6	81.6 ±0.6	50.1 ±0.9	37.4 ±3.5	29.8 ±1.8	60.0
ReSample	70.0 ±1.0	74.1 ±2.2	61.0 ±0.6	66.8 ±0.5	81.0 ±1.7	77.7 ±1.1	30.6 ±2.3	67.5 ±0.3	82.6 ±0.6	55.0 ±0.2	74.3 ±0.4	31.4 ±0.6	64.3
ReWeight	71.9 ±0.6	69.6 ±0.2	59.3 ±1.1	64.2 ±1.9	83.1 ±0.7	76.8 ±0.9	25.0 ±0.0	67.0 ±0.4	84.0 ±0.1	61.4 ±1.3	73.7 ±1.0	27.7 ±1.1	63.6
SqrtReWeight	71.0 ±1.4	66.9 ±2.2	68.6 ±1.1	63.8 ±2.4	82.6 ±0.4	76.8 ±0.9	32.8 ±3.5	68.0 ±0.4	83.1 ±0.2	61.2 ±0.6	68.5 ±1.6	27.7 ±1.1	64.2
CBLoss	74.4 ±1.2	65.4 ±1.4	67.3 ±0.2	63.6 ±2.4	83.1 ±0.0	76.8 ±0.9	31.7 ±3.6	67.6 ±0.3	84.0 ±0.1	50.2 ±0.9	74.0 ±0.7	27.7 ±1.1	63.8
Focal	71.6 ±0.8	56.9 ±3.4	61.9 ±1.1	62.4 ±2.0	81.0 ±0.4	71.9 ±1.2	30.6 ±2.3	68.7 ±0.4	70.9 ±9.8	50.0 ±0.9	42.1 ±4.0	26.9 ±0.6	57.9
LDAM	70.9 ±1.7	57.0 ±4.1	28.4 ±7.7	65.5 ±0.8	83.6 ±0.4	76.7 ±0.5	31.7 ±3.6	66.6 ±0.6	81.0 ±0.3	50.1 ±0.9	36.0 ±0.7	24.3 ±0.8	56.0
BSoftmax	74.1 ±0.9	69.6 ±1.2	58.3 ±1.1	63.6 ±2.4	82.6 ±0.4	76.1 ±2.0	35.6 ±1.8	67.6 ±0.6	83.8 ±0.3	58.6 ±1.8	73.8 ±1.0	28.6 ±1.4	64.4
DFR	89.0 ±0.2	73.7 ±0.8	64.4 ±0.1	63.8 ±0.0	81.4 ±0.1	74.4 ±1.8	38.0 ±3.8	67.1 ±0.4	80.2 ±0.0	60.8 ±0.4	75.8 ±0.3	26.3 ±0.4	66.2
CRT	76.3 ±0.8	69.6 ±0.7	67.8 ±0.3	65.4 ±0.2	83.1 ±0.0	78.2 ±0.5	33.3 ±0.0	68.1 ±0.1	83.4 ±0.0	61.8 ±0.1	74.6 ±0.4	31.1 ±0.1	66.1
ReWeightCRT	76.3 ±0.2	70.7 ±0.6	64.7 ±0.2	65.2 ±0.2	85.1 ±0.4	77.5 ±0.7	33.3 ±0.0	67.9 ±0.1	83.4 ±0.0	53.1 ±2.3	75.1 ±0.2	33.1 ±0.1	65.4

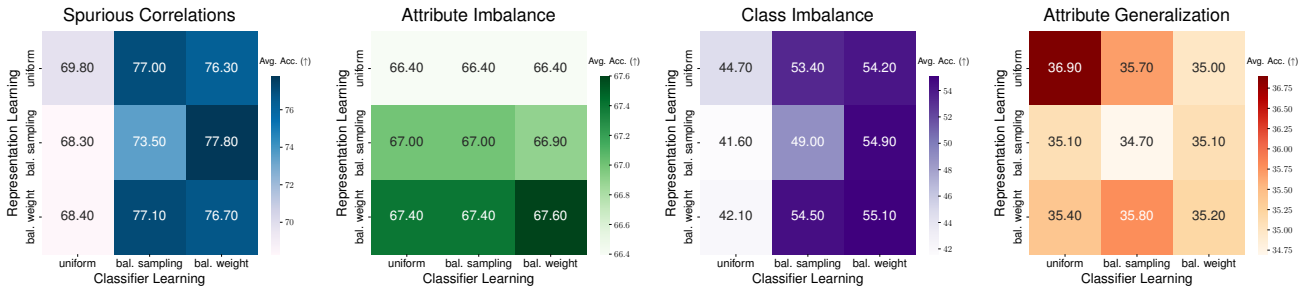


Figure 3. Averaged worst-group accuracy of different manners for representation learning and classifier learning under different shifts. Within each shift type, we average the results across datasets that belong to this shift to report the final accuracy. As observed, balanced classifier learning substantially improves the results for **SC** and **CI**, while balanced representation learning gives reasonable gains for **AI**; Yet, no stratified learning manners lead to performance gains under **AG** compared to vanilla ERM. Experimental details are in Sec. 5.3.

Table 4. Relative improvements over ERM when using stratified balanced representation or classifier learning under different shifts.

	SC	AI	CI	AG
REPRESENTATION	−0.3	+1.1	−0.2	−0.4
CLASSIFIER	+8.1	+0.0	+11.9	−0.4

rithms can provide consistent worst-group gains over ERM even in the absence of validation attributes, indicating that progress has been made for tackling these two specific shifts. Interestingly however, when it comes to *attribute imbalance*, little improvement is observed across datasets. In addition, the performance becomes even worse for *attribute generalization*. These findings stress that current advances are only made for specific shifts (i.e., SC and CI), while no progress has been made for the more challenging shifts such as AG.

Methods that decouple representation and classifier are more effective. When further zoom into the performance across all datasets in Table 3, a set of methods that decouple the training of representation and classifier (Izmailov et al., 2022; Kang et al., 2020) achieve remarkable gains over all other algorithms (highlighted in gray). As prior works also confirmed (Izmailov et al., 2022), features learned by ERM seem to be good enough under spurious correlations. These findings inspire us to further understand the role of *representation* and *classifier* in subpopulation shift, especially their behaviors under different subgroup shifts.

5.3. The Role of Representation and Classifier

We are motivated to explore the role of representation and classifier in subpopulation shift. In particular, we separate the whole network into two parts: the feature extractor and the classifier. We then employ three training strategies for representation and classifier learning, respectively: (1) *uniform*, which follows the normal ERM training; (2) *balanced sampling*, where balanced samples are drawn from each group (class if attribute not available) during training, and (3) *re-weighting*, where we re-weight all the samples by the inverse of the sample size of their groups (classes). Note that classifier re-balancing resembles CRT (Kang et al., 2020) and DFR (Izmailov et al., 2022). We train models following the above settings across all datasets, and average the results over datasets according to the type of shift.

Representation & classifier quality play different roles under different shifts. As Fig. 3 reveals, for **SC** and **CI**, balanced classifier learning (i.e., both re-sampling and re-weighting) can substantially improve the performance when fixing the representation, whereas different representation

Table 5. Test-set worst-group accuracy difference (%) between each selection strategy on each dataset, relative to the oracle which selects the best worst-group accuracy. Complete results across all datasets and all selection strategies are provided in Appendix D.3.

Selection Strategy	CelebA	CheXpert	CivilComments	MIMIC-CXR	MIMICNotes	MetaShift	Avg
Max Worst-Class Accuracy	-5.0 ±6.3	**-0.4** ±0.8	**-3.2** ±5.2	**-0.9** ±1.0	**-0.1** ±0.5	**-1.5** ±3.0	**-1.8**
Max Balanced Accuracy	**-4.4** ±5.4	-1.3 ±2.5	-3.5 ±5.8	-2.9 ±4.9	-2.3 ±6.2	-1.7 ±3.0	-2.7
Min Class Accuracy Diff	-6.1 ±9.1	-1.9 ±5.3	-4.1 ±8.0	-1.9 ±5.0	-0.3 ±1.2	-2.2 ±4.6	-2.7
Max Worst-Class F1	-13.4 ±10.4	-5.4 ±6.7	**-3.2** ±3.8	-2.5 ±2.2	-4.4 ±8.7	-1.8 ±3.3	-5.1
Max Overall AUROC	-12.2 ±10.3	-10.4 ±13.0	-8.2 ±9.0	-6.6 ±9.9	-10.0 ±16.5	-3.2 ±7.0	-8.4
Max Overall Accuracy	-18.6 ±12.0	-30.9 ±24.9	-13.7 ±9.5	-5.1 ±6.3	-19.9 ±26.0	-1.9 ±3.3	-15.0

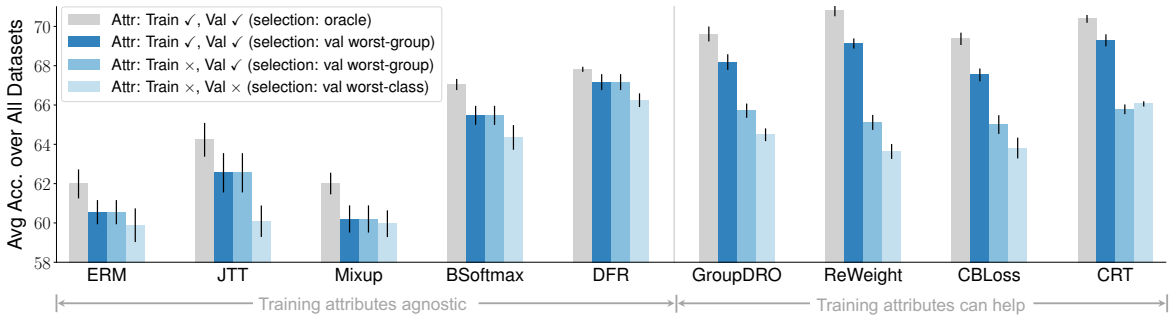


Figure 4. Averaged worst-group accuracy of various algorithms under different model selection and attribute availability settings.

learning schemes do not lead to notable gains when fixing the classifier learning manner. Interestingly, for **AI**, balancing the classifier does not lead to better performance, while balanced representation schemes can bring notable gains.

ERM features are not sufficient for subpopulation shift. Unlike recent works that claim ERM features are sufficient for out-of-distribution generalization (Izmailov et al., 2022; Rosenfeld et al., 2022), our above intriguing findings suggest that features learned via ERM may only be good enough for **certain** shifts. Concretely, improving the feature extractor still leads to notable gains especially for **AI**. The results in turn well explain the performance differences in Fig. 1, that SOTA algorithms with two-stage training do not improve worst-case accuracy under **AI** or **AG**.

Stratified balanced learning does not outperform ERM under AG. Finally, no stratified learning manners lead to performance gains under **AG**. As Table 4 summarizes, both stratified representation and classifier learning manners even exhibit negative gains for datasets that require **AG**. This reveals the intrinsic limitation of SOTA algorithms (Izmailov et al., 2022) against diverse types of subpopulation shift.

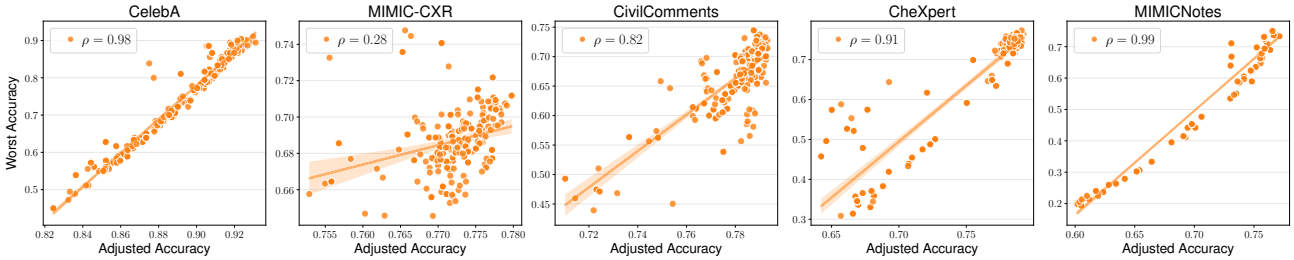
5.4. On Model Selection and Attribute Availability

Model selection (e.g., choice of hyperparameters, training checkpoints) and attribute availability affect subpopulation shift evaluation considerably, especially given that almost all SOTA algorithms need access to a group-annotated validation set for model selection (Idrissi et al., 2022). We study this problem in-depth, where we follow three settings mentioned earlier (i.e., the availability of both *training* and *validation* attributes), and summarize the results in Fig. 4.

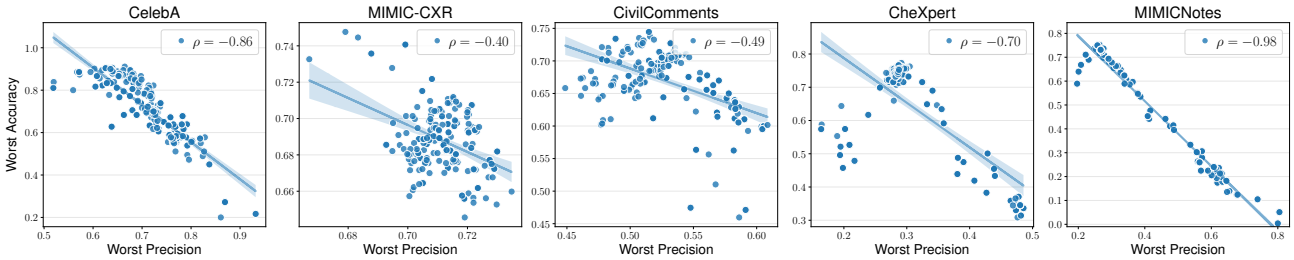
The importance of training attribute availability relies on algorithm properties. As Fig. 4 verifies, when training attribute is available, it can greatly boost the performance of algorithms that need group information (e.g., GroupDRO), while it does not bring benefits for attribute-agnostic methods (e.g., ERM, JTT).

Validation attribute may not be necessary once you have a good selection metric. We further investigate the performance without validation attributes. It is widely known that SOTA subpopulation shift methods rely on group labels for validation. Surprisingly however, we observe a relatively small accuracy drop over all methods when using a simple *worst-class accuracy* (degenerated from *worst-group* as attributes are unknown in validation) as selection metric. Specifically, comparing the last two bars across all methods in Fig. 4, the average accuracy drop is less than merely 2%. This striking finding contrasts with the literature, where large degradation (over 20%) is observed when using *average* accuracy as the metric without validation attributes. This suggests that if carefully choosing a metric for model selection, we can achieve minimal worst-group accuracy loss even in the absence of any attribute information.

Simple selection criterion using worst-class accuracy is surprisingly effective even without validation attribute. We examine different strategies for choosing when to stop during model training when no attribute annotations are available in both training and validation. We select six representative datasets and six representative selection strategies, respectively (full results across all datasets and all selection



(a) **Accuracy on the line:** Adjusted accuracy is *positively* correlated with WGA.



(b) **Accuracy on the inverse line:** Worst-case precision is *negatively* correlated with WGA.

Figure 5. Fundamental tradeoff between WGA and other evaluation metrics. Complete results for all metrics are in Appendix D.4.

strategies are in Appendix D.3). For each model, we utilize each stopping criterion over the validation set metrics computed throughout training, to determine its corresponding stopping point. We evaluate a variety of selection criteria in this way for a large variety of methods trained on each dataset. We compare each strategy with the oracle selection criteria, summarizing our results in Table 5. We observe that simply stopping when the *worst-class accuracy* reaches a maxima achieves the best worst-group accuracy on average. As expected, any selection criterion based on overall performance (e.g., accuracy, AUROC) performs much worse.

5.5. Metrics Beyond Worst-Group Accuracy

Worst-group accuracy (WGA) has long been treated as the gold-standard for assessing the model performance in subpopulation shift. Recent studies also discovered that WGA and model average performance are linearly correlated, a phenomenon called "*Accuracy on the line*" (Izmailov et al., 2022; Miller et al., 2021). However, WGA essentially assesses the worst-case (top-1) recall conditioned on attribute (Yang et al., 2022), which does not reflect other important metrics such as worst-case precision and calibration error. Whether models with high WGA will also perform better across these metrics remains unknown. Therefore, we further examine the relationship between WGA and other evaluation metrics that proposed in our benchmark.

Intrinsic tradeoff: Accuracy can be on the inverse line. Interestingly, we observe that not all metrics are positively correlated with WGA. In particular, we show scatter plots of WGA *vs.* other metrics for representative datasets. As Fig. 5(a) confirms, adjusted accuracy is linearly correlated with WGA, which is well aligned with existing observations (Izmailov et al., 2022). Interestingly however, for worst-case precision, the *positive* correlation does not hold anymore; instead, we observe a strong ***negative*** linear correlation, indicating an intrinsic tradeoff between WGA and worst-case precision. We show in Appendix D.4 that many metrics also possess such "*accuracy on the **inverse** line*" property, further verifying the inherent tradeoff between testing metrics.

Fundamental limitations of WGA as the only metric. The above observations highlight the complex relationship between WGA and other metrics: Certain metrics display high positive correlation, while many others show the opposite case. This finding uncovers the fundamental limitation of using only WGA to assess model performance in subpopulation shift: A well performed model with high WGA can however have low worst-case precision, which is alarming especially in critical applications such as medical diagnosis (e.g., `CheXpert`). Our observations emphasize the need for more realistic evaluation metrics in subpopulation shift.

5.6. Further Analysis

Impact of model architecture (Appendix D.5). We study the effect of different model architectures on subpopulation shift across various datasets and modalities. In particular, we employ ResNets and vision transformers (ViTs) for the image modality, and five different transformer-based language models for the text modality. We observe that on text datasets, base BERT models are already competitive over other architecture variants (Table 13). Yet, the results on image datasets are mixed when comparing the worst-group performance for ResNets and ViTs (Tables 14 and 15).

Impact of pretraining methods (Appendix D.5). We investigate how different pretraining methods affect the model performance under subpopulation shift. We consider both *supervised* and *self-supervised* pretraining using various SOTA methods. Similar to previous findings (Izmailov et al., 2022), we observe that supervised pretraining outperforms self-supervised counterparts for most of the experiments. The results may also suggest that better self-supervised schemes could be developed for tackling subgroup shifts.

Impact of pretraining datasets (Appendix D.5). Finally, we investigate whether increasing the pretraining dataset size could lead to better subgroup performance. We leverage ImageNet-21K (Ridnik et al., 2021) and SWAG (Singh et al., 2022) in addition to the default ImageNet-1K. Interestingly, we find consistent and significant worst-group performance gains when going from ImageNet-1K to ImageNet-21K to SWAG, indicating that larger and more diverse pretraining datasets seem to increase worst-group performance.

6. Conclusion

We systematically study the subpopulation shift problem, formalize a unified framework to define and quantify different types of subpopulation shift, and further set up a comprehensive benchmark for realistic evaluation. Our benchmark includes 20 SOTA methods and 12 real-world datasets across different domains. Based on over 10K trained models, we reveal several intriguing properties in subpopulation shift that have implications for future research, including divergent performance on different shifts, model selection criteria, and metrics to evaluate against. We hope our benchmark and findings will promote realistic and rigorous evaluations and inspire new advances in subpopulation shift.

Acknowledgements

This work was supported in part by the MIT-IBM Watson Research Collaboration grant, and a grant from Quanta Computing.

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Anthony, L. F. W., Kanding, B., and Selvan, R. Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. *arXiv preprint arXiv:2007.03051*, 2020.

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

Beltagy, I., Lo, K., and Cohan, A. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

Borkan, D., Dixon, L., Sorensen, J., Thain, N., and Vasserman, L. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pp. 491–500, 2019.

Cai, T., Gao, R., Lee, J., and Lei, Q. A theory of label propagation for subpopulation shift. In *International Conference on Machine Learning*, pp. 1170–1182. PMLR, 2021.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, 2019.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

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, 2019.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Creager, E., Jacobsen, J.-H., and Zemel, R. Environment inference for invariant learning. In *International Conference on Machine Learning*, pp. 2189–2200. PMLR, 2021.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. Class-balanced loss based on effective number of samples. In *CVPR*, 2019.

DeGrave, A. J., Janizek, J. D., and Lee, S.-I. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Duchi, J. and Namkoong, H. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 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*, 2018.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Gowda, S., Joshi, S., Zhang, H., and Ghassemi, M. Pulling up by the causal bootstraps: Causal data augmentation for pre-training debiasing. *arXiv preprint arXiv:2108.12510*, 2021.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

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

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. In *ICLR*, 2021.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

Han, Z., Liang, Z., Yang, F., Liu, L., Li, L., Bian, Y., Zhao, P., Wu, B., Zhang, C., and Yao, J. Umix: Improving importance weighting for subpopulation shift via uncertainty-aware mixup. *arXiv preprint arXiv:2209.08928*, 2022.

Hashimoto, T., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pp. 1929–1938. PMLR, 2018.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, 2016.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Idrissi, B. Y., Arjovsky, M., Pezeshki, M., and Lopez-Paz, D. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, pp. 336–351. PMLR, 2022.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 590–597, 2019.

Izmailov, P., Kirichenko, P., Gruver, N., and Wilson, A. G. On feature learning in the presence of spurious correlations. *arXiv preprint arXiv:2210.11369*, 2022.

Japkowicz, N. The class imbalance problem: Significance and strategies. In *Proc. of the Int'l Conf. on Artificial Intelligence*, volume 56, pp. 111–117. Citeseer, 2000.

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

Johnson, A. E., Pollard, T. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Peng, Y., Lu, Z., Mark, R. G., Berkowitz, S. J., and Horng, S. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

Joshi, N., Pan, X., and He, H. Are all spurious features in natural language alike? an analysis through a causal lens. *arXiv preprint arXiv:2210.14011*, 2022.

Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., and Kalantidis, Y. Decoupling representation and classifier for long-tailed recognition. *ICLR*, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Kornblith, S., Shlens, J., and Le, Q. V. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2661–2671, 2019.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. Connecting language and vision using crowdsourced dense image annotations. *Visual genome*, 2016.

Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., and Chi, E. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740, 2020.

Li, H., Pan, S. J., Wang, S., and Kot, A. C. Domain generalization with adversarial feature learning. In *CVPR*, 2018.

Li, T., Cao, P., Yuan, Y., Fan, L., Yang, Y., Feris, R., Indyk, P., and Katabi, D. Targeted supervised contrastive learning for long-tailed recognition. *arXiv preprint arXiv:2111.13998*, 2021.

Li, Z., Evtimov, I., Gordo, A., Hazirbas, C., Hassner, T., Ferrer, C. C., Xu, C., and Ibrahim, M. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. *arXiv preprint arXiv:2212.04825*, 2022.

Liang, W. and Zou, J. Metashift: A dataset of datasets for evaluating contextual distribution shifts and training conflicts. *arXiv preprint arXiv:2202.06523*, 2022.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection. In *ICCV*, pp. 2980–2988, 2017.

Liu, E. Z., Haghgoo, B., Chen, A. S., Raghunathan, A., Koh, P. W., Sagawa, S., Liang, P., and Finn, C. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pp. 6781–6792. PMLR, 2021.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019a.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., and Yu, S. X. Large-scale long-tailed recognition in an open world. In *CVPR*, 2019b.

Maleki, F., Muthukrishnan, N., Ovens, K., Reinhold, C., and Forghani, R. Machine learning algorithm validation: from essentials to advanced applications and implications for regulatory certification and deployment. *Neuroimaging Clinics*, 30(4):433–445, 2020.

Martinez, N., Bertran, M., and Sapiro, G. Minimax pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*, pp. 6755–6764. PMLR, 2020.

Martinez, N. L., Bertran, M. A., Papadaki, A., Rodrigues, M., and Sapiro, G. Blind pareto fairness and subgroup robustness. In *International Conference on Machine Learning*, pp. 7492–7501. PMLR, 2021.

Mehta, R., Albiero, V., Chen, L., Evtimov, I., Glaser, T., Li, Z., and Hassner, T. You only need a good embeddings extractor to fix spurious correlations. *arXiv preprint arXiv:2212.06254*, 2022.

Menon, A. K., Rawat, A. S., and Kumar, S. Overparameterisation and worst-case generalisation: friend or foe? In *International Conference on Learning Representations*, 2020.

Miller, G. A. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

Miller, J. P., Taori, R., Raghunathan, A., Sagawa, S., Koh, P. W., Shankar, V., Liang, P., Carmon, Y., and Schmidt, L. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pp. 7721–7735. PMLR, 2021.

Nam, J., Cha, H., Ahn, S., Lee, J., and Shin, J. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33: 20673–20684, 2020.

Nam, J., Kim, J., Lee, J., and Shin, J. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. *arXiv preprint arXiv:2204.02070*, 2022.

Paul, S. and Chen, P.-Y. Vision transformers are robust learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2071–2081, 2022.

Pfohl, S. R., Zhang, H., Xu, Y., Foryciarz, A., Ghassemi, M., and Shah, N. H. A comparison of approaches to improve worst-case predictive model performance over patient subpopulations. *Scientific reports*, 12(1):1–13, 2022.

Puli, A. M., Zhang, L. H., Oermann, E. K., and Ranganath, R. Out-of-distribution generalization in the presence of nuisance-induced spurious correlations. In *International Conference on Learning Representations*, 2021.

Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset shift in machine learning*. Mit Press, 2008.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Ren, J., Yu, C., Ma, X., Zhao, H., Yi, S., et al. Balanced meta-softmax for long-tailed visual recognition. In *NeurIPS*, 2020.

Ridnik, T., Ben-Baruch, E., Noy, A., and Zelnik-Manor, L. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021.

Rosenfeld, E., Ravikumar, P., and Risteski, A. Domain-adjusted regression or: Erm may already learn features sufficient for out-of-distribution generalization. *arXiv preprint arXiv:2202.06856*, 2022.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252, 2015.

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. *arXiv preprint arXiv:1911.08731*, 2019.

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 *ICLR*, 2020.

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 preprint arXiv:2008.04859*, 2020.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

Seyyed-Kalantari, L., Zhang, H., McDermott, M. B., Chen, I. Y., and Ghassemi, M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 27 (12):2176–2182, 2021.

Simon, H. A. Spurious correlation: A causal interpretation. *Journal of the American statistical Association*, 49(267): 467–479, 1954.

Singh, M., Gustafson, L., Adcock, A., de Freitas Reis, V., Gedik, B., Kosaraju, R. P., Mahajan, D., Girshick, R., Dollár, P., and Van Der Maaten, L. Revisiting weakly supervised pre-training of visual perception models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 804–814, 2022.

Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., and Beyer, L. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

Sun, B. and Saenko, K. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*, 2016.

Tang, K., Tao, M., Qi, J., Liu, Z., and Zhang, H. Invariant feature learning for generalized long-tailed classification. In *ECCV*, pp. 709–726. Springer, 2022.

Vapnik, V. N. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The caltech-ucsd birds-200-2011 dataset. 2011.

Wang, T., Yue, Z., Huang, J., Sun, Q., and Zhang, H. Self-supervised learning disentangled group representation as feature. *Advances in Neural Information Processing Systems*, 34:18225–18240, 2021.

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., et al. Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine*, 25(9):1337–1340, 2019.

Wiles, O., Gowal, S., Stimberg, F., Alvise-Rebuffi, S., Ktena, I., Dvijotham, K., and Cemgil, T. A fine-grained analysis on distribution shift. *arXiv preprint arXiv:2110.11328*, 2021.

Williams, A., Nangia, N., and Bowman, S. R. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

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.

Xiao, T., Li, H., Ouyang, W., and Wang, X. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, 2016.

Yang, Y. and Xu, Z. Rethinking the value of labels for improving class-imbalanced learning. In *NeurIPS*, 2020.

Yang, Y., Zha, K., Chen, Y.-C., Wang, H., and Katabi, D. Delving into deep imbalanced regression. In *ICML*, 2021.

Yang, Y., Wang, H., and Katabi, D. On multi-domain long-tailed recognition, imbalanced domain generalization and beyond. In *European Conference on Computer Vision (ECCV)*, 2022.

Yang, Y., Liu, X., Wu, J., Borac, S., Katabi, D., Poh, M.-Z., and McDuff, D. Simper: Simple self-supervised learning of periodic targets. In *International Conference on Learning Representations*, 2023.

Yao, H., Wang, Y., Li, S., Zhang, L., Liang, W., Zou, J., and Finn, C. Improving out-of-distribution robustness via selective augmentation. *arXiv preprint arXiv:2201.00299*, 2022.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pp. 12310–12320. PMLR, 2021.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.

Zhang, H., Lu, A. X., Abdalla, M., McDermott, M., and Ghassemi, M. Hurtful words: quantifying biases in clinical contextual word embeddings. In *proceedings of the ACM Conference on Health, Inference, and Learning*, pp. 110–120, 2020.

Zhang, X., Zhou, L., Xu, R., Cui, P., Shen, Z., and Liu, H. Nico++: Towards better benchmarking for domain generalization. *arXiv preprint arXiv:2204.08040*, 2022.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

Zong, Y., Yang, Y., and Hospedales, T. Medfair: Benchmarking fairness for medical imaging. *arXiv preprint arXiv:2210.01725*, 2022.

A. Limitations and Broader Impacts

Limitations. We acknowledge several limitations of our benchmark and analyses. First, we have used 12 real-world predictive datasets in our benchmark. However, real-world data can have many complexities including potential mislabelling in both attributes and labels. We do not consider this effect, though it would be interesting to examine it in a synthetic setting. Moreover, prior work has shown that in the case of multiple spurious attributes, reducing reliance on one can increase reliance on another (Li et al., 2022). We only consider a single attribute in this benchmark, though an evaluation of this effect in the context of model selection criteria would be an interesting direction of future research.

Potential Negative Impacts. There are several potential negative social impacts of our work. First, we assume throughout the work that we would like to have models that are robust to subpopulation shift. However, in practice, this comes at the cost of overall accuracy on the training distribution. There may be cases where the practitioner would like to maximize overall accuracy regardless of spurious correlations, and thus subpopulation shift methods would worsen overall performance and potentially cause excess harm. Next, we recognize that the large grid of deep models trained for our evaluations likely resulted in a significant carbon footprint (Anthony et al., 2020). However, we hope that the insights provided in this paper will reduce the number of models and training steps (and therefore carbon emissions) required by future practitioners. Finally, we have constructed several models in this paper that utilize clinical data for clinical predictive tasks. We do not advocate for blind deployment of these models in any way, as there are many issues that need to be verified and resolved before their deployment, such as real-world clinical testing, privacy, fairness, interpretability, and regulatory requirements (Maleki et al., 2020; Wiens et al., 2019).

B. Details of the Subpopulation Shift Benchmark

B.1. Dataset Details

We explore subpopulation shift using **12** real-world datasets from a variety of domains including computer vision, natural language processing, and healthcare applications. We provide example inputs for each dataset in Table 6 and Table 7. Note that we omit showing examples for `MIMIC-CXR`, `MIMICNotes`, and `CXRMultisite` to comply with the *PhysioNet Credentialed Health Data Use Agreement*. Below, we provide detailed descriptions for each dataset in our benchmark.

`Waterbirds` **(Wah et al., 2011).** `Waterbirds` is a commonly used binary classification image dataset in the spurious correlations setting, constructed by placing images from the Caltech-UCSD Birds-200-2011 (CUB) dataset (Wah et al., 2011) over backgrounds from the Places dataset (Zhou et al., 2017). The task is to classify whether a bird is a landbird or a waterbird, where the spurious attribute is the background (water or land). We use standard train/val/test splits given by prior work (Idrissi et al., 2022).

`CelebA` **(Liu et al., 2015).** `CelebA` is a binary classification image dataset consisting of over 200,000 celebrity face images. The task, which is also used widely in the spurious correlations literature, is to predict the hair color of the person (blond *vs.* non-blond), where the spurious correlation is the gender. We also use standard dataset splits from prior work (Idrissi et al., 2022). The dataset is licensed under the *Creative Commons Attribution 4.0 International* license.

`MetaShift` **(Liang & Zou, 2022).** `MetaShift` is a general method of creating image datasets from the Visual Genome project (Krishna et al., 2016). Here, we make use of the pre-processed Cat *vs.* Dog dataset, where the goal is to distinguish between the two animals. The spurious attribute is the image background, where cats and more likely to be indoors, and dogs are more likely to be outdoors. We use the "unmixed" version generated from the authors' codebase.

`CivilComments` **(Borkan et al., 2019).** `CivilComments` is a binary classification text dataset, where the goal is to predict whether a internet comment contains toxic language. The spurious attribute is whether the text contains reference to eight demographic identities (*male, female, LGBTQ, Christian, Muslim, other religions, Black,* and *White*). We use the standard splits provided by the WILDS benchmark (Koh et al., 2021).

`MultiNLI` **(Williams et al., 2017).** `MultiNLI` is a text classification dataset with 3 classes, where the target is the natural language inference relationship between the premise and the hypothesis (neutral, contradiction, or entailment). The spurious attribute is whether negation appears in the text, as negation is highly correlated with the contradiction label. We use standard train/val/test splits given by prior work (Idrissi et al., 2022).

Table 6. Example inputs for **image datasets** in our benchmark. We omit showing samples for `MIMIC-CXR` and `CXRMultisite` to comply with the *PhysioNet Credentialed Health Data Use Agreement*.

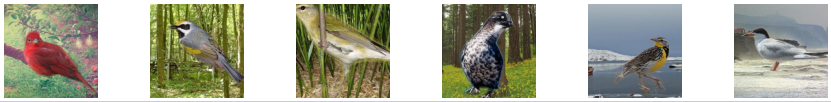
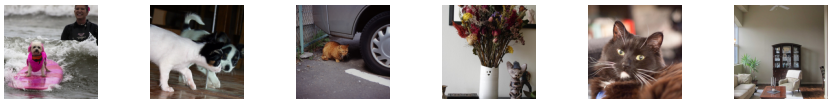
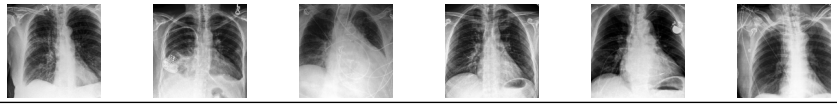
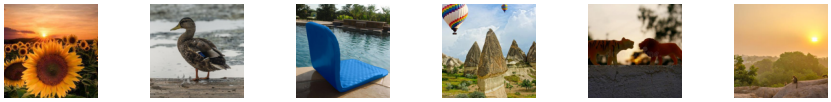
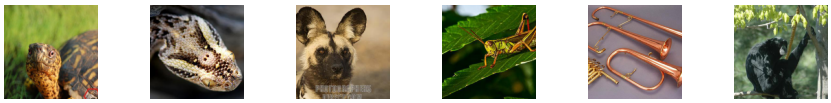
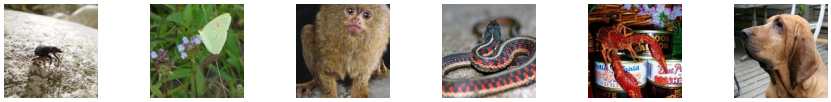
Dataset	Examples
`Waterbirds`	
`CelebA`	
`MetaShift`	
`CheXpert`	
`NICO++`	
`ImageNetBG`	
`Living17`	

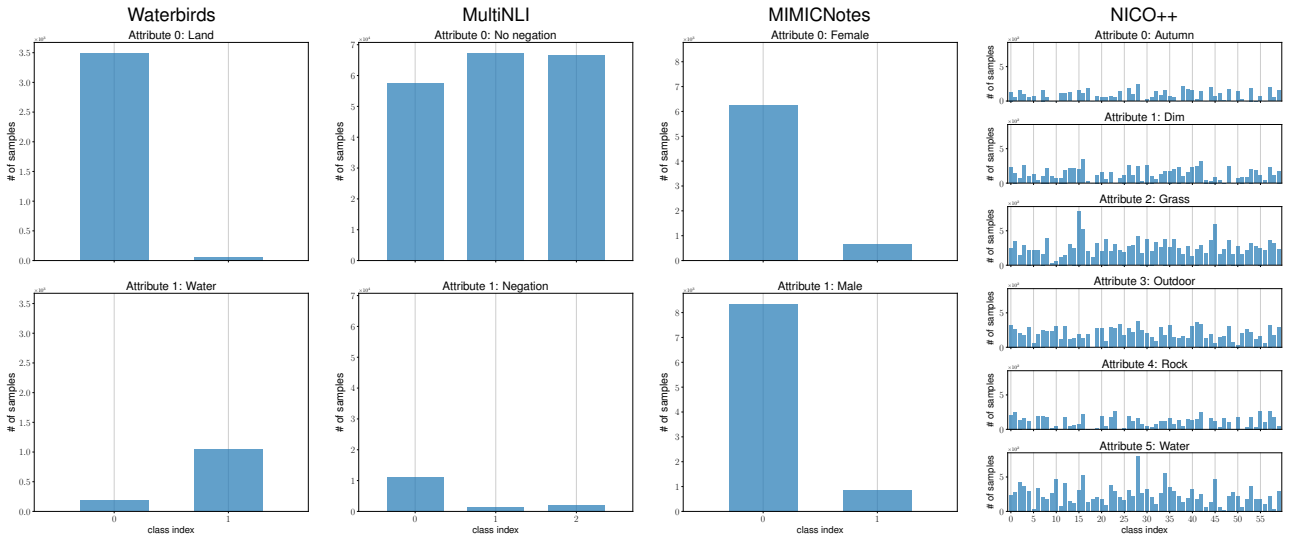
Table 7. Example inputs for **text datasets** in our benchmark. We omit showing samples for `MIMICNotes` to comply with the *PhysioNet Credentialed Health Data Use Agreement*.

Dataset	Examples
`CivilComments`	"Munchins looks like a munchins. The man who dont want to show his taxes, will tell you everything..." "The democratic party removed the filibuster to steamroll its agenda. Suck it up boys and girls." "so you dont use oil? no gasoline? no plastic? man you ignorant losers are pathetic."
`MultiNLI`	"The analysis proves that there is no link between PM and bronchitis." "Postal Service were to reduce delivery frequency." "The famous tenements (or lands) began to be built."

`MIMIC-CXR` **(Johnson et al., 2019).** `MIMIC-CXR` is a chest X-ray dataset originating from the Beth Israel Deaconess Medical Center from Boston, Massachusetts containing over 300,000 images. We use "No Finding" as the label, where a positive label means that the patient has no illness. Inspired by prior work (Seyyed-Kalantari et al., 2021), we use the intersection of race (*White, Black, Other*) and gender as attributes. We randomly split the dataset into 85% train, 5% validation, and 10% test splits.

`CheXpert` **(Irvin et al., 2019).** `CheXpert` is a chest X-ray dataset originating from the Stanford University Medical center containing over 200,000 images. We use the same data processing setup as `MIMIC-CXR`.

`CXRMultisite` **(Puli et al., 2021).** `CXRMultisite` is a dataset proposed by Puli et al. (2021) which combines MIMIC-CXR (Johnson et al., 2019) and CheXpert (Irvin et al., 2019) to create a semi-synthetic spurious correlation. The task is to predict pneumonia, and the dataset is constructed such that 90% of the patients with pnuemonia are from MIMIC-CXR, and 90% of the healthy patients are from CheXpert. Thus, the site where the image was taken is the spurious correlation. We



(a) Spurious correlations. (b) Attribute imbalance. (c) Class imbalance. (d) Attribute generalization.

Figure 6. Typical label distributions for different types of subpopulation shift.

create this correlation by subsampling. We randomly split the dataset into 85% train, 5% validation, and 10% test splits.

`MIMICNotes` **(Johnson et al., 2016).** `MIMICNotes` is a dataset used in a prior work (Chen et al., 2019) showing differences in error rate between demographic groups in predicting mortality from clinical notes in MIMIC-III (Johnson et al., 2016). Following their work, we reproduce their dataset which consists of featurizing the first 48 hours of clinical text from a patient's hospital stay using the top 5,000 TF-IDF features. We use gender as the attribute.

`NICO++` **(Zhang et al., 2022).** `NICO++` is a large-scale benchmark for domain generalization. Here, we use data from Track 1 (the common context generalization) of their challenge. We only use their training dataset, which consists of 60 classes and 6 common attributes (*autumn, dim, grass, outdoor, rock, water*). To transform this dataset into the attribute generalization setting, we select all (attribute, label) pairs with less than 75 samples, and remove them from our training split, so they are only used for validation and testing. For each (attribute, label) pair, we use 25 samples for validation and 50 samples for testing, and use the remaining data as training samples.

`ImageNetBG` **(Xiao et al., 2020).** `ImageNetBG` is a benchmark created with the goal of evaluating the reliance of ImageNet classifiers on the background. The authors first created a subset of ImageNet with 9 classes (ImageNet-9), and annotated bounding boxes so that backgrounds can be removed. In our setup, we train models on the original IN-9L (with backgrounds), and evaluate our model on `MIXED-RAND`. Note that attribute (i.e., the label of the background) is not available for this dataset. This can be thought of as an attribute generalization setting, as we do not observe test backgrounds during training.

`Living17` **(Santurkar et al., 2020).** `Living17` is a dataset created as part of the BREEDS benchmark for subpopulation shift. Their setup is slightly different from a traditional subpopulation shift setting, where subpopulations are defined using a WordNet hierarchy, and the goal is to generalize to unseen subclasses in the same hierarchy level. As such, it is difficult to define the notion of an "attribute" in this setting. In particular, the `Living17` dataset consists of images of living objects across 17 classes. We train our models on the source subclasses and evaluate them on the target subclasses.

Label distribution for different types of subpopulation shift. Finally, we provide typical label distributions for different subpopulation shift types in Fig. 6. As highlighted, different shifts exhibit distinct types of label distributions, resulting in different properties in learning. For `NICO++` (Fig. 6(d)), certain attributes have no training samples in certain classes.

B.2. Algorithm Details

Our benchmark contains a large number of algorithms that span different learning strategies. We group them according to their categories, and provide detailed descriptions for each algorithm below.

- *Vanilla:* The empirical risk minimization (**ERM**) (Vapnik, 1999) minimizes the sum of errors across all samples.

- *Subgroup robust methods:* Group distributionally robust optimization (**GroupDRO**) (Sagawa et al., 2020) performs ERM while increasing the importance of groups with larger errors. **CVaRDRO** (Duchi & Namkoong, 2018) proposes a variant of GroupDRO that dynamically weights data samples that have the highest losses. **LfF** (Nam et al., 2020) trains two models simultaneously, where the first model is biased and the second one is debiased by re-weighting the gradient of the loss. Just train twice (**JTT**) (Liu et al., 2021) first trains an ERM model to identify minority groups in the training set and then trains a second ERM model with the identified samples being re-weighted. **LISA** (Yao et al., 2022) learns invariant predictors through data interpolation within and across attributes. Deep feature re-weighting (**DFR**) (Izmailov et al., 2022) first trains an ERM model, then retrains the last layer of the model using a balanced validation set with group annotations.

- *Data augmentation:* **Mixup** (Zhang et al., 2018) performs ERM on linear interpolations of randomly sampled training examples and their labels.

- *Domain-invariant representation learning:* Invariant risk minimization (**IRM**) (Arjovsky et al., 2019) learns a feature representation such that the optimal linear classifier on top of that representation matches across domains. Deep correlation alignment (**CORAL**) (Sun & Saenko, 2016) matches the mean and covariance of feature distributions. Maximum mean discrepancy (**MMD**) (Li et al., 2018) matches the MMD (Gretton et al., 2012) of feature distributions. Note that all methods in this category require group annotations during training.

- *Imbalanced learning:* **ReSample** (Japkowicz, 2000) and **ReWeight** (Japkowicz, 2000) simply re-sample or re-weight the inputs according to the number of samples per class. Focal loss (**Focal**) (Lin et al., 2017) reduces the relative loss for well-classified samples and focuses on difficult samples. Class-balanced loss (**CBLoss**) (Cui et al., 2019) proposes re-weighting by the inverse effective number of samples. The LDAM loss (**LDAM**) (Cao et al., 2019) employs a modified marginal loss that favors minority samples more. Balanced-Softmax (**BSoftmax**) (Ren et al., 2020) extends Softmax to an unbiased estimation that considers the number of samples in each class. Classifier re-training (**CRT**) (Kang et al., 2020) decomposes the representation and classifier learning into two stages, where it fine-tunes the classifier using class-balanced sampling with representation fixed in the second stage. **ReWeightCRT** (Kang et al., 2020) is a re-weighting variant of CRT.

B.3. Evaluation Metrics

We describe in detail all the evaluation metrics we used in our experiments.

Average & Worst Accuracy. The average accuracy is defined as the accuracy over all samples. For worst-group accuracy (WGA), we compute the accuracy over all subgroups in the test set and report the worst one. When viewing each class as a subgroup, WGA degenerates to the worst-class accuracy.

Average & Worst Precision. Precision is defined as $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$, where TP is the number of true positives and FP the number of false positives. Average precision simply takes the average precision score over all classes, whereas the worst precision reports the lowest precision value across classes.

Average & Worst F1-score. The F1-score is defined as the harmonic mean of precision and recall. Average F1-score simply takes the average F1-score over all classes, whereas the worst F1-score reports the lowest value across all classes.

Adjusted Accuracy. Adjusted accuracy is defined as the average accuracy on a group-balanced dataset, which accounts for the data imbalance over subgroups.

Balanced Accuracy. Balanced accuracy is defined as the average of recall obtained on each class, taking the imbalance over classes into account.

AUROC. Following the common evaluation practice for the medical datasets used in our benchmark (Irvin et al., 2019; Johnson et al., 2019), we also include the area under the receiver operating characteristic curve (AUROC) for evaluation.

ECE (Guo et al., 2017). The expected calibration error (ECE) is defined as the difference in expected accuracy and expected confidence, which measures how close the output pseudo-probabilities match with the actual probabilities of a correct prediction (lower the better).

B.4. Model Selection Protocol

There has been an increasing interest in model selection within the literature on out-of-distribution generalization (Gulrajani & Lopez-Paz, 2021). In subpopulation shift, model selection becomes essential especially when attributes are completely unknown in both training and validation set. Significant drop (over 20%) in worst-group test accuracy has been reported if using the highest *average* validation accuracy as the model selection criterion without any group annotations (Izmailov et al., 2022).

Our benchmark provides different model selection strategies based on various evaluation metrics as described in Appendix B.3. Throughout the paper, we mainly use *worst-group accuracy* as the metric for model selection (which degenerates to *worst-class accuracy* when attributes are unknown in the validation set). Nevertheless, one can specify any aforementioned metric during model selection stage for experimenting with different selection strategies.

C. Experimental Settings

C.1. Implementation Details

Following (Gulrajani & Lopez-Paz, 2021; Izmailov et al., 2022), we use pretrained ResNet-50 model (He et al., 2016) as the backbone network for image datasets (except for `Living17`, which we train from scratch), and use pretrained BERT model (Idrissi et al., 2022) for all text datasets. We employ a three-layer MLP for `MIMICNotes` dataset given its simplicity. For all image datasets, we follow standard pre-processing steps (Idrissi et al., 2022): resize and center crop the image to 224×224 pixels, and perform normalization using the ImageNet channel statistics. Following the literature (Idrissi et al., 2022; Izmailov et al., 2022), we use the AdamW optimizer (Kingma & Ba, 2015) for all text datasets, and use SGD with momentum for all image datasets. We train all models for 5,000 steps on `Waterbirds` and `MetaShift`, 10,000 steps on `MIMICNotes` and `ImageNetBG`, 20,000 steps on `CheXpert` and `CXRMultisite`, and 30,000 steps on all other datasets to ensure convergence.

C.2. Hyperparameters Search Protocol

For a fair evaluation across different algorithms, following the training protocol in (Gulrajani & Lopez-Paz, 2021), for each algorithm we conduct a random search of 16 trials over a joint distribution of its all hyperparameters. We then use the validation set to select the best hyperparameters for each algorithm, fix them and rerun the experiments under 3 different random seeds to report the final average results (and standard deviation). Such process ensures the comparison is best-versus-best, and the hyperparameters are optimized for all algorithms.

We detail the hyperparameter choices for each algorithm in Table 8.

D. Additional Analysis and Studies

D.1. Quantifying the Degree of Different Shifts

In order to quantify the degree of each shift for each dataset relative to others, we use several simple metrics (see Table 9, Table 10, and Table 11). For spurious correlations, we use:

- The Mutual Information (**MI**) between A and Y, $I(A;Y)$.
- The Normalized Mutual Information (**NMI**) between A and Y, where norm $I(A;Y) = 1$ indicates that the two are perfectly correlated:
$$\text{norm } I(A;Y) = \frac{2I(A;Y)}{H(Y) + H(A)}.$$
- **Cramer's V**, which is an association measure based on the Chi-squared test statistic. It has a range of $[0, 1]$, where 1 indicates perfect correlation.
- **Tschuprow's T**, which is closely related to Cramer's V. It also has a range of $[0, 1]$.

Note that we only examine the correlation between A and Y, but not the degree of effectiveness to which A can be inferred from X. This is an important component, as the model can not take advantage of the spurious correlation if it could not

Table 8. Hyperparameters search space for all experiments.

Condition	Parameter	Default value	Random distribution
General:			
ResNet	learning rate	0.001	$10^{\text{Uniform}(-4,-2)}$
	batch size	108	$2^{\text{Uniform}(6,7)}$
BERT	learning rate	0.00001	$10^{\text{Uniform}(-5.5,-4)}$
	batch size	32	$2^{\text{Uniform}(3,5.5)}$
	dropout	0.5	$\text{RandomChoice}([0, 0.1, 0.5])$
MLP	learning rate	0.001	$10^{\text{Uniform}(-4,-2)}$
	batch size	256	$2^{\text{Uniform}(7,10)}$
Algorithm-specific:			
IRM	lambda	100	$10^{\text{Uniform}(-1,5)}$
	iterations of penalty annealing	500	$10^{\text{Uniform}(0,4)}$
GroupDRO	eta	0.01	$10^{\text{Uniform}(-3,-1)}$
Mixup	alpha	0.2	$10^{\text{Uniform}(0,4)}$
CVaRDRO	alpha	0.1	$10^{\text{Uniform}(-2,0)}$
JTT	first stage step fraction	0.5	$\text{Uniform}(0.2, 0.8)$
	lambda	10	$10^{\text{Uniform}(0,2.5)}$
LISA	alpha	2	$10^{\text{Uniform}(-1,1)}$
	p_select	0.5	$\text{Uniform}(0, 1)$
LfF	q	0.7	$\text{Uniform}(0.05, 0.95)$
DFR	regularization	0.1	$10^{\text{Uniform}(-2,0.5)}$
CORAL, MMD	gamma	1	$10^{\text{Uniform}(-1,1)}$
Focal	gamma	1	$0.5 * 10^{\text{Uniform}(0,1)}$
CBLoss	beta	0.9999	$1 - 10^{\text{Uniform}(-5,-2)}$
LDAM	max_m	0.5	$10^{\text{Uniform}(-1,-0.1)}$
	scale	30	$\text{RandomChoice}([10, 30])$

be learnt easily. However, we would expect that most attributes (e.g., words in text, image backgrounds) should be easily inferred from the inputs for the datasets we examine.

For attribute and class imbalance, we use the following metrics (shown for the class imbalance case):

- **Entropy**: $H(Y)$.
- **Normalized Entropy**, where norm $H(Y) = 1$ means that the distribution is uniform (i.e., no imbalance):

$$\text{norm } H(Y) = \frac{H(Y)}{\log |\text{supp}(Y)|}.$$

- Difference between the probability of the most frequent class and the probability of the least frequent class ($p_{\text{max}} - p_{\text{min}}$).

For attribute generalization, we simply examine whether there exist any subpopulations in the test set which do not appear during training.

Table 9. Metrics for quantifying the degree of *spurious correlations*.

Dataset	MI$^{\uparrow}$	NMI$^{\uparrow}$	Cramer$^{\uparrow}$	Tschuprow$^{\uparrow}$
Waterbirds	0.37	0.67	0.87	0.87
CelebA	0.06	0.11	0.31	0.31
MetaShift	0.09	0.13	0.41	0.41
CivilComments	0.02	0.02	0.19	0.11
MultiNLI	0.03	0.04	0.25	0.21
MIMIC-CXR	0.01	0.01	0.15	0.10
MIMICNotes	$< 1e^{-4}$	$< 1e^{-4}$	0.01	0.01
CXRMultisite	0.03	0.13	0.32	0.32
CheXpert	$< 1e^{-3}$	$< 1e^{-3}$	0.03	0.02
NICO++	0.11	0.04	0.20	0.11
ImageNetBG	–	–	–	–
Living17	–	–	–	–

Table 10. Metrics for quantifying the degree of *attribute imbalance*.

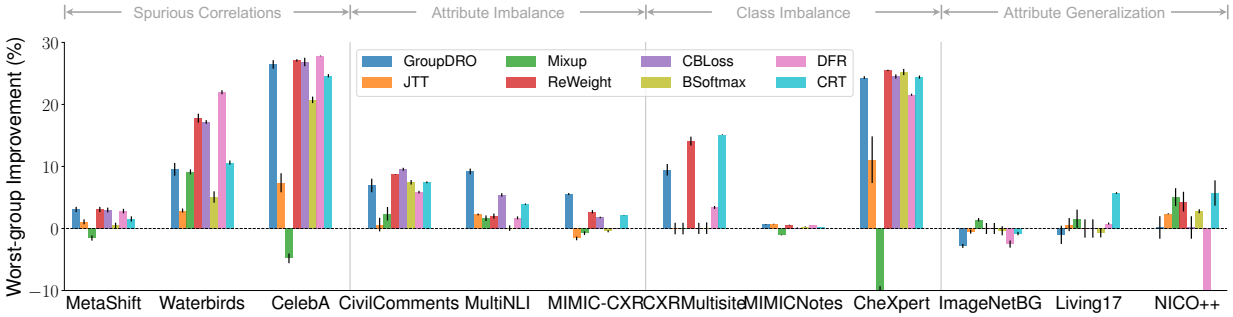
Dataset	Entropy$^{\downarrow}$	N. Entropy$^{\downarrow}$	$p_{\text{max}} - p_{\text{min}}{}^{\uparrow}$
Waterbirds	0.82	0.82	0.48
CelebA	0.98	0.98	0.16
MetaShift	0.99	0.99	0.14
CivilComments	2.78	0.93	0.20
MultiNLI	0.37	0.37	0.86
MIMIC-CXR	2.33	0.90	0.27
MIMICNotes	0.99	0.99	0.14
CXRMultisite	0.51	0.51	0.77
CheXpert	2.20	0.85	0.32
NICO++	2.47	0.96	0.17
ImageNetBG	–	–	–
Living17	–	–	–

Table 11. Metrics for quantifying the degree of *class imbalance*.

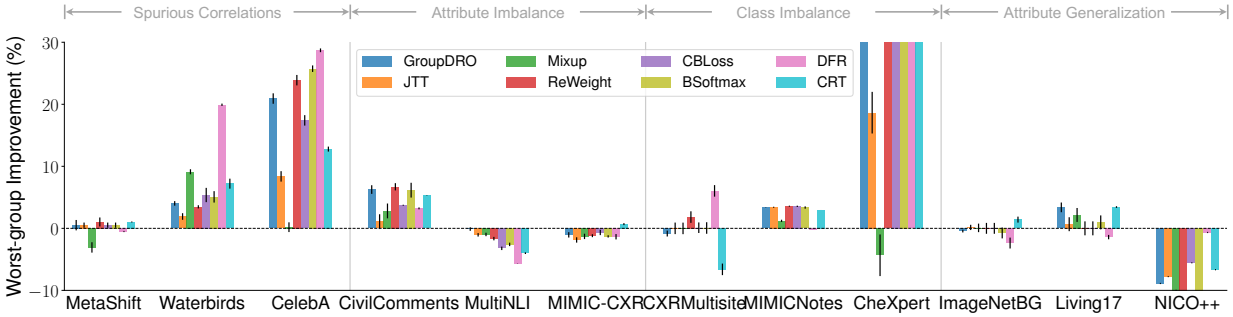
Dataset	Entropy$^{\downarrow}$	N. Entropy$^{\downarrow}$	$p_{\text{max}} - p_{\text{min}}{}^{\uparrow}$
Waterbirds	0.78	0.78	0.54
CelebA	0.61	0.61	0.70
MetaShift	0.99	0.99	0.13
CivilComments	0.67	0.67	0.65
MultiNLI	1.58	0.99	0.001
MIMIC-CXR	0.97	0.97	0.20
MIMICNotes	0.45	0.45	0.81
CXRMultisite	0.12	0.12	0.97
CheXpert	0.47	0.47	0.80
NICO++	5.81	0.98	0.03
ImageNetBG	3.17	1	0
Living17	4.09	1	0

D.2. Improvements across Different Shifts & Settings

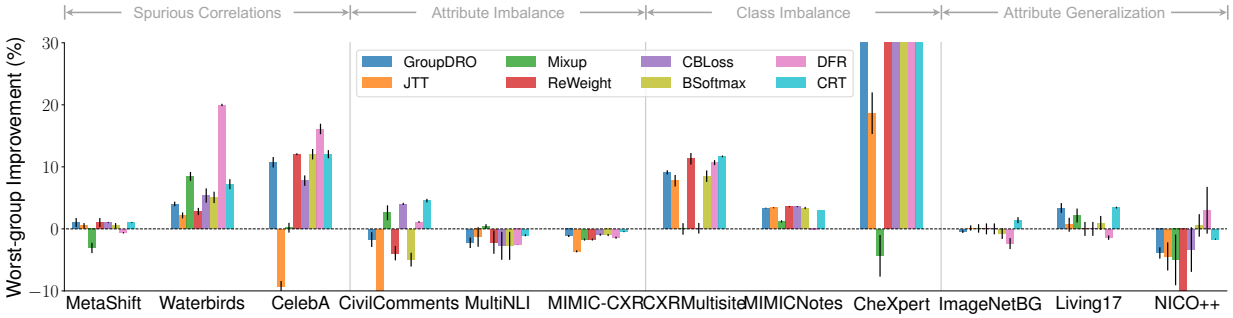
We show in Fig. 7 the complete results on worst-group performance improvements over ERM under different settings. As can be observed from all figures, algorithmic advances have been made for *spurious correlations* and *class imbalance*, where consistent improvements can be obtained across different training & validation settings. Yet, small overall improvements are observed for *attribute imbalance*, while almost no performance gains can be obtained for *attribute generalization*, indicating the limitation of SOTA algorithms on tackling these types of subpopulation shift.



(a) Train & validation attributes both known (*oracle selection*).



(b) Train attributes unknown, but validation attributes known (*worst-group accuracy selection*).



(c) Train & validation attributes both unknown (*worst-class accuracy selection*).

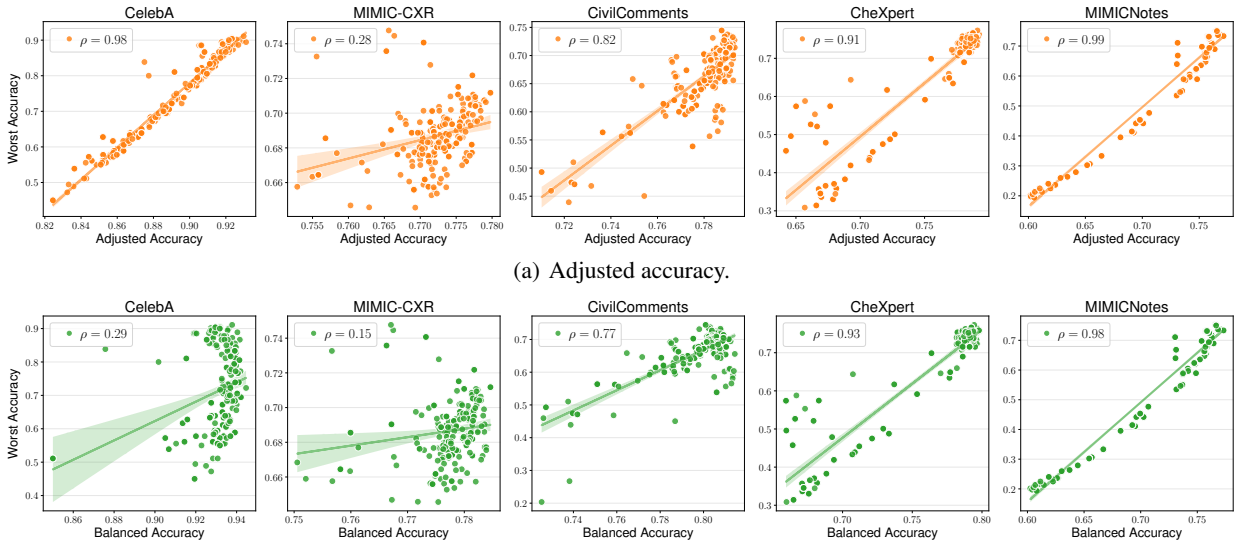
Figure 7. Complete results on worst-group performance improvements over ERM under different settings.

Table 12. Test-set worst-group accuracy difference (%) between each selection strategy on each dataset, relative to the oracle which selects the best test-set worst-group accuracy. Note that we have only defined AUPRC and Brier score for the binary classification case.

Selection Strategy	CXRMultisite	CelebA	CheXpert	CivilComments	ImageNetBG	Living17	MIMIC-CXR	MIMICNotes	MetaShift	MultiNLI	NICO++	Waterbirds	Avg
Max Worst-Class Accuracy	-6.9 ±10.7	-5.0 ±6.3	**-0.4** ±0.8	**-3.2** ±5.2	**-0.7** ±1.3	**-1.6** ±2.3	**-0.9** ±1.0	**-0.1** ±0.5	**-1.5** ±3.0	**-1.9** ±2.9	**-5.3** ±5.6	**-0.8** ±1.4	**-2.4**
Max Balanced Accuracy	-6.9 ±10.7	-4.4 ±5.4	-1.3 ±2.5	-3.5 ±5.8	-0.9 ±1.6	-4.5 ±5.4	-2.9 ±4.9	-2.3 ±6.2	-1.7 ±3.0	-3.7 ±3.9	-7.0 ±5.8	-1.3 ±1.9	-3.4
Min Class Accuracy Diff	**-6.2** ±10.3	-6.1 ±9.1	-1.9 ±5.3	-4.1 ±8.0	-2.8 ±13.0	-5.1 ±10.0	-1.9 ±5.0	-0.3 ±1.2	-2.2 ±4.6	-5.7 ±8.6	-27.2 ±15.4	-2.4 ±4.8	-5.5
Max Worst-Class F1	-7.7 ±11.3	-13.4 ±10.4	-5.4 ±6.7	-3.2 ±3.8	-0.8 ±1.2	-3.5 ±4.4	-2.5 ±2.2	-4.4 ±8.7	-1.8 ±3.3	-2.3 ±3.0	-6.7 ±6.3	-2.6 ±3.5	-4.5
Max Macro Avg F1	-8.2 ±11.6	-14.3 ±10.6	-7.7 ±9.8	-5.1 ±4.7	-0.9 ±1.5	-4.4 ±5.3	-2.8 ±4.5	-8.2 ±13.2	-1.8 ±2.9	-3.3 ±3.4	-7.0 ±5.8	-3.1 ±4.0	-5.6
Min Per-Class Recall Stdev.	-6.2 ±10.3	-6.1 ±9.1	-1.9 ±5.3	-4.1 ±8.0	-2.3 ±11.5	-5.5 ±9.1	-1.9 ±5.0	-0.3 ±1.2	-2.2 ±4.6	-5.6 ±8.7	-29.7 ±14.3	-2.4 ±4.8	-5.7
Max Weighted Avg Precision	-8.3 ±11.5	-13.5 ±10.1	-6.3 ±11.1	-5.7 ±8.6	-0.8 ±1.3	-7.5 ±7.8	-4.3 ±6.4	-12.6 ±21.5	-3.3 ±8.0	-3.4 ±4.7	-6.8 ±5.5	-4.9 ±10.1	-6.5
Max Overall AUROC	-10.0 ±12.5	-12.2 ±10.3	-10.4 ±13.0	-8.2 ±9.0	-1.1 ±2.1	-5.5 ±6.7	-6.6 ±9.9	-10.0 ±16.5	-3.2 ±7.0	-4.4 ±5.8	-6.9 ±6.3	-2.6 ±6.1	-6.7
Max Overall AUPRC	-10.0 ±12.5	-13.0 ±10.3	-11.6 ±11.9	-8.1 ±8.9	-	-	-7.3 ±10.2	-9.6 ±16.3	-2.7 ±6.2	-	-	-4.0 ±9.5	-8.3
Min Overall BCE	-8.2 ±11.5	-18.1 ±13.2	-18.7 ±16.4	-13.1 ±12.3	-0.9 ±1.6	-7.2 ±7.3	-7.2 ±12.0	-14.3 ±20.7	-3.7 ±7.7	-6.2 ±7.8	-7.6 ±6.1	-12.5 ±18.4	-9.8
Max Per-class Precision	-8.2 ±11.7	**-3.0** ±8.9	-6.8 ±12.5	-14.8 ±24.3	-7.6 ±18.4	-19.3 ±15.9	-9.4 ±12.7	-12.6 ±22.4	-9.9 ±17.4	-6.6 ±10.1	-14.8 ±11.8	-5.3 ±12.4	-9.8
Max Overall Accuracy	-8.2 ±11.4	-18.6 ±12.0	-30.9 ±24.9	-13.7 ±9.5	-0.9 ±1.6	-4.5 ±5.4	-5.1 ±6.3	-19.9 ±26.0	-1.9 ±3.3	-3.7 ±3.9	-7.1 ±5.8	-7.2 ±11.7	-10.2
Min Overall Brier Score	-8.2 ±11.5	-18.8 ±13.1	-19.6 ±16.6	-13.5 ±12.3	-	-	-7.1 ±12.0	-15.1 ±21.6	-2.7 ±5.3	-	-	-6.9 ±11.0	-11.5
Min Overall ECE	-8.2 ±11.5	-20.5 ±15.7	-20.3 ±17.4	-14.4 ±13.5	-16.9 ±33.6	-28.8 ±19.6	-12.3 ±18.2	-16.2 ±22.7	-20.9 ±28.8	-24.6 ±19.0	-20.0 ±14.3	-11.0 ±17.9	-17.9

D.3. Model Selection without Validation Attributes

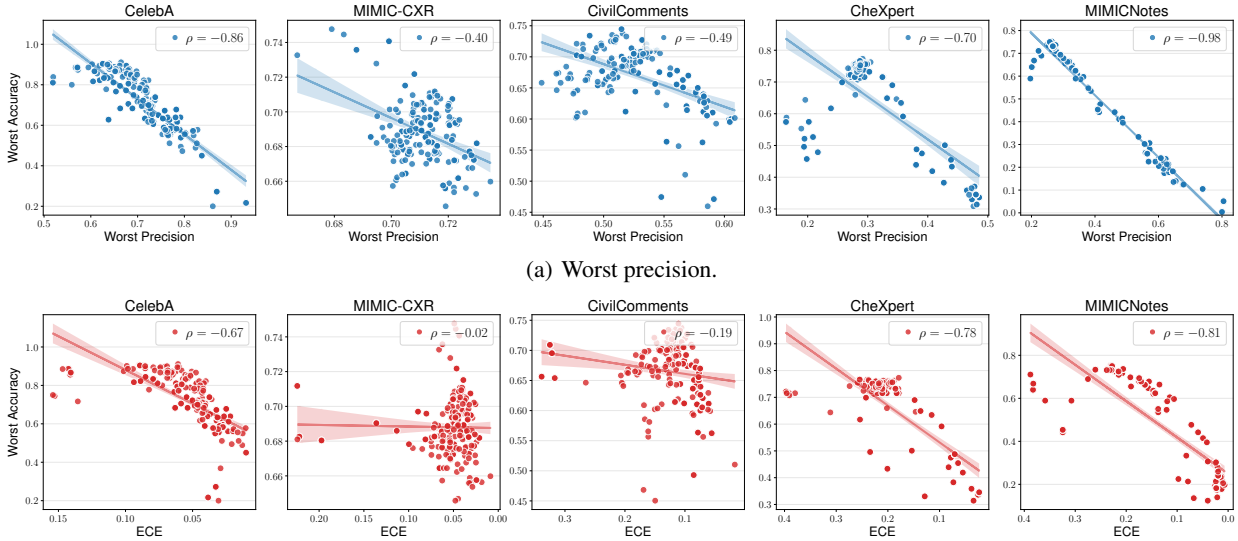
In the main paper, we examine the feasibility of different metrics for model selection without group-annotated validation data. We further confirm this in Table 12 by showing the results for more selection strategies with all metrics across all



(a) Adjusted accuracy.

(b) Balanced accuracy.

Figure 8. **Accuracy on the line**. We show metrics that are *positively* correlated with worst-group accuracy.



(a) Worst precision.

(b) ECE.

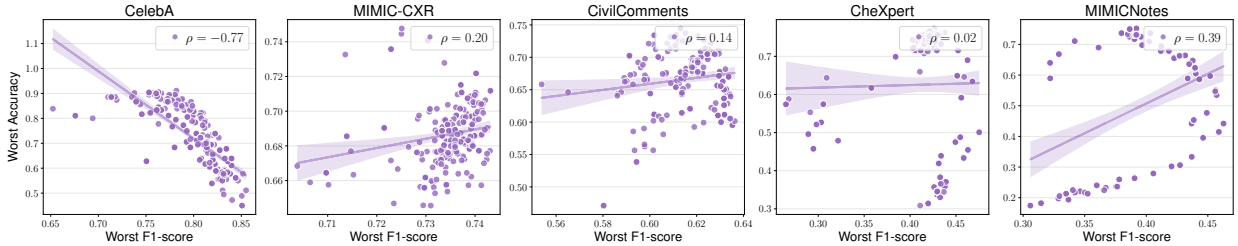
Figure 9. **Accuracy on the inverse line**. We show metrics that are *negatively* correlated with worst-group accuracy.

datasets in our benchmark. Specifically, when using *worst-class accuracy* as the model selection criterion, on average we achieve only **2.4%** degrade of worst-group accuracy compared to oracle selection method. The selection criterion also performs the best over all other selection metrics on 10 out of 12 datasets, indicating its effectiveness for reliable model selection without *any* attribute information.

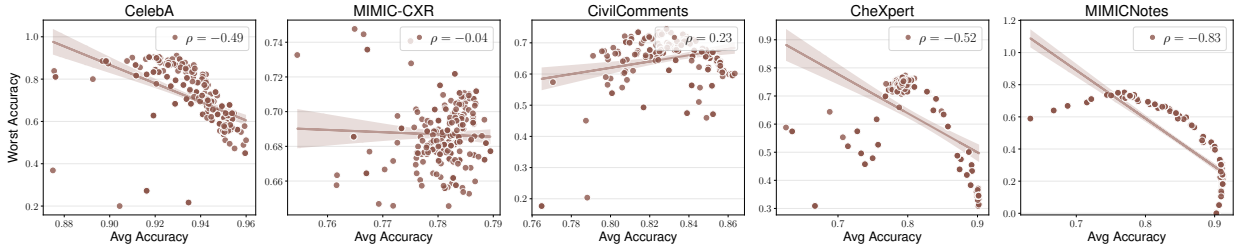
D.4. Rethinking Evaluation Metrics in Subpopulation Shift

We provide complete results on the correlation between worst-group accuracy (WGA) and other metrics we consider in our benchmark.

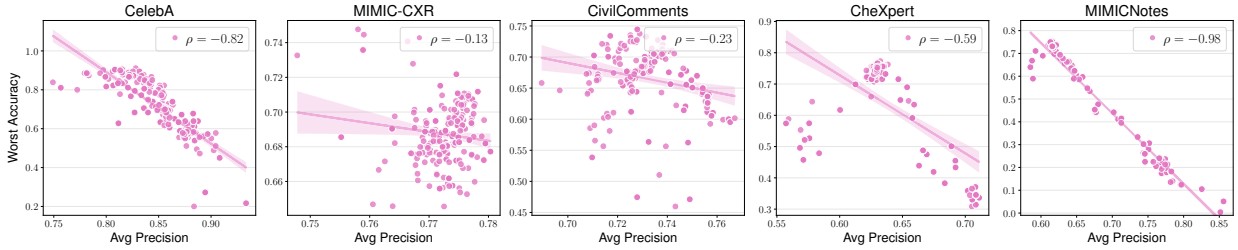
Accuracy on the line. In the main paper we show that certain metrics exhibit high linear correlation with WGA. We further show in Fig. 8 with a full list of metrics that exhibit consistent positive correlation across diverse datasets. Specifically, both adjusted accuracy and balanced accuracy display the "*accuracy on the line*" property, which has also been confirmed in prior work (Izmailov et al., 2022).



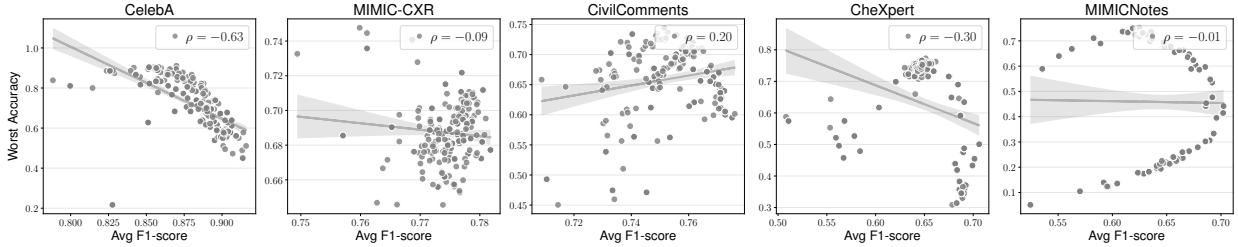
(a) Worst F1-score.



(b) Average accuracy.



(c) Average precision.



(d) Average F1-score.

Figure 10. **Accuracy not on the line**. We show metrics that do not demonstrate consistent correlations across datasets with worst-group accuracy.

Accuracy on the inverse line. More interestingly, we further establish the intrinsic tradeoff between WGA and certain metrics. Fig. 9 shows that both worst-case precision and ECE exhibit clear *negative* correlation with WGA, demonstrating the fundamental tradeoff between WGA and several important metrics in subpopulation shift. These intriguing observations highlight the need for considering more realistic evaluation metrics in subpopulation shift beyond just using WGA.

Accuracy not on the line. Finally, we display also other metrics that do not show either positive or negative correlation with WGA (Fig. 10). As observed, the correlation between these metrics and WGA shows inconsistent behavior across datasets. Interestingly, this phenomenon also indicates the potential bad performance on these metrics when merely optimizing for better WGA. We leave the exploration of other metrics and the rationale behind these behaviors for future work.

D.5. Impact of Architecture, Pretraining Method, and Pretraining Dataset

In this section, we examine the impact of model architecture and the source of the initial model weights on the worst group accuracy. Similar to the experiments above, we consider the following settings:

- **Known Attributes.** Attributes are known in both training and validation, and validation set worst-group accuracy is used as the model selection criteria.
- **Unknown Attributes.** Attributes are unknown during training and validation. Following our findings in Sec. 5.4, we use worst-class accuracy as the model selection criteria.

We experiment with ERM, JTT, and DFR as representative methods; `CivilComments` as the representative text dataset, and `Waterbirds`, `CheXpert`, and `NICO++` as representative image datasets.

For the text modality, we consider the following architectures and initial weights:

- **BERT**$_{\text{BASE}}$ (Devlin et al., 2018): A contextual language model based on the transformer architecture pretrained on BookCorpus and English Wikipedia data using the masked language model and next sentence prediction tasks.
- **SciBERT** (Beltagy et al., 2019): Same architecture as BERT$_{\text{BASE}}$, but pretrained on scientific papers from Semantic Scholar, and has higher reported performance on scientific NLP tasks.
- **DistilBERT** (Sanh et al., 2019): A knowledge distilled (Hinton et al., 2015) version of BERT$_{\text{BASE}}$ with 40% fewer parameters, pretrained using the same datasets as BERT$_{\text{BASE}}$.
- **GPT-2** (Radford et al., 2019): An autoregressive language model based on the transformer decoder, pretrained using text from webpages upvoted on Reddit.
- **RoBERTa**$_{\text{BASE}}$ (Liu et al., 2019a): Same architecture as BERT$_{\text{BASE}}$, but pretrained with a more efficient procedure and using a collection of corpora much larger than BERT$_{\text{BASE}}$.

For the image modality, we consider **ResNet-50** (He et al., 2016) and vision transformers (**ViT-B**) (Steiner et al., 2021). We consider model weights initialized with the following pretraining methods that span supervised and self-supervised manners:

- **Supervised** pretraining (Kornblith et al., 2019).
- **SimCLR** (Chen et al., 2020): Self-supervised contrastive pretraining using image augmentations.
- **Barlow** Twins (Zbontar et al., 2021): Self-supervised pretraining via redundancy reduction.
- **DINO** (Caron et al., 2021): Self-distillation with no labels.
- **CLIP** (Radford et al., 2021): Using associated text as supervision. We select only the vision encoder.

We consider model weights initialized using the above pretraining methods on the following pretraining datasets:

- **ImageNet-1K** (Deng et al., 2009): 1.2 million images belonging to 1,000 classes, introduced as part of the ILSVRC2012 visual recognition challenge (Russakovsky et al., 2015).
- **ImageNet-21K** (Ridnik et al., 2021): A superset of ImageNet-1K, consisting of 14 million images belonging to 21,841 classes.
- **SWAG** (Singh et al., 2022): 3.6 billion images collected from public Instagram posts, weakly supervised using their associated hashtags.
- **LAION-2B** (Schuhmann et al., 2022): 2.32 billion English image-text pairs constructed from Common Crawl.
- **OpenAI-CLIP** (Radford et al., 2021): 400 million image-text pairs collected by OpenAI in training their CLIP model.

As model weights for many combinations of the above architectures, pretraining methods, and pretraining datasets are not available, we only experiment with the subset of combinations of weights that exist in public repositories.

Based on our experimental results on `CivilComments` (Table 13), we find that BERT$_{\text{BASE}}$ is competitive in performance, even outperforming its successor RoBERTa$_{\text{BASE}}$ on many tasks. In addition, DistilBERT and GPT-2 exhibits much worse performance especially on ERM models.

Based on our experimental results on image datasets (Tables 14 and 15), we find the following:

- **Optimal architecture is dataset dependent.** Contrary to prior work (Paul & Chen, 2022), we find mixed results when comparing the worst-group performance for ResNet and ViT-B. Specifically, ResNets seem to work better on `CheXpert` and `Waterbirds`, while vision transformers work better on `NICO++`.
- **Supervised pretraining outperforms others.** Similar to prior work (Izmailov et al., 2022), we find that supervised pretraining outperforms self-supervised learning for the most part, though some self-supervised pretraining methods are still competitive. The results also warrant better self-supervised schemes for subgroup shifts (Yang et al., 2023).
- **Larger pretraining datasets yield better results.** The biggest impact on worst-group accuracy by far appears to be the dataset on which the initial model weights are derived. This is especially true for `NICO++` and `Waterbirds`, where going from ImageNet-1K to ImageNet-21K to SWAG almost always leads to a significant increase in worst-group accuracy, indicating that larger and more diverse pretraining datasets seem to increase performance. The effectiveness of SWAG-pretrained ViTs on `Waterbirds` has also been discussed in prior work (Mehta et al., 2022).

Table 13. Test-set worst-group accuracy on `CivilComments` for different text architectures and pretraining methods.

Arch	Unknown Attributes			Known Attributes		
	ERM	JTT	DFR	ERM	JTT	DFR
BERT	**65.6**	**69.6**	62.4	**66.2**	65.0	**69.7**
SciBERT	61.1	58.3	**62.5**	61.1	58.3	68.0
DistilBERT	51.8	55.1	61.8	59.6	66.2	67.6
GPT-2	14.7	49.0	51.7	14.7	49.0	51.9
RoBERTa	61.0	58.0	61.6	63.1	**66.7**	68.2

Table 14. Test-set worst-group accuracy for three image datasets with *known attributes*, varying the model architecture and source of model initial weights. Best results of each column are in **bold** and the second best are underlined.

Arch	Pretrain Method	Pretrain Dataset	CheXpert			NICO++			Waterbirds			Avg
			ERM	JTT	DFR	ERM	JTT	DFR	ERM	JTT	DFR	
ResNet	Barlow	ImageNet-1K	46.2	66.0	74.7	40.0	40.0	20.0	67.3	72.4	88.3	57.2
	DINO	ImageNet-1K	43.0	71.5	72.8	39.5	40.0	4.0	72.9	72.5	89.1	56.1
	SimCLR	ImageNet-1K	47.9	**72.3**	**74.8**	30.0	30.0	16.0	70.1	68.1	81.2	54.5
	Supervised	ImageNet-1K	**59.2**	61.7	72.2	25.0	30.0	20.0	76.5	74.3	**90.2**	56.6
	Supervised	ImageNet-21K	51.4	68.0	70.0	40.0	46.0	**40.0**	74.5	75.9	**90.2**	61.8
ViT-B	CLIP	Laion-2B	49.2	58.5	69.1	33.3	40.0	33.3	39.6	46.9	75.5	49.5
	CLIP	OpenAI-CLIP	42.2	55.8	68.8	33.3	40.0	30.0	40.4	40.4	78.2	47.7
	DINO	ImageNet-1K	43.4	71.8	72.4	30.0	40.0	32.0	63.9	64.6	**90.2**	56.5
	Supervised	ImageNet-1K	40.4	64.5	70.1	20.0	33.3	0.0	51.2	52.6	80.4	45.8
	Supervised	ImageNet-21K	47.5	69.1	69.1	48.0	**50.0**	18.0	69.9	73.8	87.2	59.2
	Supervised	SWAG	48.7	67.3	72.5	**50.0**	**50.0**	34.0	**82.7**	**81.2**	87.5	**63.8**

Table 15. Test-set worst-group accuracy for three image datasets with *unknown attributes*, varying the model architecture and source of model initial weights. Best results of each column are in **bold** and the second best are underlined.

Arch	Pretrain Method	Pretrain Dataset	CheXpert			NICO++			Waterbirds			Avg
			ERM	JTT	DFR	ERM	JTT	DFR	ERM	JTT	DFR	
ResNet	Barlow	ImageNet-1K	46.2	66.0	73.7	33.3	40.0	40.0	67.3	72.4	89.8	58.7
	DINO	ImageNet-1K	43.0	71.5	73.3	39.5	40.0	12.0	72.9	72.5	87.9	57.0
	SimCLR	ImageNet-1K	47.9	**72.3**	74.6	30.0	30.0	26.0	70.1	69.0	79.2	55.5
	Supervised	ImageNet-1K	**59.2**	61.7	**75.4**	40.0	30.0	33.3	67.0	74.3	89.6	58.9
	Supervised	ImageNet-21K	45.3	69.3	69.9	40.0	40.0	40.0	74.5	75.9	88.3	60.4
ViT-B	CLIP	Laion-2B	49.2	58.5	69.7	30.0	30.0	40.0	45.2	46.9	78.4	49.8
	CLIP	OpenAI-CLIP	42.2	57.4	70.4	33.3	40.0	40.0	26.5	44.4	77.4	48.0
	DINO	ImageNet-1K	43.4	69.4	72.3	40.0	41.2	37.5	63.9	64.6	**90.0**	58.0
	Supervised	ImageNet-1K	40.4	69.5	71.5	33.3	33.3	16.7	49.4	52.6	81.2	49.8
	Supervised	ImageNet-21K	47.5	69.7	71.3	**50.0**	**50.0**	38.0	69.9	73.8	88.9	62.1
	Supervised	SWAG	52.5	63.8	71.3	**50.0**	**50.0**	**50.0**	**82.7**	**81.2**	88.6	**65.6**

E. Complete Results

We provide complete evaluation results in this section. As confirmed earlier, model selection and attribute availability play critical roles in subpopulation shift evaluation. To provide a thorough analysis, we investigate the following three settings:

- **Attributes are known in both training & validation (Appendix E.1).** When attributes are known in both training and validation set, which corresponds to the most ideal scenario, we use "*test set worst-group accuracy*" as an oracle selection method to identify the best possible performance for each algorithm.
- **Attributes are unknown in training, but known in validation (Appendix E.2).** When attributes are still known in validation, we use "*validation set worst-group accuracy*" to select models. We ignore algorithms that require attribute information in the training set (i.e., IRM, MMD, CORAL) when reporting results under this setting.
- **Attributes are unknown in both training & validation (Appendix E.3).** When attributes are completely unknown, we still use "*validation set worst-group accuracy*" for model selection, which however degenerates to "*worst-class accuracy*". We again ignore algorithms that require attribute information in the training set.

E.1. Attributes Known in Both Training & Validation

E.1.1. Waterbirds

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	84.1 ±1.7	69.1 ±4.7	77.4 ±2.0	60.7 ±3.2	79.4 ±2.1	69.5 ±2.9	83.1 ±2.0	83.1 ±2.0	91.0 ±1.4	12.9 ±1.7
Mixup	89.5 ±0.4	78.2 ±0.4	83.9 ±0.6	71.6 ±1.3	85.9 ±0.4	78.8 ±0.6	88.9 ±0.3	88.9 ±0.3	94.7 ±0.2	7.0 ±0.6
GroupDRO	88.8 ±1.8	78.6 ±1.0	83.6 ±2.4	70.9 ±5.0	85.3 ±2.0	78.1 ±2.6	88.5 ±0.8	88.5 ±0.8	95.5 ±0.5	9.1 ±2.2
IRM	88.4 ±0.1	74.5 ±1.5	82.5 ±0.2	69.5 ±0.6	84.3 ±0.1	76.4 ±0.1	87.1 ±0.3	87.1 ±0.3	94.0 ±0.3	9.5 ±0.2
CVaRDRO	89.8 ±0.4	75.5 ±2.2	84.5 ±0.7	73.2 ±1.7	86.1 ±0.3	79.0 ±0.4	88.5 ±0.3	88.5 ±0.3	95.4 ±0.2	8.2 ±0.2
JTT	88.8 ±0.6	72.0 ±0.3	83.1 ±0.8	71.2 ±1.5	84.7 ±0.6	76.9 ±0.8	86.9 ±0.3	86.9 ±0.3	94.1 ±0.1	9.0 ±0.4
LfF	87.0 ±0.3	75.2 ±0.7	80.7 ±0.3	66.2 ±0.5	82.8 ±0.3	74.3 ±0.5	86.2 ±0.3	86.2 ±0.3	93.3 ±0.3	9.4 ±0.5
LISA	92.8 ±0.2	88.7 ±0.6	88.4 ±0.4	79.5 ±0.8	90.0 ±0.3	84.8 ±0.4	92.0 ±0.1	92.0 ±0.1	97.0 ±0.1	5.4 ±0.3
MMD	93.0 ±0.1	83.9 ±1.4	89.5 ±0.4	83.1 ±1.1	90.0 ±0.1	84.5 ±0.1	90.5 ±0.2	90.5 ±0.2	96.2 ±0.1	6.4 ±0.5
ReSample	89.4 ±0.9	77.7 ±1.2	84.0 ±1.4	72.1 ±3.1	85.7 ±1.0	78.4 ±1.4	88.3 ±0.4	88.3 ±0.4	95.2 ±0.3	8.0 ±1.1
ReWeight	91.8 ±0.2	86.9 ±0.7	87.1 ±0.3	77.5 ±0.8	88.7 ±0.2	82.7 ±0.3	90.7 ±0.1	90.7 ±0.1	95.8 ±0.1	7.0 ±0.6
SqrtReWeight	88.7 ±0.3	78.6 ±0.1	82.8 ±0.4	69.6 ±1.1	84.9 ±0.3	77.3 ±0.4	88.1 ±0.2	88.1 ±0.2	94.5 ±0.1	8.2 ±0.7
CBLoss	91.3 ±0.7	86.2 ±0.3	86.5 ±1.1	76.4 ±2.3	88.2 ±0.7	82.0 ±1.0	90.4 ±0.1	90.4 ±0.1	95.7 ±0.0	8.2 ±1.4
Focal	89.3 ±0.2	71.6 ±0.8	83.7 ±0.3	72.4 ±0.5	85.2 ±0.3	77.5 ±0.4	87.1 ±0.3	87.1 ±0.3	94.2 ±0.2	6.9 ±0.1
LDAM	87.3 ±0.5	71.0 ±1.8	81.2 ±0.6	67.7 ±1.5	83.0 ±0.5	74.4 ±0.6	85.7 ±0.2	85.7 ±0.2	93.3 ±0.2	13.7 ±2.2
BSoftmax	88.4 ±1.3	74.1 ±0.9	82.7 ±1.6	70.1 ±3.0	84.4 ±1.5	76.5 ±2.1	87.0 ±1.0	87.0 ±1.0	94.1 ±1.0	9.8 ±1.2
DFR	92.3 ±0.2	91.0 ±0.3	87.5 ±0.3	77.5 ±0.6	89.5 ±0.2	84.1 ±0.3	92.1 ±0.1	92.1 ±0.1	97.4 ±0.1	7.1 ±0.6
CRT	90.5 ±0.0	79.7 ±0.3	85.3 ±0.0	74.5 ±0.0	87.0 ±0.1	80.3 ±0.1	89.3 ±0.1	89.3 ±0.1	95.7 ±0.0	7.9 ±0.1
ReWeightCRT	91.2 ±0.1	78.4 ±0.1	86.4 ±0.2	76.8 ±0.3	87.7 ±0.1	81.2 ±0.2	89.4 ±0.1	89.4 ±0.1	95.8 ±0.1	6.3 ±0.2

E.1.2. CelebA

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	95.1 ±0.2	62.6 ±1.5	87.5 ±0.5	76.4 ±1.2	90.1 ±0.3	83.1 ±0.5	86.9 ±0.2	93.4 ±0.2	98.4 ±0.1	3.3 ±0.5
Mixup	95.4 ±0.1	57.8 ±0.8	88.4 ±0.3	78.5 ±0.7	90.6 ±0.2	83.8 ±0.3	85.8 ±0.2	93.1 ±0.1	98.4 ±0.1	2.5 ±0.2
GroupDRO	91.4 ±0.6	89.0 ±0.7	80.4 ±0.8	61.5 ±1.7	84.9 ±0.8	74.9 ±1.2	92.6 ±0.1	93.3 ±0.2	98.1 ±0.0	8.0 ±0.9
IRM	94.7 ±0.8	63.0 ±2.5	87.0 ±1.9	75.3 ±3.9	89.6 ±1.1	82.2 ±1.8	86.9 ±0.5	93.3 ±0.3	98.5 ±0.0	3.4 ±1.3
CVaRDRO	95.2 ±0.1	64.1 ±2.8	88.4 ±0.6	78.6 ±1.4	90.1 ±0.1	83.0 ±0.2	86.7 ±0.9	92.2 ±0.7	98.2 ±0.1	2.6 ±0.3
JTT	90.4 ±2.3	70.0 ±10.2	80.5 ±4.2	62.5 ±8.7	83.4 ±3.3	72.6 ±5.1	86.4 ±1.6	90.3 ±1.1	93.2 ±2.2	4.1 ±1.4
LfF	81.1 ±5.6	53.0 ±4.3	71.8 ±4.1	45.2 ±8.3	73.2 ±5.6	59.0 ±7.3	78.3 ±3.0	85.3 ±2.9	94.1 ±1.2	27.9 ±5.5
LISA	92.6 ±0.1	86.5 ±1.2	82.2 ±0.2	65.1 ±0.4	86.6 ±0.2	77.6 ±0.3	92.0 ±0.3	94.0 ±0.1	98.5 ±0.0	7.7 ±0.3
MMD	92.5 ±0.7	24.4 ±2.0	91.4 ±1.4	90.1 ±2.2	79.8 ±2.1	63.7 ±3.9	68.5 ±0.7	74.3 ±2.1	96.0 ±0.9	3.6 ±0.2
ReSample	92.0 ±0.8	87.4 ±0.8	81.4 ±1.2	63.6 ±2.6	85.6 ±1.0	76.0 ±1.5	92.0 ±0.2	93.1 ±0.1	98.1 ±0.0	7.4 ±1.1
ReWeight	91.9 ±0.5	89.7 ±0.2	81.2 ±0.8	63.2 ±1.8	85.4 ±0.7	75.7 ±1.0	92.6 ±0.2	93.0 ±0.2	98.0 ±0.1	7.9 ±0.9
SqrtReWeight	93.6 ±0.1	82.4 ±0.5	84.0 ±0.2	69.0 ±0.3	87.9 ±0.2	79.6 ±0.3	91.2 ±0.1	93.8 ±0.1	98.4 ±0.1	5.8 ±0.2
CBLoss	91.2 ±0.7	89.4 ±0.7	80.2 ±1.1	61.0 ±2.3	84.6 ±1.0	74.5 ±1.6	92.6 ±0.2	93.2 ±0.3	98.0 ±0.1	8.4 ±1.0
Focal	94.9 ±0.3	59.1 ±2.0	87.5 ±0.8	76.7 ±1.7	89.7 ±0.4	82.4 ±0.6	85.6 ±0.5	92.5 ±0.4	98.2 ±0.1	3.2 ±0.4
LDAM	94.5 ±0.2	59.6 ±2.4	86.5 ±0.8	74.7 ±1.9	89.0 ±0.2	81.3 ±0.3	85.6 ±0.8	92.3 ±0.7	98.0 ±0.1	28.3 ±2.7
BSoftmax	91.9 ±0.1	83.3 ±0.5	81.1 ±0.2	62.9 ±0.4	85.6 ±0.2	76.1 ±0.3	91.1 ±0.2	93.9 ±0.1	98.6 ±0.0	8.4 ±0.2
DFR	91.9 ±0.1	90.4 ±0.1	81.2 ±0.2	63.2 ±0.3	85.5 ±0.1	75.8 ±0.2	92.3 ±0.0	93.1 ±0.1	97.9 ±0.0	8.9 ±0.1
CRT	92.7 ±0.1	87.2 ±0.3	82.4 ±0.1	65.7 ±0.2	86.5 ±0.1	77.4 ±0.1	91.8 ±0.1	93.4 ±0.0	98.2 ±0.0	6.6 ±0.1
ReWeightCRT	92.5 ±0.2	87.2 ±0.3	82.1 ±0.3	65.1 ±0.6	86.3 ±0.2	77.1 ±0.3	91.8 ±0.0	93.4 ±0.0	98.2 ±0.0	7.1 ±0.3

E.1.3. CIVILCOMMENTS

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	85.4 ±0.2	63.7 ±1.1	75.4 ±0.2	57.8 ±0.6	77.0 ±0.0	63.1 ±0.1	77.7 ±0.2	79.2 ±0.3	90.0 ±0.0	8.1 ±0.2
Mixup	84.9 ±0.3	66.1 ±1.3	74.8 ±0.4	56.4 ±0.9	76.6 ±0.2	62.7 ±0.2	77.9 ±0.3	79.3 ±0.3	89.7 ±0.0	8.4 ±1.0
GroupDRO	81.8 ±0.6	70.6 ±1.2	72.0 ±0.5	49.6 ±1.1	74.2 ±0.5	60.3 ±0.6	78.5 ±0.2	79.9 ±0.2	88.8 ±0.2	12.2 ±0.9
IRM	85.5 ±0.0	63.2 ±0.8	75.5 ±0.1	57.8 ±0.2	77.1 ±0.0	63.3 ±0.1	77.8 ±0.1	79.4 ±0.1	89.9 ±0.1	7.4 ±0.6
CVaRDRO	83.5 ±0.3	68.7 ±1.3	73.5 ±0.3	52.8 ±0.6	75.9 ±0.2	62.4 ±0.2	78.6 ±0.2	80.7 ±0.1	89.8 ±0.1	32.9 ±0.4
JTT	83.3 ±0.1	64.3 ±1.5	72.8 ±0.1	52.4 ±0.3	74.8 ±0.1	60.3 ±0.2	76.8 ±0.2	78.4 ±0.2	88.2 ±0.1	10.2 ±0.3
LfF	65.5 ±5.6	51.0 ±6.1	60.4 ±3.5	31.2 ±5.6	58.5 ±5.0	41.9 ±5.6	64.8 ±4.2	65.6 ±4.5	69.2 ±6.5	26.4 ±2.4
LISA	82.7 ±0.1	73.7 ±0.3	72.6 ±0.1	51.1 ±0.2	75.0 ±0.1	61.1 ±0.1	78.7 ±0.2	80.1 ±0.1	89.1 ±0.1	11.7 ±0.3
MMD	84.6 ±0.2	54.5 ±1.4	73.9 ±0.4	56.7 ±0.7	74.4 ±0.4	58.2 ±0.7	73.6 ±0.6	74.9 ±0.5	86.1 ±0.7	5.0 ±1.5
ReSample	82.2 ±0.0	73.3 ±0.5	72.4 ±0.0	50.2 ±0.1	74.8 ±0.0	61.1 ±0.0	79.2 ±0.0	80.6 ±0.0	89.3 ±0.1	12.2 ±0.2
ReWeight	82.5 ±0.0	72.5 ±0.0	72.6 ±0.1	50.8 ±0.1	75.0 ±0.1	61.4 ±0.1	79.1 ±0.1	80.6 ±0.1	89.5 ±0.0	12.0 ±0.2
SqrtReWeight	83.3 ±0.5	71.7 ±0.4	73.3 ±0.4	52.5 ±1.0	75.7 ±0.4	62.0 ±0.4	78.9 ±0.1	80.4 ±0.1	89.7 ±0.0	10.3 ±0.8
CBLoss	82.9 ±0.1	73.3 ±0.2	72.9 ±0.1	51.5 ±0.2	75.4 ±0.1	61.7 ±0.1	79.2 ±0.1	80.6 ±0.1	89.6 ±0.1	11.1 ±0.3
Focal	85.5 ±0.2	62.0 ±1.0	75.5 ±0.4	58.5 ±0.8	76.8 ±0.3	62.5 ±0.4	76.9 ±0.4	78.4 ±0.4	89.1 ±0.3	6.7 ±0.4
LDAM	81.9 ±2.2	37.4 ±8.1	69.6 ±3.5	49.9 ±5.9	69.7 ±3.4	50.6 ±5.5	67.5 ±4.0	70.0 ±3.3	79.7 ±4.2	21.1 ±0.3
BSoftmax	83.8 ±0.0	71.2 ±0.4	73.8 ±0.0	53.5 ±0.0	76.1 ±0.0	62.5 ±0.0	78.7 ±0.1	80.4 ±0.0	89.8 ±0.0	10.3 ±0.1
DFR	83.3 ±0.0	69.6 ±0.2	73.2 ±0.0	52.3 ±0.1	75.6 ±0.0	61.8 ±0.0	78.1 ±0.0	80.2 ±0.0	89.5 ±0.0	16.6 ±0.3
CRT	83.8 ±0.0	71.1 ±0.1	73.8 ±0.0	53.5 ±0.0	76.1 ±0.0	62.5 ±0.0	78.6 ±0.0	80.4 ±0.0	89.4 ±0.0	11.2 ±0.3
ReWeightCRT	83.8 ±0.1	71.0 ±0.1	73.8 ±0.1	53.5 ±0.2	76.1 ±0.0	62.4 ±0.0	78.5 ±0.0	80.4 ±0.1	89.6 ±0.0	10.7 ±0.1

E.1.4. MULTINLI

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	80.9 ±0.1	66.8 ±0.5	81.1 ±0.1	76.0 ±0.2	80.9 ±0.1	77.8 ±0.1	79.7 ±0.0	80.9 ±0.1	93.6 ±0.1	8.1 ±0.3
Mixup	81.4 ±0.3	68.5 ±0.6	81.6 ±0.3	76.0 ±0.5	81.4 ±0.3	78.0 ±0.2	80.1 ±0.3	81.4 ±0.3	93.6 ±0.1	9.4 ±0.9
GroupDRO	81.1 ±0.3	76.0 ±0.7	81.4 ±0.3	74.7 ±0.5	81.1 ±0.3	77.8 ±0.1	80.8 ±0.3	81.1 ±0.3	93.7 ±0.1	9.8 ±1.0
IRM	77.8 ±0.6	63.6 ±1.3	78.3 ±0.5	71.0 ±1.1	77.9 ±0.6	74.8 ±0.4	76.6 ±0.5	77.8 ±0.6	91.5 ±0.3	11.2 ±1.8
CVaRDRO	75.1 ±0.1	63.0 ±1.5	76.2 ±0.2	65.6 ±0.2	75.2 ±0.1	72.1 ±0.2	74.2 ±0.4	75.1 ±0.1	86.3 ±0.2	41.4 ±0.1
JTT	80.9 ±0.5	69.1 ±0.1	81.3 ±0.4	74.3 ±1.1	81.0 ±0.5	77.6 ±0.5	80.0 ±0.4	80.9 ±0.5	93.7 ±0.2	7.0 ±1.5
LfF	71.7 ±1.1	63.6 ±2.9	71.8 ±1.1	68.7 ±0.7	71.7 ±1.1	68.5 ±1.8	70.8 ±1.4	71.7 ±1.1	87.0 ±0.8	4.4 ±0.6
LISA	80.3 ±0.4	73.3 ±1.0	80.4 ±0.4	75.9 ±0.3	80.3 ±0.4	76.7 ±0.4	79.8 ±0.5	80.3 ±0.4	92.7 ±0.2	4.3 ±0.4
MMD	78.8 ±0.1	69.1 ±1.5	79.3 ±0.2	71.7 ±0.7	78.9 ±0.1	75.5 ±0.1	78.0 ±0.4	78.8 ±0.1	91.7 ±0.1	11.6 ±0.3
ReSample	77.2 ±0.2	72.3 ±0.8	77.6 ±0.0	70.7 ±1.0	77.3 ±0.1	73.8 ±0.1	77.6 ±0.3	77.2 ±0.2	90.9 ±0.0	10.8 ±0.2
ReWeight	81.0 ±0.2	68.8 ±0.4	81.1 ±0.2	76.0 ±0.7	81.0 ±0.2	77.4 ±0.1	79.6 ±0.1	81.0 ±0.2	93.5 ±0.1	8.1 ±0.1
SqrtReWeight	80.7 ±0.3	69.5 ±0.7	81.0 ±0.3	74.6 ±0.5	80.8 ±0.3	77.5 ±0.3	79.9 ±0.4	80.7 ±0.3	93.4 ±0.2	9.2 ±1.0
CBLoss	80.6 ±0.1	72.2 ±0.3	80.8 ±0.1	74.9 ±0.3	80.6 ±0.1	77.5 ±0.1	80.1 ±0.1	80.6 ±0.1	93.4 ±0.1	7.5 ±0.5
Focal	80.7 ±0.2	69.4 ±0.7	81.2 ±0.2	73.7 ±0.6	80.8 ±0.2	77.3 ±0.1	79.6 ±0.2	80.7 ±0.2	93.6 ±0.1	4.4 ±1.0
LDAM	80.7 ±0.3	69.6 ±1.6	81.1 ±0.1	73.9 ±0.9	80.8 ±0.2	77.4 ±0.2	79.7 ±0.3	80.7 ±0.3	93.5 ±0.1	33.4 ±0.3
BSoftmax	80.9 ±0.1	66.9 ±0.4	81.1 ±0.1	75.9 ±0.3	80.9 ±0.1	77.7 ±0.1	79.7 ±0.0	80.9 ±0.1	93.6 ±0.1	8.1 ±0.2
DFR	81.7 ±0.0	68.5 ±0.2	82.1 ±0.0	75.6 ±0.2	81.7 ±0.0	77.9 ±0.0	81.2 ±0.0	81.7 ±0.0	93.2 ±0.0	8.8 ±0.3
CRT	81.9 ±0.0	70.7 ±0.1	82.2 ±0.0	75.9 ±0.1	82.0 ±0.0	78.3 ±0.0	81.1 ±0.0	81.9 ±0.0	93.9 ±0.0	11.5 ±0.0
ReWeightCRT	81.3 ±0.0	69.0 ±0.2	81.4 ±0.0	77.0 ±0.1	81.3 ±0.0	77.6 ±0.0	80.5 ±0.0	81.3 ±0.0	93.7 ±0.0	6.9 ±0.1

E.1.5. METASHIFT

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	91.3 ±0.3	82.6 ±0.4	91.2 ±0.3	90.6 ±0.4	91.2 ±0.3	90.7 ±0.3	89.3 ±0.3	91.2 ±0.3	97.3 ±0.2	6.3 ±0.9
Mixup	91.6 ±0.3	81.0 ±0.8	91.7 ±0.3	90.6 ±0.2	91.6 ±0.3	91.0 ±0.3	89.4 ±0.3	91.6 ±0.3	97.3 ±0.1	2.3 ±0.1
GroupDRO	91.0 ±0.1	85.6 ±0.4	90.9 ±0.1	90.0 ±0.6	90.9 ±0.1	90.4 ±0.0	89.8 ±0.1	91.0 ±0.1	97.5 ±0.0	3.2 ±0.5
IRM	91.8 ±0.4	83.0 ±0.1	91.8 ±0.4	90.5 ±1.0	91.7 ±0.4	91.3 ±0.4	89.7 ±0.5	91.7 ±0.4	97.6 ±0.2	5.3 ±0.2
CVaRDRO	92.1 ±0.2	84.6 ±0.0	92.1 ±0.2	90.8 ±0.6	92.1 ±0.2	91.6 ±0.2	90.4 ±0.2	92.1 ±0.2	97.7 ±0.0	4.9 ±0.3
JTT	91.2 ±0.5	83.6 ±0.4	91.3 ±0.6	89.3 ±1.1	91.1 ±0.5	90.6 ±0.4	89.6 ±0.8	91.1 ±0.5	97.4 ±0.0	5.9 ±0.7
LfF	80.2 ±0.3	73.1 ±1.6	80.5 ±0.3	77.2 ±1.3	80.1 ±0.3	78.8 ±0.3	80.3 ±0.6	80.1 ±0.1	90.6 ±0.6	8.3 ±1.5
LISA	89.5 ±0.4	84.1 ±0.4	89.6 ±0.4	88.4 ±0.3	89.5 ±0.5	88.8 ±0.6	88.5 ±0.3	89.5 ±0.5	96.0 ±0.1	25.4 ±0.2
MMD	89.4 ±0.1	85.9 ±0.7	89.5 ±0.2	88.3 ±0.2	89.3 ±0.1	88.4 ±0.1	89.4 ±0.0	89.2 ±0.1	95.4 ±0.3	3.2 ±0.3
ReSample	91.2 ±0.1	85.6 ±0.4	91.1 ±0.1	90.8 ±0.1	91.1 ±0.1	90.5 ±0.1	90.0 ±0.2	91.1 ±0.1	97.4 ±0.1	5.2 ±0.2
ReWeight	91.7 ±0.4	85.6 ±0.4	91.8 ±0.4	90.2 ±0.6	91.7 ±0.3	91.1 ±0.3	90.6 ±0.5	91.6 ±0.3	97.5 ±0.1	4.2 ±0.2
SqrtReWeight	91.5 ±0.2	84.6 ±0.7	91.5 ±0.2	89.7 ±0.2	91.5 ±0.2	91.1 ±0.2	89.7 ±0.3	91.6 ±0.2	97.7 ±0.0	3.6 ±0.6
CBLoss	91.7 ±0.4	85.5 ±0.4	91.8 ±0.4	90.2 ±0.7	91.6 ±0.3	91.1 ±0.3	90.6 ±0.4	91.6 ±0.3	97.5 ±0.1	4.1 ±0.2
Focal	91.7 ±0.2	81.5 ±0.0	91.7 ±0.2	91.1 ±0.6	91.7 ±0.2	91.2 ±0.2	89.5 ±0.2	91.7 ±0.2	97.7 ±0.0	5.2 ±1.6
LDAM	91.5 ±0.1	83.6 ±0.4	91.5 ±0.1	90.7 ±0.3	91.5 ±0.1	90.9 ±0.1	89.8 ±0.1	91.5 ±0.1	97.5 ±0.1	10.8 ±0.6
BSoftmax	91.6 ±0.2	83.1 ±0.7	91.6 ±0.2	89.8 ±0.3	91.6 ±0.2	91.2 ±0.2	89.4 ±0.3	91.7 ±0.1	97.7 ±0.0	4.0 ±0.6
DFR	88.4 ±0.3	85.4 ±0.4	88.4 ±0.3	86.8 ±0.3	88.4 ±0.3	87.8 ±0.4	87.7 ±0.3	88.5 ±0.3	95.6 ±0.1	5.7 ±0.2
CRT	91.3 ±0.2	84.1 ±0.4	91.3 ±0.2	90.2 ±0.2	91.3 ±0.2	90.8 ±0.2	89.6 ±0.2	91.3 ±0.2	97.3 ±0.0	7.4 ±0.1
ReWeightCRT	91.2 ±0.1	85.6 ±0.4	91.1 ±0.1	90.1 ±0.1	91.2 ±0.1	90.7 ±0.0	89.8 ±0.1	91.2 ±0.0	96.8 ±0.1	7.8 ±0.1

E.1.6. IMAGENETBG

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	88.4 ±0.1	81.0 ±0.9	88.5 ±0.1	80.4 ±0.1	88.4 ±0.1	81.3 ±0.1	88.4 ±0.1	88.4 ±0.1	99.0 ±0.0	5.8 ±0.2
Mixup	88.5 ±0.3	82.4 ±0.3	88.7 ±0.2	80.4 ±1.3	88.5 ±0.3	81.9 ±0.8	88.5 ±0.3	88.5 ±0.3	98.8 ±0.1	3.3 ±1.0
GroupDRO	87.3 ±0.2	78.2 ±0.3	87.5 ±0.2	77.9 ±1.1	87.3 ±0.2	80.0 ±0.9	87.3 ±0.2	87.3 ±0.2	98.9 ±0.0	4.3 ±0.5
IRM	88.7 ±0.1	81.3 ±0.3	88.8 ±0.1	81.6 ±0.2	88.7 ±0.1	81.7 ±0.1	88.7 ±0.1	88.7 ±0.1	99.1 ±0.0	5.2 ±0.1
CVaRDRO	88.2 ±0.1	80.7 ±1.1	88.4 ±0.1	78.6 ±1.9	88.3 ±0.1	80.7 ±0.5	88.2 ±0.1	88.2 ±0.1	99.0 ±0.0	4.9 ±0.4
JTT	87.2 ±0.1	80.5 ±0.3	87.5 ±0.2	78.0 ±0.7	87.2 ±0.1	80.2 ±0.6	87.2 ±0.1	87.2 ±0.1	98.9 ±0.0	2.4 ±0.5
LfF	85.3 ±0.3	76.7 ±0.5	85.6 ±0.3	74.0 ±2.2	85.3 ±0.3	75.8 ±1.3	85.3 ±0.3	85.3 ±0.3	98.5 ±0.0	2.6 ±0.4
LISA	86.2 ±0.3	76.1 ±0.8	86.3 ±0.3	75.5 ±1.0	86.2 ±0.3	77.1 ±0.5	86.2 ±0.3	86.2 ±0.3	98.3 ±0.1	4.2 ±0.2
MMD	88.2 ±0.2	80.8 ±0.5	88.4 ±0.2	80.0 ±1.1	88.2 ±0.2	80.7 ±0.3	88.2 ±0.2	88.2 ±0.2	99.0 ±0.0	5.8 ±0.2
ReSample	88.5 ±0.2	81.0 ±0.4	88.7 ±0.2	79.9 ±1.0	88.5 ±0.2	81.5 ±0.4	88.5 ±0.2	88.5 ±0.2	99.0 ±0.0	6.0 ±0.2
ReWeight	88.4 ±0.1	81.0 ±0.9	88.5 ±0.1	80.4 ±0.1	88.4 ±0.1	81.3 ±0.1	88.4 ±0.1	88.4 ±0.1	99.0 ±0.0	5.8 ±0.2
SqrtReWeight	88.3 ±0.1	80.1 ±0.2	88.4 ±0.1	80.5 ±0.5	88.3 ±0.1	80.9 ±0.4	88.3 ±0.1	88.3 ±0.1	99.0 ±0.0	5.3 ±0.3
CBLoss	88.4 ±0.1	81.0 ±0.9	88.5 ±0.1	80.4 ±0.1	88.4 ±0.1	81.3 ±0.1	88.4 ±0.1	88.4 ±0.1	99.0 ±0.0	5.8 ±0.2
Focal	87.2 ±0.1	78.4 ±0.1	87.3 ±0.2	78.7 ±0.7	87.2 ±0.1	78.9 ±0.5	87.2 ±0.1	87.2 ±0.1	98.8 ±0.0	4.4 ±1.1
LDAM	88.0 ±0.1	80.1 ±0.3	88.3 ±0.0	80.1 ±0.6	88.1 ±0.1	81.4 ±0.3	88.0 ±0.1	88.0 ±0.1	98.7 ±0.1	48.3 ±1.9
BSoftmax	88.3 ±0.1	80.7 ±0.7	88.4 ±0.1	79.4 ±0.9	88.3 ±0.1	80.8 ±0.4	88.3 ±0.1	88.3 ±0.1	99.0 ±0.0	6.0 ±0.2
DFR	87.2 ±0.2	78.5 ±0.6	87.2 ±0.3	78.2 ±1.2	87.2 ±0.2	78.8 ±0.9	87.2 ±0.2	87.2 ±0.2	98.8 ±0.0	9.9 ±1.3
CRT	88.4 ±0.1	80.2 ±0.3	88.4 ±0.1	80.4 ±0.8	88.3 ±0.1	80.7 ±0.3	88.4 ±0.1	88.4 ±0.1	99.0 ±0.0	4.5 ±0.5
ReWeightCRT	88.6 ±0.0	79.4 ±0.2	88.7 ±0.0	81.6 ±0.7	88.6 ±0.0	81.5 ±0.2	88.6 ±0.0	88.6 ±0.0	99.1 ±0.0	4.5 ±0.8

E.1.7. NICO++

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	84.5 ±0.5	37.6 ±2.0	85.5 ±0.3	54.5 ±2.8	84.6 ±0.4	65.8 ±1.1	84.0 ±0.5	84.3 ±0.5	99.3 ±0.0	10.4 ±0.1
Mixup	84.0 ±0.6	42.7 ±1.4	85.2 ±0.5	53.0 ±1.6	84.2 ±0.6	63.4 ±1.1	83.7 ±0.6	83.9 ±0.6	99.3 ±0.0	2.5 ±1.0
GroupDRO	83.2 ±0.4	37.8 ±1.8	84.5 ±0.4	55.5 ±1.0	83.3 ±0.4	63.6 ±0.6	82.7 ±0.4	83.0 ±0.4	99.3 ±0.0	8.7 ±0.6
IRM	84.4 ±0.7	40.0 ±0.0	85.1 ±0.5	63.0 ±2.0	84.4 ±0.6	65.9 ±1.2	83.9 ±0.7	84.3 ±0.7	99.4 ±0.0	7.0 ±1.4
CVaRDRO	83.6 ±0.6	36.7 ±2.7	85.0 ±0.4	55.7 ±2.3	83.8 ±0.6	64.3 ±1.5	83.2 ±0.6	83.5 ±0.6	99.4 ±0.0	7.9 ±1.1
JTT	85.1 ±0.3	40.0 ±0.0	86.0 ±0.3	54.8 ±2.7	85.2 ±0.3	65.4 ±1.8	84.7 ±0.3	85.0 ±0.3	99.4 ±0.0	10.2 ±0.2
LfF	78.3 ±0.4	30.4 ±1.3	80.7 ±0.2	45.6 ±1.3	78.6 ±0.4	52.5 ±0.6	78.0 ±0.3	78.3 ±0.4	99.2 ±0.0	1.4 ±0.3
LISA	84.7 ±0.3	42.7 ±2.2	85.7 ±0.2	54.7 ±1.4	84.8 ±0.3	65.4 ±1.2	84.2 ±0.3	84.6 ±0.3	99.2 ±0.0	11.9 ±1.6
MMD	84.9 ±0.1	40.7 ±0.5	85.8 ±0.1	57.0 ±1.2	85.0 ±0.1	66.3 ±0.7	84.5 ±0.1	84.8 ±0.1	99.4 ±0.0	9.2 ±0.4
ReSample	84.8 ±0.3	40.0 ±0.0	85.8 ±0.3	58.6 ±2.6	84.9 ±0.4	65.4 ±1.7	84.4 ±0.4	84.7 ±0.4	99.4 ±0.0	8.8 ±0.2
ReWeight	85.7 ±0.2	41.9 ±1.6	86.6 ±0.1	57.3 ±3.8	85.8 ±0.1	65.0 ±1.7	85.3 ±0.2	85.6 ±0.2	99.4 ±0.0	9.8 ±0.3
SqrtReWeight	84.7 ±0.7	40.0 ±0.0	85.7 ±0.4	57.5 ±1.3	84.8 ±0.6	65.7 ±1.5	84.2 ±0.6	84.6 ±0.7	99.4 ±0.0	8.1 ±1.1
CBLoss	84.5 ±0.4	37.8 ±1.8	85.2 ±0.5	61.1 ±0.8	84.5 ±0.5	66.1 ±1.4	84.0 ±0.4	84.3 ±0.4	99.4 ±0.0	8.3 ±1.2
Focal	83.8 ±1.4	36.7 ±2.7	85.0 ±1.1	54.2 ±3.7	83.9 ±1.4	63.8 ±3.0	83.3 ±1.4	83.6 ±1.4	99.4 ±0.1	4.8 ±0.7
LDAM	82.8 ±0.4	42.0 ±0.9	84.4 ±0.3	51.1 ±2.7	83.0 ±0.4	62.0 ±1.6	82.4 ±0.4	82.7 ±0.4	98.7 ±0.1	68.7 ±2.2
BSoftmax	84.0 ±0.5	40.4 ±0.3	84.8 ±0.3	61.4 ±1.1	84.1 ±0.4	65.2 ±1.1	83.7 ±0.5	84.0 ±0.5	99.4 ±0.0	7.0 ±1.2
DFR	75.6 ±0.5	23.7 ±0.7	77.4 ±0.4	37.7 ±3.2	75.8 ±0.4	46.0 ±2.6	75.3 ±0.5	75.5 ±0.5	98.6 ±0.0	19.4 ±0.5
CRT	85.2 ±0.3	43.3 ±2.7	85.7 ±0.2	64.6 ±0.6	85.2 ±0.3	69.2 ±0.3	84.7 ±0.3	85.0 ±0.3	99.4 ±0.0	7.9 ±0.6
ReWeightCRT	85.0 ±0.1	23.3 ±1.4	85.5 ±0.1	61.6 ±0.3	85.0 ±0.1	67.0 ±0.1	84.3 ±0.1	84.8 ±0.1	99.3 ±0.0	3.6 ±0.3

E.1.8. MIMIC-CXR

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	78.2 ±0.1	68.9 ±0.3	77.3 ±0.1	71.1 ±0.1	77.5 ±0.1	73.6 ±0.1	77.2 ±0.0	77.8 ±0.1	85.2 ±0.1	3.4 ±0.2
Mixup	78.3 ±0.0	68.1 ±0.9	77.4 ±0.0	71.6 ±0.2	77.5 ±0.0	73.4 ±0.1	77.2 ±0.1	77.8 ±0.1	85.1 ±0.1	3.6 ±0.2
GroupDRO	76.9 ±0.3	74.4 ±0.2	76.1 ±0.2	68.7 ±0.5	76.3 ±0.2	72.7 ±0.1	76.7 ±0.1	76.9 ±0.2	83.7 ±0.1	4.7 ±0.1
IRM	78.2 ±0.0	67.7 ±0.2	77.3 ±0.0	71.4 ±0.1	77.5 ±0.0	73.5 ±0.1	77.2 ±0.1	77.8 ±0.1	85.2 ±0.1	3.4 ±0.2
CVaRDRO	78.3 ±0.1	68.6 ±0.4	77.4 ±0.1	71.1 ±0.3	77.7 ±0.1	73.9 ±0.0	77.4 ±0.0	78.1 ±0.0	85.1 ±0.0	7.8 ±0.3
JTT	78.1 ±0.0	67.3 ±0.7	77.1 ±0.0	71.4 ±0.2	77.3 ±0.0	73.2 ±0.1	77.0 ±0.0	77.5 ±0.1	84.9 ±0.0	3.4 ±0.1
LfF	73.3 ±0.9	62.6 ±2.6	72.3 ±1.0	65.2 ±1.0	72.4 ±1.0	67.7 ±1.4	72.4 ±1.1	72.8 ±1.1	79.3 ±1.3	12.3 ±0.7
LISA	77.9 ±0.1	70.4 ±0.2	77.0 ±0.1	70.6 ±0.3	77.2 ±0.1	73.3 ±0.1	77.2 ±0.1	77.6 ±0.1	84.9 ±0.1	4.0 ±0.6
MMD	76.8 ±0.2	68.0 ±0.6	75.9 ±0.2	70.2 ±0.4	76.0 ±0.2	71.5 ±0.3	76.0 ±0.3	76.2 ±0.2	83.4 ±0.2	8.8 ±2.0
ReSample	78.1 ±0.1	71.9 ±0.2	77.3 ±0.1	70.7 ±0.3	77.5 ±0.1	73.8 ±0.2	77.6 ±0.1	78.0 ±0.1	85.0 ±0.1	5.5 ±0.8
ReWeight	78.2 ±0.1	71.6 ±0.3	77.4 ±0.1	70.9 ±0.3	77.6 ±0.1	73.8 ±0.1	77.6 ±0.1	78.0 ±0.1	85.1 ±0.1	4.2 ±0.2
SqrtReWeight	78.2 ±0.2	70.3 ±0.2	77.3 ±0.2	71.0 ±0.3	77.5 ±0.2	73.6 ±0.2	77.3 ±0.3	77.9 ±0.2	85.2 ±0.2	4.1 ±0.2
CBLoss	78.4 ±0.1	70.7 ±0.1	77.5 ±0.1	71.6 ±0.2	77.7 ±0.1	73.8 ±0.1	77.6 ±0.1	78.0 ±0.1	85.2 ±0.0	4.1 ±0.4
Focal	78.3 ±0.1	68.7 ±0.4	77.4 ±0.1	70.8 ±0.2	77.6 ±0.1	73.9 ±0.0	77.4 ±0.1	78.1 ±0.0	85.4 ±0.0	10.1 ±0.6
LDAM	77.7 ±0.6	68.6 ±1.1	76.8 ±0.6	70.4 ±0.9	77.0 ±0.6	73.1 ±0.7	76.9 ±0.6	77.4 ±0.6	84.6 ±0.6	22.0 ±0.2
BSoftmax	77.8 ±0.2	68.4 ±0.2	76.9 ±0.2	70.2 ±0.3	77.1 ±0.2	73.3 ±0.2	77.0 ±0.2	77.6 ±0.2	84.9 ±0.2	5.0 ±0.2
DFR	78.0 ±0.0	68.9 ±0.0	77.1 ±0.0	70.9 ±0.0	77.3 ±0.0	73.3 ±0.0	77.0 ±0.0	77.6 ±0.0	84.9 ±0.0	7.0 ±0.1
CRT	78.5 ±0.0	71.0 ±0.0	77.6 ±0.0	71.5 ±0.1	77.9 ±0.0	74.0 ±0.0	77.7 ±0.0	78.2 ±0.0	85.4 ±0.0	4.1 ±0.1
ReWeightCRT	78.5 ±0.0	70.8 ±0.0	77.6 ±0.0	71.5 ±0.1	77.8 ±0.0	73.9 ±0.0	77.7 ±0.0	78.2 ±0.0	85.4 ±0.0	4.3 ±0.0

E.1.9. MIMICNOTES

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	91.1 ±0.1	18.7 ±2.7	77.6 ±0.9	63.1 ±2.0	63.2 ±1.6	31.2 ±3.1	59.7 ±1.3	59.9 ±1.3	85.3 ±0.1	2.1 ±0.8
Mixup	91.1 ±0.0	22.7 ±3.2	76.8 ±0.7	61.2 ±1.6	65.1 ±1.6	35.0 ±3.2	61.5 ±1.6	61.7 ±1.7	85.4 ±0.0	2.0 ±0.8
GroupDRO	76.1 ±0.7	72.6 ±0.5	61.3 ±0.1	25.7 ±0.4	61.8 ±0.4	38.6 ±0.3	76.2 ±0.3	76.5 ±0.2	85.0 ±0.1	22.2 ±0.6
IRM	91.0 ±0.0	22.5 ±2.5	76.3 ±0.5	60.1 ±1.2	65.2 ±1.2	35.3 ±2.4	61.5 ±1.2	61.7 ±1.2	85.3 ±0.0	1.9 ±0.2
CVaRDRO	90.9 ±0.1	23.0 ±4.6	76.5 ±1.2	60.6 ±2.9	64.6 ±2.4	34.0 ±4.8	61.4 ±2.3	61.6 ±2.4	85.1 ±0.1	4.2 ±1.6
JTT	71.3 ±3.7	65.9 ±2.8	60.3 ±0.7	23.4 ±1.9	58.6 ±2.3	36.0 ±1.8	75.5 ±0.4	75.6 ±0.4	84.9 ±0.1	27.5 ±3.9
LfF	84.0 ±1.2	62.7 ±2.1	64.6 ±0.7	33.6 ±1.6	67.1 ±0.8	43.6 ±0.8	74.7 ±0.4	74.7 ±0.5	85.1 ±0.0	12.5 ±1.2
LISA	85.2 ±1.4	58.0 ±3.1	65.5 ±0.9	35.7 ±2.1	68.0 ±0.9	44.5 ±0.9	74.0 ±0.6	74.2 ±0.7	85.3 ±0.0	15.5 ±1.5
MMD	91.2 ±0.1	23.0 ±0.5	76.8 ±0.3	61.0 ±0.5	65.9 ±0.5	36.5 ±0.9	61.9 ±0.4	62.1 ±0.4	85.3 ±0.0	1.4 ±0.1
ReSample	80.4 ±1.8	68.0 ±3.0	63.0 ±0.8	29.7 ±1.8	64.9 ±1.2	41.6 ±1.2	75.8 ±0.4	76.1 ±0.4	85.3 ±0.0	18.8 ±2.2
ReWeight	84.8 ±0.8	60.5 ±2.5	65.2 ±0.6	34.8 ±1.3	67.8 ±0.5	44.4 ±0.5	74.5 ±0.5	74.7 ±0.5	85.2 ±0.0	14.1 ±0.9
SqrtReWeight	90.1 ±0.3	37.2 ±4.5	71.8 ±1.1	49.9 ±2.5	69.1 ±0.9	43.7 ±1.9	67.6 ±1.7	67.8 ±1.7	85.2 ±0.1	4.2 ±1.0
CBLoss	83.2 ±1.2	63.3 ±2.2	64.1 ±0.6	32.5 ±1.5	66.6 ±0.8	43.0 ±0.8	74.8 ±0.4	74.9 ±0.5	85.2 ±0.1	14.7 ±1.3
Focal	91.0 ±0.0	19.1 ±2.3	77.1 ±0.6	62.1 ±1.4	63.6 ±1.3	31.9 ±2.6	59.9 ±1.1	60.2 ±1.1	85.3 ±0.1	8.1 ±0.7
LDAM	90.6 ±0.1	5.3 ±2.4	84.4 ±0.8	78.1 ±1.7	52.5 ±2.1	10.0 ±4.1	52.7 ±1.2	52.7 ±1.2	84.9 ±0.1	28.9 ±1.0
BSoftmax	76.9 ±0.9	73.1 ±1.0	61.7 ±0.2	26.5 ±0.6	62.5 ±0.5	39.3 ±0.4	76.6 ±0.2	76.7 ±0.2	85.4 ±0.0	23.5 ±1.1
DFR	43.1 ±19.8	6.7 ±5.5	51.9 ±2.8	7.3 ±3.0	28.3 ±9.1	7.2 ±5.8	53.4 ±2.8	53.4 ±2.8	84.5 ±0.0	40.1 ±0.3
CRT	82.1 ±3.5	56.2 ±13.8	65.9 ±3.5	36.8 ±8.2	63.4 ±0.5	37.5 ±1.4	70.9 ±4.0	71.0 ±4.0	84.3 ±0.0	28.3 ±4.3
ReWeightCRT	83.5 ±2.6	58.7 ±6.8	64.6 ±1.5	33.9 ±3.5	66.1 ±1.5	42.0 ±1.2	72.9 ±1.5	73.0 ±1.5	84.3 ±0.0	28.9 ±2.2

E.1.10. CXRMULTISITE

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	98.3 ±0.0	0.0 ±0.0	49.2 ±0.0	0.0 ±0.0	49.6 ±0.0	0.0 ±0.0	50.0 ±0.0	50.0 ±0.0	93.1 ±0.1	0.3 ±0.1
Mixup	98.3 ±0.0	0.0 ±0.0	49.2 ±0.0	0.0 ±0.0	49.6 ±0.0	0.0 ±0.0	50.0 ±0.0	50.0 ±0.0	92.9 ±0.1	0.3 ±0.0
GroupDRO	84.0 ±10.9	19.3 ±13.7	55.7 ±2.0	12.7 ±4.1	51.2 ±5.0	12.4 ±2.9	55.9 ±2.5	59.7 ±2.5	79.4 ±2.9	29.3 ±6.3
IRM	77.5 ±17.0	8.8 ±7.2	49.6 ±0.3	0.7 ±0.5	42.4 ±5.9	1.3 ±1.0	51.1 ±0.9	51.8 ±1.5	64.2 ±7.3	47.3 ±1.1
CVaRDRO	98.3 ±0.0	0.0 ±0.0	61.2 ±4.9	24.0 ±9.8	50.7 ±0.7	2.2 ±1.5	50.2 ±0.2	50.6 ±0.4	93.0 ±0.0	0.9 ±0.3
JTT	94.1 ±0.9	0.0 ±0.0	59.0 ±0.7	18.5 ±1.4	62.9 ±0.8	28.9 ±1.2	55.2 ±0.9	82.2 ±2.4	93.2 ±0.1	6.4 ±0.5
LfF	9.9 ±6.7	5.4 ±4.4	17.4 ±13.5	0.6 ±0.5	8.5 ±5.6	1.2 ±1.0	50.5 ±0.4	51.7 ±1.4	60.6 ±1.6	82.6 ±12.8
LISA	98.3 ±0.0	0.0 ±0.0	49.2 ±0.0	0.0 ±0.0	49.6 ±0.0	0.0 ±0.0	50.0 ±0.0	50.0 ±0.0	90.3 ±0.0	8.9 ±1.6
MMD	87.4 ±8.9	12.8 ±10.4	49.6 ±0.4	0.8 ±0.6	47.0 ±2.1	1.4 ±1.2	50.6 ±0.5	52.0 ±1.7	56.5 ±1.9	15.4 ±12.5
ReSample	96.4 ±0.3	1.1 ±0.5	57.9 ±0.3	17.1 ±0.5	59.8 ±0.1	21.4 ±0.3	54.0 ±0.1	63.4 ±1.2	89.7 ±0.1	4.1 ±0.2
ReWeight	88.0 ±5.3	19.4 ±7.9	52.9 ±0.7	6.9 ±1.4	52.0 ±2.2	10.8 ±2.1	56.7 ±1.7	64.1 ±4.4	75.7 ±2.4	37.2 ±4.1
SqrtReWeight	98.0 ±0.1	0.0 ±0.0	65.5 ±0.3	32.4 ±0.6	60.7 ±1.8	22.5 ±3.7	53.4 ±0.7	58.9 ±2.1	92.9 ±0.1	4.1 ±0.9
CBLoss	98.0 ±0.0	0.0 ±0.0	64.7 ±0.3	30.9 ±0.6	59.2 ±1.0	19.4 ±1.9	52.5 ±0.3	56.9 ±1.0	92.5 ±0.0	6.0 ±0.8
Focal	98.3 ±0.0	0.0 ±0.0	55.4 ±5.1	12.5 ±10.2	49.7 ±0.1	0.3 ±0.2	50.0 ±0.0	50.1 ±0.1	93.2 ±0.0	11.5 ±0.6
LDAM	98.3 ±0.0	0.0 ±0.0	49.2 ±0.0	0.0 ±0.0	49.6 ±0.0	0.0 ±0.0	50.0 ±0.0	50.0 ±0.0	92.9 ±0.0	33.3 ±0.0
BSoftmax	89.1 ±0.2	0.5 ±0.1	56.2 ±0.1	12.5 ±0.2	58.1 ±0.2	22.0 ±0.3	50.4 ±0.0	90.0 ±0.1	92.9 ±0.1	19.9 ±1.3
DFR	79.3 ±8.8	22.2 ±9.9	54.9 ±2.7	10.9 ±5.6	48.2 ±3.5	9.0 ±1.4	55.5 ±1.8	63.9 ±4.7	78.9 ±5.3	41.5 ±1.9
CRT	87.0 ±2.7	17.2 ±6.5	54.2 ±1.4	9.1 ±2.8	54.2 ±2.9	15.5 ±4.2	58.2 ±1.5	73.4 ±3.0	81.5 ±4.1	35.1 ±3.4
ReWeightCRT	82.5 ±6.3	27.8 ±11.4	56.0 ±3.5	13.0 ±7.1	51.8 ±4.1	13.6 ±4.5	58.7 ±2.1	66.8 ±2.8	81.1 ±4.8	29.9 ±8.6

E.1.11. CHEXPERT

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	86.9 ±0.5	50.2 ±3.8	66.3 ±0.3	37.6 ±0.8	68.6 ±0.2	44.6 ±0.5	72.5 ±1.1	72.8 ±1.1	84.8 ±0.3	8.6 ±0.4
Mixup	81.9 ±6.2	37.4 ±3.5	63.5 ±5.0	33.9 ±9.3	62.5 ±5.9	35.7 ±7.6	63.8 ±4.7	64.1 ±4.6	76.1 ±8.5	16.1 ±9.0
GroupDRO	78.9 ±0.3	74.5 ±0.2	62.8 ±0.1	28.3 ±0.3	64.4 ±0.2	41.7 ±0.3	78.4 ±0.1	79.0 ±0.1	86.0 ±0.1	21.1 ±1.0
IRM	89.8 ±0.3	34.4 ±1.7	70.1 ±0.7	46.5 ±1.6	68.6 ±0.1	42.7 ±0.3	67.5 ±0.6	67.5 ±0.6	85.8 ±0.3	4.4 ±1.2
CVaRDRO	66.2 ±2.7	57.9 ±0.4	56.4 ±0.5	17.7 ±1.0	52.9 ±1.6	27.9 ±1.1	66.1 ±0.6	67.0 ±0.6	73.0 ±0.6	40.4 ±0.0
JTT	73.0 ±1.9	61.3 ±4.9	58.6 ±1.1	21.6 ±1.7	57.9 ±1.7	32.8 ±2.4	69.8 ±2.6	71.0 ±2.5	77.6 ±2.3	26.3 ±2.0
LfF	22.3 ±10.2	13.7 ±9.8	37.3 ±5.8	9.0 ±0.7	19.5 ±8.3	8.8 ±3.9	46.2 ±2.9	46.2 ±3.1	30.5 ±10.1	65.7 ±10.2
LISA	79.2 ±0.8	75.6 ±0.6	63.1 ±0.4	28.8 ±0.8	64.8 ±0.6	42.3 ±0.7	78.8 ±0.3	79.4 ±0.1	86.5 ±0.1	21.5 ±1.2
MMD	86.9 ±0.5	50.2 ±3.8	66.3 ±0.3	37.6 ±0.8	68.6 ±0.2	44.6 ±0.5	72.5 ±1.1	72.8 ±1.1	84.8 ±0.3	8.6 ±0.4
ReSample	79.0 ±0.8	75.3 ±0.5	62.8 ±0.3	28.4 ±0.7	64.5 ±0.6	41.7 ±0.7	78.4 ±0.0	78.7 ±0.1	85.7 ±0.2	20.1 ±1.4
ReWeight	78.7 ±0.4	75.7 ±0.1	62.7 ±0.1	28.2 ±0.3	64.3 ±0.3	41.6 ±0.3	78.5 ±0.1	78.9 ±0.0	86.3 ±0.0	20.9 ±0.5
SqrtReWeight	82.1 ±1.5	70.0 ±2.3	64.3 ±0.7	31.8 ±1.7	66.7 ±1.1	44.1 ±1.2	77.7 ±0.5	78.3 ±0.5	86.5 ±0.1	18.8 ±2.2
CBLoss	79.1 ±0.1	74.7 ±0.3	62.7 ±0.0	28.3 ±0.1	64.3 ±0.1	41.4 ±0.1	77.9 ±0.0	78.4 ±0.1	85.7 ±0.1	22.1 ±0.5
Focal	89.3 ±0.3	42.1 ±4.0	69.6 ±0.4	44.7 ±1.1	69.8 ±0.4	45.5 ±1.0	70.4 ±1.1	70.4 ±1.3	86.5 ±0.1	16.1 ±1.7
LDAM	90.1 ±0.0	36.4 ±0.3	70.6 ±0.1	47.5 ±0.1	68.9 ±0.2	43.3 ±0.3	67.3 ±0.3	67.6 ±0.2	86.0 ±0.1	32.3 ±0.3
BSoftmax	79.1 ±0.4	75.4 ±0.5	63.0 ±0.2	28.6 ±0.4	64.7 ±0.3	42.1 ±0.4	78.4 ±0.2	79.2 ±0.1	86.4 ±0.1	23.9 ±0.2
DFR	78.2 ±0.4	71.7 ±0.2	62.4 ±0.2	27.6 ±0.4	63.8 ±0.3	40.9 ±0.3	77.5 ±0.1	78.6 ±0.0	85.5 ±0.0	39.5 ±0.1
CRT	79.1 ±0.2	74.6 ±0.3	62.8 ±0.1	28.4 ±0.2	64.4 ±0.2	41.6 ±0.3	78.0 ±0.2	78.6 ±0.2	85.8 ±0.2	21.2 ±0.3
ReWeightCRT	80.4 ±0.0	76.0 ±0.1	63.5 ±0.0	29.8 ±0.1	65.6 ±0.1	43.0 ±0.1	78.8 ±0.1	79.1 ±0.1	86.3 ±0.0	20.2 ±0.1

E.1.12. LIVING17

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	28.2 ±1.5	8.7 ±1.0	29.5 ±1.1	9.0 ±1.2	27.8 ±1.3	10.0 ±1.5	28.2 ±1.5	28.2 ±1.5	77.9 ±1.2	53.5 ±2.5
Mixup	29.8 ±1.8	9.3 ±1.4	32.5 ±2.3	9.5 ±1.3	29.8 ±1.8	10.1 ±1.5	29.8 ±1.8	29.8 ±1.8	78.3 ±1.2	34.6 ±1.7
GroupDRO	27.2 ±1.5	9.7 ±0.7	29.8 ±0.7	7.8 ±0.9	27.3 ±1.1	9.1 ±0.8	27.2 ±1.5	27.2 ±1.5	77.8 ±0.6	55.8 ±0.5
IRM	28.2 ±1.5	8.7 ±1.0	29.5 ±1.1	9.0 ±1.2	27.8 ±1.3	10.0 ±1.5	28.2 ±1.5	28.2 ±1.5	77.9 ±1.2	53.5 ±2.5
CVaRDRO	28.3 ±0.7	8.3 ±0.7	30.0 ±0.7	7.9 ±0.2	27.9 ±0.8	8.0 ±0.2	28.3 ±0.7	28.3 ±0.7	81.0 ±0.1	33.2 ±4.1
JTT	28.8 ±1.1	8.7 ±1.0	29.8 ±1.8	8.3 ±1.5	28.3 ±1.4	9.1 ±1.6	28.8 ±1.1	28.8 ±1.1	80.2 ±1.1	38.0 ±5.8
LfF	26.2 ±1.1	8.7 ±0.3	28.3 ±0.9	8.8 ±1.1	26.0 ±1.1	9.3 ±0.6	26.2 ±1.1	26.2 ±1.1	76.6 ±0.7	56.4 ±3.3
LISA	29.8 ±0.9	11.3 ±0.3	32.0 ±0.4	9.4 ±0.4	29.9 ±0.7	10.4 ±0.3	29.8 ±0.9	29.8 ±0.9	78.2 ±0.6	30.3 ±0.6
MMD	26.6 ±1.8	8.3 ±0.3	28.9 ±1.1	9.5 ±1.1	26.5 ±1.5	9.5 ±0.9	26.6 ±1.8	26.6 ±1.8	78.5 ±1.0	48.4 ±6.7
ReSample	30.7 ±2.1	10.3 ±2.3	33.1 ±1.2	10.5 ±1.3	30.7 ±2.0	11.2 ±1.4	30.7 ±2.1	30.7 ±2.1	80.9 ±0.4	47.5 ±3.1
ReWeight	28.2 ±1.5	8.7 ±1.0	29.5 ±1.1	9.0 ±1.2	27.8 ±1.3	10.0 ±1.5	28.2 ±1.5	28.2 ±1.5	77.9 ±1.2	53.5 ±2.5
SqrtReWeight	28.2 ±1.5	8.7 ±1.0	29.5 ±1.1	9.0 ±1.2	27.8 ±1.3	10.0 ±1.5	28.2 ±1.5	28.2 ±1.5	77.9 ±1.2	53.5 ±2.5
CBLoss	28.2 ±1.5	8.7 ±1.0	29.5 ±1.1	9.0 ±1.2	27.8 ±1.3	10.0 ±1.5	28.2 ±1.5	28.2 ±1.5	77.9 ±1.2	53.5 ±2.5
Focal	28.0 ±1.2	8.0 ±0.5	28.8 ±1.3	7.8 ±1.1	27.1 ±1.0	8.3 ±1.0	28.0 ±1.2	28.0 ±1.2	79.5 ±1.1	48.6 ±1.0
LDAM	24.7 ±0.8	7.0 ±0.5	28.3 ±0.6	6.0 ±0.4	24.5 ±0.6	6.7 ±0.3	24.7 ±0.8	24.7 ±0.8	78.1 ±1.2	9.7 ±2.7
BSoftmax	27.5 ±0.8	8.7 ±0.7	28.6 ±1.0	8.5 ±0.7	27.0 ±0.8	9.4 ±1.0	27.5 ±0.8	27.5 ±0.8	78.1 ±1.0	54.7 ±3.1
DFR	29.0 ±0.2	10.0 ±0.0	31.6 ±0.3	10.8 ±0.5	28.8 ±0.2	11.6 ±0.5	29.0 ±0.2	29.0 ±0.2	82.8 ±0.0	3.4 ±0.4
CRT	33.9 ±0.1	10.7 ±0.3	34.5 ±0.2	10.0 ±0.2	33.3 ±0.1	10.3 ±0.2	33.9 ±0.1	33.9 ±0.1	83.2 ±0.1	32.8 ±1.4
ReWeightCRT	33.7 ±0.1	7.7 ±0.3	33.9 ±0.1	15.3 ±0.6	33.1 ±0.1	11.5 ±0.4	33.7 ±0.1	33.7 ±0.1	82.5 ±0.0	41.4 ±0.2

E.1.13. Overall

Algorithm	Waterbirds	CelebA	CivilComments	MultiNLI	MetaShift	ImageNetBG	NICO++	MIMIC-CXR	MIMICNotes	CXRMultisite	CheXpert	Living17	Avg
ERM	69.1 ±4.7	62.6 ±1.5	63.7 ±1.1	66.8 ±0.5	82.6 ±0.4	81.0 ±0.9	37.6 ±2.0	68.9 ±0.3	83.1 ±0.1	50.1 ±0.9	50.2 ±3.8	28.2 ±1.5	62.0
Mixup	78.2 ±0.4	57.8 ±0.8	66.1 ±1.3	68.5 ±0.6	81.0 ±0.8	82.4 ±0.3	42.7 ±1.4	68.1 ±0.9	82.0 ±0.4	50.1 ±0.9	37.4 ±3.5	29.8 ±1.8	62.0
GroupDRO	78.6 ±1.0	89.0 ±0.7	70.6 ±1.2	76.0 ±0.7	85.6 ±0.4	78.2 ±0.3	37.8 ±1.8	74.4 ±0.2	83.7 ±0.1	59.6 ±1.0	74.5 ±0.2	27.2 ±1.5	69.6
IRM	74.5 ±1.5	63.0 ±2.5	63.2 ±0.8	63.6 ±1.3	83.0 ±0.1	81.3 ±0.3	40.0 ±0.0	67.7 ±0.2	83.2 ±0.2	47.9 ±1.1	34.4 ±1.7	28.2 ±1.5	60.8
CVaRDRO	75.5 ±2.2	64.1 ±2.8	68.7 ±1.3	63.0 ±1.5	84.6 ±0.0	80.7 ±1.1	36.7 ±2.7	68.6 ±0.4	81.9 ±0.1	50.2 ±0.9	57.9 ±0.4	28.3 ±0.7	63.3
JTT	72.0 ±0.3	70.0 ±10.2	64.3 ±1.5	69.1 ±0.1	83.6 ±0.4	80.5 ±0.3	40.0 ±0.0	67.3 ±0.7	83.8 ±0.1	50.1 ±0.9	61.3 ±4.9	28.8 ±1.1	64.2
LfF	75.2 ±0.7	53.0 ±4.3	51.0 ±6.1	63.6 ±2.9	73.1 ±1.6	76.7 ±0.5	30.4 ±1.3	62.6 ±2.6	84.0 ±0.1	50.1 ±0.9	13.7 ±9.8	26.2 ±1.1	55.0
LISA	88.7 ±0.6	86.5 ±1.2	73.7 ±0.3	73.3 ±1.0	84.1 ±0.4	76.1 ±0.8	42.7 ±2.2	70.4 ±0.2	83.6 ±0.2	48.9 ±1.3	75.6 ±0.6	29.8 ±0.9	69.5
MMD	83.9 ±1.4	24.4 ±2.0	54.5 ±1.4	69.1 ±1.5	85.9 ±0.7	80.8 ±0.5	40.7 ±0.5	68.0 ±0.6	82.0 ±0.5	50.1 ±0.9	50.2 ±3.8	26.6 ±1.8	59.7
ReSample	77.7 ±1.2	87.4 ±0.8	73.3 ±0.5	72.3 ±0.8	85.6 ±0.4	81.0 ±0.4	40.0 ±0.0	71.9 ±0.2	83.9 ±0.1	59.0 ±1.1	75.3 ±0.5	30.7 ±2.1	69.8
ReWeight	86.9 ±0.7	89.7 ±0.2	72.5 ±0.0	68.8 ±0.4	85.6 ±0.4	81.0 ±0.9	41.9 ±1.6	71.6 ±0.3	83.6 ±0.3	64.2 ±0.7	75.7 ±0.1	28.2 ±1.5	70.8
SqrtReWeight	78.6 ±0.1	82.4 ±0.5	71.7 ±0.4	69.5 ±0.7	84.6 ±0.7	80.1 ±0.2	40.0 ±0.0	70.3 ±0.2	83.1 ±0.1	50.0 ±1.0	70.0 ±2.3	28.2 ±1.5	67.4
CBLoss	86.2 ±0.3	89.4 ±0.7	73.3 ±0.2	72.2 ±0.3	85.5 ±0.4	81.0 ±0.9	37.8 ±1.8	70.7 ±0.1	83.2 ±0.1	50.1 ±0.9	74.7 ±0.3	28.2 ±1.5	69.4
Focal	71.6 ±0.8	59.1 ±2.0	62.0 ±1.0	69.4 ±0.7	81.5 ±0.0	78.4 ±0.1	36.7 ±2.7	68.7 ±0.4	71.1 ±9.9	50.0 ±0.9	42.1 ±4.0	28.0 ±1.2	59.9
LDAM	71.0 ±1.8	59.6 ±2.4	37.4 ±8.1	69.6 ±1.6	83.6 ±0.4	80.1 ±0.3	42.0 ±0.9	68.6 ±1.1	81.0 ±0.3	50.0 ±0.0	36.4 ±0.3	24.7 ±0.8	58.7
BSoftmax	74.1 ±0.9	83.3 ±0.5	71.2 ±0.4	66.9 ±0.4	83.1 ±0.7	80.7 ±0.7	40.4 ±0.3	68.4 ±0.2	83.4 ±0.3	50.1 ±1.1	75.4 ±0.5	27.5 ±0.8	67.0
DFR	91.0 ±0.3	90.4 ±0.1	69.6 ±0.2	68.5 ±0.2	85.4 ±0.4	78.5 ±0.6	23.7 ±0.7	68.9 ±0.0	83.6 ±0.0	53.5 ±0.2	71.7 ±0.2	29.0 ±0.2	67.8
CRT	79.7 ±0.3	87.2 ±0.3	71.1 ±0.1	70.7 ±0.1	84.1 ±0.4	80.2 ±0.3	43.3 ±2.7	71.0 ±0.0	83.4 ±0.0	65.2 ±0.1	74.6 ±0.3	33.9 ±0.1	70.4
ReWeightCRT	78.4 ±0.1	87.2 ±0.3	71.0 ±0.1	69.0 ±0.2	85.6 ±0.4	79.4 ±0.2	23.3 ±1.4	70.8 ±0.0	83.4 ±0.0	60.9 ±0.8	76.0 ±0.1	33.7 ±0.1	68.2

E.2. Attributes Unknown in Training, but Known in Validation

E.2.1. Waterbirds

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	84.1 ±1.7	69.1 ±4.7	77.4 ±2.0	60.7 ±3.2	79.4 ±2.1	69.5 ±2.9	83.1 ±2.0	83.1 ±2.0	91.0 ±1.4	12.9 ±1.7
Mixup	89.5 ±0.4	78.2 ±0.4	83.9 ±0.6	71.6 ±1.3	85.9 ±0.4	78.8 ±0.6	88.9 ±0.3	88.9 ±0.3	94.7 ±0.2	7.0 ±0.6
GroupDRO	86.9 ±0.9	73.1 ±0.4	80.7 ±1.1	66.1 ±2.2	82.8 ±0.9	74.4 ±1.2	86.3 ±0.5	86.3 ±0.5	94.0 ±0.3	10.5 ±0.8
CVaRDRO	89.9 ±0.4	75.5 ±2.2	84.5 ±0.7	73.2 ±1.7	86.2 ±0.3	79.0 ±0.4	88.5 ±0.3	88.5 ±0.3	95.4 ±0.2	8.3 ±0.2
JTT	88.9 ±0.6	71.0 ±0.5	83.2 ±0.8	71.5 ±1.7	84.7 ±0.6	76.8 ±0.8	86.8 ±0.2	86.8 ±0.2	94.2 ±0.1	9.1 ±0.3
LfF	86.5 ±0.6	74.7 ±1.0	80.1 ±0.7	64.7 ±1.4	82.3 ±0.6	73.8 ±0.8	86.2 ±0.3	86.2 ±0.3	93.5 ±0.2	9.9 ±0.8
LISA	89.5 ±0.4	78.2 ±0.4	83.9 ±0.6	71.6 ±1.3	85.9 ±0.4	78.8 ±0.6	88.9 ±0.3	88.9 ±0.3	94.7 ±0.2	7.0 ±0.6
ReSample	86.2 ±0.5	70.0 ±1.0	79.8 ±0.6	64.9 ±1.4	81.7 ±0.5	72.7 ±0.7	85.0 ±0.2	85.0 ±0.2	92.8 ±0.1	11.3 ±0.3
ReWeight	86.9 ±0.5	72.5 ±0.3	80.7 ±0.6	66.1 ±1.3	82.7 ±0.5	74.2 ±0.6	86.1 ±0.1	86.1 ±0.1	93.9 ±0.1	10.6 ±0.3
SqrtReWeight	89.7 ±0.4	71.3 ±1.4	84.3 ±0.6	73.6 ±0.9	85.7 ±0.6	78.2 ±0.8	87.4 ±0.5	87.4 ±0.5	94.5 ±0.4	8.8 ±0.4
CBLoss	86.8 ±0.6	74.4 ±1.2	80.4 ±0.7	65.5 ±1.3	82.6 ±0.7	74.0 ±1.0	86.2 ±0.6	86.2 ±0.6	93.5 ±0.4	11.3 ±0.4
Focal	89.3 ±0.2	71.6 ±0.8	83.7 ±0.3	72.4 ±0.5	85.2 ±0.3	77.5 ±0.4	87.1 ±0.3	87.1 ±0.3	94.2 ±0.2	6.9 ±0.1
LDAM	87.3 ±0.5	71.0 ±1.8	81.2 ±0.6	67.7 ±1.5	83.0 ±0.5	74.4 ±0.6	85.7 ±0.2	85.7 ±0.2	93.3 ±0.2	13.7 ±2.2
BSoftmax	88.4 ±1.2	74.1 ±0.9	82.6 ±1.6	69.9 ±2.9	84.4 ±1.5	76.4 ±2.0	87.0 ±1.0	87.0 ±1.0	94.0 ±0.9	9.9 ±1.2
DFR	92.2 ±0.2	89.0 ±0.2	87.7 ±0.3	78.4 ±0.5	89.2 ±0.2	83.6 ±0.3	91.2 ±0.1	91.2 ±0.1	96.8 ±0.1	6.8 ±0.4
CRT	89.2 ±0.1	76.3 ±0.8	83.5 ±0.1	71.3 ±0.4	85.3 ±0.1	77.8 ±0.1	87.9 ±0.1	87.9 ±0.1	94.8 ±0.0	9.2 ±0.2
ReWeightCRT	89.4 ±0.3	76.3 ±0.2	83.8 ±0.3	71.9 ±0.7	85.6 ±0.3	78.1 ±0.4	88.0 ±0.2	88.0 ±0.2	94.9 ±0.1	8.8 ±0.2

E.2.2. CelebA

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	95.0 ±0.1	57.6 ±0.8	87.4 ±0.2	76.3 ±0.3	89.9 ±0.1	82.7 ±0.2	85.6 ±0.2	93.0 ±0.0	98.4 ±0.0	2.9 ±0.1
Mixup	95.4 ±0.1	57.8 ±0.8	88.4 ±0.3	78.5 ±0.7	90.6 ±0.2	83.8 ±0.3	85.8 ±0.2	93.1 ±0.1	98.4 ±0.1	2.5 ±0.2
GroupDRO	92.4 ±0.2	78.5 ±1.1	81.9 ±0.3	64.5 ±0.5	86.3 ±0.2	77.1 ±0.4	90.1 ±0.3	93.9 ±0.1	98.5 ±0.0	7.4 ±0.5
CVaRDRO	95.1 ±0.1	62.2 ±3.1	87.8 ±0.2	77.1 ±0.5	90.2 ±0.1	83.2 ±0.1	86.8 ±0.7	93.2 ±0.2	98.4 ±0.1	3.0 ±0.2
JTT	88.1 ±0.5	66.0 ±11.9	75.2 ±1.1	54.1 ±1.0	76.7 ±3.3	60.5 ±6.6	81.7 ±5.5	82.7 ±6.7	91.3 ±0.2	5.0 ±1.0
LfF	81.1 ±5.6	53.0 ±4.3	71.8 ±4.1	45.2 ±8.3	73.2 ±5.6	59.0 ±7.3	78.3 ±3.0	85.3 ±2.9	94.1 ±1.2	27.9 ±5.5
LISA	95.4 ±0.1	57.8 ±0.8	88.4 ±0.3	78.5 ±0.7	90.6 ±0.2	83.8 ±0.3	85.8 ±0.2	93.1 ±0.1	98.4 ±0.1	2.5 ±0.2
ReSample	92.2 ±0.4	82.2 ±1.2	81.5 ±0.6	63.7 ±1.3	85.9 ±0.5	76.6 ±0.8	90.8 ±0.1	93.8 ±0.1	98.5 ±0.0	7.4 ±0.8
ReWeight	92.0 ±0.4	81.5 ±0.9	81.3 ±0.6	63.2 ±1.4	85.7 ±0.6	76.3 ±0.8	90.7 ±0.2	93.8 ±0.1	98.4 ±0.1	7.8 ±0.8
SqrtReWeight	93.7 ±0.2	72.0 ±2.2	84.3 ±0.5	69.4 ±1.1	88.1 ±0.4	80.0 ±0.6	89.0 ±0.5	94.0 ±0.0	98.4 ±0.0	5.3 ±0.4
CBLoss	93.8 ±0.3	75.0 ±2.4	84.5 ±0.7	69.9 ±1.6	88.3 ±0.5	80.3 ±0.8	89.8 ±0.5	94.0 ±0.1	98.5 ±0.0	5.1 ±0.6
Focal	94.9 ±0.3	59.1 ±2.0	87.5 ±0.8	76.7 ±1.7	89.7 ±0.4	82.4 ±0.6	85.6 ±0.5	92.5 ±0.4	98.2 ±0.1	3.2 ±0.4
LDAM	94.5 ±0.2	59.3 ±2.3	86.5 ±0.8	74.7 ±1.9	89.1 ±0.2	81.4 ±0.3	85.6 ±0.8	92.5 ±0.7	98.2 ±0.1	28.0 ±2.6
BSoftmax	91.9 ±0.1	83.3 ±0.5	81.1 ±0.2	62.9 ±0.4	85.6 ±0.2	76.1 ±0.3	91.1 ±0.2	93.9 ±0.1	98.6 ±0.0	8.4 ±0.2
DFR	91.2 ±0.1	86.3 ±0.3	80.0 ±0.1	61.0 ±0.2	84.4 ±0.1	74.1 ±0.1	90.8 ±0.0	92.6 ±0.0	97.9 ±0.0	14.1 ±0.0
CRT	94.1 ±0.1	70.4 ±0.4	85.1 ±0.2	71.3 ±0.5	88.6 ±0.1	80.6 ±0.2	88.5 ±0.1	93.5 ±0.1	98.4 ±0.0	4.5 ±0.2
ReWeightCRT	94.2 ±0.1	71.1 ±0.5	85.3 ±0.1	71.8 ±0.3	88.7 ±0.1	80.9 ±0.1	88.7 ±0.1	93.6 ±0.1	98.4 ±0.0	4.6 ±0.2

E.2.3. CivilComments

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	85.4 ±0.2	63.2 ±1.2	75.4 ±0.3	57.6 ±0.7	77.0 ±0.1	63.2 ±0.1	77.7 ±0.1	79.4 ±0.2	89.8 ±0.1	7.8 ±0.4
Mixup	84.9 ±0.3	66.1 ±1.3	74.8 ±0.3	56.2 ±0.8	76.7 ±0.2	62.8 ±0.3	78.0 ±0.1	79.6 ±0.1	89.8 ±0.1	9.0 ±0.6
GroupDRO	83.1 ±0.3	69.5 ±0.7	73.2 ±0.2	52.0 ±0.6	75.7 ±0.2	62.2 ±0.2	79.1 ±0.1	80.9 ±0.1	89.9 ±0.1	12.3 ±0.5
CVaRDRO	83.5 ±0.3	68.7 ±1.3	73.5 ±0.3	52.8 ±0.6	75.9 ±0.2	62.4 ±0.2	78.6 ±0.2	80.7 ±0.1	89.8 ±0.1	32.9 ±0.4
JTT	83.3 ±0.1	64.3 ±1.5	72.8 ±0.1	52.4 ±0.3	74.8 ±0.1	60.3 ±0.2	76.8 ±0.2	78.4 ±0.2	88.2 ±0.1	10.2 ±0.3
LfF	68.2 ±5.0	50.3 ±5.9	62.9 ±3.5	34.1 ±5.4	61.5 ±4.7	45.8 ±5.5	68.5 ±4.4	69.8 ±4.8	75.0 ±6.5	30.8 ±2.7
LISA	84.9 ±0.3	66.1 ±1.3	74.8 ±0.3	56.2 ±0.8	76.7 ±0.2	62.8 ±0.3	78.0 ±0.1	79.6 ±0.1	89.8 ±0.1	9.0 ±0.6
ReSample	82.5 ±0.6	68.2 ±0.7	72.7 ±0.5	51.0 ±1.3	75.0 ±0.5	61.2 ±0.5	78.4 ±0.1	80.3 ±0.1	89.3 ±0.1	13.8 ±1.0
ReWeight	83.1 ±0.1	69.9 ±0.6	73.2 ±0.1	52.0 ±0.1	75.6 ±0.1	62.1 ±0.1	78.8 ±0.1	80.7 ±0.1	89.8 ±0.0	11.0 ±0.2
SqrtReWeight	83.6 ±0.1	70.1 ±0.3	73.6 ±0.1	52.9 ±0.2	76.0 ±0.1	62.4 ±0.2	78.7 ±0.2	80.7 ±0.2	89.9 ±0.1	10.1 ±0.1
CBLoss	84.1 ±0.7	67.0 ±0.1	74.1 ±0.7	54.3 ±1.6	76.2 ±0.5	62.6 ±0.5	78.4 ±0.1	80.2 ±0.2	90.0 ±0.0	9.1 ±1.0
Focal	85.6 ±0.3	61.9 ±1.1	75.6 ±0.4	58.5 ±0.9	77.0 ±0.3	62.9 ±0.5	77.3 ±0.3	78.7 ±0.3	89.4 ±0.4	7.7 ±0.4
LDAM	81.8 ±2.2	37.0 ±7.9	69.4 ±3.4	49.7 ±5.8	69.4 ±3.3	49.9 ±5.2	67.1 ±3.9	69.5 ±3.2	79.0 ±3.8	21.0 ±0.2
BSoftmax	83.0 ±0.4	69.4 ±1.2	73.1 ±0.3	51.8 ±0.7	75.5 ±0.3	61.9 ±0.3	78.7 ±0.3	80.6 ±0.3	89.7 ±0.2	12.1 ±0.9
DFR	81.3 ±0.0	66.5 ±0.2	71.2 ±0.0	48.6 ±0.0	73.4 ±0.0	59.0 ±0.0	76.8 ±0.0	78.8 ±0.0	86.7 ±0.1	19.6 ±0.1
CRT	83.0 ±0.0	68.5 ±0.0	73.0 ±0.0	51.7 ±0.1	75.4 ±0.0	61.8 ±0.0	78.6 ±0.0	80.6 ±0.0	89.4 ±0.1	12.5 ±0.1
ReWeightCRT	83.4 ±0.0	68.2 ±0.4	73.3 ±0.0	52.5 ±0.1	75.7 ±0.0	62.0 ±0.0	78.3 ±0.0	80.4 ±0.0	89.4 ±0.0	11.5 ±0.2

E.2.4. MultiNLI

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	80.9 ±0.3	69.5 ±0.3	81.3 ±0.3	73.7 ±0.4	81.0 ±0.3	77.6 ±0.3	79.7 ±0.3	80.9 ±0.3	93.6 ±0.1	10.7 ±0.5
Mixup	81.4 ±0.3	68.5 ±0.6	81.6 ±0.3	76.0 ±0.5	81.4 ±0.3	78.0 ±0.2	80.1 ±0.3	81.4 ±0.3	93.6 ±0.1	9.4 ±0.9
GroupDRO	81.0 ±0.0	69.3 ±1.5	81.3 ±0.1	74.9 ±0.6	81.1 ±0.1	77.7 ±0.1	79.5 ±0.3	81.0 ±0.0	93.8 ±0.0	8.7 ±0.4
CVaRDRO	75.1 ±0.1	63.0 ±1.5	76.2 ±0.2	65.6 ±0.2	75.2 ±0.1	72.1 ±0.2	74.2 ±0.4	75.1 ±0.1	86.3 ±0.2	41.4 ±0.1
JTT	81.4 ±0.1	68.4 ±0.6	81.6 ±0.1	75.7 ±0.3	81.5 ±0.1	78.1 ±0.1	80.2 ±0.2	81.4 ±0.1	93.9 ±0.0	9.4 ±0.5
LfF	71.7 ±1.1	63.6 ±2.9	71.8 ±1.1	68.7 ±0.7	71.7 ±1.1	68.5 ±1.8	70.8 ±1.4	71.7 ±1.1	87.0 ±0.8	4.4 ±0.6
LISA	81.4 ±0.3	68.5 ±0.6	81.6 ±0.3	76.0 ±0.5	81.4 ±0.3	78.0 ±0.2	80.1 ±0.3	81.4 ±0.3	93.6 ±0.1	9.4 ±0.9
ReSample	81.4 ±0.3	67.5 ±0.4	81.7 ±0.3	74.7 ±0.6	81.4 ±0.3	77.8 ±0.3	79.9 ±0.2	81.4 ±0.3	93.8 ±0.1	11.3 ±1.3
ReWeight	79.2 ±0.4	67.8 ±1.2	79.5 ±0.3	73.0 ±0.5	79.3 ±0.4	75.7 ±0.3	78.4 ±0.2	79.2 ±0.4	92.5 ±0.1	13.1 ±1.5
SqrtReWeight	80.9 ±0.1	66.6 ±0.4	81.1 ±0.1	76.0 ±0.3	80.9 ±0.1	77.7 ±0.1	79.6 ±0.0	80.9 ±0.1	93.6 ±0.1	8.1 ±0.2
CBLoss	81.1 ±0.2	66.2 ±0.7	81.2 ±0.2	76.6 ±0.3	81.1 ±0.2	77.8 ±0.1	79.7 ±0.0	81.1 ±0.2	93.7 ±0.1	8.9 ±0.4
Focal	80.7 ±0.2	69.3 ±0.8	81.2 ±0.2	73.5 ±0.5	80.8 ±0.2	77.4 ±0.2	79.5 ±0.1	80.7 ±0.2	93.6 ±0.1	4.5 ±1.1
LDAM	80.7 ±0.3	69.6 ±1.6	81.1 ±0.1	73.9 ±0.9	80.8 ±0.2	77.4 ±0.2	79.7 ±0.3	80.7 ±0.3	93.5 ±0.1	33.4 ±0.3
BSoftmax	80.9 ±0.1	66.9 ±0.4	81.1 ±0.1	75.9 ±0.3	80.9 ±0.1	77.7 ±0.1	79.7 ±0.0	80.9 ±0.1	93.6 ±0.1	8.1 ±0.2
DFR	80.2 ±0.0	63.8 ±0.0	80.3 ±0.0	75.2 ±0.0	80.3 ±0.0	76.2 ±0.0	78.5 ±0.0	80.2 ±0.0	92.9 ±0.0	5.7 ±0.0
CRT	80.2 ±0.0	65.4 ±0.1	80.3 ±0.0	74.4 ±0.1	80.2 ±0.0	76.4 ±0.0	78.6 ±0.0	80.2 ±0.0	92.9 ±0.0	14.9 ±0.1
ReWeightCRT	80.2 ±0.0	65.3 ±0.1	80.3 ±0.0	74.4 ±0.0	80.2 ±0.0	76.4 ±0.0	78.6 ±0.0	80.2 ±0.0	92.9 ±0.0	15.0 ±0.2

E.2.5. MetaShift

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	91.5 ±0.2	82.1 ±0.8	91.5 ±0.1	90.7 ±0.3	91.4 ±0.2	90.9 ±0.2	89.4 ±0.3	91.4 ±0.2	97.5 ±0.1	5.7 ±0.4
Mixup	91.4 ±0.2	79.0 ±0.8	91.4 ±0.2	90.9 ±0.2	91.3 ±0.2	90.8 ±0.2	88.7 ±0.1	91.3 ±0.2	97.2 ±0.0	1.7 ±0.2
GroupDRO	91.5 ±0.3	82.6 ±1.1	91.5 ±0.3	90.8 ±0.3	91.5 ±0.3	91.0 ±0.3	89.5 ±0.4	91.5 ±0.3	97.5 ±0.1	5.9 ±0.5
CVaRDRO	91.5 ±0.2	82.6 ±1.1	91.5 ±0.2	90.7 ±0.5	91.5 ±0.2	91.0 ±0.2	89.5 ±0.4	91.5 ±0.2	97.5 ±0.1	7.8 ±2.1
JTT	91.5 ±0.2	82.6 ±0.4	91.5 ±0.2	91.0 ±0.1	91.5 ±0.2	90.9 ±0.3	89.6 ±0.1	91.4 ±0.3	97.6 ±0.1	6.5 ±0.1
LfF	80.3 ±0.4	72.6 ±1.2	80.6 ±0.3	77.5 ±1.4	80.2 ±0.3	78.9 ±0.4	80.4 ±0.6	80.2 ±0.2	90.7 ±0.6	8.0 ±1.3
LISA	91.4 ±0.2	79.0 ±0.8	91.4 ±0.2	90.9 ±0.2	91.3 ±0.2	90.8 ±0.2	88.7 ±0.1	91.3 ±0.2	97.2 ±0.0	1.7 ±0.2
ReSample	92.1 ±0.3	80.5 ±1.5	92.1 ±0.3	91.4 ±0.4	92.1 ±0.3	91.6 ±0.3	89.5 ±0.1	92.1 ±0.3	97.5 ±0.1	6.7 ±0.4
ReWeight	91.2 ±0.5	83.1 ±0.7	91.1 ±0.5	90.2 ±0.4	91.1 ±0.5	90.6 ±0.6	89.3 ±0.5	91.1 ±0.5	97.4 ±0.2	6.6 ±0.8
SqrtReWeight	91.1 ±0.2	82.1 ±0.8	91.1 ±0.2	90.5 ±0.2	91.1 ±0.2	90.5 ±0.2	89.1 ±0.3	91.1 ±0.2	97.4 ±0.1	6.7 ±0.3
CBLoss	91.2 ±0.1	82.6 ±0.4	91.2 ±0.1	90.7 ±0.2	91.2 ±0.1	90.6 ±0.1	89.3 ±0.2	91.1 ±0.1	97.4 ±0.0	6.7 ±0.1
Focal	91.6 ±0.2	81.0 ±0.4	91.6 ±0.3	91.2 ±0.4	91.6 ±0.2	91.0 ±0.2	89.5 ±0.2	91.5 ±0.2	97.6 ±0.0	3.3 ±0.4
LDAM	91.7 ±0.0	83.6 ±0.4	91.7 ±0.0	90.9 ±0.3	91.7 ±0.0	91.2 ±0.0	90.0 ±0.1	91.7 ±0.0	97.5 ±0.1	9.9 ±0.7
BSoftmax	91.2 ±0.3	82.6 ±0.4	91.1 ±0.3	90.2 ±0.2	91.2 ±0.3	90.7 ±0.3	89.2 ±0.3	91.2 ±0.3	97.4 ±0.1	6.8 ±0.3
DFR	90.5 ±0.4	81.5 ±0.0	90.5 ±0.4	89.2 ±0.3	90.5 ±0.4	90.0 ±0.4	88.2 ±0.3	90.6 ±0.4	96.7 ±0.0	3.2 ±0.2
CRT	91.5 ±0.0	83.1 ±0.0	91.4 ±0.0	90.6 ±0.1	91.4 ±0.0	90.9 ±0.0	89.5 ±0.0	91.4 ±0.0	97.3 ±0.0	6.8 ±0.0
ReWeightCRT	91.3 ±0.1	85.1 ±0.4	91.2 ±0.1	90.1 ±0.3	91.2 ±0.1	90.8 ±0.1	89.7 ±0.2	91.3 ±0.1	96.8 ±0.1	8.1 ±0.1

E.2.6. ImageNetBG

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	87.7 ±0.1	76.8 ±0.9	87.9 ±0.1	78.3 ±1.3	87.7 ±0.1	80.7 ±0.4	87.7 ±0.1	87.7 ±0.1	99.0 ±0.0	5.1 ±0.1
Mixup	87.9 ±0.2	76.9 ±0.7	88.4 ±0.1	76.6 ±2.6	88.0 ±0.2	80.5 ±1.0	87.9 ±0.2	87.9 ±0.2	98.7 ±0.0	4.7 ±1.6
GroupDRO	87.7 ±0.1	76.4 ±0.2	87.9 ±0.1	76.2 ±0.5	87.6 ±0.1	81.1 ±0.3	87.7 ±0.1	87.7 ±0.1	99.0 ±0.0	5.1 ±0.1
CVaRDRO	87.8 ±0.2	74.8 ±0.8	88.0 ±0.1	79.9 ±0.9	87.8 ±0.2	79.5 ±0.3	87.8 ±0.2	87.8 ±0.2	99.0 ±0.0	5.6 ±0.2
JTT	87.6 ±0.4	77.0 ±0.4	87.8 ±0.3	78.3 ±3.0	87.5 ±0.3	80.4 ±0.6	87.6 ±0.4	87.6 ±0.4	99.0 ±0.0	3.7 ±0.2
LfF	84.7 ±0.5	70.1 ±1.4	85.4 ±0.3	72.1 ±3.1	84.7 ±0.5	76.2 ±0.6	84.7 ±0.5	84.7 ±0.5	98.6 ±0.0	1.8 ±0.4
LISA	87.9 ±0.2	76.9 ±0.7	88.4 ±0.1	76.6 ±2.6	88.0 ±0.2	80.5 ±1.0	87.9 ±0.2	87.9 ±0.2	98.7 ±0.0	4.7 ±1.6
ReSample	88.2 ±0.4	77.7 ±1.1	88.4 ±0.4	79.7 ±1.0	88.2 ±0.4	80.6 ±1.0	88.2 ±0.4	88.2 ±0.4	99.0 ±0.0	5.4 ±0.5
ReWeight	87.7 ±0.1	76.8 ±0.9	87.9 ±0.1	78.3 ±1.3	87.7 ±0.1	80.7 ±0.4	87.7 ±0.1	87.7 ±0.1	99.0 ±0.0	5.1 ±0.1
SqrtReWeight	87.7 ±0.1	76.8 ±0.9	87.9 ±0.1	78.3 ±1.3	87.7 ±0.1	80.7 ±0.4	87.7 ±0.1	87.7 ±0.1	99.0 ±0.0	5.1 ±0.1
CBLoss	87.7 ±0.1	76.8 ±0.9	87.9 ±0.1	78.3 ±1.3	87.7 ±0.1	80.7 ±0.4	87.7 ±0.1	87.7 ±0.1	99.0 ±0.0	5.1 ±0.1
Focal	86.7 ±0.2	71.9 ±1.2	87.1 ±0.2	74.2 ±1.3	86.6 ±0.2	77.6 ±0.7	86.7 ±0.2	86.7 ±0.2	98.9 ±0.0	2.8 ±0.6
LDAM	88.2 ±0.1	76.7 ±0.5	88.5 ±0.1	77.6 ±1.1	88.1 ±0.1	81.3 ±0.4	88.2 ±0.1	88.2 ±0.1	98.8 ±0.0	45.9 ±0.7
BSoftmax	87.7 ±0.1	76.1 ±2.0	88.0 ±0.1	77.7 ±1.3	87.7 ±0.1	80.4 ±0.8	87.7 ±0.1	87.7 ±0.1	99.0 ±0.0	5.6 ±0.5
DFR	86.8 ±0.5	74.4 ±1.8	86.9 ±0.5	78.9 ±1.6	86.7 ±0.5	78.1 ±1.3	86.8 ±0.5	86.8 ±0.5	98.8 ±0.1	8.9 ±1.4
CRT	88.3 ±0.1	78.2 ±0.5	88.3 ±0.1	82.7 ±0.4	88.3 ±0.1	80.9 ±0.2	88.3 ±0.1	88.3 ±0.1	99.1 ±0.0	5.6 ±0.2
ReWeightCRT	88.4 ±0.1	77.5 ±0.7	88.5 ±0.1	82.1 ±0.4	88.4 ±0.1	81.2 ±0.3	88.4 ±0.1	88.4 ±0.1	99.1 ±0.0	5.4 ±0.2

E.2.7. NICO++

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	85.3 ±0.3	40.0 ±0.0	86.2 ±0.2	59.2 ±1.1	85.4 ±0.3	66.0 ±0.6	84.8 ±0.3	85.2 ±0.3	99.4 ±0.0	9.6 ±0.2
Mixup	85.5 ±0.2	30.0 ±4.1	86.5 ±0.1	55.7 ±2.5	85.7 ±0.2	66.5 ±1.6	85.0 ±0.2	85.4 ±0.2	99.1 ±0.0	2.0 ±0.3
GroupDRO	83.0 ±0.3	31.1 ±0.9	84.5 ±0.2	54.4 ±3.9	83.2 ±0.3	64.9 ±0.9	82.6 ±0.3	82.9 ±0.3	99.3 ±0.0	7.4 ±0.5
CVaRDRO	84.4 ±0.6	31.7 ±3.6	85.5 ±0.5	57.1 ±1.6	84.6 ±0.6	66.6 ±1.5	84.0 ±0.6	84.3 ±0.6	99.4 ±0.0	7.8 ±1.3
JTT	85.4 ±0.2	32.2 ±0.9	86.2 ±0.2	57.8 ±4.6	85.5 ±0.2	65.1 ±2.8	84.9 ±0.2	85.3 ±0.2	99.4 ±0.0	10.2 ±0.2
LfF	78.5 ±0.6	28.3 ±1.7	80.9 ±0.3	44.3 ±0.8	78.8 ±0.6	54.4 ±1.3	78.1 ±0.6	78.4 ±0.6	99.2 ±0.0	1.8 ±0.2
LISA	85.5 ±0.2	30.0 ±4.1	86.5 ±0.1	55.7 ±2.5	85.7 ±0.2	66.5 ±1.6	85.0 ±0.2	85.4 ±0.2	99.1 ±0.0	2.0 ±0.3
ReSample	84.8 ±0.4	23.3 ±1.4	85.7 ±0.3	58.9 ±2.7	84.9 ±0.4	65.5 ±1.1	84.2 ±0.4	84.7 ±0.4	99.3 ±0.0	10.1 ±0.0
ReWeight	85.8 ±0.1	25.0 ±0.0	86.6 ±0.1	59.8 ±1.1	85.9 ±0.1	69.1 ±0.5	85.3 ±0.1	85.7 ±0.1	99.4 ±0.0	9.4 ±0.2
SqrtReWeight	85.4 ±0.2	35.6 ±1.8	86.4 ±0.1	57.5 ±1.4	85.6 ±0.1	66.7 ±1.2	84.9 ±0.2	85.3 ±0.2	99.4 ±0.0	9.3 ±0.0
CBLoss	85.1 ±0.4	34.4 ±2.4	85.9 ±0.3	56.9 ±1.1	85.2 ±0.4	64.9 ±0.1	84.7 ±0.4	85.1 ±0.4	99.4 ±0.0	9.1 ±0.7
Focal	85.1 ±0.5	29.4 ±2.0	86.0 ±0.3	58.1 ±2.2	85.3 ±0.4	65.0 ±0.8	84.6 ±0.5	85.0 ±0.5	99.4 ±0.0	4.6 ±1.1
LDAM	84.7 ±0.4	26.7 ±1.4	85.6 ±0.3	60.4 ±1.4	84.8 ±0.4	65.3 ±0.7	84.2 ±0.4	84.6 ±0.4	98.9 ±0.1	62.6 ±2.0
BSoftmax	85.2 ±0.3	29.4 ±2.0	85.9 ±0.2	57.6 ±2.1	85.3 ±0.3	66.9 ±0.7	84.8 ±0.3	85.1 ±0.3	99.4 ±0.0	8.4 ±0.9
DFR	82.5 ±0.0	39.3 ±2.4	83.3 ±0.0	55.3 ±1.3	82.6 ±0.0	63.7 ±0.9	82.1 ±0.0	82.4 ±0.0	99.2 ±0.0	11.7 ±0.1
CRT	85.7 ±0.0	33.3 ±0.0	86.1 ±0.0	64.4 ±0.2	85.7 ±0.0	69.1 ±0.1	85.2 ±0.0	85.6 ±0.0	99.4 ±0.0	4.7 ±0.0
ReWeightCRT	85.8 ±0.1	33.3 ±0.0	86.1 ±0.1	66.1 ±0.5	85.8 ±0.1	69.9 ±0.2	85.4 ±0.1	85.7 ±0.1	99.4 ±0.0	6.2 ±0.2

E.2.8. MIMIC-CXR

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	78.6 ±0.2	68.5 ±0.4	77.7 ±0.2	72.2 ±0.5	77.8 ±0.2	73.7 ±0.2	77.5 ±0.2	78.1 ±0.2	85.5 ±0.2	2.6 ±0.2
Mixup	78.5 ±0.1	67.2 ±1.1	77.5 ±0.1	72.3 ±0.5	77.6 ±0.0	73.4 ±0.1	77.2 ±0.1	77.8 ±0.1	85.3 ±0.1	2.4 ±0.7
GroupDRO	78.3 ±0.1	67.4 ±0.9	77.4 ±0.1	71.9 ±0.2	77.6 ±0.1	73.4 ±0.1	77.2 ±0.1	77.8 ±0.1	85.1 ±0.0	3.7 ±0.4
CVaRDRO	78.4 ±0.3	67.5 ±0.1	77.4 ±0.3	72.1 ±0.6	77.6 ±0.3	73.4 ±0.2	77.2 ±0.2	77.7 ±0.2	84.9 ±0.2	5.7 ±0.3
JTT	78.2 ±0.1	66.6 ±0.8	77.3 ±0.1	71.9 ±0.3	77.4 ±0.1	73.2 ±0.1	77.0 ±0.1	77.6 ±0.1	85.0 ±0.1	3.3 ±0.1
LfF	73.9 ±1.1	62.1 ±2.4	72.9 ±1.1	66.5 ±1.1	73.0 ±1.2	68.0 ±1.6	72.9 ±1.2	73.2 ±1.2	79.9 ±1.5	11.3 ±0.3
LISA	78.5 ±0.1	67.2 ±1.1	77.5 ±0.1	72.3 ±0.5	77.6 ±0.0	73.4 ±0.1	77.2 ±0.1	77.8 ±0.1	85.3 ±0.1	2.4 ±0.7
ReSample	78.7 ±0.1	68.9 ±0.3	77.8 ±0.1	72.2 ±0.1	78.0 ±0.1	74.0 ±0.1	77.7 ±0.1	78.2 ±0.1	85.4 ±0.1	3.6 ±0.1
ReWeight	78.0 ±0.1	67.4 ±0.3	77.1 ±0.0	70.9 ±0.2	77.3 ±0.0	73.3 ±0.0	77.0 ±0.0	77.6 ±0.0	84.9 ±0.0	4.1 ±0.6
SqrtReWeight	78.5 ±0.2	68.9 ±0.5	77.6 ±0.2	71.6 ±0.3	77.7 ±0.2	73.8 ±0.2	77.4 ±0.2	78.1 ±0.2	85.4 ±0.1	3.6 ±0.4
CBLoss	78.6 ±0.1	67.8 ±0.6	77.6 ±0.1	72.3 ±0.4	77.8 ±0.1	73.6 ±0.2	77.4 ±0.1	78.0 ±0.1	85.4 ±0.1	3.2 ±0.3
Focal	78.3 ±0.1	67.3 ±1.2	77.4 ±0.1	71.6 ±0.1	77.5 ±0.1	73.5 ±0.2	77.2 ±0.2	77.8 ±0.1	85.3 ±0.1	10.0 ±0.6
LDAM	78.5 ±0.1	68.2 ±1.4	77.6 ±0.1	72.0 ±0.4	77.7 ±0.1	73.6 ±0.3	77.5 ±0.2	78.0 ±0.2	85.3 ±0.1	22.2 ±0.1
BSoftmax	78.2 ±0.2	67.2 ±0.2	77.3 ±0.2	71.9 ±0.5	77.4 ±0.2	73.2 ±0.1	77.1 ±0.1	77.6 ±0.1	85.1 ±0.1	3.6 ±0.5
DFR	78.3 ±0.0	67.1 ±0.4	77.3 ±0.0	72.0 ±0.2	77.5 ±0.0	73.2 ±0.1	77.1 ±0.0	77.6 ±0.1	85.0 ±0.0	20.0 ±0.1
CRT	78.3 ±0.0	69.1 ±0.2	77.4 ±0.0	71.1 ±0.1	77.6 ±0.0	73.7 ±0.0	77.4 ±0.0	78.0 ±0.0	85.3 ±0.0	4.7 ±0.1
ReWeightCRT	77.9 ±0.0	68.9 ±0.0	77.0 ±0.0	70.5 ±0.0	77.3 ±0.0	73.4 ±0.0	77.0 ±0.0	77.7 ±0.0	85.0 ±0.0	4.9 ±0.3

E.2.9. MIMICNOTES

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	91.1 ±0.1	24.2 ±2.8	76.4 ±0.9	60.3 ±2.1	65.8 ±1.2	36.5 ±2.5	62.2 ±1.3	62.4 ±1.4	85.2 ±0.1	2.2 ±0.1
Mixup	91.1 ±0.0	22.7 ±3.2	76.8 ±0.7	61.2 ±1.6	65.1 ±1.6	35.0 ±3.2	61.5 ±1.6	61.7 ±1.7	85.4 ±0.0	2.0 ±0.8
GroupDRO	83.2 ±2.4	62.6 ±6.3	64.7 ±1.4	33.7 ±3.4	66.4 ±1.3	42.8 ±1.1	74.3 ±1.5	74.4 ±1.5	85.1 ±0.1	13.8 ±3.2
CVaRDRO	90.2 ±0.0	0.0 ±0.0	45.1 ±0.0	0.0 ±0.0	47.4 ±0.0	0.0 ±0.0	50.0 ±0.0	50.0 ±0.0	71.6 ±0.4	40.2 ±0.0
JTT	71.3 ±3.7	65.9 ±2.8	60.3 ±0.7	23.4 ±1.9	58.6 ±2.3	36.0 ±1.8	75.5 ±0.4	75.6 ±0.4	84.9 ±0.1	27.5 ±3.9
LfF	84.0 ±1.2	62.7 ±2.1	64.6 ±0.7	33.6 ±1.6	67.1 ±0.8	43.6 ±0.8	74.7 ±0.4	74.7 ±0.5	85.1 ±0.0	12.5 ±1.2
LISA	91.1 ±0.0	22.7 ±3.2	76.8 ±0.7	61.2 ±1.6	65.1 ±1.6	35.0 ±3.2	61.5 ±1.6	61.7 ±1.7	85.4 ±0.0	2.0 ±0.8
ReSample	81.4 ±1.5	67.1 ±2.6	63.3 ±0.6	30.5 ±1.5	65.4 ±1.0	42.0 ±1.0	75.4 ±0.3	75.6 ±0.4	85.1 ±0.0	17.4 ±1.9
ReWeight	82.7 ±0.7	65.5 ±1.3	63.8 ±0.4	31.7 ±0.9	66.3 ±0.5	42.8 ±0.5	75.3 ±0.2	75.4 ±0.3	85.2 ±0.1	15.9 ±0.7
SqrtReWeight	90.3 ±0.2	35.7 ±4.0	72.4 ±0.9	51.3 ±2.1	68.7 ±0.9	42.8 ±2.1	66.7 ±1.6	66.8 ±1.6	85.2 ±0.1	3.7 ±0.7
CBLoss	78.2 ±1.0	72.3 ±1.3	61.9 ±0.3	27.3 ±0.8	63.2 ±0.7	39.8 ±0.6	76.1 ±0.2	76.2 ±0.2	85.0 ±0.0	20.6 ±1.3
Focal	91.0 ±0.0	19.1 ±2.3	77.1 ±0.6	62.1 ±1.4	63.6 ±1.3	31.9 ±2.6	59.9 ±1.1	60.2 ±1.1	85.3 ±0.1	8.1 ±0.7
LDAM	90.6 ±0.1	5.3 ±2.4	84.4 ±0.8	78.1 ±1.7	52.5 ±2.1	10.0 ±4.1	52.7 ±1.2	52.7 ±1.2	84.9 ±0.1	28.9 ±1.0
BSoftmax	76.9 ±0.9	73.1 ±1.0	61.7 ±0.2	26.5 ±0.6	62.5 ±0.5	39.3 ±0.4	76.6 ±0.2	76.7 ±0.2	85.4 ±0.0	23.5 ±1.1
DFR	69.2 ±1.3	67.3 ±1.7	58.8 ±0.2	21.0 ±0.5	56.5 ±0.8	33.1 ±0.5	73.1 ±0.0	73.1 ±0.0	81.0 ±0.0	38.4 ±0.1
CRT	77.8 ±0.0	73.1 ±0.0	61.6 ±0.0	26.7 ±0.0	62.8 ±0.0	39.2 ±0.0	75.9 ±0.0	75.9 ±0.0	84.3 ±0.0	23.0 ±0.1
ReWeightCRT	81.2 ±2.8	63.9 ±7.6	63.6 ±1.6	31.4 ±3.9	64.7 ±1.6	40.7 ±1.3	73.8 ±1.6	73.9 ±1.6	84.3 ±0.0	26.5 ±2.4

E.2.10. CXRMULTISITE

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	98.3 ±0.0	0.0 ±0.0	49.2 ±0.0	0.0 ±0.0	49.6 ±0.0	0.0 ±0.0	50.0 ±0.0	50.0 ±0.0	93.1 ±0.1	0.3 ±0.1
Mixup	98.3 ±0.0	0.0 ±0.0	49.2 ±0.0	0.0 ±0.0	49.6 ±0.0	0.0 ±0.0	50.0 ±0.0	50.0 ±0.0	92.9 ±0.1	0.3 ±0.0
GroupDRO	90.2 ±0.1	0.0 ±0.0	56.7 ±0.1	13.6 ±0.1	59.2 ±0.1	23.7 ±0.2	50.2 ±0.1	90.4 ±0.0	92.8 ±0.2	13.5 ±0.7
CVaRDRO	98.3 ±0.0	0.0 ±0.0	61.2 ±4.9	24.0 ±9.8	50.7 ±0.7	2.2 ±1.5	50.2 ±0.2	50.6 ±0.4	93.0 ±0.0	0.9 ±0.3
JTT	94.1 ±0.9	0.0 ±0.0	59.0 ±0.7	18.5 ±1.4	62.9 ±0.8	28.9 ±1.2	55.2 ±0.9	82.2 ±2.4	93.2 ±0.1	6.4 ±0.5
LfF	9.9 ±6.7	5.4 ±4.4	17.4 ±13.5	0.6 ±0.5	8.5 ±5.6	1.2 ±1.0	50.5 ±0.4	51.7 ±1.4	60.6 ±1.6	82.6 ±12.8
LISA	98.3 ±0.0	0.0 ±0.0	49.2 ±0.0	0.0 ±0.0	49.6 ±0.0	0.0 ±0.0	50.0 ±0.0	50.0 ±0.0	92.9 ±0.1	0.3 ±0.0
ReSample	88.3 ±1.6	0.1 ±0.1	55.9 ±0.6	12.0 ±1.2	57.3 ±1.4	20.9 ±1.8	50.3 ±0.5	88.1 ±0.7	92.3 ±0.1	13.0 ±2.8
ReWeight	89.5 ±0.0	0.3 ±0.1	56.4 ±0.0	13.0 ±0.0	58.5 ±0.0	22.7 ±0.0	50.5 ±0.2	90.3 ±0.0	93.2 ±0.1	17.7 ±1.7
SqrtReWeight	94.5 ±0.4	0.0 ±0.0	59.4 ±0.2	19.3 ±0.6	63.7 ±0.3	30.2 ±0.4	56.3 ±0.1	82.3 ±1.6	93.3 ±0.0	6.0 ±0.2
CBLoss	1.7 ±0.0	0.0 ±0.0	0.8 ±0.0	0.0 ±0.0	1.7 ±0.0	0.0 ±0.0	50.0 ±0.0	50.0 ±0.0	62.7 ±1.5	98.3 ±0.0
Focal	98.3 ±0.0	0.0 ±0.0	55.4 ±5.1	12.5 ±10.2	49.7 ±0.1	0.3 ±0.2	50.0 ±0.0	50.1 ±0.1	93.2 ±0.0	11.5 ±0.6
LDAM	98.3 ±0.0	0.0 ±0.0	49.2 ±0.0	0.0 ±0.0	49.6 ±0.0	0.0 ±0.0	50.0 ±0.0	50.0 ±0.0	93.1 ±0.1	0.3 ±0.1
BSoftmax	89.1 ±0.2	0.5 ±0.1	56.2 ±0.1	12.5 ±0.2	58.1 ±0.2	22.0 ±0.3	50.4 ±0.0	90.0 ±0.1	92.9 ±0.1	19.9 ±1.3
DFR	89.7 ±0.1	0.6 ±0.1	56.5 ±0.0	13.2 ±0.1	58.7 ±0.1	23.0 ±0.1	50.4 ±0.0	90.4 ±0.0	92.8 ±0.1	47.3 ±0.1
CRT	90.4 ±0.1	1.1 ±0.5	56.9 ±0.0	13.9 ±0.1	59.5 ±0.1	24.2 ±0.1	51.2 ±0.1	90.2 ±0.1	93.3 ±0.0	15.7 ±0.9
ReWeightCRT	89.9 ±0.1	1.4 ±0.6	56.6 ±0.0	13.4 ±0.1	59.0 ±0.1	23.3 ±0.1	51.1 ±0.3	90.4 ±0.0	93.1 ±0.1	15.5 ±0.7

E.2.11. CHEXPERT

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	88.6 ±0.7	41.7 ±3.4	68.3 ±1.1	42.3 ±2.4	68.7 ±0.2	43.8 ±0.7	69.5 ±1.3	70.0 ±1.5	85.4 ±0.4	5.0 ±1.1
Mixup	81.9 ±6.2	37.4 ±3.5	63.5 ±5.0	33.9 ±9.3	62.5 ±5.9	35.7 ±7.6	63.8 ±4.7	64.1 ±4.6	76.1 ±8.5	16.1 ±9.0
GroupDRO	79.0 ±0.3	75.1 ±0.6	62.9 ±0.1	28.5 ±0.3	64.6 ±0.2	42.0 ±0.3	78.6 ±0.0	79.3 ±0.1	86.3 ±0.1	22.0 ±0.4
CVaRDRO	73.7 ±1.0	50.2 ±1.8	57.3 ±0.1	20.1 ±0.3	56.8 ±0.4	29.9 ±0.2	65.7 ±0.6	67.0 ±0.4	72.9 ±0.4	40.4 ±0.0
JTT	75.2 ±0.8	60.4 ±4.8	59.4 ±1.1	23.0 ±1.5	59.6 ±1.4	34.4 ±2.3	70.7 ±2.6	72.0 ±2.5	79.0 ±2.5	24.4 ±0.5
LfF	22.3 ±10.2	13.7 ±9.8	37.3 ±5.8	9.0 ±0.7	19.5 ±8.3	8.8 ±3.9	46.2 ±2.9	46.2 ±3.1	30.5 ±10.1	65.7 ±10.2
LISA	81.9 ±6.2	37.4 ±3.5	63.5 ±5.0	33.9 ±9.3	62.5 ±5.9	35.7 ±7.6	63.8 ±4.7	64.1 ±4.6	76.1 ±8.5	16.1 ±9.0
ReSample	77.6 ±0.4	73.0 ±0.6	62.3 ±0.1	27.2 ±0.3	63.4 ±0.3	40.6 ±0.3	78.0 ±0.3	78.7 ±0.3	85.9 ±0.4	19.1 ±1.0
ReWeight	79.2 ±0.5	73.8 ±1.0	62.9 ±0.3	28.6 ±0.5	64.6 ±0.4	42.0 ±0.5	78.5 ±0.3	79.0 ±0.2	86.2 ±0.1	20.6 ±0.4
SqrtReWeight	83.5 ±0.3	68.5 ±1.6	65.0 ±0.2	33.3 ±0.4	67.9 ±0.2	45.5 ±0.3	77.8 ±0.4	78.5 ±0.3	86.3 ±0.3	15.6 ±1.1
CBLoss	80.0 ±0.5	74.0 ±0.7	63.3 ±0.2	29.4 ±0.5	65.2 ±0.4	42.6 ±0.5	78.6 ±0.2	79.0 ±0.2	86.1 ±0.3	19.6 ±0.6
Focal	89.3 ±0.3	42.1 ±4.0	69.6 ±0.4	44.7 ±1.1	69.8 ±0.4	45.5 ±1.0	70.4 ±1.1	70.4 ±1.3	86.5 ±0.1	16.1 ±1.7
LDAM	90.1 ±0.0	34.5 ±1.5	70.6 ±0.1	47.6 ±0.1	68.6 ±0.3	42.5 ±0.7	66.8 ±0.5	67.0 ±0.5	85.5 ±0.3	31.7 ±0.4
BSoftmax	79.5 ±0.2	74.2 ±1.1	63.2 ±0.1	29.0 ±0.2	65.0 ±0.2	42.5 ±0.2	78.7 ±0.2	79.6 ±0.1	86.6 ±0.1	21.9 ±0.2
DFR	78.9 ±0.2	75.4 ±0.6	62.9 ±0.1	28.5 ±0.1	64.5 ±0.1	42.0 ±0.1	78.9 ±0.1	79.3 ±0.0	86.0 ±0.1	26.2 ±0.5
CRT	79.1 ±0.1	74.0 ±0.2	62.9 ±0.0	28.6 ±0.1	64.6 ±0.1	42.0 ±0.1	78.7 ±0.1	79.1 ±0.1	86.2 ±0.0	21.9 ±0.1
ReWeightCRT	79.0 ±0.0	73.9 ±0.2	62.9 ±0.0	28.5 ±0.0	64.5 ±0.0	41.9 ±0.0	78.8 ±0.1	79.2 ±0.1	86.3 ±0.1	22.3 ±0.1

E.2.12. LIVING17

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	27.7 ±1.1	5.7 ±1.5	28.2 ±1.0	8.2 ±1.6	27.1 ±1.1	6.9 ±1.7	27.7 ±1.1	27.7 ±1.1	77.3 ±1.3	59.6 ±0.5
Mixup	29.8 ±1.8	8.7 ±1.4	30.9 ±2.2	9.5 ±1.3	29.3 ±1.9	9.7 ±1.5	29.8 ±1.8	29.8 ±1.8	78.2 ±1.2	34.1 ±1.9
GroupDRO	31.1 ±1.0	6.0 ±1.4	32.1 ±0.9	9.6 ±0.2	30.8 ±0.7	7.6 ±0.8	31.1 ±1.0	31.1 ±1.0	80.1 ±0.8	53.5 ±0.8
CVaRDRO	27.3 ±1.6	4.0 ±0.5	29.2 ±1.7	5.1 ±0.8	26.5 ±1.5	4.8 ±0.6	27.3 ±1.6	27.3 ±1.6	81.0 ±0.2	28.8 ±6.8
JTT	28.3 ±1.1	5.7 ±2.2	31.1 ±0.9	8.0 ±1.7	28.3 ±1.3	7.2 ±2.3	28.3 ±1.1	28.3 ±1.1	81.0 ±0.8	36.9 ±3.2
LfF	26.4 ±1.3	7.0 ±1.2	28.3 ±0.8	9.6 ±1.8	26.1 ±1.2	8.7 ±1.7	26.4 ±1.3	26.4 ±1.3	76.6 ±0.6	61.0 ±0.7
LISA	29.8 ±1.8	8.7 ±1.4	30.9 ±2.2	9.5 ±1.3	29.3 ±1.9	9.7 ±1.5	29.8 ±1.8	29.8 ±1.8	78.2 ±1.2	34.1 ±1.9
ReSample	31.4 ±0.6	6.7 ±1.5	33.0 ±0.6	11.0 ±0.6	31.0 ±0.6	8.3 ±1.2	31.4 ±0.6	31.4 ±0.6	81.0 ±0.7	46.6 ±3.1
ReWeight	27.7 ±1.1	5.7 ±1.5	28.2 ±1.0	8.2 ±1.6	27.1 ±1.1	6.9 ±1.7	27.7 ±1.1	27.7 ±1.1	77.3 ±1.3	59.6 ±0.5
SqrtReWeight	27.7 ±1.1	5.7 ±1.5	28.2 ±1.0	8.2 ±1.6	27.1 ±1.1	6.9 ±1.7	27.7 ±1.1	27.7 ±1.1	77.3 ±1.3	59.6 ±0.5
CBLoss	27.7 ±1.1	5.7 ±1.5	28.2 ±1.0	8.2 ±1.6	27.1 ±1.1	6.9 ±1.7	27.7 ±1.1	27.7 ±1.1	77.3 ±1.3	59.6 ±0.5
Focal	26.9 ±0.6	5.3 ±0.3	28.8 ±1.0	7.1 ±1.0	27.0 ±0.7	6.3 ±0.5	26.9 ±0.6	26.9 ±0.6	78.7 ±0.5	49.9 ±1.9
LDAM	24.3 ±0.8	4.0 ±0.8	28.0 ±1.2	6.6 ±1.4	24.0 ±0.8	5.1 ±0.3	24.3 ±0.8	24.3 ±0.8	79.1 ±1.0	12.4 ±0.5
BSoftmax	28.6 ±1.4	6.7 ±0.7	30.7 ±1.0	8.2 ±0.8	28.3 ±1.3	7.3 ±0.5	28.6 ±1.4	28.6 ±1.4	78.0 ±1.0	56.5 ±1.7
DFR	26.3 ±0.4	6.0 ±0.9	27.4 ±0.3	8.6 ±0.8	25.7 ±0.2	7.5 ±1.1	26.3 ±0.4	26.3 ±0.4	79.4 ±0.1	13.8 ±0.4
CRT	31.1 ±0.1	6.3 ±0.3	31.8 ±0.0	7.5 ±0.3	30.5 ±0.1	6.8 ±0.3	31.1 ±0.1	31.1 ±0.1	80.3 ±0.1	49.6 ±1.7
ReWeightCRT	33.1 ±0.1	9.3 ±0.3	33.4 ±0.1	11.3 ±0.3	32.6 ±0.0	10.8 ±0.2	33.1 ±0.1	33.1 ±0.1	82.0 ±0.0	40.0 ±0.4

E.2.13. OVERALL

Algorithm	Waterbirds	CelebA	CivilComments	MultiNLI	MetaShift	ImageNetBG	NICO++	MIMIC-CXR	MIMICNotes	CXRMultisite	CheXpert	Living17	Avg
ERM	69.1 ±4.7	57.6 ±0.8	63.2 ±1.2	69.5 ±0.3	82.1 ±0.8	76.8 ±0.9	40.0 ±0.0	68.5 ±0.4	80.4 ±0.2	50.1 ±0.9	41.7 ±3.4	27.7 ±1.1	60.5
Mixup	78.2 ±0.4	57.8 ±0.8	66.1 ±1.3	68.5 ±0.6	79.0 ±0.8	76.9 ±0.7	30.0 ±4.1	67.2 ±1.1	81.6 ±0.6	50.1 ±0.9	37.4 ±3.5	29.8 ±1.8	60.2
GroupDRO	73.1 ±0.4	78.5 ±1.1	69.5 ±0.7	69.3 ±1.5	82.6 ±1.1	76.4 ±0.2	31.1 ±0.9	67.4 ±0.9	83.7 ±0.0	49.2 ±0.5	75.1 ±0.6	31.1 ±1.0	65.6
CVaRDRO	75.5 ±2.2	62.2 ±3.1	68.7 ±1.3	63.0 ±1.5	82.6 ±1.1	74.8 ±0.8	31.7 ±3.6	67.5 ±0.1	65.6 ±1.5	50.2 ±0.9	50.2 ±1.8	27.3 ±1.6	59.9
JTT	71.0 ±0.5	66.0 ±11.9	64.3 ±1.5	68.4 ±0.6	82.6 ±0.4	77.0 ±0.4	32.2 ±0.9	66.6 ±0.8	83.8 ±0.1	50.1 ±0.9	60.4 ±4.8	28.3 ±1.1	62.6
LfF	74.7 ±1.0	53.0 ±4.3	50.3 ±5.9	63.6 ±2.9	72.6 ±1.2	70.1 ±1.4	28.3 ±1.7	62.1 ±2.4	84.1 ±0.0	50.1 ±0.9	13.7 ±9.8	26.4 ±1.3	54.1
LISA	78.2 ±0.4	57.8 ±0.8	66.1 ±1.3	68.5 ±0.6	79.0 ±0.8	76.9 ±0.7	30.0 ±4.1	67.2 ±1.1	81.6 ±0.6	50.1 ±0.9	37.4 ±3.5	29.8 ±1.8	60.2
ReSample	70.0 ±1.0	82.2 ±1.2	68.2 ±0.7	67.5 ±0.4	80.5 ±1.5	77.7 ±1.1	23.3 ±1.4	68.9 ±0.3	82.4 ±0.5	47.8 ±2.5	73.0 ±0.6	31.4 ±0.6	64.4
ReWeight	72.5 ±0.3	81.5 ±0.9	69.9 ±0.6	67.8 ±1.2	83.1 ±0.7	76.8 ±0.9	25.0 ±0.0	67.4 ±0.3	84.0 ±0.1	51.9 ±2.3	73.8 ±1.0	27.7 ±1.1	65.1
SqrtReWeight	71.3 ±1.4	72.0 ±2.2	70.1 ±0.3	66.6 ±0.4	82.1 ±0.8	76.8 ±0.9	35.6 ±1.8	68.9 ±0.5	83.1 ±0.2	50.2 ±0.9	68.5 ±1.6	27.7 ±1.1	64.4
CBLoss	74.4 ±1.2	75.0 ±2.4	67.0 ±0.1	66.2 ±0.7	82.6 ±0.4	76.8 ±0.9	34.4 ±2.4	67.8 ±0.6	83.9 ±0.1	50.2 ±0.9	74.0 ±0.7	27.7 ±1.1	65.0
Focal	71.6 ±0.8	59.1 ±2.0	61.9 ±1.1	69.3 ±0.8	81.0 ±0.4	71.9 ±1.2	29.4 ±2.0	67.3 ±1.2	70.3 ±9.6	50.0 ±0.9	42.1 ±4.0	26.9 ±0.6	58.4
LDAM	71.0 ±1.8	59.3 ±2.3	37.0 ±7.9	69.6 ±1.6	83.6 ±0.4	76.7 ±0.5	26.7 ±1.4	68.2 ±1.4	81.0 ±0.3	50.1 ±0.9	34.5 ±1.5	24.3 ±0.8	56.8
BSoftmax	74.1 ±0.9	83.3 ±0.5	69.4 ±1.2	66.9 ±0.4	82.6 ±0.4	76.1 ±2.0	29.4 ±2.0	67.2 ±0.2	83.7 ±0.3	50.1 ±1.1	74.2 ±1.1	28.6 ±1.4	65.5
DFR	89.0 ±0.2	86.3 ±0.3	66.5 ±0.2	63.8 ±0.0	81.5 ±0.0	74.4 ±1.8	39.3 ±2.4	67.1 ±0.4	80.2 ±0.0	56.1 ±3.5	75.4 ±0.6	26.3 ±0.4	67.2
CRT	76.3 ±0.8	70.4 ±0.4	68.5 ±0.0	65.4 ±0.1	83.1 ±0.0	78.2 ±0.5	33.3 ±0.0	69.1 ±0.2	83.4 ±0.0	56.5 ±3.5	74.0 ±0.2	31.1 ±0.1	65.8
ReWeightCRT	76.3 ±0.2	71.1 ±0.5	68.2 ±0.4	65.3 ±0.1	85.1 ±0.4	77.5 ±0.7	33.3 ±0.0	68.9 ±0.0	83.4 ±0.0	55.4 ±4.9	73.9 ±0.2	33.1 ±0.1	66.0

E.3. Attributes Unknown in Both Training & Validation

E.3.1. WATERBIRDS

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	84.1 ±1.7	69.1 ±4.7	77.4 ±2.0	60.7 ±3.2	79.4 ±2.1	69.5 ±2.9	83.1 ±2.0	83.1 ±2.0	91.0 ±1.4	12.9 ±1.7
Mixup	89.2 ±0.6	77.5 ±0.7	83.5 ±0.8	70.6 ±1.9	85.6 ±0.7	78.4 ±0.9	88.9 ±0.2	88.9 ±0.2	94.9 ±0.1	8.0 ±1.1
GroupDRO	86.9 ±0.9	73.1 ±0.4	80.7 ±1.1	66.1 ±2.2	82.8 ±0.9	74.4 ±1.2	86.3 ±0.5	86.3 ±0.5	94.0 ±0.3	10.5 ±0.8
CVaRDRO	89.9 ±0.4	75.5 ±2.2	84.5 ±0.7	73.2 ±1.7	86.2 ±0.3	79.0 ±0.4	88.5 ±0.3	88.5 ±0.3	95.4 ±0.2	8.3 ±0.2
JTT	88.9 ±0.6	71.2 ±0.5	83.2 ±0.8	71.4 ±1.6	84.7 ±0.6	76.8 ±0.8	86.8 ±0.2	86.8 ±0.2	94.2 ±0.1	9.2 ±0.3
LfF	86.6 ±0.5	75.0 ±0.7	80.3 ±0.6	65.1 ±1.1	82.5 ±0.5	74.0 ±0.7	86.3 ±0.3	86.3 ±0.3	93.4 ±0.2	10.0 ±0.8
LISA	89.2 ±0.6	77.5 ±0.7	83.5 ±0.8	70.6 ±1.9	85.6 ±0.7	78.4 ±0.9	88.9 ±0.2	88.9 ±0.2	94.9 ±0.1	8.0 ±1.1
ReSample	86.2 ±0.5	70.0 ±1.0	79.8 ±0.6	64.9 ±1.4	81.7 ±0.5	72.7 ±0.7	85.0 ±0.2	85.0 ±0.2	92.8 ±0.1	11.3 ±0.3
ReWeight	86.2 ±0.6	71.9 ±0.6	79.9 ±0.7	64.3 ±1.6	82.1 ±0.6	73.5 ±0.7	86.2 ±0.1	86.2 ±0.1	94.0 ±0.1	10.8 ±0.4
SqrtReWeight	89.4 ±0.4	71.0 ±1.4	83.9 ±0.6	72.8 ±0.9	85.3 ±0.6	77.7 ±0.9	87.2 ±0.6	87.2 ±0.6	94.4 ±0.5	9.0 ±0.5
CBLoss	86.8 ±0.6	74.4 ±1.2	80.4 ±0.7	65.5 ±1.3	82.6 ±0.7	74.0 ±1.0	86.2 ±0.6	86.2 ±0.6	93.5 ±0.4	11.3 ±0.4
Focal	89.3 ±0.2	71.6 ±0.8	83.7 ±0.3	72.4 ±0.5	85.2 ±0.3	77.5 ±0.4	87.1 ±0.3	87.1 ±0.3	94.2 ±0.2	6.9 ±0.1
LDAM	87.9 ±0.2	70.9 ±1.7	81.9 ±0.3	69.1 ±0.8	83.6 ±0.1	75.2 ±0.1	86.0 ±0.2	86.0 ±0.2	93.5 ±0.1	11.8 ±1.7
BSoftmax	88.4 ±1.2	74.1 ±0.9	82.6 ±1.6	69.9 ±2.9	84.4 ±1.5	76.4 ±2.0	87.0 ±1.0	87.0 ±1.0	94.0 ±0.9	9.9 ±1.2
DFR	92.2 ±0.2	89.0 ±0.2	87.6 ±0.3	78.3 ±0.6	89.2 ±0.2	83.5 ±0.4	91.2 ±0.1	91.2 ±0.1	96.8 ±0.0	6.9 ±0.4
CRT	89.2 ±0.1	76.3 ±0.8	83.5 ±0.1	71.3 ±0.4	85.3 ±0.1	77.8 ±0.1	87.9 ±0.1	87.9 ±0.1	94.8 ±0.0	9.2 ±0.2
ReWeightCRT	89.4 ±0.3	76.3 ±0.2	83.8 ±0.3	71.9 ±0.7	85.6 ±0.3	78.1 ±0.4	88.0 ±0.2	88.0 ±0.2	94.9 ±0.1	8.8 ±0.2

E.3.2. CELEBA

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	95.0 ±0.1	57.6 ±0.8	87.4 ±0.2	76.3 ±0.3	89.9 ±0.1	82.7 ±0.2	85.6 ±0.2	93.0 ±0.0	98.4 ±0.0	2.9 ±0.1
Mixup	95.4 ±0.1	57.8 ±0.8	88.4 ±0.3	78.5 ±0.7	90.6 ±0.2	83.8 ±0.3	85.8 ±0.2	93.1 ±0.1	98.4 ±0.1	2.5 ±0.2
GroupDRO	94.4 ±0.0	68.3 ±0.9	85.8 ±0.1	72.7 ±0.1	89.2 ±0.1	81.7 ±0.1	88.4 ±0.2	93.9 ±0.1	98.6 ±0.0	4.1 ±0.0
CVaRDRO	95.1 ±0.1	60.2 ±3.0	87.7 ±0.3	76.9 ±0.6	90.1 ±0.1	83.1 ±0.2	86.3 ±0.7	93.1 ±0.1	98.4 ±0.0	3.1 ±0.1
JTT	95.9 ±0.0	48.3 ±1.5	90.5 ±0.1	82.9 ±0.3	91.4 ±0.1	85.2 ±0.1	83.4 ±0.4	92.4 ±0.2	98.6 ±0.0	1.3 ±0.1
LfF	81.1 ±5.6	53.0 ±4.3	71.8 ±4.1	45.2 ±8.3	73.2 ±5.6	59.0 ±7.3	78.3 ±3.0	85.3 ±2.9	94.1 ±1.2	27.9 ±5.5
LISA	95.4 ±0.1	57.8 ±0.8	88.4 ±0.3	78.5 ±0.7	90.6 ±0.2	83.8 ±0.3	85.8 ±0.2	93.1 ±0.1	98.4 ±0.1	2.5 ±0.2
ReSample	94.1 ±0.1	74.1 ±2.2	85.0 ±0.1	71.1 ±0.2	88.6 ±0.1	80.7 ±0.2	89.5 ±0.5	93.8 ±0.1	98.4 ±0.0	4.8 ±0.1
ReWeight	94.1 ±0.1	69.6 ±0.2	85.1 ±0.1	71.2 ±0.2	88.7 ±0.1	81.0 ±0.2	88.6 ±0.0	94.0 ±0.1	98.5 ±0.0	4.6 ±0.0
SqrtReWeight	94.0 ±0.2	66.9 ±2.2	85.0 ±0.4	71.0 ±0.9	88.6 ±0.3	80.7 ±0.4	87.9 ±0.5	93.9 ±0.1	98.4 ±0.0	4.8 ±0.2
CBLoss	94.4 ±0.0	65.4 ±1.4	85.9 ±0.1	72.9 ±0.2	89.2 ±0.1	81.7 ±0.1	87.6 ±0.3	93.8 ±0.1	98.5 ±0.0	4.4 ±0.1
Focal	94.9 ±0.3	56.9 ±3.4	87.4 ±0.7	76.4 ±1.5	89.7 ±0.4	82.4 ±0.7	85.2 ±0.8	92.6 ±0.3	98.3 ±0.1	3.1 ±0.4
LDAM	94.7 ±0.3	57.0 ±4.1	86.7 ±0.9	74.8 ±2.0	89.5 ±0.5	82.1 ±0.8	85.5 ±0.9	93.2 ±0.2	98.4 ±0.0	30.7 ±0.5
BSoftmax	94.5 ±0.1	69.6 ±1.2	85.9 ±0.2	72.9 ±0.4	89.4 ±0.2	82.0 ±0.3	88.8 ±0.3	94.2 ±0.1	98.6 ±0.0	4.6 ±0.0
DFR	93.6 ±0.0	73.7 ±0.8	84.1 ±0.1	69.4 ±0.2	87.8 ±0.1	79.4 ±0.1	89.0 ±0.2	93.2 ±0.0	98.2 ±0.0	14.8 ±0.5
CRT	94.1 ±0.1	69.6 ±0.7	85.1 ±0.3	71.4 ±0.5	88.6 ±0.2	80.7 ±0.3	88.4 ±0.1	93.6 ±0.0	98.4 ±0.0	4.6 ±0.2
ReWeightCRT	94.2 ±0.1	70.7 ±0.6	85.4 ±0.1	71.9 ±0.3	88.8 ±0.1	81.1 ±0.2	88.7 ±0.1	93.6 ±0.0	98.4 ±0.0	4.7 ±0.1

E.3.3. CIVILCOMMENTS

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	85.4 ±0.2	63.2 ±1.2	75.4 ±0.3	57.6 ±0.7	77.0 ±0.1	63.2 ±0.1	77.7 ±0.1	79.4 ±0.2	89.8 ±0.1	7.8 ±0.4
Mixup	84.6 ±0.1	65.8 ±1.5	74.4 ±0.2	55.3 ±0.3	76.4 ±0.1	62.6 ±0.2	78.0 ±0.2	79.7 ±0.0	89.7 ±0.1	9.3 ±0.4
GroupDRO	81.2 ±0.3	61.5 ±1.8	71.9 ±0.2	48.6 ±0.5	74.2 ±0.3	60.9 ±0.3	78.9 ±0.1	81.3 ±0.1	89.8 ±0.1	15.7 ±0.6
CVaRDRO	81.6 ±0.7	62.9 ±3.8	72.1 ±0.5	49.3 ±1.2	74.4 ±0.6	60.9 ±0.6	78.4 ±0.4	80.8 ±0.1	89.6 ±0.1	31.9 ±0.1
JTT	79.0 ±1.8	51.0 ±4.2	69.7 ±1.3	45.5 ±2.8	71.4 ±1.6	56.6 ±1.8	75.0 ±0.8	77.7 ±0.8	86.5 ±1.0	14.0 ±1.6
LfF	69.1 ±4.3	42.2 ±7.2	62.6 ±3.2	33.9 ±4.5	62.0 ±4.3	45.7 ±5.2	67.2 ±4.0	69.7 ±4.7	75.0 ±6.6	27.9 ±1.6
LISA	84.6 ±0.1	65.8 ±1.5	74.4 ±0.2	55.3 ±0.3	76.4 ±0.1	62.6 ±0.2	78.0 ±0.2	79.7 ±0.0	89.7 ±0.1	9.3 ±0.4
ReSample	80.4 ±0.2	61.0 ±0.6	71.2 ±0.1	47.2 ±0.2	73.4 ±0.1	59.8 ±0.2	78.3 ±0.1	80.7 ±0.1	89.3 ±0.1	17.0 ±0.6
ReWeight	80.6 ±0.3	59.3 ±1.1	71.5 ±0.2	47.6 ±0.5	73.7 ±0.2	60.3 ±0.2	78.7 ±0.1	81.3 ±0.0	89.9 ±0.1	14.9 ±0.5
SqrtReWeight	82.9 ±0.5	68.6 ±1.1	73.0 ±0.4	51.7 ±0.9	75.4 ±0.4	61.8 ±0.4	78.6 ±0.2	80.6 ±0.2	89.8 ±0.1	10.9 ±0.7
CBLoss	84.0 ±0.8	67.3 ±0.2	74.0 ±0.7	54.1 ±1.8	76.2 ±0.6	62.5 ±0.5	78.5 ±0.2	80.3 ±0.3	90.0 ±0.0	9.5 ±1.4
Focal	85.6 ±0.3	61.9 ±1.1	75.6 ±0.4	58.5 ±0.9	77.0 ±0.3	62.9 ±0.5	77.3 ±0.3	78.7 ±0.3	89.4 ±0.4	7.7 ±0.4
LDAM	80.2 ±2.1	28.4 ±7.7	67.8 ±3.3	46.2 ±5.6	68.3 ±3.1	48.9 ±5.0	66.1 ±3.6	69.5 ±3.2	77.4 ±4.0	20.7 ±0.5
BSoftmax	80.3 ±0.2	58.3 ±1.1	71.3 ±0.1	47.2 ±0.3	73.5 ±0.2	60.0 ±0.1	78.4 ±0.1	81.1 ±0.1	89.8 ±0.1	16.5 ±0.8
DFR	80.7 ±0.0	64.4 ±0.1	70.9 ±0.0	47.6 ±0.1	73.1 ±0.0	58.7 ±0.0	76.8 ±0.0	79.0 ±0.0	86.9 ±0.0	20.4 ±0.1
CRT	82.7 ±0.1	67.8 ±0.3	72.8 ±0.1	51.1 ±0.2	75.2 ±0.1	61.6 ±0.1	78.7 ±0.0	80.7 ±0.0	89.5 ±0.1	13.0 ±0.1
ReWeightCRT	82.4 ±0.0	64.7 ±0.2	72.6 ±0.0	50.5 ±0.1	75.0 ±0.0	61.4 ±0.0	78.4 ±0.0	80.7 ±0.0	89.5 ±0.0	12.6 ±0.1

E.3.4. MULTINLI

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	81.0 ±0.3	66.4 ±2.3	81.3 ±0.3	74.2 ±0.4	81.1 ±0.3	77.6 ±0.2	79.4 ±0.8	81.0 ±0.3	93.6 ±0.1	11.1 ±0.6
Mixup	81.7 ±0.1	66.8 ±0.3	81.9 ±0.1	75.8 ±0.1	81.8 ±0.1	78.3 ±0.1	80.1 ±0.2	81.7 ±0.1	93.7 ±0.0	10.9 ±0.2
GroupDRO	81.1 ±0.2	64.1 ±0.8	81.3 ±0.2	74.8 ±0.1	81.1 ±0.2	77.9 ±0.1	79.4 ±0.2	81.1 ±0.2	93.7 ±0.1	9.1 ±0.9
CVaRDRO	75.4 ±0.2	48.2 ±3.4	75.8 ±0.3	68.9 ±0.1	75.5 ±0.3	72.0 ±0.1	72.3 ±0.8	75.4 ±0.2	87.3 ±0.2	41.8 ±0.2
JTT	81.4 ±0.0	65.1 ±1.6	81.7 ±0.0	75.0 ±0.1	81.5 ±0.0	77.8 ±0.1	79.8 ±0.3	81.4 ±0.0	93.9 ±0.0	9.4 ±0.4
LfF	71.4 ±1.6	57.3 ±5.7	71.5 ±1.7	67.1 ±2.0	71.4 ±1.7	68.4 ±2.7	69.3 ±2.9	71.4 ±1.6	86.6 ±1.4	6.2 ±0.5
LISA	81.7 ±0.1	66.8 ±0.3	81.9 ±0.1	75.8 ±0.1	81.8 ±0.1	78.3 ±0.1	80.1 ±0.2	81.7 ±0.1	93.7 ±0.0	10.9 ±0.2
ReSample	81.5 ±0.0	66.8 ±0.5	81.9 ±0.1	74.2 ±0.2	81.6 ±0.0	78.0 ±0.1	79.9 ±0.1	81.5 ±0.0	93.9 ±0.0	12.2 ±0.6
ReWeight	79.4 ±0.2	64.2 ±1.9	79.6 ±0.2	72.9 ±0.3	79.4 ±0.2	75.8 ±0.1	78.0 ±0.2	79.4 ±0.2	92.6 ±0.1	14.2 ±0.6
SqrtReWeight	80.6 ±0.2	63.8 ±2.4	80.8 ±0.2	75.1 ±0.2	80.6 ±0.2	77.5 ±0.2	78.8 ±0.5	80.6 ±0.2	93.5 ±0.1	7.8 ±0.2
CBLoss	80.6 ±0.3	63.6 ±2.4	80.8 ±0.2	75.1 ±0.3	80.6 ±0.3	77.5 ±0.2	78.7 ±0.5	80.6 ±0.3	93.5 ±0.1	7.8 ±0.2
Focal	80.9 ±0.2	62.4 ±2.0	81.2 ±0.2	74.3 ±0.1	81.0 ±0.2	77.4 ±0.2	78.7 ±0.3	80.9 ±0.2	93.7 ±0.1	5.3 ±0.8
LDAM	80.9 ±0.1	65.5 ±0.8	81.1 ±0.1	74.6 ±0.3	80.9 ±0.1	77.4 ±0.0	79.2 ±0.2	80.9 ±0.1	93.5 ±0.0	33.2 ±0.4
BSoftmax	80.6 ±0.2	63.6 ±2.4	80.8 ±0.2	75.1 ±0.2	80.7 ±0.2	77.6 ±0.2	78.7 ±0.5	80.6 ±0.2	93.5 ±0.1	7.8 ±0.2
DFR	80.2 ±0.0	63.8 ±0.0	80.3 ±0.0	75.2 ±0.0	80.3 ±0.0	76.2 ±0.0	78.5 ±0.0	80.2 ±0.0	92.9 ±0.0	5.8 ±0.0
CRT	80.2 ±0.0	65.4 ±0.2	80.3 ±0.0	74.3 ±0.0	80.2 ±0.0	76.4 ±0.0	78.6 ±0.0	80.2 ±0.0	92.8 ±0.0	14.9 ±0.1
ReWeightCRT	80.2 ±0.0	65.2 ±0.2	80.3 ±0.0	74.4 ±0.0	80.2 ±0.0	76.4 ±0.0	78.6 ±0.0	80.2 ±0.0	92.9 ±0.0	14.7 ±0.1

E.3.5. METASHIFT

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	91.2 ±0.4	82.1 ±0.8	91.1 ±0.4	90.4 ±0.5	91.1 ±0.4	90.6 ±0.4	89.1 ±0.5	91.1 ±0.4	97.4 ±0.1	6.1 ±0.7
Mixup	91.4 ±0.1	79.0 ±0.8	91.4 ±0.1	90.6 ±0.4	91.4 ±0.1	90.9 ±0.1	88.5 ±0.2	91.4 ±0.1	97.3 ±0.0	1.9 ±0.4
GroupDRO	91.5 ±0.3	83.1 ±0.7	91.4 ±0.3	90.1 ±0.6	91.4 ±0.3	91.0 ±0.3	89.4 ±0.5	91.5 ±0.3	97.6 ±0.1	5.5 ±1.0
CVaRDRO	91.2 ±1.0	83.5 ±0.5	91.2 ±1.0	89.4 ±1.8	91.1 ±1.0	90.7 ±0.9	89.2 ±1.0	91.2 ±0.9	97.5 ±0.2	13.0 ±5.9
JTT	91.2 ±0.1	82.6 ±0.4	91.1 ±0.1	90.6 ±0.2	91.1 ±0.1	90.6 ±0.1	89.2 ±0.1	91.1 ±0.1	97.6 ±0.0	7.2 ±0.2
LfF	80.4 ±0.4	72.3 ±1.3	80.7 ±0.3	76.9 ±1.7	80.4 ±0.4	79.7 ±0.4	80.5 ±0.6	80.6 ±0.3	91.5 ±0.1	8.5 ±1.1
LISA	91.4 ±0.1	79.0 ±0.8	91.4 ±0.1	90.6 ±0.4	91.4 ±0.1	90.9 ±0.1	88.5 ±0.2	91.4 ±0.1	97.3 ±0.0	1.9 ±0.4
ReSample	92.2 ±0.3	81.0 ±1.7	92.1 ±0.3	91.4 ±0.4	92.2 ±0.3	91.7 ±0.3	89.6 ±0.1	92.2 ±0.2	97.5 ±0.1	6.8 ±0.4
ReWeight	91.5 ±0.4	83.1 ±0.7	91.5 ±0.4	90.6 ±0.3	91.5 ±0.4	91.0 ±0.4	89.5 ±0.4	91.5 ±0.4	97.5 ±0.1	5.8 ±0.6
SqrtReWeight	91.3 ±0.1	82.6 ±0.4	91.2 ±0.1	90.3 ±0.2	91.2 ±0.1	90.7 ±0.2	89.2 ±0.2	91.3 ±0.1	97.5 ±0.1	5.6 ±1.0
CBLoss	91.4 ±0.1	83.1 ±0.0	91.3 ±0.1	90.4 ±0.4	91.4 ±0.1	90.9 ±0.2	89.4 ±0.1	91.4 ±0.2	97.4 ±0.1	6.3 ±0.4
Focal	91.6 ±0.2	81.0 ±0.4	91.7 ±0.2	90.9 ±0.6	91.6 ±0.2	91.1 ±0.2	89.4 ±0.2	91.6 ±0.2	97.6 ±0.0	4.9 ±1.2
LDAM	91.6 ±0.1	83.6 ±0.4	91.6 ±0.0	90.9 ±0.3	91.6 ±0.1	91.1 ±0.1	89.9 ±0.1	91.6 ±0.1	97.5 ±0.1	9.5 ±1.0
BSoftmax	91.3 ±0.3	82.6 ±0.4	91.3 ±0.3	89.9 ±0.3	91.3 ±0.3	90.9 ±0.3	89.2 ±0.3	91.4 ±0.3	97.5 ±0.1	5.7 ±0.8
DFR	90.2 ±0.2	81.4 ±0.1	90.2 ±0.2	88.1 ±0.5	90.2 ±0.2	89.8 ±0.2	88.0 ±0.2	90.3 ±0.2	96.7 ±0.0	3.2 ±0.2
CRT	91.5 ±0.0	83.1 ±0.0	91.4 ±0.0	90.6 ±0.1	91.4 ±0.0	90.9 ±0.0	89.5 ±0.0	91.4 ±0.0	97.3 ±0.0	6.8 ±0.0
ReWeightCRT	91.3 ±0.1	85.1 ±0.4	91.2 ±0.1	90.1 ±0.3	91.2 ±0.1	90.8 ±0.1	89.7 ±0.2	91.3 ±0.1	96.8 ±0.1	8.1 ±0.1

E.3.6. IMAGENETBG

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	87.7 ±0.1	76.8 ±0.9	87.9 ±0.1	78.3 ±1.3	87.7 ±0.1	80.7 ±0.4	87.7 ±0.1	87.7 ±0.1	99.0 ±0.0	5.1 ±0.1
Mixup	87.9 ±0.2	76.9 ±0.7	88.4 ±0.1	76.6 ±2.6	88.0 ±0.2	80.5 ±1.0	87.9 ±0.2	87.9 ±0.2	98.7 ±0.0	4.7 ±1.6
GroupDRO	87.7 ±0.1	76.4 ±0.2	87.9 ±0.1	76.2 ±0.5	87.6 ±0.1	81.1 ±0.3	87.7 ±0.1	87.7 ±0.1	99.0 ±0.0	5.1 ±0.1
CVaRDRO	87.8 ±0.2	74.8 ±0.8	88.0 ±0.1	79.9 ±0.9	87.8 ±0.2	79.5 ±0.3	87.8 ±0.2	87.8 ±0.2	99.0 ±0.0	5.6 ±0.2
JTT	87.6 ±0.4	77.0 ±0.4	87.8 ±0.3	78.3 ±3.0	87.5 ±0.3	80.4 ±0.6	87.6 ±0.4	87.6 ±0.4	99.0 ±0.0	3.7 ±0.2
LfF	84.7 ±0.5	70.1 ±1.4	85.4 ±0.3	72.1 ±3.1	84.7 ±0.5	76.2 ±0.6	84.7 ±0.5	84.7 ±0.5	98.6 ±0.0	1.8 ±0.4
LISA	87.9 ±0.2	76.9 ±0.7	88.4 ±0.1	76.6 ±2.6	88.0 ±0.2	80.5 ±1.0	87.9 ±0.2	87.9 ±0.2	98.7 ±0.0	4.7 ±1.6
ReSample	88.2 ±0.4	77.7 ±1.1	88.4 ±0.4	79.7 ±1.0	88.2 ±0.4	80.6 ±1.0	88.2 ±0.4	88.2 ±0.4	99.0 ±0.0	5.4 ±0.5
ReWeight	87.7 ±0.1	76.8 ±0.9	87.9 ±0.1	78.3 ±1.3	87.7 ±0.1	80.7 ±0.4	87.7 ±0.1	87.7 ±0.1	99.0 ±0.0	5.1 ±0.1
SqrtReWeight	87.7 ±0.1	76.8 ±0.9	87.9 ±0.1	78.3 ±1.3	87.7 ±0.1	80.7 ±0.4	87.7 ±0.1	87.7 ±0.1	99.0 ±0.0	5.1 ±0.1
CBLoss	87.7 ±0.1	76.8 ±0.9	87.9 ±0.1	78.3 ±1.3	87.7 ±0.1	80.7 ±0.4	87.7 ±0.1	87.7 ±0.1	99.0 ±0.0	5.1 ±0.1
Focal	86.7 ±0.2	71.9 ±1.2	87.1 ±0.2	74.2 ±1.3	86.6 ±0.2	77.6 ±0.7	86.7 ±0.2	86.7 ±0.2	98.9 ±0.0	2.8 ±0.6
LDAM	88.2 ±0.1	76.7 ±0.5	88.5 ±0.1	77.6 ±1.1	88.1 ±0.1	81.3 ±0.4	88.2 ±0.1	88.2 ±0.1	98.8 ±0.0	45.9 ±0.7
BSoftmax	87.7 ±0.1	76.1 ±2.0	88.0 ±0.1	77.7 ±1.3	87.7 ±0.1	80.4 ±0.8	87.7 ±0.1	87.7 ±0.1	99.0 ±0.0	5.6 ±0.5
DFR	86.8 ±0.5	74.4 ±1.8	86.9 ±0.5	78.9 ±1.6	86.7 ±0.5	78.1 ±1.3	86.8 ±0.5	86.8 ±0.5	98.8 ±0.1	8.9 ±1.4
CRT	88.3 ±0.1	78.2 ±0.5	88.3 ±0.1	82.7 ±0.4	88.3 ±0.1	80.9 ±0.2	88.3 ±0.1	88.3 ±0.1	99.1 ±0.0	5.6 ±0.2
ReWeightCRT	88.4 ±0.1	77.5 ±0.7	88.5 ±0.1	82.1 ±0.4	88.4 ±0.1	81.2 ±0.3	88.4 ±0.1	88.4 ±0.1	99.1 ±0.0	5.4 ±0.2

E.3.7. NICO++

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	85.3 ±0.3	35.0 ±4.1	86.3 ±0.2	56.3 ±4.9	85.4 ±0.3	64.9 ±2.3	84.9 ±0.3	85.2 ±0.3	99.4 ±0.0	9.4 ±0.1
Mixup	85.4 ±0.1	30.0 ±4.1	86.2 ±0.1	62.2 ±0.5	85.5 ±0.1	69.4 ±0.4	85.0 ±0.1	85.3 ±0.1	99.2 ±0.0	2.3 ±0.5
GroupDRO	83.7 ±0.5	31.1 ±0.9	84.6 ±0.4	58.2 ±1.5	83.8 ±0.5	66.7 ±0.9	83.3 ±0.4	83.6 ±0.5	99.3 ±0.0	7.0 ±0.2
CVaRDRO	85.8 ±0.1	27.8 ±2.3	86.5 ±0.1	61.5 ±0.7	85.8 ±0.1	67.3 ±0.5	85.2 ±0.1	85.7 ±0.1	99.4 ±0.0	9.9 ±0.0
JTT	85.7 ±0.1	30.6 ±2.3	86.4 ±0.1	60.9 ±1.2	85.8 ±0.1	66.7 ±0.4	85.2 ±0.1	85.6 ±0.1	99.4 ±0.0	9.5 ±0.2
LfF	78.7 ±0.6	28.8 ±2.0	81.0 ±0.3	45.1 ±1.2	79.0 ±0.6	54.2 ±1.4	78.4 ±0.5	78.6 ±0.6	99.2 ±0.0	1.5 ±0.3
LISA	85.4 ±0.1	30.0 ±4.1	86.2 ±0.1	62.2 ±0.5	85.5 ±0.1	69.4 ±0.4	85.0 ±0.1	85.3 ±0.1	99.2 ±0.0	2.3 ±0.5
ReSample	84.9 ±0.2	30.6 ±2.3	85.6 ±0.1	62.9 ±1.6	84.9 ±0.2	67.1 ±0.3	84.3 ±0.2	84.8 ±0.2	99.3 ±0.0	10.2 ±0.4
ReWeight	85.5 ±0.2	25.0 ±0.0	86.4 ±0.1	59.0 ±1.3	85.6 ±0.2	67.6 ±0.4	84.9 ±0.2	85.4 ±0.2	99.4 ±0.0	9.7 ±0.0
SqrtReWeight	85.5 ±0.1	32.8 ±3.5	86.5 ±0.1	55.9 ±2.1	85.6 ±0.1	66.4 ±1.3	85.0 ±0.1	85.4 ±0.1	99.4 ±0.0	9.4 ±0.1
CBLoss	85.9 ±0.1	31.7 ±3.6	86.6 ±0.1	59.3 ±3.0	86.0 ±0.1	67.2 ±0.7	85.4 ±0.1	85.8 ±0.0	99.4 ±0.0	10.1 ±0.1
Focal	85.7 ±0.1	30.6 ±2.3	86.5 ±0.1	58.5 ±0.6	85.8 ±0.1	66.7 ±0.8	85.2 ±0.1	85.6 ±0.1	99.5 ±0.0	6.3 ±0.3
LDAM	85.4 ±0.4	31.7 ±3.6	86.1 ±0.3	62.4 ±1.0	85.5 ±0.4	68.1 ±1.3	84.9 ±0.5	85.3 ±0.4	99.1 ±0.0	56.9 ±1.4
BSoftmax	85.8 ±0.0	35.6 ±1.8	86.4 ±0.1	60.7 ±1.4	85.8 ±0.0	69.2 ±0.6	85.3 ±0.0	85.7 ±0.0	99.4 ±0.0	9.4 ±0.1
DFR	82.7 ±0.1	38.0 ±3.8	83.2 ±0.1	58.5 ±1.6	82.7 ±0.1	65.1 ±0.2	82.4 ±0.1	82.6 ±0.1	99.2 ±0.0	11.7 ±0.2
CRT	85.8 ±0.1	33.3 ±0.0	86.1 ±0.0	65.3 ±0.6	85.8 ±0.1	69.9 ±0.4	85.3 ±0.1	85.6 ±0.1	99.4 ±0.0	6.0 ±0.6
ReWeightCRT	85.8 ±0.1	33.3 ±0.0	86.1 ±0.1	64.3 ±0.7	85.8 ±0.1	69.7 ±0.4	85.4 ±0.1	85.7 ±0.1	99.4 ±0.0	6.1 ±0.8

E.3.8. MIMIC-CXR

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	78.3 ±0.1	68.6 ±0.2	77.4 ±0.1	71.2 ±0.1	77.6 ±0.1	73.7 ±0.1	77.3 ±0.1	78.0 ±0.1	85.4 ±0.1	3.8 ±0.2
Mixup	77.9 ±0.3	66.8 ±0.6	77.0 ±0.3	70.7 ±0.6	77.2 ±0.3	73.2 ±0.3	76.9 ±0.3	77.5 ±0.3	84.9 ±0.3	3.6 ±0.5
GroupDRO	77.9 ±0.0	67.4 ±0.5	77.1 ±0.0	70.2 ±0.1	77.4 ±0.0	73.7 ±0.0	77.3 ±0.0	77.9 ±0.0	85.2 ±0.1	5.5 ±0.3
CVaRDRO	78.2 ±0.1	68.0 ±0.2	77.3 ±0.1	70.7 ±0.2	77.6 ±0.1	73.8 ±0.1	77.3 ±0.1	78.0 ±0.1	85.1 ±0.0	6.8 ±0.7
JTT	77.4 ±0.3	64.9 ±0.3	76.5 ±0.3	70.1 ±0.7	76.7 ±0.2	72.8 ±0.2	76.4 ±0.2	77.1 ±0.2	84.5 ±0.2	4.2 ±0.4
LfF	73.2 ±0.9	62.2 ±2.4	72.3 ±1.0	65.1 ±0.9	72.5 ±1.0	67.8 ±1.5	72.4 ±1.1	72.9 ±1.1	79.3 ±1.3	11.3 ±0.8
LISA	77.9 ±0.3	66.8 ±0.6	77.0 ±0.3	70.7 ±0.6	77.2 ±0.3	73.2 ±0.3	76.9 ±0.3	77.5 ±0.3	84.9 ±0.3	3.6 ±0.5
ReSample	78.4 ±0.1	67.5 ±0.3	77.6 ±0.1	70.8 ±0.1	77.8 ±0.1	74.2 ±0.1	77.6 ±0.1	78.3 ±0.1	85.4 ±0.1	5.3 ±0.0
ReWeight	77.6 ±0.0	67.0 ±0.4	76.8 ±0.0	69.7 ±0.0	77.1 ±0.0	73.4 ±0.0	76.9 ±0.1	77.6 ±0.0	84.9 ±0.0	5.2 ±0.4
SqrtReWeight	78.3 ±0.0	68.0 ±0.4	77.5 ±0.0	70.6 ±0.0	77.8 ±0.0	74.2 ±0.0	77.6 ±0.0	78.3 ±0.0	85.6 ±0.0	5.2 ±0.3
CBLoss	78.3 ±0.2	67.6 ±0.3	77.4 ±0.2	70.8 ±0.2	77.7 ±0.2	74.0 ±0.2	77.5 ±0.1	78.2 ±0.2	85.5 ±0.1	4.7 ±0.3
Focal	78.3 ±0.1	68.7 ±0.4	77.4 ±0.1	70.8 ±0.2	77.6 ±0.1	73.9 ±0.0	77.4 ±0.1	78.1 ±0.0	85.4 ±0.0	10.1 ±0.6
LDAM	78.0 ±0.1	66.6 ±0.6	77.2 ±0.2	70.4 ±0.2	77.4 ±0.2	73.7 ±0.2	77.2 ±0.2	77.9 ±0.2	85.2 ±0.1	22.5 ±0.2
BSoftmax	78.0 ±0.1	67.6 ±0.6	77.1 ±0.1	70.4 ±0.1	77.3 ±0.1	73.6 ±0.1	77.2 ±0.1	77.8 ±0.1	85.1 ±0.1	5.2 ±0.4
DFR	78.3 ±0.0	67.1 ±0.4	77.4 ±0.0	72.1 ±0.2	77.5 ±0.0	73.2 ±0.1	77.1 ±0.0	77.6 ±0.1	85.1 ±0.0	20.0 ±0.1
CRT	78.0 ±0.0	68.1 ±0.1	77.2 ±0.0	70.1 ±0.0	77.4 ±0.0	73.9 ±0.0	77.3 ±0.0	78.0 ±0.0	85.3 ±0.0	6.2 ±0.0
ReWeightCRT	77.8 ±0.0	67.9 ±0.1	77.0 ±0.0	70.0 ±0.0	77.2 ±0.0	73.6 ±0.0	77.0 ±0.0	77.8 ±0.0	85.0 ±0.0	5.4 ±0.0

E.3.9. MIMICNOTES

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	91.1 ±0.1	24.2 ±2.8	76.4 ±0.9	60.3 ±2.1	65.8 ±1.2	36.5 ±2.5	62.2 ±1.3	62.4 ±1.4	85.2 ±0.1	2.2 ±0.1
Mixup	91.1 ±0.0	22.7 ±3.2	76.8 ±0.7	61.2 ±1.6	65.1 ±1.6	35.0 ±3.2	61.5 ±1.6	61.7 ±1.7	85.4 ±0.0	2.0 ±0.8
GroupDRO	83.2 ±2.4	62.6 ±6.3	64.7 ±1.4	33.7 ±3.4	66.4 ±1.3	42.8 ±1.1	74.3 ±1.5	74.4 ±1.5	85.1 ±0.1	13.8 ±3.2
CVaRDRO	90.2 ±0.0	0.0 ±0.0	45.1 ±0.0	0.0 ±0.0	47.4 ±0.0	0.0 ±0.0	50.0 ±0.0	50.0 ±0.0	71.6 ±0.4	40.2 ±0.0
JTT	71.3 ±3.7	65.9 ±2.8	60.3 ±0.7	23.4 ±1.9	58.6 ±2.3	36.0 ±1.8	75.5 ±0.4	75.6 ±0.4	84.9 ±0.1	27.5 ±3.9
LfF	84.0 ±1.2	62.7 ±2.1	64.6 ±0.7	33.6 ±1.6	67.1 ±0.8	43.6 ±0.8	74.7 ±0.4	74.7 ±0.5	85.1 ±0.0	12.5 ±1.2
LISA	91.1 ±0.0	22.7 ±3.2	76.8 ±0.7	61.2 ±1.6	65.1 ±1.6	35.0 ±3.2	61.5 ±1.6	61.7 ±1.7	85.4 ±0.0	2.0 ±0.8
ReSample	81.4 ±1.5	67.1 ±2.6	63.3 ±0.6	30.5 ±1.5	65.4 ±1.0	42.0 ±1.0	75.4 ±0.3	75.6 ±0.4	85.1 ±0.0	17.4 ±1.9
ReWeight	82.7 ±0.7	65.5 ±1.3	63.8 ±0.4	31.7 ±0.9	66.3 ±0.5	42.8 ±0.5	75.3 ±0.2	75.4 ±0.3	85.2 ±0.1	15.9 ±0.7
SqrtReWeight	90.3 ±0.2	35.7 ±4.0	72.4 ±0.9	51.3 ±2.1	68.7 ±0.9	42.8 ±2.1	66.7 ±1.6	66.8 ±1.6	85.2 ±0.1	3.7 ±0.7
CBLoss	78.2 ±1.0	72.3 ±1.3	61.9 ±0.3	27.3 ±0.8	63.2 ±0.7	39.8 ±0.6	76.1 ±0.2	76.2 ±0.2	85.0 ±0.0	20.6 ±1.3
Focal	91.0 ±0.0	19.1 ±2.3	77.1 ±0.6	62.1 ±1.4	63.6 ±1.3	31.9 ±2.6	59.9 ±1.1	60.2 ±1.1	85.3 ±0.1	8.1 ±0.7
LDAM	90.6 ±0.1	5.3 ±2.4	84.4 ±0.8	78.1 ±1.7	52.5 ±2.1	10.0 ±4.1	52.7 ±1.2	52.7 ±1.2	84.9 ±0.1	28.9 ±1.0
BSoftmax	76.9 ±0.9	73.1 ±1.0	61.7 ±0.2	26.5 ±0.6	62.5 ±0.5	39.3 ±0.4	76.6 ±0.2	76.7 ±0.2	85.4 ±0.0	23.5 ±1.1
DFR	69.2 ±1.3	67.3 ±1.7	58.8 ±0.2	21.0 ±0.5	56.5 ±0.8	33.1 ±0.5	73.1 ±0.0	73.1 ±0.0	81.0 ±0.0	38.4 ±0.1
CRT	77.8 ±0.0	73.1 ±0.0	61.6 ±0.0	26.7 ±0.0	62.8 ±0.0	39.2 ±0.0	75.9 ±0.0	75.9 ±0.0	84.3 ±0.0	23.0 ±0.1
ReWeightCRT	81.2 ±2.8	63.9 ±7.6	63.6 ±1.6	31.4 ±3.9	64.7 ±1.6	40.7 ±1.3	73.8 ±1.6	73.9 ±1.6	84.3 ±0.0	26.5 ±2.4

E.3.10. CXRMULTISITE

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	98.3 ±0.0	0.0 ±0.0	49.2 ±0.0	0.0 ±0.0	49.6 ±0.0	0.0 ±0.0	50.0 ±0.0	50.0 ±0.0	93.1 ±0.1	0.3 ±0.1
Mixup	98.3 ±0.0	0.0 ±0.0	49.2 ±0.0	0.0 ±0.0	49.6 ±0.0	0.0 ±0.0	50.0 ±0.0	50.0 ±0.0	92.9 ±0.1	0.3 ±0.0
GroupDRO	90.2 ±0.1	0.0 ±0.0	56.7 ±0.1	13.6 ±0.1	59.2 ±0.1	23.7 ±0.2	50.2 ±0.1	90.4 ±0.0	92.8 ±0.2	13.5 ±0.7
CVaRDRO	98.3 ±0.0	0.0 ±0.0	61.2 ±4.9	24.0 ±9.8	50.7 ±0.7	2.2 ±1.5	50.2 ±0.2	50.6 ±0.4	93.0 ±0.0	0.9 ±0.3
JTT	94.1 ±0.9	0.0 ±0.0	59.0 ±0.7	18.5 ±1.4	62.9 ±0.8	28.9 ±1.2	55.2 ±0.9	82.2 ±2.4	93.2 ±0.1	6.4 ±0.5
LfF	9.9 ±6.7	5.4 ±4.4	17.4 ±13.5	0.6 ±0.5	8.5 ±5.6	1.2 ±1.0	50.5 ±0.4	51.7 ±1.4	60.6 ±1.6	82.6 ±12.8
LISA	98.3 ±0.0	0.0 ±0.0	49.2 ±0.0	0.0 ±0.0	49.6 ±0.0	0.0 ±0.0	50.0 ±0.0	50.0 ±0.0	92.9 ±0.1	0.3 ±0.0
ReSample	88.3 ±1.6	0.1 ±0.1	55.9 ±0.6	12.0 ±1.2	57.3 ±1.4	20.9 ±1.8	50.3 ±0.5	88.1 ±0.7	92.3 ±0.1	13.0 ±2.8
ReWeight	89.5 ±0.0	0.3 ±0.1	56.4 ±0.0	13.0 ±0.0	58.5 ±0.0	22.7 ±0.0	50.5 ±0.2	90.3 ±0.0	93.2 ±0.1	17.7 ±1.7
SqrtReWeight	94.5 ±0.4	0.0 ±0.0	59.4 ±0.2	19.3 ±0.6	63.7 ±0.3	30.2 ±0.4	56.3 ±0.1	82.3 ±1.6	93.3 ±0.0	6.0 ±0.2
CBLoss	1.7 ±0.0	0.0 ±0.0	0.8 ±0.0	0.0 ±0.0	1.7 ±0.0	0.0 ±0.0	50.0 ±0.0	50.0 ±0.0	62.7 ±1.5	98.3 ±0.0
Focal	98.3 ±0.0	0.0 ±0.0	55.4 ±5.1	12.5 ±10.2	49.7 ±0.1	0.3 ±0.2	50.0 ±0.0	50.1 ±0.1	93.2 ±0.0	11.5 ±0.6
LDAM	98.3 ±0.0	0.0 ±0.0	49.2 ±0.0	0.0 ±0.0	49.6 ±0.0	0.0 ±0.0	50.0 ±0.0	50.0 ±0.0	93.1 ±0.1	0.3 ±0.1
BSoftmax	89.1 ±0.2	0.5 ±0.1	56.2 ±0.1	12.5 ±0.2	58.1 ±0.2	22.0 ±0.3	50.4 ±0.0	90.0 ±0.1	92.9 ±0.1	19.9 ±1.3
DFR	89.7 ±0.1	0.6 ±0.1	56.5 ±0.0	13.2 ±0.1	58.7 ±0.1	23.0 ±0.1	50.4 ±0.0	90.4 ±0.0	92.8 ±0.1	47.3 ±0.1
CRT	90.4 ±0.1	1.1 ±0.5	56.9 ±0.0	13.9 ±0.1	59.5 ±0.1	24.2 ±0.1	51.2 ±0.1	90.2 ±0.1	93.3 ±0.0	15.7 ±0.9
ReWeightCRT	89.9 ±0.1	1.4 ±0.6	56.6 ±0.0	13.4 ±0.1	59.0 ±0.1	23.3 ±0.1	51.1 ±0.3	90.4 ±0.0	93.1 ±0.1	15.5 ±0.7

E.3.11. CHEXPERT

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	88.6 ±0.7	41.7 ±3.4	68.3 ±1.1	42.3 ±2.4	68.7 ±0.2	43.8 ±0.7	69.5 ±1.3	70.0 ±1.5	85.4 ±0.4	5.0 ±1.1
Mixup	81.9 ±6.2	37.4 ±3.5	63.5 ±5.0	33.9 ±9.3	62.5 ±5.9	35.7 ±7.6	63.8 ±4.7	64.1 ±4.6	76.1 ±8.5	16.1 ±9.0
GroupDRO	79.2 ±0.2	74.7 ±0.3	62.9 ±0.1	28.6 ±0.2	64.7 ±0.2	42.0 ±0.2	78.4 ±0.2	79.1 ±0.1	86.2 ±0.0	21.3 ±0.8
CVaRDRO	73.7 ±1.0	50.2 ±1.8	57.3 ±0.1	20.1 ±0.3	56.8 ±0.4	29.9 ±0.2	65.7 ±0.6	67.0 ±0.4	72.9 ±0.4	40.4 ±0.0
JTT	75.2 ±0.8	60.4 ±4.8	59.4 ±1.1	23.0 ±1.5	59.6 ±1.4	34.4 ±2.3	70.7 ±2.6	72.0 ±2.5	79.0 ±2.5	24.4 ±0.5
LfF	22.3 ±10.2	13.7 ±9.8	37.3 ±5.8	9.0 ±0.7	19.5 ±8.3	8.8 ±3.9	46.2 ±2.9	46.2 ±3.1	30.5 ±10.1	65.7 ±10.2
LISA	81.9 ±6.2	37.4 ±3.5	63.5 ±5.0	33.9 ±9.3	62.5 ±5.9	35.7 ±7.6	63.8 ±4.7	64.1 ±4.6	76.1 ±8.5	16.1 ±9.0
ReSample	79.6 ±0.6	74.3 ±0.4	63.1 ±0.3	29.0 ±0.6	65.0 ±0.5	42.3 ±0.6	78.3 ±0.3	79.0 ±0.2	86.3 ±0.2	20.1 ±1.4
ReWeight	79.6 ±0.5	73.7 ±1.0	63.1 ±0.2	29.0 ±0.5	65.0 ±0.4	42.4 ±0.5	78.4 ±0.3	79.1 ±0.2	86.2 ±0.1	21.0 ±0.7
SqrtReWeight	83.5 ±0.3	68.5 ±1.6	65.0 ±0.2	33.3 ±0.4	67.9 ±0.2	45.5 ±0.3	77.8 ±0.4	78.5 ±0.3	86.3 ±0.3	15.6 ±1.1
CBLoss	80.0 ±0.5	74.0 ±0.7	63.3 ±0.2	29.4 ±0.5	65.2 ±0.4	42.6 ±0.5	78.6 ±0.2	79.0 ±0.2	86.1 ±0.3	19.6 ±0.6
Focal	89.3 ±0.3	42.1 ±4.0	69.6 ±0.4	44.7 ±1.1	69.8 ±0.4	45.5 ±1.0	70.4 ±1.1	70.4 ±1.3	86.5 ±0.1	16.1 ±1.7
LDAM	90.0 ±0.0	36.0 ±0.7	70.6 ±0.1	47.3 ±0.2	69.1 ±0.1	43.6 ±0.1	67.8 ±0.2	67.9 ±0.1	86.1 ±0.1	32.3 ±0.2
BSoftmax	79.9 ±0.2	73.8 ±1.0	63.3 ±0.0	29.4 ±0.1	65.3 ±0.1	42.8 ±0.0	78.6 ±0.2	79.5 ±0.1	86.6 ±0.1	21.3 ±0.4
DFR	79.1 ±0.0	75.8 ±0.3	63.0 ±0.0	28.6 ±0.0	64.7 ±0.0	42.1 ±0.0	78.8 ±0.0	79.3 ±0.0	86.0 ±0.0	25.6 ±0.1
CRT	79.3 ±0.1	74.6 ±0.4	62.9 ±0.1	28.7 ±0.1	64.7 ±0.1	42.0 ±0.1	78.5 ±0.1	79.0 ±0.1	86.1 ±0.1	21.7 ±0.1
ReWeightCRT	79.3 ±0.1	75.1 ±0.2	63.0 ±0.0	28.7 ±0.0	64.7 ±0.0	42.0 ±0.0	78.6 ±0.1	79.0 ±0.1	86.2 ±0.0	21.7 ±0.1

E.3.12. LIVING17

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1	Worst F1	Adjusted Acc.	Balanced Acc.	AUROC	ECE
ERM	27.7 ±1.1	5.7 ±1.5	28.2 ±1.0	8.2 ±1.6	27.1 ±1.1	6.9 ±1.7	27.7 ±1.1	27.7 ±1.1	77.3 ±1.3	59.6 ±0.5
Mixup	29.8 ±1.8	8.7 ±1.4	30.9 ±2.2	9.5 ±1.3	29.3 ±1.9	9.7 ±1.5	29.8 ±1.8	29.8 ±1.8	78.2 ±1.2	34.1 ±1.9
GroupDRO	31.1 ±1.0	6.0 ±1.4	32.1 ±0.9	9.6 ±0.2	30.8 ±0.7	7.6 ±0.8	31.1 ±1.0	31.1 ±1.0	80.1 ±0.8	53.5 ±0.8
CVaRDRO	27.3 ±1.6	4.0 ±0.5	29.2 ±1.7	5.1 ±0.8	26.5 ±1.5	4.8 ±0.6	27.3 ±1.6	27.3 ±1.6	81.0 ±0.2	28.8 ±6.8
JTT	28.3 ±1.1	5.7 ±2.2	31.1 ±0.9	8.0 ±1.7	28.3 ±1.3	7.2 ±2.3	28.3 ±1.1	28.3 ±1.1	81.0 ±0.8	36.9 ±3.2
LfF	26.4 ±1.3	7.0 ±1.2	28.3 ±0.8	9.6 ±1.8	26.1 ±1.2	8.7 ±1.7	26.4 ±1.3	26.4 ±1.3	76.6 ±0.6	61.0 ±0.7
LISA	29.8 ±1.8	8.7 ±1.4	30.9 ±2.2	9.5 ±1.3	29.3 ±1.9	9.7 ±1.5	29.8 ±1.8	29.8 ±1.8	78.2 ±1.2	34.1 ±1.9
ReSample	31.4 ±0.6	6.7 ±1.5	33.0 ±0.6	11.0 ±0.6	31.0 ±0.6	8.3 ±1.2	31.4 ±0.6	31.4 ±0.6	81.0 ±0.7	46.6 ±3.1
ReWeight	27.7 ±1.1	5.7 ±1.5	28.2 ±1.0	8.2 ±1.6	27.1 ±1.1	6.9 ±1.7	27.7 ±1.1	27.7 ±1.1	77.3 ±1.3	59.6 ±0.5
SqrtReWeight	27.7 ±1.1	5.7 ±1.5	28.2 ±1.0	8.2 ±1.6	27.1 ±1.1	6.9 ±1.7	27.7 ±1.1	27.7 ±1.1	77.3 ±1.3	59.6 ±0.5
CBLoss	27.7 ±1.1	5.7 ±1.5	28.2 ±1.0	8.2 ±1.6	27.1 ±1.1	6.9 ±1.7	27.7 ±1.1	27.7 ±1.1	77.3 ±1.3	59.6 ±0.5
Focal	26.9 ±0.6	5.3 ±0.3	28.8 ±1.0	7.1 ±1.0	27.0 ±0.7	6.3 ±0.5	26.9 ±0.6	26.9 ±0.6	78.7 ±0.5	49.9 ±1.9
LDAM	24.3 ±0.8	4.0 ±0.8	28.0 ±1.2	6.6 ±1.4	24.0 ±0.8	5.1 ±0.3	24.3 ±0.8	24.3 ±0.8	79.1 ±1.0	12.4 ±0.5
BSoftmax	28.6 ±1.4	6.7 ±0.7	30.7 ±1.0	8.2 ±0.8	28.3 ±1.3	7.3 ±0.5	28.6 ±1.4	28.6 ±1.4	78.0 ±1.0	56.5 ±1.7
DFR	26.3 ±0.4	6.0 ±0.9	27.4 ±0.3	8.6 ±0.8	25.7 ±0.2	7.5 ±1.1	26.3 ±0.4	26.3 ±0.4	79.4 ±0.1	13.8 ±0.4
CRT	31.1 ±0.1	6.3 ±0.3	31.8 ±0.0	7.5 ±0.3	30.5 ±0.1	6.8 ±0.3	31.1 ±0.1	31.1 ±0.1	80.3 ±0.1	49.6 ±1.7
ReWeightCRT	33.1 ±0.1	9.3 ±0.3	33.4 ±0.1	11.3 ±0.3	32.6 ±0.0	10.8 ±0.2	33.1 ±0.1	33.1 ±0.1	82.0 ±0.0	40.0 ±0.4

E.3.13. OVERALL

Algorithm	Waterbirds	CelebA	CivilComments	MultiNLI	MetaShift	ImageNetBG	NICO++	MIMIC-CXR	MIMICNotes	CXRMultisite	CheXpert	Living17	Avg
ERM	69.1 ±4.7	57.6 ±0.8	63.2 ±1.2	66.4 ±2.3	82.1 ±0.8	76.8 ±0.9	35.0 ±4.1	68.6 ±0.2	80.4 ±0.2	50.1 ±0.9	41.7 ±3.4	27.7 ±1.1	59.9
Mixup	77.5 ±0.7	57.8 ±0.8	65.8 ±1.5	66.8 ±0.3	79.0 ±0.8	76.9 ±0.7	30.0 ±4.1	66.8 ±0.6	81.6 ±0.6	50.1 ±0.9	37.4 ±3.5	29.8 ±1.8	60.0
GroupDRO	73.1 ±0.4	68.3 ±0.9	61.5 ±1.8	64.1 ±0.8	83.1 ±0.7	76.4 ±0.2	31.1 ±0.9	67.4 ±0.5	83.7 ±0.1	59.2 ±0.3	74.7 ±0.3	31.1 ±1.0	64.5
CVaRDRO	75.5 ±2.2	60.2 ±3.0	62.9 ±3.8	48.2 ±3.4	83.5 ±0.5	74.8 ±0.8	27.8 ±2.3	68.0 ±0.2	65.6 ±1.5	50.2 ±0.9	50.2 ±1.8	27.3 ±1.6	57.8
JTT	71.2 ±0.5	48.3 ±1.5	51.0 ±4.2	65.1 ±1.6	82.6 ±0.4	77.0 ±0.4	30.6 ±2.3	64.9 ±0.3	83.8 ±0.1	57.9 ±2.1	60.4 ±4.8	28.3 ±1.1	60.1
LfF	75.0 ±0.7	53.0 ±4.3	42.2 ±7.2	57.3 ±5.7	72.3 ±1.3	70.1 ±1.4	28.8 ±2.0	62.2 ±2.4	84.0 ±0.1	50.1 ±0.9	13.7 ±9.8	26.4 ±1.3	52.9
LISA	77.5 ±0.7	57.8 ±0.8	65.8 ±1.5	66.8 ±0.3	79.0 ±0.8	76.9 ±0.7	30.0 ±4.1	66.8 ±0.6	81.6 ±0.6	50.1 ±0.9	37.4 ±3.5	29.8 ±1.8	60.0
ReSample	70.0 ±1.0	74.1 ±2.2	61.0 ±0.6	66.8 ±0.5	81.0 ±1.7	77.7 ±1.1	30.6 ±2.3	67.5 ±0.3	82.6 ±0.6	55.0 ±0.2	74.3 ±0.4	31.4 ±0.6	64.3
ReWeight	71.9 ±0.6	69.6 ±0.2	59.3 ±1.1	64.2 ±1.9	83.1 ±0.7	76.8 ±0.9	25.0 ±0.0	67.0 ±0.4	84.0 ±0.1	61.4 ±1.3	73.7 ±1.0	27.7 ±1.1	63.6
SqrtReWeight	71.0 ±1.4	66.9 ±2.2	68.6 ±1.1	63.8 ±2.4	82.6 ±0.4	76.8 ±0.9	32.8 ±3.5	68.0 ±0.4	83.1 ±0.2	61.2 ±0.6	68.5 ±1.6	27.7 ±1.1	64.2
CBLoss	74.4 ±1.2	65.4 ±1.4	67.3 ±0.2	63.6 ±2.4	83.1 ±0.0	76.8 ±0.9	31.7 ±3.6	67.6 ±0.3	84.0 ±0.1	50.2 ±0.9	74.0 ±0.7	27.7 ±1.1	63.8
Focal	71.6 ±0.8	56.9 ±3.4	61.9 ±1.1	62.4 ±2.0	81.0 ±0.4	71.9 ±1.2	30.6 ±2.3	68.7 ±0.4	70.9 ±9.8	50.0 ±0.9	42.1 ±4.0	26.9 ±0.6	57.9
LDAM	70.9 ±1.7	57.0 ±4.1	28.4 ±7.7	65.5 ±0.8	83.6 ±0.4	76.7 ±0.5	31.7 ±3.6	66.6 ±0.6	81.0 ±0.3	50.1 ±0.9	36.0 ±0.7	24.3 ±0.8	56.0
BSoftmax	74.1 ±0.9	69.6 ±1.2	58.3 ±1.1	63.6 ±2.4	82.6 ±0.4	76.1 ±2.0	35.6 ±1.8	67.6 ±0.6	83.8 ±0.3	58.6 ±1.8	73.8 ±1.0	28.6 ±1.4	64.4
DFR	89.0 ±0.2	73.7 ±0.8	64.4 ±0.1	63.8 ±0.0	81.4 ±0.1	74.4 ±1.8	38.0 ±3.8	67.1 ±0.4	80.2 ±0.0	60.8 ±0.4	75.8 ±0.3	26.3 ±0.4	66.2
CRT	76.3 ±0.8	69.6 ±0.7	67.8 ±0.3	65.4 ±0.2	83.1 ±0.0	78.2 ±0.5	33.3 ±0.0	68.1 ±0.1	83.4 ±0.0	61.8 ±0.1	74.6 ±0.4	31.1 ±0.1	66.1
ReWeightCRT	76.3 ±0.2	70.7 ±0.6	64.7 ±0.2	65.2 ±0.2	85.1 ±0.4	77.5 ±0.7	33.3 ±0.0	67.9 ±0.1	83.4 ±0.0	53.1 ±2.3	75.1 ±0.2	33.1 ±0.1	65.4