“Why Here and not There?”: Diverse Contrasting Explanations of Dimensionality Reduction

André Artelt[1,2][a], Alexander Schulz[1][b] and Barbara Hammer[1][c]
[1] *Faculty of Technology, Bielefeld University, Bielefeld, Germany*
[2] *University of Cyprus, Nicosia, Cyprus*

Keywords: XAI, Dimensionality Reduction, Data Visualization, Counterfactual Explanations, Data Mining.

Abstract: Dimensionality reduction is a popular preprocessing and a widely used tool in data mining. Transparency, which is usually achieved by means of explanations, is nowadays a widely accepted and crucial requirement of machine learning based systems like classifiers and recommender systems. However, transparency of dimensionality reduction and other data mining tools have not been considered much yet, still it is crucial to understand their behavior – in particular practitioners might want to understand why a specific sample got mapped to a specific location. In order to (locally) understand the behavior of a given dimensionality reduction method, we introduce the abstract concept of contrasting explanations for dimensionality reduction, and apply a realization of this concept to the specific application of explaining two dimensional data visualization.

1 INTRODUCTION

Transparency of machine learning (ML) based system, applied in the real world, is nowadays a widely accepted requirement – the importance of transparency was also recognized by the policy makers and therefore made its way into legal regulations like the EU’s GDPR (parliament and council, 2016). A popular way of achieving transparency is by means of explanations (Molnar, 2019) which then gave rise to the field of eXplainable AI (XAI) (Samek et al., 2017; Tjoa and Guan, 2019). Although a lot of different explanation methodologies for ML based systems have been developed (Molnar, 2019; Tjoa and Guan, 2019), it is important to realize that it is still somewhat unclear what exactly makes up a good explanation (Doshi-Velez and Kim, 2017; Offert, 2017). Therefore one must carefully pick the right explanation in the right situation, as there are (potentially) different target users with different goals (Ribera and Lapedriza, 2019) – e.g. ML engineers need explanations that help them to improve the system, while lay users need trust building explanations. Popular explanations methods (Molnar, 2019; Tjoa and Guan, 2019) are feature relevance/importance methods (Fisher et al., 2018), and examples based methods (Aamodt and Plaza., 1994) which use a set or a single example for explaining the behavior of the system. Instances of example based methods are contrasting explanations like counterfactual explanations (Wachter et al., 2017; Verma et al., 2020) and prototypes & criticisms (Kim et al., 2016).

Dimensionality reduction methods are a popular tool in data mining, e.g. for data visualization, an often used preprocessing in ML pipelines (Gisbrecht and Hammer, 2015) and are also used for inspecting trained models (Schulz et al., 2021; Lapuschkin et al., 2019). Similar to other ML methods, dimensionality reduction methods itself are not easy to understand – i.e. a high-dimensional sample is “somehow” mapped to a low-dimensional sample without providing any explanation/reason of this mapping. A ML pipeline can not be transparent if it contains non-transparent preprocessings like dimensionality reduction, and a proper and responsible use of data analysis tools such as data visualization is not possible if the inner working of the tool is not understood. Therefore, we argue that there is a need for understanding dimensionality reduction methods – we aim to provide such an understanding by means of contrasting explanations.

Related Work. In the context of explaining dimensionality reduction, only little work exists so far. Some approaches (Schulz and Hammer, 2015; Schulz et al., 2014) aim to infer global feature importance for a given data projection. Another work (Bibal et al., 2020) estimates feature importance locally for a vicin-

[a] https://orcid.org/0000-0002-2426-3126
[b] https://orcid.org/0000-0002-0739-612X
[c] https://orcid.org/0000-0002-0935-5591

Artelt, A., Schulz, A. and Hammer, B.
"Why Here and not There?": Diverse Contrasting Explanations of Dimensionality Reduction.
DOI: 10.5220/0011618300003411
In *Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2023)*, pages 27-38
ISBN: 978-989-758-626-2; ISSN: 2184-4313

ity around a projected data point, using locally linear models. A recent paper (Bardos et al., 2022) proposes to use local feature importance explanations by computing a local linear approximation for each reduced dimension, extracting feature importances from the weight vectors. Further, saliency map approaches such as the layer-wise relevance propagation (LRP) (Bach et al., 2015) could in principle be applied to a parametric dimensionality reduction mapping in order to obtain locally relevant features. However, these approaches do not provide contrasting explanations, in which we are interested in this work.

Our Contributions. First, we make a conceptional contribution by proposing a general formalization of diverse counterfactual explanations for explaining dimensionality reduction methods. Second, we propose concrete realizations of this concept for four popular representatives of parametric dimensionality reduction method classes: PCA (linear mappings), SOM (Kohonen, 1990) (topographic mappings), autoencoders (Goodfellow et al., 2016) (neural networks) and parametric t-SNE (Van Der Maaten, 2009) (parametric extensions of neighbor embeddings). Finally, we empirically evaluate them in the particular usecase of two-dimensional data visualization.

The remainder of this work is structured as follows: First (Section 2) we review the necessary foundations of dimensionality reduction and contrasting explanations. Next (Section 3.1), we propose and formalize diverse counterfactual explanations for explaining dimensionality reduction – we first propose a general concept (Section 3.1), and then propose practical realizations for popular parametric dimensionality reduction methods (Sections 3.2,3.3). We empirically evaluate our proposed explanations in Section 4 where we consider two-dimensional data visualization as a popular application of dimensionality reduction. Finally, this work closes with a summary and conclusion in Section 5.

2 FOUNDATIONS

2.1 Dimensionality Reduction

The common setting for dimensionality reduction (DR) is that data $\vec{x}_i, i = 1,\ldots,m$ are given in a high-dimensional input space $\mathcal{X}$ – we will assume $\mathcal{X} = \mathbb{R}^d$ in the following. The goal is to project them to lower-dimensional points $\vec{y}_i, i = 1,\ldots,m$ in $\mathbb{R}^{d'}$ – where for data visualization often $d' = 2$ –, such that as much structure as possible is preserved. The precise mathematical formalization of the term “structure preservation” is then one of the key differences between different DR methods in literature (Van Der Maaten et al., 2009; Lee and Verleysen, 2007; Bunte et al., 2012).

One major view for grouping DR methods is whether they provide an explicit function $\phi : \mathcal{X} \to \mathbb{R}^{d'}$ for projection, where the parameters of ϕ are adjusted by the according DR method, or whether no such functional form is assumed by the approach. The former methods are referred to as parametric and the latter ones as non-parametric (Van Der Maaten et al., 2009; Gisbrecht and Hammer, 2015).

Since we require parametric mappings in our work, we recap a few of the most popular parametric DR approaches in the following. However, since there do exist successful extensions for non-parametric approaches to also provide a parametric function, we will consider one of them here as well. We will consider these approaches again in our experiments.

2.1.1 Linear Methods

The most classical DR methods are based on a linear functional form:

$$\phi(\vec{x}) = \mathbf{A}\vec{x} + \vec{b} \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{d' \times d}$ and $\vec{b} \in \mathbb{R}^{d'}$. Particular instances are Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and also the mappings obtained by metric learning approaches such as the Large Margin Nearest Neighbor (LMNN) method (Gisbrecht and Hammer, 2015). These constitute different cost function based approaches for estimating the parameters of $\phi(\cdot)$, but in the end result in such a linear parametric mapping Eq. (1).

2.1.2 Topographic Mappings

A class of non-linear DR approaches is given by topographic mappings such as the Self Organizing Map (SOM) and the Generative Topographic Mapping (GTM). We consider the SOM as one representative of this class of methods in the following. The SOM (Kohonen, 1990) consists of a set of prototypes $\vec{p}_{\vec{z}} \in \mathbb{R}^d$ which are mapped to an index set I, $\phi : \mathbb{R}^d \to I$ – e.g. the prototypes are arranged as a two-dimensional grid: $I \subset \mathbb{N}^2$. The dimensionality reduction maps a given input $\vec{x}$ to the index of the closest prototype:

$$\phi(\vec{x}) = \underset{\vec{z} \in I}{\arg\min}\, \|\vec{x} - \vec{p}_{\vec{z}}\|_2. \tag{2}$$

2.1.3 Autoencoder

An autoencoder (AE) $f_\theta : \mathbb{R}^d \to \mathbb{R}^d$ is a neural network consisting of an encoder, mapping the input to a smaller representation (also called the bottleneck)

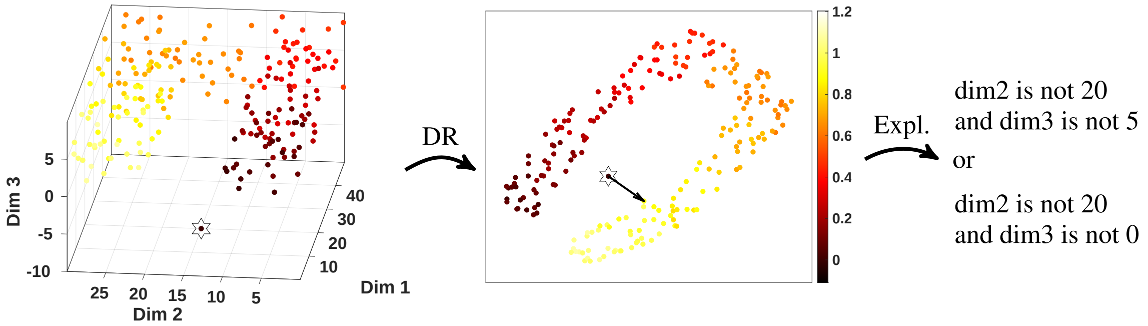

Figure 1: Illustration of the investigated topic: a 'high-dimensional' data set (left, with an outlier marked as a star) is mapped to two dimensions (middle), where the question 'why is the central point mapped here and not there' is asked (indicated by the arrow). Possible explanations are depicted (right).

and a decoder, mapping it back to the original input (Goodfellow et al., 2016):

$$f_\theta(\vec{x}) = (\mathrm{dec}_\theta \circ \mathrm{enc}_\theta)(\vec{x}), \quad (3)$$

which are trained to optimize the reconstruction loss. A (typically non-linear) dimensionality reduction $\phi(\cdot)$ based on this approach consists of the encoder mapping:

$$\phi(\vec{x}) = \mathrm{enc}_\theta(\vec{x}) \quad (4)$$

2.1.4 Neighbor Embeddings

The class of neighbor embedding methods constitutes a set of non-parametric approaches that are considered as the most successful or state-of-the-art techniques in many cases (Kobak and Berens, 2019; Becht et al., 2019; Gisbrecht and Hammer, 2015). Instances are the very popular t-Distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) approaches (Van der Maaten and Hinton, 2008; McInnes et al., 2018). In the following, we consider t-SNE as a representative of this class of methods and, among its parametric extensions (Van Der Maaten, 2009; Gisbrecht et al., 2015), the approach Parametric t-SNE (Van Der Maaten, 2009).

Parametric t-SNE, uses a neural network $f_\theta : \mathbb{R}^d \to \mathbb{R}^{d'}$ for mapping a given input $\vec{x}$ to a lower-dimensional domain:

$$\phi(\vec{x}) = f_\theta(\vec{x}), \quad (5)$$

with respect to the t-SNE cost function.

While there do exist more families of DR approaches, such as manifold embeddings (including MVU and LLE) or discriminative/supervised DR, it would exceed the scope of the present work to investigate all possible choices.

2.2 Contrasting Explanations

Contrasting explanations state a change to some features of a given input such that the resulting data point causes a different behavior of the system/model than the original input does. Counterfactual explanations (often just called *counterfactuals*) are the most prominent instance of contrasting explanations (Molnar, 2019). One can think of a counterfactual explanation as a recommendation of actions that change the model's behavior/prediction. One reason why counterfactual explanations are so popular is that there exists evidence that explanations used by humans are often contrasting in nature (Byrne, 2019) – i.e. people often ask questions like *"What would have to be different in order to observe a different outcome?"*. It was also shown that such questions are useful to learn about an unknown functionality and exploit this knowledge to achieve some goals (Kuhl et al., 2022a; Kuhl et al., 2022b).

A prominent example for illustrating the concept of a counterfactual explanation is the example of loan application: *Imagine you applied for a loan at a bank. Now, the bank rejects your application and you would like to know why. In particular, you would like to know what would have to be different so that your application would have been accepted. A possible explanation might be that you would have been accepted if you had earned 500$ more per month and if you had not had a second credit card.*

Unfortunately, many explanation methods (including counterfactual explanations) are lacking uniqueness: Often there exists more than one possible & valid explanation – this is called "Rashomon effect" (Molnar, 2019) – and in such cases, it is not clear which or how many of the possible explanations should be presented to the user. See Figure 2 where we illustrate the concept of a counterfactual explanation, including the existing of multiple possible and valid counterfactuals. Most approaches ignore this

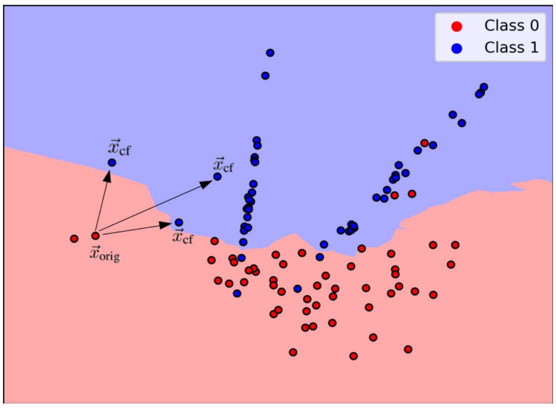

Figure 2: “Rashomon effect”: Illustration of multiple possible counterfactual explanations $\vec{x}_{\text{cf}}$ of a given sample $\vec{x}_{\text{orig}}$ for a binary classifier.

problem, however, there exist a few approaches that propose to compute multiple diverse counterfactuals to make the user aware that there exist different possible explanations (Rodriguez et al., 2021; Russell, 2019; Mothilal et al., 2020). In order to keep the explanation (suggested changes) simple – i.e. we are looking for low-complexity explanations that are easy to understand – an obvious strategy is to look for a small number of changes so that the resulting sample (counterfactual) is similar/close to the original sample. This is aimed to be captured by Definition 1.

Definition 1 ((Closest) Counterfactual Explanation (Wachter et al., 2017))**.** *Assume a prediction function (e.g. a classifier) $h : \mathbb{R}^d \to \mathcal{Y}$ is given. Computing a counterfactual $\vec{x}_{cf} \in \mathbb{R}^d$ for a given input $\vec{x} \in \mathbb{R}^d$ is phrased as an optimization problem:*

$$\underset{\vec{x}_{cf} \in \mathbb{R}^d}{\arg\min}\ \ell\left(h(\vec{x}_{cf}), y_{cf}\right) + C \cdot \theta(\vec{x}_{cf}, \vec{x}) \qquad (6)$$

where $\ell(\cdot)$ denotes a loss function, $\vec{y}_{cf}$ the target prediction, $\theta(\cdot)$ a penalty for dissimilarity of $\vec{x}_{cf}$ and $\vec{x}$, and $C > 0$ denotes the regularization strength.

The counterfactuals from Definition 1 are also called *closest counterfactuals* because the optimization problem Eq. (6) tries to find an explanation $\vec{x}_{\text{cf}}$ that is as close as possible to the original sample $\vec{x}$. However, other aspects like plausibility and actionability are ignored in Definition 1, but are covered in other work (Looveren and Klaise, 2019; Artelt and Hammer, 2020; Artelt and Hammer, 2021). In this work, we refer to counterfactuals in the spirit of Definition 1. Note that counterfactual explanations also exist in the causality domain (Pearl, 2010). Here the knowledge of a structural causal model (SCM), describing the interaction of features, is assumed. This work is not based in the causality domain and we only consider counterfactual explanations as proposed by (Wachter et al., 2017).

3 COUNTERFACTUAL EXPLANATIONS OF DIMENSIONALITY REDUCTION

In this section, we propose counterfactual explanations of dimensionality reduction – i.e. explaining why a specific point was mapped to some location instead of a requested different location. As it is the nature of counterfactual explanations, the explanations state how we have to (minimally) change the original sample such that it gets mapped to some requested location – see Figure 1 for an illustrative example.

We argue that this type of explanation is in particular very well suited for explaining data visualization which is a common application of dimensionality reduction in data mining (Gisbrecht and Hammer, 2015; Lee and Verleysen, 2007; Kaski and Peltonen, 2011) – e.g. data is mapped to a two-dimensional space which is then depicted in a scatter plot. For instance, we could utilize such explanations to explain outliers in the data visualization: I.e. explaining why a point got mapped far away from the other points instead of close to the other ones – a counterfactual explanation states how to change the outlier such that it is no longer an outlier in the visualization, which would allow us to learn something about the particular reasons why this point was flagged as an outlier in the visualization. See Figure 3 for an illustrative example where we explain anomalous pressure measurements in a water distribution network: We consider the hydraulically isolated “Area A” in the L-Town network (Vrachimis et al., 2020) where 29 pressure sensors are installed – we simulate a sensor failure (constant added the original pressure value) in node *n105*. We pick an outlier $\vec{x}$ (see left plot in Figure 3) and compute a counterfactual explanation for each normal data point as a target mapping – i.e. asking which sensor measurements must be changed so that the overall measurement vector $\vec{x}_{\text{cf}}$ is mapped to the specified location in the data visualization. When aggregating all explanations by summing up and normalizing the suggested changes for each sensor (see right plot in Figure 3), we are able to identify the faulty sensors and thereby “explain” the outlier.

Note that existing explanation methods for explaining dimensionality reduction methods, which usually focus on feature importances (see Section 1), can not provide such an explanation – i.e. answer-

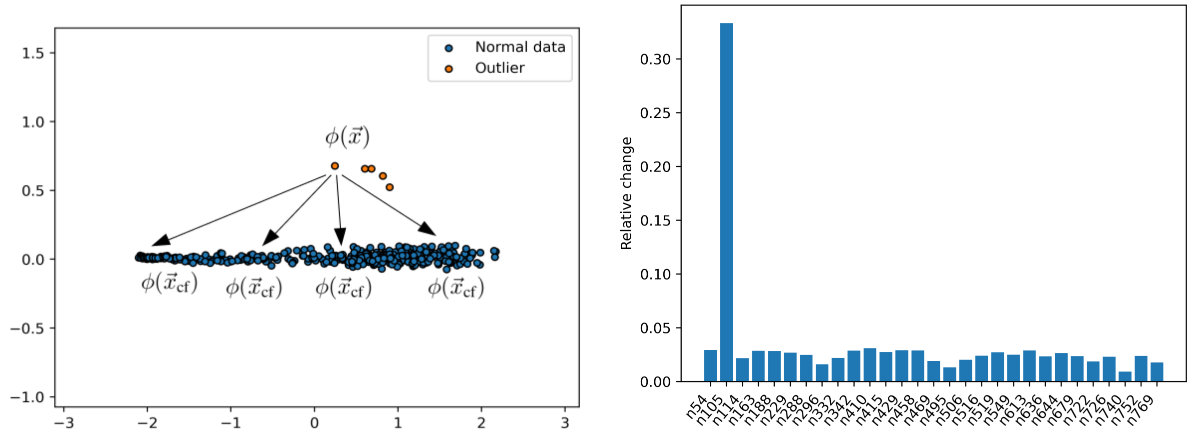

Figure 3: Explaining anomalous pressure measurements – Left: Two dimensional data visualization of 29-dimensional pressure measurements; Right: Normalized amount of suggested changes per sensor – the faulty sensor is suggested to change the most and is therefore correctly identified.

ing contrasting questions like "Why was the point mapped here and not there". This is because they only highlight important features but do not suggest any changes or magnitude of changes that would yield a different (requested) mapping. However, as already mentioned in Section 2.2, the Rashomon effect states that there might exist many possible explanations why a particular point was mapped far away from the others – we therefore aim for a set of diverse counterfactual explanations in order to learn the most about the observed mapping and provide different possibilities for actionable recourse.

First, we formalize the general concept of (diverse) counterfactual explanations of dimensionality reduction in Section 3.1. Next, we consider some popular parametric dimensionality reduction methods, and propose methods for efficiently computing single counterfactuals (see Section 3.2) and diverse counterfactuals (see Section 3.3).

3.1 General Modeling

We assume the DR method is given as a mapping

$$\phi : \mathbb{R}^d \to \mathbb{R}^{d'} \tag{7}$$

with $d > d'$.

A counterfactual explanation of a sample $\vec{x} \in \mathbb{R}^d$ is a sample $\vec{x}_{\mathrm{cf}}$ in the original domain (i.e. $\mathbb{R}^d$) that differs in a few features only from the given original sample $\vec{x}$, but is mapped to a requested location $\vec{y}_{\mathrm{cf}} \in \mathbb{R}^{d'}$ which is different from the mapping of the original sample $\vec{x}$. We formalize this in Definition 2.

Definition 2 (Counterfactual Explanation of Dimensionality Reduction)**.** *For a given DR method* $\phi(\cdot)$ *Eq.* (7), *a counterfactual explanation* $\vec{x}_{cf} \in \mathbb{R}^d, \vec{y}_{cf} \in \mathbb{R}^{d'}$ *of a specific sample* $\vec{x} \in \mathbb{R}^d$ *is given as a solution to the following multi-criteria optimization problem:*

$$\min_{\vec{x}_{cf} \in \mathbb{R}^d} \Big(\|\vec{x} - \vec{x}_{cf}\|_0, \|\phi(\vec{x}_{cf}) - \vec{y}_{cf}\|_p \Big), \tag{8}$$

where p defines the norm that is used. As discussed in Section 2.2, there usually exists more than one possible explanation ("Rashomon effect") – clearly this is the case for dimensionality reduction as well because dimensionality reduction is a many-to-one mapping (i.e. multiple points are mapped to the same location). In this context, a set of diverse (i.e. highly different) explanations would provide more information than a single explanation only. We therefore extend Definition 2 to a set of diverse counterfactuals explanations instead of a single one:

Definition 3 (Diverse Counterfactual Explanations of Dimensionality Reduction)**.** *For a given DR method* $\phi(\cdot)$ *Eq.* (7), *a set of diverse counterfactual explanations* $\{\vec{x}^i_{cf} \in \mathbb{R}^d\}, \vec{y}_{cf} \in \mathbb{R}^{d'}$ *of a specific sample* $\vec{x} \in \mathbb{R}^d$ *is given as a solution to the following multi-criteria optimization problem:*

$$\min_{\{\vec{x}^i_{cf} \in \mathbb{R}^d\}} \Big(\|\vec{x} - \vec{x}^i_{cf}\|_0, \|\phi(\vec{x}^i_{cf}) - \vec{y}_{cf}\|_p, \psi(\vec{x}^i_{cf}, \vec{x}^j_{cf}) \Big) \tag{9}$$

where $\psi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ *denotes a function measuring the pair-wise diversity of two given counterfactuals – i.e. returning a small value if the two counterfactuals are very different and a larger value otherwise.*

The term "diversity" itself is somewhat fuzzy and different use-cases might require different definitions of *diverse counterfactuals*. In this work we utilize a very general definition of diversity, namely the number of overlapping features – i.e. diverse counterfactuals should not change the same features:

$$\psi(\vec{x}_{\text{cf}}^{j},\vec{x}_{\text{cf}}^{k}) = \sum_{i=1}^{d} \mathbb{1}\left((\vec{\delta}_{\text{cf}}^{j})_i \neq 0 \wedge (\vec{\delta}_{\text{cf}}^{k})_i \neq 0 \right) \tag{10}$$

where $\vec{\delta}_{\text{cf}}^{j} = \vec{x}_{\text{cf}}^{j} - \vec{x}$ and $\mathbb{1}(\cdot)$ denotes the indicator function that returns 1 if the boolean expression is true and 0 otherwise.

3.2 Method Specific Computation of a Single Counterfactual

Hereinafter, we propose practical relaxations for computing a single counterfactual explanation of different parametric dimensionality reduction methods (see Definition 2). While Definition 2 does not make any assumptions on the dimensionality reduction $\phi(\cdot)$, we now assume a parametric dimensionality reduction in order to get tractable optimization problems.

Note that in all cases, we approximate the 0-norm with the 1-norm for measuring closeness between the original sample $\vec{x}$ and the counterfactual $\vec{x}_{\text{cf}}$. Furthermore, we use $p = 2$, the 2-norm for measuring the distance between the mapping of $\vec{x}_{\text{cf}}$ and the requested mapping $\vec{y}_{\text{cf}}$.

3.2.1 Linear Methods

In the case of linear mappings as defined in Section 2.1.1, we phrase the computation of a single counterfactual explanations as the following convex quadratic program:

$$\begin{aligned} &\underset{\vec{x}_{\text{cf}} \in \mathbb{R}^d}{\arg\min}\ \|\vec{x} - \vec{x}_{\text{cf}}\|_1 + C \cdot \xi \\ &\text{s.t. } \|\mathbf{A}\vec{x}_{\text{cf}} + \vec{b} - \vec{y}_{\text{cf}}\|_2^2 \leq \xi \\ &\xi \geq 0 \end{aligned} \tag{11}$$

where $C > 0$ acts as a regularization strength balancing between the two objectives in Eq. (8) – the regularization is necessary because it is numerically difficult (or even impossible) to find a counterfactual $\vec{x}_{\text{cf}}$ that yields the exact mapping $\phi(\vec{x}_{\text{cf}}) = \vec{y}_{\text{cf}}$, we therefore have to specify how much difference we are willing to tolerate. Note that convex quadratic programs can be solved efficiently (Boyd and Vandenberghe, 2004).

3.2.2 Self Organizing Map

Similar to linear methods, we phrase the computation of a single counterfactual explanations for SOMs (Section 2.1.2) as the following convex quadratic program, which again can be solved efficiently using standard solvers from convex optimization (Boyd and Vandenberghe, 2004):

$$\begin{aligned} &\underset{\vec{x}_{\text{cf}} \in \mathbb{R}^d}{\arg\min}\ \|\vec{x} - \vec{x}_{\text{cf}}\|_1 \\ &\text{s.t. } \|\vec{x}_{\text{cf}} - \vec{p}_{\vec{y}_{\text{cf}}}\|_2^2 + \varepsilon \leq \|\vec{x}_{\text{cf}} - \vec{p}_{\vec{z}}\|_2^2 \quad \forall \vec{z} \in I \end{aligned} \tag{12}$$

where $\varepsilon > 0$ makes sure that the set of feasible solutions is closed.

3.2.3 Autoencoder

For autoencoders (AEs) as discussed in Section 2.1.3, we utilize the penalty method to merge the two objectives from Eq. (8) into a single objective:

$$\underset{\vec{x}_{\text{cf}} \in \mathbb{R}^d}{\arg\min}\ \|\vec{x} - \vec{x}_{\text{cf}}\|_1 + C \cdot \|\text{enc}_\theta(\vec{x}_{\text{cf}}) - \vec{y}_{\text{cf}}\|_2 \tag{13}$$

where the hyperparameter $C > 0$ acts as a regularization strength.

Assuming continuous differentiability of the encoder $\text{enc}_\theta(\cdot)$, we can solve Eq. (13) using a gradient based method. However, due to the non-linearity of $\text{enc}_\theta(\cdot)$, we might find a local optimum only.

3.2.4 Parametric t-SNE

Although the neural network $f_\theta(\cdot)$ of parametric t-SNE (Section 2.1.4) is trained in a completely different way compared to an autoencoder based dimensionality reduction, the final modeling is the same and consequently, everything from the case of autoencoder based DR applies here as well:

$$\underset{\vec{x}_{\text{cf}} \in \mathbb{R}^d}{\arg\min}\ \|\vec{x} - \vec{x}_{\text{cf}}\|_1 + C \cdot \|f_\theta(\vec{x}_{\text{cf}}) - \vec{y}_{\text{cf}}\|_2 \tag{14}$$

3.3 Computation of Diverse Counterfactuals

In this section, we propose an algorithm for computing diverse counterfactual explanations (see Definition 3) of the four DR methods considered in the previous section.

Regarding the formalization of diversity Eq. (10), instead of using Eq. (10) directly, we propose a more stricter version in order to get a continuous function which then yields tractable optimization problems, similar to the ones we proposed in the previous section: In order to compute a set of diverse counterfactuals instead of a single counterfactual, we utilize our proposed methods for computing a single counterfactual explanations from Section 3.2 and extend these with a mechanism to forbid or punish changes in black-listed features. We then first compute a single counterfactual explanations using the

Algorithm 1: Computation of Diverse Counterfactuals.

```
Input: Original input $\vec{x}$, Target location $\vec{y}_{\text{cf}}$, $k \geq 1$: number of diverse counterfactuals, Dimensionality reduction $\phi(\cdot)$
Output: Set of diverse counterfactuals $\mathcal{R} = \{\vec{x}^{i}_{\text{cf}}\}$
$\mathcal{F} = \{\}$    ▷ Initialize set of black-listed features
$\mathcal{R} = \{\}$    ▷ Initialize set of diverse counterfactuals
for $i = 1, \dots, k$ do    ▷ Compute $k$ diverse counterfactuals
    $\vec{x}^{i}_{\text{cf}} = \text{CF}_{\phi}(\vec{x}, \vec{y}_{\text{cf}}, \mathcal{F})$    ▷ Compute next counterfactual
    $\mathcal{R} = \mathcal{R} \cup \{\vec{x}^{i}_{\text{cf}}\}$
    $\mathcal{F} = \mathcal{F} \cup \{j \mid (\vec{x}_{\text{cf}} - \vec{x})_j \neq 0\}$    ▷ Update set of black-listed features
end for
```

methodology proposed in Section 3.2 and then iteratively compute another counterfactual explanation but black-listing all features that have been changed in the previous counterfactuals – this procedure is illustrated as pseudo-code in Algorithm 1.

Black-Listing Features. We assume we are given an ordered set $\mathcal{F}$ of black-listed features. In case of convex programs (e.g. linear methods and SOM), we consider black-listed features $\mathcal{F}$ by means of an additional affine equality constraint:

$$\mathbf{M}\vec{x}_{\text{cf}} = \vec{m} \tag{15}$$

where $\mathbf{M} \in \mathbb{R}^{|\mathcal{F}| \times d}, \vec{m} \in \mathbb{R}^{|\mathcal{F}|}$ with

$$(\mathbf{M})_{i,j} = \begin{cases} 1 & \text{if } (\mathcal{F})_i = j \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

and $\vec{m}_k = (\vec{x})_{(\mathcal{F})_k}$.

Whereas in all other cases (e.g. autoencoder and parametric t-SNE), where we minimize a (non-convex) cost function, we replace the counterfactual $\vec{x}_{\text{cf}}$ in the optimization problem with an affine mapping undoing any potential changes in black-listed features – i.e. black-listed features can be changed but have no effect on the final counterfactual because they are reset to their original value:

$$\|\phi(\mathbf{M}\vec{x}_{\text{cf}} + \vec{m}) - \vec{y}_{\text{cf}}\|_2 \tag{17}$$

where $\mathbf{M} \in \mathbb{R}^{d \times d}, \vec{m} \in \mathbb{R}^{d}$ with

$$\begin{aligned} (\mathbf{M})_{i,j} &= \begin{cases} 1 & \text{if } i = j \text{ and } i \notin \mathcal{F} \\ 0 & \text{otherwise} \end{cases} \\ (\vec{m})_i &= \begin{cases} (\vec{x})_i & \text{if } i \in \mathcal{F} \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{18}$$

Note that in both cases, the complexity and type of optimization problem does not change – e.g. convex programs remain convex programs.

For convenience, we use $\text{CF}_{\phi}(\vec{x}, \vec{y}_{\text{cf}}, \mathcal{F})$ to denote the computation of a counterfactual $(\vec{x}_{\text{cf}}, \vec{y}_{\text{cf}})$ of a DR method $\phi(\cdot)$ at a given sample $\vec{x}$ subject to a set $\mathcal{F}$ of black-listed features.

4 EXPERIMENTS

We empirically evaluate our proposed explanation methodology of DR methods on the specific use-case of data visualization – i.e. dimensionality reduction to two dimensions. All experiments are implemented in Python and are publicly available on GitHub[1].

4.1 Data

We run all our experiments on a set of different ML benchmark data sets – all data sets are standardized:

Diabetes. The "Diabetes Data Set" (N/A, 1994) is a labeled data set containing recordings from diabetes patients. The data set contains 442 samples and 10 real valued scaled features in $[-.2, .2]$ such as body mass index, age in years and average blood pressure. The labels are integers in $[25, 346]$ denoting a quantitative measure of disease progression one year after baseline.

Breast Cancer. The "Breast Cancer Wisconsin (Diagnostic) Data Set" (William H. Wolberg, 1995) is used for classifying breast cancer samples into benign and malignant (i.e. binary classification). The data set contains 569 samples and 30 numerical features such as area, smoothness and compactness.

Toy. An an artificial, self created, toy data set containing 500 ten dimensional samples. Each feature is distributed according to a normal distribution whereby we choose a different random mean for each feature - by this we can guarantee that, in contrast to the other data sets, the features are independent of each other. The binary labelling of the samples is done by splitting the data into two clusters using k-means.

4.2 Model Agnostic Algorithm for Comparison

We compare Algorithm 1 to a general model agnostic algorithm (ModelAgnos) for computing diverse

[1] https://github.com/HammerLabML/ContrastingExplanationDimRed

counterfactual explanations where we select samples from the training data set $\mathcal{D}$ that minimize a weighted combination of Eq. (9) – i.e. we make use of the penalty method to solve the multi-objective optimization problem Eq. (9) without making any further assumption on the dimensionality reduction $\phi(\cdot)$:

$$\min_{\{\vec{x}^i_{\text{cf}} \in \mathcal{D}\}} C_1 \cdot \|\vec{x} - \vec{x}^i_{\text{cf}}\|_1 + C_2 \cdot \|\phi(\vec{x}^i_{\text{cf}}) - \vec{y}_{\text{cf}}\|_2 + C_3 \cdot \psi(\vec{x}^i_{\text{cf}}, \vec{x}^j_{\text{cf}}) \tag{19}$$

where $C_1, C_2, C_3 > 0$ denote regularization coefficients that allow us to balance between the different objectives, and $\psi(\cdot)$ is implemented as stated in Eq. (10). By limiting the set of feasible solutions to the training data set, we can guarantee plausibility of the resulting counterfactual explanations – note that plausibility of the counterfactuals generated by Algorithm 1 can not be guaranteed.

4.3 Setup

For each data set and each of the four parametric DR methods (PCA, Autoencoder, SOM, parametric t-SNE) from Section 3.2, we fit the DR method to the entire data set and compute for each sample in the data set a set of three diverse counterfactual explanations – we evaluate and compare the counterfactuals[2] computed by our proposed Algorithm 1 with those from the model agnostic algorithm (see Section 4.2). For the requested target location $\vec{y}_{\text{cf}}$ – recall that in a counterfactual explanation we ask for a change that would lead to a different specified mapping $\vec{y}_{\text{cf}}$ instead of the original mapping $\vec{y}$ – we consider two scenarios:

- *Perturbations:* Choose the mapping of the original sample $\vec{x}$ after perturbing three random features – the same type of perturbation is applied to these three features.
- *Without any perturbations:* Choose the mapping of a different sample (with a different label) from the training data set as $\vec{y}_{\text{cf}}$.

Regarding the perturbations, we consider the following ones:

- *Shift*: A constant is added to the feature value.
- *Gaussian*: Gaussian noise is added to feature value.

Note that we evaluate each perturbation separately. Furthermore, note that these perturbations could be interpreted as sensor failures and are therefore highly relevant to practice.

4.4 Evaluation

For all experimental scenarios, we monitor and evaluate some quantitative measurements:

- *CfSparse*: Sparsity of the counterfactual explanations – i.e. how many (percentage) of the available features are used in the explanation, smaller values are better.
- *CfDist*: Euclidean distance between the mapping of the counterfactual $\phi(\vec{x}_{\text{cf}})$ and the requested mapping $\vec{y}_{\text{cf}}$ – i.e. this can be interpreted as a measurement of the error of counterfactual explanations, smaller values are better.
- *CfDiv*: Diversity of the counterfactual explanations – i.e. the number of overlapping features between the diverse explanations (see Eq. (10) in Section 3.1), smaller values are better.

For the scenarios where we apply a perturbation to the original sample, we also record the recall of the identified perturbed features in the counterfactual explanations – i.e. checking if the used features in the explanation coincide with the perturbed features. By this, we try to measure the usefulness of our explanations for identifying relevant features – however, since dimensionality reduction is a many-to-one mapping, we consider recall only because we do not expect to observe a high precision due to the Rashomon effect.

Note that each experiment is repeated 100 times in order to get statistically reliable estimates of the quantitative measurements.

4.5 Results

The results of the scenario without any perturbations – i.e. randomly selecting the target sample from the training set – are shown in Table 1 and the results of the scenarios with perturbations are shown in Tables 6,2 – note that, due to space constraints, the latter one is put in the appendix.

We observe that Algorithm 1, on average, achieves much sparser and more diverse explanations than the mode agnostic algorithm (Section 4.2) does. Only in case of SOM, the sparsity is often a bit worse than those from the baseline – this might be due to numerical instabilities of the mathematical program Eq. (12). In particular, while Algorithm 1 almost always yields completely diverse explanations, the model agnostic algorithm fails completely – this highlights the strength of our proposed Algorithm 1 for computing diverse explanations. Furthermore, both methods are able to yield counterfactual explanations that are very close to the requested target location. In most cases Algorithm 1 yields counterfactuals that are closer to

[2] All hyperparameters (regularization strength) C_i $\forall i$ are set to 1.

Table 1: Quantitative results: *No perturbation* – all numbers are rounded to two decimal places, best scores are highlighted in **bold-face**.

	DataSet	CfSparse ↓		CfDiv ↓		CfDist ↓	
		Algo 1	ModelAgnos	Algo 1	ModelAgnos	Algo 1	ModelAgnos
Linear	Diabetes	**0.21 ± 0.0**	0.55 ± 0.0	**0.0 ± 0.0**	7.07 ± 0.86	**0.2 ± 0.2**	1.33 ± 0.39
	Breast cancer	**0.15 ± 0.01**	0.66 ± 0.0	**0.0 ± 0.0**	29.6 ± 0.89	**0.36 ± 1.47**	2.64 ± 1.5
	Toy	**0.21 ± 0.0**	0.67 ± 0.0	**0.0 ± 0.0**	9.99 ± 0.01	**0.03 ± 0.04**	0.85 ± 0.12
SOM	Diabetes	0.88 ± 0.02	**0.61 ± 0.01**	**0.0 ± 0.0**	9.63 ± 21.16	**0.01 ± 0.15**	3.2 ± 2.73
	Breast cancer	0.91 ± 0.02	**0.66 ± 0.0**	**0.0 ± 0.04**	29.71 ± 0.62	**0.4 ± 5.22**	4.07 ± 3.91
	Toy	0.84 ± 0.03	**0.68 ± 0.0**	**0.0 ± 0.0**	10.6 ± 11.73	**0.01 ± 0.17**	3.77 ± 3.18
AE	Diabetes	**0.14 ± 0.03**	0.51 ± 0.0	**0.0 ± 0.01**	6.15 ± 0.7	0.28 ± 0.08	**0.23 ± 0.04**
	Breast cancer	**0.03 ± 0.01**	0.65 ± 0.0	**0.0 ± 0.05**	28.95 ± 0.92	0.36 ± 0.16	**0.31 ± 0.12**
	Toy	**0.14 ± 0.03**	0.67 ± 0.0	**0.0 ± 0.02**	9.98 ± 0.02	0.3 ± 0.09	**0.18 ± 0.02**
t-SNE	Diabetes	**0.33 ± 0.0**	0.58 ± 0.0	**0.0 ± 0.0**	8.12 ± 1.11	5.35 ± 7.43	**3.0 ± 1.83**
	Breast cancer	**0.32 ± 0.01**	0.67 ± 0.0	**0.23 ± 4.74**	29.86 ± 0.17	8.52 ± 11.59	**4.32 ± 2.16**
	Toy	**0.33 ± 0.0**	0.67 ± 0.0	**0.0 ± 0.0**	10.0 ± 0.0	1.92 ± 0.81	**1.06 ± 0.21**

the target location, only in case of parametric t-SNE the model agnostic algorithm yields "better" counterfactuals – however, in both cases the variance is quite large which indicates instabilities of the learned dimensionality reduction. Note that, since the three evaluation metrics 4.4 are contradictory, it can be misleading to evaluate the performance under each metric separately without looking at the other metrics at the same time – e.g. a method might yield very sparse counterfactuals but their distance to the requesting mappings is very large. In order to compensate the contradictory nature of the evaluation metrics, we suggest to also consider a ranking over the three metrics when assessing the performance of the two proposed algorithms for computing counterfactuals – we give such a ranking in Tables 3,4,5. According to these rankings, Algorithm 1 outperforms the model agnostic algorithm in many cases or is at at least as good as the model agnostic method but never worse. While the recall of the baseline is very good across all DR methods and data sets, the recall of Algorithm 1 is often very good as well, however, there exist some cases (in particular the breast cancer data set) where the recall drops significantly compared to the model agnostic algorithm.

5 CONCLUSION

In this work, we proposed the abstract concept of contrasting explanations for locally explaining dimensionality reduction methods – we considered two-dimensional data visualization as a popular example application. In order to deal with the Rashomon effect – i.e. the fact that there exist more than one possible and valid explanation – we considered a set of diverse explanations instead of a single explanation. Furthermore, we also proposed an implementation of this concept using counterfactual explanations and proposed modelings and algorithms for efficiently computing diverse counterfactual explanations of different parametric dimensionality reduction methods. We empirically evaluated different aspects of our proposed algorithms on different standard benchmark data sets – we observe that our proposed methods consistently yield good results.

Based on this initial work, there are a couple of potential extensions and directions for future research:

Depending on the domain and application, it might be necessary to guarantee plausibility of the counterfactuals – i.e. making sure that the counterfactual $\vec{x}_{\text{cf}}$ is reasonable and plausible in the data domain. Implausibility or a lack of realism of the counterfactual $\vec{x}_{\text{cf}}$ might hinder successful recourse in practice. In this work, we ignored the aspect of plausibility and it might happen that the computed counterfactuals $\vec{x}_{\text{cf}}$ are not always realistic samples from the data domain – only in case of our model agnostic algorithm (see Section 4.2) we can guarantee plausibility because we only consider samples from the training data set $\mathcal{D}$ as potential counterfactuals $\vec{x}_{\text{cf}}$. In future work, a first approach could be to add plausibility constraints to our proposed modelings (see Section 3.1) like it was done for counterfactual explanations of classifiers (Artelt and Hammer, 2020; Artelt and Hammer, 2021; Looveren and Klaise, 2019).

Another crucial aspects of transparency & explainability is the human. In particular, quantitative evaluation of algorithmic properties do not necessary coincide with a human evaluation (Kuhl et al., 2022a). Therefore we suggest to conduct a user-study to evaluate how "useful" our proposed explanation actually are – in particular it would be of interest to compare normal vs. plausible explanations, and to compare diverse explanations vs. a single explanations.

Table 2: Quantitative results: *Shift perturbation* – all numbers are rounded to two decimal places, best scores are highlighted in **bold-face**.

	DataSet	CfSparse ↓		CfDiv ↓		CfDist ↓		Recall ↑	
		Algo 1	ModelAgnos	Algo 1	ModelAgnos	Algo 1	ModelAgnos	Algo 1	ModelAgnos
Linear	Diabetes	**0.23 ± 0.0**	0.56 ± 0.0	**0.0 ± 0.0**	7.44 ± 0.82	**0.28 ± 0.37**	2.86 ± 1.42	0.72 ± 0.07	**0.95 ± 0.02**
	Breast cancer	**0.11 ± 0.01**	0.66 ± 0.0	**0.0 ± 0.0**	29.48 ± 1.02	**0.02 ± 0.02**	1.56 ± 0.26	0.42 ± 0.09	**1.0 ± 0.0**
	Toy	**0.21 ± 0.0**	0.67 ± 0.0	**0.0 ± 0.0**	10.0 ± 0.0	**0.01 ± 0.03**	2.34 ± 2.21	0.8 ± 0.04	**1.0 ± 0.0**
SOM	Diabetes	0.8 ± 0.04	**0.6 ± 0.01**	**0.0 ± 0.01**	9.31 ± 20.54	**0.01 ± 0.09**	3.74 ± 3.72	0.94 ± 0.02	**0.96 ± 0.01**
	Breast cancer	0.67 ± 0.07	**0.66 ± 0.0**	**0.01 ± 0.18**	29.68 ± 0.84	**0.31 ± 4.04**	4.6 ± 13.18	0.95 ± 0.02	**1.0 ± 0.0**
	Toy	0.77 ± 0.04	**0.67 ± 0.0**	**0.0 ± 0.01**	10.47 ± 9.14	**0.02 ± 0.25**	3.67 ± 3.3	0.92 ± 0.03	**1.0 ± 0.0**
AE	Diabetes	**0.14 ± 0.03**	0.51 ± 0.0	**0.0 ± 0.03**	6.21 ± 0.71	0.78 ± 0.6	**0.74 ± 0.63**	0.43 ± 0.24	**0.9 ± 0.03**
	Breast cancer	**0.04 ± 0.01**	0.65 ± 0.0	**0.0 ± 0.04**	28.98 ± 0.94	**0.47 ± 0.24**	0.51 ± 0.24	0.13 ± 0.1	**1.0 ± 0.0**
	Toy	**0.16 ± 0.03**	0.67 ± 0.0	**0.0 ± 0.01**	9.97 ± 0.03	0.62 ± 0.26	**0.57 ± 0.39**	0.48 ± 0.25	**1.0 ± 0.0**
t-SNE	Diabetes	**0.33 ± 0.0**	0.58 ± 0.0	**0.01 ± 0.09**	8.06 ± 0.91	6.13 ± 8.14	**4.4 ± 5.63**	**1.0 ± 0.0**	0.96 ± 0.01
	Breast cancer	**0.32 ± 0.0**	0.66 ± 0.0	**0.08 ± 1.42**	29.65 ± 0.95	3.14 ± 2.2	**2.0 ± 0.55**	0.97 ± 0.02	**1.0 ± 0.0**
	Toy	**0.33 ± 0.0**	0.67 ± 0.0	**0.0 ± 0.0**	10.0 ± 0.0	2.81 ± 1.46	**1.87 ± 0.97**	**1.0 ± 0.0**	**1.0 ± 0.0**

Table 3: Ranking of results from Table 1 – counting the number of metrics where the method yields the best score, best scores are highlighted in **bold-face**.

	DataSet	Algo 1	ModelAgnos
Linear	Diabetes	**3/3**	0/3
	Breast cancer	**3/3**	0/3
	Toy	**3/3**	0/3
SOM	Diabetes	**2/3**	1/3
	Breast cancer	**2/3**	1/3
	Toy	**2/3**	1/3
AE	Diabetes	**2/3**	1/3
	Breast cancer	**2/3**	1/3
	Toy	**2/3**	1/3
t-SNE	Diabetes	**2/3**	1/3
	Breast cancer	**2/3**	1/3
	Toy	**2/3**	1/3

Table 4: Ranking of results from Table 2 – counting the number of metrics where the method yields the best score, best scores are highlighted in **bold-face**.

	DataSet	Algo 1	ModelAgnos
Linear	Diabetes	**3/4**	1/4
	Breast cancer	**3/4**	1/4
	Toy	**3/4**	1/4
SOM	Diabetes	**2/4**	**2/4**
	Breast cancer	**2/4**	**2/4**
	Toy	**2/4**	**2/4**
AE	Diabetes	**2/4**	**2/4**
	Breast cancer	**3/4**	1/4
	Toy	**2/4**	**2/4**
t-SNE	Diabetes	**3/4**	1/4
	Breast cancer	**2/4**	**2/4**
	Toy	**3/4**	1/4

ACKNOWLEDGEMENTS

We gratefully acknowledge funding from the VW-Foundation for the project *IMPACT* funded in the frame of the funding line *AI and its Implications for Future Society*.

REFERENCES

Aamodt, A. and Plaza., E. (1994). Case-based reasoning: Foundational issues, methodological variations, and systemapproaches. *AI communications*.

Artelt, A. and Hammer, B. (2020). Convex density constraints for computing plausible counterfactual explanations. 29th International Conference on Artificial Neural Networks (ICANN).

Artelt, A. and Hammer, B. (2021). Convex optimization for actionable \& plausible counterfactual explanations. *CoRR*, abs/2105.07630.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

Bardos, A., Mollas, I., Bassiliades, N., and Tsoumakas, G. (2022). Local explanation of dimensionality reduction. *arXiv preprint arXiv:2204.14012*.

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W., Ng, L. G., Ginhoux, F., and Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*, 37(1):38–44.

Bibal, A., Vu, V. M., Nanfack, G., and Frénay, B. (2020). Explaining t-sne embeddings locally by adapting lime. In *ESANN*, pages 393–398.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, New York, NY, USA.

Bunte, K., Biehl, M., and Hammer, B. (2012). A general framework for dimensionality-reducing data visualization mapping. *Neural Computation*, 24(3):771–804.

Byrne, R. M. J. (2019). Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning. In *IJCAI-19*.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning.

Fisher, A., Rudin, C., and Dominici, F. (2018). All Models are Wrong but many are Useful: Variable Importance for Black-Box, Proprietary, or Misspecified Prediction Models, using Model Class Reliance. *arXiv e-prints*, page arXiv:1801.01489.

Gisbrecht, A. and Hammer, B. (2015). Data visualization by nonlinear dimensionality reduction. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(2):51–73.

Gisbrecht, A., Schulz, A., and Hammer, B. (2015). Parametric nonlinear dimensionality reduction using kernel t-sne. *Neurocomputing*, 147:71–82.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Kaski, S. and Peltonen, J. (2011). Dimensionality reduction for data visualization [applications corner]. *IEEE signal processing magazine*, 28(2):100–104.

Kim, B., Koyejo, O., and Khanna, R. (2016). Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in Neural Information Processing Systems 29*.

Kobak, D. and Berens, P. (2019). The art of using t-sne for single-cell transcriptomics. *Nature communications*, 10(1):1–14.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480.

Kuhl, U., Artelt, A., and Hammer, B. (2022a). Keep your friends close and your counterfactuals closer: Improved learning from closest rather than plausible counterfactual explanations in an abstract setting. *arXiv preprint arXiv:2205.05515*.

Kuhl, U., Artelt, A., and Hammer, B. (2022b). Let's go to the alien zoo: Introducing an experimental framework to study usability of counterfactual explanations for machine learning. *arXiv preprint arXiv:2205.03398*.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K.-R. (2019). Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1–8.

Lee, J. A. and Verleysen, M. (2007). *Nonlinear dimensionality reduction*, volume 1. Springer.

Looveren, A. V. and Klaise, J. (2019). Interpretable counterfactual explanations guided by prototypes. *CoRR*, abs/1907.02584.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Molnar, C. (2019). *Interpretable Machine Learning*.

Mothilal, R. K., Sharma, A., and Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617.

N/A (1994). Diabetes data set. https://www4.stat.ncsu.edu /~boos/var.select/diabetes.html.

Offert, F. (2017). "i know it when i see it". visualization and intuitive interpretability.

parliament, E. and council (2016). General data protection regulation: Regulation (eu) 2016/679 of the european parliament.

Pearl, J. (2010). Causal inference. *Causality: objectives and assessment*, pages 39–58.

Ribera, M. and Lapedriza, A. (2019). Can we do better explanations? a proposal of user-centered explainable ai. In *IUI Workshops*, volume 2327, page 38.

Rodriguez, P., Caccia, M., Lacoste, A., Zamparo, L., Laradji, I., Charlin, L., and Vazquez, D. (2021). Beyond trivial counterfactual explanations with diverse valuable explanations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1056–1065.

Russell, C. (2019). Efficient search for diverse coherent explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 20–28.

Samek, W., Wiegand, T., and Müller, K. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *CoRR*, abs/1708.08296.

Schulz, A., Gisbrecht, A., and Hammer, B. (2014). Relevance learning for dimensionality reduction. In *ESANN*, pages 165–170. Citeseer.

Schulz, A. and Hammer, B. (2015). Metric learning in dimensionality reduction. In *ICPRAM (1)*, pages 232–239.

Schulz, A., Hinder, F., and Hammer, B. (2021). Deepview: visualizing classification boundaries of deep neural networks as scatter plots using discriminative dimensionality reduction. In *Proceedings of IJCAI*, pages 2305–2311.

Tjoa, E. and Guan, C. (2019). A survey on explainable artificial intelligence (XAI): towards medical XAI. *CoRR*, abs/1907.07374.

Van Der Maaten, L. (2009). Learning a parametric embedding by preserving local structure. In *Artificial intelligence and statistics*, pages 384–391. PMLR.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Van Der Maaten, L., Postma, E., Van den Herik, J., et al. (2009). Dimensionality reduction: a comparative. *J Mach Learn Res*, 10(66-71):13.

Verma, S., Dickerson, J., and Hines, K. (2020). Counterfactual explanations for machine learning: A review.

Vrachimis, S. G., Eliades, D. G., Taormina, R., Ostfeld, A., Kapelan, Z., Liu, S., Kyriakou, M., Pavlou, P., Qiu, M., and Polycarpou, M. M. (2020). Battledim: Battle of the leakage detection and isolation methods.

Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841.

William H. Wolberg, W. Nick Street, O. L. M. (1995). Breast cancer wisconsin (diagnostic) data set. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic).

APPENDIX

Results of the Empirical Evaluation

Table 5: Ranking of results from Table 6 – counting the number of metrics where the method yields the best score, best scores are highlighted in **bold-face**.

	DataSet	Algo 1	ModelAgnos
Linear	Diabetes	**3/4**	1/4
	Breast cancer	**3/4**	1/4
	Toy	**3/4**	1/4
SOM	Diabetes	**2/4**	**2/4**
	Breast cancer	**2/4**	**2/4**
	Toy	**2/4**	**2/4**
AE	Diabetes	**2/4**	**2/4**
	Breast cancer	**2/4**	**2/4**
	Toy	**2/4**	**2/4**
t-SNE	Diabetes	**3/4**	1/4
	Breast cancer	**2/4**	**2/4**
	Toy	**3/4**	1/4

Table 6: Quantitative results: *Gaussian perturbation* – all numbers are rounded to two decimal places, best scores are highlighted in **bold-face**.

	DataSet	CfSparse ↓		CfDiv ↓		CfDist ↓		Recall ↑	
		Algo 1	ModelAgnos	Algo 1	ModelAgnos	Algo 1	ModelAgnos	Algo 1	ModelAgnos
Linear	Diabetes	**0.22 ± 0.0**	0.55 ± 0.0	**0.0 ± 0.0**	7.14 ± 0.93	**0.57 ± 2.23**	4.24 ± 49.13	0.7 ± 0.07	**0.93 ± 0.02**
	Breast cancer	**0.09 ± 0.0**	0.66 ± 0.0	**0.0 ± 0.0**	29.43 ± 1.03	**0.05 ± 0.21**	1.35 ± 1.7	0.36 ± 0.09	**1.0 ± 0.0**
	Toy	**0.22 ± 0.0**	0.67 ± 0.0	**0.0 ± 0.0**	10.0 ± 0.0	**0.16 ± 0.62**	2.99 ± 12.53	0.57 ± 0.12	**1.0 ± 0.0**
SOM	Diabetes	0.79 ± 0.05	**0.6 ± 0.01**	**0.0 ± 0.01**	9.15 ± 19.24	**0.02 ± 0.2**	3.84 ± 3.93	0.92 ± 0.03	**0.94 ± 0.02**
	Breast cancer	0.6 ± 0.07	**0.66 ± 0.0**	**0.01 ± 0.2**	29.66 ± 0.9	**0.28 ± 3.74**	3.87 ± 14.26	0.95 ± 0.02	**1.0 ± 0.0**
	Toy	0.74 ± 0.06	**0.68 ± 0.0**	**0.0 ± 0.01**	10.55 ± 10.66	**0.01 ± 0.09**	3.68 ± 3.82	0.9 ± 0.04	**1.0 ± 0.0**
AE	Diabetes	**0.14 ± 0.03**	0.51 ± 0.0	**0.0 ± 0.01**	6.23 ± 0.69	1.09 ± 5.81	**1.07 ± 3.78**	0.43 ± 0.24	**0.91 ± 0.03**
	Breast cancer	**0.04 ± 0.01**	0.65 ± 0.0	**0.0 ± 0.03**	28.98 ± 0.94	0.48 ± 0.51	**0.38 ± 0.22**	0.11 ± 0.09	**1.0 ± 0.0**
	Toy	**0.14 ± 0.03**	0.67 ± 0.0	**0.0 ± 0.01**	9.98 ± 0.02	0.99 ± 3.21	**0.5 ± 0.99**	0.42 ± 0.24	**1.0 ± 0.0**
t-SNE	Diabetes	**0.33 ± 0.0**	0.56 ± 0.0	**0.01 ± 0.07**	7.78 ± 1.11	4.69 ± 11.37	**3.78 ± 8.53**	**1.0 ± 0.0**	0.93 ± 0.02
	Breast cancer	**0.32 ± 0.0**	0.66 ± 0.0	**0.05 ± 0.78**	29.6 ± 0.97	1.94 ± 1.99	**1.41 ± 0.64**	0.96 ± 0.04	**1.0 ± 0.0**
	Toy	**0.33 ± 0.0**	0.67 ± 0.0	**0.0 ± 0.02**	10.0 ± 0.0	2.16 ± 1.37	**1.52 ± 0.85**	**1.0 ± 0.0**	**1.0 ± 0.0**