Disentangled Representation with Causal Constraints for Counterfactual Fairness

Ziqi Xu  
University of South Australia  
Adelaide, Australia  
ziqi.xu@mymail.unisa.edu.au

Jixue Liu  
University of South Australia  
Adelaide, Australia  
jixue.liu@unisa.edu.au

Debo Cheng  
University of South Australia  
Adelaide, Australia  
dobo.cheng@unisa.edu.au

Jiuyong Li  
University of South Australia  
Adelaide, Australia  
jiuyong.li@unisa.edu.au

Lin Liu  
University of South Australia  
Adelaide, Australia  
lin.liu@unisa.edu.au

Ke Wang  
Simon Fraser University  
Burnaby, Canada  
wangk@cs.sfu.ca

ABSTRACT

Much research has been devoted to the problem of learning fair representations; however, they do not explicitly relate the relationship between latent representations. In many real-world applications, there may be causal relationships between latent representations. Furthermore, most fair representation learning methods focus on group-level fairness and are based on correlations, ignoring the causal relationships underlying the data. In this work, we theoretically demonstrate that using the structured representations enable downstream predictive models to achieve counterfactual fairness, and then we propose the Counterfactual Fairness Variational AutoEncoder (CF-VAE) to obtain structured representations with respect to domain knowledge. The experimental results show that the proposed method achieves better fairness and accuracy performance than the benchmark fairness methods.

1 INTRODUCTION

Machine learning algorithms have gradually penetrated into our life [34] and have been applied to decision-making for credit scoring [25], crime prediction [23] and loan assessment [10]. The fairness of these decisions and their impact on individuals or society have become an increasing concern. Some extreme unfair incidents have appeared in recent years. For example, COMPAS, a decision support model that estimates the risk of a defendant becoming a recidivist was found to predict higher risk for black people and lower risk for white people [4]; Google Photos are classifying black people as primates [56]; Facebook users receive a recommendation prompt when watching a video featuring blacks, asking them if they’d like to continue to watch videos about primates [31]. These incidents indicate that the machine learning models may become a source of unfairness, which may lead to serious social problems. Since most models are trained with data, which will lead to unfair decisions due to discrimination in the training data. Therefore, the key issue for solving unfair decisions becomes whether we can eliminate these biases embedded in the data through algorithms [27].

To obtain unbiased decisions, many methods [11, 15, 30, 32, 33, 36, 42, 44, 51] are proposed to learn fair representations through two competing goals: encoding data as much as possible, while eliminating any information that transfers through the sensitive attributes. To separate the information from sensitive attributes, various extensions of Variational Autoencoder (VAE) consider minimising the mutual information among latent representations [11, 30, 36, 42]. For example, Creager et al. [11] introduced disentanglement loss into the VAE objective function to decompose observed attributes into sensitive latents and non-sensitive latents to achieve subgroup level fairness; Park et al. [36] improved the above methods and proposed the mutual attribute latent (MAL) to retain only beneficial information for fair predictions.

All existing works of learning fair representations make the assumption that observed attributes in the real-world can be represented by a number of latent representations. Nevertheless, the latent representations may have causal relationships among them. Let us consider an example where we aim to predict a person’s salary using some observed attributes. Following the domain knowledge, we know that a person’s salary is determined by two semantic concepts, intelligence and career respectively. We also note that a person’s intelligence determines their career with high probability, which can be expressed as a conceptual level causal graph $G_c$, that is, \textit{Intelligence} $\rightarrow$ \textit{Career}. Figure 1a shows the causal graph that is learnt from the collected data, while the data itself is biased since the set of sensitive attributes $A$ can affect the target attribute $Y$.

The existing methods [11, 30] follow Figure 1b to achieve fair predictions. Specifically, this method uses $Z_a$ to represent the “concept” as mentioned while ensuring $Z_a$ do not contain sensitive information that transfer through the path $A \rightarrow X$. However, this method may not satisfy the domain knowledge since there are causal relationships within these “concepts”. Therefore, we need a method as shown in Figure 1c that not only ensures the representation of observed attributes with no sensitive information but also retains causal relationships with respect to domain knowledge.

We note that our method builds on the premise that $G_c$ is available, and we believe this assumption is valid. Fairness issues require humans to guide algorithms, and the causal graph should be given by humans rather than given by machine learning [6]. Compared with the complete version of the causal graph (i.e., a causal graph containing causal relationships between all observed attributes), $G_c$ only covers the relationship between these “concepts” and is easier to obtain expert consensus.

On the measurement of fairness, all fair representation learning methods use fairness metrics based on correlation, including the VAE-based methods [11, 30, 36, 42]. It is well known that correlation or more generally association does not imply causation. Recent studies [38, 39] have shown that quantifying fairness based on correlation may produce higher deviations. Counterfactual fairness
We use upper case letters to represent attributes and boldfaced upper case letters to denote the set of attributes. We use lower case letters to represent the values of the set of attributes. The values of attributes are represented using lower case letters.

Let $A$ be the set of sensitive attributes, which should not be used for predictive models; $X$ be the set of other observed attributes, which may have causal relationships with $A$; $V$ be the set of all observed attributes, i.e., $V = \{A, X\}$; $Y$ be the target attribute that may have causal relationships with attributes in $A$ and $X$. We use $\tilde{Y}(\cdot)$ to represent the predictor.

$G_c$ is the conceptual level causal graph and represents domain knowledge. The nodes shown in $G_c$ are “concepts”, each of which represents a set of observed attributes that have similar meanings. Each “concept” has causal relationships with the other “concepts”. For example, Intelligence is a “concept” in $G_c$ and it may represent several observed attributes that have similar meanings, including GPA, Education level and Major.

We define that $Z_a$ is the representation of $A$; $Z_x$ is the representation of $X$ without embedding causal relationships; $Z'_x$ is a structured version of $Z_a$ under the causal constraints of domain knowledge and does not contain sensitive information.

### 2.1 Counterfactual Fairness

In this paper, a causal graph is used to represent a causal mechanism. In a causal graph, a directed edge, such as $V_j \rightarrow V_i$ denotes that $V_j$ is a parent (i.e., direct cause) and we use $pa_i$ to denote the set of parents of $V_i$. We follow Pearl’s [39] notation and define a causal model as a triple $(U, V, F)$. $U$ is a set of the latent background attributes, which are the factors not caused by any attributes in the set $V = \{A, X\}$; $F$ is a set of deterministic functions, $V_i = f_i(pa_i, U_{pa_i})$, such that $pa_i \subseteq \{V_i\}$ and $U_{pa_i} \subseteq U$. Such equations are also known as structural equations [3]. Besides, some commonly used definitions in graphical causal modelling, such as faithfulness, $d$-separation and causal path can be found in [39, 40, 43].

With the causal model $(U, V, F)$, we have the following definition of counterfactual fairness:

**Definition 1 (Counterfactual Fairness [26]).** Predictor $\tilde{Y}(\cdot)$ is counterfactually fair if under any context $X = x$ and $A = a$,

$$P(\tilde{Y}_{A=a}(U) = y \mid X = x, A = a) = P(\tilde{Y}_{A-a}(U) = y \mid X = x, A = a),$$

(1)
for all $y$ and any value $a$ attainable by $A$.

Counterfactual fairness is considered to be related to individual fairness [26]. Individual fairness means that similar individuals should receive similar predicted outcomes. The concept of individual fairness when measuring the similarity of the individual is unknowable, which is similar to the unknowable distance between the real-world and the counterfactual world in counterfactual fairness [28].

2.2 Variational Autoencoder

Variational Autoencoder (VAE) was proposed by Kingma and Welling [24], which was originally applied to image dimensionality reduction. The objective of VAE is to maximise the Evidence Lower Bound (ELBO) $\mathcal{M}$, and derived as follows:

$$\log p(V) \geq \mathbb{E}_{q(Z|V)} [\log p(V|Z) + \log p(Z) - \log q(V|Z)] =: \mathcal{M}_{\text{VAE}}.$$ (2)

which can also be rewritten as follows:

$$\mathcal{M}_{\text{VAE}} = \mathbb{E}_{q(Z|V)} [\log p(V|Z)] - D_{KL} [q(Z|V)||p(Z)].$$ (3)

where $V$ denotes the set of all observed attributes and $Z$ denotes the set of learnt representations. The encoder $q$ encodes $V$ into $Z$, and the decoder $p$ reconstructs $V$ from $Z$.

The first part in Equation 3 can be considered as reconstruction error, i.e., the loss between the reconstructed and the original $V$. The second part is the distribution distance between the Gaussian prior $p(Z) = N(0, I)$ and the $Z$ encoded with $V$. The training process of a VAE is to learn the parameters in $q$ and $p$ through the neural networks.

Higgins et al. [18] modified the above VAE objective function by adding a hyperparameter $\beta$ that balances latent channel capacity and independence constraints with reconstruction accuracy. Then, they devised a protocol to quantitatively compare the degree of disentanglement learnt by different models and argued that each dimension of a correctly disentangled representation should capture no more than one semantically meaningful concept. The ELBO of $\beta$-VAE is defined as:

$$\mathcal{M}_{\text{\beta-VAE}} = \mathbb{E}_{q(Z|V)} [\log p(V|Z)] - \beta D_{KL} [q(Z|V)||p(Z)].$$ (4)

2.3 Disentangled Representation with Causal Constraints for Counterfactual Fairness

We discuss what types of representations enable downstream predictive models to achieve counterfactual fairness. Following the work in [16], we define the three steps for counterfactual inference.

Definition 2 (Counterfactual Inference [16]). Given a causal model $(U, V, F)$ and evidence $W$, where $W \subset V$, the counterfactual inference is the computation of probabilities $P(Y_{A \rightarrow a}(U|W = w))$.

- **Abduction**: for a given prior on $U$, compute the posterior distribution of $U$ given the evidence $W = w$;
- **Action**: substitute the equations for $A$ with the intervention values $a$, resulting in the modified set of equations $F_a$;
- **Prediction**: compute the implied distribution on the remaining elements of $V$ using $F_a$ and the posterior $P(U|W = w)$.

Following the work in [26], the implication of counterfactual fairness is described as follows:

**Proposition 3.1 (Implication of Counterfactual Fairness [26]).** Let $G$ be the causal graph of the given model $(U, V, F)$. If there exists $W$ be any non-descendant of $A$, then downstream predictor $\tilde{Y}(W)$ will be counterfactually fair.

We extend Proposition 3.1 to the fair representation learning and present the following theorem. We follow the similar proof process in work [26] to prove this theorem.

**Theorem 3.2.** Given the causal graph $G$, $Z_a$ is the representation of sensitive attributes $A$, $Z_x$ is the structured representation of the other observed attributes $X$ with respect to the conceptual level causal graph $G_c$. We have $\tilde{Y}(Z_a)$ satisfy counterfactual fairness.

**Proof.** Given the causal graph $G$ as shown in Figure 2, there is not a parent node of $A$ in $X$, and there is not a child node of $Y$ in $X$. $X$ contains four subsets: $X^A$ is the subset of other observed attributes that are descendants of $A$ and parents of $Y$; $X^*\setminus A$ is the subset of other observed attributes that are only parents of $Y$; $X^N$ is the subset of other observed attributes that are not relationships with $A$ and $Y$; $X_d$ is the subset of other observed attributes that are only descendants of $A$. After perfect representation learning, we obtain $Z_a$ and $Z_{x'}$.

We prove that $Z_a$ is not the descendant of $A$ with the following two subsets. For the first subsets $\{X^A, X^*, X_d\}$, there are seven paths between $A$ and $Z_a$, including $A \rightarrow X^A \leftarrow Z_a$, $A \rightarrow X^A \rightarrow Y \leftarrow Z_a$, $A \rightarrow X^* \rightarrow Y \leftarrow Z_a$, $A \rightarrow X_d \rightarrow Y \leftarrow Z_a$, $A \rightarrow Y \leftarrow Z_a$, $A \rightarrow Y \leftarrow X_d \leftarrow Z_a$, and $A \rightarrow Y \leftarrow X_d$. These seven paths are blocked by $\emptyset$ (i.e., $A$ and $Z_a$ are d-separated by $\emptyset$), since each path contains a collider either $X^A$ or $Y$ or $X_d$. For second subset $X^N$, there is no path connecting $X^N$ and $Y$. Hence, $Z_a$ is not the descendant of $A$. Therefore, $\tilde{Y}(Z_a)$ is counterfactually fair based on Proposition 3.1.

We use Figure 2 to show whether the following predictors satisfy counterfactual fairness.

3 PROPOSED METHOD

In this section, we first theoretically demonstrate that learning counterfactually fair representations are feasible. Then, we propose the Counterfactual Fairness Variational AutoEncoder (CF-VAE) to obtain the structured representations for predictors to achieve counterfactual fairness.
We first discuss the causal constraints and then explain the loss function of CF-VAE in detail. The architecture of CF-VAE is shown in Figure 3.

3.2 CF-VAE

We first discuss the causal constraints and then explain the loss function of CF-VAE in detail. The architecture of CF-VAE is shown in Figure 3.

3.2.1 Learning Representations with Causal Constraints. We aim to retain causal relationships between "concepts" through a more easily accessible conceptual level causal graph \( \mathcal{G}_c \) and embed these relationships in representations. Following the works in [48, 57], we transform these causal relationships in the form of an adjacency matrix \( C \) as causal constraints to construct \( \mathcal{Z}_x \) and feed them to predictive models.

Many researchers study fair decision-making under the method of causality, in which an accessible causal graph is an important assumption. Zhang et al. [54] used the PC algorithm [19, 43] to learn the causal relationships of the dataset itself and used the causal graph to restrict the transfer of sensitive information along the specific paths. Learning causal graphs from observational data has been shown to be feasible in unbiased data, but in fairness computing, the dataset may be biased and the PC algorithm may not be suitable. Other researches [8, 35] assume that the complete version of causal graph is accessible and evolved from domain knowledge. In our paper, we adopt the second approach and further weaken it. We assume that \( \mathcal{G}_c \) only covers the relationships between "concepts," not all observed attributes, which is easy to obtain consensus from experts.

To formalise causal relationships, we consider \( n \) "concepts" in the dataset, which means \( \mathcal{Z}_x \) should have the same dimension as "concepts". The "concepts" in observations are causally structured by \( G_c \) with an adjacency matrix \( C \). For simplicity, in this paper, the causal constraints are exactly implemented by a linear structural equation model as follows:

\[
\mathcal{Z}_x' = (I - C^T)^{-1}\mathcal{Z}_x.
\]

where \( I \) is the identity matrix. \( \mathcal{Z}_x \) is obtained from the encoder, \( \mathcal{Z}_x' \) is constructed from \( \mathcal{Z}_x \) and \( C \). \( C \) is obtained from \( \mathcal{G}_c \) with respect to domain knowledge. The parameters in \( C \) indicate that there are corresponding edges, and the values of the parameters indicate the weight of the causal relationships. It is worth noting that if the parameter value is zero, it means that such an edge does not exist, i.e., no causal relationship between these two "concepts".

As mentioned above, \( \mathcal{Z}_x \) is obtained from the encoder, we cannot guarantee that each attribute inside is independent. To ensure the independence of each attribute in \( \mathcal{Z}_x \), we follow the Factor-VAE in work [21] and employ the total correction regularisation (TCR) in our loss function. TCR also encourages the correctness of structured \( \mathcal{Z}_x \) with respect to domain knowledge since there are no correlations in \( \mathcal{Z}_x \) before adding causal constraints. The TCR for our proposed CF-VAE is defined as:

\[
\mathcal{L}_{\text{TCR}} = yD_{KL}[q(Z_x)||\prod_{i=1}^{D_{\mathcal{Z}_x}} q(Z_{x_i})],
\]

where \( y \) is the weight value, \( D_{\mathcal{Z}_x} \) is dimension of \( \mathcal{Z}_x \).
We can make their inner product close to zero, then the latent representations are closer to being orthogonal. The orthogonality promoting regularisation based on the pairwise cosine similarity is defined as:

\[ CS(E_1, E_2) = \frac{E_1^T E_2}{\|E_1\|_2 \|E_2\|_2}, \]  

where \( \| \cdot \|_2 \) is the \( l_2 \) norm.

To encourage orthogonality between two vectors \( E_1 \) and \( E_2 \), we can make their inner product \( E_1^T E_2 \) close to zero and their \( l_2 \) norm \( \|E_1\|_2, \|E_2\|_2 \) close to one [47]. The orthogonality promoting regularisation (OPR) for our proposed CF-VAE is defined as:

\[ L_{OPR} = \frac{1}{B} \sum_{i=1}^{B} CS(Z_{a_i}, Z'_{x_i}), \]

where \( B \) denotes the batch size for neural network.

In conclusion, the loss function of our proposed CF-VAE is defined as:

\[ L_{CF-VAE} = -M' + L_{TCR} + L_{OPR}. \]  

### 4 EXPERIMENTS

In this section, we conduct extensive experiments to evaluate CF-VAE on real-world datasets. Before showing the detailed results, we first present the details of selected methods and the evaluation metrics.

#### 4.1 Framework Comparison

The proposed CF-VAE is considered as a pre-processing technique to address fairness issues since it obtains structured representations for downstream predictive models to achieve counterfactual fairness. Hence, we compare CF-VAE with traditional and VAE-based pre-processing methods. For traditional methods, we select baselines including ReWeighting (RW) [20], Disparate Impact Remover (DIR) [14] and Optimized Preprocessing (OP) [5]. Both of them are available in AIF360 [1]. For VAE-based methods, we compare with VFAE [30] and FFVAE [11]. Both of them are implemented in Pytorch [37]. We also obtain the Full model for comparison, which uses all attributes in the dataset to make predictions.

We do not choose the basic VAE (e.g., VAE [24], \( \beta \)-VAE [18], and Factor-VAE [21]) for comparison in this experiment, since they are not optimised for fairness problems. In addition, we do not use VAE-based inference models [8, 22, 41, 48] for comparison, because the purpose of these inference models is to generate counterfactual data or to estimate effects, which is different from our goals.

We select several well-known predictive models to simulate the downstream prediction process. Linear Regression (LR), Stochastic Gradient Descent Regression (SGD\( r \)) and Multi-layer Perceptron Regression (MLP\( r \)) are used for regression tasks; Logistic Regression (LR\( C \)), Stochastic Gradient Descent Classification (SGD\( C \)) and Multi-layer Perceptron Classification (MLP\( C \)) are used for classification tasks. For each predictive model, we run 10 times and record the mean and error of the results for evaluation metrics, which are explained in detail in Section 4.2.
The situation test has already been widely used in United States law school, and their first-year average grade (FYA). The school scores (LSAT), their grade point average (GPA) collected prior to information of 21,790 law students, including their entrance exam information from 163 law schools in the United States. It contains scores of sensitive attributes. We take this matched pair as the input to the predictive model, and the predictive model is fair if the predictions of the matched pair are the same as the original pair.

We define unfairness score (UFS) to measure the result of the situation test. Specifically, the form of score differs for different predictive models. For regression tasks, we define UFS_R that measure the bias between prediction results for the matched pair and the original pair; For classification tasks, UFS_C is defined as how many individuals’ prediction results are changed after intervening the values of sensitive attributes. UFS_R and UFS_C are described as follows:

\[
\text{UFS}_R = \frac{1}{N} \sum_{i=1}^{N} \left( \hat{y}_{\text{A} \rightarrow \hat{a}}(Z_x^i) - \hat{y}_{\text{A} \rightarrow \hat{a}}(Z_{x'}^i) \right)^2;
\]

\[
\text{UFS}_C = \frac{1}{N} \sum_{i=1}^{N} \text{xor} \left( \hat{y}_{\text{A} \rightarrow \hat{a}}(Z_x^i), \hat{y}_{\text{A} \rightarrow \hat{a}}(Z_{x'}^i) \right),
\]

where \( N \) is the number of samples for evaluation.

The lower UFS value means that the predictive models achieve higher individual fairness.

4.2.2 Accuracy. We evaluate the performance on prediction with the following metrics. For regression tasks, we use Root Mean Square Error (RMSE) to compare the error between prediction results and target attributes’ values. For classification tasks, we use accuracy to evaluate various predictive models.

4.3 Law School

The law school dataset comes from a survey [46] of admissions information from 163 law schools in the United States. It contains information of 21,790 law students, including their entrance exam scores (LSAT), their grade point average (GPA) collected prior to law school, and their first-year average grade (FYA). The school expects to predict if the applicants will have a high FYA. Gender and race are sensitive attributes in this dataset, and the school also wants to ensure that predictions are not affected by sensitive attributes. However, LSAT, GPA and FYA scores may be biased due to socio-environmental factors. The process of CF-VAE for the Law school dataset is shown in Figure 4a.

4.3.1 Implementation Details. We divide the Law school dataset into 70% training set for training the representation models, 30% testing set for evaluating the accuracy of the predictive models, and inverting the values of sensitive attributes in the testing set to generate the auditing set for evaluating the fairness of the predictive models.

We use the same \( G_c \) as shown in work [26] to model latent “concepts” of GPA and LSAT. Since \( \text{known}_1 \) and \( \text{known}_2 \) have no causal relationship, the parameters in adjacency matrix \( C \) are set to zero. As a results, we set the \( D_\gamma = 2 \) and set the weight value in \( L_{TCR} \) as \( \gamma = 10 \).

4.3.2 Fairness. The purpose is to demonstrate our method can achieve better fairness performance than other VAE-based methods. As shown in Table 1, since the Full model uses sensitive attributes to make predictions, inverting sensitive attributes has the highest impact on the individual’s prediction results, which means that the model is unfair. RW, DBC and OP achieves fair predictions by modifying the dataset compared to the Full model. Both VFAE and FFVAE disentangle the sensitive attributes with latent representations, so the influence of inverting the sensitive attributes on the prediction results is small. Our method achieves the lowest UFS_R, 0.013, 0.025, and 0.044 for LR, SGD, and MLP respectively, which means CF-VAE disentangle \( Z'_s \) and \( Z_s \) more precisely.

4.3.3 Accuracy. The accuracy results are shown in Table 1. The Full model is unfair and it uses sensitive information to more accurately predict FYA and thus achieves the highest accuracy. The proposed CF-VAE achieves the best fairness aware accuracy in all predictive models than other methods. Our method not only achieves counterfactual fairness for downstream predictors but also flexible for choosing predictive models.
The process of CF-VAE is shown in Figure 4b. The Adult dataset comes from the UCI repository [12] contains 10 attributes including race, age, education information, marital information as well as capital gain and loss for 48,842 individuals. We divide the Adult dataset into 70% training set for training representations with sensitive attributes. After that, we get 45,222 individuals and the downstream tasks’ goal is to predict whether the individual’s income is above $50,000. We set race, age, sex and native country as A; marital, relationship, work class, work hour, occupation and education level as X. We divide the Adult dataset into 70% training set for training representation models, 30% testing set for evaluating the accuracy of the predictive models, and select 10,000 individuals with female that income below 50K as the auditing set.

We use the same $G_c$ as shown in previous research [8, 35] to model the latent “concepts”. We set the $D_{x_c} = 3$ and set the weight value in $L_{TCR}$ as $y = 10$. The adjacency matrix $C$ is defined as:

$$C = \begin{bmatrix} 0 & \lambda_{12} & \lambda_{13} \\ 0 & 0 & \lambda_{23} \\ 0 & 0 & 0 \end{bmatrix}$$

Then, we construct $Z'_x$ from $Z_x$ and C as follows:

$$Z'_{x_1} = Z_{x_1}; \quad Z'_{x_2} = \lambda_{12} Z_{x_1} + Z_{x_2}; \quad Z'_{x_3} = \lambda_{13} Z_{x_1} + \lambda_{23} Z_{x_2} + Z_{x_3}.$$  

4.4.2 Fairness. The fairness results are shown in Table 2, the Full model achieves the worst UFS$_C$, since it use A to predict income. Both baseline fairness models and other VAE-based methods improve fairness to a certain extent. The proposed CF-VAE achieves the best UFS$_C$ only 3.1%, 2.0% and 2.4% of individuals’ results are affected by sensitive attributes’ values inversions in LR$_C$, SGD$_C$ and MLP$_C$, respectively. Our method achieves better fairness performance than other methods, since it remains causal relationships in latent representations with respect to $G_c$ and disentangles structured representations with sensitive attributes.

4.4.3 Accuracy. The Full model uses all observed attributes for predictions. It is worth noting that the Full model does not achieve the 85% accuracy shown in [12], because we omit capital gain and loss, and achieve similar accuracy as shown in the work [35]. The accuracy results are shown in Table 2. In order to achieve fairness, VFAE and FFVAE lose about 2% of their accuracy performance. The proposed CF-VAE not only guarantees the fairness performance but also retains the causal relationships to improve accuracy. CF-VAE loses less information than other VAE-base models and achieves the best fairness aware accuracy performance in all predictive models, i.e., 80.1%, 79.4% and 82.0% in LR$_C$, SGD$_C$ and MLP$_C$, respectively.

### Table 1: The results for Law School dataset. The best fairness aware RMSE and the best UFS$_A$ are shown in bold.

| Model | Accuracy (RMSE) $\downarrow$ | Fairness (UFS$_A$) $\downarrow$ |
|-------|------------------------------|-------------------------------|
|       | LR$_R$                       | SGD$_R$                      | MLP$_R$          |
| Full  | 0.865 ± 0.007                | 0.867 ± 0.007                | 0.865 ± 0.007   |
| RW    | 0.955 ± 0.013                | 0.956 ± 0.012                | 0.953 ± 0.012   |
| DIR   | 0.943 ± 0.009                | 0.944 ± 0.009                | 0.941 ± 0.010   |
| OP    | 0.959 ± 0.011                | 0.960 ± 0.011                | 0.956 ± 0.010   |
| VFAE  | 0.932 ± 0.007                | 0.933 ± 0.007                | 0.934 ± 0.007   |
| FFVFAE| 0.933 ± 0.005                | 0.934 ± 0.004                | 0.935 ± 0.005   |
| CF-VAE| 0.931 ± 0.006                | 0.932 ± 0.006                | 0.932 ± 0.006   |

### Table 2: The results for Adult dataset. The best fairness aware accuracy and the best UFS$_C$ are shown in bold.

| Model | Accuracy $\uparrow$ | Fairness (UFS$_C$) $\downarrow$ |
|-------|---------------------|-------------------------------|
|       | LR$_C$              | SGD$_C$                      | MLP$_C$          |
| Full  | 0.802 ± 0.002       | 0.803 ± 0.004                | 0.831 ± 0.004   |
| RW    | 0.797 ± 0.001       | 0.792 ± 0.002                | 0.819 ± 0.001   |
| DIR   | 0.800 ± 0.001       | 0.793 ± 0.003                | 0.817 ± 0.001   |
| OP    | 0.780 ± 0.002       | 0.779 ± 0.003                | 0.783 ± 0.002   |
| VFAE  | 0.785 ± 0.001       | 0.781 ± 0.003                | 0.819 ± 0.004   |
| FFVFAE| 0.785 ± 0.003       | 0.782 ± 0.001                | 0.814 ± 0.005   |
| CF-VAE| 0.801 ± 0.002       | 0.794 ± 0.004                | 0.820 ± 0.002   |

4.4 Adult

The Adult dataset comes from the UCI repository [12] contains 14 attributes including race, age, education information, marital information as well as capital gain and loss for 48,842 individuals. We follow the same procedure in [7] to generate synthetic datasets and conduct an ablation study to validate the contribution of each component in our method as shown in Table 3.

The Full model uses all the observed attributes to train the predictors. The predictors achieve the best accuracy but the worst fairness
performance as shown in the first row in Table 3. VFAE is the basic VAE-based unsupervised fair representation learning method. We set it to be the baseline in the second row in Table 3. The third row is CF-VAE without adding causal constraints, which achieves similar results as VFAE since both methods remove sensitive information from the learnt representations.

Then, we employ causal constraints and add TCR (γ = 10) in the loss function. As shown in the fourth row in Table 3, this step retains causal relationships in latent representations and improves both accuracy and fairness performance than previous rows. The last step is to encourage \( Z' \) and \( Z_a \) are disentangled by adding OPR. Our proposed CF-VAE achieves the best accuracy performance and UFS\(_R\) among most predictive models as shown in the last row in Table 3.

5 RELATED WORKS

The machine learning literature has increasingly focused on exploring how algorithms can protect marginalised populations from unfair treatment. An important research area is how to quantify fairness, which can be divided into two categories, the statistical framework and the causal framework.

In the statistical framework, Demographic parity was defined by Zemel et al. [51], which is used to measure group level fairness. Other similar metrics include equalised odds [17], predictive rate parity [50]. Dwork et al. [13] proposed a measurement to quantify individual level fairness, that is, similar individuals should have similar treatments, and they use distance functions to measure how similar between individuals. In the causal framework, the (conditional) average causal effect is used to quantify fairness between groups [29]; Natural direct and natural indirect effects are used to quantify specific fairness [35, 52, 55]; When unfair causal paths are identified by domain knowledge, Chiappa [8] used the path-specific causal effects to quantify fairness on approved paths. For more related works, please refer to the literature review [9, 34, 53].

Our work is related to learning fair representations, which aims to encode data information into a lower space while removing sensitive information, and remaining causal relationships with respect to domain knowledge for building counterfactually fair predictive models. VAE [24] and β-VAE [18], as introduced in Section 2.2, have inspired several studies in fair representation learning. Louizos et al. [30] first introduced VAE for learning fair representation to disentangle the sensitive information and non-sensitive information, they proposed a semi-supervised method to encourage disentanglement by using “Maximum Mean Discrepancy” (MMD). However, Zemel et al. [51], Gitiaux and Rangwala [15] argued that in real-world applications, the organisations that collect the data cannot predict the downstream uses of the data and the models that might be used. Due to this, there are many following up works focusing on unsupervised learning fair representation. For example, Creager et al. [11] proposed an algorithm that can achieve group level fairness by adding demographic parity as a constraint in objection function; Song et al. [42] developed an information theory-based method for learning maximally expressive representations subject to fairness constraints that allows users to control the fairness of representations by specifying limits on unfairness.

Our approach combines counterfactual fairness and unsupervised representation learning to provide the proper representations to help predictive models achieve counterfactual fairness. We extend the definition of counterfactual fairness [26] to the representation learning. Based on the current literature review, our work is the first method to use VAE-based techniques for unsupervised representation and satisfy counterfactual fairness. Furthermore, we innovatively embed domain knowledge into representations by adding causal constraints with respect to domain knowledge.

6 CONCLUSION

In this paper, we investigate unsupervised counterfactually fair representation learning and propose a novel method named CF-VAE which considers causal relationships with respect to domain knowledge. We theoretically demonstrate that the structured representations obtained by CF-VAE enable predictive models to achieve counterfactual fairness. Experimental results on real-world datasets show that CF-VAE achieves better accuracy and fairness performance on downstream predictive models than the benchmark fairness methods. Ablation study on synthetic datasets shows that causal constraints with total correction regularisation achieve better accuracy performance and orthogonality promoting regularisation encourages disentanglement with sensitive attributes.

ACKNOWLEDGMENTS

Supported by Australian Research Council (DP200101210) and Postgraduate Research Scholarship of University of South Australia.
