Supervised Algorithmic Fairness in Distribution Shifts: A Survey

Minglai Shao\textsuperscript{1}, Dong Li\textsuperscript{2}, Chen Zhao\textsuperscript{3}, Xintao Wu\textsuperscript{4}, Yujie Lin\textsuperscript{1}, Qin Tian\textsuperscript{2}

\textsuperscript{1}School of New Media and Communication, Tianjin University, China
\textsuperscript{2}College of Intelligence and Computing, Tianjin University, China
\textsuperscript{3}Department of Computer Science, Baylor University, USA
\textsuperscript{4}Electrical Engineering and Computer Science Department, University of Arkansas, USA

\{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 \textit{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 \textit{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 \textit{et al.}, 2022] refers to certain sensitive population subgroups becoming more or less probable during inference. Dependence shift [Roh \textit{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 \textit{et al.}, 2023b] and data augmentation [Pham \textit{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 \textit{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 \textit{et al.}, 2022] can decrease the correlation between representation and sensitive attributes. It can also enable the transfer of fairness across different domains [Schumann \textit{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-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. $\mathcal{Z} \subseteq \mathcal{Z}$ is a sensitive space. $\mathcal{Y} \subseteq \mathcal{Z}$ 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 $P_{XZY} := P(X, Z, Y)$ on $\mathcal{X} \times \mathcal{Z} \times \mathcal{Y}$. A dataset sampled i.i.d. from a domain $P_{XZY}$ is represented as $D = \{\{x_i, z_i, y_i\}\}_{i=1}^{|D|}$, where $x, z, y$ are the realizations of random variables $X, Z, Y$ in the corresponding spaces. A classifier parameterized by $\theta \in \Theta$ in a the space $F$ is denoted as $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$. We denote $E_{src}$ and $E_{tgt}$ 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 $f_\theta(X)$ and the sensitive attribute $Z$. A crucial aspect of generalizing model fairness is to regulate the $(f_\theta(X), Y)$-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** 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.

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

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.

$$d(f_\theta(x_i), f_\theta(x_j)) \leq \delta \cdot D[x_i, x_j], \quad i \neq j \quad (2)$$

where $\delta > 0$ is a constant, $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ 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.

$$P(f_\theta(z \leftarrow x, U = u)(Y = y|X = x, Z = z)) = P(f_\theta(U = u)(Y = y|X = x, Z = z), \forall y \text{ and } z \neq z') \quad (3)$$

where $U$ denotes latent variables in a given causal model and $f_\theta(z \leftarrow x, U = u)(Y = y|X = x, Z = z)$ 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 $\{P_{XZY}\}_{i=1}^S$, where $S = |E_{src}|$ represents the number of source domains, the objective of learning fairness under distribution shift is to train a fair classifier $f_\theta$ using source data. This classifier can generalize well to a distinct target domain $P_{XZY}$ with respect to predicted model utility and fairness, where $t \in E_{tgt}$ and $P_{XZY} \neq P_{XZY}^{\prime}, \forall s \in E_{src}$.

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

$$\min_{\theta \in \Theta} \mathbb{E}_{P_{XZY}^{\prime}, s \in E_{src}}[\ell(f_\theta(x^s), Y^s), s.t. \mathbb{E}_{P_{XZY}^{\prime}, s \in E_{src}}[g(f_\theta(x^s), Z^s)] \leq \epsilon \quad (4)$$

where $g : \mathcal{Y} \times \mathcal{Z} \rightarrow \mathbb{R}$ denotes a fairness notion, describing the dependence between model outcomes and sensitive attributes.
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 $f_\theta$ 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 Sec. 2.2 and Fig. 1 with illustrative examples. **Covariate shift** [Shimodaira, 2000] refers to changes in the marginal distribution of the input variable $X$, formally $P^s_X \neq P^t_X, \forall s \in E_{src}$. 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 $P^s_Y \neq P^t_Y, \forall s \in E_{src}$. 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 $f_\theta$, 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 $P^s_{Y|X} \neq P^t_{Y|X}, \forall s \in E_{src}$. 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 $P^t_{Y|Z}$ 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 $P^s_Z \neq P^t_Z, \forall s \in E_{src}$. 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 $f_\theta$ 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, $P^s_{Y|Z} \neq P^t_{Y|Z}$ and $P^s_Z = P^t_Z$, or $P^s_{Z|Y} \neq P^t_{Z|Y}$ and $P^s_Y = P^t_Y$. 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, $f_\theta$ may fail.

Other types of shifts. The aforementioned distribution shifts can be further categorized based on whether they un-
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: \( \mathbb{P}_{R[Y]}^i = \mathbb{P}_{R[Y]}^j \) and \( \mathbb{P}_{R[Y,Z]}^i = \mathbb{P}_{R[Y,Z]}^j \) 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., 2023]. 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 weighting function to estimate accuracy metrics [Kallus and Zhou, 2018]. To improve fair training under dependence shift, a decoupling framework was introduced, wherein preprocessing 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 $\mathbb{Z}$ and domain labels $\mathbb{E}$, fairness transfer can be achieved by aligning the distributions between $\mathbb{P}_Z$ and $\mathbb{P}_E$, as well as $\mathbb{P}_E$ and $\mathbb{P}_E$ [Schumel 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

**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 relevant 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 the 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. Corolored-Colored-MNIST (ccMNIST) [Arjovsky et al., 2019] is created by colorizing digits and the backgrounds of the MNIST 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.

**Graph datasets.** Pokecs [Dai and Wang, 2022] comprises 134,563 graphs extracted from a popular social network in Slovakia, organized according to the users’ respective provinces. Different domains within the dataset are composed of users from two major regions within the corresponding provinces, with the region being treated as a sensitive attribute. Bail-Bs dataset, consisting of nodes representing defendants on bail, is derived from commonly used fair-related graphs obtained from Bail, with race being a sensitive attribute. Credit-Cs is derived from Credit, wherein nodes represent credit card users, following the same partitioning scheme as Bail-Bs. The task involves categorizing customers’ credit risks as high or low based on age as the sensitive attribute.

<table>
<thead>
<tr>
<th>Type</th>
<th>Dataset</th>
<th># Samples</th>
<th>Feature Dimension</th>
<th>Reference</th>
<th>Shifts</th>
<th>Domain Characteristic</th>
<th>Sensitive Attribute</th>
<th>Class Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tabular</td>
<td>UCI Adult</td>
<td>48,842</td>
<td>14</td>
<td>[Du and Wu, 2021] [Han et al., 2023] [Iosifidis and Ntoutsi, 2020] [Yoon et al., 2020] [Rezaei et al., 2021]</td>
<td>Cov., Lab., Dem.</td>
<td>Cove., gender &amp; race, recidivism</td>
<td>gender &amp; race, income</td>
<td>InD / OOD</td>
</tr>
<tr>
<td></td>
<td>COMPAS</td>
<td>6,167</td>
<td>9</td>
<td>[Han et al., 2023] [Biswas and Mukherjee, 2021]</td>
<td>Cov., Lab., Dem.</td>
<td>Cove., gender &amp; race, recidivism</td>
<td>race, recidivism</td>
<td>race</td>
</tr>
<tr>
<td></td>
<td>German</td>
<td>1,000</td>
<td>20</td>
<td>[Yoon et al., 2020]</td>
<td>Cov., Dem.</td>
<td>Cove., gender &amp; age, credit</td>
<td>gender &amp; age, credit</td>
<td>credit</td>
</tr>
<tr>
<td></td>
<td>NYSF</td>
<td>311,367</td>
<td>16</td>
<td>[Iosifidis and Ntoutsi, 2020]</td>
<td>Cov.</td>
<td>Cove., years, gender, race, weapon found</td>
<td>race, weapon found</td>
<td>race</td>
</tr>
<tr>
<td></td>
<td>UTKFace</td>
<td>70,000</td>
<td>28 × 28 × 3</td>
<td>[Creager et al., 2021] [Zha et al., 2023a] [Zhao et al., 2023b]</td>
<td>Cov., Dep.</td>
<td>Dep. (Y, Z)-correlation, digit color, digit group</td>
<td>background color</td>
<td>digit group</td>
</tr>
<tr>
<td>Image</td>
<td>rcMNIST</td>
<td>10,000</td>
<td>28 × 28 × 3</td>
<td>[Zha et al., 2023a]</td>
<td>Cov. &amp; Dep.</td>
<td>Dep. (Y, Z)-correlation, background color, digit group</td>
<td>background color</td>
<td>digit group</td>
</tr>
<tr>
<td></td>
<td>ccMNIST</td>
<td>70,000</td>
<td>28 × 28 × 3</td>
<td>[Zha et al., 2023a]</td>
<td>Cov. &amp; Dep.</td>
<td>Dep. (Y, Z)-correlation, background color, digit group</td>
<td>background color</td>
<td>digit group</td>
</tr>
<tr>
<td></td>
<td>FairFace</td>
<td>4,793</td>
<td>not fixed</td>
<td>[Creager et al., 2021]</td>
<td>Cov. &amp; Dep.</td>
<td>Dep. (Y, Z)-correlation, background color, digit group</td>
<td>background color</td>
<td>digit group</td>
</tr>
<tr>
<td>Text</td>
<td>Pokecs</td>
<td>134,656</td>
<td>265</td>
<td>[Li et al., 2024]</td>
<td>Cov.</td>
<td>Cove., province, region, working field</td>
<td>region</td>
<td>region</td>
</tr>
<tr>
<td></td>
<td>Bail-Bs</td>
<td>120,24</td>
<td>18</td>
<td>[Li et al., 2024]</td>
<td>Cov.</td>
<td>Cove., community, race, bail</td>
<td>race</td>
<td>bail</td>
</tr>
<tr>
<td></td>
<td>Credit-Cs</td>
<td>16,009</td>
<td>13</td>
<td>[Li et al., 2024]</td>
<td>Cov.</td>
<td>Cove., community, age, risk</td>
<td>age, risk</td>
<td>risk</td>
</tr>
</tbody>
</table>

Table 3: Common datasets for fairness under distribution shifts.
4.2 Evaluation Metrics for Assessing Fairness

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

\[
\Delta_{EOp} = |P(f_0(X) = 1|Z = 1, Y = 1) - P(f_0(X) = 1|Z = 0, Y = 1)|
\]

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 \(x\) to its \(k\)-nearest neighbors, \(kNN(x)\), consistency is formulated:

\[
\text{consistency} = 1 - \frac{1}{N_k} \sum_n \left| f_0(x_n) - \sum_{j \in kNN(x_n)} f_0(x_j) \right|
\]

where \(N\) represents the total number of samples, and \(x_n\) 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 \(P(y_a) = P(Y = y|do(A = a))\). 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

\[
\text{TCE}(z, z') = |P(Y = y|do(A = a)) - P(Y = y)|.
\]

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

\[
\text{CE}(z, z'|o) = |P(y|do(o)) - P(y|do(o'))|.
\]

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.** 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 \(P_X\) 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 \(P_{Y|X}\) when there are shifts in \(P_X\). The variation of \(P_{Y|X}\) in an unknown domain is referred to as *conditional shift* [Liu et al., 2021]. Consequently, certain domain alignment approaches [Liu et al., 2021] have proposed aligning the class-conditional distribution \(P_{X|Y}\) under the assumption that \(P_Y\) remains unchanged. We believe fairness issues under conditional shift are a promising area for exploration.

**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.

7 Conclusion

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.

Acknowledgments

This work is supported by the National Natural Science Foundation of China program (NSFC #62272338).

Contribution Statement

Minglai Shao, Dong Li and Chen Zhao contribute equally. Yujie Lin and Minglai Shao are corresponding authors. This work is supervised by Chen Zhao and Minglai Shao.
References


[Lin et al., 2024] Yujie Lin, Chen Zhao, Minglai Shao, Baoluo Meng, Xujiang Zhao, and Haifeng Chen. Towards counterfactual fairness-aware domain generalization in changing environments. IJCAI, 2024.


