Causal Fair Metric: Bridging Causality, Individual Fairness, and Adversarial Robustness

A Preprint

Ahmad-Reza Ehyaei
Max Planck Institute for Intelligent Systems, Tübingen, Germany
ahmad.ehyaei@tuebingen.mpg.de

Golnoosh Farnadi
McGill University; University of Montréal; MILA , Montréal, Canada
farnadig@mila.quebec

Samira Samadi
Max Planck Institute for Intelligent Systems, Tübingen, Germany
ssamadi@tuebingen.mpg.de

October 30, 2023

Abstract

Adversarial perturbation is used to expose vulnerabilities in machine learning models, while the concept of individual fairness aims to ensure equitable treatment regardless of sensitive attributes. Despite their initial differences, both concepts rely on metrics to generate similar input data instances. These metrics should be designed to align with the data's characteristics, especially when it is derived from causal structure and should reflect counterfactuals proximity. Previous attempts to define such metrics often lack general assumptions about data or structural causal models. In this research, we introduce a causal fair metric formulated based on causal structures that encompass sensitive attributes. For robustness analysis, the concept of protected causal perturbation is presented. Additionally, we delve into metric learning, proposing a method for metric estimation and deployment in real-world problems. The introduced metric has applications in the fields adversarial training, fair learning, algorithmic recourse, and causal reinforcement learning.

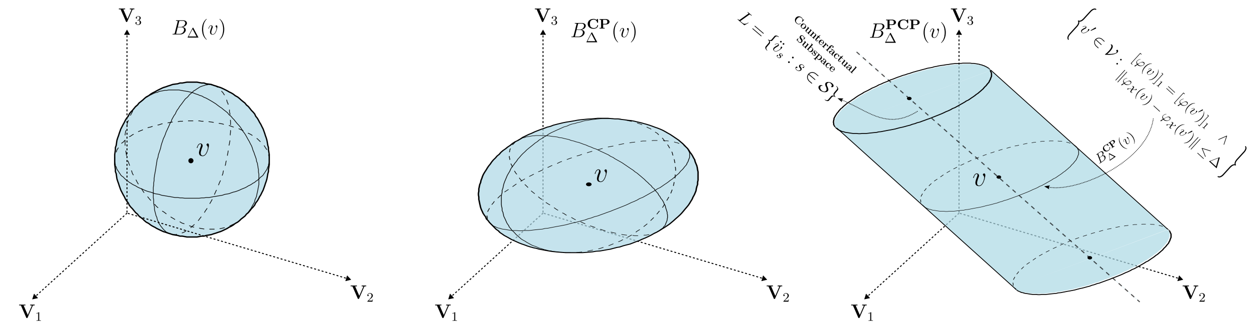
***K*eywords** Structural Causal Model · Individual Fairness · Adversarial Robustness · Metric Learning

1 Introduction

The concept of *individual fairness*, as delineated by Dwork et al. (2012), along with notions like group fairness, provides pivotal frameworks to understand fairness in the domain of machine learning and algorithmic decision-making. This notion emphasizes the equitable treatment of analogous individuals to preclude discrimination stemming from individual attributes. The formulation of individual fairness, as articulated through either the Lipschitz formulation Dwork et al. (2012) or the $\epsilon - \delta$ approach John et al. (2020), necessitates the development and evaluation of a *fair metric*. Such metrics serve as pivotal quantitative tools for assessing the conformity of algorithms to the tenets of individual fairness.

Adversarial perturbation, as outlined by Goodfellow et al. (2014) and Madry et al. (2017), involves the purposeful manipulation of input data to uncover machine learning model vulnerabilities or assess robustness. It closely intersects with metrics, quantifying how alterations in input data impact model performance, often employing distance metrics to quantify differences between original and perturbed inputs.

When confronted with a causal structure underlying the data, conventional metrics such as the Euclidean norm which does not capture causal relations, prove to be inadequate Kilbertus et al. (2017). This intricacy becomes apparent when


$\mathbf{V}_3$
$B_\Delta(v)$
v
$\mathbf{V}_1$
$\mathbf{V}_2$
$\mathbf{V}_3$
$B_\Delta^{\mathbf{CP}}(v)$
v
$\mathbf{V}_1$
$\mathbf{V}_2$
$L = \{\ddot{v}_s : s \in \mathcal{S}\}$
Counterfactual Subspace
$B_\Delta^{\mathbf{PCP}}(v)$
$\mathbf{V}_3$
$\{v' \in \mathcal{V} : [\varphi(v)]_1 = [\varphi(v')]_1 \wedge \|\varphi_\mathcal{X}(v) - \varphi_\mathcal{X}(v')\| \leq \Delta\}$
$B_\Delta^{\mathbf{CP}}(v)$
v
$\mathbf{V}_1$
$\mathbf{V}_2$

(a) $\{v' \in \mathcal{V} : \|v - v'\| \leq \Delta\}$ (b) $\{v' \in \mathcal{V} : \|\varphi(v) - \varphi(v')\| \leq \Delta\}$ (c) $\{v' \in \mathcal{V} : \|\varphi_\mathcal{X}(v) - \varphi_\mathcal{X}(v')\| \leq \Delta\}$

Figure 1: To illustrate the transition from a basic perturbation to a protected causal perturbation ball, consider a simple linear SCM with Euclidean norm in both exogenous and endogenous spaces. (a) Perturbation ball without causality and sensitive feature protection, (b) Perturbation ball with causality, assuming no sensitive features, (c) Perturbation ball with causality, considering V_1 as a sensitive feature.

aiming to achieve equitable treatment concerning sensitive attributes. In this context, an ideal metric should yield minimal values for counterfactual instances associated with each data point.

Prior studies often simplified counterfactual computations by altering sensitive feature levels. In the recent work by Dominguez-Olmedo et al. (2022), they incorporated adversarial perturbations into structural causal models, primarily focusing on continuous features, possibly overlooking fairness concerns. Conversely, Ehyaei et al. (2023b) introduced a fairness metric based on causal functional structure, designed to align with sensitive attribute protection. However, their methodology is confined to specific causal structures, and their fairness metric lacks a well-established axiomatic foundation.

In this study, we aim to bridge the gap between fairness metrics in causal structures and sensitive attributes. We begin by identifying suitable properties that align with our objectives and subsequently derive this metric from observational data. In § 3, we introduce a definition for a *causal fair metric* applicable to any structural causal model, effectively addressing both causality and the protection of sensitive attributes. In § 4, we employ the causal fair metric to generate an adversarial perturbation, referred to as a *protected causal perturbation* and investigate the geometric properties and attributes of this adversarial perturbation.

Constructing a causal fair metric typically requires knowledge of structural causal models, which is often unavailable in many real-world applications. To overcome this limitation, we aim to derive the metric from data. In § 5, we illustrate that relying solely on observational or interventional data is insufficient for learning the causal fair metric. To address the absence of SCMs, an alternative approach involves metric learning using tagged distance data. These tags indicate proximity values or labels indicating data point closeness. By discussing the requirements of other methods, we focus on deep metric learning due to its compatibility with the structure of causal fair metrics. To enhance practicality, we employ contrastive and triplet deep metric learning scenarios.

In § 6, we conduct experiments on synthetic and real-world datasets to empirically validate our theoretical findings. We regard the Siamese metric learning method as our baseline across various scenarios. The results indicate that knowing the structure of the causal fair metric improves learning performance in deep metric network design scenarios. Furthermore, our empirical analysis reveals that in terms of balancing applicability and accuracy, label-based metric approaches are more practical and align better with the notion of protected causal perturbation. In summary, our main contributions are:

- **Causal Fair Metric (§3):** We present a causal fair metric that incorporates both causal considerations and the protection of sensitive attributes.
- **Metric Representation (§3):** We demonstrate how causal fair metric can be embedded in exogenous space.
- **Protected Causal Perturbation (§4):** We utilize the causal fair metric to generate adversarial perturbation within causal structures while addressing fairness concerns.
- **Impossibility of Metric Learning (§5):** We show that, Without SCM assumptions, learning a fair metric from observational or interventional distributions is not guaranteed.
- **Metric Learning (§5):** We introduce a learning algorithm designed to extract a causal fair metric from empirical data, with a focus on both causality and fairness considerations.

Related Work. Previous research has aimed to define adversarial perturbation and learn fair metrics. Notably, Ehyaei et al. (2023b) worked on constructing a fair metric in the presence of causal structures and sensitive attributes. However, their metric is limited to an additive noise model and lacks a full characterization of its properties. Our work is inspired by their approach. In Mukherjee et al. (2020), the authors attempted fair metric learning, but their method didn't heavily rely on causal structure. They assumed an embedding into a space where sensitive attributes form a linear subspace but didn't clarify its connection to SCM. Moreover, they assumed knowledge of the embedding map during metric learning. In Ilvento (2020), submetrics were developed for learning metrics for individual fairness using human judgments. Under specific assumptions about point distribution and representative point selection, these submetrics maintained accuracy relative to the true metric. However, this work didn't address the impact of sensitive attributes, which often compromise metric properties.

Spectral-based metric estimation methods, akin to those in Zhang et al. (2016) and Olson (2022), often require specific embedding kernel forms or observations of all pairwise distances $d(v_i, v_j)$ for guaranteed metric convergence. Fair representation learning Zemel et al. (2013); McNamara et al. (2017); Ruoss et al. (2020) aims to map individuals to prototypes, eliminating protected attributes, while preserving performance-relevant information during training. Another non-linear metric estimator, is the tree-based approach, proposed by Demirović and Stuckey (2021). They introduced a novel algorithm using bi-objective optimization to compute decision trees that are provably optimal for non-linear metrics. Online learning algorithms, as in Bechavod et al. (2020); Gillen et al. (2018), ensure a finite number of fairness constraint violations and bounded regret, relying on some metric-based assumptions.

Various spectral, probabilistic, and deep metric learning methods are discussed in Ghojogh et al. (2022, 2023); Suárez et al. (2021); Francis and Raimond (2021). To the best of our knowledge, none of the existing algorithms address the integration of causal structure and sensitive attributes in metric learning.

2 Background

Structural Causal Model. A *structural causal model* (SCM) for a set of n random variables $\mathbf{V} = \{\mathbf{V}_i\}_{i=1}^n$ is represented by the tuple $\mathcal{M} = \langle \mathcal{G}, \mathbf{V}, \mathbf{U}, \mathbb{F}, \mathbf{P}_{\mathcal{U}} \rangle$ Pearl (2009), where:

- The set $\mathbb{F} = \{\mathbf{V}_i := f_i(\mathbf{V}_{\text{Pa}(i)}, \mathbf{U}_i)\}_{i=1}^n$ contains *structural equations*, with each equation f_i denoting the causal connection between the endogenous variable $\mathbf{V}_i$, its direct causal parents $\mathbf{V}_{\text{Pa}(i)}$ from $\mathcal{G}$, and an exogenous variable $\mathbf{U} = \{\mathbf{U}_i\}_{i=1}^n$ signifying unobservable background influences. In this work, we suppose $\mathcal{G}$ is a directed acyclic graph.
- The distribution $\mathbb{P}_{\mathbf{U}}$ of exogenous noise variables factorizes, $\mathbb{P}_{\mathbf{U}} = \prod_{i=1}^n \mathbb{P}_{\mathbf{U}_i}$, due to the assumption of *causal sufficiency*.

Under acyclicity, each instance $u \in \mathcal{U}$ of the exogenous space $\mathcal{U}$ uniquely determined by $v \in \mathcal{V}$ with the reduced-form mapping $g : \mathcal{U} \to \mathcal{V}$, where g is obtained by iteratively substituting the structural equations $\mathbb{F}$ following the causal graph's topological order $\mathcal{G}$. The SCM entails a unique joint distribution $\mathbb{P}_X$ over the endogenous variables through the reduced-form mapping, $\mathbb{P}_{\mathbf{V}}(\mathbf{V} = v) := \mathbb{P}_{\mathbf{U}}(\mathbf{U} = g^{-1}(v))$ where g^{-1} is the preimage of g.

Causal Identifiability. Discovering true causal connections among variables solely from observational data typically necessitates additional assumptions about the structural functions $\mathbb{F}$. One identifiable family of SCMs is the additive noise model (ANM) Hoyer et al. (2009), represented by $\mathbf{V} = f(\mathbf{V}) + \mathbf{U}$. In ANMs, obtaining the relationship from u to v is straightforward when considering I as the identity function ($I(v) = v$), then g is obtained by $g = (I - f)^{-1}$. Post-nonlinear models Zhang and Hyvarinen (2012) and location-scale noise models Immer et al. (2023) are another identifiable SCM families.

Interventions. SCMs facilitate modeling and assessing the impact of external manipulation on the system represented by the intervention Peters et al. (2017). Two main intervention types are *hard interventions* and *soft interventions*. In hard interventions (expressed as $\mathcal{M}^{do(\mathbf{V}_{\mathcal{I}}:=\theta)}$), a subset $\mathcal{I} \subseteq \{1, \ldots, n\}$ of features $\mathbf{V}_{\mathcal{I}}$ is forcibly fixed to a constant $\theta \in \mathbb{R}^{|\mathcal{I}|}$ by excluding relevant parts of the structural equations:

$$\mathbb{F}^{do(\mathbf{V}_{\mathcal{I}}:=\theta)} = \begin{cases} \mathbf{V}_i := \theta_i & \text{if } i \in \mathcal{I} \\ \mathbf{V}_i := f_i(\mathbf{V}_{\mathbf{Pa}(i)}, \mathbf{U}_i) & \text{otherwise} \end{cases}$$

Hard interventions disrupt the causal connections between affected variables and their ancestral components in the causal graph, whereas soft interventions maintain all causal relationships while adjusting the structural equation functions. For example, *additive (shift) intervention* Eberhardt and Scheines (2007), denoted as $\mathcal{M}^{do(\mathbf{V}_{\mathcal{I}} \mathrel{+}= \delta)}$, modify features $\mathbf{V}$ using a perturbation vector $\delta \in \mathbb{R}^n$ with equations $\left\{V_i := f_i\left(\mathbf{V}_{\mathbf{Pa}(i)}, \mathbf{U}_i\right) + \delta_i\right\}_{i=1}^n$.

Counterfactuals. *Counterfactual* is a hypothetical scenario that represents what would have happened if certain interventions or changes were applied to the variables in the SCM. The counterfactual outcome $\mathbf{CF}(v, \theta)$ for a specific variable $\mathbf{V}_{\mathcal{I}}$ under the hard intervention $do(\mathbf{V}_{\mathcal{I}} := \theta)$ can be computed using the modified structural equations as $g^{\theta}(g^{-1}(v))$, where g^{θ} represents the altered reduced-form mapping $\mathcal{M}^{do(\mathbf{V}_{\mathcal{I}} := \theta)}$ after the intervention.

Sensitive Attribute. A sensitive attribute, like race, holds ethical or legal significance in decision-making, such as in hiring, lending, or criminal justice, determining equitable treatment or outcomes for individuals or groups. Let $\mathbf{S} \in \{\mathbf{V}_1, \ldots, \mathbf{V}_n\}$ represent a sensitive attribute with domain $\mathcal{S}$ (discrete or continuous). For each instance $v \in \mathcal{V}$, the set of *counterfactual twins* regarding the sensitive feature $\mathbf{S}$ is obtained by $\ddot{\mathbf{v}} = \{\ddot{v}_s = \mathrm{CF}(v, s) : s \in \mathcal{S}\}$.

Individual Fairness. *Individual fairness*, as introduced by Dwork et al. (2012), ensures equitable treatment for individuals with comparable predefined metric similarities. Two formulations, including the Lipschitz mapping-based formulation Dwork et al. (2012):

$$d_{\mathcal{Y}}(h(v), h(v')) \leq L\, d_{\mathcal{V}}(v, v') \quad \forall v, v' \in \mathcal{V}$$

and the ϵ-δ formulation John et al. (2020):

$$\forall v, v' \in \mathcal{V} \quad d_{\mathcal{V}}(v, v') \leq \delta \quad \Longrightarrow \quad d_{\mathcal{Y}}(h(v), h(v')) \leq \epsilon$$

have been proposed. Where, $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ are metrics for the input and output spaces, respectively, with h as the classifier and $L \in \mathbb{R}_{+}$. The core of the definition lies in the *fair metric* $d_{\mathcal{X}}$, which gauges individual similarity based on pertinent attributes.

Counterfactual fairness, introduced by Kusner et al. (2017), embodies fairness through causal models, contrasting an individual's outcomes in the real world with those in a hypothetical scenario involving a distinct levels of sensitive features. The classifier h is considered counterfactually fair if it meets the condition $h(\ddot{v}_s) = h(\ddot{v}_{s'}) \quad \forall s, s' \in \mathcal{S}$.

3 Fair Metric

The fair metric, widely used in previous studies, lacks clarity when applied in problems involving both causal structures and sensitive attributes. Prior works, such as Yurochkin et al. (2019) and Mukherjee et al. (2020), have addressed the inclusion of sensitive attributes in impartial evaluation metrics. For instance, Yurochkin et al. (2019) introduced a fair metric denoted as $d_{\mathcal{X}}(x_1, x_2) = \langle (x_1 - x_2), \Sigma(x_1, x_2) \rangle$, where Σ is defined as $\Sigma = (I - P_{\mathrm{ran}(S)})$, with the matrix S encompassing k dimensions corresponding to sensitive features. However, due to limited consideration of causality, our goal is to enhance the fair metric by integrating it with causal frameworks. Incorporating a causal structure alongside a sensitive attribute requires a dissimilarity function that remains zero for counterfactual twins and maintains stability against minor perturbations of non-sensitive attributes. These characteristics are fundamental to metric formulation.

In Dominguez-Olmedo et al. (2022) and Ehyaei et al. (2023a), additive interventions are employed as a form of perturbation in additive noise models. In their work, adding noise in ANMs is easily interpretable as perturbing the noise. However, in arbitrary SCMs, this direct interpretation is not applicable. For interpretative clarity, we opt for soft interventions that perturb exogenous variable values instead of additive interventions.

Definition 1 (Additive Noise Intervention) *In additive noise intervention, δ is added to the noise term to manipulate the SCM:*

$$\mathbb{F}^{do(\mathbf{V}_{\mathcal{I}} \mathrel{+}= \delta)} = \left\{V_i := f_i\left(\mathbf{V}_{\mathbf{Pa}(i)}, \mathbf{U}_i + \delta_i\right)\right\}_{i=1}^n$$

The counterfactual CF(v, δ) for additive noise interventions is defined under the intervention $do(\mathbf{V}_{\mathcal{I}} \mathrel{+}= \delta)$.

Through the aforementioned intervention, we establish the definition of a causal fair metric.

Definition 2 (Causal Fair Metric) *Let $d : \mathcal{V} \times \mathcal{V} \to \mathbb{R}_{\geq 0}$ represent a metric [1] defined on the feature space $\mathcal{V}$, generated by a SCM $\mathcal{M}$. Let $\mathbf{S}$ denote a sensitive attributes, and $\mathcal{I}$ represent the index set of sensitive features within the SCM. The metric is called a causal fair metric if it adheres to the following properties:*

- *For all $v \in \mathcal{V}$ and $s \in \mathcal{S}$, the metric is zero only for twin pairs, i.e., $d(v, \ddot{v}_s) = 0$.*
- *For all $v \in \mathcal{V}$, with respect to δ, the function $d(v, CF(v, \delta))$ is continuous. Here, $CF(v, \delta)$ refers to an additive noise intervention, and $\delta \in \mathbb{R}^n$ such that $\delta_{|\mathcal{I}} = 0$.*

The first property is essential when considering causal structures and sensitive features, implying that the distance between an instance and its counterfactual should be zero. The second property ensures the metric exhibits local causality, making it apt for defining adversarial perturbations and facilitating gradient-based computations.

Constructing a metric on $\mathcal{M}$ is challenging due to feature relationships in the dissimilarity function. Using the causal sufficiency principle, which ensures features independence in the exogenous space, allows to define individual similarity functions per each feature. These metrics could be merged by using a product metric in the exogenous space. And finally, the metric extends to the feature space via a push-forward metric i.e., $d_{\mathcal{V}}(v, v') = d_{\mathcal{U}}(g^{-1}(v), g^{-1}(v'))$. However, this idea is insufficient for measuring distances between an instance v and its counterfactual twins. The subsequent example highlights this limitation.

Example 1 *Consider the SCM denoted as $\mathcal{M}$ with the following structural equations:*

$$\mathbb{F} = \begin{cases} V_1 := 2 * (U_1 - 0.5), & U_1 \sim \mathcal{B}(0.5) \\ V_2 := V_1 * U_2, & U_2 \sim \mathcal{B}(0.5) \end{cases}$$

where $\mathcal{B}(p)$ represents the Bernoulli distribution. For $\mathcal{M}$ the reduced-form mappings g and g^{-1} as follows: $g(u_1, u_2) = (2(u_1 - 0.5), 2(u_1 - 0.5) \cdot u_2)$ and $g^{-1}(v_1, v_2) = (0.5 \cdot v_1 + 0.5, \frac{v_2}{v_1})$, respectively. To establish a metric on the exogenous space, we employ the Euclidean norm. Through a pull-back procedure, we define the metric on the exogenous space as $d_{\mathcal{V}}(v_1, v_2) = \|g^{-1}(v_1) - g^{-1}(v_2)\|_2$. The feature space comprises the elements $\mathcal{V} = \{(-1, 0), (-1, -1), (1, 0), (1, 1)\}$. Taking V_2 as a sensitive attribute, it encompasses values of -1, 0, and 1. To fulfill property (i), we must evaluate $d_{\mathcal{V}}(v, \ddot{v}_s)$. Let's examine the case where $v = (1, 1)$ and its counterfactual twin $\ddot{v}_{-1} = (1, -1)$. In this case, it's clear that the point $(1, -1)$ falls outside the feature space, indicating that the use of the pull-back approach for fair metric construction does not consistently yield viable outcomes.

To address the issue encountered in the previous example, we embed the feature space into a semi-latent space introduced by Ehyaei et al. (2023b) to better account for counterfactual counterparts and interventions in the presence of sensitive attributes.

Definition 3 (Semi-latent Space Ehyaei et al. (2023b)) *Consider SCM $\mathcal{M}$ with sensitive features indexed by $\mathcal{I}$. We define the semi-latent space $\mathcal{Q}$ as a combination of observed sensitive features $\mathbf{V}_i$ with distribution $\mathbb{P}_{\mathbf{V}_i}$ where $i \in \mathcal{I}$, and latent variables $\mathbf{U}_j$ for other features with distribution $\mathbb{P}_{\mathbf{U}_j}$. Let $v = (v_1, v_2, \ldots, v_n)$ be an instance in the observed space and $u = (u_1, u_2, \ldots, u_n) = g^{-1}(v)$ be the corresponding instance in the latent space. The mapping $\varphi : \mathcal{V} \to \mathcal{Q}$ transforms v to the semi-latent space $q = (q_1, q_2, \ldots, q_n) = \varphi(v)$, where q_i is defined as follows:*

$$q_i := \begin{cases} v_i & i \in \mathcal{I} \\ u_i & i \notin \mathcal{I} \end{cases}$$

The inverse function $v = \varphi^{-1}(q)$ is determined as follows:

$$v_i := \begin{cases} q_i & i \in \mathcal{I} \\ f_i(v_{\boldsymbol{pa}(i)}) + q_i & i \notin \mathcal{I} \end{cases}$$

The metric construction in the semi-latent space is simpler compared to the feature space due to the independence of its components. This independence arises from the sufficiency assumption for $\mathbf{U}_i$ and the intervention assumption for $X_{\mathcal{I}}$ in the structural causal model. Let $(\mathcal{Q}_i, d_i)$ represent the metric space for the semi-latent space. We can define the dissimilarity function for all Q_i using a product metric, similar to the Euclidean example i.e., $d(x, y) =$

[1] The concept of a causal metric does not adhere to its rigorous mathematical definition, as it evidently lacks the positivity property inherent in distance metrics.

$\sqrt{\sum_{i=1}^{n} d_i(x_i, y_i)^2}$. We aim to ascertain the specific formulations of the causal fair metric, as delineated in Definition 2. Before that, we consider the decomposition of semi-latent space $\mathcal{Q} = \mathcal{S} \times \mathcal{X}$ to sensitive subspace $\mathcal{S}$ and non-sensitive subspace of exogenous space that is denoted by $\mathcal{X}$.

Proposition 1 *Let $\mathrm{d} : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ be a causal fair metric, then d can be written as a form:*

$$d(v, v') = d_{\mathcal{X}}(\varphi_{\mathcal{X}}(v), \varphi_{\mathcal{X}}(v')) \tag{1}$$

where $\varphi_{\mathcal{X}}(v) = P_{\mathcal{X}}(\varphi(v))$, φ is the mapping from feature space to semi-latent space, $P_{\mathcal{X}}$ is a projection on the non-sensitive subspace of exogenous space, and $\mathrm{d}_{\mathcal{X}}$ represents the metric defined on the non-sensitive subspace of $\mathcal{Q}$, which exhibits continuity along its diagonal in relation to each of its components.

In the scenario where the semi-latent space metric is defined in terms of an inner product, Proposition 1 results a widely recognized form, which is the kernelized version of the Mahalanobis distance:

$$d(v, v') = \langle (\varphi(v) - \varphi(v')), \Sigma(\varphi(v) - \varphi(v')) \rangle \tag{2}$$

where, Σ is the projection matrix on non-sensitive exogenous space.

4 Protected Causal Perturbation

An adversarial perturbation ball, often called an attack ball, is a critical concept in machine learning. It defines a geometric region in the input space, representing potential data modifications that remain classified in the same category by the model. This concept assesses the model's vulnerability to input changes, particularly in adversarial attack scenarios, where perturbations can deceive the model. Metrics play a vital role in quantifying these perturbations by measuring the distances between original and modified data points. To extend this concept further, we apply fair causal metrics to define adversarial perturbations.

Definition 4 (Protected Causal Perturbation Ball (PCP)) *Consider an SCM $\mathcal{M}$ that includes sensitive attributes, and let d represent its causal fair metric. We define the PCP ball with radius Δ for an instance v as follows:*

$$B^{\textit{PCP}}_{\Delta}(v) = \{v' \in \mathcal{V} : d(v, v') \leq \Delta\} \tag{3}$$

where Δ is a non-negative real number.

We will study the perturbation ball's shape alterations upon introducing causal structures and sensitive feature protection. Figure 1 illustrates the counterfactual ball's evolution with these properties. Consider the closed ball $B^{\mathcal{X}}_{\Delta}$ in space $\mathcal{X}$, defined by: $B^{\mathcal{X}}_{\Delta}(x) = \{x' \in \mathcal{X} : d_{\mathcal{X}}(x, x') \leq \Delta\}$. Equation 3 yields a straightforward expression: $B^{\mathbf{PCP}}_{\Delta}(v) = \varphi^{-1}_{\mathcal{X}}(B^{\mathcal{X}}_{\Delta}(\varphi(v)))$. However, due to the non-bijective nature of $\varphi_{\mathcal{X}}$ in the presence of sensitive features, $B^{\mathbf{PCP}}_{\Delta}$ and $B^{\mathcal{X}}_{\Delta}$ are not isomorphic. Let $B^{\mathbf{CP}}_{\Delta}$ be the subset of $B^{\mathbf{PCP}}_{\Delta}$ that solely captures the causal structure, excluding the sensitive protected property: $\{v' \in \mathcal{V} : P^{\perp}_{\mathcal{X}}(\varphi(v)) = P^{\perp}_{\mathcal{X}}(\varphi(v')) \ \wedge \ \varphi_{\mathcal{X}}(v') \in B^{\mathcal{X}}_{\Delta}(\varphi_{\mathcal{X}}(v)\}$ It can be observed that $B^{\mathbf{CP}}_{\Delta}$ is isomorphic to $B^{\mathcal{X}}_{\Delta}$. We can now demonstrate that $B^{\mathbf{PCP}}_{\Delta}$ is constructed as a union of causal captured balls over instances twinned with v.

Proposition 2 *Let $B^{\textit{PCP}}_{\Delta}(v)$ represent the PCP ball around the instance v with a radius of Δ. It can be decomposed as:*

$$B^{\textit{PCP}}_{\Delta}(v) = \bigcup_{s \in \mathcal{S}} B^{\textit{CP}}_{\Delta}(\ddot{v}_s), \tag{4}$$

where $\mathcal{S}$ represents the level set of sensitive features (which may be continuous or discrete). $B^{\textit{PCP}}_{\Delta}$ exhibits invariance under twins, meaning that for all $s \in \mathcal{S}$, we have $B^{\textit{PCP}}_{\Delta}(v) = B^{\textit{PCP}}_{\Delta}(\ddot{v}_s)$.

The PCP definition, along with the causal fair metric property, captures the counterfactual proximity definition. The subsequent lemma demonstrates that a PCP with a diameter of 0 represents the set of twins.

Proposition 3 *Let $\mathbf{S}$ denote the protected features, and let d be the causal fair metric. The set of counterfactual twins corresponds to the PCP with a zero radius i.e., $\ddot{\mathbb{V}} = \lim_{\Delta \to 0} B^{\textit{PCP}}_{\Delta}(v)$.*

5 Fair Metric Learning

From Equation 1, a fair metric can be formulated using structural equations and a metric for non-sensitive exogenous variables. Practical implementation requires deriving the metric from data, addressing unknowns: sensitive features, embedding function φ, and the metric for non-sensitive exogenous features. Assuming known sensitive features and domain expert-provided dissimilarity functions for independent exogenous components. Under this assumption, if we have knowledge of the functional structures, we can construct φ and consequently derive a causal fair metric. Notably, the essence of the causal fair metric revolves around counterfactuals. Therefore, a central question arises: Is it feasible to estimate counterfactuals from observational data? The below example that is adapted from chapter 6.19 in Peters et al. (2017) investigate the possibility of this idea.

Example 2 *Let $\mathcal{M}_A$ and $\mathcal{M}_B$ be two SCM with below structural equations respectively:*

$$\begin{cases} V_1 := U_1, & \text{in} \quad \{\mathcal{M}_A, \mathcal{M}_B\} \\ V_2 := V_1(1 - U_2), & \text{in} \quad \{\mathcal{M}_A, \mathcal{M}_B\} \\ V_3 := \mathbb{I}_{V_1 \neq V_2}(\mathbb{I}_{U_3 > 0} V_1 + \mathbb{I}_{U_3 = 0} V_2) + \mathbb{I}_{V_1 = V_2} U_3 & \text{in} \quad \mathcal{M}_A \\ V_3 := \mathbb{I}_{V_1 \neq V_2}(\mathbb{I}_{U_3 > 0} V_1 + \mathbb{I}_{U_3 = 0} V_2) + \mathbb{I}_{V_1 = V_2}(N - U_3) & \text{in} \quad \mathcal{M}_B \end{cases}$$

where, U_1 and U_2 have a Bernoulli distribution with a 0.5 probability, and U_3 has a uniform distribution spanning from 0 to a constant value N. Consider the instance $v = (1, 0, 0)$, with V_1 denoted as the sensitive feature. The counterfactuals for v with respect to $\mathcal{M}_A$ and $\mathcal{M}_A$ are $(0, 0, 0)$ and $(0, 0, N)$, respectively.

Both SCMs share identical causal graphs, observational distributions, and intervention distributions for all possible interventions. Consequently, there exist no randomized trials or observational data capable of discerning between $\mathcal{M}_A$ and $\mathcal{M}_B$. Hence, when our interest lies in counterfactual statements, additional assumptions become imperative. Therefore, Example 2 establishes the following proposition.

Proposition 4 (Lack of Guarantee for Metric Estimation) *If the set of descendants of intervened variables is non-empty, estimating a causal fair metric, from observational data or with a causal graph, necessitates knowledge of the true structural equations, irrespective of data quantity or type.*

Proposition 4 posits that, absent prior knowledge of the structural causal model, data-driven metric learning becomes infeasible. Given that SCM knowledge often remains elusive in practical applications, an alternative approach emerges: direct estimation of the causal-fair metric from distance-tagged data. Various approaches exist for metric learning, with categorizations including spectral, probabilistic, and deep metric learning Ghojogh et al. (2022). Spectral methods focus on eigenvalue decomposition, adopting a geometric perspective to represent the manifold in a reduced-dimensional subspace. While, probabilistic approach assumes the presence of a low-dimensional latent variable upon which the high-dimensional variable is dependent, necessitating the inference and discovery of this latent variable.

Both spectral and probabilistic metric learning techniques employ the generalized Mahalanobis distance, denoted as Equation 2, with a predetermined kernel such as the Gaussian kernel or a kernel that is learned, aiming to optimize the dissimilarity matrix Ghojogh et al. (2022). Conversely, deep metric learning employs neural networks to estimate the embedding function. The network's objective is to minimize distances between similar points and maximize distances between dissimilar ones.

The deep metric methodology is consistent with Proposition 1, which posits the existence of an embedding $\varphi_{\mathcal{X}} : \mathbb{R}^n \to \mathbb{R}^k$. The causal fair metric is represented as $d_\varphi(v, w) = d_{\mathcal{X}}(\varphi_{\mathcal{X}}(v), \varphi_{\mathcal{X}}(w))$, where k denotes the dimensionality of the non-sensitive exogenous space. This formulation yields three primary key insights: the dimensionality of the embedding space, the assurance of coordinate independence in embedding space, and insights regarding $d_{\mathcal{X}}$. These insights support the development of specialized deep learning methods for causal fair metric learning.

We examine feed-forward neural networks of depth $L \geq 1$, defined by layer widths $d_1, \ldots, d_L$ where $d_0 = n$, $d_L = k$, and possess element-wise activation functions $\sigma_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_i}$. The network's transformation is articulated as:

$$\varphi_{\mathbf{W}}(v) = \sigma_L(\mathbf{W}_L \times \sigma_{L-1}(\mathbf{W}_{L-1} \times \cdots \sigma_1(\mathbf{W}_1 \times v) \cdots)) \tag{5}$$

Consequently, the causal fair metric can be expressed as $d(v, w) = d_{\mathcal{X}}(\varphi_{\mathbf{W}}(v), \varphi_{\mathbf{W}}(w))$, where $\mathbf{W} = (\mathbf{W}_1, \ldots, \mathbf{W}_L)$ denotes a tuple of matrices. Each matrix $\mathbf{W}_i \in \mathbb{R}^{d_i \times d_{i-1}}$, and $d_{\mathcal{X}}$ is a known metric. This matrix tuple family is symbolized as $\mathcal{W}$, thus allowing the representation of the family of non-linear functions as $\Phi = \{\varphi_{\mathbf{W}} : \mathbf{W} \in \mathcal{W}\}$.

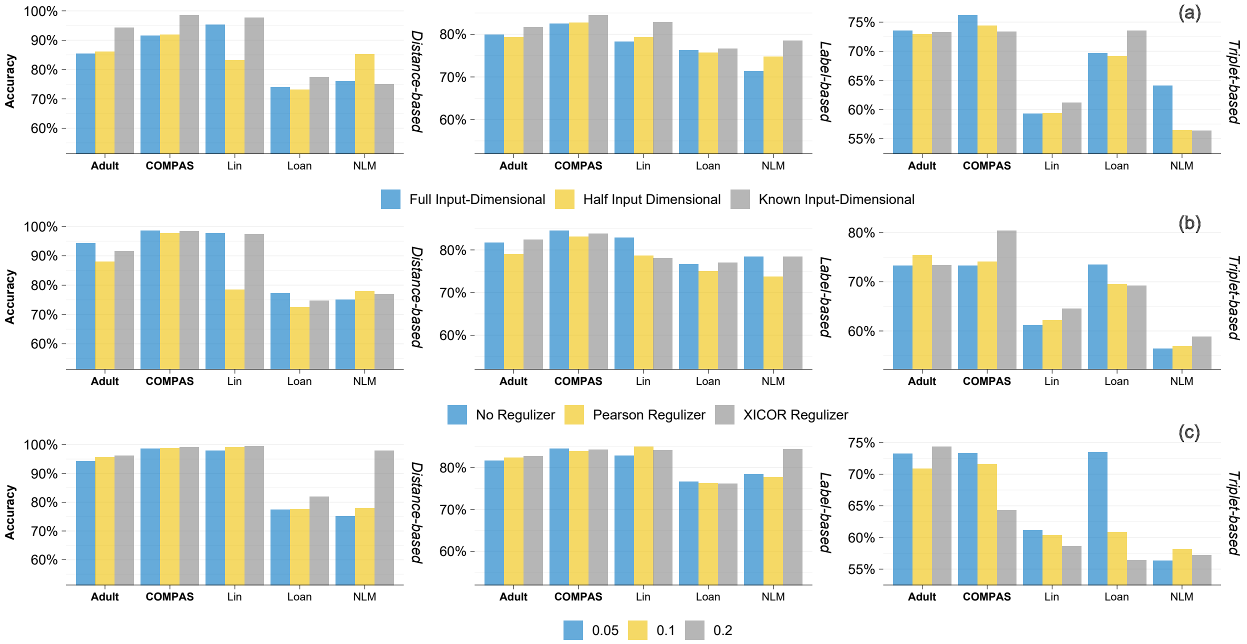


Figure 2: This figure illustrates the impact of the causal metric assumptions on the accuracy of deep metric models: (a) The accuracy performance, when comparing methods based on knowledge of embedding layer sizes and embedding space metric, indicates that, with the exception of the Triplet-based scenario, there is an enhancement in prediction accuracy. (b) It is evident that in simpler models, the network effectively learns the embedding space properties. However, in situations with imprecise learning data about the metric, such as the Triplet-based scenario or non-linear SCM, the incorporation of decorrelation methods enhances accuracy. (c) The stability of our methods is evident as the radius size increases, with notable improvements in accuracy for larger radii in distance-based and label-based scenarios. An additional performance metric is provided in the Appendix.

To evaluate the causal fair metric's influence on deep learning, indicators are needed for convergence under varied conditions. Kozdoba and Mannor (2021) adopted the approach from Bartlett et al. (2017) in accordance with metric learning principles. Their results are based on the following norm definitions: The spectral norm of a matrix $W \in \mathbb{R}^{s\times t}$ is denoted as $\|W\|$. Additionally, $\|W\|_{2,1}$ is introduced as the sum of the ℓ_2 norms of each column in matrix W, where $W_{.,i}$ represents the i-th column of the matrix.

Proposition 5 (Kozdoba and Mannor (2021)) *Consider a feed-forward network with L layers described in Equation 5. Assuming that activation functions ρ_i are λ_i-Lipschitz, and the feature space $\mathcal{V}$ is bounded with $\|v\|_2 \leq B$ for all $v \in \mathcal{V}$, the Rademacher complexity of Φ for a family of matrix tuples $\mathcal{W}$ is bounded as follows:*

$$\mathcal{R}(\Phi) \leq \bar{O}\left(\frac{1}{\sqrt{n}}B^2\left(\prod_{i=1}^{n}\lambda_i\|\mathcal{W}_i\|\right)^2\left(\sum_{i=1}^{L}\frac{\|\mathcal{W}_i\|_{2,1}^{\frac{2}{3}}}{\|\mathcal{W}_i\|^{\frac{2}{3}}}\right)^{\frac{3}{2}}\right)$$

Here, $\|\mathcal{W}_i\|$ represents the supremum norm over W in $\mathcal{W}$ for W_i, and $\|\mathcal{W}_i\|_{2,1}$ is the supremum over W in $\mathcal{W}$ for W_i with respect to the $\ell_{2,1}$ norm.

The aforementioned proposition asserts that deep metric learning can discern embeddings irrespective of dimension or metric. However, the numerical analysis section demonstrates how causal fair metric assumptions enhance estimations compared to general metric learning methods.

6 Numerical Experiments

In this section, we seek empirical validation for the metric learning techniques presented in § 5. We contrast our enhanced deep learning methods, integrating causal structure and sensitive information, with standard deep metric learning. To this end, we employ deep learning for the estimation of the embedding function and adopt the Siamese metric learning as the baseline. This can be stratified into three distinct scenarios relevant to the acquisition of a causal fair metric:

- **Distance-based**: Utilizing distance-tagged triplets (v_i, v'_i, d_i), where d_i indicates the non-negative real number representing $d(v_i, v'_i)$. We apply the Huber function for learning loss:

$$\ell(v_i, v'_i, d_i) = \begin{cases} \frac{1}{2}(d_i - d(\varphi(v_i), \varphi(v'_i)))^2 & |d_i - d(\varphi(v_i), \varphi(v'_i))| \leq \delta \\ \delta(|d_i - d(\varphi(v_i), \varphi(v'_i))| - \frac{1}{2}\delta) & \text{otherwise} \end{cases}$$

- **Label-based**: Utilizing a Siamese network Chicco (2021) with a contrastive loss for triplets (v_i, v'_i, y_i), where $y_i \in \{0, 1\}$ signifies proximity between points, the loss function $\ell(v_i, v'_i, y_i)$ is $(1 - y_i)d(\varphi(v_i), \varphi(v'_i)) + y_i[-d(\varphi(v_i), \varphi(v'_i)) + m]_+$. where $m > 0$ is the marginal and $[.]_+ := \max(., 0)$ is standard Hinge loss.
- **Triplet-based**: In this approach, tuples $(v_i^1, v_i^2, v_i^3, y_i)$ are considered, where y_i denotes the closeness of v_i^1 to v_i^2 compared to v_i^1 and v_i^3. Embedding is trained using a Siamese network with the triplet loss function $\ell(v_i^1, v_i^2, v_i^3, y_i) = [d(\varphi(v_i^1), \varphi(v_i^2)) - d(\varphi(v_i^1), \varphi(v_i^3)) + m]_+$.

For the tabular experimental data's embedding network, we employ a feed-forward network with 100-node layers using *PReLU* activation. We consider three embedding layer dimensions: known dimension, half of the input, and equivalent to the input. We probe the network's depth under two conditions: 5 and 14 hidden layers, and assess the effects of a known metric in the exogenous space by contrasting cases with known and unknown metrics.

We examine embedding coordinate independence by incorporating a decorrelation loss function Patil and Purcell (2022), based on the *Frobenius* norm of the difference identity and Pearson correlation matrix. Furthermore, instead of Pearson correlation, we explore the effect of non-parametric correlation by using *XIcor* Chatterjee (2021) on training performance.

In numerical experiments, a key challenge is the lack of datasets reflecting causal structure in metric learning. Thus, for real-world datasets like Adult Kohavi and Becker (1996) and COMPAS Washington (2018), we first fit the causal structure like as Nabi and Shpitser (2018) paper. We also examine synthetic datasets for Linear (LIN) and Non-linear (NLM) SCMs, alongside a semi-synthetic Loan dataset following Karimi et al. (2020). For each SCM, three metric learning data scenarios are generated using the inherent SCM structure. Using the PCP ball with radii $\Delta = 0.05, 0.1$, and 0.2 for contrastive label creation. We generate 10,000 samples and evaluate deep metric learning over 100 iterations with different random seeds.

To evaluate prediction proximity label performance, we use classifiers metrics such as accuracy (*Acc*), Matthews correlation coefficient (*MCC*), false-negative (*FN*), and false-positive (*FP*) rates. For assessing embedding kernel learning, we employ root mean square error (*RMSE*) and mean absolute error (*MAE*). In both label-based and triplet-based contexts, we apply continuous metrics for kernel learning. In distance-based contexts, we integrate label predictions using radius Δ for label-based analysis, ensuring consistent performance indicators across scenarios. The codes for the numerical analysis can be accessed on GitHub.

Main Results Our simulation confirms that despite Proposition 4 ensuring deep learning convergence for various layer sizes without knowledge of the embedding space metric, performance in distance and label-based scenarios as shown in Figure 2 is enhanced by understanding the metric and embedding space dimensionality (refer to Table 2 and Figure 4 in the appendix).

Figure 2 illustrates that in the distance-based and label-based scenarios, embedding learning is generally effective. Consequently, in these two cases, the utilization of the decorrelation loss function does not yield noticeable improvements. However, in the triplet scenario, where some partial information about the metric is available, the use of this loss function leads to enhanced results (See Table 3 and Figure 6).

Figure 2 shows embedding metric performance declining with increased radius size, while label-based algorithms effectively discern data point proximity (refer to Figure 5). Table 4 guides the optimization of embedding network layers. A five-layer network optimally fits both label-based and distance-based scenarios, while triplet-based scenarios benefit from deeper configurations.

In short, upon analyzing various learning methodologies, it becomes evident that when precise distance-based data is accessible, distance-based metric learning emerges as the optimal strategy. However, practical implementations often challenge this ideal. Under such constraints, the label-based technique serves as a compelling alternative for metric approximation. Enhanced accuracy in estimation is observed when the embedding space's dimensions and metric information are integrated, as corroborated by Table 1 and Figure 3. A notable merit of the label-based method is its

		Real-World Data								**Synthetic Data**											
		Adult				**COMPAS**				**Lin**				**Loan**				**NLM**			
Δ	Loss Function	Acc↑	FN↓	MAE ↓	RMSE ↓	Acc	FN	MAE	RMSE	Acc	FN	MAE	RMSE	Acc	FN	MAE	RMSE	Acc	FN	MAE	RMSE
0.05	Distance-based	**0.943**	0.045	**0.014**	**0.021**	**0.986**	0.011	**0.003**	**0.005**	**0.978**	0.016	**0.004**	**0.006**	**0.773**	0.121	**0.043**	**0.061**	0.751	0.150	**0.037**	**0.053**
	Label-based	0.817	**0.001**	0.047	0.064	0.845	**0.000**	0.050	0.066	0.828	**0.000**	0.048	0.064	0.766	**0.020**	0.048	0.070	**0.785**	**0.000**	0.051	0.072
	Triplet-based	0.733	0.133	0.087	0.115	0.733	0.119	0.088	0.116	0.612	0.210	0.086	0.114	0.735	0.136	0.088	0.116	0.563	0.227	0.169	0.208
0.10	Distance-based	**0.956**	0.032	**0.024**	**0.039**	**0.988**	0.008	**0.005**	**0.008**	**0.990**	0.007	**0.004**	**0.005**	**0.775**	0.121	**0.087**	**0.127**	**0.779**	0.144	**0.064**	**0.096**
	Label-based	0.824	**0.000**	0.092	0.126	0.840	**0.000**	0.100	0.134	0.850	**0.000**	0.098	0.131	0.762	**0.033**	0.099	0.139	0.777	**0.000**	0.104	0.144
	Triplet-based	0.709	0.145	0.174	0.230	0.716	0.124	0.175	0.231	0.603	0.191	0.175	0.231	0.608	0.196	0.175	0.230	0.582	0.208	0.174	0.212
0.20	Distance-based	**0.961**	0.028	**0.048**	**0.076**	**0.992**	0.006	**0.010**	**0.016**	**0.994**	0.005	**0.005**	**0.008**	**0.819**	0.118	**0.154**	**0.224**	**0.978**	0.016	**0.017**	**0.028**
	Label-based	0.827	**0.000**	0.184	0.254	0.843	**0.000**	0.201	0.268	0.841	**0.000**	0.186	0.248	0.762	**0.037**	0.197	0.275	0.844	**0.001**	0.197	0.260
	Triplet-based	0.744	0.125	0.348	0.460	0.643	0.178	0.350	0.462	0.587	0.213	0.352	0.465	0.564	0.223	0.355	0.468	0.572	0.213	0.302	0.406

Table 1: The table presents results from a numerical experiment comparing various learning scenarios. We evaluate these scenarios based on their accuracy (**Acc** - higher values indicate better performance), false negative error (**FN** - lower values are better), root mean square error (**RMSE** - lower values are better), and mean average error (**MAE** - lower values are better). The optimal learning scenario for each dataset and perturbation radius is highlighted in bold. Notably, we exclude the correlation function and employ an embedding network with 5 hidden layers, considering known embedding dimensions and space metrics.

diminished false negative rate relative to other strategies, suggesting its efficacy in approximating the genuine metric. In contexts demanding a fair metric, such as robust learning, this reduced false negative rate ensures adherence to robustness criteria, fostering a sturdier model. Consequently, the label-based methodology outperforms its counterparts. In scenarios where label data procurement is unattainable, the triplet approach, complemented by a decorrelation loss function and deeper networks, facilitates the inference of the embedding function.

7 Discussion and Future Work

In this study, we present a methodology for constructing and learning a fair metric that incorporates causal structures while safeguarding sensitive attributes. We delineate the advantages and disadvantages of our proposed method relative to other baseline deep metric learning approaches. A primary challenge we encountered is the absence of real datasets for metric learning that originate from a causal structure. Given the importance of fair metrics in designing equitable systems, we anticipate the collection of such data for future research endeavors.

In our study, we did not delve into the specifics of the embedding network architecture. Instead, we employed a straightforward feed-forward network, which is better suited for our tabular data. This choice aligns with Proposition 4, assisting in discerning the impacts of various assumptions.

One critique of our methodology is its similarity to numerous metric learning approaches that lack theoretical guarantees regarding estimator performance and potentially grapple with the issue of local minima. In forthcoming work, by imposing constraints on the structure of the causal fair metric, we aim to introduce metric learning methodologies underpinned by explicit convergence theorems.

The notion of protected causal perturbation is applicable in analyzing fairness and robustness in fields assuming data emerges from a causal framework, including algorithmic recourse, causal bandits, causal reinforcement learning, and other causal ML models.

References

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Yahav Bechavod, Christopher Jung, and Steven Z Wu. Metric-free individual fairness in online learning. *Advances in neural information processing systems*, 33:11214–11225, 2020.

Sourav Chatterjee. A new coefficient of correlation. *Journal of the American Statistical Association*, 116(536): 2009–2022, 2021.

Davide Chicco. Siamese neural networks: An overview. *Artificial neural networks*, pages 73–94, 2021.

Emir Demirović and Peter J Stuckey. Optimal decision trees for nonlinear metrics. In *Proceedings of the AAAI conference on artificial intelligence*, 2021. Volume 35, Number 5, Pages 3733–3741.

Ricardo Dominguez-Olmedo, Amir H Karimi, and Bernhard Schölkopf. On the adversarial robustness of causal algorithmic recourse. In *International Conference on Machine Learning*, pages 5324–5342. PMLR, 2022.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

Frederick Eberhardt and Richard Scheines. Interventions and causal inference. *Philosophy of Science*, 74(5):981–995, 2007.

Ahmad-Reza Ehyaei, Amir-Hossein Karimi, Bernhard Schölkopf, and Setareh Maghsudi. Robustness implies fairness in causal algorithmic recourse. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 984–1001, 2023a.

Ahmad-Reza Ehyaei, Kiarash Mohammadi, Amir-Hossein Karimi, Samira Samadi, and Golnoosh Farnadi. Causal adversarial perturbations for individual fairness and robustness in heterogeneous data spaces. *arXiv preprint arXiv:2308.08938*, 2023b.

Deena P Francis and Kumudha Raimond. Major advancements in kernel function approximation. *Artificial Intelligence Review*, 54:843–876, 2021.

Benyamin Ghojogh, Ali Ghodsi, Fakhri Karray, and Mark Crowley. Spectral, probabilistic, and deep metric learning: Tutorial and survey. *arXiv preprint arXiv:2201.09267*, 2022.

Benyamin Ghojogh, Mark Crowley, Fakhri Karray, and Ali Ghodsi. *Elements of dimensionality reduction and manifold learning*. Springer Nature, 2023.

Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. Online learning with an unknown fairness metric. *Advances in neural information processing systems*, 31, 2018.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696, 2009.

Christina Ilvento. Metric learning for individual fairness. In *1st Symposium on Foundations of Responsible Computing*, 2020.

Alexander Immer, Christoph Schultheiss, Julia E Vogt, Bernhard Schölkopf, Peter Bühlmann, and Alexander Marx. On the identifiability and estimation of causal location-scale noise models. In *International Conference on Machine Learning*, pages 14316–14332. PMLR, 2023.

Philips George John, Deepak Vijaykeerthy, and Diptikalyan Saha. Verifying individual fairness in machine learning models. In *Conference on Uncertainty in Artificial Intelligence*, pages 749–758. PMLR, 2020.

Amir-Hossein Karimi, Julius Von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *Advances in neural information processing systems*, 33: 265–277, 2020.

Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30, 2017.

Ronny Kohavi and Barry Becker. Uci adult data set. *UCI Meachine Learning Repository*, 5, 1996.

Mark Kozdoba and Shie Mannor. Two regimes of generalization for non-linear metric learning. *OpenReview, ICLR 2022, https://openreview.net/forum?id=zPLQSnfd14w*, 2021.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4069–4079, 2017.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Daniel McNamara, Cheng Soon Ong, and Robert C Williamson. Provably fair representations. *arXiv preprint arXiv:1710.04394*, 2017.

Debarghya Mukherjee, Mikhail Yurochkin, Moulinath Banerjee, and Yuekai Sun. Two simple ways to learn individual fairness metrics from data. In *International Conference on Machine Learning*, pages 7097–7107. PMLR, 2020.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. Volume 32, Number 1.

Conlan Olson. *Algorithmic Fairness, Metric Embedding, and Metric Learning*. PhD thesis, Harvard University, 2022.

Pranita Patil and Kevin Purcell. Decorrelation-based deep learning for bias mitigation. *Future Internet*, 14(4):110, 2022.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Anian Ruoss, Mislav Balunovic, Marc Fischer, and Martin Vechev. Learning certified individually fair representations. *Advances in neural information processing systems*, 33:7584–7596, 2020.

Juan Luis Suárez, Salvador García, and Francisco Herrera. A tutorial on distance metric learning: Mathematical foundations, algorithms, experimental analysis, prospects and challenges. *Neurocomputing*, 425:300–322, 2021.

Anne L Washington. How to argue with an algorithm: Lessons from the compas-propublica debate. *Colo. Tech. LJ*, 17: 131, 2018.

Mikhail Yurochkin, Amanda Bower, and Yuekai Sun. Training individually fair ml models with sensitive subspace robustness. *arXiv preprint arXiv:1907.00020*, 2019.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

Kun Zhang and Aapo Hyvarinen. On the identifiability of the post-nonlinear causal model. *arXiv preprint arXiv:1205.2599*, 2012.

Luwan Zhang, Grace Wahba, and Ming Yuan. Distance shrinkage and euclidean embedding via regularized kernel estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(4):849–867, 2016.

.1 Additional Background

Definition 5 ((Pseudo-) Metric Space) *A metric space (X, d) is defined as a set X accompanied by a non-negative real-valued function $d : X \times X \longrightarrow \mathbb{R}_{\geq 0}$, which is referred to as a metric. This metric function d adheres to the subsequent properties for any $x, y, z \in X$:*

- ***Non-negativity**: $d(x, y) \geq 0$, and $d(x, y) = 0$ if and only if $x = y$.*
- ***Symmetry**: $d(x, y) = d(y, x)$.*
- ***Triangle inequality**: $d(x, z) \leq d(x, y) + d(y, z)$.*

When the positivity condition, i.e., $d(x, y) = 0$ if and only if $x = y$ is relaxed, the d is called pseudometric (or semi-metric).

Definition 6 (The pull-back $\&$ push-forward metric) *let $f : \mathcal{U} \to \mathcal{V}$ be a mapping between the metric spaces $(\mathcal{U}, d_{\mathcal{U}})$ and $(\mathcal{V}, d_{\mathcal{V}})$. The push-forward metric d induced by the function f is defined as:*

$$d(u_1, u_2) = d_{\mathcal{V}}(f(u_1), f(u_2)); \quad u_1, u_2 \in \mathcal{U}$$

Similarly, the pull-back metric on the space $\mathcal{U}$ is defined as:

$$d(v_1, v_2) = d_{\mathcal{U}}(f^{-1}(v_1), f^{-1}(v_2)); \quad v_1, v_2 \in \mathcal{V}$$

These definitions allow us to relate distances in $\mathcal{U}$ and $\mathcal{V}$ via the mapping f and its inverse f^{-1}.

.2 Proofs

Proposition 1.

Let consider a causal fair metric denoted as $d : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$, with an associated embedding $\varphi : \mathcal{V} \to \mathcal{Q}$, mapping from the feature space to a semi-latent space. We define d^* as the pull-back metric of d onto $\mathcal{Q}$:

$$d^*(q_1, q_2) = d(\varphi^{-1}(q_1), \varphi^{-1}(q_2))$$

d^* possesses metric properties, and we aim to elucidate which properties it inherits from Definition 2. We consider a decomposition of $\mathcal{Q}$ into $\mathcal{S} \times \mathcal{X}$, and let $q = \varphi^{-1}(v)$, where $v \in \mathcal{V}$. Utilizing this decomposition, we express q as (s, x). Property (i) of the causal fair metric implies:

$$d(v, \ddot{v}_{s'}) = d^*((s, x), (s', x)) = 0 \qquad \forall s' \in \mathcal{S}$$

This property implies that d^* is invariant with respect to the sensitive part $\mathcal{S}$. To demonstrate this, we assert that for any two points $q_1 = (s_1, x_1)$ and $q_2 = (s_2, x_2)$, along with an arbitrary $s_0 \in \mathcal{S}$, the following equality holds:

$$d^*((s_1, x_1), (s_2, x_2)) = d^*((s_0, x_1), (s_0, x_2))$$

By utilizing the triangle property of d^*, we can establish:

$$\begin{aligned} d^*((s_1, x_1),&(s_2, x_2)) \leq d^*((s_0, x_1), (s_2, x_2)) + \cancelto{0}{d^*((s_1, x_1), (s_0, x_1))} \implies \\ &d^*((s_1, x_1), (s_2, x_2)) \leq d^*((s_0, x_1), (s_2, x_2)) \end{aligned}$$

The distance $d^*((s_1, x_1), (s_0, x_1))$ is zero due to the first property. Similarly, it can be shown that:

$$\begin{aligned} d^*((s_0, x_1),&(s_2, x_2)) \leq d^*((s_1, x_1), (s_2, x_2)) + \cancelto{0}{d^*((s_1, x_1), (s_0, x_1))} \implies \\ &d^*((s_0, x_1), (s_2, x_2)) \leq d^*((s_1, x_1), (s_2, x_2)) \end{aligned}$$

it concludes that:

$$d^*((s_0, x_1), (s_2, x_2)) = d^*((s_1, x_1), (s_2, x_2))$$

similarly we can show:

$$d^*((s_0, x_1), (s_2, x_2)) = d^*((s_0, x_1), (s_0, x_2))$$

This last equation implies that d^* is invariant with respect to the sensitive subspace. If we consider $d_{\mathcal{X}}$ as the induced metric of d^* on the sensitive subspace $\mathcal{X}$, then we can express:

$$d^*((s_1, x_1), (s_2, x_2)) = d_{\mathcal{X}}(x_1, x_2)$$

The second property of Definition 2 can be expressed in a simplified form based on $d_{\mathcal{X}}$. It implies that for every $x \in \mathcal{X}$, the distance $d_{\mathcal{X}}(x, x + \delta)$, where $\delta \in \mathbb{R}^{\dim(\mathcal{X})}$, is continuous with respect to δ. This continuity implies that $d_{\mathcal{X}}$ is continuous along each component on its diagonal, i.e., (x, x).

Finally, if we replace x with $P_{\mathcal{X}}(\varphi(v))$, where $P_{\mathcal{X}}$ is the projection operator onto the subspace $\mathcal{X}$ within $\mathcal{Q}$, we obtain:

$$d(v, w) = d_{\mathcal{X}}(P_{\mathcal{X}}(\varphi(v)), P_{\mathcal{X}}(\varphi(w)))$$

This equation completes the proof.

Proposition 2.

The proof is straightforward when we write out the definitions. Let $\varphi(v) = (s, x)$ represent the embedding of the variable v in the semi-latent space. To begin, we can demonstrate how the semi-latent space enables us to describe the counterfactual of instance v with respect to the hard action $do(\mathbf{S}{:=}s')$ as follows:

$$\varphi^{-1}(\varphi(v) \odot_I s') = \varphi^{-1}((s, x) \odot_I s') = \varphi^{-1}((s', x)) = \mathbf{CF}(v, do(\mathbf{S}{:=}s')) = \ddot{v}_{s'}$$

In above Equation, we use the symbol $v \odot_I \theta$ to represent a masking operator that modifies the values of the entries corresponding to set I in vector v by replacing them with θ. The validity of the last line in Equation .2 is based on the definition of the semi-latent space embedding.

By the definition 4, the $B_{\Delta}^{\textbf{PCP}}(v)$ is equal to:

$$
\begin{aligned}
B^{\mathbf{PCP}}_{\Delta}(v) &= \{v' \in \mathcal{V} : d(v,v') \leq \Delta\} = \{v' \in \mathcal{V} : d_{\mathcal{X}}(P_{\mathcal{X}}(\varphi(v)), P_{\mathcal{X}}(\varphi(v'))) \leq \Delta\} = \\
&\{v' \in \mathcal{V} : d_{\mathcal{X}}(x,x') \leq \Delta\} = \bigcup_{s\in\mathcal{S}} \{v' \in \mathcal{V} : \varphi(v') = (s,x') \wedge d_{\mathcal{X}}(x,x') \leq \Delta\} = \\
&\bigcup_{s\in\mathcal{S}} \{v' \in \mathcal{V} : P^{\perp}_{\mathcal{X}}(\varphi(v')) = s \ \wedge \ d_{\mathcal{X}}(\varphi_{\mathcal{X}}(\ddot{v}_s), \varphi_{\mathcal{X}}(v')) \leq \Delta\} = \\
&\bigcup_{s\in\mathcal{S}} \{v' \in \mathcal{V} : P^{\perp}_{\mathcal{X}}(\varphi(\ddot{v}_s)) = P^{\perp}_{\mathcal{X}}(\varphi(v')) \ \wedge \ \varphi_{\mathcal{X}}(v') \in B^{\mathcal{X}}_{\Delta}(\varphi_{\mathcal{X}}(\ddot{v}_s)\} = \\
&\bigcup_{s\in\mathcal{S}} B^{\mathbf{CP}}_{\Delta}(\ddot{v}_s)
\end{aligned}
$$

The last equation completes the proof.

Proposition 3.

To present the result, we must first prove the following lemma:

Lemma 1 *Let d be a causal metric, and let $d_{\mathcal{X}}$ be the corresponding embedding metric on the non-sensitive part of the exogenous space. For the closed ball $B^{\mathcal{X}}_{\Delta}$, we have:*

$$\lim_{\Delta\to 0} B^{\mathcal{X}}_{\Delta}(x) = x$$

Proof 1 *We establish the aforementioned lemma by means of a proof by contradiction. Let us assume that there exists another point, denoted as $x' \neq x$, within the set* $\lim_{\Delta\to 0} B^{\mathcal{X}}_{\Delta}(x)$*. Consequently, we have $d_{\mathcal{X}}(x,x') = 0$. If we consider $v' = \varphi^{-1}((s,x'))$, then for v', we have $d(v,v') = 0$, since $v' \notin \{\ddot{v}_s\}$. However, this contradicts the property inherent to one of the causal fair metrics.*

By utilizing above lemma and Proposition 2, we can represent the result as follows:

$$
\begin{aligned}
B^{\mathbf{PCP}}_{0}(v) &= \lim_{\Delta\to 0} B^{\mathbf{PCP}}_{\Delta}(v) = \lim_{\Delta\to 0} \bigcup_{s\in\mathcal{S}} B^{\mathbf{CP}}_{\Delta}(\ddot{v}_s) = \bigcup_{s\in\mathcal{S}} \lim_{\Delta\to 0} B^{\mathbf{CP}}_{\Delta}(\ddot{v}_s) = \\
&\bigcup_{s\in\mathcal{S}} \lim_{\Delta\to 0} \{v' \in \mathcal{V} : \varphi(v') = (s',x') \wedge s' = s \ \wedge \ x' \in B^{\mathcal{X}}_{\Delta}(x)\} = \\
&\bigcup_{s\in\mathcal{S}} \{v' \in \mathcal{V} : \varphi(v') = (s',x') \wedge s' = s \ \wedge \ x' \in \lim_{\Delta\to 0} B^{\mathcal{X}}_{\Delta}(x)\} = \\
&\bigcup_{s\in\mathcal{S}} \{v' \in \mathcal{V} : \varphi(v') = (s,x)\} = \bigcup_{s\in\mathcal{S}} \ddot{v}_s
\end{aligned}
$$

.3 Synthetic Data Models

In the § 6, we detail the structural equations employed to formulate the SCMs for both LIN and NLM models. The protected feature, denoted as $\mathbf{S}$, and the non-sensitive variables represented by $\mathbf{X}_i$ are derived based on the subsequent structural equations:

- linear SCM (LIN):

$$
\mathbb{F} = \begin{cases} S := U_S, & U_S \sim \mathcal{B}(0.5) \\ X_1 := 2S + U_1, & U_1 \sim \mathcal{N}(0,1) \\ X_2 := S - X_1 + U_2, & U_2 \sim \mathcal{N}(0,1) \end{cases}
$$

- Non-linear Model (NLM)

$$
\mathbb{F} = \begin{cases} S := U_S, & U_S \sim \mathcal{B}(0.5) \\ X_1 := 2S^2 + U_1, & U_1 \sim \mathcal{N}(0,1) \\ X_2 := S - X_1^2 + U_2, & U_2 \sim \mathcal{N}(0,1) \end{cases}
$$

Where $\mathcal{B}(p)$ represents Bernoulli random variables characterized by a probability p, and $\mathcal{N}(\mu,\sigma^2)$ denotes normal random variables, which are defined by a mean of μ and a variance of σ^2.

.4 Semi-Synthetic Data Model

This semi-synthetic dataset encompasses gender, age, education, loan amount, duration, income, and saving variables, governed by the following structural equations:

$$\mathbb{F} = \begin{cases} G := U_G, & U_G \sim \mathcal{B}(0.5) \\ A := -35 + U_A, & U_A \sim \mathcal{G}(10, 3.5) \\ E := -0.5 + \left(1 + e^{-\left(-1+0.5G+(1+e^{-0.1A})^{-1}+U_E\right)}\right)^{-1}, & U_E \sim \mathcal{N}(0, 0.25) \\ L := 1 + 0.01(A-5)(5-A) + G + U_L, & U_L \sim \mathcal{N}(0, 4) \\ D := -1 + 0.1A + 2G + L + U_D, & U_D \sim \mathcal{N}(0, 9) \\ I := -4 + 0.1(A+35) + 2G + GE + U_I, & U_I \sim \mathcal{N}(0, 4) \\ S := -4 + 1.5\mathbb{I}_{\{I>0\}}I + U_S, & U_S \sim \mathcal{N}(0, 25) \end{cases}$$

Where $\mathcal{B}$ and $\mathcal{G}$ represent the Bernoulli and Gamma distributions, respectively. G is considered as a sensitive attribute.

.5 Real-World Data

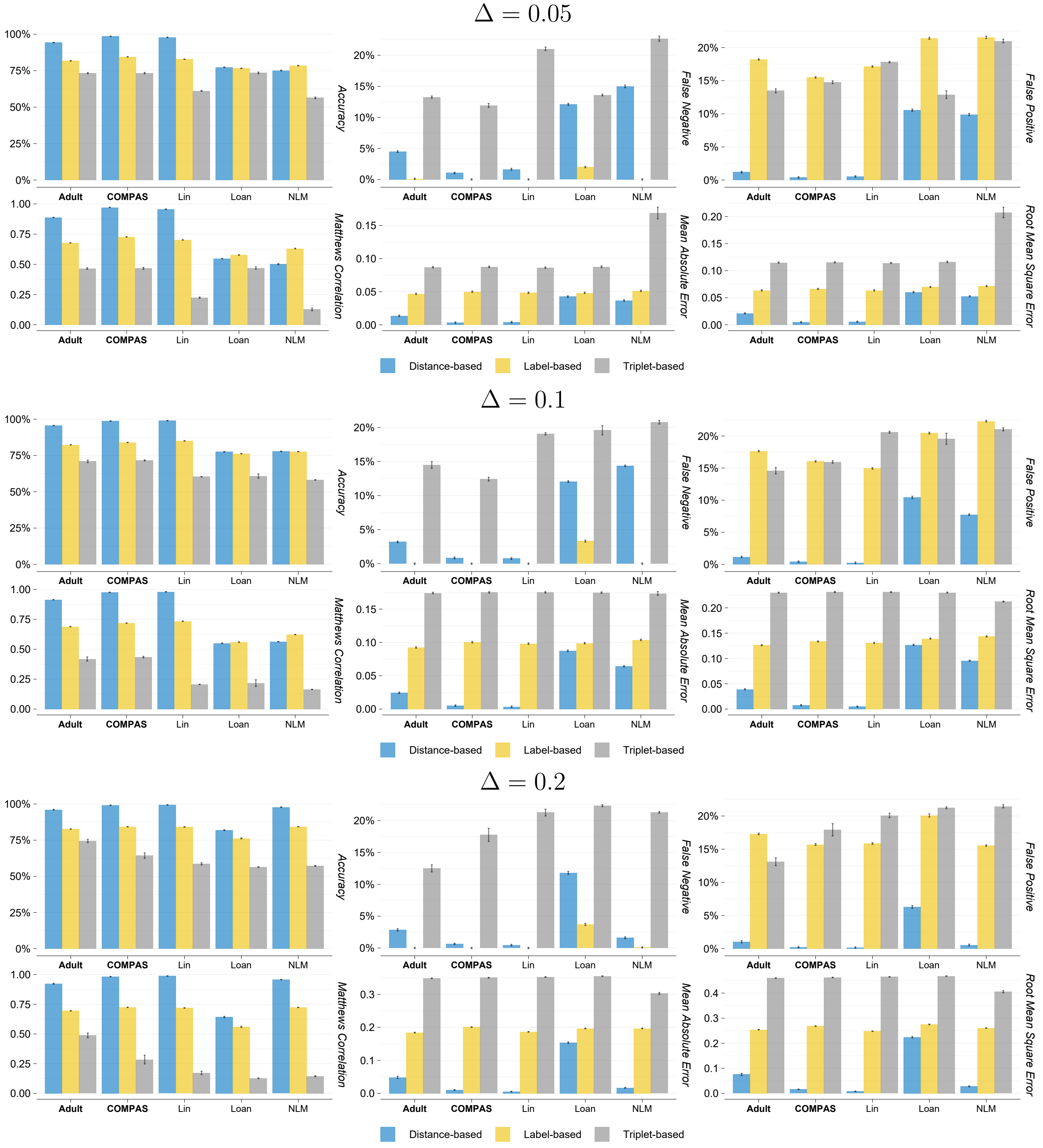
In our study, we employed the Adult Kohavi and Becker (1996) and COMPAS Washington (2018) datasets, constructing an SCM from the causal graph by Nabi and Shpitser (2018). For the Adult dataset, we considered features like **sex**, **age**, and **education-num**, with sex as a sensitive attribute. For COMPAS, features included **age**, **race**, and **priors count**, with sex as the sensitive attribute.

.6 Hyperparameter Tuning

In our experimental setup, we generated 10,000 samples for each SCM model. The data was divided into batches of 1,000, and the learning process spanned 100 epochs. The coefficient of the decorrelation regularizer was set to 0.1. Furthermore, in the contrastive label-based scenario, the margin was set equal to the radius of the experiment, while in the triplet-based scenario, the margin was set to zero to have more sensitivity for metric learning.

.7 Additional Numerical Results

In the subsequent tables and figures, additional numerical analysis results are presented to support the assertions of this study. Their explanations can be found in § 6.



$\Delta = 0.05$

$\Delta = 0.1$

$\Delta = 0.2$

Figure 3: Performance Comparison. This figures illustrate the outcomes of the experiment detailed in Table 1 with different radius size. Error bars are included to represent the variability. Two performance metrics, Matthews correlation coefficient (**MCC** - higher values indicate superior performance) and false-positive rate (**FP** - lower values indicate better performance), are considered for the analysis.

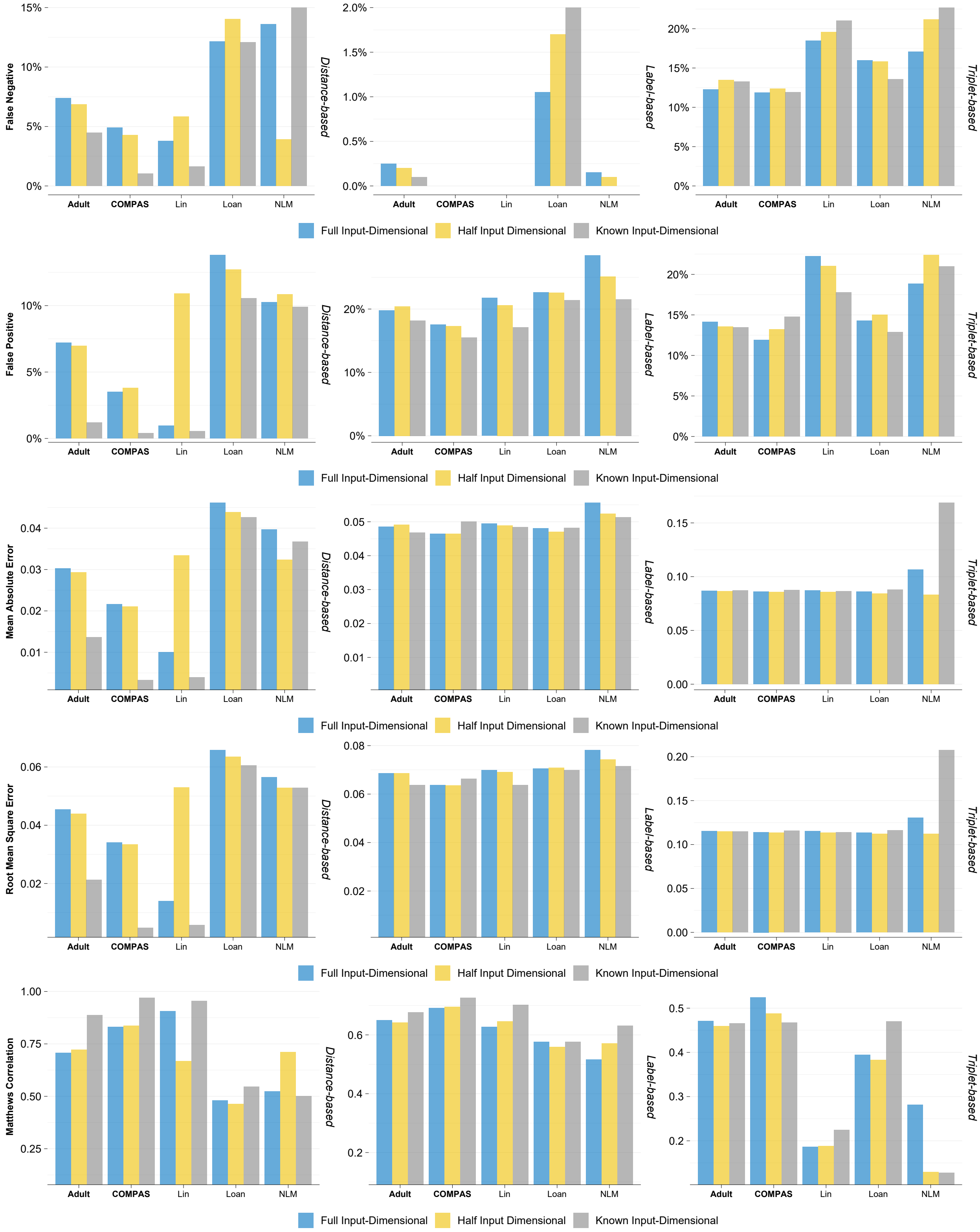


Figure 4: The performance metric of the learning scenario with varying knowledge regarding the embedding layer size.

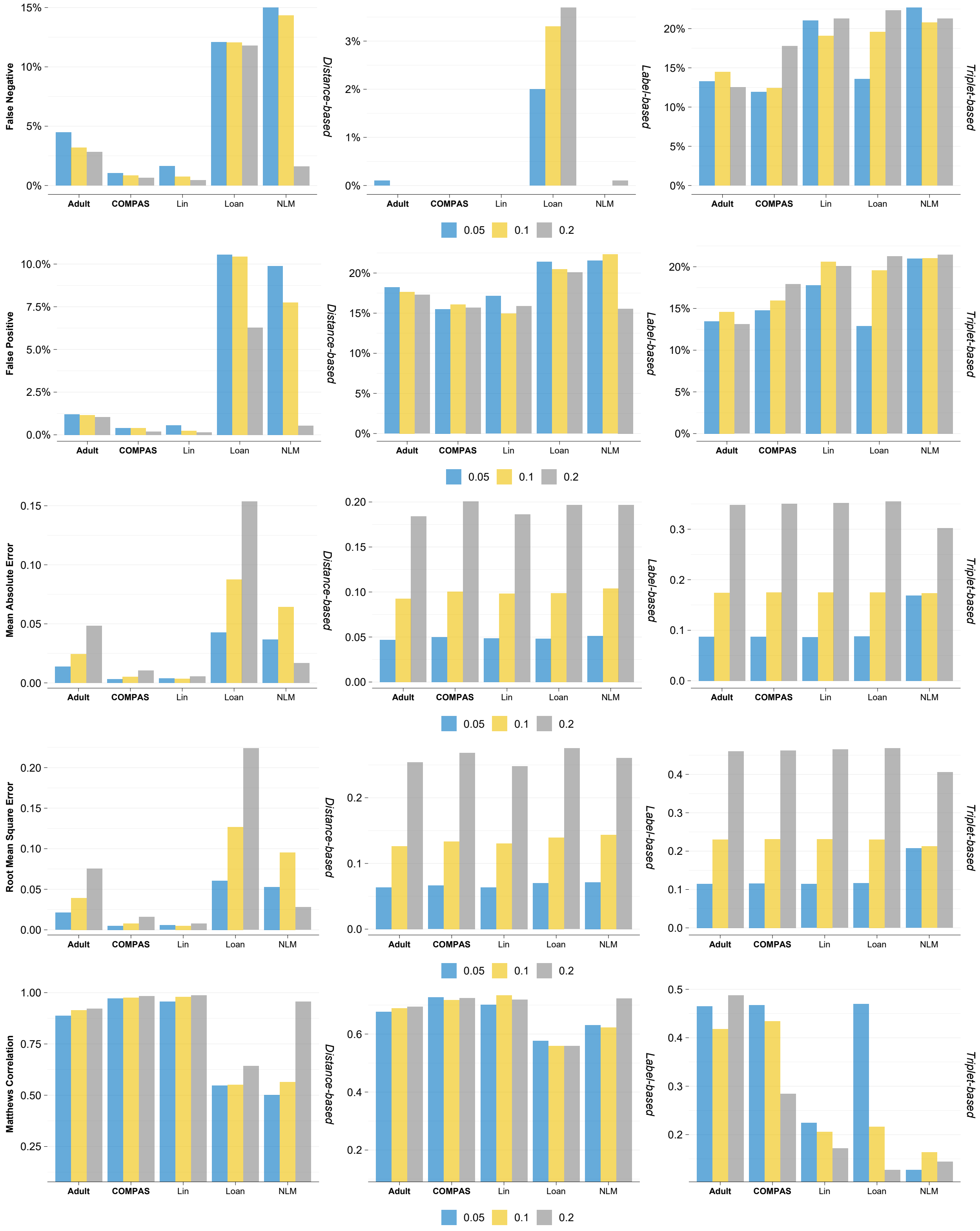


Figure 5: The performance metric of the learning scenarios under varying values of perturbation radius.

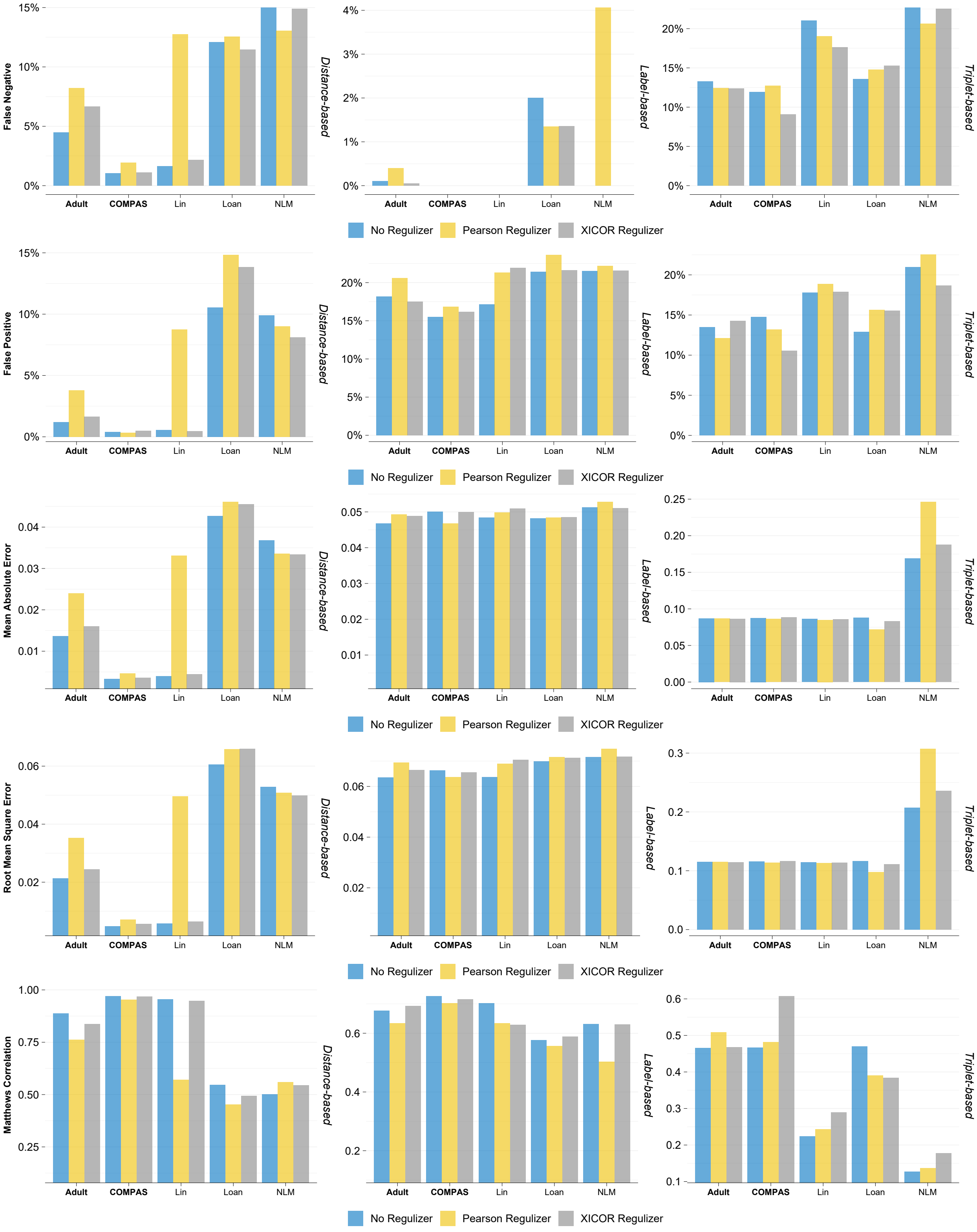


Figure 6: The performance metric of learning scenarios under the influence of different decorrelation regularizers.

		Embedding Layer Dimension		
Loss Function	Performance Metric	Full Input-Dimensional	Half Input Dimensional	Known Input-Dimensional
Distance-based	Accuracy ↑	0.914 ± 0.081	0.854 ± 0.053	0.924 ± 0.091
	False Negative ↓	0.05 ± 0.042	0.062 ± 0.032	0.047 ± 0.053
	False Positive ↓	0.035 ± 0.04	0.085 ± 0.032	0.029 ± 0.039
	Matthews Correlation ↑	0.829 ± 0.161	0.709 ± 0.106	0.849 ± 0.182
	Mean Absolute Error ↓	0.032 ± 0.028	0.069 ± 0.038	0.029 ± 0.038
	Root Mean Square Error ↓	0.048 ± 0.044	0.11 ± 0.061	0.044 ± 0.056
Label-based	Accuracy ↑	0.808 ± 0.038	0.788 ± 0.024	0.819 ± 0.032
	False Negative ↓	0.001 ± 0.003	0.002 ± 0.005	0.005 ± 0.012
	False Positive ↓	0.191 ± 0.037	0.21 ± 0.022	0.176 ± 0.024
	Matthews Correlation ↑	0.665 ± 0.063	0.632 ± 0.041	0.68 ± 0.06
	Mean Absolute Error ↓	0.113 ± 0.061	0.112 ± 0.059	0.113 ± 0.06
	Root Mean Square Error ↓	0.156 ± 0.081	0.158 ± 0.083	0.153 ± 0.081
Triplet-based	Accuracy ↑	0.686 ± 0.088	0.673 ± 0.096	0.642 ± 0.065
	False Negative ↓	0.155 ± 0.044	0.161 ± 0.048	0.181 ± 0.038
	False Positive ↓	0.159 ± 0.045	0.165 ± 0.049	0.177 ± 0.029
	Matthews Correlation ↑	0.372 ± 0.175	0.346 ± 0.193	0.284 ± 0.13
	Mean Absolute Error ↓	0.202 ± 0.107	0.203 ± 0.11	0.206 ± 0.104
	Root Mean Square Error ↓	0.267 ± 0.142	0.268 ± 0.145	0.271 ± 0.138

Table 2: The table displays the average performance metrics for comparing scenarios with knowledge of embedding dimensions and their corresponding metrics against scenarios with no knowledge of the embedding space, with input dimension either at full or half capacity. Green cell highlights denote superior performance, while smaller values indicate the standard deviation of the estimations.

		Decorrelation Regularizer Function		
Loss Function	Performance Metric	-	Pearson	XICOR
Distance-based	Accuracy ↑	0.924 ±0.091	0.89 ±0.107	0.918 ±0.095
	False Negative ↓	0.047 ±0.053	0.066 ±0.06	0.05 ±0.052
	False Positive ↓	0.029 ±0.039	0.044 ±0.05	0.032 ±0.046
	Matthews Correlation ↑	0.849 ±0.182	0.78 ±0.212	0.837 ±0.189
	Mean Absolute Error ↓	0.029 ±0.038	0.04 ±0.049	0.032 ±0.042
	Root Mean Square Error ↓	0.044 ±0.056	0.059 ±0.072	0.047 ±0.062
Label-based	Accuracy ↑	0.819 ±0.032	0.809 ±0.043	0.812 ±0.033
	False Negative ↓	0.005 ±0.012	0.007 ±0.014	0.005 ±0.012
	False Positive ↓	0.176 ±0.024	0.184 ±0.033	0.183 ±0.026
	Matthews Correlation ↑	0.68 ±0.06	0.659 ±0.083	0.67 ±0.063
	Mean Absolute Error ↓	0.113 ±0.06	0.114 ±0.061	0.114 ±0.061
	Root Mean Square Error ↓	0.153 ±0.081	0.156 ±0.083	0.156 ±0.083
Triplet-based	Accuracy ↑	0.642 ±0.065	0.666 ±0.068	0.651 ±0.069
	False Negative ↓	0.181 ±0.038	0.167 ±0.036	0.176 ±0.041
	False Positive ↓	0.177 ±0.029	0.167 ±0.033	0.173 ±0.03
	Matthews Correlation ↑	0.284 ±0.13	0.332 ±0.137	0.302 ±0.138
	Mean Absolute Error ↓	0.206 ±0.104	0.209 ±0.104	0.207 ±0.103
	Root Mean Square Error ↓	0.271 ±0.138	0.275 ±0.138	0.271 ±0.136

Table 3: The table displays average performance metrics for various scenarios, considering the presence of different decorrelation policies. Green cells highlight the best performance, while smaller values represent standard deviations of the estimates.

		Network Layer	
Loss Function	Performance Metric	CIFNet 14 Layers	CIFNet 5 Layers
Distance-based	Accuracy ↑	0.846 ± 0.066	0.924 ± 0.091
	False Negative ↓	0.078 ± 0.039	0.047 ± 0.053
	False Positive ↓	0.076 ± 0.035	0.029 ± 0.039
	Matthews Correlation ↑	0.693 ± 0.131	0.849 ± 0.182
	Mean Absolute Error ↓	0.07 ± 0.049	0.029 ± 0.038
	Root Mean Square Error ↓	0.104 ± 0.07	0.044 ± 0.056
Label-based	Accuracy ↑	0.799 ± 0.028	0.819 ± 0.032
	False Negative ↓	0.008 ± 0.009	0.005 ± 0.012
	False Positive ↓	0.192 ± 0.027	0.176 ± 0.024
	Matthews Correlation ↑	0.644 ± 0.049	0.68 ± 0.06
	Mean Absolute Error ↓	0.108 ± 0.055	0.113 ± 0.06
	Root Mean Square Error ↓	0.154 ± 0.078	0.153 ± 0.081
Triplet-based	Accuracy ↑	0.543 ± 0.032	0.642 ± 0.065
	False Negative ↓	0.306 ± 0.102	0.181 ± 0.038
	False Positive ↓	0.151 ± 0.081	0.177 ± 0.029
	Matthews Correlation ↑	0.091 ± 0.063	0.284 ± 0.13
	Mean Absolute Error ↓	0.204 ± 0.109	0.206 ± 0.104
	Root Mean Square Error ↓	0.269 ± 0.144	0.271 ± 0.138

Table 4: In order to determine the optimal number of layers required for the best estimation of the embedding function, a comparison was conducted between two networks containing 5 and 14 layers, respectively.