Learning for Counterfactual Fairness from Observational Data

Jing Ma
University of Virginia
jm3mr@virginia.edu

Ruocheng Guo
Bytedance Research
ruocheng.guo@bytedance.com

Aidong Zhang
University of Virginia
aidong@virginia.edu

Jundong Li
University of Virginia
jundong@virginia.edu

ABSTRACT

Fairness-aware machine learning has attracted a surge of attention in many domains, such as online advertising, personalized recommendation, and social media analysis in web applications. Fairness-aware machine learning aims to eliminate biases of learning models against certain subgroups described by certain protected (sensitive) attributes such as race, gender, and age. Among many existing fairness notions, counterfactual fairness is a popular notion defined from a causal perspective. It measures the fairness of a predictor by comparing the prediction of each individual in the original world and that in the counterfactual worlds in which the value of the sensitive attribute is modified. A prerequisite for existing methods to achieve counterfactual fairness is the prior human knowledge of the causal model for the data. However, in real-world scenarios, the underlying causal model is often unknown, and acquiring such human knowledge could be very difficult. In these scenarios, it is risky to directly trust the causal models obtained from information sources with unknown reliability and even causal discovery methods, as incorrect causal models can consequently bring biases to the predictor and lead to unfair predictions. In this work, we address the problem of counterfactually fair prediction from observational data without given causal models by proposing a novel framework CLAIRE. Specifically, under certain general assumptions, CLAIRE effectively mitigates the biases from the sensitive attribute with a representation learning framework based on counterfactual data augmentation and an invariant penalty. Experiments conducted on both synthetic and real-world datasets validate the superiority of CLAIRE in both counterfactual fairness and prediction performance.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Mathematics of computing** → **Causal networks**; • **Applied computing** → *Law, social and behavioral sciences.*

KEYWORDS

Counterfactual Fairness; Causal Model; Sensitive Attributes

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08.
https://doi.org/10.1145/3580305.3599408

ACM Reference Format:
Jing Ma, Ruocheng Guo, Aidong Zhang, and Jundong Li. 2023. Learning for Counterfactual Fairness from Observational Data. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3580305.3599408

1 INTRODUCTION

Recent years have witnessed a rapid development of machine learning based prediction [10, 14, 44] in various high-impact applications such as personalized recommendation [36, 51], ranking in searches [17, 40], and social media analysis [1, 32]. Recent literatures [7] have shown that the predictions based on traditional machine learning often exhibit biases against certain demographic subgroups that are described by certain protected attributes (a.k.a. sensitive attributes) such as race, gender, age, and sexual orientation. Thus, how to develop a *fair* predictor has attracted a surge of attentions [5, 9, 20, 22, 49, 54, 55]. Among them, the seminal work of *counterfactual fairness* [30] makes use of the causal mechanism to model how discrimination is exhibited, and eliminates it at the individual level based on the Pearl's causal structural models [39]. The intuition of counterfactual fairness is to encourage the predictions made from different versions of the same individual to be equal. For example, the predictions for "in an online talent search, how would a certain candidate be ranked if this candidate had been a male/female?" should be identical to achieve the notion of counterfactual fairness.

A prerequisite of existing methods to achieve counterfactual fairness is the prior human knowledge of causal models. A causal model [38, 39] typically consists of a causal graph and the corresponding structural equations that describe the causal relationships among different variables. Existing works on counterfactual fairness [30, 42, 52, 53] overwhelmingly rely on the assumption that the underlying causal model is (at least partially) known and correct, in order to mitigate the biases across different sensitive subgroups. However, existing work often suffers from the following major limitation: In real world, the underlying causal model is often unknown, especially when the data is high-dimensional [6, 50]. The construction of a trustworthy causal model often requires knowledge from domain experts, which is expensive in both time and labor. In addition, it is extremely challenging to validate the correctness of the obtained causal model. Without external guidance of human knowledge, other existing works mostly rely on causal discovery techniques [23, 26, 31, 38, 46, 47] to learn the causal model from observational data, but these methods can suffer from various mistakes in discovering the causal relations, and thus lead the predictor to pick up biased information of the sensitive attribute [37].

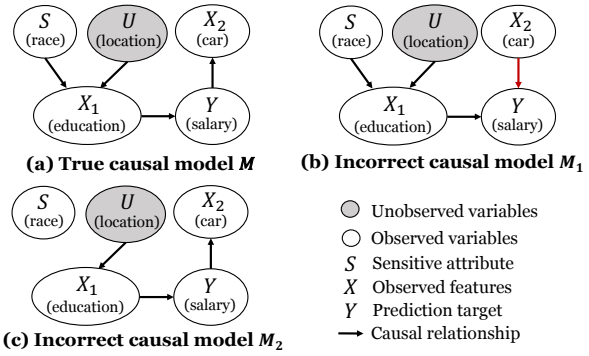


Figure 1: An illustrative example of incorrect causal models.

Here, the toy example in Fig. 1 intuitively explains two scenarios with incorrect causal models. Fig. 1(a) shows an example of a true causal model (often determined by domain experts) in which we aim to predict the salary (prediction target Y) of people in different races (described by the sensitive attribute S). We assume that the level of education (observed feature X_1) of each person is a cause, and the salary also influences the type of car each person would like to purchase (observed feature X_2). Unobserved variables U (e.g., geographic location) could also have a causal effect on the observed variables. To learn a counterfactually fair predictor, most existing works [30, 42] utilize a given causal model, and only use those variables which are not causally influenced by the sensitive attribute (i.e., non-descendants of S) for prediction. We now consider two cases when the given causal model is incorrect: 1) Consider an incorrect causal model $\mathcal{M}_1$ in Fig. 1(b), where the direction of the causal relation $Y \rightarrow X_2$ is reversed (highlighted in red). Note that X_2 is causally influenced by S in the true causal model $\mathcal{M}$. If a predictor is based on $\mathcal{M}_1$, X_2 would be directly used in prediction, and thus it violates counterfactual fairness with biases from the sensitive attribute. 2) Consider another incorrect causal model M_2, where an existing causal relation $S \rightarrow X_1$ in the true causal model $\mathcal{M}$ is ignored. Predictors based on M_2 would directly use X_1 in prediction, which results in biases. Unfortunately, causal models are quite common to be incorrectly assumed or discovered [26, 31, 38, 46].

To address the aforementioned issues of insufficient human knowledge of causal model, we study a novel problem of *learning counterfactually fair predictor with unknown causal models*. Although it is in principle impossible to achieve counterfactual fairness without any causal model [30], we take initial explorations to mitigate the unfairness based on certain general assumptions, and circumvent the prerequisite of explicit prior knowledge. However, this studied problem remains a daunting task mainly due to the following challenges: 1) In order to achieve counterfactual fairness, the causal effect from the sensitive attribute S to the prediction must be removed [30, 42], but an unknown causal model brings challenges to track the influence of the sensitive attribute and eliminate the biases; 2) There might exist unobserved variables which can be used to predict the target (e.g., "geographic location" in the salary prediction example), but without a correct causal model, it is harder to capture these unobserved variables for prediction due to the lack of prior knowledge regarding these variables. 3) Many factors (e.g., failure in obtaining correct causal relations) may lead to unfair predictions, but it is difficult to exclude their influence without a correct causal model. In a nutshell, all of these challenges are essential due to the lack of counterfactual data.

To tackle these challenges, we propose a novel framework — ***C**ounterfactua**L**ly f**AI**r and invariant p**RE**dictor (CLAIRE)*, which learns counterfactually fair representations for target prediction. To remove the biases from sensitive attributes without any given causal model (challenge 1), we develop a counterfactual data augmentation module to implicitly capture the causal relations in data, and generate counterfactuals for each individual with different sensitive attribute values. In this way, CLAIRE can learn fair representations by using a counterfactual fairness constraint to minimize the difference between the predictions made on the original data and on its counterfactuals. To capture the unobserved variables which can help counterfactually fair prediction (challenge 2), CLAIRE maps the observed variables to a latent representation space to encode the unobserved variables that can facilitate the prediction. The aforementioned counterfactual fairness constraint can preserve those unobserved variables which are not biased. To further reduce the factors which potentially impede counterfactual fairness (challenge 3), we exclude the variables with *spurious correlations* to the target (i.e., variables that appear to be causal to the target but are not, e.g., X_2 in Fig. 1(a)) from the learned representations. Spurious correlations can easily lead to incorrect causal models. Besides, removing these variables can often benefit model prediction performance, as shown in [3]. We summarize our main contributions as follows:

- **Problem:** We study an important problem of learning counterfactually fair predictor from observational data. We analyze its importance, challenges, and impacts.
- **Algorithm:** We propose a novel framework CLAIRE for this problem. Specifically, we learn fair representations based on counterfactual data augmentation. Besides, we exclude spurious correlations to further reduce potential biases.
- **Experiments:** We conduct extensive experiments to evaluate our framework on synthetic and real-world datasets. The results show that CLAIRE outperforms the existing baselines.

2 PRELIMINARIES

2.1 Notations

In this paper, we use upper-cased letters, e.g., X, to denote random variables, lower-cased letters, e.g., x, to denote specific values. $P(X)$ refers to the probabilistic function of X. We use X, S, U, Y to represent the observed non-sensitive features/attributes, sensitive attribute, unobserved variables, prediction label/target for any instance, respectively. Specifically, we use X^s, Y^s to denote the corresponding features and target of any instance with the observation of a specific sensitive attribute value $S = s$, where $s \in \mathcal{S}$, and $\mathcal{S}$ is the space of the sensitive attribute value. $\widehat{Y}$ denotes the predicted label (for classification tasks) or target (for regression tasks).

2.2 Counterfactual Fairness

Counterfactual fairness [30] is an individual-level fairness notion based on the causal mechanism. It is built upon the Pearl's causal framework [39], which is defined as a triple (U, V, F) such that:

- U is the set of latent variables, which are often assumed to be exogenous and consequently independent of each other;
- V is a set of observed variables, which are endogenous and determined by variables in $U \cup V$;
- $F = \{f_1(\cdot), f_2(\cdot), ..., f_{|V|}(\cdot)\}$ is a set of functions (referred to as *structural equations*) which describe the causal relationships among the above variables. For each variable $V_i \in V$, $V_i = f_i(pa_i, U_{pa_i})$, where "$pa_i \subseteq V \setminus V_i$" and "$U_{pa_i} \subseteq U$" are variables that directly determine V_i.

A causal model is associated with a *causal graph*, which is a directed acyclic graph (DAG). Each node in the causal graph corresponds to a variable in the causal model, and each directed edge represents a causal relationship. For example, for observed variables A, B, the value of the *counterfactual* "what would A have been if B had been set to b?" is denoted by $A_{B\leftarrow b}$.

Based on a given causal model, a predictor uses a function $\widehat{Y} = f(X, S)$ to make the prediction for each instance. The predictor is *counterfactually fair* [30] if under any context $X = x$ and $S = s$,

$$P(\widehat{Y}_{S\leftarrow s} = y | X = x, S = s) = P(\widehat{Y}_{S\leftarrow s'} = y | X = x, S = s), \quad (1)$$

for all y and $s' \neq s$. Here $\widehat{Y}_{S\leftarrow s} = f(X_{S\leftarrow s}, s)$ denotes the prediction made on the counterfactuals when the value of S had been set to s.

2.3 Biases under Incorrect Causal Models

To achieve the notion of counterfactual fairness, existing works often [30, 42] follow a two-step process: 1) First, they use the observed data to fit the causal model and infer the posterior distribution $P(U|X, S)$ of unobserved variables U; 2) Second, they train a counterfactually fair predictor based on the fitted causal model. In particular, this step can be achieved in different ways: an initial work [30] trains the predictor with only unobserved variables U and the non-descendants of S as input. We refer to this method as CFP-U. Another work [42] considers a counterfactual fairness objective $|f(X_{S\leftarrow s}, s) - f(X_{S\leftarrow s'}, s')|$ for each instance, aiming to minimize the difference between the predictions made on different counterfactuals of the sensitive attribute. We refer to this method as CFP-O. In this subsection, we use some simple examples to show the biases in the prediction of these existing counterfactual fairness methods when the given causal model is incorrect.

Example 1. First, we consider the case when the counterfactual fairness methods have been given an incorrect causal model as shown in Fig. 1(b). In the aforementioned salary prediction example, the ground truth causal model $\mathcal{M}$ is shown in Fig. 1(a). It indicates that people's salary can causally influence their choices of cars to purchase. In this example, we let the causal model $\mathcal{M}$ be as follows:

$$P(S = 1) = 0.5, P(S = 0) = 0.5, \epsilon_1, \epsilon_y, \epsilon_2 \sim \mathcal{N}(0, 1),$$
$$X_1 \leftarrow S + U + \epsilon_1, Y \leftarrow X_1 + \epsilon_y, X_2 \leftarrow Y + \epsilon_2.$$

X_2 is correlated with Y because it is Y's child node, but this correlation may lead the model to incorrectly take X_2 as one of Y's parent nodes, as the incorrect causal model $\mathcal{M}_1$ shown in Fig. 1(b). Then the goal of counterfactual fairness: $P(\widehat{Y}^{\mathcal{M}_1}_{S\leftarrow s} | X = x, S = s) = P(\widehat{Y}^{\mathcal{M}_1}_{S\leftarrow s'} | X = x, S = s)$ defined on $\mathcal{M}_1$ is different from what is defined on the true causal model $\mathcal{M}$. Based on the incorrect causal model $\mathcal{M}_1$, CFP-U will take X_2 as an input to the predictor, but X_2 contains biased information because it is actually a descendant of the sensitive attribute, thus it will bring bias into prediction. For CFP-O, if we assume a linear predictor $\widehat{Y} = W_1X_1 + W_2X_2 + W_SS$, then the fairness penalty on the incorrect causal model would be:

$$\mathbb{E}(|f(X^{\mathcal{M}_1}_{S\leftarrow 1}, 1) - f(X^{\mathcal{M}_1}_{S\leftarrow 0}, 0)|) = |W_1 + W_S|,$$

while the fairness penalty based on the true causal model would be:

$$\mathbb{E}(|f(X^{\mathcal{M}}_{S\leftarrow 1}, 1) - f(X^{\mathcal{M}}_{S\leftarrow 0}, 0)|) = |W_1 + W_2 + W_S|.$$

Such difference can lead to inappropriate learning results for the parameters in the predictor. As the fairness penalty based on the incorrect causal model has no constraint on W_2, the predictor can not exclude the biases contained in X_2.

Example 2. We now consider another case of incorrect causal model shown in Fig. 1(c). In the salary prediction example, consider that the dataset contains a majority sensitive subgroup $S = 0$ (e.g., race A) and a minority sensitive subgroup $S = 1$ (e.g., race B). The ground-truth causal model is assumed to be as below:

$$P(S = 1) = 0.1, P(S = 0) = 0.9, \epsilon_1, \epsilon_y, \epsilon_2 \sim \mathcal{N}(0, 1),$$
$$X_1 \leftarrow S + U + \epsilon_1, Y \leftarrow X_1 + \epsilon_y, X_2 \leftarrow Y + \epsilon_2.$$

As the subgroup $S = 1$ is underrepresented, the fitted causal model may miss the causal relation $S \rightarrow X_1$ for $S = 1$, i.e., the fitted causal model is biased (as the causal model $\mathcal{M}_2$ shown in Fig. 1(c)). Then for CFP-U, X_1 and X_2 will be taken as input for prediction because they are considered to be non-descendants of S, but as X_1 and X_2 are actually biased because they are descendants of S, the predictor will also be biased consequently. Let us take the predictor $\widehat{Y} = X_1$ for example. The predictor makes prediction $\widehat{Y}_{S\leftarrow 0} = X_{1,S\leftarrow 0} = U + \epsilon_1$ and $\widehat{Y}_{S\leftarrow 1} = X_{1,S\leftarrow 1} = U + \epsilon_1 + 1$ in when $S \leftarrow 0$ and $S \leftarrow 1$, respectively, and this is obviously not counterfactually fair. For CFP-O, the fairness penalty on this biased causal model $\mathcal{M}_2$ is:

$$\mathbb{E}(|f(X^{\mathcal{M}_2}_{S\leftarrow 1}, 1) - f(X^{\mathcal{M}_2}_{S\leftarrow 0}, 0)|) = |W_S|,$$

while the fairness penalty based on the true causal model $\mathcal{M}$ is:

$$\mathbb{E}(|f(X^{\mathcal{M}}_{S\leftarrow 1}, 1) - f(X^{\mathcal{M}}_{S\leftarrow 0}, 0)|) = |W_1 + W_2 + W_S|.$$

Such difference may lead to inappropriate use of X_1 and X_2, and thus bring biases to the predictor.

As a summary, existing counterfactual fairness machine learning methods heavily rely on given causal models, and would result in biases when the given causal models are incorrect.

3 THE PROPOSED FRAMEWORK

In this section, we introduce the proposed framework CLAIRE, which targets at achieving counterfactual fairness without relying on explicit prior knowledge about the causal model. To achieve this goal, CLAIRE learns counterfactually fair representations with counterfactual data augmentation, and then makes predictions based on the learned representations.

3.1 Assumptions and Examples

Before technical details, we first present the key concepts and assumptions of CLAIRE, and then use general examples of causal models (Fig. 2) to describe the information needed in CLAIRE.

Previous works of counterfactual fairness [30] have discussed three levels of required prior knowledge about the causal model:

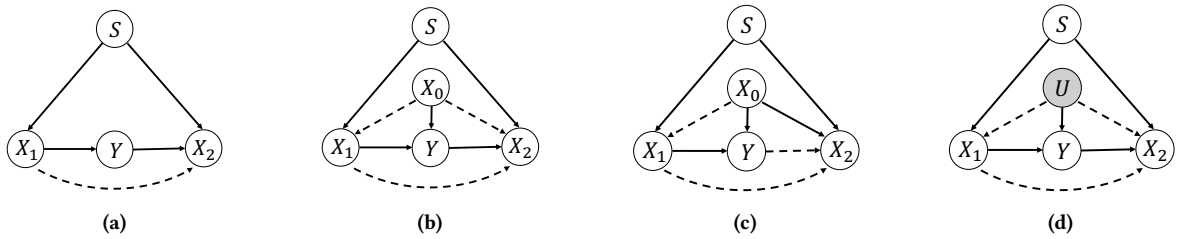


Figure 2: Case studies of different kinds of variables in causal models. Each white (gray) node denotes an observed (unobserved) variable, each arrow denotes a causal relationship, and each dashed arrow denotes a possible causal relationship. S, Y, U denotes the sensitive attribute, the prediction target, and the unobserved variable, respectively. X_1 is a causal variable of Y and is a descendent of S, X_0 is a causal variable of Y and is non-descendent of S, and X_2 is a variable with spurious correlations to Y.

1) Level 1 only requires to know which observed features are non-descendants of the sensitive attribute, and only uses them for prediction; 2) Level 2 postulates and infers the unobserved variables with partial prior knowledge of the causal model, and also uses them for prediction; 3) Level 3 makes assumptions on the causal model (e.g., additive noise model [24]), postulates the complete causal model, and then uses the inferred unobserved/observed non-descendants of the sensitive attribute for prediction. These three levels make increasingly stronger assumptions on the underlying causal model. But even the first level still requires to figure out which variables are non-descendants of the sensitive attribute. In this work, we aim to propose a principled way for counterfactually fair prediction without relying on the prior knowledge of the causal model. The main assumptions in our framework are listed as follows:

Assumption 1. *The sensitive attribute is not causally influenced by any other variables. This is a common assumption in most of existing fairness works [7, 30, 42], as the commonly-used sensitive attributes such as race and gender usually do not have any causes.*

Assumption 2. *If a variable X_c directly affects Y (i.e., an edge $X_c \rightarrow Y$ exists in the causal model), we assume $P(Y|X_c)$ is stable across different sensitive subgroups, but for the variables X_s which do not causally affect Y, $P(Y|X_s)$ may be unstable in different sensitive subgroups. This assumption and its variants are widely used in invariant learning [2, 3].*

As the ground truth causal model can be complicated, to investigate more general settings, we consider several different types of variables in the causal model, including descendant and non-descendant variables of S, causal and non-causal variables of Y, and observed and unobserved variables. Here we conduct several case studies on the causal model, and each corresponds to a causal graph shown in Fig. 2. Suppose there is a ground truth causal model $\mathcal{M}$, we call the variables in $\mathcal{M}$ which causally affect the prediction target Y (i.e., Y is the descendant of such variables) as *causal variables* of Y. In all the causal models in Fig. 2, X_1 is a causal variable of Y, but it is also a descendant of S, thus it can not be directly used for counterfactually fair prediction. As shown in Fig. 2(b) and (c), X_0 is also a causal variable of Y, and is non-descendant of S, thus X_0 is supposed to be used for fair prediction. X_2 is not a causal variable of Y, but it has statistically spurious correlations to Y. The reason may be that X_2 is Y's descendant, as shown in Fig. 2(b), or X_2 and Y are affected by some common variables, as shown in Fig. 2(c). As discussed in [3, 11], the spurious correlations between X_2 and Y often vary across different sensitive subgroups and thus degrade the model prediction performance. Besides, if these non-causal variables are also descendants of sensitive attribute, incorporating them into prediction would also impede counterfactual fairness. Therefore, in our framework, we exclude these non-causal variables to further avoid potential biases. Above cases are all about observed variables, for those unobserved variables which are causative to Y, such as U in Fig. 2(d), we try to better capture these unobserved variables by utilizing the observed variables which have correlations with them.

Overall, in our framework, we *learn representations to capture the causal variables which are not influenced by the sensitive attribute.*

3.2 Overview of CLAIRE Framework

Existing counterfactual fairness works [30, 42] involve counterfactual inference for predictor training, but it is often infeasible in real-world applications due to the lack of a correct causal model, especially when the data is noisy and high-dimensional [6]. Without enough knowledge about the causal model, inferring the unobserved variables and learning a fair predictor can be quite challenging. Here, we define the goal of our framework with respect to counterfactual fairness, and show an overview of the methodology.

Based on the aforementioned preliminaries, we know that the key point of this problem is to capture the information which elicits a fair predictor, such as the causal variables that are non-descendants of S. In our framework, we use the observed features to learn a representation $Z = \Phi(X)$ which captures the fair information, and then build a predictor $\widehat{Y} = g(Z)$ on top of it. In the implementation, we learn the representations Z in the following ways: (1) To capture the causal variables of Y, we leverage the invariant risk minimization loss [3] to exclude those non-causal variables with unstable spurious correlations to Y. (2) To avoid taking the biases from the sensitive attribute into prediction, we develop a counterfactual data augmentation module, and encourage the learned representation to achieve the following goal: for any $s \neq s'$, and any x, $P(\Phi(x_{S \leftarrow s})) = P(\Phi(x_{S \leftarrow s'}))$. Intuitively, it means that for each individual with observed features x and sensitive attribute value s, the distributions of the representations learned from its original version and its counterfactuals should be the same.

Algorithm 1 shows an overview of our framework, including counterfactual data augmentation and fair representation learning. Detailed techniques will be introduced in the following subsections.

Algorithm 1: The proposed CLAIRE framework

Data: Instances of observable variables $\{X, S, Y\}$
Result: Counterfactually fair predictor $\widehat{Y} = f(X, S)$
/* 1. Counterfactual Data Augmentation */
Train a VAE with encoder $\Psi(\cdot)$ and decoder $D(\cdot)$ with loss function in Eq. (3) (CLAIRE-M) or Eq. (4) (CLAIRE-A)
for *each instance of random variables* $\{X, S, Y\}$ **do**
  Generate K samples $H^1, ..., H^K$ with $H = \Psi(X, Y)$
  for $s \in \mathcal{S}$ **do**
    $X_s^{CF}, Y_s^{CF} = \textsc{Aggregate}(D(H^1, s), ..., D(H^K, s))$
  end
end
/* 2. Fair representation learning */
Train a model $f = g \circ \Phi$ consisting of a representation learner $\Phi(\cdot)$ and a predictor $g(\cdot)$
for *each instance of random variables* $\{X, S, Y\}$ **do**
  $Z = \Phi(X)$, $\widehat{Y} = g(Z)$
  for $s \in \mathcal{S}$ **do**
    $Z_s^{CF} = \Phi(X_s^{CF})$, $\widehat{Y}_s^{CF} = g(Z_s^{CF})$
  end
  Back-propagation with loss function in Eq. (7)
end

3.3 Counterfactual Data Augmentation

The lack of counterfactual data is the essential challenge to achieve counterfactual fairness. Thus, we pretrain a counterfactual data augmentation module to generate counterfactuals for each instance by manipulating its sensitive attribute. Then, the augmented counterfactuals together with original data are utilized to learn fair representations. The counterfactual data augmentation module is based on a variational auto-encoder (VAE) [28] with an encoder-decoder structure. Specifically, the encoder in the VAE takes $\{X, Y\}$ as input, encodes them into a latent embedding space, and then the decoder reconstructs the original data $\{X, Y\}$ with the embeddings H (notice that the embedding H is different from the representation Z introduced in the previous subsection. H is the output of the bottleneck layer of the VAE in counterfactual data augmentation to generate counterfactuals) and sensitive attribute S. Note that S is only used as an input of the decoder to enable counterfactual generation in later steps. The reconstruction loss $\mathcal{L}_r$ is:

$$\mathcal{L}_r = \mathbb{E}_{q(H|X,Y)}[-\log(p(X, Y|H, S))] + \mathrm{KL}[q(H|X, Y)\|p(H)], \quad (2)$$

where $p(H)$ is a prior distribution, e.g., standard normal distribution $\mathcal{N}(0, I)$. $\mathrm{KL}[\cdot\|\cdot]$ is the Kullback-Leibler (KL) divergence.

To generate counterfactuals with the embeddings H and a manipulated sensitive attribute value later, we need to capture more "fair" generative factors (i.e., those generative factors which are not causal influenced by S) in the embeddings, i.e., in encoder, we remove the causal influence of the sensitive attribute on the embedding H. Based on Assumption 1, if there is no dependency between the embeddings and sensitive attribute, then the embeddings encode no descendants of sensitive attributes. Now, we introduce two different implementations to remove the causal effect of S on H by minimizing the dependency between them. These implementations include the distribution matching based CLAIRE (CLAIRE-M) and the adversarial learning based CLAIRE (CLAIRE-A).

Distribution matching based CLAIRE. To remove the influence of the sensitive attribute, we use the distribution matching technique [33, 45] on the embeddings for different sensitive subgroups. We refer this implementation as CLAIRE-M. In particular, we minimize the Maximum Mean Discrepancy (MMD) [33, 45] among the embedding distributions of different sensitive subgroups.

The loss function of training the counterfactual data augmentation model with distribution matching is as below:

$$\min \mathcal{L}_r + \alpha \frac{1}{N_p} \sum\nolimits_{s \neq s'} MMD(P(H|s), P(H|s')), \quad (3)$$

where $N_p = \frac{|\mathcal{S}| \times (|\mathcal{S}|-1)}{2}$ is the number of pairs of different sensitive attribute values, and $|\mathcal{S}|$ is the number of different sensitive attribute values. The second term is the distribution matching penalty, which aims to achieve $P(H|S = s) = P(H|S = s')$ for all pairs of different sensitive subgroups (s, s'). Here $\alpha \geq 0$ is a hyperparameter which controls the importance of the distribution balancing term.

Adversarial Learning based CLAIRE. We also propose an adversarial learning based implementation, referred as CLAIRE-A. In this implementation, we train a discriminator $h(\cdot)$ which uses the embeddings to distinguish instances that bear different values of the sensitive attribute. The objective function is as below:

$$\min_{\Psi(\cdot)} \max_{h(\cdot)} \mathcal{L}_r + \alpha' \frac{1}{|\mathcal{S}|} \sum\nolimits_{s \in \mathcal{S}} \mathbb{E}_{X^s, S^s}[\log P(h(H) = s)], \quad (4)$$

where $\Psi(\cdot)$ is the encoder. The first term is the aforementioned reconstruction loss. The second term calculates the probability that the discriminator makes correct predictions for each instance's sensitive attribute. Therefore, the sensitive attribute predictor $h(\cdot)$ is playing an adversarial game with the encoder $\Psi(\cdot)$. In this way, the embeddings are encouraged to exclude the information related to the sensitive attribute. Here $\alpha' \geq 0$ is a hyperparameter to control the weight of the sensitive attribute discriminator. The minimax problem is optimized with an alternating gradient descent process.

3.4 Fair Representation Learning

3.4.1 Counterfactually Fair Representations. With the counterfactual data augmentation module, we generate counterfactuals by feeding the embeddings H and a sensitive attribute value s' different from the original one s into the decoder $D(\cdot)$, and taking the output $(X_{s'}^{CF}, Y_{s'}^{CF}) = D(H, s')$ as the counterfactuals corresponding to $S \leftarrow s'$. For each instance and each sensitive attribute value, we generate K samples of embeddings $(H^1, ..., H^K)$, and aggregate the corresponding counterfactuals by an operation $\textsc{Aggregate}(\cdot)$ (e.g., mean). For notation simplicity, we still denote the aggregated counterfactual data as $(X_{s'}^{CF}, Y_{s'}^{CF}) = \textsc{Aggregate}(D(H^1, s'), ..., D(H^K, s'))$. Based on these counterfactuals, we train a representation learner $\Phi(\cdot)$ which maps instance features X into representations: $Z = \Phi(X)$, and we use a predictor $g(\cdot)$ to make predictions based on Z.

To learn counterfactually fair representations Z, we add a counterfactual fairness constraint to mitigate the discrepancy between the representations learned from original data and its corresponding counterfactuals. The constraint is formulated as:

$$\mathcal{L}_c = \frac{1}{|\mathcal{S}| - 1} \sum_{s' \neq s} d(Z, Z_{s'}^{CF}) = \frac{1}{|\mathcal{S}| - 1} \sum_{s' \neq s} d(\Phi(X), \Phi(X_{s'}^{CF})), \quad (5)$$

where $X_{s'}^{CF}$ is the counterfactual generated in counterfactual data augmentation corresponding to $S \leftarrow s'$, and $d(\cdot,\cdot)$ is a distance metric such as cosine distance to measure the discrepancy between two representations.

3.4.2 Invariant Representations. As aforementioned, the non-causal variables which have spurious correlations to the target Y are likely to degrade the model prediction performance, and may also incorporate potential biases from sensitive attributes to prediction. It has been shown in [3] that the relationships from these variables to Y often vary across different domains, e.g., different sensitive subgroups. Therefore, to exclude the influence of such non-causal variables on the learned representations and capture the causal variables of Y, we leverage the invariant risk minimization (IRM) loss [3] for the sensitive subgroup s as below:

$$\mathcal{L}_{IRM}^{s} = R^s(g \circ \Phi) + \lambda \left\| \nabla_{w|w=1.0} R^s(w \cdot (g \circ \Phi)) \right\|_2^2, \tag{6}$$

where $\mathcal{L}_{IRM}^{s}$ is the IRM loss in the sensitive subgroup s, the first term $R^s(g \circ \Phi) = \mathbb{E}[\mathcal{L}(g(\Phi(X^s, S^s)), Y^s)]$ is the prediction loss under sensitive subgroup s, and w is a scalar and is fixed as $w = 1.0$. According to [3], the gradient of $R^s(w \cdot (g \circ \Phi))$ w.r.t. w can reflect the "invariance" of the learned representations. Therefore, in the above formulation, the second term measures the invariance of the relationship between the representations and the target across different sensitive groups. Here, λ is a hyperparameter for the trade-off between the prediction performance and the level of invariance. The IRM loss aims to ensure that the predictor can be optimal in all the different sensitive subgroups, thus the unstable spurious correlations varying across sensitive subgroups can be excluded.

To put it all together, the overall loss function for fair representation learning is as follows:

$$\mathcal{L} = \frac{1}{|\mathcal{S}|} \sum\nolimits_{s \in \mathcal{S}} \mathcal{L}_{IRM}^{s} + \beta \mathcal{L}_c, \tag{7}$$

where β is the weight of the counterfactual fairness constraint. More implementation details can be found in Appendix A.

4 EXPERIMENTAL EVALUATIONS

In this section, we conduct extensive experiments to evaluate the proposed framework CLAIRE on two real-world datasets and one synthetic dataset. Before showing the detailed results, we first present the details of used datasets and the experimental settings.

4.1 Datasets

Law School. This dataset contains academic information of students in 163 law schools. Our goal is to predict each student's first year average grade (FYA), and this is a regression task. We take *race* as their sensitive attribute, and take grade-point average (GPA) and entrance exam scores (LSAT) as two observed features. Here, we select persons in races of white, black, and asian. The dataset contains 20, 412 instances. We use the level-2 causal model in [30] as the true causal model with causal graph shown in Fig. 3(a).

Adult. UCI Adult income dataset[1] contains census data for different adults and the target here is to predict whether their income exceeds 50K/yr. We take *race* as the sensitive attribute S, and their *income* as the prediction label Y. This is a binary classification task. We select persons in the races of white, black, and Asian-Pac-Islander. In addition to the sensitive attribute of *race*, we use other 5 attributes for prediction. The dataset contains 31, 979 instances. Here, we follow [52] and consider the causal model used by them as the ground truth. The causal graph is shown in Fig. 3(b).

Synthetic Dataset. Here, we use a ground truth causal model to generate the synthetic data. The true causal graph is shown in Fig. 4(a), containing a sensitive attribute S with four different categorical values $\{0, 1, 2, 3\}$, an unobserved variable U, a causal variable X_0 which is non-descendant of S, a causal variable X_1 which is descendant of S, and a variable X_2 which is the descendant of Y. The structural equations are as follows:

$$\begin{aligned} S \sim \text{Catgorical}(\pi), U \sim \mathcal{N}(0, \sigma_U^2), X_0 = \mathcal{N}(0, \sigma_0^2), \\ X_1 = W_S S + U + \mathcal{N}(0, \sigma_{S,1}^2), Y = X_1 + X_0 + \mathcal{N}(0, \sigma_{S,Y}^2), \\ X_2 = Y + \mathcal{N}(0, \sigma_{S,2}^2), \end{aligned} \tag{8}$$

where $\pi = \{0.5, 0.4, 0.05, 0.05\}$, $\sigma_U = \sigma_0 = 1$, $\sigma_{S,*}$ and W_S are set as $\{0.5, 1.0, 1.5, 2.0\}$ and $\{0.1, 0.2, 1.0, 2.0\}$ respectively for four values of sensitive attribute. In this dataset, the spurious correlation $X_2 \rightarrow Y$ and the imbalanced distribution of sensitive subgroups may lead to incorrect causal models, as shown in [37]. We will further investigate the impact of these two situations in Section 4.4.

4.2 Experimental Settings

Baselines. To investigate the effectiveness of our framework in learning counterfactually fair predictors from observational data, we compare the proposed framework with multiple state-of-the-art methods. First, we briefly introduce all the compared baseline methods and their settings:

- **Constant Predictor:** A predictor which has constant output for any input. We obtain this constant predictor by finding a constant which can minimize the mean squared error (MSE) loss on the training data.
- **Full Predictor:** Full predictor takes *all* the observed attributes (except the attribute used as label) as input for prediction.
- **Unaware Predictor:** Unaware predictor is based on the notion of fairness through unawareness [20]. It takes all features except the sensitive attribute as input to predict the label.
- **Counterfactual Fairness Predictor:** We use two different counterfactual fairness predictors here, including **CFP-U** [30] and **CFP-O** [42]. These methods require a given causal model.

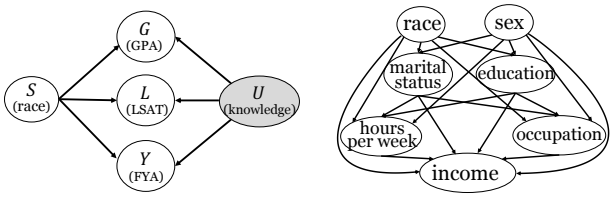
For baselines full/unaware/counterfactual fair predictors, we use linear regression for regression and logistic regression for classification. More details of baselines can be found in Appendix B.

Evaluation Metrics. Generally, the evaluation metrics consider two different aspects: prediction performance and counterfactual fairness. To measure the model prediction performance, we employ the commonly used metrics – Root Mean Square Error (RMSE) and mean absolute error (MAE) for regression tasks and accuracy for classification tasks. To evaluate different methods with respect to counterfactual fairness, we compare the distribution divergence of the predictions made on different counterfactuals generated by

[1] https://archive.ics.uci.edu/ml/datasets/adult

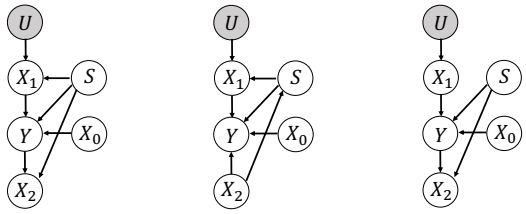
Table 1: Results comparison of different predictors on two real-world datasets. Our method CLAIRE can achieve the best performance in counterfactual fairness with competitive prediction performance.

Method	Law school				Adult		
	RMSE (↓)	MAE (↓)	MMD (↓)	Wass(↓)	Accuracy (↑)	MMD (↓)	Wass (↓)
Constant	0.952 ± 0.003	0.772 ± 0.002	**0.000 ± 0.000**	**0.000 ± 0.000**	0.745 ± 0.001	**0.000 ± 0.000**	**0.000 ± 0.000**
Full	**0.896 ± 0.004**	**0.723 ± 0.003**	259.744 ± 5.213	65.656 ± 1.326	**0.815 ± 0.002**	50.513 ± 3.283	5.217 ± 0.582
Unaware	0.909 ± 0.002	0.734 ± 0.004	39.144 ± 3.248	10.093 ± 1.254	0.809 ± 0.003	16.832 ± 2.377	1.983 ± 0.462
CFP-U (true)	0.932 ± 0.003	0.738 ± 0.002	4.307 ± 0.003	0.019 ± 0.001	0.745 ± 0.002	3.582 ± 0.007	0.025 ± 0.002
CFP-O (true)	0.929 ± 0.004	0.735 ± 0.003	4.325 ± 0.002	0.020 ± 0.012	0.748 ± 0.003	3.623 ± 0.004	0.029 ± 0.004
CLAIRE-M (ours)	0.909 ± 0.002	0.733 ± 0.003	4.297 ± 0.002	0.019 ± 0.001	0.778 ± 0.002	3.552 ± 0.021	0.023 ± 0.002
CLAIRE-A (ours)	0.910 ± 0.002	0.734 ± 0.002	4.289 ± 0.002	0.018 ± 0.001	0.780 ± 0.003	3.547 ± 0.007	0.023 ± 0.002



(a) Law school **(b) Adult**

Figure 3: The ground truth causal models of two real-world datasets Law School and Adult.



(a) $\mathcal{M}$ **(b) $\mathcal{M}_1$** **(c) $\mathcal{M}_2$**

Figure 4: The true causal model ($\mathcal{M}$) and two incorrect causal models ($\mathcal{M}_1$ and $\mathcal{M}_2$) of the synthetic dataset.

Table 2: Study on synthetic data about the adverse effects of incorrect causal model $\mathcal{M}_1$.

Method	RMSE	MAE	MMD	Wass
CFP-U (true)	1.34 ± 0.01	0.88 ± 0.01	8.42 ± 0.70	3.07 ± 0.01
CFP-U (false)	1.30 ± 0.01	0.83 ± 0.02	10.11 ± 0.52	3.79 ± 0.03
CFP-O (true)	1.32 ± 0.01	0.87 ± 0.01	8.48 ± 0.83	3.32 ± 0.02
CFP-O (false)	**1.29 ± 0.01**	**0.81 ± 0.01**	10.94 ± 0.61	3.84 ± 0.02
CLAIRE-M	1.32 ± 0.01	0.87 ± 0.02	7.52 ± 0.08	2.63 ± 0.02
CLAIRE-A	1.31 ± 0.01	0.85 ± 0.03	**7.49 ± 0.05**	**2.58 ± 0.01**

the ground truth causal model. If a predictor is counterfactually fair, the distributions of the predictions under different ground-truth counterfactuals are expected to be the same. Here, we use two distribution distance metrics (including Wasserstein-1 distance (Wass) [41] and Maximum Mean Discrepancy (MMD) [33, 45]) to measure the distribution divergence. We compute the divergence of prediction distributions in every pair of counterfactuals ($S \leftarrow s$ and $S \leftarrow s'$ for any $s \neq s'$), then take the average value as the final result. The smaller the average values of MMD and Wass are, the better a predictor performs in counterfactual fairness. For the synthetic data, the ground truth causal model is known, while for the real-world datasets, we adopt the widely accepted causal models as mentioned in Section 4.1.

Table 3: Study on synthetic data regarding the adverse effects of incorrect causal model $\mathcal{M}_2$.

Method	$S \leftarrow 0$ and $S \leftarrow 1$		$S \leftarrow 0$ and $S \leftarrow 2$	
	MMD	WASS	MMD	Wass
CFP-U (true)	6.05 ± 0.02	1.10 ± 0.02	7.97 ± 0.03	2.55 ± 0.02
CFP-U (false)	6.63 ± 0.09	1.24 ± 0.04	9.33 ± 1.00	3.62 ± 0.01
CFP-O (true)	6.34 ± 0.07	1.13 ± 0.03	8.31 ± 0.98	2.84 ± 0.03
CFP-O (false)	6.83 ± 0.08	1.35 ± 0.05	9.92 ± 1.01	3.98 ± 0.02
CLAIRE-M	6.12 ± 0.04	1.13 ± 0.02	7.94 ± 0.06	2.52 ± 0.01
CLAIRE-A	**6.05 ± 0.03**	**1.11 ± 0.03**	**7.42 ± 0.04**	**2.49 ± 0.01**

Hyperparameter Settings. For all these three datasets, we split the training/validation/test set as 60%/20%/20%. All the presented results are on the test data. We set the number of training epochs as 500, the representation dimension as 10, $\alpha = 2.0$, $\alpha' = 1.0$, $K = 20$, $\beta = 5.0$, and $\lambda = 1.0$.

4.3 Experimental Results on Real-world Data

To assess the superiority of the proposed framework CLAIRE, we compare its two implementations CLAIRE-M and CLAIRE-A against other predictors on two real-world datasets Law School and Adult. We show the ground truth causal models of these two datasets in Fig. 3 although our proposed framework and its variants do not rely on the causal model. Table 1 presents the performance of different methods regarding prediction and counterfactual fairness. The best results are shown in **bold**, and the runner-up results are underlined. Generally speaking, existing methods which are not designed for counterfactual fairness have higher MMD and Wass, although they can use the biased features to achieve better prediction performance. We make the following observations from Table 1:

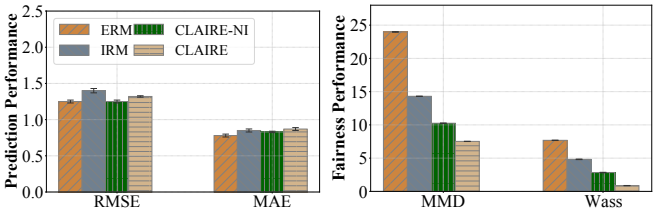
- Among all the compared methods, the constant predictor has the worst performance in prediction as it lacks capability to distinguish different instances. However, it always satisfies counterfactual fairness because it has constant output.
- The full predictor performs well in prediction, as it utilizes all the features (both sensitive and non-sensitive). But the use of sensitive attribute also brings biases to the prediction, as demonstrated by its high values on fairness metrics.
- The unaware predictor removes certain biases by ignoring the sensitive attribute, but it cannot exclude the implicit biases caused by inappropriate usage of the descendants of the sensitive attribute.
- Both CFP-U and CFP-O infer the latent variables based on the given causal model, so they perform well if the given causal model is correct.
- Our proposed CLAIRE consistently outperform other baselines (except the constant predictor) under different fairness metrics, and also have better prediction performance than many other fairness-aware baselines (including CFP-U and CFP-O). It implies that CLAIRE can achieve a good balance between prediction performance and counterfactual fairness.
- The variants CLAIRE-M and CLAIRE-A generally have similar performance, but CLAIRE-A is slightly better in fairness, it may benefit from the effectiveness of its adversarial learning mechanism in removing the sensitive information.

4.4 Experimental Results on Synthetic Data

The above experiments on real-world datasets have demonstrated the superiority of CLAIRE. Here, we perform further studies on the synthetic dataset to show the impact of incorrect causal models.

Incorrect causal model $\mathbf{M_1}$. In this experiment, we use the synthetic data to showcase the impact of an incorrect causal model as the example shown in Fig. 4(b). The true causal model of the synthetic data is shown in Fig. 4(a). Here, causal relations regarding X_2 in $\mathcal{M}_1$ are reversed. As all the baselines (except CFP-U and CFP-O) do not rely on the causal model for prediction, so their results are not influenced by the correctness of the causal model. Here, we investigate the influence of the incorrect causal model on CFP-U and CFP-O and compare their performance with our proposed framework. From the results shown in Table 2, we find the fairness of CFP-U and CFP-O are obviously affected by the incorrect causal model. Although CFP-U and CFP-O with incorrect causal model have slightly better performance in prediction, that is because based on the incorrect causal model, they may take X_2 into prediction, which however, brings biases for prediction. Our proposed framework does not assume the existence of any given causal model for prediction. The counterfactual data augmentation enables us to eliminate the influence of sensitive attributes to the prediction. Furthermore, the learned invariant representations in CLAIRE exclude the adverse impacts of non-causal variables with spurious correlations and leverage the causal variables to learn representations, thus X_2 is encouraged to be excluded from prediction.

Incorrect causal model $\mathbf{M_2}$. Now, we use the synthetic data to showcase the impact of another incorrect causal model as shown in Fig. 4(c). As described in Section 4.1, we set the parameter W_S in Eq. (8), which determines the relation $S \rightarrow X_1$, to be small on the majority sensitive subgroups ($S = 0, 1$) but relatively large on the minority sensitive subgroups ($S = 2, 3$). Here, the incorrect causal model misses the causal relation $S \rightarrow X_1$ (as shown in Fig. 4(c)). We compare the prediction differences between pairs of different counterfactuals generated by the true causal model shown in Fig. 4(a). The results are shown in Table 3, where we select two pairs of counterfactuals: ($S \leftarrow 0$ and $S \leftarrow 1$) and ($S \leftarrow 0$ and $S \leftarrow 2$). As W_S is small when $S = 0$ and $S = 1$, the biased causal model would not bring too much bias from the sensitive attribute to the prediction in the two counterfactuals ($S \leftarrow 0$ and $S \leftarrow 1$), so the discrepancy between this pair is relatively lower than the other pair. But for the counterfactuals of $S \leftarrow 2$ (and also $S \leftarrow 3$), CFP-U and CFP-O suffer more from the biased causal model. As observed in Table 3, when CFP-U and CFP-O are under the biased causal model, the prediction discrepancy between the pair of counterfactuals ($S \leftarrow 0$ and $S \leftarrow 2$) becomes larger than the case when CFP-U and CFP-O are under the true causal model. Similar observations can also be found in the pair ($S \leftarrow 2$ and $S \leftarrow 3$), as shown in Appendix C. Our framework outperforms the baselines due to the following key factors: the fair generative factors captured in counterfactual data augmentation remove the influence of the observed sensitive attribute to the generated counterfactuals. Therefore, the counterfactual fairness constraint mitigates the influence of sensitive attribute on the learned representations, and makes our framework suffer less from imbalanced sensitive subgroups.



(a) Prediction **(b) Fairness**

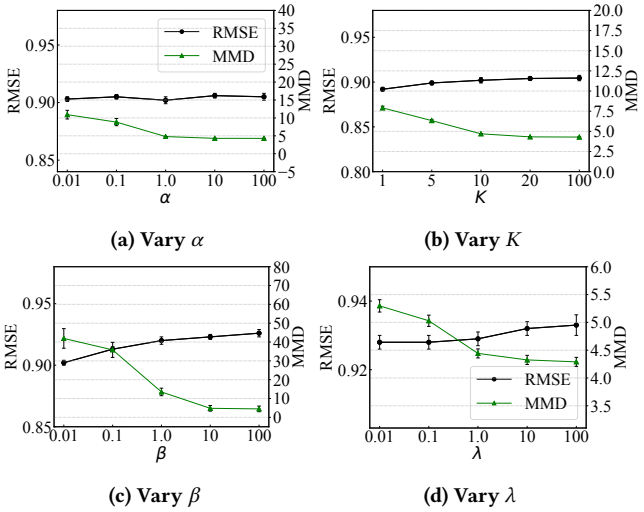
Figure 5: Ablation Study on Synthetic Dataset.

4.5 Ablation Study

To evaluate the effectiveness of each component in our method, we provide ablation study with the following variants: 1) **Empirical Risk Minimization (ERM):** ERM can be considered as a variant of our proposed framework CLAIRE. Here, we only use the empirical risk minimization loss (the first term of Eq. (6)) in prediction without the counterfactual fairness constraint and invariant penalty by setting $\beta = 0$ and $\lambda = 0$. 2) **Invariant Risk Minimization (IRM) [3]:** Here, we remove the counterfactual fairness constraint in our framework by setting $\beta = 0$. 3) **CLAIRE-NI:** As the third variant of our proposed framework, we remove the invariant penalty by setting $\lambda = 0$ in CLAIRE. From the results shown in Fig. 5, the counterfactual data augmentation and invariant penalty both contribute to the overall fairness performance.

4.6 Parameter Study

We set the hyperparameter $\alpha \in \{0.01, 0.1, 1.0, 10, 100\}$, the sampling number $K \in \{1, 5, 10, 20, 100\}$, $\beta \in \{0.01, 0.1, 1.0, 10, 100\}$,



(a) Vary α (b) Vary K

(c) Vary β (d) Vary λ

Figure 6: Performance of CLAIRE with different settings of hyperparameters.

$\lambda \in \{0.01, 0.1, 1.0, 10, 100\}$, and compare the performance of our proposed framework in Fig. 6. Here we only show the results of CLAIRE-M on the law school dataset, as similar patterns can be observed in CLAIRE-A and other datasets. As observed in Fig. 6(a), α controls the "fairness" of the embedding in counterfactual data augmentation. Larger values of α can improve the counterfactual fairness of the framework, and have no obvious impact on the prediction performance. With larger K in Fig. 6(b), the performance of counterfactual fairness also improves because more samples are generated in counterfactual data augmentation. β controls the importance of counterfactual fairness constraint, λ controls the invariance penalty of the representations. As shown in Fig. 6(c), with the increase of β, the framework focuses more on removing the biases from the sensitive attribute, which may sacrifice some information to predict the target, and thus results in higher RMSE, but can achieve better fairness. As shown in Fig. 6(d), with the increase of λ, the framework may exclude more variables with unstable relationships to the target across different sensitive subgroups, it may thus lose some information specific to each sensitive subgroup, but can also contribute to better fairness. From the observations, the framework achieves a good trade-off on the prediction performance and counterfactual fairness with proper parameter settings.

5 RELATED WORK

Counterfactual Fairness. Recently, aside from traditional statistical fairness notions [4, 12, 13, 16, 22, 54, 55], causal-based fairness notions [30, 35, 42] have attracted a surge of attentions because of its strong capability of modeling how the discrimination is exhibited. Among them, the notion of counterfactual fairness [30] assesses fairness at the individual level. Most of the existing counterfactual fairness studies [18, 30, 53] are based on a given ground-truth causal model or rely on causal discovery methods [26, 38, 46]. Multi-world fairness [42] considers the situation when the ground-truth causal model cannot be decided, but it still requires a candidate set containing causal models which may be true, and proposes an optimization based method to achieve counterfactual fairness with the average of the causal models in the candidate set. Many methods based on traditional causal discovery are limited in certain scenarios, such as low-dimensional and linear settings. Recent studies [19, 27, 56] provide more discussion about counterfactual fairness under different assumptions and scenarios. But in conclusion, most of the above methods require much explicit prior knowledge of the causal model to remove the influence of the sensitive attribute on the prediction, and lack discussion of the impact of incorrect causal models.

Invariant Risk Minimization. Invariant risk minimization (IRM) [3] and its variants [2, 11, 21, 25, 29, 34] are originally proposed for out-of-distribution (OOD) generalization [29, 43]. It is based on the theorem that the representations of causal features elicit the existence of an optimal predictor across different domains. From a causal perspective, IRM identifies these causal features and excludes those features with spurious correlations as these correlations are not robust across different domains. The connections between fairness and IRM are discussed in [3, 15, 48]. IRM can learn representations to capture causal features which have invariant relationships to the prediction target. However, the representations may still contain the information of domains (e.g., different sensitive attributes), which may cause biases to prediction. Our work investigate to bridge this gap between IRM and counterfactual fairness.

6 CONCLUSION

In this work, we study a novel problem of learning counterfactually fair predictors from observational data with unknown causal models. We propose a principled framework CLAIRE. More specifically, we specify this framework by learning counterfactually fair representations for each instance, and make predictions based on the representations. To learn fair representations, a variational auto-encoder based counterfactual data augmentation module is developed to generate counterfactual data with different values of sensitive attribute for each instance. We further reduce potential biases by applying the invariant penalty in each sensitive subgroup to exclude the variables with spurious correlations to the target. We evaluate the proposed framework under both real-world benchmark datasets and synthetic data. Extensive experimental results validate the superiority of the proposed framework over existing fairness predictors in different aspects. Overall, this paper provides insights for promoting counterfactual fairness in a more realistic scenario without given correct causal models, and also shows the impact of incorrect causal models. In the future, more research work on counterfactual fairness in real-world cases, such as missing and noisy data, is worth further exploration.

ACKNOWLEDGEMENTS

Jing Ma, Aidong Zhang, and Jundong Li are supported by the National Science Foundation under grants (IIS-1955151, IIS-2006844, IIS-2008208, IIS-2106913, IIS-2144209, IIS-2223769, CNS-2154962, CNS-2213700, BCS-2228534, and CCF-2217071), the Commonwealth Cyber Initiative awards (VV-1Q23-007 and HV-2Q23-003), the JP Morgan Chase Faculty Research Award, the Cisco Faculty Research Award, the Jefferson Lab subcontract 23-D0163, the UVA 3 Cavaliers seed grant, and the 4-VA collaborative research grant.

REFERENCES

[1] Carlos Aguirre, Keith Harrigian, and Mark Dredze. 2021. Gender and racial fairness in depression research using social media. *arXiv preprint* (2021).

[2] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. 2020. Invariant risk minimization games. In *ICML*. 145–155.

[3] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint* (2019).

[4] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2017. Fairness in machine learning. *NeurIPS tutorial* 1 (2017), 2017.

[5] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4–1.

[6] Alexandre Belloni, Victor Chernozhukov, Iván Fernández-Val, and Christian Hansen. 2017. Program evaluation and causal inference with high-dimensional data. *Econometrica* 85, 1 (2017), 233–298.

[7] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. Fairness in criminal justice risk assessments: the state of the art. *Sociological Methods & Research* (2018), 0049124118782533.

[8] Eli Bingham, Jonathan P Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D Goodman. 2019. Pyro: Deep universal probabilistic programming. *JMLR* 20, 1 (2019), 973–978.

[9] Sarah Bird, Krishnaram Kenthapadi, Emre Kiciman, and Margaret Mitchell. 2019. Fairness-aware machine learning: Practical challenges and lessons learned. In *WSDM*. 834–835.

[10] Tim Brennan, William Dieterich, and Beate Ehret. 2009. Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior* 36, 1 (2009), 21–40.

[11] Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. 2020. Invariant rationalization. In *ICML*. 1448–1458.

[12] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[13] Alexandra Chouldechova and Aaron Roth. 2018. The frontiers of fairness in machine learning. *arXiv preprint* (2018).

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

[15] Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. 2020. Environment inference for invariant learning. In *ICML Workshop on Uncertainty and Robustness*.

[16] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. COMPAS risk scales: demonstrating accuracy equity and predictive parity. *Northpoint Inc* 7, 7.4 (2016), 1.

[17] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *SIGKDD*. 2221–2231.

[18] Vincent Grari, Sylvain Lamprier, and Marcin Detyniecki. 2020. Adversarial learning for counterfactual fairness. *arXiv preprint* (2020).

[19] Vincent Grari, Sylvain Lamprier, and Marcin Detyniecki. 2022. Adversarial learning for counterfactual fairness. *Machine Learning* (2022), 1–23.

[20] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2016. The case for process fairness in learning: feature selection for fair decision making. In *NeurIPS Symposium on Machine Learning and the Law*.

[21] Ruocheng Guo, Pengchuan Zhang, Hao Liu, and Emre Kiciman. 2021. Out-of-distribution Prediction with Invariant Risk Minimization: The Limitation and An Effective Fix. *arXiv preprint* (2021).

[22] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *NeurIPS*. 3315–3323.

[23] David Heckerman, Christopher Meek, and Gregory Cooper. 1999. A Bayesian approach to causal discovery. *Computation, causation, and discovery* 19 (1999), 141–166.

[24] Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, Bernhard Schölkopf, et al. 2008. Nonlinear causal discovery with additive noise models. In *NeurIPS*, Vol. 21. Citeseer, 689–696.

[25] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. 2020. Domain extrapolation via regret minimization. *arXiv preprint* (2020).

[26] Markus Kalisch and Peter Bühlmann. 2007. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *JMLR* 8, Mar (2007), 613–636.

[27] Hyemi Kim, Seungjae Shin, JoonHo Jang, Kyungwoo Song, Weonyoung Joo, Wanmo Kang, and Il-Chul Moon. 2021. Counterfactual fairness with disentangled causal effect variational autoencoder. In *AAAI*, Vol. 35. 8128–8136.

[28] Diederik P Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. *stat* 1050 (2014), 1.

[29] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Remi Le Priol, and Aaron Courville. 2020. Out-of-distribution generalization via risk extrapolation (rex). *arXiv preprint* (2020).

[30] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *NeurIPS*. 4066–4076.

[31] Thuc Le, Tao Hoang, Jiuyong Li, Lin Liu, Huawen Liu, and Shu Hu. 2016. A fast PC algorithm for high dimensional causal discovery with multi-core PCs. *IEEE/ACM TCBB* (2016).

[32] Sabina Leonelli, Rebecca Lovell, Benedict W Wheeler, Lora Fleming, and Hywel Williams. 2021. From FAIR data to fair data use: Methodological data fairness in health-related social media research. *Big Data & Society* 8, 1 (2021).

[33] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. 2015. Learning transferable features with deep adaptation networks. In *ICML*. 97–105.

[34] Divyat Mahajan, Shruti Tople, and Amit Sharma. 2020. Domain generalization using causal matching. *arXiv preprint* (2020).

[35] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2020. Survey on Causal-based Machine Learning Fairness Notions. *arXiv preprint* (2020).

[36] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *CIKM*. 2243–2251.

[37] Meike Nauta, Doina Bucur, and Christin Seifert. 2019. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction* 1, 1 (2019), 312–340.

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

[39] Judea Pearl et al. 2009. Causal inference in statistics: an overview. *Statistics surveys* 3 (2009), 96–146.

[40] Evaggelia Pitoura, Georgia Koutrika, and Konstantinos Stefanidis. 2020. Fairness in rankings and recommenders. (2020).

[41] Ludger Rüschendorf. 1985. The Wasserstein distance and approximation theorems. *Probability Theory and Related Fields* 70, 1 (1985).

[42] Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. 2017. When worlds collide: integrating different counterfactual assumptions in fairness. In *NeurIPS*. 6414–6423.

[43] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2019. Distributionally robust neural networks for group shifts: on the importance of regularization for worst-case generalization. *arXiv preprint* (2019).

[44] Steven Schwartz et al. 2004. Fair admissions to higher education: recommendations for good practice. *London: Higher Education Steering Group* (2004).

[45] Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *ICML*. 3076–3085.

[46] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. 2000. *Causation, prediction, and search*. MIT press.

[47] Peter Spirtes and Kun Zhang. 2016. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, Vol. 3. SpringerOpen, 1–28.

[48] Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. 2021. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. *arXiv preprint* (2021).

[49] Christina Wadsworth, Francesca Vera, and Chris Piech. 2018. Achieving fairness through adversarial learning: an application to recidivism prediction. *arXiv preprint* (2018).

[50] Y Samuel Wang and Mathias Drton. 2020. High-dimensional causal discovery under non-Gaussianity. *Biometrika* 107, 1 (2020), 41–59.

[51] Le Wu, Lei Chen, Pengyang Shao, Richang Hong, Xiting Wang, and Meng Wang. 2021. Learning Fair Representations for Recommendation: A Graph-based Perspective. In *WWW*. 2198–2208.

[52] Yongkai Wu, Lu Zhang, and Xintao Wu. 2019. Counterfactual fairness: unidentification, bound and algorithm. In *IJCAI*. 1438–1444.

[53] Depeng Xu, Yongkai Wu, Shuhan Yuan, Lu Zhang, and Xintao Wu. 2019. Achieving causal fairness through generative adversarial networks. In *IJCAI*. 1452–1458.

[54] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *WWW*. 1171–1180.

[55] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *ICML*. 325–333.

[56] Aoqi Zuo, Susan Wei, Tongliang Liu, Bo Han, Kun Zhang, and Mingming Gong. 2022. Counterfactual Fairness with Partially Known Causal Graph. *arXiv preprint* (2022).

A IMPLEMENTATION DETAILS

We use two fully connected layers in neural networks to implement $\Phi(\cdot)$, $g(\cdot)$ and $h(\cdot)$, respectively. The softmax function is used on top of $h(\cdot)$ when the sensitive attribute is categorical. LeakyRelu is used as activation functions in our framework. We aggregate the counterfactuals with mean operation, and we use mean square error (MSE) to compute the target prediction loss. For CLAIRE-M, we adopt the implementation of MMD from [33], and the optimization problem can be solved by traditional stochastic gradient descent algorithms. For CLAIRE-A, following [11], the minimax optimization problem is conducted with an alternating gradient descent process. We use cosine distance to implement $d(\cdot, \cdot)$.

B DETAILS OF EXPERIMENT SETTINGS

B.1 Full Introduction of Baselines

- **Constant Predictor:** A predictor which has constant output can obviously satisfy counterfactual fairness. We obtain this constant predictor by finding a constant which can minimize the mean squared error (MSE) loss on the training data.
- **Full Predictor:** Full predictor takes *all* the observed attributes (except the attribute used as label) as input for prediction. We use linear regression for the regression task and logistic regression for the classification task.
- **Unaware Predictor:** Unaware predictor is based on the notion of fairness through unawareness [20]. It takes all features except the sensitive attribute as input to predict the label through linear regression for the regression task and logistic regression for the classification task.
- **Counterfactual Fairness Predictor:** We use two different counterfactual fairness predictors here: 1) As introduced in [30], the predictor infers the latent variables and uses them along with the observed variables which are non-descendants of the sensitive attributes; 2) As described in [42], the predictor takes the input of both sensitive and non-sensitive attributes, with a fairness term added in the loss function which minimize the difference of the predictions made on two counterfactuals. We refer to these two methods as CFP-U and CFP-O, respectively. We follow the original implementations in [30, 42], where CFP-U uses linear regression for the regression task and logistic regression for the classification task, and CFP-O is implemented with neural networks.

B.2 Detailed Experimental Setup

We use Pyro [8] to implement the causal models. The number of sampling in the counterfactual generation is set as 500. For the baselines CFP-U and CFP-O, the epochs for the causal model training is set as $2,000$ and the learning rate is set as 0.001. All the presented results are averaged over ten executions of experiments.

C MORE EXPERIMENTAL RESULTS

Table 4 shows the discrepancy of predictions made on different counterfactuals. In addition to the two pairs of counterfactuals ($S \leftarrow 0$ and $S \leftarrow 1$) and ($S \leftarrow 0$ and $S \leftarrow 2$) shown in Table 3, Table 4 also shows the results in pair ($S \leftarrow 2$ and $S \leftarrow 3$). Generally, the observation on the pair ($S \leftarrow 2$ and $S \leftarrow 3$) is similar to the aforementioned observation on the pair ($S \leftarrow 0$ and $S \leftarrow 2$).

Table 4: Study on synthetic data regarding the adverse effects of incorrect causal model $\mathcal{M}_2$.

Method	$S \leftarrow 2$ and $S \leftarrow 3$	
	MMD	WASS
CFP-U (true)	8.407 ± 0.810	2.900 ± 0.092
CFP-U (false)	10.317 ± 1.011	3.780 ± 0.052
CFP-O (true)	8.793 ± 0.927	3.136 ± 0.040
CFP-O (false)	10.337 ± 1.002	3.864 ± 0.030
CLAIRE-M	8.108 ± 0.024	2.860 ± 0.004
CLAIRE-A	**7.902 ± 0.055**	**2.761 ± 0.005**