Metrics to Ascertain the Plausibility and Faithfulness of Counterfactual Explanations

Ali Faruk Yücel[1]

Supervisor(s): Prof. Cynthia Liem[1]**, Patrick Altmeyer**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Ali Faruk Yücel
Final project course: CSE3000 Research Project
Thesis committee: Prof. Cynthia Liem, Patrick Altmeyer, Prof. Bernd Dudzik

Abstract

Counterfactual Explanations (CE) are essential for understanding the predictions of black-box models by suggesting minimal changes to input features that would alter the output. Despite their importance in Explainable AI (XAI), there is a lack of standardized metrics to assess the plausibility and faithfulness of these explanations. This paper reviews evaluation procedures in literature and proposes novel formal metrics for evaluating the plausibility and faithfulness of counterfactual explanations, addressing the existing limitations. Plausibility is defined as the coherence of explanations with the true data-generating process, while faithfulness refers to the accuracy of explanations in representing the model's reasoning. We discuss the shortcomings of existing evaluation procedures and metrics for measuring plausibility and faithfulness and consequently compare our proposed metrics with existing ones, highlighting their advantages and disadvantages. The proposed metrics are then empirically validated through experiments across multiple models and datasets, demonstrating their model-agnostic nature and reliability. Our findings indicate that the proposed metrics provide a correct and reliable means to quantify the plausibility and faithfulness of counterfactual explanations, thereby allowing one to gauge their feasibility and trustworthiness consistently.

1 Introduction

Counterfactual Explanations (CEs) are "what if" scenarios that allow for better understanding of the behavior of black-box models. These explanations suggest minimal changes to input features that would alter the output. Within the ontology of Explainable AI (XAI), CE is classified as a *post-hoc*, and *model-agnostic* explainability method with *local scoop* under *example-based explanations* [1].

A core focus of CEs is *outcome explanation problem* rather than *model explanation, model inspection, or transparent-box design problems* [2].

The importance of counterfactual explanations lies in their ability to provide actionable insights to end-users, allowing them to understand and potentially contest decisions made by AI systems [3]. However, the challenge remains in quantifying the plausibility and faithfulness of these explanations.

Despite the growing body of research that produces explainability methods, there have been few works (less than 5% of studied papers in the cited survey [1]) on evaluating these methods and quantifying their relevance. There is a notable gap in standardized metrics and robust evaluation frameworks that can universally assess the plausibility and faithfulness of counterfactual explanations across different domains. Therefore, the "primary target of future work should be developing formalized, rigorous evaluation metrics". [1] With that in mind, this paper focuses on developing evaluation metrics to measure the *plausibility* and *faithfulness* of counterfactual explanations which are critical aspects that determine their *feasibility* and *trustworthiness* respectively.

Colloquially, a counterfactual explanation is considered *plausible* if it is coherent with human reasoning and understanding [4]. Furthermore, it is considered *faithful* if it accurately represents the reasoning of the underlying model.

The existing definition of plausibility does not specify what constitutes coherence with human reasoning and understanding. It can be argued that the plausibility of an explanation must inherently reflect the faithfulness of that explanation to the underlying model's decision-making process if it is to be considered trustworthy. Otherwise, a plausible but unfaithful explanation would be no different than a con-artist; persuasive, but its logic falls apart once the facts and logical implications between them are examined more closely.

An explanation is only necessary when an unexpected outcome occurs. This indicates an inability to uncover the causality relation that lead to the unexpected outcome. Therefore, when talking about explanations, we inherently are talking about hidden causality relations. Humans instinctively evaluate explanations based on a hierarchical list of criteria that prioritize certain attributes over others. These criteria, listed in order of perceived importance, include [5]:

1. *Coherence (Logical Consistency):* The degree to which an explanation adheres to logical principles and is free from internal contradictions.
2. *Simplicity:* The simplicity or parsimony of an explanation, where simpler explanations are often preferred.
3. *Generality:* The applicability of the explanation across different contexts or situations.
4. *Truthfulness:* The factual accuracy of the explanation, reflecting its alignment with reality.
5. *Probability:* The likelihood of the explanation being true based on antecedents.

In other words, a coherent and simple explanation might be chosen over a truthful but not so coherent one by humans when uncovering causality relations. As important as simplicity is, the two most important factors when it comes to uncovering *causality relations* are *truthfulness* and *logical consistency*, with truthfulness taking precedence.

The aforementioned list of *instinctual evaluation mechanisms* also support the claim that, as the plausibility of explanations were prioritized in the CE generation process, the faithfulness of these explanations to the underlying model declined [4]. The term "truthfulness" in this context refers to the faithfulness of an explanation, which denotes its accuracy in representing the model's reasoning processes. Similarly, the term "coherence" refers to the plausibility of an explanation. Given that truthfulness does not rank highly on instinctual evaluations, it is fair to assume that the more appealing an explanation is to the general public, the

higher the weight of *coherence*, *simplicity* and *generality* over *truthfulness*.

To bridge the preceding discussion to the bulk of our research, we introduce the concept of a data manifold. In literature, a data manifold typically refers to the underlying structure of a data distribution. In our work, we refine this definition to address two separate distributions. For plausibility evaluation, data manifold refers to the true conditional distribution of samples that belong to a selected target class. For faithfulness evaluation, it refers to the learned conditional distribution of samples belonging to that same target class. In essence, plausibility and faithfulness evaluation is about quantifying the degree of closeness of a counterfactual to the appropriate data manifold. Accompanied by these definitions we present the objective of our paper.

The aim of this paper is to holistically analyse and decide **how to evaluate the plausibility and faithfulness of counterfactual explanations.** To that end, we answer three sub-questions:

1. *(SQ1) What are the shortcomings of methods to quantify the data manifold?*
2. *(SQ2) Which metrics are used to quantify the degree of closeness of a counterfactual to the data manifold?*
3. *(SQ3) What novel metrics could be used as proxies to estimate the degree of closeness of a counterfactual to the data manifold?*

Evaluation of plausibility and faithfulness of a CE is systematically analyzed through a three-step process in our work:

1. Determination of boundary values for plausibility and faithfulness by examining the generator of the CE
2. Quantification of the data manifold
3. Quantification of the degree of closeness to the data manifold using metrics

The most important parts for evaluation lie primarily in steps 2 and 3. Although our research touches upon the possible effect of CE generators putting an inherent limit on how plausible or faithful a CE can be, the primary focus of our research questions is directed towards quantification methods of the data manifold and metrics to quantify the degree of closeness to that data manifold.

Our contribution to the current body of knowledge is twofold. First, we conduct a critical literature review, identifying shortcomings of existing methods to quantify the data manifold and current metrics utilized to assess plausibility and faithfulness. Second, we propose novel metrics for evaluating the plausibility and faithfulness of counterfactual explanations and empirically validate these metrics through experiments across multiple datasets and models, demonstrating their reliability in various scenarios.

Overall, our paper is structured as follows: Section 3 reviews the related work and theoretical background, setting the stage for our contributions. Section 2 classifies the methodology types employed in this paper, followed by Section 4, where the process of development for the proposed evaluation metrics is detailed. Section 5 presents the experimental setup and results. Section 6 discusses the implications of our findings. Section 8 touches upon the methods utilized such that this research is replicable. Finally, Section 9 concludes the paper, summarizing the key points and contributions, and ultimately outlining potential future research directions.

2 Methodology

This work employs an amalgamation of quantitative and qualitative strategies to address the research questions. The study design is approximately distributed as one half quantitative strategies and one half qualitative strategies. The methodology is divided into two parts: a critical literature review [6] and a theoretical study focusing on the development of proposed evaluation metrics with empirical support. SQ1 and SQ2 are answered as part of Section 3, whereas SQ3 is answered with Section 4.

This work follows standardized guidelines [7], and SALSA [6] framework alongside the *snowball sampling method* to filter out important papers. A critical review was conducted to isolate the most significant papers that are relevant to our research questions.

As per the definition of a critical literature review, we searched for the most significant methods and metrics within the field related to our work. For appraisal, we mainly evaluated papers according to their contribution rather than a comprehensive quality assessment. The synthesis is narrative and conceptual and does not include tabular accompaniment since it is not a systematic review [6]. Lastly, our analysis involved indicating the consensus (or lack thereof) among the surveyed papers, along with prevalent themes between methods and metrics.

To get a comprehensive coverage, we initiated our search from Google Scholar. Moving forward, we have utilized an extensive list of databases including IEEE Xplore, Scopus, Web of Science, DBLP, O'Reilly for Higher Education, and arXiv. Furthermore, we consulted specific academic journals and proceedings of conferences to ensure alignment with the thematic focus of our study.

3 Related Work

Plausibility and faithfulness evaluation of a CE is related to three main factors. The first of these factors is the CE generation method as this may set an upper or lower bound on the plausibility or faithfulness of a CE, thereby possibly rendering further evaluation unnecessary. The second, is the quantification of true and learned conditional distributions, which are respectively utilized in the assessment of plausibility and faithfulness. Lastly, the third factor is the metric to compute the distance to the true or learned distributions. In this section, we define the necessary terms for our analysis and then present a critical literature review on the influence of the first and second factors on plausibility and faithfulness assessment along with existing metrics that are currently in use for the third factor.

3.1 Preliminaries

Formal definitions necessary for our analysis are given below.

Definition 1 (Counterfactual explanation)**.** Given a classifier b that outputs the decision $y = b(x)$ for an instance x, a counterfactual explanation consists of an instance x' such that the decision for b on x' is different from y, i.e., $b(x') \neq y$ (we take $b(x') = y^+$), and such that the difference between x and x' is *minimal* [8], i.e. a solution to the optimization Equation 1, where d is a distance function. Converting the objective into a differentiable, unconstrained form yields two terms [9], see Equation 2.

$$\arg\min_{x'} d(x, x') \text{ subject to } b(x') = y' \tag{1}$$

$$\arg\min_{x'} \max_{\lambda} \lambda(b(x') - y')^2 + d(x, x') \tag{2}$$

Counterfactual explanations are often plainly called counterfactuals. CEs often serve a dual function: they explain the outcomes of black-box models and provide an algorithmic method of recourse for individuals seeking to improve their attained outcomes.

Definition 2 (Counterfactual explainer)**.** A counterfactual explainer is a function f_k that takes as input a classifier b (often a black-box model), a set X of known instances, and a given instance of interest x. With its application $C = f_k(x, b, X)$, it returns a set $C = \{x'_1, \ldots, x'_h\}$ of $h \leq k$ valid counterfactual examples, where k is the number of counterfactuals requested [8].

Most of the counterfactual explainers (i.e. CE generators, which is the term we prefer to use) in the literature are designed as f_1 functions to return a single valid counterfactual. If $C = \emptyset$, i.e., $h = 0$, it means that the explainer was not able to find any valid counterfactual [8].

Definition 3 (Plausibility)**.** Plausibility is the degree to which generated counterfactuals adhere to the true data-generating process (DGP). Formally, let $\mathcal{X} \mid \mathbf{y}^+ = p(\mathbf{x} \mid \mathbf{y}^+)$ denote the true conditional distribution of $\mathbf{x}$ (samples) in the target class $\mathbf{y}^+$. For $\mathbf{x}'$ to be considered a plausible counterfactual, we need $\mathbf{x}' \sim \mathcal{X} \mid \mathbf{y}^+$ [10].

Take note that the formal definition of plausibility quantifies the colloquial definition, i.e. coherence with human reasoning and understanding by measuring compliance with the true data-generating process (DGP).

REVISE In Arnaud et. al.'s paper, 'realism' has an equal definition to 'plausibility' and so 'realistic' counterfactuals are synonymously used in place of 'plausible' counterfactuals.

The FACE paper adopts the term 'feasible' when talking about counterfactuals that are 'realistic' to achieve. Even though no formal definition of 'feasible' is given in the paper, the synonymous use of 'coherence with the underlying data distribution' clearly indicates that the definition of 'feasible' is the same as 'plausible' and it is transitively the same as 'realistic'.

For instance, an example of a 'closest possible' CE that is not feasible would be a customer whose loan application has been rejected would (probably) disregard a counterfactual explanation conditioned on him being 10 years younger.

Definition 4 (Faithfulness)**.** Faithfulness is the degree to which counterfactuals are consistent with what the model has learned about the data. Formally, let $\mathcal{X}_\theta \mid \mathbf{y}^+ = p_\theta(\mathbf{x} \mid \mathbf{y}^+)$ denote the learned conditional distribution of $\mathbf{x}$ (samples) in the target class $\mathbf{y}^+$, where θ denotes the parameters of model M_θ. For $\mathbf{x}'$ to be considered a faithful counterfactual, we need $\mathbf{x}' \sim \mathcal{X}_\theta \mid \mathbf{y}^+$ [10].

Necessary Conditions for Evaluation

Evaluation of plausibility and faithfulness of counterfactuals both necessitate the quantification of some conditional distribution of samples in the target class ($\mathcal{X}_s \mid \mathbf{y}^+$, $s \in \{\theta, \epsilon\}$, ϵ signifies the empty string). For plausible counterfactuals, quantification of the true conditional distribution $p(\mathbf{x} \mid \mathbf{y}^+)$ is required, whereas, for faithful counterfactuals, quantification of the learned, posterior conditional distribution $p_\theta(\mathbf{x} \mid \mathbf{y}^+)$ is required.

Once distributions are quantified, the objective is to decide whether or not a counterfactual $\mathbf{x}'$ is sampled from that distribution. More importantly, it is imperative that the proposed metric allows for comparison with regards to the degree of plausibility or faithfulness of two counterfactuals. In essence, a proxy for estimating the distance of counterfactual $\mathbf{x}'$ to the *data manifold* ($\mathcal{X} \mid \mathbf{y}^+$ or $\mathcal{X}_\theta \mid \mathbf{y}^+$) is desired.

3.2 Methods to Quantify the Data Manifold & Their Shortcomings

Estimation of the true or learned conditional distribution is typically made as part of a CE generator. In this section, we elaborate on the methods of quantification of the data manifold that various CE generators utilize with the aim of producing plausible or faithful counterfactuals.

Plausibility: Quantify the True Distribution

When it comes to quantifying the true conditional distribution of samples and subsequently producing plausible counterfactuals, the FACE paper [11] has significant contributions. The method proposed in this paper constructs a graph over data points with edge weights determined by density-weighted metrics. To then find 'feasible' counterfactuals, which are counterfactuals that come from high-density regions in the underlying data distribution, making them 'plausible', it calculates the shortest path distances which were defined based on one of three types of density-weighted metrics: KDE, k-NN, or ϵ-graph [11].

Joshi et. al. (2019) propose using a Variational Autoencoder (VAE) to traverse a latent embedding that condenses the DGP instead of searching for CEs in feature space [12]. Methods that rely on surrogate models to estimate the input data distribution are able to generate plausible but not necessarily faithful counterfactuals[10].

In contrast, Schut et al. (2021) propose minimizing predictive uncertainty to generate plausible counterfactuals without explicitly modeling the input distribution [13]. This method assumes that the black-box model provides well-calibrated predictive uncertainty estimates. Unlike surrogate model-based methods, black-box methods operate directly on the true input distribution without attempting to model or codify it. However, these methods

necessitate predictive uncertainty estimates, which can be computationally expensive and impractical for most deep learning models.

Faithfulness: Quantify the Learned Distribution
Quantification of the posterior conditional distribution requires the utilization of the black-box model that learns the dataset. In Altmeyer et al.'s work [10], a method from energy-based modelling called Stochastic Gradient Langevin Dynamics (SGLD) is used to quantify $p_\theta(\mathbf{x} \mid \mathbf{y}^+)$. Energy-Constrained Conformal Counterfactuals (ECCCo) is introduced in their work. By leveraging energy-based modeling and conformal prediction, ECCCo generates counterfactuals that are faithful to the model's learned behavior and plausible when appropriate. The paper shows that for models with accessible gradients, ECCCo can achieve state-of-the-art performance without relying on surrogate models by leveraging properties defining the black-box model itself [10].

3.3 Closeness to Data Manifold

Instead of a single objective optimization problem, if we formulate the search for CEs as a multi-objective optimization problem we can use metrics that represent closeness to the data manifold and by minimizing these metrics we can arrive at plausible or faithful counterfactuals.

For optimization-based CE generation, there are two ways to *ensure* closeness of a generated CE to the data manifold which follow as a corollary from the definitions of a CE and the data manifold:

1. The optimization function used to generate CEs (Equation 2) is updated with an addition of a loss function such that it prefers counterfactuals within the data manifold. We update Equation 3 with the addition of $l(x'; \mathbf{X}_{y+}) = l(x'; \mathcal{X}|\mathbf{y}^+)$ to represent the loss function.

$$\arg\min_{x'} \max_{\lambda} \lambda(b(x')-y')^2+d(x,x')+l(x';\mathbf{X}_{y+}) \quad (3)$$

2. The optimization function is not altered but generated counterfactuals are filtered with a metric that quantifies their closeness to the data manifold.

The loss function from the first item is the same as the metric from the second item. This metric represents a function that calculates the distance of a counterfactual to a data manifold.

Metrics to Quantify Closeness to Data Manifold
IM1 and IM2 metrics [14] were introduced by Arnaud et. al. to measure the *realism* of a counterfactual, which has the same formal definition as *plausibility*. IM1 is preferred over IM2 because IM2 scores are not significantly different for "out-of-distribution" data than in-distribution data [13]. IM1 is calculated as the ratio of reconstruction errors of two autoencoders.

$$\text{IM1} = \frac{\|x' - \text{AE}_{y'}(x')\|_2^2}{\|x' - \text{AE}_y(x')\|_2^2 + \epsilon} \quad (4)$$

Where x' is the counterfactual instance, $AE_{y'}$ is an autoencoder trained only on instances of the counterfactual class y', and AE_y is an autoencoder trained only on instances of the original class y. A lower value for IM1 signifies that x' can be better reconstructed by the autoencoder which is trained on the counterfactual class y' than by the autoencoder that has only seen instances of the original class y [14]. This implies that x' resides closer to the data manifold of counterfactual class y' compared to y, hence a lower value is considered more realistic/plausible.

The overarching name for the criteria used to measure faithfulness are commonly called erasure-based criteria. These criteria systematically remove or 'erase' parts of the input data until the model's output is changed. The process of removing parts of the input unfortunately increases the likelihood of input data to fall out of the distribution that the model was trained on, and may result in inaccurate faithfulness evaluation. The *comprehensiveness* and *sufficiency* scores are introduced [15] as formal generalizations of erasure-based criteria.

$$\text{Comp.} = \frac{1}{N}\sum_{i=1}^{N}\left(p(y_i \mid x_i) - p(y_i \mid x_i \setminus e_i)\right) \quad (5)$$

$$\text{Suff.} = \frac{1}{N}\sum_{i=1}^{N}\left(p(y_i \mid x_i) - p(y_i \mid e_i)\right) \quad (6)$$

Where e_i are selected features of x_i and $x_i \setminus e_i$ is the remaining features in x_i after features in e_i are deleted.

One of the metrics provided in literature to measure plausibility is the inverse of the implausibility metric given in Equation 7. Faithfulness is similarly measured by the inverse of the unfaithfulness metric given in Equation 8. Distance function in both of these metrics is L2-norm (Euclidean distance).

$$\text{impl}(\mathbf{x}', \mathbf{X}_{y+}) = \frac{1}{|\mathbf{X}_{y+}|}\sum_{\mathbf{x}\in\mathbf{X}_{y+}} \text{dist}(\mathbf{x}', \mathbf{x}) \quad (7)$$

$$\text{unfaith}(\mathbf{x}', \hat{\mathbf{X}}_{\theta,y+}) = \frac{1}{|\hat{\mathbf{X}}_{\theta,y+}|}\sum_{\mathbf{x}\in\hat{\mathbf{X}}_{\theta,y+}} \text{dist}(\mathbf{x}', \mathbf{x}) \quad (8)$$

Where $\mathbf{X}_{y+}$ represents the true conditional distribution of samples that are in target class y^+ and $\hat{\mathbf{X}}_{\theta,y+}$ represents the estimated posterior distribution using Stochastic Gradient Langevin Dynamics (SGLD) in Altmeyer et al.'s work. [10] The novel metrics we introduce aim to measure plausibility and faithfulness of a CE based on Definitions 3 and 4 and not approximations of Equations 7 and 8.

Diffusion Distance is derived from diffusion maps, which are commonly used for dimensionality reduction and feature extraction. [16]. It measures the *connectivity* between points in a dataset. It is robust to noise, and can handle non-linear manifolds. However, quite recently, diffusion distance has been applied as a loss function to generate plausible counterfactuals [17]. Therefore, we have decided to conduct our experiments with an unexplored method.

4 Metrics to Ascertain the Plausibility and Faithfulness of Counterfactuals

We propose two approaches to measuring plausibility and faithfulness that follow logically from their formal definitions. The first one is a direct corollary from the formal definitions. It is to use a Goodness-of-Fit test which verifies whether a counterfactual instance is sampled from a distribution. The second option is to define proxy metrics to estimate the distance of a counterfactual to the data manifold. The most appropriate of these metrics, which was determined to be the Local Outlier Factor, is evaluated (following the steps described below) on selected datasets and counterfactuals produced by the referred CE generators in Section 5.

4.1 Goodness-of-Fit Test

By applying a Goodness-of-Fit test, we can assess if a counterfactual is likely to have been sampled from the true or learned conditional distribution, thus providing a measure of its plausibility or faithfulness.

Among the Goodness-of-Fit tests, the Kolmogorov-Smirnov (K-S) test is particularly advantageous for our work because it does not rely on parametric assumptions about the underlying distribution of the data, if it were, this would have lead to issues such as the inability to conclude the correctness of a model if the null hypothesis is not rejected. Specifically, for parametric tests, "if we reject the null hypothesis, we conclude that the model [in our case, the distribution being tested for goodness-of-fit] should not be used [in other words, the counterfactual does not come from the distribution]. However, if we do not reject the null hypothesis, we cannot definitively conclude that the model is correct" [18]; it could simply be that the test did not have enough power to detect a difference. Additionally, the K-S test is less biased for moderate sample sizes and light-tailed distributions, more sensitive to deviations from the center of a distribution [19], and its critical values are distribution-free [20] as opposed to the Anderson-Darling (AD) test.

Proposal: Kolmogorov-Smirnov (K-S) Test

The K-S test assesses whether a sample comes from a specified distribution or whether two samples come from the same distribution. It can be utilized to evaluate the similarities of the distribution of individual features of a counterfactual with the rest of the data points in the manifold. The procedure for applying the K-S test is given below:

One-Sample K-S Test: The one-sample K-S test compares the empirical distribution function (EDF) of a sample (often a set of counterfactual instances) with the cumulative distribution function (CDF) of a reference distribution (often the true distribution). Implemented as follows: Define **Null Hypothesis** H_0 and **Alternative Hypothesis** H_1. For instance, to test for plausibility: H_0 = The counterfactual comes from the true conditional distribution $(X|y+)$. $H_1 = \neg H_0$.

The K-S statistic D is defined as the least upper bound (supremum) between the EDF $F_n(x)$ and the CDF $F(x)$:

$$D = \sup_x |F_n(x) - F(x)|$$

where $F_n(x)$ is the EDF of the sample and $F(x)$ is the CDF of the reference distribution.

Two-Sample K-S Test: The only difference between the one-sample and two-sample is instead of the CDF $F(x)$, a second EDF is used which is calculated with the second sample, called $G_m(x)$:

$$D_{n,m} = \sup_x |F_n(x) - G_m(x)|$$

Determine the P-value Once the type of test is determined, the test statistic D is compared with the critical value from the K-S distribution table or the p-value is calculated. If the p-value is less than the significance level (e.g., 0.05), reject the null hypothesis.

While this process may appear to be a multi-step evaluation procedure, it is fundamentally a single metric. The detailed description of the calculation steps serves to enhance clarity and facilitate practical implementation.

Case Specific Modifications As our data is composed of multivariate vectors, instead of checking each of the counterfactual vector's features against a univariate distribution, the correct way to calculate this test statistic is to check the vector as a whole against a multivariate distribution because we are interested in the joint distribution of all features rather than the marginal distribution of each feature independently.

4.2 Proxy Metrics

Proxy metrics are alternative methods used to estimate the closeness of a counterfactual to the data manifold. The Local Outlier Factor (LOF) is chosen as the proposed proxy metrics due to several key reasons: LOF considers the local density around each point, making it suitable for datasets with varying densities. By using reachability distances (RDs), LOF reduces the impact of statistical fluctuations for points that are close together. The statistical fluctuations refer to the small variations in distance in dense regions of a dataset, possibly resulting in noisy results when trying to identify outliers. By setting RD to the actual distance for sparse areas and defaulting RD to the k-distance for dense areas, LOF smooths out these small variations. It takes into account only k other instances, which brings about computational efficiency and robustness to outliers, as only 2k distances need to be calculated instead of distances to the entire dataset as done in Equations 7 and 8, and instances farther away do not unduly influence the average distance. Most crucially, LOF provides an outlier score rather than a binary label, allowing for quantification of distance to the data manifold.

Proposal: Modified Local Outlier Factor

We propose a new loss function based on Gower Distance (GD) and Local Outlier Factor (LOF) to quantify the closeness of a counterfactual to the data manifold. LOF is a commonly used outlier score that measures how unusual a given instance is by using k-nearest neighbour (k-NN) and our modification to LOF is detailed in the next subsection.

LOF is an anomaly detection algorithm that identifies outliers by comparing the local density of a data point to the local densities of its neighbors [21], which makes it particularly effective for identifying anomalies in datasets

with varying densities. The LOF algorithm operates in several key steps:

First, the Reachability Distance (RD) of a point p with respect to another point o is calculated as follows:

$$\text{RD}(p, o) = \max(k\text{-distance}(o), d(p, o))$$

where $d(p, o)$ is the actual distance between p and o. Often, Euclidean distance is utilized here as the distance metric. The k-distance(o), is the distance between o and its k-th nearest neighbor. This distance provides an estimate of the density around o. The Reachability Distance is used to mitigate the effects of statistical fluctuations for points that are close together.

The Local Reachability Density (LRD) of a point p is then calculated as the inverse of the average Reachability Distance from the k-nearest neighbors of p:

$$\text{LRD}(p) = \frac{|N_k(p)|}{\sum_{o \in N_k(p)} \text{RD}(p, o)}$$

where $N_k(p)$ is the set of k-nearest neighbors of p. The $|N_k(p)|$ is always greater than or equal to k. For $|N_k(p)|$ to be greater than k, two points must be exactly as far from the point p. In practice, for datasets with more than 3 features, this is rarely the case. Therefore, it is also acceptable to assume $|N_k(p)| = k$. A higher LRD value indicates a denser region.

Finally, the Local Outlier Factor (LOF) of a point p is computed as the ratio of the average Local Reachability Density of the k-nearest neighbors of p to the Local Reachability Density of p:

$$\text{LOF}(p) = \frac{\sum_{o \in N_k(p)} \text{LRD}(o)}{|N_k(p)| \, \text{LRD}(p)}$$

A LOF value around 1 indicates that the point p is in a region of similar density to its neighbors. A LOF value greater than 1 indicates that p has a lower density than its neighbors, making it an outlier candidate. The higher the LOF value, the more abnormal the point is considered to be. In our case, the point p is the counterfactual x'.

Proposal Modification: Gower Distance

Normally, LOF utilizes Euclidean distance (L2-Norm) as its distance function when calculating reachability distance and existing metrics from literature such as Equation 7 and 8 likewise use Euclidean distance to quantify plausibility and faithfulness of a counterfactual. There are a couple of important issues with this. First of all, Euclidean distance is highly sensitive to features with large values. This requires the data to be scaled appropriately to ensure that variables with larger rangers do not dominate the distance calculation, leading to biases. Unfortunately, the scaling process is sometimes overlooked when Euclidean distance is disguised as part of other metrics.

Secondly, Euclidean distance is designed for numerical data and cannot handle categorical or binary data types. This limitation necessitates the use of one-hot encoding for categorical data, increasing the dimensionality and sparsity of the dataset, which in turn exacerbates the Curse of Dimensionality [22]. The inflated space caused by one-hot encoding often makes the distance calculation less meaningful and more computationally intensive.

Thirdly, Euclidean distance does not handle missing values well. Missing values often require pre-processing steps to exclude the incomplete records, however, this is not as significant of an issue in practice as datasets are usually cleaned before conducting any experiments.

Lastly, Euclidean distance is meaningful only 'locally' [23]. As the dimensions of the data increase, the concept of proximity or nearest neighbour becomes less meaningful [24]. The distances between the nearest and farthest points tend to converge, making it difficult to differentiate between close and distant points. Data points that were close together in lower-dimensions become more separated as number of dimensions increase. Higher dimensions cause the volume of the space to grow exponentially, making the data points appear more isolated.

These are the main reasons why we propose utilizing Gower distance with the most significant one being the last. When combined with Local Outlier Factor algorithm, Gower distance addresses all of the aforementioned issues. Gower distance automatically normalizes the contribution of each variable to the overall distance calculation, ensuring that features with different scales do not dominate the calculation. By definition, it is able to handle mixed data types and it automatically excludes missing value pairs and scales the distance accordingly. That being said, the formal definition of Gower distance is as follows:

Gower distance d_G between two samples i and j is defined as:

$$d_G(i, j) = 1 - s_G(i, j)$$

where $s_G(i, j)$ is the Gower similarity coefficient, calculated as:

$$s_G(i, j) = \frac{\sum_{k=1}^{n} w_{ijk} s_{ijk}}{\sum_{k=1}^{n} w_{ijk}}$$

Here:

- n is the number of features.
- s_{ijk} is the similarity between the i-th and j-th samples for the k-th feature.
- w_{ijk} is a weight assigned to the k-th feature, which can be 0 or 1 depending on whether the feature is considered in the calculation.

The similarity s_{ijk} for each type of feature is calculated as follows. For numerical features:

$$s_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{R_k}$$

where R_k is the range of the k-th feature. For categorical features, the similarity is binary. If the categories match, the score is 1, otherwise 0. The scoring for binary features is the same as categorical features.

5 Experimental Setup

In this section, we detail the experimental procedures, the datasets utilized, the architecture of the deep learning models, the software packages employed for training, and the metrics applied for evaluation.

5.1 Hardware Setup

The experiments were conducted on a MacBook Pro with M2 Max Chip and 96GB of memory and several Kaggle Notebooks. It took the entire experimental loop per base model 35 ± 10 minutes to calculate implausibility scores and K-S test p-values. The loop took 55 ± 10 minutes for the LOF score calculation.

5.2 Datasets & Models

We have opted to use simpler deep models due to performance and time constraints. As our work is a combination of critical literature review and theoretical proposal with experimental validation as merely a part of the theoretical proposal, we did not have enough time to conduct experiments with more complex models. Consequently, state-of-the-art architectures such as SAINT or DeepFM [25] are not considered. The foundational models for all datasets in this study are composed of a single hidden layer consisting of fully connected perceptrons. For the hidden layer, the Rectified Linear Unit (ReLU) activation function is used, while the output layer utilizes the softmax function. Cross-entropy serves as the loss function for these models. This architecture was selected to align with the classification focus of our research, as per the requirements of a counterfactual explanation. Table 1 offers comprehensive details regarding the training procedures and architectural specifics of the Artificial Neural Networks (ANN).

For implementation, we use FLUX.JL for the ANNs. Our experiments were conducted on three datasets,each from a different real-life situation where the evaluation of plausibility and faithfulness of counterfactuals are crucial for algorithmic recourse: German Credit [26], California Housing [27], and Adult Census Income [28]. They are all part of the TAIJADATA.JL repository. The German Credit dataset provides 1000 samples with 700 samples with class 1, 300 with class 0. As this is not balanced we tried undersampling the majority class to mitigate biases of models (600 samples after undersampling). However, the average loss increased dramatically (from 0.45 on average for 10 models to 0.75) and as there are only 1000 samples in total, we decided not to undersample.

We concentrated on black-box models, as explainability is less of an issue for most other machine learning models as they tend to be quite transparent. The models utilized in our experiments include fully connected artificial neural networks (ANNs), and dropout neural networks.

For every dataset and every basis model, we have trained 10 different models for 10 epochs on the given number of samples.

5.3 Counterfactual Generators

We employed four counterfactual generators for our experiments: Generic, DiCE with $\lambda_2 = 0.5$ (denoted as *ddp diversity* penalty), DiCE with $\lambda_2 = 1$, and ClaPROAR.

DiCE (Diverse Counterfactual Explanations) generator is designed to produce diverse counterfactuals, which is crucial for evaluating the reliability of our metrics across different scenarios. The diversity-proximity trade-off in DiCE is managed by the parameter λ, where a higher λ accentuates diversity over proximity. Proximity in DiCE, defined as the closeness of a counterfactual to the original instance, is closely related to plausibility, as plausible counterfactuals should be close to the data manifold of instances belonging to the target class.

ClaPROAR generator uses a model loss penalty that calculates the loss between the model's prediction for the counterfactual instance and the target value. This penalty allows ClaPROAR to produce more faithful counterfactuals. In contrast, the proximity penalty in DiCE helps generate more plausible counterfactuals by increasing the likelihood that they are close to the data manifold.

1. Generic: A baseline generator
2. DiCE with $\lambda_2 = 0.5$. Balances diversity and proximity, enhancing the plausibility of counterfactuals.
3. DiCE with $\lambda_2 = 1$ Prioritizes diversity, potentially at the cost of proximity. However, allows us to test our metrics on counterfactual instances that embody a wide range of values for their features.
4. ClaPROAR (Classifier Preserving ROAR): Focuses on faithfulness by incorporating a *model loss penalty*, which makes it more likely that the generated counterfactuals are consistent with the model's learned behavior.

6 Results & Discussion

In this section the results from the experiments conducted with baseline and proposed metrics are presented and discussed. For our baseline metric we have opted to use Equation 7 to measure the plausibility values.

For the communication of scores, we have opted to round all reported scores to four significant figures. This level of precision was deemed appropriate given the relatively small differences observed between scores across experimental conditions.

The average LOF scores (representing plausibility or faithfulness, depending on which conditional distribution is used) are reported similar to implausibility scores.

Table 2 shows average implausibility scores, where lower values indicate more plausible counterfactuals. ClaPROAR achieves the lowest implausibility scores relatively frequently, suggesting it generates the most plausible counterfactuals. The Tables 3 and 4 report Local Outlier Factor (LOF) scores using Euclidean and Gower distances, respectively, where lower scores denote greater closeness to the data manifold. ClaPROAR shows the lowest LOF scores, reinforcing its effectiveness in generating plausible and faithful counterfactuals. The use of Gower distance, which handles mixed data types better than Euclidean distance, provides a more robust assessment of counterfactuals' adherence to the data manifold. Overall, ClaPROAR emerges as the most reliable method for generating plausible and faithful counterfactuals, while DiCE with $\lambda_2 = 0.5$ offers a good balance between diversity and proximity.

Dataset	Input	Hidden	Output	Activations	Epochs	Train Batch	Samples
German Credit	20	40	2	ReLU, Softmax	10	16	200
California Housing	8	12	2	ReLU, Softmax	10	16	200
Adult Income	14	28	2	ReLU, Softmax	10	16	300

Table 1: Structure and training parameters of the fully connected ANNs used for different datasets

Model	Generator	German Credit	California Housing	Adult Income
Neural Network	Generic	6.185 ± 0.290	3.274 ± 0.017	5.054 ± 0.025
	DiCE ($\lambda_2 = 0.5$)	6.181 ± 0.268	3.268 ± 0.024	5.059 ± 0.287
	DiCE ($\lambda_2 = 1$)	6.210 ± 0.038	3.264 ± 0.018	5.018 ± 0.035
	ClaPROAR	6.117 ± 0.017	3.274 ± 0.038	4.945 ± 0.044
Dropout	Generic	6.181 ± 0.442	3.278 ± 0.288	4.933 ± 0.184
	DiCE ($\lambda_2 = 0.5$)	6.181 ± 0.413	3.270 ± 0.237	4.921 ± 0.316
	DiCE ($\lambda_2 = 1$)	6.179 ± 0.042	3.248 ± 0.016	5.053 ± 0.017
	ClaPROAR	6.103 ± 0.015	3.177 ± 0.047	5.020 ± 0.057

Table 2: Average implausibility score per model and dataset for different generators. Dash (–) means computation was prohibited by circumstances

Model	Generator	German Credit	California Housing	Adult Income
Neural Network	Generic	0.972 ± 0.010	0.465 ± 0.010	0.702 ± 0.024
	DiCE ($\lambda_2 = 0.5$)	0.971 ± 0.010	0.462 ± 0.010	0.701 ± 0.027
	DiCE ($\lambda_2 = 1$)	0.971 ± 0.010	0.548 ± 0.023	0.722 ± 0.015
	ClaPROAR	0.991 ± 0.002	0.461 ± 0.107	0.796 ± 0.144
Dropout	Generic	0.975 ± 0.010	0.484 ± 0.019	0.694 ± 0.009
	DiCE ($\lambda_2 = 0.5$)	0.975 ± 0.010	0.480 ± 0.017	0.694 ± 0.010
	DiCE ($\lambda_2 = 1$)	0.974 ± 0.010	0.486 ± 0.166	0.714 ± 0.008
	ClaPROAR	0.992 ± 0.002	0.639 ± 0.194	0.783 ± 0.223

Table 3: Average LOF score (using L2 Norm as distance) per model and dataset for various generators. Dash (–) means computation was prohibited by circumstances

Model	Generator	German Credit	California Housing	Adult Income
Neural Network	Generic	0.972 ± 0.010	0.465 ± 0.009	0.709 ± 0.018
	DiCE ($\lambda_2 = 0.5$)	0.972 ± 0.010	0.464 ± 0.011	0.710 ± 0.017
	DiCE ($\lambda_2 = 1$)	0.972 ± 0.010	0.480 ± 0.009	0.702 ± 0.019
	ClaPROAR	0.991 ± 0.002	0.546 ± 0.022	0.770 ± 0.018
Dropout	Generic	0.975 ± 0.010	0.482 ± 0.020	0.703 ± 0.01
	DiCE ($\lambda_2 = 0.5$)	0.974 ± 0.010	0.483 ± 0.022	0.701 ± 0.009
	DiCE ($\lambda_2 = 1$)	0.974 ± 0.010	0.492 ± 0.021	0.700 ± 0.010
	ClaPROAR	0.992 ± 0.001	0.638 ± 0.016	0.765 ± 0.023

Table 4: Average LOF score (using Gower distance) per model and dataset for various generators. Dash (–) means computation was prohibited by circumstances

7 Limitations and Future Work

Initially, we intended to incorporate ensemble neural networks into our study. However, due to performance constraints and subsequent time limitations, we were unable to implement this approach. Our survey of over 20 packages in both Julia and Python revealed a lack of multivariate Kolmogorov-Smirnov (K-S) test implementations. This prompted us to develop our own. However, once again time constraints resulted in a rudimentary implementation, and we have reservations about the accuracy of the data gathered from this metric.

Additionally, we had considered using Euclidean distance for calculating the k-distance, given its local nature. However, we ultimately decided against this approach as the Euclidean distance is highly sensitive to large values in any given feature, which could potentially skew the results when taking the maximum value of two measurements (as is done when calculating Reachability Distance). It's worth noting that using Euclidean distance for k-distance within the Local Outlier Factor (LOF) algorithm may have lead to inaccuracies, particularly for features with large ranges. We strongly urge the reader to conduct further studies on how to implement a multivariate K-S test and further probe into utilizing IM1 scores to calculate faithfulness.

8 Responsible Research

A key ethical aspect in need of discussion is the use of real-world datasets. This approach ensures that our findings are relevant and applicable to real-world scenarios, but it also necessitates strict adherence to ethical guidelines regarding data privacy and consent. All datasets used in our experiments have MIT free licences and were anonymized to protect the privacy of individuals, and we ensured compliance with all relevant data protection regulations. The results of preprocessing and exploratory data analysis of the datasets were stored in folders which were excluded from version tracking.

To address potential biases and ensure fairness, we carefully curated our datasets to be representative and free from underlying prejudices. Undersampling was performed for the majority class for all datasets in order to balance the dataset and mitigate biases of the trained models.

The source code for generating counterfactual explanations and running the experiments is made publicly available on GitHub to ensure full reproducibility of our research. We also include detailed documentation of all parameter settings and experimental procedures in Section 5, allowing other researchers to replicate our study accurately and precisely. By averaging existing and proxy metric scores over multiple experimental runs we have further mitigated outliers and ensured that our results are reproducible by providing our serialized models.

We make sure to detail the computing process of each of our proposals with case specific modifications such that reproducibility is trivial. This methodical approach allows for precise quantification and reproducibility.

9 Conclusion

This study aimed to explore ways of evaluating the plausibility and faithfulness of counterfactual explanations (CEs) in machine learning models. The primary research questions addressed were: (1) identifying the shortcomings of existing methods to quantify the conditional distribution of a sample in the target class, (2) reviewing which existing metrics are used to quantify the degree of closeness of a counterfactual to the data manifold, and (3) proposing novel metrics as proxies to estimate the distance of a counterfactual to the data manifold.

Our research led to several key conclusions. First, we found that many existing methods of quantifying conditional distributions rely on surrogate models, which can produce plausible but not necessarily faithful explanations. This highlights a significant gap in ensuring that CEs accurately represent the underlying model's reasoning. Second, we proposed the use of the Kolmogorov-Smirnov (K-S) test and Local Outlier Factor (LOF) with Gower Distance as effective metrics for evaluating the plausibility and faithfulness of CEs. These metrics combined provide a good starting point for assessing how closely a counterfactual aligns with the true data distribution.

In order to address the limitations mentioned in the field of evaluation methods, we developed formal metrics to ascertain the plausibility and faithfulness of CEs, addressing a critical gap in the literature. The proposed LOF with Gower distance metric has addressed significant shortcomings as mentioned in Section 4.

Our work described a first step in proposing formal metrics and methods to evaluate the plausibility and faithfulness of CEs. Despite the contributions, several open issues and areas for improvement remain. Future research should explore methods to reduce dependency on surrogate models to quantify closeness to the data manifold, ensuring that CEs remain faithful to the original model. The computational cost of some proposed metrics, particularly those involving autoencoders, can be high. Optimizing these methods for efficiency without compromising accuracy is a crucial area for future work. While our metrics were validated on specific datasets, further research is needed to test their applicability across a wider range of datasets and model types. Incorporating user feedback also represents a critical area for further research. This approach can provide valuable insights into the practical utility and trustworthiness of CEs, ensuring that the evaluations of CEs are not only theoretically sound but also align with user evaluations and expectations.

In conclusion, this study proposes a new method and formal metrics for evaluating the plausibility and faithfulness of counterfactual explanations, contributing to the development of more feasible, transparent and trustworthy AI systems. Future research should continue to propose novel metrics, refine the formerly proposed metrics and explore new approaches to enhance the reliability and applicability of CEs in diverse real-world scenarios.

References

[1] A. Adadi **and** M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, **journvol** 6, **pages** 52 138–52 160, 2018. DOI: 10.1109/access.2018.2870052.

[2] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti **and** D. Pedreschi, "A survey of methods for explaining black box models," *ACM Computing Surveys*, **journvol** 51, **number** 5, **pages** 1–42, 2018. DOI: 10.1145/3236009.

[3] S. Wachter, B. Mittelstadt **and** C. Russell, *Counterfactual Explanations Without Opening The Black Box: Automated Decisions and the GDPR*, **url**: https://arxiv.org/pdf/1711.00399.

[4] C. A. Lakkaraju, "Faithfulness vs. plausibility: On the (un)reliability of explanations from large language models," **url**: https://arxiv.org/html/2402.04614v2.

[5] G. Weiss, *Multiagent Systems*. MIT Press, 2013.

[6] M. J. Grant **and** A. Booth, "A typology of reviews: An analysis of 14 review types and associated methodologies," *Health Information amp; Libraries Journal*, **journvol** 26, **number** 2, **pages** 91–108, **may** 2009. DOI: 10.1111/j.1471-1842.2009.00848.x.

[7] A. García-Holgado, S. Marcos-Pablos **and** F. J. García-Peñalvo, "Guidelines for performing systematic research projects reviews," *International Journal of Interactive Multimedia and Artificial Intelligence*, **journvol** 6, **page** 9, 2 2020. DOI: 10.9781/ijimai.2020.05.005.

[8] R. Guidotti, "Counterfactual explanations and how to find them: Literature review and benchmarking," *Data Mining and Knowledge Discovery*, 2022. DOI: 10.1007/s10618-022-00831-6.

[9] S. Verma, V. Boonsanong, M. V. Hoang, K. E. Hines, J. P. Dickerson **and** C. Shah, "Counterfactual explanations and algorithmic recourses for machine learning: A review," 2020. DOI: 10.48550/arxiv.2010.10596.

[10] P. Altmeyer, A. van Deursen **and** C. C. s. Liem, "Explaining black-box models through counterfactuals," *Proceedings of the JuliaCon Conferences*, **journvol** 1, **number** 1, **page** 130, 2023. DOI: 10.21105/jcon.00130. **url**: https://doi.org/10.21105/jcon.00130.

[11] R. Poyiadzi, K. Sokol, R. Santos-Rodríguez, T. D. Bie **and** P. A. Flach, "Face," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020. DOI: 10.1145/3375627.3375850.

[12] S. Joshi, O. Koyejo, W. Vijitbenjaronk, B. Kim **and** J. Ghosh, "Towards realistic individual recourse and actionable explanations in black-box decision making systems," 2019. DOI: 10.48550/arxiv.1907.09615.

[13] L. Schut, O. Key, R. McGrath **andothers**, "Generating interpretable counterfactual explanations by implicit minimisation of epistemic and aleatoric uncertainties," 2021. DOI: 10.48550/arxiv.2103.08951.

[14] A. V. Looveren **and** J. Klaise, "Interpretable counterfactual explanations guided by prototypes," 2019. DOI: 10.48550/arxiv.1907.02584.

[15] J. DeYoung, S. Jain, N. F. Rajani **andothers**, "Eraser: A benchmark to evaluate rationalized nlp models," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. DOI: 10.18653/v1/2020.acl-main.408.

[16] R. R. Coifman **and** M. Hirn, "Diffusion maps for changing data," *Applied and Computational Harmonic Analysis*, **journvol** 36, **pages** 79–107, 1 2014. DOI: 10.1016/j.acha.2013.03.001.

[17] M. Domnich **and** R. Vicente, *Enhancing counterfactual explanation search with diffusion distance and directional coherence*, **april** 2024. **url**: https://arxiv.org/abs/2404.12810.

[18] **in***All of Statistics: A Concise Course in Statistical Inference* Springer, **pages** 169–169.

[19] **url**: http://www.jaqm.ro/issues/volume-6,issue-3/pdfs/1_engmann_cousineau.pdf.

[20] *Anderson-darling and shapiro-wilk tests*. **url**: https://www.itl.nist.gov/div898/handbook/prc/section2/prc213.htm.

[21] V. Jayaswal, *Local outlier factor (lof)-algorithm for outlier identification*, **november** 2020. **url**: https://towardsdatascience.com/local-outlier-factor-lof-algorithm-for-outlier-identification-8efb887d9843.

[22] **url**: https://builtin.com/data-science/curse-dimensionality.

[23] **url**: https://www.kdnuggets.com/2020/03/diffusion-map-manifold-learning-theory-implementation.html.

[24] S. Victoroff, *Is euclidean distance meaningful for high dimensional data?* **september** 2021. **url**: https://indicodata.ai/blog/is-euclidean-distance-meaningful-for-high-dimensional-data/.

[25] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk **and** G. Kasneci, "Deep neural networks and tabular data: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, **journvol** 35, **pages** 7499–7519, 6 2024. DOI: 10.1109/tnnls.2022.3229161.

[26] H. Hofmann, *Statlog (German Credit Data)*, UCI Machine Learning Repository, DOI: https://doi.org/10.24432/C5NC77, 1994.

[27] *California Housing*, DOI: https://www.kaggle.com/datasets/camnugent/california-housing-prices, 1990.

[28] B. Becker **and** R. Kohavi, *Adult*, UCI Machine Learning Repository, DOI: https://doi.org/10.24432/C5XW20, 1996.