

Adapting to Shifting Correlations with Unlabeled Data Calibration

Minh Nguyen¹, Alan Q. Wang¹, Heejong Kim², and Mert R. Sabuncu^{1,2}

¹ Cornell Tech, New York NY 10044, USA {bn244,aw847,msabuncu}@cornell.edu

² Weill Cornell Medicine, New York NY 10065, USA
hek4004@med.cornell.edu

Abstract. Distribution shifts between sites can seriously degrade model performance since models are prone to exploiting unstable correlations. Thus, many methods try to find features that are stable across sites and discard unstable features. However, unstable features might have complementary information that, if used appropriately, could increase accuracy. More recent methods try to adapt to unstable features at the new sites to achieve higher accuracy. However, they make unrealistic assumptions or fail to scale to multiple confounding features. We propose Generalized Prevalence Adjustment (GPA for short), a flexible method that adjusts model predictions to the shifting correlations between prediction target and confounders to safely exploit unstable features. GPA can infer the interaction between target and confounders in new sites using unlabeled samples from those sites. We evaluate GPA on several real and synthetic datasets, and show that it outperforms competitive baselines.

Keywords: Domain generalization · Invariance · EM · Prevalence

1 Introduction

Real-world data from multiple sites (domains) often diverge from the independent and identically distributed (i.i.d.) assumption. Distribution shifts at different sites can cause correlations between (confounding) variables to vary significantly (*i.e.*, instability). For example, different hospitals may use different imaging equipment, resulting in images appearing differently. Some hospitals may adopt specific triaging procedures, resulting in correlations between diagnoses and other variables. ML models often struggle to generalize to different sites (*i.e.*, their performance does not transfer), since models are prone to shortcut learning [17] (*i.e.*, exploit correlations between confounding variables). There have been several examples of shortcut learning in healthcare [2, 13, 40]. Foundation models [24, 37, 42, 52] trained with large amounts of diverse data can generalize out-of-the-box, in many difficult problems. However, generalization to medical imaging is harder because the data distributions and tasks differ significantly from natural images. Besides, training foundation models for medical imaging is also challenging, as gathering diverse data is difficult and costly.

One way to achieve out-of-domain (OOD) generalization is to avoid unstable features and only classify using stable features. Methods such as DANN [15],

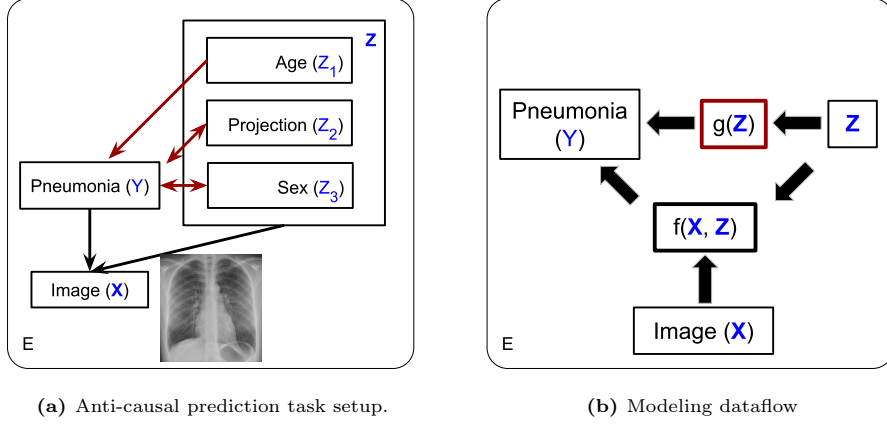


Fig. 1: Left: The generative distribution of image $\mathbf{X}$ (*i.e.*, $P(\mathbf{X}|\mathbf{Y}, \mathbf{Z})$) is stable at different sites while the correlations between prediction target Y and confounding variables $\mathbf{Z}$ can vary. Red edges are unstable (*i.e.* generative mechanisms vary with sites) while black edges are stable. Right: The generative mechanism inspires our two-part modeling approach, where one model (*i.e.*, $f(\mathbf{X}, \mathbf{Z})$) learns the stable mechanism, while another model learns the mechanism that varies with sites (*i.e.*, $g(\mathbf{Z})$). At a new site, only $g(\mathbf{Z})$ needs to be estimated from unlabeled data, while $f(\mathbf{X}, \mathbf{Z})$ can be reused.

CORAL [49], and IRM [4] ensure that no site (domain)-specific features are used for classification. These methods assume that using solely stable features is sufficient to achieve good classification results. However, unstable features might have complementary information, so eliminating them may lower predictive performance. Thus, exploiting unstable features appropriately may improve both in-domain and OOD performance.

Recent adaptive methods try to exploit unstable features at new sites to achieve higher accuracy. To exploit unstable features, it is often necessary to understand the data-generating process and where instability arises. Figure 1a shows a common anti-causal learning setup [46], where the distribution of the input images $\mathbf{X}$ conditioned on the target label Y and confounding variables $\mathbf{Z}$ is stable across sites. However, the joint distribution of Y and $\mathbf{Z}$ may vary from site to site. TTLISA [50] is an adaptive method that estimates this joint distribution at new sites using unlabeled data and uses the estimated joint distribution to re-weight model predictions. However, TTLISA requires marginalizing out the confounding variables and may not be suitable for high-dimensional or continuous $\mathbf{Z}$. CoPA [36] is another adaptive method that estimates the conditional prevalence (*i.e.*, $P(Y|\mathbf{Z})$) instead of the joint distribution. However, CoPA requires access to pairs of $(Y, \mathbf{Z})$ at new sites to estimate the conditional prevalence, which may be unrealistic. Moreover, CoPA also assumes that $\mathbf{Z}$ is observed at new sites, further limiting its applicability.

We propose Generalized Prevalence Adjustment (GPA for short), an adaptive method to achieve better OOD generalization in the anti-causal learning

setup. Similar to CoPA, GPA also learns a stable adaptive predictor [48] by modeling the data-generation process using two separate estimators: one for the stable mechanism and the other for the shifting $P(Y|\mathbf{Z})$ distribution (*i.e.*, f and g in Fig. 1b respectively). However, unlike CoPA, GPA can infer the interaction between the target and the confounders at new sites using only unlabeled samples from those sites. Specifically, GPA can estimate the conditional and marginal prevalence (*i.e.*, $P(Y|\mathbf{Z})$ and $P(Y)$) from unlabeled data at test time. Furthermore, GPA can better predict for samples without $\mathbf{Z}$ at test time even for high-dimensional $\mathbf{Z}$. Our experiments on synthetic and real data show GPA outperforming competitive baselines.

2 Related Work

The challenges of data distribution shifts and leveraging knowledge from one domain to improve performance in another have been addressed with several frameworks, grounded in distinct assumptions [18]. Transfer learning assumes access to labeled data from test site [56]. Domain adaptation relies on the availability of unlabeled data from test sites [38]. Domain generalization, on the other hand, does not rely on any data from the test site [4, 39]. Our approach can benefit from test-site data, even though it does not require it. Hence, our approach is more closely aligned with domain generalization.

Domain-invariant representation learning [27, 32, 51, 54, 58] to find stable features across sites can benefit OOD generalization. However, some domain-invariant learning methods may fail in the presence of label-shift [4, 12, 58]. Some notable methods for domain generalization include adversarial learning [15, 28], CORAL [49], IRM [4], and DRO [45]. IRM [4], which finds invariant causal predictors [21, 35, 39], has been influential in tackling distribution shifts via causal understanding of data generation. More recent methods such as IWDANN [12] and LAMDA [25] try to tackle both domain adaptation and the label-shift problem. However, these were developed for only 2 sites (1 source and 1 target) and might perform poorly in a multi-site setup [36]. Furthermore, some methods are also not applicable when the links between Y and $\mathbf{Z}$ are unstable but causal [36].

Methods relying exclusively on stable features can exhibit limited performance because they ignore unstable features. Recent approaches have emerged to incorporate both types of features to achieve higher accuracy. However, these methods make certain assumptions about the data generation. For example, SFB [14] (i) learns both stable and conditionally-independent unstable features; and (ii) uses the stable-feature predictions to adapt the unstable-feature predictions in the test site using unlabeled data. SFB assumes that stable and unstable features are independent conditioned on the target. Both CoPA [36] and TTLA [50] assume that the distribution of images conditioned on labels and confounders is stable. CoPA [36] models the instability in the conditional distribution $P(Y|\mathbf{Z})$ and additionally assumes that this can be estimated at test site. TTLA [36] models the instability in the joint distribution $P(Y, \mathbf{Z})$ and

estimates this distribution using unlabeled data at test time using Expectation Maximization (EM) [31].

Saerens *et al.*'s landmark work popularized the use EM to adapt to label-shift [44]; yet, EM may not work well if models are poorly calibrated [19]. BBSL [30] and RLLS [5] were proposed as alternatives that can work even when the predictions are not calibrated. However, when calibrated properly, EM-based adaption can perform effectively [3].

3 Conditional Prevalence Adjustment (CoPA)

In this section, we briefly review the CoPA algorithm [36], as our approach will build on it. We use uppercase letters (e.g., $\mathbf{X}, Y$) to denote random variables, while lowercase letters (e.g., $y, \mathbf{z}$) will denote values. Bolded letters (e.g., $\mathbf{X}, \mathbf{z}$) denote high-dimensional variables or values. For brevity, we denote distributions at a specific site e , *i.e.* $P(\cdot|\cdot, E=e)$, as $P_e(\cdot|\cdot)$.

Since $P(\mathbf{X}|Y, \mathbf{Z})$ is assumed to be stable, for any two sites a and b :

$$P_a(\mathbf{X}|Y, \mathbf{Z}) = P(\mathbf{X}|Y, \mathbf{Z}) = P_b(\mathbf{X}|Y, \mathbf{Z}) \quad (1)$$

By applying Bayes' rule to Equation (1):

$$\frac{P_a(Y|\mathbf{X}, \mathbf{Z})P_a(\mathbf{X}|\mathbf{Z})}{P_a(Y|\mathbf{Z})} = \frac{P_b(Y|\mathbf{X}, \mathbf{Z})P_b(\mathbf{X}|\mathbf{Z})}{P_b(Y|\mathbf{Z})} \quad (2)$$

$$\Rightarrow P_b(Y|\mathbf{X}, \mathbf{Z}) = P_b(Y|\mathbf{Z}) \frac{P_a(Y|\mathbf{X}, \mathbf{Z})}{P_a(Y|\mathbf{Z})} \frac{P_a(\mathbf{X}|\mathbf{Z})}{P_b(\mathbf{X}|\mathbf{Z})}. \quad (3)$$

By collapsing all terms that do not depend on Y , Equation (3) can be rewritten as:

$$P_b(Y|\mathbf{X}, \mathbf{Z}) = \text{Norm} \left(P_b(Y|\mathbf{Z}) \frac{P_a(Y|\mathbf{X}, \mathbf{Z})}{P_a(Y|\mathbf{Z})} \right), \quad (4)$$

where **Norm** is the normalization operation so that probabilities sum up to 1. More detailed justification for this step can be found in Appendix A.1. Equation (4) implies that the ratio $P(Y|\mathbf{X}, \mathbf{Z})/P(Y|\mathbf{Z})$ is invariant across sites. Let $f_\theta(\mathbf{X}, \mathbf{Z})$ be an estimator of this ratio parameterized by θ . Let $g_\phi^b(\mathbf{Z})$ be an estimator of $P_b(Y|\mathbf{Z})$ parameterized by ϕ .

During training, CoPA first estimates g_ϕ^e using $(y, \mathbf{z})$ pairs for all training sites e . Then, keeping all g_ϕ^e frozen, CoPA performs maximum likelihood estimation on Eq. (4) with respect to θ . At a new test site b , CoPA first estimates $g_\phi^b(\mathbf{z})$ using $(y, \mathbf{z})$ pairs. Then, given $(\mathbf{x}, \mathbf{z})$ as input, a prediction for y is found by:

$$\hat{y}_b(\mathbf{x}, \mathbf{z}) = \underset{Y}{\operatorname{argmax}} \hat{P}_b(Y|\mathbf{x}, \mathbf{z}) = \operatorname{argmax} \text{Norm}(g_\phi^b(\mathbf{z})f_\theta(\mathbf{x}, \mathbf{z})), \quad (5)$$

where the final **argmax** is taken over the elements of the vector.

4 Proposed Method

The CoPA algorithm outlined in Section 3 makes three assumptions: (1) a stable mechanism for generating $\mathbf{X}$ from label Y and confounders $\mathbf{Z}$, (2) access to confounders $\mathbf{Z}$ at training sites and novel test sites, and (3) access to the conditional prevalence $P_b(Y|\mathbf{Z})$ at novel test sites. In this work, we extend the CoPA algorithm by relaxing these assumptions. In Section 4.2, we will remove assumption (2) and its necessity of $\mathbf{Z}$ at test sites by proposing to let the model learn to predict with missing $\mathbf{Z}$ using input knockout [34]. In Section 4.1, we will remove assumption (3) and the necessity of Y at test sites via the use of the Expectation-Maximization (EM) algorithm [31].

We refer to this algorithm as Generalized Prevalence Adjustment (GPA).

4.1 Conditional Prevalence Estimation without Y

CoPA finds $g_\phi^b(\mathbf{Z})$, the estimator of $P_b(Y|\mathbf{Z})$, *directly* using $(y, \mathbf{z})$ pairs from b . GPA estimates $g_\phi^b(\mathbf{Z})$ *indirectly* using $(\mathbf{x}, \mathbf{z})$ pairs and the EM algorithm [31, 44]. Given data $\{\mathbf{x}^n, y^n, \mathbf{z}^n\}_{n=1}^N$ at a new site, the complete-data log-likelihood is:

$$\ell_b(\mathbf{X}, Y, \mathbf{Z}; \phi) = \sum_{n=1}^N \sum_{i=1}^{|\mathcal{Y}|} \mathbb{I}[y^n=i] \log P_b^\phi(\mathbf{x}^n, y^n=i, \mathbf{z}^n) \quad (6)$$

$$= \sum_{n=1}^N \sum_{i=1}^{|\mathcal{Y}|} \mathbb{I}[y^n=i] \log \left\{ P_b(\mathbf{z}^n) P_b^\phi(y^n=i|\mathbf{z}^n) P(\mathbf{x}^n|y^n=i, \mathbf{z}^n) \right\} \quad (7)$$

$$= \sum_{n=1}^N \sum_{i=1}^{|\mathcal{Y}|} \mathbb{I}[y^n=i] \log P_b^\phi(y^n=i|\mathbf{z}^n) + C, \quad (8)$$

where $\mathbb{I}$ denotes the indicator function. In the last step, we combine all terms which do not include ϕ into a constant C (we drop this term hereafter). Since the true labels y^n are unknown in our setting, we replace the indicator function with its conditional expectation, based on the latest parameter estimate ϕ^t at the t 'th iteration:

$$\mathbb{E}[\mathbb{I}[y^n=i] | \mathbf{x}^n, \mathbf{z}^n; \phi^t] = P_b^{\phi^t}(y^n=i | \mathbf{x}^n, \mathbf{z}^n) = [\text{Norm}(g_{\phi^t}^b(\mathbf{z}^n) f_\theta(\mathbf{x}^n, \mathbf{z}^n))]_i, \quad (9)$$

where $[\cdot]_i$ denotes the i 'th element of the vector.

At the $(t+1)$ 'th iteration, the objective is to maximize $Q(\phi | \phi^t)$ with respect to ϕ , where Q is the expected value of the log-likelihood function of ϕ :

$$Q(\phi | \phi^t) = \mathbb{E}_{Y|\mathbf{X}, \mathbf{Z}, \phi^t}[\ell_b(\mathbf{X}, Y, \mathbf{Z}; \phi)] \quad (10)$$

$$= \sum_{n=1}^N \sum_{i=1}^{|\mathcal{Y}|} P_b^{\phi^t}(y^n=i | \mathbf{x}^n, \mathbf{z}^n) \log P_b^\phi(y^n=i | \mathbf{z}^n) \quad (11)$$

$$= \sum_{n=1}^N \sum_{i=1}^{|\mathcal{Y}|} [\text{Norm}(g_{\phi^t}^b(\mathbf{z}^n) f_\theta(\mathbf{x}^n, \mathbf{z}^n)) \log g_\phi^b(\mathbf{z}^n)]_i. \quad (12)$$

Input: $D^b = \{(\mathbf{x}^n, \mathbf{z}^n)\}_{n=1}^N, f_\theta, T$
Output: g_ϕ^b
Randomly initialize ϕ
for $t = 0, \dots, T-1$ **do**
 $L \leftarrow 0$
 for $n = 1, \dots, N$ **do**
 $\hat{y} \leftarrow \text{Norm}(g_\phi^b(\mathbf{z}^n) f_\theta(\mathbf{x}^n, \mathbf{z}^n))$
 $L \leftarrow L - \sum_i [\hat{y}.\text{detach}() \odot \log g_\phi^b(\mathbf{z}^n)]_i$
 end
 $L.\text{backward}()$
end

Algorithm 1: $P(Y|\mathbf{Z})$ estimation at a new site using $(\mathbf{x}, \mathbf{z})$ pairs and EM. $\odot$ denotes element-wise multiplication. The $\text{detach}()$ call stops gradient flow.

Input: $D^b = \{\mathbf{x}^n\}_{n=1}^N, f_\theta, T$
Output: g_ϕ^b
Randomly initialize ϕ
for $t = 0, \dots, T-1$ **do**
 $L \leftarrow 0$
 for $n = 1, \dots, N$ **do**
 $\hat{y} \leftarrow \text{Norm}(g_\phi^b(\mathbf{z}_0) f_\theta(\mathbf{x}^n, \mathbf{z}_0))$
 $L \leftarrow L - \sum_i [\hat{y}.\text{detach}() \odot \log g_\phi^b(\mathbf{z}_0)]_i$
 end
 $L.\text{backward}()$
end

Algorithm 2: $P(Y)$ estimation at a new site using $(\mathbf{x})$ samples and EM. $\odot$ denotes element-wise multiplication. The $\text{detach}()$ call stops gradient flow. $\mathbf{z}_0$ is a place-holder value used by input knockout.

We implement a stochastic gradient-based optimizer to maximize Q with respect to ϕ (see Algorithm 1). This can be shown to be a generalized version of the EM algorithm [41] for estimating $P_b(Y|\mathbf{X})$ without direct access to Y . In standard EM, the m-step is a closed-form solution. Algorithm 1 is closer to a generalized version of EM since it uses numerical optimization in the m-step so optimum may not be guaranteed. That said, the choice of optimizer in Algorithm 1 influences the behavior of the m-step. When using SGD, $\text{backward}()$ takes one gradient step, so Algorithm 1 is closer to generalized EM. When using LBFGS, $\text{backward}()$ fully maximizes the expectation before the next e-step, so Algorithm 1 is closer to standard EM. Note that a similar approach can be taken to estimate $P_b(Y)$ (see Algorithm 2). That derivation is provided in Appendix A.2.

4.2 Predicting Y without $\mathbf{Z}$ using Input Knockout

Using Eq. (5) to predict Y requires access to $\mathbf{Z}$. However, $\mathbf{Z}$ may be fully or partially missing at some test sites. One way to remove this requirement is marginalizing over $\mathbf{Z}$, *i.e.* $P_b(Y|\mathbf{X}) = \int_{\mathbf{Z}} P_b(Y|\mathbf{X}, \mathbf{Z}) P_b(\mathbf{Z}|\mathbf{X})$. CoPA [36] approximates

Table 1: Summary of CoPA and GPA and their ablated variants.

Name	Input	Need $P_b(Y \mathbf{Z})$	Need $P_b(Y)$	Predict w/o $\mathbf{Z}$	Estimate g_ϕ
CoPA	$(\mathbf{X}, \mathbf{Z})$	yes	no	n/a	Directly
CoPA*	$(\mathbf{X})$	no	yes	Monte-Carlo	Directly
GPA	$(\mathbf{X}, \mathbf{Z})$	no	no	n/a	Indirectly (Alg. 1)
GPA*	$(\mathbf{X})$	no	no	Input knockout	Indirectly (Alg. 2)

this marginalization over $\mathbf{Z}$ by using a Monte-Carlo strategy and assuming that (1) $\mathbf{Z} \mid \mathbf{X}$ is uniformly distributed and (2) $P_b(Y)$ is a good approximation for $P_b(Y|\mathbf{Z})$.

$$\hat{P}_b(Y|\mathbf{X}) \approx \text{Norm}\left(\sum_{\mathbf{Z}} \hat{P}_b(Y|\mathbf{X}, \mathbf{Z})\right) \approx \text{Norm}\left(\sum_{\mathbf{Z}} \text{Norm}(P_b(Y)f_\theta(\mathbf{X}, \mathbf{Z}))\right) \quad (13)$$

CoPA’s approximation in Equation (13) may be intractable or may yield a poor estimate for high-dimensional $\mathbf{Z}$ or continuous $\mathbf{Z}$. Instead, GPA learns to predict Y when $\mathbf{Z}$ is missing. Specifically, during training, the values of $\mathbf{Z}$ are randomly replaced with a default value $\mathbf{z}_0$. Then, during inference,

$$\hat{P}_b(Y|\mathbf{X}) \approx \hat{P}_b(Y|\mathbf{X}, \mathbf{z}_0) = \text{Norm}(g_\phi^b(\mathbf{z}_0)f_\theta(\mathbf{X}, \mathbf{z}_0)) \quad (14)$$

$$\hat{P}_b(Y) \approx \hat{P}_b(Y|\mathbf{z}_0) = g_\phi^b(\mathbf{z}_0), \quad (15)$$

which we refer to as marginalization with input knockout. This “dropout-like” scheme to the input forces the neural networks to model conditional distributions with regard to any possible set of Z variables [6, 7, 23, 29, 34]. The default value $\mathbf{z}_0$ should be outside the support (range) of $\mathbf{Z}$ so that the neural networks know whether to output the full conditional distribution or the marginal distribution [26, 34, 57]. After the model has learned to predict with missing $\mathbf{Z}$, we can also use the model to infer the marginal distribution $P_b(Y)$ as shown in Algorithm 2.

4.3 The GPA Algorithm

Table 1 lists the different instantiations of the algorithms. CoPA* and GPA* are the ablated versions of CoPA and GPA respectively when $\mathbf{Z}$ is withheld from the input (see Section 4.2). Using EM to find the maximum likelihood estimates requires the model to be well-calibrated to achieve good performance [3, 16]. Since neural networks are often not well-calibrated [19], most EM approaches relying on probability outputs from neural networks need a calibration step [3, 16, 50]. We calibrate f_θ using validation data similar to [14, 50]. Algorithm 3 outlines the steps in GPA.

Input: Data from training sites e , $D^e = \{(\mathbf{x}^n, y^n, \mathbf{z}^n)\}^e$
Input: Test data from new site b , D^b
Output: Predicted labels $\{y^n\}^b$

1. Fit $g_\phi^e(\mathbf{Z})$ for each training site e using $\{(y^n, \mathbf{z}^n)\}^e$ and input knockout
2. Fit $f_\theta(\mathbf{X}, \mathbf{Z})$ using Eq. (4) and input knockout
3. Calibrate $f_\theta(\mathbf{X}, \mathbf{Z})$ using held-out validation data
4. **if** D^b has $\mathbf{z}^n$ **then**
 - Estimate $g^b(\mathbf{Z})$ / $P_b(Y|\mathbf{Z})$ using Algorithm 1
 - Predict y^n using Eq. (4)
- else**
 - Estimate $g^b(\mathbf{z}_0)$ / $P_b(Y)$ using Algorithm 2
 - Predict y^n using Eq. (14)

end

Algorithm 3: GPA algorithm**Table 2:** Overview of GPA compared to baselines.

	IRM	DRO	SFB	TTLA	CoPA	GPA (Ours)
Works with single training site?	✗	✓	✓	✓	✓	✓
Works with multiple training sites?	✓	✓	✓	✗	✓	✓
Can Z be continuous?	n/a	✗	n/a	✗	✓	✓
Can Z be high-dimensional?	n/a	✓	n/a	?	✓	✓
Don't need Z at test?	✓	✓	✓	✓	✗	✓
Don't need extra Y at test?	✓	✓	✓	✓	✗	✓

5 Experiments

5.1 Baselines

We compared GPA against ERM, IRM [4], DANN [15], CORAL [49], DRO [45], and CoPA [36]. CoPA was described above in Section 3, and assumes that the labels y are available in the new test site. In the ablation results, CoPA* refers to the variant of CoPA that does not assume access to z at test site and marginalizes out $\mathbf{Z}$ via Monte-Carlo sampling. We provide a description of the baselines below:

- Empirical Risk Minimization (ERM) is the standard scheme that trains a neural network to predict Y , given $\mathbf{X}$ by minimizing cross-entropy loss on all training data and ignoring site information.
- Invariant Risk Minimization (IRM) learns representations that are invariant across different sites or environments. IRM aims to find features that lead to similar predictions across all sites, thus improving generalization to unseen environments.
- Domain-Adversarial Neural Network (DANN) trains a feature extractor to produce features that are indistinguishable between source and target domains, using a domain classifier that tries to distinguish between the two. The goal is to learn representations that are domain-invariant, enhancing the model's ability to generalize across domains.

- Correlation Alignment (CORAL) minimizes domain shift by aligning the second-order statistics (covariances) of source and target domain feature distributions. CORAL seeks to adjust the source domain features to have similar distributions to the target domain, without requiring explicit domain labels.
- Distributionally Robust Optimization (DRO) aims to improve the worst-case performance across a set of potential distributions, enhancing model robustness to shifts and outliers. As unstable correlations include spurious correlations, methods designed for spurious correlations, such as IRM/DRO could be competitive.

We train a couple of ERM variants: one where $\mathbf{X}$ is the only input and another, indicated as ERM^Z , trained with $(\mathbf{x}, \mathbf{z})$ input pairs. In the MNIST experiment, we also have a third ERM model, ERM^c that was trained with grayscale input images, where the spurious color feature was removed. When a baseline depends on assumptions that may be invalid in some experimental setup, we exclude that baseline from the experiment (see Table 2). For example, when there is a single training site, IRM is not applicable. Hence, IRM is excluded from the chest X-ray experiment. For each method, we average results from 5 different runs using different random seeds. Since CORAL and DANN can use unlabeled data to learn invariant representation, they were provided unlabeled data from validation and test sets.

5.2 Implementation Details

The f_θ and g_ϕ are implemented as neural networks. For the architectural details of the networks, please see Appendix C. When g_ϕ is optimized using EM, we used LBFGS as the optimizer and set the maximum number of EM iterations to be 5 (*i.e.*, $T = 5$ in Algorithm 1 and Algorithm 2). For the input knockout in Section 4.2, $\mathbf{z}_0$ is chosen to be outside the range of values of $\mathbf{Z}$. Specifically, for a categorical Z variable with $|\mathcal{Z}|$ possible values, z_0 is $|\mathcal{Z}| + 1$. For a continuous variable Z , the range of value of Z is remapped to be between 1 and 2 and z_0 is chosen to be 0 (*e.g.*, if Z lies in the range $[0, 100]$, a value of 50 is mapped to 1.5).

Two common calibration approaches that show good empirical results are bias-corrected temperature scaling (BCTS) and vector scaling (VS) [3]. We adopt VS [3] to calibrate our model by minimizing the negative log-likelihood on the held-out data from the validation site. The validation site data is also used to select the best model snapshot.

We use F1-score instead of accuracy to evaluate performance due to the class imbalance in the datasets. Standard errors over these runs are indicated with error bars in the figures.

5.3 Simulation - Color MNIST

Data We synthesized digit images based on the MNIST dataset [1]. The image X is determined by Y (randomly sampled from image set with labels $\{5, 6, 7, 8, 9\}$

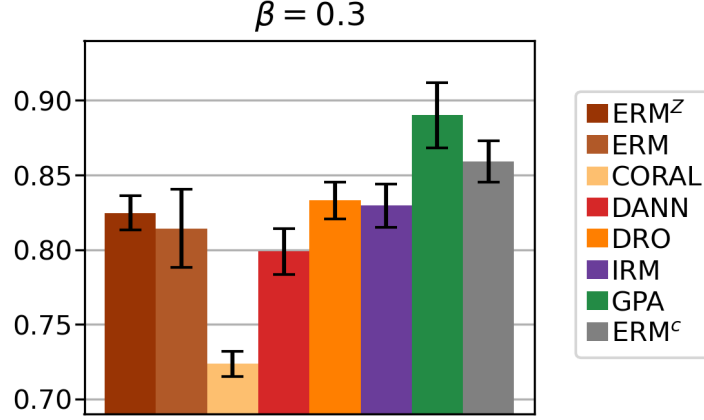


Fig. 2: F1-score at test site. GPA outperforms all the baseline approaches.

when $Y = 1$ and from $\{0, 1, 2, 3, 4\}$ when $Y = 0$) and the color is a confounding variable Z (red for $Z = 1$ and green for $Z = 0$). The Y and Z labels were created following Equation 16-18. Additional results with different causal relationships between Y and Z are shown in Appendix B. $\text{Unif}(0, 1)$ indicates a random variable following a uniform distribution on $(0, 1)$. The value of α is set at 0.3. β is a coefficient within the range $(0, 1)$ that represents “site” in real datasets.

$$S \leftarrow \text{Unif}(0, 1) \quad (16)$$

$$Y \leftarrow \mathbb{I}[\beta S + (1 - \beta)\alpha > 0.5] \quad (17)$$

$$Z \leftarrow \mathbb{I}[\beta S + (1 - \beta)\text{Unif}(0, 1) > 0.5] \quad (18)$$

Two sets, each comprising 10k samples, were used for training, generated with $\beta = 0.9$ and $\beta = 0.7$ respectively. We used 0.5k independent samples created with $\beta = 0.5$ for validation and 1k non-overlapping samples with $\beta = 0.3$ for testing.

Results Figure 2 compares the performance of GPA against baselines, demonstrating that GPA performs the best. Generally, we find that all domain adaptation methods (CoPA and GPA) outperform ERM. GPA is slightly worse than CoPA, since GPA has to indirectly estimate the prevalence, whereas CoPA has direct access to y ’s in the test sample. Furthermore, we compare GPA with and without $\mathbf{Z}$ at test-time by marginalizing it out with input knockout (see Fig. 3). We observed that this decreases performance by 1-2% as expected, but it still outperforms ERM while requiring the same amount of data. In this simulation, Z is a single binary variable and $P_b(Y)$ is a good approximation of $P_b(Y|\mathbf{Y})$, therefore CoPA’s Monte-Carlo approximation is likely reasonable. The Monte-Carlo approximation might be less effective for multivariate Z (see Section 5.4).

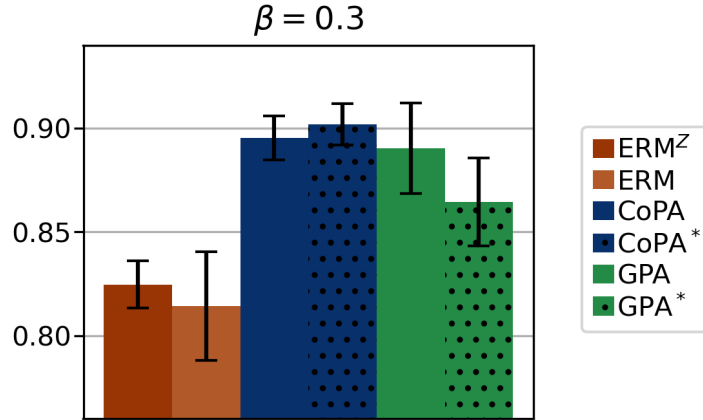


Fig. 3: Color MNIST experiment ablation. F1-scores are shown. * indicates that $\mathbf{Z}$ is missing at test site for that method.

5.4 Real Data Experiments

ISIC Data We replicate the experimental setup used in [36]. We used multi-center skin cancer dataset from the International Skin Imaging Collaboration (ISIC) archive, amounting 70k data samples in total [9–11, 20, 43, 47, 53]. Each sample comprises an input image X , a binary target label Y (melanoma or not) and confounding variables Z . We considered three Z variables correlated with Y : (1) *Age*, (2) *Anatomical Site*, and (3) *Sex*. *Anatomical Site* and *Sex* are categorical variables while *Age* is a continuous variable. Since *Age* is a continuous variable, methods such as DRO or TTSLA which only work for discrete grouping is not directly applicable unless the variable is discretized. However, discretizing *Age* may cause information loss that can be undesirable because *Age* could be predictive of the target (as a causal parent of the target). We grouped the samples using spatio-temporal information (site) and split the sites into the train/validation/test sites as shown in Table 4.

CXR Data We used the Chest X-Ray (CXR) data that consists of CXR8 [55], CheXpert [22], PadChest [8], and VinDR [33]. Each sample includes an input image X , a binary target label Y (pneumonia or not) and confounding variables Z . Three Z variables are considered (1) *Age*, (2) *Projection* (AP, PA, or LL), and (3) *Sex*. *Age* is a continuous variable while *Projection* and *Sex* are categorical variables. The table 3 shows the train/validation/test sites along with their associated marginal prevalence. The VinDR dataset has partially missing $\mathbf{Z}$ which can be quite common in practice since different sites may collect different amounts of meta data. This dataset is a good benchmark for evaluating GPA when $\mathbf{Z}$ is missing. The example images of each dataset are available in the Appendix for reference.

Table 3: Different sites and corresponding marginal prevalence ($P(Y|E)$) in CXR. Underlined: validation site, **bolded**: test site.

Site (E)	CXR8	<u>CheXpert</u>	PadChest	VinDR
No. of samples	26202	5886	4592	13369
$P(Y=1 E)$	0.049	0.635	0.082	0.053
Sex $\in \mathbf{Z}$	Yes	Yes	Yes	Yes
PA/AP $\in \mathbf{Z}$	Yes	Yes	Yes	Yes
Age $\in \mathbf{Z}$	Yes	Yes	Yes	No

Table 4: Different sites in ISIC. The effect of label-shift (change in $P(Y|E)$) is very pronounced between sites. Underlined: validation site, **bolded**: test sites. BCN: Barcelona, MA: Massachusetts, NY: New York, QLD: Queensland, SYD: Sydney, WIE: Vienna

Site (E)	BCN1	BCN2	MA	NY1	<u>NY2</u>	NY3	QLD	SYD	WIE1	WIE2
Samples	7063	7311	9251	11108	1814	3186	8449	1884	7818	4374
$P(Y=1 E)$	0.404	0.024	0.000	0.019	0.146	0.208	0.001	0.071	0.142	0.009

Results Fig. 4 shows the performance of algorithms on test sites NY3 and SYD in the ISIC dataset. Similarly, Fig. 6 shows the performance of algorithms on PadChest and VinDR datasets. We observe that GPA using the EM algorithm and Z as input is comparable to or outperforms baseline algorithms. Across all experiments, we find that GPA generally outperforms CoPA, indicating that the proposed EM strategy leads to better g estimators and potentially less overfitting. Within GPA, we also compare variants with and without Z at adaptation time (solid vs. textured boxes) (See Figs. 5 and 7). Generally, we find that using input knockout as a means to marginalize over Z is an effective strategy, as the difference between variants with and without Z is small.

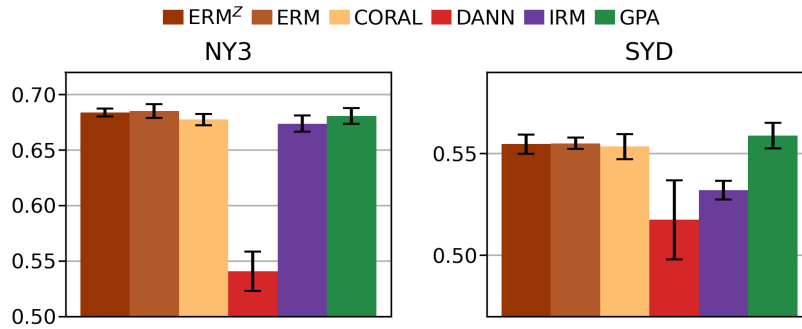


Fig. 4: F1-score at test sites in ISIC experiment. GPA is on par or better than the baseline methods.

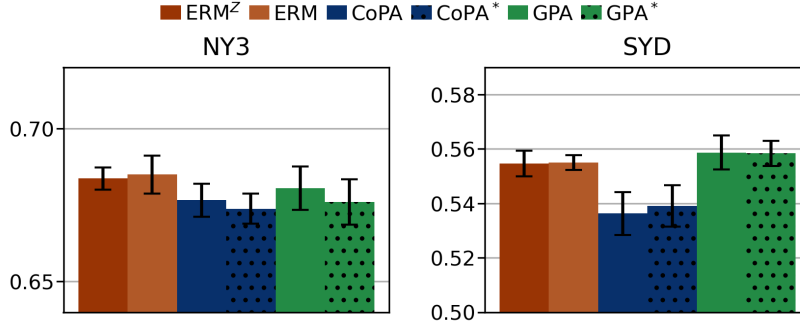


Fig. 5: ISIC experiment ablation. F1 scores are shown. .* indicates that $\mathbf{Z}$ is missing at test site for that method.

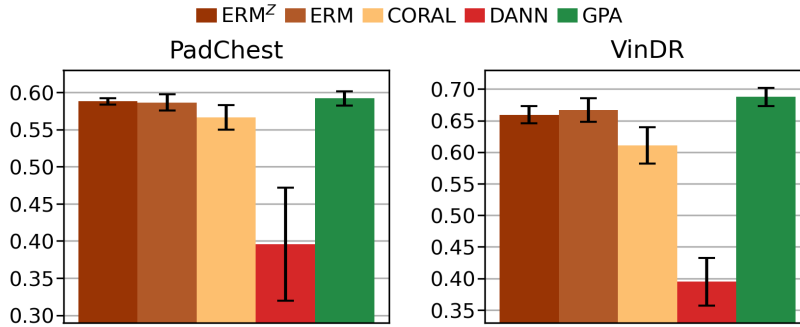


Fig. 6: F1-score at test sites in CXR experiment. GPA outperforms the baseline methods at both test sites.

6 Discussion

In this work, we assume that the mechanism that generates $\mathbf{X}$ from Y and $\mathbf{Z}$ is stable and doesn't change across sites. This stable generation assumption requires knowing all $\mathbf{Z}$ (*i.e.*, it must represent the sufficient adjustment set) which can cause confounding. In the event that not all confounders are captured, it remains to be seen how this can impact performance. One solution may be to learn some latent $\mathbf{Z}$, which is implicitly done in algorithms like IRM and more generally in the causal discovery literature. Another limitation of the algorithm is the requirement that the full set of $\mathbf{Z}$ must be available across all sites during training. Handling the scenario where only a subset of $\mathbf{Z}$ s are available may require modifications to training, for example by drawing from ideas in the missingness literature.

In this work, we fit separate g for each environment. This may be suboptimal in the sense that each g has limited training data. Future work might explore fitting a single model for g , which would allow for better support coverage at the

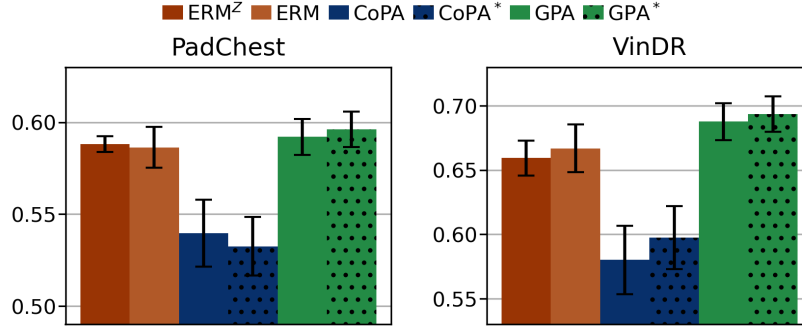


Fig. 7: CXR experiment ablation. F1-scores are shown of different methods when $\mathbf{Z}$ is missing.

potential cost of environment-specific adaptation. Another question of interest is the cost of adaptation. Future work will have to characterize how much unlabeled data is sufficient to perform well in a test site.

7 Conclusion

We present Generalized Prevalence Adjustment, GPA, which is a flexible method for OOD generalization. GPA adaptively adjusts model predictions to the shifting correlations between prediction target and confounders to safely exploit unstable features. It can infer the interaction between target and confounders in new sites using unlabeled samples from those sites. Our experimental results demonstrate the effectiveness of our model on a synthetic dataset and two real-world datasets.

Acknowledgement

Funding for this project was in part provided by the NIH grants R01AG053949, R01AG064027 and R01AG070988, and the NSF CAREER 1748377 grant.

References

1. The mnist database of handwritten digits. <http://yann.lecun.com/exdb/mnist/> 9
2. AlBadawy, E.A., Saha, A., Mazurowski, M.A.: Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. *Medical physics* **45**(3), 1150–1158 (2018) 1
3. Alexandari, A., Kundaje, A., Shrikumar, A.: Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In: *ICML*. pp. 222–232. PMLR (2020) 4, 7, 9

4. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization. Eprint [arXiv:1907.02893](#) (2019) **2**, **3**, **8**
5. Azizzadenesheli, K., Liu, A., Yang, F., Anandkumar, A.: Regularized learning for domain adaptation under label shifts. In: ICLR (2018) **4**
6. Belghazi, M., Oquab, M., Lopez-Paz, D.: Learning about an exponential amount of conditional distributions. NeurIPS **32** (2019) **7**
7. Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., Drouin, A.: Differentiable causal discovery from interventional data. NeurIPS **33**, 21865–21877 (2020) **7**
8. Bustos, A., Pertusa, A., Salinas, J.M., de la Iglesia-Vayá, M.: Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical image analysis **66**, 101797 (2020) **11**
9. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). Eprint [arXiv:1902.03368](#) (2019) **11**
10. Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: Proceedings of ISBI. pp. 168–172. IEEE (2018) **11**
11. Combalia, M., Codella, N.C., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A.C., Puig, S., et al.: Bcn20000: Dermoscopic lesions in the wild. Eprint [arXiv:1908.02288](#) (2019) **11**
12. Tachet des Combes, R., Zhao, H., Wang, Y.X., Gordon, G.J.: Domain adaptation with conditional distribution matching and generalized label shift. In: Proceedings of NeurIPS. vol. 33, pp. 19276–19289 (2020) **3**
13. DeGrave, A.J., Janizek, J.D., Lee, S.I.: Ai for radiographic covid-19 detection selects shortcuts over signal. Nature Machine Intelligence **3**(7), 610–619 (2021) **1**
14. Eastwood, C., Singh, S., Nicolicioiu, A.L., Vlastelica, M., von Kügelgen, J., Schölkopf, B.: Spuriousity Didn’t Kill the Classifier: Using Invariant Predictions to Harness Spurious Features. In: NeurIPS (2023) **3**, **7**
15. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. JMLR **17**(1), 2096–2030 (2016) **1**, **3**, **8**
16. Garg, S., Wu, Y., Balakrishnan, S., Lipton, Z.: A unified view of label shift estimation. NeurIPS **33**, 3290–3300 (2020) **7**
17. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020) **1**
18. Gulrajani, I., Lopez-Paz, D.: In search of lost domain generalization. In: Proceedings of ICLR (2021) **3**
19. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: ICML. pp. 1321–1330. PMLR (2017) **4**, **7**
20. Gutman, D., Codella, N.C., Celebi, E., Helba, B., Marchetti, M., Mishra, N., Halpern, A.: Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). Eprint [arXiv:1605.01397](#) (2016) **11**
21. Heinze-Deml, C., Peters, J., Meinshausen, N.: Invariant causal prediction for non-linear models. Journal of Causal Inference **6**(2) (2018) **3**

22. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghighi, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of AAAI. vol. 33, pp. 590–597 (2019) [11](#)
23. Ke, N.R., Bilaniuk, O., Goyal, A., Bauer, S., Larochelle, H., Schölkopf, B., Mozer, M.C., Pal, C., Bengio, Y.: Neural causal structure discovery from interventions. TMLR (2023) [7](#)
24. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023) [1](#)
25. Le, T., Nguyen, T., Ho, N., Bui, H., Phung, D.: Lamda: Label matching deep domain adaptation. In: Proceedings of ICML. pp. 6043–6054. PMLR (2021) [3](#)
26. Li, S.C.X., Jiang, B., Marlin, B.: Misgan: Learning from incomplete data with generative adversarial networks. In: ICLR (2019) [7](#)
27. Li, Y., Gong, M., Tian, X., Liu, T., Tao, D.: Domain generalization via conditional invariant representations. In: Proceedings of AAAI. vol. 32 (2018) [3](#)
28. Li, Y., Tian, X., Gong, M., Liu, Y., Liu, T., Zhang, K., Tao, D.: Deep domain generalization via conditional invariant adversarial networks. In: Proceedings of ECCV. pp. 624–639 (2018) [3](#)
29. Lippe, P., Cohen, T., Gavves, E.: Efficient neural causal discovery without acyclicity constraints. In: ICLR (2022) [7](#)
30. Lipton, Z., Wang, Y.X., Smola, A.: Detecting and correcting for label shift with black box predictors. In: Proceedings of ICML. pp. 3122–3130. PMLR (2018) [4](#)
31. Moon, T.: The expectation-maximization algorithm. IEEE Signal Processing Magazine **13**(6), 47–60 (1996). <https://doi.org/10.1109/79.543975> [4](#), [5](#)
32. Muandet, K., Balduzzi, D., Schölkopf, B.: Domain generalization via invariant feature representation. In: Proceedings of ICML. pp. 10–18. PMLR (2013) [3](#)
33. Nguyen, H.Q., Lam, K., Le, L.T., Pham, H.H., Tran, D.Q., Nguyen, D.B., Le, D.D., Pham, C.M., Tong, H.T., Dinh, D.H., et al.: Vindr-cxr: An open dataset of chest x-rays with radiologist’s annotations. Scientific Data **9**(1), 429 (2022) [11](#)
34. Nguyen, M., Karaman, B.K., Kim, H., Wang, A.Q., Liu, F., Sabuncu, M.R.: Knock-out: A simple way to handle missing inputs. arXiv preprint arXiv:2405.20448 (2024) [5](#), [7](#)
35. Nguyen, M., Sabuncu, M.R.: Efficient identification of direct causal parents via invariance and minimum error testing (2024) [3](#)
36. Nguyen, M., Wang, A.Q., Kim, H., Sabuncu, M.R.: Robust Learning via Conditional Prevalence Adjustment (2024) [2](#), [3](#), [4](#), [7](#), [8](#), [11](#)
37. OpenAI: GPT-4 Technical Report. arXiv preprint arXiv:2303.08774 (2023) [1](#)
38. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering **22**(10), 1345–1359 (2010) [3](#)
39. Peters, J., Bühlmann, P., Meinshausen, N.: Causal inference by using invariant prediction: identification and confidence intervals. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **78**(5), 947–1012 (2016) [3](#)
40. Pooch, E.H., Ballester, P., Barros, R.C.: Can we trust deep learning based diagnosis? the impact of domain shift in chest radiograph classification. In: Thoracic Image Analysis Workshop. pp. 74–83. Springer (2020) [1](#)
41. Prescher, D.: A tutorial on the expectation-maximization algorithm including maximum-likelihood estimation and em training of probabilistic context-free grammars (2005) [6](#)

42. 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: ICML. pp. 8748–8763. PMLR (2021) [1](#)
43. Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M., Dusza, S., Guitera, P., Gutman, D., et al.: A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific data* **8**(1), 1–8 (2021) [11](#)
44. Saerens, M., Latinne, P., Decaestecker, C.: Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural computation* **14**(1), 21–41 (2002) [4](#), [5](#)
45. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In: Proceedings of ICLR (2019) [3](#), [8](#)
46. Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., Mooij, J.: On causal and anticausal learning. In: Proceedings of ICML. pp. 1255–1262 (2012) [2](#)
47. Scope, A., Marghoob, A., Chen, C., Lieb, J., Weinstock, M., Halpern, A., Group, S.S.: Dermoscopic patterns and subclinical melanocytic nests in normal-appearing skin. *British Journal of Dermatology* **160**(6), 1318–1321 (2009) [11](#)
48. Subbaswamy, A., Chen, B., Saria, S.: A unifying causal framework for analyzing dataset shift-stable learning algorithms. *Journal of Causal Inference* **10**(1), 64–89 (2022) [3](#)
49. Sun, B., Saenko, K.: Deep coral: Correlation alignment for deep domain adaptation. In: Proceedings of ECCV. pp. 443–450. Springer (2016) [2](#), [3](#), [8](#)
50. Sun, Q., Murphy, K., Ebrahimi, S., D’Amour, A.: Beyond Invariance: Test-Time Label-Shift Adaptation for Distributions with "Spurious" Correlations. In: NeurIPS (2023) [2](#), [3](#), [7](#)
51. Tanwani, A.: Dirl: Domain-invariant representation learning for sim-to-real transfer. In: Proceedings of CoRL. pp. 1558–1571. PMLR (2021) [3](#)
52. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023) [1](#)
53. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data* **5**(1), 1–9 (2018) [11](#)
54. Wang, A.Q., Nguyen, M., Sabuncu, M.R.: Learning invariant representations with a nonparametric nadaraya-watson head. In: NeurIPS (2023) [3](#)
55. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of CVPR. pp. 2097–2106 (2017) [11](#)
56. Wilson, G., Cook, D.J.: A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)* **11**(5), 1–46 (2020) [3](#)
57. Yoon, J., Jordon, J., Schaar, M.: Gain: Missing data imputation using generative adversarial nets. In: ICML. pp. 5689–5698. PMLR (2018) [7](#)
58. Zhao, H., Des Combes, R.T., Zhang, K., Gordon, G.: On learning invariant representations for domain adaptation. In: Proceedings of ICML. pp. 7523–7532. PMLR (2019) [3](#)