Supervised Algorithmic Fairness in Distribution Shifts: A Survey

Yujie Lin[1*], **Dong Li**[1*], **Chen Zhao**[2*], **Xintao Wu**[3], **Qin Tian**[1], **Minglai Shao**[1]

[1]Tianjin University
[2]Baylor University
[3]University of Arkansas
{linyujie_22, ld2022244154, tianqin123, shaoml}@tju.edu.cn,
chen_zhao@baylor.edu, xintaowu@uark.edu

Abstract

Supervised fairness-aware machine learning under distribution shifts is an emerging field that addresses the challenge of maintaining equitable and unbiased predictions when faced with changes in data distributions from source to target domains. In real-world applications, machine learning models are often trained on a specific dataset but deployed in environments where the data distribution may shift over time due to various factors. This shift can lead to unfair predictions, disproportionately affecting certain groups characterized by sensitive attributes, such as race and gender. In this survey, we provide a summary of various types of distribution shifts and comprehensively investigate existing methods based on these shifts, highlighting six commonly used approaches in the literature. Additionally, this survey lists publicly available datasets and evaluation metrics for empirical studies. We further explore the interconnection with related research fields, discuss the significant challenges, and identify potential directions for future studies.

1 Introduction

Fairness in machine learning has emerged as a crucial consideration in real-world applications, recognizing the societal impact of algorithmic decision-making. The significance of fairness is particularly evident in scenarios, such as hiring processes, loan approvals, and criminal justice systems, where biased algorithms can perpetuate and exacerbate existing inequalities. Fairness in machine learning refers to the equitable treatment of individuals, irrespective of their sensitive characteristics, such as race and gender. It emphasizes the need to mitigate algorithmic discrimination and promote equal opportunities in model outcomes. Nevertheless, achieving fairness is not devoid of challenges, especially in the presence of distribution shifts. These shifts can pose significant hurdles as models trained on source distributions may not generalize well to target data distributions, potentially exacerbating biases and undermining the intended fairness objectives. Addressing these challenges is essential for advancing the responsible and ethical deployment of machine learning systems in the real world.

There are two main lines of distribution shifts: general and fairness-specific distribution shifts. The former focuses on shifts involving the input features and labels. Covariate shift [Shimodaira, 2000] refers to variations due to differences between the set of marginal distributions over instances. Label shift [Wang *et al.*, 2003], just as its name implies, indicates the changes in the distribution of the class variable. Concept shift [Widmer and Kubat, 1996] refers to "functional relation change" [Yamazaki *et al.*, 2007] due to the change amongst the instance-conditional distributions. On the other hand, approaches addressing fairness-specific shifts consider sensitive attributes, recognizing their significant correlation with fair training. Demographic shift [Giguere *et al.*, 2022] refers to certain sensitive population subgroups becoming more or less probable during inference. Dependence shift [Roh *et al.*, 2023] captures the correlation change between labels and sensitive attributes. Within these distribution shifts, the fairness of the trained model is directly impacted and may deteriorate when adapted to target domains.

To enhance the performance of fairness generalization under distribution shifts, in this survey, we thoroughly examine existing supervised fairness-aware machine learning methods and highlight six commonly used approaches in the literature. Methods falling under feature disentanglement [Zhao *et al.*, 2023b] and data augmentation [Pham *et al.*, 2023] focus on capturing invariance across various domains. Examining causal graphs and paths [Yao and Liu, 2023] aids in comprehending how the model's predictions for different sensitive groups are generated, thereby identifying and addressing potential unfair factors. Adjusting weights of instances or features is a pre-processing method [Roh *et al.*, 2023] that enhances model performance by rectifying unfairness in the target domain. The objective of robust optimization [Du and Wu, 2021] is to minimize the worst-case loss over various subsets of the training set or other well-defined perturbation sets around the data. Regularization-based methods [An *et al.*, 2022] can decrease the correlation between representation and sensitive attributes. It can also enable the transfer of fairness across different domains [Schumann *et al.*, 2019].

Our main contributions of this survey are summarized:

- We summarize a list of different types of distribution shifts and illustrate the effectiveness of generalizing a fairness-

[*] Equal contribution

aware classifier from source to target domains in the context of each distribution shift.

- We categorize existing methods based on different distribution shifts and highlight six main approaches commonly used for handling such shifts.
- We compile a list of the publicly available datasets and survey the literature to identify the most commonly used evaluation metric for quantifying fairness.
- We point out the significant challenges and explore several future directions for study fairness under distribution shifts.

2 Background

Let $\mathcal{X} \subseteq \mathbb{R}^p$ denotes a feature space. $\mathcal{Z} \subset \mathbb{Z}$ is a sensitive space. $\mathcal{Y} \subset \mathbb{Z}$ is defined as an output or a label space. In this survey, we narrow the scope and only concentrate on classification tasks. A domain is defined as a joint distribution $\mathbb{P}_{XZY} := \mathbb{P}(X, Z, Y)$ on $\mathcal{X} \times \mathcal{Z} \times \mathcal{Y}$. A dataset sampled *i.i.d.* from a domain $\mathbb{P}_{XZY}$ is represented as $\mathcal{D} = \{(\mathbf{x}_i, z_i, y_i)\}_{i=1}^{|\mathcal{D}|}$, where $\mathbf{x}, z, y$ are the realizations of random variables X, Z, Y in the corresponding spaces. A classifier parameterized by $\boldsymbol{\theta} \in \Theta$ in a the space $\mathcal{F}$ is denoted as $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Y}$. We denote $\mathcal{E}_{src}$ and $\mathcal{E}_{tgt}$ as sets of domain labels for source and target domains, respectively.

2.1 Supervised Algorithmic Fairness

Algorithmic fairness in supervised machine learning refers to the concern of ensuring that the predictions and decision-making made by a model do not perpetuate existing biases and mitigate discriminatory practices. A fair algorithm aims to seek a classifier trained using data sampled from a source domain that can be generalized to a target domain, ensuring robust predictive performance in model utility (*e.g.,* accuracy) while maintaining fair dependence between outcomes $f_{\boldsymbol{\theta}}(X)$ and the sensitive attribute Z. A crucial aspect of generalizing model fairness is to regulate the $(f_{\boldsymbol{\theta}}(X), Y)$-correlation assessed by fairness notions, which can be broadly categorized into three types: group fairness (**G**), individual fairness (**I**), and counterfactual fairness (**CF**).

Group fairness[1] aims to ensure equitable outcomes for different demographic groups characterized by sensitive attributes. This is often expressed through the lens of demographic parity (DP) and equalized odds (EO) [Hardt *et al.*, 2016], where the conditional probability of a positive outcome for each class is equal across different sensitive subgroups.

$$\begin{aligned} \text{DP:} \quad & \mathbb{P}(f_{\boldsymbol{\theta}}(X)=1|Z=1) = \mathbb{P}(f_{\boldsymbol{\theta}}(X)=1|Z=0) \\ \text{EO:} \quad & \mathbb{P}(f_{\boldsymbol{\theta}}(X)=1|Z=1, Y=y) \\ & = \mathbb{P}(f_{\boldsymbol{\theta}}(X)=1|Z=0, Y=y), \forall y \end{aligned} \tag{1}$$

Although numerous works showcase the effectiveness of generalizing fairness to target domains using group fairness, they fall short in capturing the goal of treating individual people in a fair manner [Dwork *et al.*, 2012]. This limitation has given rise to individual fairness.

Individual fairness [Dwork *et al.*, 2012] focuses on treating similar individuals similarly. In contrast to group fairness, which aims to achieve fairness at the statistical level , individual fairness defines a metric with respect to the particular context and requires that instances that are similar according to such metric receive similar outcomes.

$$d[f_{\boldsymbol{\theta}}(\mathbf{x}_i), f_{\boldsymbol{\theta}}(\mathbf{x}_j)] \leq \delta \cdot D[\mathbf{x}_i, \mathbf{x}_j], \quad i \neq j \tag{2}$$

where $\delta > 0$ is a constant, $d : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and $D : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ denote suitable distance metrics in output and feature spaces, respectively.

Counterfactual fairness [Kusner *et al.*, 2017] is often expressed through the use of causal inference and potential outcomes. It involves evaluating whether a model's predictions would remain fair if certain characteristics of an individual were different while keeping other factors constant. This approach aims to uncover and address unfairness in predictive models by exploring hypothetical situations to ensure equitable outcomes.

$$\begin{aligned} & \mathbb{P}(f_{\boldsymbol{\theta}, Z \leftarrow z}(U) = y | X = \mathbf{x}, Z = z) \\ = & \mathbb{P}(f_{\boldsymbol{\theta}, Z \leftarrow z'}(U) = y | X = \mathbf{x}, Z = z), \forall y \text{ and } z \neq z' \end{aligned} \tag{3}$$

where U denotes latent variables in a given causal model and $f_{\boldsymbol{\theta}, Z \leftarrow z}(U)$ represents the solution for outcomes for a given $U = u$ where the equation for Z are placed with $Z = z$.

In traditional fair machine learning, while many approaches [Corbett-Davies and Goel, 2018] have proven successful, their methods typically assume that source and target data originate from identical distributions. However, it is more realistic to expect methods to operate in non-stationary environments, where the distributions for source and target domains undergo shifts.

2.2 Learning Fairness Under Distribution Shifts

In the context of distribution shift, given access to one or more distinct source domains $\{\mathbb{P}^s_{XZY}\}_{s=1}^S$, where $S = |\mathcal{E}_{src}|$ represents the number of source domains, the objective of learning fairness under distribution shift is to train a fair classifier $f_{\boldsymbol{\theta}}$ using source data. This classifier can generalize well to a distinct target domain $\mathbb{P}^t_{XZY}$ with respect to predicted model utility and fairness, where $t \in \mathcal{E}_{tgt}$ and $\mathbb{P}^t_{XZY} \neq \mathbb{P}^s_{XZY}, \forall s \in \mathcal{E}_{src}$.

Problem 1 (Learning Fairness under Distribution Shifts)**.** *Let $\mathcal{D}_{src} = \{\mathcal{D}^s\}_{s=1}^S$ be a finite set of source data and assume that for each $s \in \mathcal{E}_{src}$, we have access to its corresponding data $\mathcal{D}^s = \{(\mathbf{x}^s_i, z^s_i, y^s_i)\}_{i=1}^{|\mathcal{D}^s|}$ sampled i.i.d from its corresponding domain $\mathbb{P}^s_{XZY}$. Given a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, the goal is to learn a fair classifier $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Y}$ using $\mathcal{D}_{src}$ that simultaneously minimizes the prediction error and mitigates the dependence between predictive outcomes and sensitive attributes. The learned $f_{\boldsymbol{\theta}}$ can be further applied to a target dataset $\mathcal{D}_{tgt}$, sampled form a distinct target domain $\mathbb{P}^t_{XZY}$ that differs from all source domains where $\mathbb{P}^t_{XZY} \neq \mathbb{P}^s_{XZY}, \forall s \in \mathcal{E}_{src}$.*

$$\min_{\boldsymbol{\theta} \in \Theta} \mathbb{E}_{\mathbb{P}^s_{XZY}, \forall s} \ell(f_{\boldsymbol{\theta}}(X^s), Y^s), \ \textit{s.t.}\ \mathbb{E}_{\mathbb{P}^s_{XZY}, \forall s} g(f_{\boldsymbol{\theta}}(X^s), Z^s) \leq \epsilon \tag{4}$$

where $g : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$ denotes a fairness notion, describing the dependence between model outcomes and sensitive

[1] For simplicity, we present notions of group fairness with binary classes and a single binary sensitive attribute.

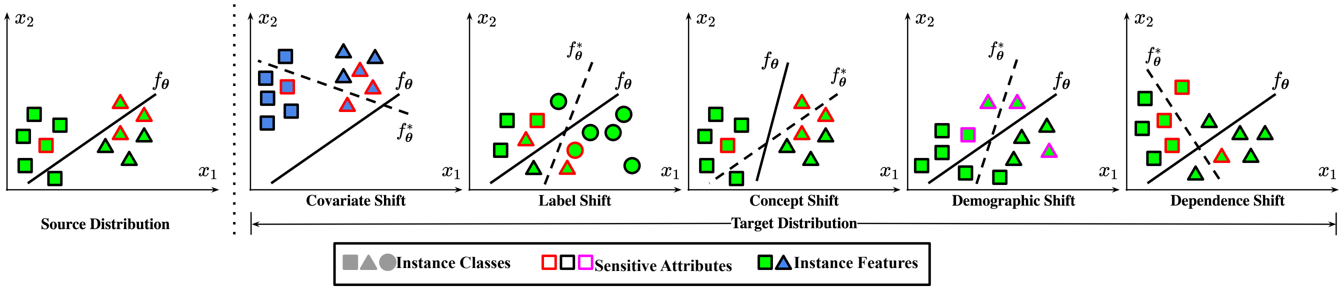

Figure 1: An illustration of fairness-aware machine learning under various distribution shifts. We consider $S = 1$ and $\mathbf{x} = [x_1, x_2]^T$ as a simple example of a two-dimensional feature vector. **(Left)** A fair classifier $f_{\boldsymbol{\theta}}$ is learned using data sampled from a source domain. **(Right)** The learned $f_{\boldsymbol{\theta}}$ is applied to data sampled from various types of shifted target domains, resulting in misclassification and unfairness. $f_{\boldsymbol{\theta}}^*$ represents the true classifier in the target domain.

Table 1: Different Types of Distribution Shifts.

Type of Shifts	**Notations**, $\forall s \in \mathcal{E}_{src}$	**References**
Covariate Shift	$\mathbb{P}_X^s \neq \mathbb{P}_X^t$	[Shimodaira, 2000]
Label Shift	$\mathbb{P}_Y^s \neq \mathbb{P}_Y^t$	[Wang *et al.*, 2003]
Concept Shift	$\mathbb{P}_{Y\|X}^s \neq \mathbb{P}_{Y\|X}^t$	[Widmer and Kubat, 1996]
Demographic Shift	$\mathbb{P}_Z^s \neq \mathbb{P}_Z^t$	[Giguere *et al.*, 2022]
Dependence Shift	$\mathbb{P}_{Y\|Z}^s \neq \mathbb{P}_{Y\|Z}^t$ and $\mathbb{P}_Z^s = \mathbb{P}_Z^t$; or $\mathbb{P}_{Z\|Y}^s \neq \mathbb{P}_{Z\|Y}^t$ and $\mathbb{P}_Y^s = \mathbb{P}_Y^t$	[Roh *et al.*, 2023]
Hybrid Shift	Any combination of the shifts above, see Tab. 2	

attributes. $\epsilon \geq 0$ *is a threshold, which specifies an upper bound on the fair dependence and trades off model utility and fairness.*

As stated in Prob. 1, a major challenge is to train a $f_{\boldsymbol{\theta}}$ that can be well generalized to a target domain from source domains under certain distribution shifts. In the following subsection, we outline various types of shifts and offer a brief analysis regarding fairness generalization for each shift.

2.3 Different Types of Distribution Shifts

An overview of different types of distribution shifts is summarized in Tab. 1 and Fig. 1 with illustrative examples.

Covariate shift [Shimodaira, 2000] refers to changes in the marginal distribution of the input variable X, formally $\mathbb{P}_X^s \neq \mathbb{P}_X^t, \forall s \in \mathcal{E}_{src}$. When the covariate shift occurs, the model's assumptions about the relationships between X and Y may no longer hold, leading to biased predictions. This bias disproportionately affects sensitive subgroups, especially if the training data does not adequately represent their protected characteristics. Consequently, the model may exhibit unfair behavior by making predictions that systematically disadvantage or advantage specific to these subgroups, reinforcing or exacerbating demographic disparities.

Label shift [Wang *et al.*, 2003] , also known as prior probability shift [Saerens *et al.*, 2002] or semantic shift [Newman, 2015], refers to changes in the distribution of the class variable Y, denoted as $\mathbb{P}_Y^s \neq \mathbb{P}_Y^t, \forall s \in \mathcal{E}_{src}$. The challenge of fairness learning under label shift primarily involves mitigating predictive bias in studies related to outlier or out-of-distribution detection. In such cases, samples with unknown classes (*i.e.,* circles in Fig. 1) are present only in the target domain. As a consequence of these samples being unknown to the learned fair classifier $f_{\boldsymbol{\theta}}$, it struggles to preserve the fair correlation between outcomes and sensitive variables while simultaneously maintaining prediction accuracy in distinguishing between known and unknown samples.

Concept shift [Widmer and Kubat, 1996] is defined as changes in among the instance-conditional distributions, expressed as $\mathbb{P}_{Y|X}^s \neq \mathbb{P}_{Y|X}^t, \forall s \in \mathcal{E}_{src}$. It happens when the distribution of Y given X changes, which presents the hardest challenge among the different types of (X, Y)-related shifts that have been tackled so far. In the presence of a structural causality graph between X, Z, and Y where $Z \rightarrow (X, Y)$ and $X \rightarrow Y$, a concept shift from source to target domains leads to a change of the conditional distribution $\mathbb{P}_{Y|Z}$ through X. As a result, the generalization of the fair classifier fails.

Demographic shift [Giguere *et al.*, 2022] refers to changes in the marginal distribution of the sensitive variable Z between source and target domains, denoted $\mathbb{P}_Z^s \neq \mathbb{P}_Z^t, \forall s \in \mathcal{E}_{src}$. In particular, a demographic shift occurs when a specific sensitive subgroup of the population becomes more or less probable in the target domain. As demonstrated in Fig. 1, specifically in the context of the demographic shift, both source and target domains consist of samples with three sensitive attributes (indicated by distinct border colors in black, red, and magenta). However, the number of samples indicated by magenta and red in the source and target domains is zero, respectively. Although $f_{\boldsymbol{\theta}}$ provides fair predictions in the source domain for the black and red groups, it might struggle to generalize fairness to the target domain in the presence of a demographic shift.

Dependence shift [Roh *et al.*, 2023] can explicitly capture the alteration in the correlation between Y and Z of the data across the source and target domains. In [Zhao *et al.*, 2023b], the authors attribute the dependence shift to the changes in the conditional distribution involving Y and z. Therefore, we formulate this shift with two similar alternative cases, $\mathbb{P}_{Y|Z}^s \neq \mathbb{P}_{Y|Z}^t$ and $\mathbb{P}_Z^s = \mathbb{P}_Z^t$, or $\mathbb{P}_{Z|Y}^s \neq \mathbb{P}_{Z|Y}^t$ and $\mathbb{P}_Y^s = \mathbb{P}_Y^t$. Notice that *Roh et al.,* [2023] indicate handling dependence shift is straightforward when it is greater in the source domain than in the target domain. Otherwise, $f_{\boldsymbol{\theta}}$ may fail.

Other types of shifts. The aforementioned distribution shifts can be further categorized based on whether they undergo a corresponding change in space. The space change across domains requires a change in the corresponding distribution, but the opposite may not hold. For example, changes in the label space $\mathcal{Y}^s \neq \mathcal{Y}^t, \forall s$ and sensitive space $\mathcal{Z}^s \neq \mathcal{Z}^t, \forall s$ indicate the introduction of new labels and new sensitive attributes, which requires label shift and demographic shift, respectively. However, these shifts with invariant cor-

Table 2: An overview of existing approaches in fairness machine learning under various types of distribution shifts.

	Reference	Distribution Shifts[2]					Spaces Change, $s \to t, \forall s \in \mathcal{E}_{src}$			$S > 1$	Accessibility of $\mathcal{D}_{tgt}$	Fairness Type[3]	Method
		Cov.	**Lab.**	**Con.**	**Dem.**	**Dep.**	$\mathcal{X}^s \neq \mathcal{X}^t$	$\mathcal{Y}^s \neq \mathcal{Y}^t$	$\mathcal{Z}^s \neq \mathcal{Z}^t$				
Covariate Shift	[Taskesen *et al.*, 2020]	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	G	RO
	[Rezaei *et al.*, 2020]	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	G	RO
	[Creager *et al.*, 2020]	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗	G	CI
	[Rezaei *et al.*, 2021]	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓	G	RW
	[Du and Wu, 2021]	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓	G	RO
	[Zhao *et al.*, 2021]	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	G	RG
	[Zhao *et al.*, 2022]	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	G	RW
	[Pham *et al.*, 2023]	✓	✗	✗	✗	✗	✓	✗	✗	✓	✗	G	DA, RG
	[Zhao *et al.*, 2023b]	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	G	FD, RG
	[Lin *et al.*, 2023a]	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	CF	FD, RG
Label Shift	[Biswas and Mukherjee, 2021]	✗	✓	✗	✗	✗	✗	✗	✗	✗	✓	G	RO
Concept Shift	[Iosifidis *et al.*, 2019]	✗	✗	✓	✗	✗	✗	✗	✗	✗	✓	G	RW
	[Iosifidis and Ntoutsi, 2020]	✗	✗	✓	✗	✗	✗	✗	✗	✗	✓	G	OT
Demographic Shift	[Schumann *et al.*, 2019]	✗	✗	✗	✓	✗	✗	✗	✓	✗	✓	G	RG
	[Giguere *et al.*, 2022]	✗	✗	✗	✓	✗	✗	✗	✓	✗	✓	G	OT
Dependence Shift	[Creager *et al.*, 2021]	✗	✗	✗	✗	✓	✗	✗	✗	✓	✗	I	RO
	[Oh *et al.*, 2022]	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	I	FD
	[Roh *et al.*, 2023]	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓	G	RW
Hybrid Shift	[Kallus and Zhou, 2018]	✓	✗	✓	✗	✗	✗	✗	✗	✗	✓	G	RW
	[Singh *et al.*, 2021]	✓	✗	✗	✓	✗	✗	✗	✗	✗	✓	G	CI
	[Schrouff *et al.*, 2022]	✓	✓	✗	✓	✗	✗	✗	✗	✗	✓	G	CI
	[An *et al.*, 2022]	✓	✓	✗	✓	✓	✓	✗	✗	✗	✓	G	RG, RW
	[Chen *et al.*, 2022]	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓	G	RO
	[Zhao *et al.*, 2023a]	✓	✗	✗	✗	✓	✓	✗	✗	✓	✗	G	FD, DA, RG
	[Han *et al.*, 2023]	✓	✓	✗	✗	✗	✗	✓	✗	✗	✗	CF	CI

[2] The abbreviations stand for covariate (Cov.), label (Lab.), concept (Con.), demographic (Dem.), and dependence (Dep.) shifts.
[3] The abbreviations stand for group (G), individual (I), and counterfactual (CF) fairness.

responding spaces refer to alterations in the proportions of instances generated from respective domains. A change in $\mathcal{X}$ denotes a shift in feature variation, such as transitioning from photo images to cartoons.

3 Methods

We classify existing methods for addressing algorithmic fairness in distribution shifts into six categories: feature disentanglement (**FD**), data augmentation (**DA**), causal inference (**CI**), reweighting (**RW**), robust optimization (**RO**), and regularization-based approaches (**RG**). Any methods not falling within these categories are designated as others (**OT**).

Feature disentanglement aims to learn latent representations of data features, enhancing their clarity and mutual independence within the model. The primary objective of this process is to enable the model to better comprehend the structure and variations present in the data. This endeavor involves disentangling the data representations into different latent spaces, facilitating a more nuanced understanding of the underlying data structure and dynamics. For the issue of fairness under distribution shifts, we obtain high-dimensional representations, denoted as $R = h(X)$, from the data, where $h : \mathcal{X} \to \mathcal{R}$ represents a high-dimensional mapping. This method aims to preserve only the semantic information unrelated to the sensitive attribute Z for the final fair decision-making process. A commonly employed approach involves utilizing VAE-based methods for disentanglement [Oh *et al.*, 2022], and some methodologies leverage structures resembling autoencoders [Zhao *et al.*, 2023a]. The advantage of feature disentanglement lies in its pronounced interpretability. However, directly evaluating the quality of the disentanglement poses challenges.

Data augmentation aims to enhance the diversity of training datasets and improve model generalization performance by systematically applying controlled transformations to the training data. The fundamental idea behind this approach is to generate new samples that are similar but slightly different from the original data by applying various transformations. Additionally, one can also employ generative models (*e.g.*, GANs) to train a generator [Pham *et al.*, 2023] initially. This generator is designed to ensure that the generated representation R satisfies the equality conditions: $\mathbb{P}^i_{R|Y} = \mathbb{P}^j_{R|Y}$ and $\mathbb{P}^i_{R|Y,Z} = \mathbb{P}^j_{R|Y,Z}$ for any source domains $\forall i, j \in \mathcal{E}_{src}, i \neq j$.

Moreover, this generator can be utilized to generate synthetic data for the purpose of data augmentation. Data augmentation can also be combined with feature disentanglement. Initially, semantic factors, domain-specific factors, and sensitive factors are disentangled from the data. By randomly sampling a new set of domain-specific and sensitive factors from their prior distributions, along with the original semantic information, a decoder is employed to generate synthetic data [Zhao *et al.*, 2023a]. While data augmentation can enhance the diversity of the training set, enabling the model to generalize better to unseen target domains and thereby improving overall performance, its efficacy is highly contingent upon the chosen transformation strategies.

Causal Inference. Causal models have been widely applied in machine learning to address issues related to model fairness. Structural Causal Models (SCMs) provide a means of explaining machine learning model predictions. Analyzing causal graphs and paths helps understand how the model's predictions for different groups are formed, thereby identifying and addressing potential unfair factors. A typical fairness causal graph comprises three relational expressions: $Z \rightarrow X$, $Z \rightarrow Y$, and $Z, X \rightarrow Y$. In the context of various distribution shifts, this is equivalent to applying *do*-operator to corresponding variables, where an intervention on X (denoted $do(X)$) serves the purpose of resetting the distribution of X. One can apply the *do*-operator separately to X, Z, and Y to generate counterfactual samples, thereby calculating training sample influence on the model's unfairness [Yao and Liu, 2023]. Similarly, one can analyze the causal dependencies among X, Z, and Y and their impact on the convergence speed of model training when transitioning to a new distribution [Lin *et al.*, 2023b]. If only the $do(Z)$ operation is performed, the investigation of counterfactual fairness issues can be conducted. By setting Z to a value different from its original state (*e.g.,* transitioning from male to female), changes in Y may be induced through the propagation of causal paths. It aims for the impact of $do(Z)$ on the predicted values to be minimized [Han *et al.*, 2023].

Reweighting. The reweighting approach is a commonly used pre-processing method in statistics and machine learning. It is employed to adjust the weights of samples or features to enhance model performance or correct biases in the dataset. These methods typically involve assigning weights to instances from the source domain to represent the overall distribution of the target domain. One related research employs a sample reweighting technique, where the impact of each instance is adjusted using the classical covariate shift instance weighing function to estimate accuracy metrics [Kallus and Zhou, 2018]. This results in the spurious fairness policy being consistently and uniformly more stringent than necessary for the disadvantaged class. To improve fair training under dependence shift, a decoupling framework was introduced, wherein pre-processing is employed to modify the correlation between sensitive attributes and labels [Roh *et al.*, 2023]. And then in-processing techniques are subsequently utilized to ensure equal weighting of samples within each (y, z)-class.

Robust optimization. A different strategy for extending the model's generalization beyond the training distribution is through robust optimization, wherein the objective is to minimize the worst-case loss over various subsets of the training set or other well-defined perturbation sets around the data. Most of these approaches operate within the framework of distributionally robust optimization (DRO). DRO aims to minimize the worst-case training loss over any uncertainty set of distribution , which is proximate to the training distribution according to a specified metric.

These methodologies are typically framed as a minimax problem, where the objective is to minimize the loss concerning the most adverse realization of perturbations within the uncertainty set . For instance, a double minimax iterative algorithm was introduced, wherein the set is defined by weighted perturbations of the empirical training distribution [Mandal *et al.*, 2020]. *Du and Wu* [2021] introduces a fairness constraint in both the minimization and maximization problems, with its uncertainty set defined as a Wasserstein ball centered around the uniform selection ratio. The uncertainty set also can be defined as the intersection of distributions within a Wasserstein ball centered at the empirical distribution [Taskesen *et al.*, 2020].

Regularization-based approaches. Fairness regularization can be broadly approached through two main strategies. The first involves introducing additional regularization terms to diminish the correlation between high-dimensional representations R and sensitive attributes Z. For example, *An et al.,* [2022] adopts adversarial learning to hinder R from encoding Z. Additionally, to ensure fairness and accuracy in the target domain under various shifts, the researchers utilize pseudo-labels generated by a teacher model as supervision for consistent training. This involves training the model to maintain consistent predictions under various transformations. Another approach involves employing regularization terms, such as Maximum Mean Discrepancy (MMD), to align distributions. In cases where the distribution of sensitive attributes $\mathbb{P}_Z$ differs between the source and target domains, with high-dimensional representations encoding sensitive attributes $\hat{Z}$ and domain labels $\hat{E}$, fairness transfer can be achieved by aligning the distributions between $\mathbb{P}_Z$ and $\mathbb{P}_{\hat{Z}}$, as well as $\mathbb{P}_E$ and $\mathbb{P}_{\hat{E}}$ [Schumann *et al.*, 2019].

Other methods. In addition to the methods mentioned above, there are other approaches to enhance the performance of fairness generalization under distribution shifts. Differing from reweighting methods employed during the preprocessing stage, *decision boundary adjustment* [Iosifidis and Ntoutsi, 2020] is introduced as a post-processing approach. This method ensures fairness under concept shift in an online learning setting. Cumulative discrimination is observed in a streaming environment, and the decision boundary is adjusted when it exceeds a threshold to meet fairness requirements. An universal training framework was developed based on the *Seldonian algorithm* [Thomas *et al.*, 2019] to ensure fairness under demographic shifts. The training data was divided into two parts: one for training candidate fair classifiers using existing fairness algorithms and the other for calculating a high-confidence upper bound on the prevalence of unfair behavior. This allowed for the evaluation of candidate models to obtain models robust to demographic shifts.

4 Datasets and Evaluation

4.1 Datasets

We summarize the most commonly used publicly available datasets in Tab. 3.

Tabular datasets. `UCI Adult` [Kohavi and others, 1996] includes nearly 48,842 adults and generates a binary income label by determining whether an individual's income exceeds \$50k. Gender is commonly chosen as the sensitive attribute. `COMPAS` [Dressel and Farid, 2018] consists of 6,167 samples, and its objective is to predict an individual's recidivism based on criminal history. The dataset assigns race as the sensitive attribute. `UCI Adult` and `COMPAS` are also employed for fair anomaly detection, where the binary label indicates whether an instance is anomalous or not. `German` [Asuncion and Newman, 2007] comprises a collection of 1,000 samples used for credit scoring prediction of whether an individual is a good or bad credit risk based on various financial features. The dataset includes a binary-sensitive attribute indicating the individual's gender along with other rel-

Table 3: Common datasets for fairness under distribution shifts.

Type	Dataset	# Samples	Feature Dimension	Reference	Shifts	Domain Characteristic	Sensitive Attribute	Class Label
Tabular	UCI Adult	48,842	14	[Du and Wu, 2021]	Cov.	features	gender	income
				[Han *et al.*, 2023]	Lab.	InD / OOD	gender	InD / OOD
				[Iosifidis and Ntoutsi, 2020]	Con.	time	gender	income
				[Yoon *et al.*, 2020]	Dem.	gender & race	gender & race	income
	COMPAS	6,167	9	[Rezaei *et al.*, 2021]	Cov.	features	race	recidivism
				[Han *et al.*, 2023]	Lab.	InD / OOD	race	InD / OOD
				[Biswas and Mukherjee, 2021]	Lab.	recidivism	race	recidivism
				[Yoon *et al.*, 2020]	Dem.	gender & race	gender & race	recidivism
	German	1,000	20	[Rezaei *et al.*, 2021]	Cov.	features	gender	credit
				[Yoon *et al.*, 2020]	Dem.	gender & age	gender & age	credit
	NYSF	311,367	16	[Iosifidis and Ntoutsi, 2020]	Con.	years	gender	sespect arrested
		685,724	112	[Kallus and Zhou, 2018]	Cov.	features	race	weapon found
				[Zhao *et al.*, 2023a]	Cov. & Dep.	city & (Y, Z)-correlation	race	stop record
Image	cMNIST	70,000	28×28×3	[Creager *et al.*, 2021]	Dep.	(Y, Z)-correlation	digit color	digit group
	rcMNIST	10,000	28×28×3	[Zhao *et al.*, 2023b]	Cov.	rotated angle	digit color	digit group
	ccMNIST	70,000	28×28×3	[Zhao *et al.*, 2023a]	Cov. & Dep.	digit color & (Y, Z)-correlation	background color	digit group
	Waterbirds	4,795	not fixed	[Creager *et al.*, 2021]	Dep.	(Y, Z)-correlation	background	bird's breed
	FairFace	108,501	224×224×3	[An *et al.*, 2022]	Cov.	data variation	race	gender
				[Zhao *et al.*, 2023a]	Cov. & Dep.	race & (Y, Z)-correlation	gender	age
	UTKFace	23,708	128×128×3	[An *et al.*, 2022]	Cov.	data variation	race	gender
Text	BOLD	23,679	N/A	[Dhamala *et al.*, 2021]	Cov.	topic	gender	sentiment

evant features. New York Stop-and-Frisk (NYSF) [Koh *et al.*, 2021] is a real dataset from the policing activities in New York City. It spans multiple years involving 685,724 records. Due to evident racial bias against African Americans, race is considered a sensitive attribute. Data in different cities serve as distinct domains, and various inspection records can be used as labels, such as stop record, weapon possession, or arrest status. NYSF is also utilized for online settings to address concept shift due to its temporal characteristics [Iosifidis and Ntoutsi, 2020]. In this context, gender is considered a sensitive attribute.

Image datasets. Colored-MNIST (cMNIST) [Arjovsky *et al.*, 2019] purposely establishes a correlation between digits and digit colors in the source data but intentionally anti-correlates them in the target one. Rotated-Colored-MNIST (rcMNIST) [Zhao *et al.*, 2023b] is extended from the cMNIST, where each digit is associated with different rotated angles representing domains. For fairness concerns, digit colors are the sensitive attribute correlated to labels. Corlored-Corlored-MNIST (ccMNIST) [Arjovsky *et al.*, 2019] is created by colorizing digits and the backgrounds of the MNIST [LeCun *et al.*, 2010] dataset. ccMNIST contains three domains characterized by a different digit color with a different correlation between the class label (same as cMNIST) and sensitive attribute (background colors). Waterbirds [Sagawa *et al.*, 2019] is created by merging bird images with backgrounds. The label indicates the bird's breed, while the sensitive attribute corresponds to the background type. Similar to cMNIST, there exists a spurious correlation between the label and sensitive attribute, generating a dependence shift between the source and target domains. FairFace [Kärkkäinen and Joo, 2019] contains 108,501 facial images across seven racial categories. A deployment for fairness in distribution shifts is to set the racial groups as domains, gender as the sensitive attribute, and age as class label [Zhao *et al.*, 2023a]. UTKFace [Zhang *et al.*, 2017] comprises 23,708 facial images. Each image is annotated with information regarding the subject's age, gender, and ethnicity. In this dataset, race is considered a sensitive attribute, while gender is the class label.

Text datasets. BOLD [Dhamala *et al.*, 2021] comprises 23,679 English text generation prompts designed to assess societal biases in open-ended language generation, specifically applied to sentiment analysis tasks. The prompts are categorized into five domains based on their topics: profession, gender, race, religion, and political ideology. Gender is considered a binary sensitive attribute within this context.

4.2 Evaluation Metrics for Assessing Fairness

Metrics for Group Fairness. Three fundamental metrics are used for evaluating group fairness. *Difference of Demographic Parity* (Δ_{DP}) [Dwork *et al.*, 2012] and *Difference of Equalized Odds* (Δ_{EO}) [Hardt *et al.*, 2016] are similar to the notions presented in Eq. (1), where Δ_{DP} and Δ_{EO} take the absolute probability difference in DP and EO. *Difference of Equalized Opportunity* (Δ_{EOp}) [Hardt *et al.*, 2016] is similar to Δ_{EO}. The difference lies in that Δ_{EOp} only requires true positive rates (TPRs) across sensitive subgroups.

$$\begin{aligned}\Delta_{EOp} =& |\mathbb{P}(f_{\boldsymbol{\theta}}(X)=1|Z=1,Y=1) \\ &-\mathbb{P}(f_{\boldsymbol{\theta}}(X)=1|Z=0,Y=1)|\end{aligned} \tag{5}$$

A value of zero for these metrics indicates fair predictions in a target domain.

Metrics for Individual Fairness. Due to the need to measure the similarity between samples, there are relatively few relevant metrics, and *consistency* [Zemel *et al.*, 2013] stands out as one of the more classical ones. Given an input $\mathbf{x}$ to its k-nearest neighbors, $kNN(\mathbf{x})$, consistency is formulated:

$$\text{consistency} = 1 - \frac{1}{Nk}\sum_{n}\left|f_{\boldsymbol{\theta}}(\mathbf{x}_n) - \sum_{j\in kNN(\mathbf{x}_n)} f_{\boldsymbol{\theta}}(\mathbf{x}_j)\right| \tag{6}$$

where N represents the total number of samples, and $\mathbf{x}_n$ denotes the n-th sample. A value close to 1 represents fairness.

Metrics for Counterfactual Fairness. As counterfactual fairness is a method defined by causal structures, one needs to use the do-operator to evaluate it. Assuming A is the intervention target of the do-operator, Y is influenced by this intervention. The post-intervention distribution of Y can be expressed as $\mathbb{P}(y_a) = \mathbb{P}(Y = y|do(A = a))$. Currently, there are primarily two types of metrics.

The *Total Causal Effect* (TCE) [Pearl, 2009] of the value change of Z from z to z' on $Y = y$ is given by

$$\mathrm{TCE}(z, z') = |\mathbb{P}(y_z) - \mathbb{P}(y_{z'})|. \tag{7}$$

Given context $O = \mathbf{o}$, the *Counterfactual Effect* (CE) [Shpitser and Pearl, 2008] of the value change of Z from z to z' on $Y = y$ is given by

$$\mathrm{CE}(z, z'|\mathbf{o}) = |\mathbb{P}(y_z|\mathbf{o}) - \mathbb{P}(y_{z'}|\mathbf{o})|. \tag{8}$$

Smaller TCE and CE indicate that the prediction results are more stable in the counterfactual generation of changing the sensitive attribute, implying greater fairness.

5 Related Research Fields

Unsupervised algorithmic fairness under distribution shifts focuses on addressing unfairness in machine learning models when data labels are not available, and the model needs to adapt to changes in the data distribution. As the currently few unsupervised fairness outlier detection methods, *Song et al.* [2021] leverage deep clustering to discover the intrinsic cluster structure and out-of-structure instances. Meanwhile, adversarial training erases the sensitive pattern for instances of fairness adaptation. *Coston et al.* [2019] investigate unsupervised domain adaptation for covariate shift between a source and target distribution, highlighting the unavailability of information from the target domain. They employ reweighting to learn weights closely approximating covariate shift weights, defined solely by non-sensitive attributes available in both domains under statistical parity constraints on the source data.

Fairness-aware outlier detection. Unlike conventional classification tasks, outlier detection aims to identify samples where covariate shifts have occurred. Under the influence of sensitive attributes, samples from certain groups are more prone to being identified as outlier samples. *Shekhar et al.* [2021] decompose the fairness in outlier detection into two issues: DP and EOp. They design regularizations to address both problems. Based on the framework of LOF, FairLOF [Deepak and Abraham, 2021] attempts to correct for such kNN neighborhood distance disparities across object groups defined over sensitive attributes. Similarly, based on LOF and adversarial networks, in contrast to FairLOF, AFLOF [Li *et al.*, 2022] learns the optimal representation of the original data while concealing the sensitive attribute in the data.

6 Challenges and Future Directions

Fairness under conditional shift. In fact, it is deemed appropriate to align $\mathbb{P}_X$ only when X is the causative factor of Y. However, it is plausible that Y serves as the cause of X, leading to an impact on $\mathbb{P}_{Y|X}$ when there are shifts in $\mathbb{P}_X$. The variation of $\mathbb{P}_{Y|X}$ in an unknown domain is referred to as *conditional shift* [Liu *et al.*, 2021]. Consequently, certain domain alignment approaches [Liu *et al.*, 2021; Li *et al.*, 2018] have proposed aligning the class-conditional distribution $\mathbb{P}_{X|Y}$, under the assumption that $\mathbb{P}_Y$ remains unchanged. We believe fairness issues under conditional shift are a promising area for exploration.

Fairness under concept shift. Existing solutions address concept shift by dynamically adjusting the learning model online. However, the model's adaptability, leading to changes in the decision boundaries of classifiers, may compromise the fairness of the model. Some approaches address this issue by modifying input data [Iosifidis *et al.*, 2019] or dynamically adjusting the model distribution online to align with fairness objectives [Iosifidis and Ntoutsi, 2020]. However, in the conventional setting, it lacks a continuous data stream, making it challenging to discern the patterns of change in $\mathbb{P}_{Y|X}$. In cases where there is a significant concept shift between the source and target domains, addressing both fairness and accuracy in the target domain becomes formidable.

Fairness in out-of-distribution detection. While existing work [Han *et al.*, 2023] ensures the absence of anomalous samples during the training process and guarantees fairness in anomaly detection during testing, out-of-distribution detection also necessitates addressing the classification of in-distribution samples. In other words, fairness concerns in out-of-distribution detection involve simultaneously considering multi-class fairness for in-distribution samples and binary-class fairness for out-of-distribution sample detection. Therefore, in the design of the model, it is imperative to balance fairness considerations for both aspects.

Multiple source domains. The majority of the investigated works in this survey under consideration typically limit their scope to a single source domain. However, in real-world scenarios, data often exhibits a multifaceted origin, stemming from diverse sources. Learning across multiple source domains enables the model to acquire more robust representations, rendering it less sensitive to variations and noise inherent in a singular source domain.

Dataset specifically designed for fairness learning under distribution shifts. While many datasets in the fields of algorithmic fairness or out-of-distribution scenarios have been utilized for fairness under distribution shifts, there is still a lack of a dataset explicitly designed for this task. Moreover, existing datasets lack a clear definition of the distribution shifts within them, creating a gap between algorithm design and experimental analysis. To address this issue more effectively, the design of a real-world dataset dedicated to this domain, capable of encompassing various types of distribution shifts, is urgently needed.

7 Conclusion

Ensuring the generalization of model fairness under distribution shifts across domains has emerged as a crucial research focus in recent years. In this survey, we initially compile a list of diverse distribution shifts, offering a brief discussion on the reasons why a predicted model may fail to adapt to each setting. Additionally, we offer a comprehensive review of existing methods addressing fairness generalization in various distribution shifts. Finally, we identify several challenges and suggest potential avenues for future research.

References

[An *et al.*, 2022] Bang An, Zora Che, Mucong Ding, and Furong Huang. Transferring fairness under distribution shifts via fair consistency regularization. *NeurIPS*, 35:32582–32597, 2022.

[Arjovsky *et al.*, 2019] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv:1907.02893*, 2019.

[Asuncion and Newman, 2007] Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

[Biswas and Mukherjee, 2021] Arpita Biswas and Suvam Mukherjee. Ensuring fairness under prior probability shifts. In *AIES*, pages 414–424, 2021.

[Chen *et al.*, 2022] Yatong Chen, Reilly Raab, Jialu Wang, and Yang Liu. Fairness transferability subject to bounded distribution shift. *NeurIPS*, 35:11266–11278, 2022.

[Corbett-Davies and Goel, 2018] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv:1808.00023*, 2018.

[Coston *et al.*, 2019] Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. Fair transfer learning with missing protected attributes. In *AIES*, pages 91–98, 2019.

[Creager *et al.*, 2020] Elliot Creager, David Madras, Toniann Pitassi, and Richard Zemel. Causal modeling for fairness in dynamical systems. In *ICML*. PMLR, 2020.

[Creager *et al.*, 2021] Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *ICML*, pages 2189–2200. PMLR, 2021.

[Deepak and Abraham, 2021] Parakkal Deepak and Savitha Sam Abraham. Fairlof: fairness in outlier detection. *Data Science and Engineering*, 2021.

[Dhamala *et al.*, 2021] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *FAccT*, pages 862–872, 2021.

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

[Du and Wu, 2021] Wei Du and Xintao Wu. Fair and robust classification under sample selection bias. In *CIKM*, pages 2999–3003, 2021.

[Dwork *et al.*, 2012] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012.

[Giguere *et al.*, 2022] Stephen Giguere, Blossom Metevier, Yuriy Brun, Bruno Castro da Silva, Philip S Thomas, and Scott Niekum. Fairness guarantees under demographic shift. In *ICLR*, 2022.

[Han *et al.*, 2023] Xiao Han, Lu Zhang, Yongkai Wu, and Shuhan Yuan. Achieving counterfactual fairness for anomaly detection. In *PAKDD*. Springer, 2023.

[Hardt *et al.*, 2016] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *NeurIPS*, 29, 2016.

[Iosifidis and Ntoutsi, 2020] Vasileios Iosifidis and Eirini Ntoutsi. Fabboo-online fairness-aware learning under class imbalance. In *International Conference on Discovery Science*, pages 159–174. Springer, 2020.

[Iosifidis *et al.*, 2019] Vasileios Iosifidis, Thi Ngoc Han Tran, and Eirini Ntoutsi. Fairness-enhancing interventions in stream classification. In *DEXA*, pages 261–276. Springer, 2019.

[Kallus and Zhou, 2018] Nathan Kallus and Angela Zhou. Residual unfairness in fair machine learning from prejudiced data. In *ICML*, pages 2439–2448. PMLR, 2018.

[Kärkkäinen and Joo, 2019] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv:1908.04913*, 2019.

[Koh *et al.*, 2021] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *ICML*, pages 5637–5664. PMLR, 2021.

[Kohavi and others, 1996] Ron Kohavi et al. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *KDD*, volume 96, pages 202–207, 1996.

[Kusner *et al.*, 2017] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *NeurIPS*, 30, 2017.

[LeCun *et al.*, 2010] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[Li *et al.*, 2018] Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *AAAI*, volume 32, 2018.

[Li *et al.*, 2022] Shu Li, Jiong Yu, Xusheng Du, Yi Lu, and Rui Qiu. Fair outlier detection based on adversarial representation learning. *Symmetry*, 14(2):347, 2022.

[Lin *et al.*, 2023a] Yujie Lin, Chen Zhao, Minglai Shao, Baoluo Meng, Xujiang Zhao, and Haifeng Chen. Pursuing counterfactual fairness via sequential autoencoder across domains. *arXiv:2309.13005*, 2023.

[Lin *et al.*, 2023b] Yujie Lin, Chen Zhao, Minglai Shao, Xujiang Zhao, and Haifeng Chen. Adaptation speed analysis for fairness-aware causal models. In *CIKM*, 2023.

[Liu *et al.*, 2021] Xiaofeng Liu, Bo Hu, Linghao Jin, Xu Han, Fangxu Xing, Jinsong Ouyang, Jun Lu, Georges EL Fakhri, and Jonghye Woo. Domain generalization under conditional and label shifts via variational bayesian inference. *arXiv:2107.10931*, 2021.

[Mandal *et al.*, 2020] Debmalya Mandal, Samuel Deng, Suman Jana, Jeannette Wing, and Daniel J Hsu. Ensuring fairness beyond the training data. *NeurIPS*, 2020.

[Newman, 2015] John Newman. Semantic shift. *The Routledge handbook of semantics*, pages 266–280, 2015.

[Oh *et al.*, 2022] Changdae Oh, Heeji Won, Junhyuk So, Taero Kim, Yewon Kim, Hosik Choi, and Kyungwoo Song. Learning fair representation via distributional contrastive disentanglement. In *KDD*, 2022.

[Pearl, 2009] Judea Pearl. *Causality*. Cambridge university press, 2009.

[Pham *et al.*, 2023] Thai-Hoang Pham, Xueru Zhang, and Ping Zhang. Fairness and accuracy under domain generalization. *arXiv:2301.13323*, 2023.

[Rezaei *et al.*, 2020] Ashkan Rezaei, Rizal Fathony, Omid Memarrast, and Brian Ziebart. Fairness for robust log loss classification. In *AAAI*, 2020.

[Rezaei *et al.*, 2021] Ashkan Rezaei, Anqi Liu, Omid Memarrast, and Brian D Ziebart. Robust fairness under covariate shift. In *AAAI*, 2021.

[Roh *et al.*, 2023] Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. Improving fair training under correlation shifts. *ICML*, 2023.

[Saerens *et al.*, 2002] Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural computation*, 14(1):21–41, 2002.

[Sagawa *et al.*, 2019] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv:1911.08731*, 2019.

[Schrouff *et al.*, 2022] Jessica Schrouff, Natalie Harris, Sanmi Koyejo, Ibrahim M Alabdulmohsin, Eva Schnider, Krista Opsahl-Ong, Alexander Brown, Subhrajit Roy, Diana Mincu, Christina Chen, et al. Diagnosing failures of fairness transfer across distribution shift in real-world medical settings. *NeurIPS*, 35:19304–19318, 2022.

[Schumann *et al.*, 2019] Candice Schumann, Xuezhi Wang, Alex Beutel, Jilin Chen, Hai Qian, and Ed H Chi. Transfer of machine learning fairness across domains. *arXiv:1906.09688*, 2019.

[Shekhar *et al.*, 2021] Shubhranshu Shekhar, Neil Shah, and Leman Akoglu. Fairod: Fairness-aware outlier detection. In *AIES*, pages 210–220, 2021.

[Shimodaira, 2000] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 2000.

[Shpitser and Pearl, 2008] Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *JMLR*, 2008.

[Singh *et al.*, 2021] Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. Fairness violations and mitigation under covariate shift. In *FAccT*, 2021.

[Song *et al.*, 2021] Hanyu Song, Peizhao Li, and Hongfu Liu. Deep clustering based fair outlier detection. In *SIGKDD*, pages 1481–1489, 2021.

[Taskesen *et al.*, 2020] Bahar Taskesen, Viet Anh Nguyen, Daniel Kuhn, and Jose Blanchet. A distributionally robust approach to fair classification. *arXiv:2007.09530*, 2020.

[Thomas *et al.*, 2019] Philip S Thomas, Bruno Castro da Silva, Andrew G Barto, Stephen Giguere, Yuriy Brun, and Emma Brunskill. Preventing undesirable behavior of intelligent machines. *Science*, 366(6468):999–1004, 2019.

[Wang *et al.*, 2003] Ke Wang, Senqiang Zhou, Chee Ada Fu, and Jeffrey Xu Yu. Mining changes of classification by correspondence tracing. In *Proceedings of the 2003 SIAM International Conference on Data Mining*. SIAM, 2003.

[Widmer and Kubat, 1996] Gerhard Widmer and Miroslav Kubat. Learning in the presence of concept drift and hidden contexts. *Machine learning*, 23:69–101, 1996.

[Yamazaki *et al.*, 2007] Keisuke Yamazaki, Motoaki Kawanabe, Sumio Watanabe, Masashi Sugiyama, and Klaus-Robert Müller. Asymptotic bayesian generalization error when training and test distributions are different. In *ICML*, pages 1079–1086, 2007.

[Yao and Liu, 2023] Yuanshun Yao and Yang Liu. Understanding unfairness via training concept influence. *arXiv:2306.17828*, 2023.

[Yoon *et al.*, 2020] Taeho Yoon, Jaewook Lee, and Woojin Lee. Joint transfer of model knowledge and fairness over domains using wasserstein distance. *IEEE Access*, 8:123783–123798, 2020.

[Zemel *et al.*, 2013] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *ICML*. PMLR, 2013.

[Zhang *et al.*, 2017] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *CVPR*, pages 5810–5818, 2017.

[Zhao *et al.*, 2021] Chen Zhao, Feng Chen, and Bhavani Thuraisingham. Fairness-aware online meta-learning. In *SIGKDD*, pages 2294–2304, 2021.

[Zhao *et al.*, 2022] Chen Zhao, Feng Mi, Xintao Wu, Kai Jiang, Latifur Khan, and Feng Chen. Adaptive fairness-aware online meta-learning for changing environments. In *SIGKDD*, pages 2565–2575, 2022.

[Zhao *et al.*, 2023a] Chen Zhao, Kai Jiang, Xintao Wu, Haoliang Wang, Latifur Khan, Christan Grant, and Feng Chen. Fairness-aware domain generalization under covariate and dependence shifts. *arXiv:2311.13816*, 2023.

[Zhao *et al.*, 2023b] Chen Zhao, Feng Mi, Xintao Wu, Kai Jiang, Latifur Khan, Christan Grant, and Feng Chen. Towards fair disentangled online learning for changing environments. In *SIGKDD*, pages 3480–3491, 2023.