This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3410540

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.0322000

Evaluation of Instance-based Explanations: An In-depth Analysis of Counterfactual Evaluation Metrics, Challenges, and the CEval Toolkit

BETÜL BAYRAK and KERSTIN BACH
Department of Computer Science, Norwegian University of Science and Technology (NTNU), Høgskoleringen 1, Trondheim, 7034, Norway
Corresponding author: Betül Bayrak (e-mail: betul.bayrak@ntnu.no).

This work was supported by the Research Council of Norway through the EXAIGON project (ID 304843) and the SFI NorwAI (ID 309834).

ABSTRACT In eXplainable Artificial Intelligence (XAI), instance-based explanations have gained importance as a method for illuminating complex models by highlighting differences or similarities between the samples and their explanations. The evaluation of these explanations is crucial for assessing their quality and effectiveness. However, the quantitative evaluation of instance-based explanation methods reveals inconsistencies and variations in terminology and metrics. Addressing this, our survey provides a unified notation for instance-based explanation evaluation metrics for instance-based explanations with a particular focus on counterfactual explanations. Further, it explores associated trade-offs, identifies areas for improvement, and offers a practical Python toolkit, CEval. Key contributions include a comprehensive survey of quantitative evaluation metrics, facilitating practical counterfactual evaluation with the package, and providing insights into explanation evaluation limitations and future directions.

INDEX TERMS Explainable artificial intelligence (XAI), instance-based explanation, contrastive explanation evaluation, counterfactual explanation evaluation

I. INTRODUCTION

The popularity of machine learning methods has sped in recent years, driven by their remarkable capabilities to solve complex problems and make decisions autonomously. However, this wave in machine learning applications has also raised concerns about the fairness, accountability, and trustworthiness of these systems, urging the need for eXplainable Artificial Intelligence (XAI) methods. Numerous XAI techniques have been proposed to shed light on the inner workings of black-box models. Notably, Hvilshøj et al. [1] have categorized them into three primary groups: saliency, surrogate, and instance-based methods. This article focuses specifically on instance-based methods, with a particular emphasis on counterfactual explanations (counterfactuals, in short).

Counterfactuals, a subset of instance-based XAI methods, enhance the understanding of decisions made by models by providing alternative scenarios and causal insights. Meeting specific requirements, such as actionability, faithfulness, diversity, and interpretability, is crucial. In line with [2]–[4], we want to underscore the importance of measuring the quality of these explanations, allowing us to assess their effectiveness and identify strengths and weaknesses and evaluation metrics aid in method comparison and application suitability, enhancing trust and utility, empowering informed decisions, and mitigating potential biases or errors in machine learning models. More technical information about instance-based explanations (counterfactuals, semi-factuals, and alter-factuals) can be found in Section II.

In the ever-developing field of XAI, there exists a multitude of methods used for both qualitative and quantitative evaluations. Our primary focus in this article is the quantitative evaluation of counterfactual methods, while our broader scope encompasses instance-based methods. Within this domain, however, we have observed inconsistencies and variations in the terminology and metrics, which are closely intertwined. Consequently, our objective is to provide a framework to converge these metrics into a unified notation, explore trade-offs and their applicability, address areas open to improvement, and facilitate the practical use of existing metrics through the development of a Python package.

To achieve this, we conducted a comprehensive survey that gathers quantitative evaluation metrics and optimization

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3410540

methods, suitable for use in evaluating instance-based explanations. It is crucial to emphasize that our survey aims at gathering a comprehensive collection of evaluation metrics rather than prescribing specific ones tailored to particular scenarios. With the diverse nature of XAI applications and the varied contexts in which these techniques are deployed, our survey refrains from assigning hierarchical rankings or subjective assessments of the metrics under consideration. Instead, it proposes a user-centric approach to users, researchers, and practitioners in selecting metrics tailored to their requirements. This approach facilitates enhanced flexibility and adaptability for the evaluation process, aligning with the diverse needs of XAI research and practice.

This article makes several noteworthy contributions to the field:

- *Unified notation:* We provide a unified notation for the diverse set of metrics used to evaluate counterfactual explanations and gather them under a single framework, thereby enhancing clarity and consistency in the field.
- *Comprehensive survey:* We conducted a thorough review of 493 articles and included 66 that offer quantitative evaluation metrics and optimization methods, categorizing these metrics into two groups: those assessing the quality of a single explanation and those evaluating the overall quality of an explainer, providing a comprehensive overview of available evaluation metrics.
- *Python package:* We develop a Python package, CEval toolkit, that makes employing existing counterfactual metrics easier for researchers and practitioners in their work, promoting practicality and accessibility in counterfactual evaluation.
- *Exploration of trade-offs:* We explore the trade-offs inherent in various counterfactual metrics, offering insights into their strengths and limitations, which can aid researchers and practitioners in selecting appropriate evaluation methods for their specific needs.
- *Addressing areas for improvement:* Our article identifies areas within the existing counterfactual evaluation metrics that are open to refinement, providing valuable guidance for future research and development for both generating high-quality instance-based explanations and evaluation of their quality.

The rest of the article is organized as follows. Section II provides a general overview of instance-based explanations and a brief overview to related work. Section III presents evaluation techniques for counterfactual explainers in the literature, while Section IV extends the overview to other types of instance-based explainers. Section V details the provided Python package, and In Section VI, we outline the advantages and limitations of the evaluated metrics, addressing concerns, deficits, and areas open for improvement. Finally, Section VII concludes the article, summarizing the key findings.

II. BACKGROUND AND RELATED WORK

A. INSTANCE-BASED EXPLANATIONS/EXPLAINERS

Instance-based explainers can be defined as explainers that provide insights into the model's decisions by highlighting the differences or similarities between the decisions and relevant instances, aiming to facilitate a clearer understanding of machine learning model behavior. Considering the literature, we can group instance-based explainers under three subgroups, Counterfactual Explainers (CE), Semifactual Explainers, and Alterfactual Explainers.

Assuming $f()$ is the prediction function, X is a set of instances to explain, and $x \in X$. $x = \langle a_1, ..., a_\rho \rangle$ while ρ is the number of features, and $y = f(x)$. $explain(x)$ is the explainer function that returns e, a set of explanations for x. $e_i \in e$ and $e_i = \{x'_i, z_i\}$ while it is a single explanation of x, where $z_i = f(x'_i)$.

1) Counterfactual Explanations

Counterfactual explanations (counterfactuals) involve the generations of hypothetical scenarios to explain the behavior of a machine learning model by providing meaningful and actionable guidance [2]. A counterfactual offers an understanding of the interpretation behind a particular prediction by proposing a hypothetical scenario, identifying minimal changes in input features necessary to change the model's outcome [5], [6]. Counterfactuals are particularly valuable in sensitive applications like healthcare or finance, where understanding and justifying model decisions is critical for trust and regulatory compliance. Various techniques have been developed in the literature to create meaningful and informative counterfactual explanations for different types of machine learning models.

In short, a counterfactual is an explanation sample that provides a hypothetical scenario to sample x while assuring $y' \neq y$.

"If you were to increase your income by an additional $500, your application would be approved."

2) Semi-factual Explanations

Similar to counterfactual explanations, semi-factual explanations aim to offer a more comprehensive understanding of model behavior, making them a promising avenue for enhancing transparency in complex machine learning systems; however, while counterfactuals propose explanations to answer the question *"what if..."*, semi-factuals use *"even if..., still ..."* kind of explanations. A semi-factual is an explanation sample that provides a hypothetical scenario to sample x while assuring there are changes in the attributes and $y' = y$.

"Even if you were to increase your income by an additional $300, your application would still not be approved."

"Even if you were to decrease your advance payment by $25,000 more, your application would still be approved."

3) Alter-factual Explanations

While counterfactuals and semi-factuals concentrate on actionable features, alter-factuals concentrate on less effective and less important features to show those attributes do not

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3410540

lead to any change in the decision-making for that instance. Similar to semi-factuals, an alter-factual is an explanation sample that provides a hypothetical scenario to sample x while assuring there are changes in the attributes and $y' = y$.

"Even if you were to change your name and birthplace, your application would still be approved."

B. RELATED WORK

In the XAI literature, both qualitative and quantitative evaluation methods for explainers have been extensively examined, as evidenced by notable contributions [6]–[9]. Additionally, several works have delved into discussions surrounding challenges, issues, and possibilities related to counterfactuals, serving as valuable guidance for this paper and contributing to the formulation of our road-map [1], [10], [11].

There is a recurrent observation in the literature regarding the insufficient evaluation of explanation methods. Adadi and Berrada [12], in 2018, noted that only 5% of XAI metrics had conducted an evaluation. In a more recent study, Nauta et al. [8] reported, in 2022, that only 33% of XAI methods were evaluated with anecdotal evidence, 58% with quantitative evaluation, and 22% with user studies. Moreover, Keane et al. [11] found that 40% of counterfactual methods had conducted an evaluation, with 21% involving user study evaluations between 2016 and 2021. Notably, the fraction of papers with evaluations increased over time.

Users stand out as the most influential group of stakeholders in XAI applications [13]–[15]. Aligned with this perspective, employing user evaluations proves to be both effective and efficient. However, it is crucial to acknowledge that user evaluations come with significant drawbacks, primarily their high cost and susceptibility to bias. To mitigate these challenges, quantitative evaluation metrics offer a valuable alternative, minimizing disadvantages and providing an easy-to-use, cost-effective medium.

III. COUNTERFACTUAL EXPLANATION EVALUATION

With a systematic search of significant academic databases, including IEEE Xplore, ACM Digital Library, and Google Scholar, titles and abstracts were screened to identify potentially relevant articles, followed by a full-text review to determine eligibility for inclusion. Articles were included if they presented quantitative evaluation metrics or optimization methods designed explicitly for instance-based explanations or applicable across this domain. The final selection of articles was determined through consensus among the authors. While striving for inclusivity, we made purposeful decisions to exclude specific metrics that we or the community deemed unreliable. Also, we intentionally included rarely mentioned metrics in some cases to underscore their potential importance based on our current understanding. Additionally, we cross-referenced the selected articles to ensure comprehensive coverage of relevant metrics and methodologies in the survey.

In this section, we provide a comprehensive overview of the quantitative evaluation metrics of CE derived from our systematic search, employing a unified notation for clarity and consistency. To facilitate readers' understanding and navigation through the equations presented in this section, we introduce Table 1 which compiles the symbols and their short descriptions used in the equations and Table 2 which summarizes presented metrics, their frequency in the literature, applicability to different types of counterfactual explainers, requirements to calculate them, and brief descriptions.

TABLE 1. Unified Notation for the Rest of the Article

Symbol	Description
x	An instance to explain, $x = \langle a_1, ..., a_\rho \rangle$.
ρ	The number of attributes.
X	Set of instances to explain, $x \in X$.
$f()$	Prediction function.
y	Prediction result for x, $y = f(x)$.
$explain()$	Explanation function, returns a set of counterfactuals for an instance.
e	Set of counterfactuals for x, $e = explain(x)$ and $e = \{e_1, ..., e_m\}$.
m	Number of counterfactuals that are provided for x.
e_i	$e_i = \{x'_i, z_i\}$, $e_i \in e$.
x'_i	$x'_i = \langle a'^i_1, ..., a'^i_\rho \rangle$
z_i	$z_i = f(x'_i)$
u_z	The number of unique class labels for the counterfactuals generated for x.
l	The number of unique class labels.
R_j	Range of j^{th} attribute.
k	Number of neighbours.
$kNN()$	$kNN(x)$ finds k Nearest Neighbours to x.
$kNLN()$	$kNLN(x)$ finds k Nearest Like Neighbours to x.
$NUN()$	Nearest Unlike Neighbour, $NUN(x)$ returns the nearest neighbour which is labelled different than x.
$dist()^*$	Distance function, quantifies the distance between two instances.
$P()^*$	Perturbation function, $P(x)$ returns perturbed version of x, x^p.
$\tau()$	$\tau(x'_i, x, j)$ copies j^{th} attribute of x to j^{th} attribute of x'_i and returns x'_i.
A	$A = \{1, 2, ..., \rho\}$
$[A]^l$	All subsets of A that have l elements. $[A]^2 = \{\{a, b\} : a, b \in A, a \neq b\}$
S	All possible subsets (power set) of A without $\emptyset$ and A (super set). $S = \bigcup_{l=1}^{\rho-1} [A]^l$
$\|S\|$	$\|S\| = 2^\rho - 2$
$\tau'()$	$\tau'(x'_i, x, s)$ copies s attributes of x to s attribute of x'_i and returns x'_i.
$\mathbb{1}_{condition}$	If *condition* is fulfilled, it returns 1.

* These functions can be defined in different ways.

Our focus extends to the systematic categorization of quantitative evaluation metrics and optimization methods into two groups. The first group is dedicated to assessing the quality of individual explanations, while the second group is tailored to evaluate the overall effectiveness of explainers.

A. MEASURING THE QUALITY OF A SINGLE EXPLANATION

The following metrics are nominated for evaluating an explanation (e) which is generated for a sample (x) and e consists of m counterfactuals ($e = \{e_1, ..., e_m\}$) and *background data* is the dataset representing the data distribution used for training the model, comprising either the training data itself or a sufficient amount of additional instances with ground truth.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3410540

1) Validity (a.k.a. success rate, hit, recourse accuracy)

A *valid* counterfactual is a counterfactual that does not belong to the same class as the sample explained (x). In the literature, validity is a commonly used metric, and there is a consensus about its definition [1], [2], [15]–[25]. Since the nature of counterfactual explanations promises counter examples as explanations, this metric can be used as an apriori evaluation.

For x, validity is calculated as Equation 1. The value interval for validity is $[0, 1]$, and the higher validity is considered better since all explanation systems aim to propose valid explanations.

$$validity = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{z_i \neq y} \tag{1}$$

In the package, we implemented it as in Equation 2, and in this way, it is guaranteed the model and the explainer agree that the counterfactual is valid.

$$validity = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{z_i \neq y} \mathbb{1}_{z_i = f(x'_i)} \tag{2}$$

Ideally, $z_i = f(x'_i)$ shall always be true. However, in the application scenarios, we need to consider inconsistencies that might caused by the model, the explainer, or other factors.

2) Proximity (a.k.a. dissimilarity, distance, cost)

Proximity is one of the most commonly used evaluation metric for CE [1], [2], [10], [15], [16], [18], [19], [21]–[23], [25]–[38], and computed as the mean of feature-wise distances (l_2 norm) between sample (x) and their counterfactuals (x'_i).

$$proximity = \frac{1}{m} \sum_{i=1}^{m} dist(x, x'_i) \tag{3}$$

Distance function, $dist()$, can be implemented in many different ways, and mostly Euclidean distance is used. However, Gower distance, proposed in 1971 [39] and a distance measure applicable to mixed types of attributes, is another opportune option [18]. As shown in Equation 4 and Equation 5, Gower distance supports numerical, categorical, and nominal data types and normalizes numeric features.

$$gower\_distance(x'_i, x) = \frac{1}{\rho} \sum_{j=1}^{\rho} \delta(a'^i_j, a_j) \tag{4}$$

$$\delta(a'^i_j, a_j) = \begin{cases} \frac{1}{R_j} |a'^i_j - a_j|, & \text{if } a_j \text{ is numerical.} \\ \mathbb{1}_{a'^i_j \neq a_j}, & \text{if } a_j \text{ is categorical or ordinal.} \end{cases} \tag{5}$$

In the package, the proximity metric is implemented with both Euclidean distance and Gower distance functions.

3) Sparsity

Sparsity measures the average number of changed attributes between an instance and its counterfactuals (l_0 norm). Sparsity metric has taken big attention in the literature [2], [15], [16], [18], [19], [21], [22], [25], [26], [30], [36], [38], [40]–[42] because it encourages to generation of concise counterfactual explanations. For example, Keane and Smyth define a *good counterfactual* as a maximum two attribute change between the instance and its counterfactual [40].

$$sparsity = \frac{1}{m\rho} \sum_{i=1}^{m} \sum_{j=1}^{\rho} \mathbb{1}_{a'^i_j \neq a_j} \tag{6}$$

4) Number of explanations

This metric counts how many counterfactuals are generated for an instance [30].

$$number\_of\_explanations = |e| \tag{7}$$

5) Diversity

Diversity for CE is a diverse set of counterfactuals for an instance to offer different actions that can be taken to flip the decision [2], [15]. However, the diversity metric is interpreted in many ways in the literature. Wachter et al. [33] suggests local optima as a source of diverse counterfactuals, while Russell [43] says in most of the problems, there is only one local minima exists and proposes a technique based on integer programming to define diversity constraints. Commonly, it is defined as average proximity between each pair of provided counterfactuals for the instance [2], [15], [16], [16], [18], [21], [25], [44]–[46].

Another way to capture diversity is to build on determinantal point processes (DPP), which has been adopted for solving subset selection problems with diversity constraints [47]. Mothilal et al. [21] used the determinant of the kernel matrix in Equation 9 to define diversity metric (see Equation 8).

$$diversity\_dpp = \det(K) \tag{8}$$

$$K = \begin{bmatrix} \frac{1}{1+dist(x'_1, x'_1)} & \frac{1}{1+dist(x'_1, x'_2)} & \cdots & \frac{1}{1+dist(x'_1, x'_m)} \\ \frac{1}{1+dist(x'_2, x'_1)} & \cdots & \cdots & \frac{1}{1+dist(x'_2, x'_m)} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{1}{1+dist(x'_m, x'_1)} & \frac{1}{1+dist(x'_m, x'_2)} & \cdots & \frac{1}{1+dist(x'_m, x'_m)} \end{bmatrix} \tag{9}$$

In the package, we implemented two different diversity metrics. The first one is Equation 9, and in the second one, we made a slight change and proposed to use *local coverage coefficient* (*lcc*), defined in Equation 10. u_z represents the number of unique class labels for the counterfactuals generated for input x, while l denotes the number of unique class labels in the application case. In Equation 11, the use of *lcc* in the metric is to capture the representative power of the generated counterfactuals.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3410540

$$lcc = \frac{u_z}{l} \tag{10}$$

$$diversity_{lcc} = lcc * diversity\_dpp \tag{11}$$

Across all discussed implementations of diversity, it is considered that maximizing diversity leads to more desired outcomes.

6) yNN

yNN metric measures the amount of support counterfactuals get from positively classified nearest neighbors in background data [19], and ideally, counterfactuals should be close to positively classified individuals, which is a desideratum formulated by Laugel et al. [10], [48].

This metric is only used for binary classification [18], [19]; however, we adopted it for multi-class classification and in the package implemented it as in Equation 12.

$$yNN = \frac{1}{mk} \sum_{i=1}^{m} \sum_{\psi_j \in kNN(x_i')} \mathbb{1}_{z_i = f(\psi_j)} \tag{12}$$

According to Pawelczyk et al. [19], the expected behavior of an explainer is to have a yNN value close to 1 because it implies that the neighborhoods around the counterfactual explanations consist of points with the same predicted label. This indicates that the neighborhoods around these points have already been reached by positively classified instances.

7) Feasibility (a.k.a. kNN distance, connectedness, implausibility)

The *feasibility* of counterfactuals is another notable metric and Ustun et al. [29] addresses the significance of feasibility and relates it to user constraints, while Dandl et al. (2020) [49] defines feasibility as a metric that considers the nearest neighbors of counterfactuals, similar to *yNN*. It quantifies how close the counterfactual example is to the nearest observations in the background data. This metric has found application in various articles [2], [8], [15], [18], [34], [49], [50] which are implemented similar to Equation 13.

$$feasibility = \frac{1}{mk} \sum_{i=1}^{m} \sum_{\psi_j \in kNN(x_i')}^{k} dist(x_i', \psi_j) \tag{13}$$

According to Redelmeier et al. (2021) [18], a smaller score indicates that the counterfactual example lies in a dense part of the training data and is more feasible. In other words, a smaller score corresponds to higher feasibility.

8) kNLN distance

kNLN distance combines *yNN* and *feasibility* metrics and calculates the average distance of the counterfactuals to their k-Nearest Like Neighbor (NLN), as described in [27]. Smaller values of *kNLN* indicate data points closer to the target class's distribution, suggesting higher accordance and a more refined alignment with the characteristics of the target class.

$$kNLN\_distance = \frac{1}{mk} \sum_{i=1}^{m} \sum_{\vartheta_j \in kNLN(x_i')}^{k} dist(x_i', \vartheta_j) \tag{14}$$

9) Relative Distance

The *relative distance* metric quantifies the ratio of the average distance between a sample and its counterfactuals to the distance between the sample and its NUN [51]. It is also known as a measure reflecting plausibility [11], [52].

$$realtive\_distance = \frac{1}{m} \sum_{i=1}^{m} \frac{dist(x_i', x)}{dist(NUN(x), x)} \tag{15}$$

The anticipated behavior for the relative distance value is less than 1 because we expect the generated counterfactuals to be closer than the existing ones. Lower values signify better performance.

10) Redundancy

The *redundancy* metric measures the number of unnecessary feature changes. In other words, how many features in the generated counterfactual do not affect the classification result [18], [19], [44], [53].

In the literature, redundancy has been formulated for changing only one feature at a time. However, as mentioned in [32], this approach is particularly applicable to independence-based methods. In our package, to measure redundancy more comprehensively and make it applicable to dependence and causality-based methods, we formulated redundancy by considering all possible combinations of feature flips, except for flipping all features simultaneously. In other words, our implementation of redundancy encompasses switching back all possible subsets of the features, excluding both the empty set and the full superset (See Equation 16).

$\tau'(x_i', x, s)$ is a function that copies s features of x to s feature of x_i' and returns x_i'. S is a set of all possible subsets (power set) of A without $\emptyset$ and A (super set) where $A = \{1, 2, ..., \rho\}$ and $|S| = 2^{\rho} - 2$.

$$redundancy = \frac{1}{m|S|} \sum_{i=1}^{m} \sum_{s \in S} \mathbb{1}_{f(\tau'(x_i', x, s)) = z_i} \tag{16}$$

Since counterfactual generation aims to propose hypothetical scenarios that alter the decision with minimal changes, a lower redundancy value indicates better results.

11) Robustness (a.k.a. stability)

Robustness is a prominent objective in counterfactual generation [54], and this metric is considered relevant for user trust [55]. Mishra et al. [56] define it as a collection of related yet distinct measures assessing the extent to which explanations for an AI model may change under certain restricted changes to the system, while similarly, Sharma et al. [37] consider robustness as the distance between the instances and

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3410540

their corresponding counterfactuals however we refer to it as proximity. In the literature, various other approaches exist to measure robustness. Guidotti's recent works define it as the ability to produce similar counterfactuals for similar instances that belong to the same class [2], [57]. Alvarez and Jaakkola's recent works [58], [59] posit that a good explanation should not only explain the given sample but also provide similar explanations for a similar sample and use local Lipschitz constant to quantify robustness. Although there is no consensus on a universal formula to measure robustness, there is consensus on the definition of robustness itself [8], [10], [22], [27], [34], [35], [38], [50], [52], [55], [59]–[66]. We have chosen one of the most flexible, common, and generalized definitions of robustness, which involves measuring the distance between the explanation of x and the explanation of a slightly perturbed version of x (refer to Equation 17) [38], [62], [67] and propose a slight change as in Equation 18 to normalize the robustness value and make it more comparable.

$$robustness = \frac{1}{m}\sum_{i=1}^{m} dist(e_i, explain(P(x))) \tag{17}$$

$$robustness_{norm} = \frac{1}{m}\sum_{i=1}^{m} \frac{dist(e_i, explain(P(x)))}{dist(x, P(x))} \tag{18}$$

A smaller robustness value indicates a more robust counterfactual. The lower, the better.

However, there is a concern in the literature regarding whether closest counterfactuals are sufficiently robust and if there exists a trade-off between the robustness and proximity metrics [62], [63]. This concern is discussed in Section VI.

12) Plausibility (a.k.a. actionability)

Plausibility is considered as a crucial requirement for informative counterfactuals [6], [36]. Karimi et al. [31] and Verma et al. [42] introduce the concept of plausible explanations, defining them as semantically meaningful, multimodal, actionable, immutable, and unbiased explanations. On the other hand, Guidotti et al. [2] emphasize that a plausible explanation should also adhere to feature-based norms, avoiding the proposal of outlier values within the features of the explanations, and Molnar [68] mentions proximity is a good proxy for plausibility.

However, Keane et al. [11] offer a more comprehensive perspective on plausibility, categorizing various definitions of the term into two groups: Plausibility-As-Proximity, which somewhat aligns with the proximity metric mentioned earlier, and Plausibility-as-More-Good-Features, which corresponds to the aspects highlighted by Karimi et al. and Verma et al. [31], [42]. In line with Keane et al., it is possible to use proximity and constraint violation metrics in accordance with the Plausibility-As-Proximity and Plausibility-as-More-Good-Features categorizations, respectively.

On the other hand, an alternative perspective suggests that conforming to the data distribution serves as a more accurate proxy for plausibility [36], [48], [50], [69], [70]. Pawelczyk et al. [64] support this approach, proposing that a plausible counterfactual is an instance from a *possible world*. In alignment with this perspective, our package implements the plausibility metric, following Laugel et al. [10], utilizing the Local Outlier Factor score (Breunig et al., 2000) [71] for outlier detection. Plausibility is formulated as in Equation 19 considering $k = 1$. In this manner, the measure quantifies the extent to which counterfactuals deviate from the ground truth instances of the same class where $NLN(x)$ is a function that returns Nearest Unlike Neighbour of x and $NUN(x)$ is a function that returns Nearest Like Neighbour of x.

$$plausibility = \frac{1}{m}\sum_{i=1}^{m} \frac{dist(x'_i, NLN(x'_i))}{dist(NLN(x'_i), NUN(NLN(x'_i)))} \tag{19}$$

13) Discriminative Power (a.k.a. dipo)

Guidotti et al. [2], in line with [21], [72], define *discriminative power* as the ability to differentiate between two distinct classes solely through a naive approach using counterfactuals and implemented the metric using following steps. Train a $1NN$ model using x and e, classify all instances in $kNUN(x) \cup kNLN(x)$ set, and calculate the accuracy score, which will be discriminative power. However, the authors underscore that, in accordance with the definition, the discriminative power relies on a subjective basis, making its quantification challenging without experiments involving human subjects.

14) Vulnerability (a.k.a. unconnectedness, unjustification)

Vulnerability refers to how susceptible the counterfactual is to manipulation. Measuring the vulnerability is approached in various ways in the literature [41], [48], [67], [73]. For example, Slack et al. [67] run adversarial attacks to show vulnerability, claim that vulnerable counterfactuals are open to manipulation, and show that even the most popular counterfactual methods are open to such manipulations. The authors point out the relationship between vulnerability and reliability and propose three ways to mitigate such threat: (I) adding noise to the initialization of the counterfactual search, (II) reducing the set of features used to compute counterfactuals, and (III) reducing the model complexity. On the other hand, with a similar approach, Laugel et al. [73] mention that the lack of robustness of the classifier causes vulnerability issues. They follow a procedure to analyze the unconnectedness of classification regions as mentioned in [48] (Local Risk Assessment (LRA) and Vulnerability Evaluation (VE)) and show that state-of-the-art post-hoc counterfactual approaches may generate justified explanations but at the expense of counterfactual proximity. And their implementation can be found at https://github.com/thibaultlaugel/truce.

15) Computational Complexity (a.k.a. cost)

Computational complexity refers to the cost for the explainer to generate a single explanation, encompassing various re-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3410540

sources such as time and memory. Efficiency is crucial for practical applications where timely and resource-efficient explanations are essential. In the literature, it is typically measured by the average time required to generate a single explanation [8], [15], [19], [27], [38], [42].

16) Constraint Violation (a.k.a. feasibility, user constraints, actionability)

The *constraint violation* metric is used to quantify the count of violated pre-defined constraints in generated counterfactuals [2], [15], [18], [19], [21], [23], [43], [44], [49]. These constraints can be set by users or experts for different reasons like bias prevention, personal preferences, and domain-related constraints.

B. MEASURING THE QUALITY OF THE EXPLANATION SYSTEMS

The following metrics are nominated for evaluating a set of explanations that are generated for a set of instances (X). To emphasize, all the metrics discussed in the previous section can be generalized for the entire system by applying them to a set of instances (sufficient number of instances) and taking the mean of the results.

1) Coverage (a.k.a. explanatory competence)

As Keane et al. mentioned [11], in the literature, a way to measure the quality of counterfactual explanation systems is to evaluate how well they cover the entire data distribution. There are two main approaches to measuring the coverage of a system. First, detecting the out-of-distribution counterfactuals like IM1 and IM2 [70], [74]. Second, quantifying the ability to generate valid counterfactuals across various types of instances [40], [49].

In the package, we follow the second idea and formulate coverage as:

$$coverage = \frac{1}{|X|}\sum_{x\in X}\mathbb{1}_{|e|>0},\ e = explain(x) \tag{20}$$

2) Rigidity (a.k.a. fidelity)

The *rigidity* metric, applicable to twin XAI systems, quantifies the extent to which the explainer can faithfully replicate the machine learning model for a specific instance [16], [75]. Bayrak and Bach [75] formulate rigidity as in Equation 21, where *acc* represents the accuracy score of the model, and *supp* denotes the fraction of instances where the explanation system and the model agree, relative to the total number of explanations. The authors mention that lower rigidity indicates better performance.

$$rigidity = \left|1 - \frac{supp}{acc}\right| \tag{21}$$

IV. APPLICABILITY FOR SEMIFACTUALS AND ALTERFACTUALS

In contrast to the extensive attention received by counterfactual explanations in the literature, other types of instance-based explainers have received comparatively less focus. Nevertheless, it is noteworthy that certain metrics can be adapted to evaluate these alternative explanation methods. *Validity*, *plausibility*, *complexity*, *diversity*, *number of explanations*, *constraint violation* and *coverage* metrics can be applied to all instance-based explainers.

A. SEMI-FACTUAL EXPLANATION EVALUATIONS

Semi-factual explanations share several key characteristics with counterfactuals. With the help of a survey conducted by Aryal and Keane [76], which gathers metrics and prior knowledge, and a recent work by Kenny and Huang [77] introduces the 'Gain' concept aiming to quantify how much a user can benefit from the explanation, we define applicable metrics for semi-factuals as follows: *Proximity*, higher values are considered better, indicating that the semi-factual should be distant from the instance. *Sparsity*, similar to counterfactuals, making changes to fewer features can be important in some cases. *yNN* and *feasibility*, similar to counterfactuals, to measure the support from background data. *Relative distance* and *kNLN distance* are applicable and also adapted as *kNUN distance* since the goal is to generate an explanation from the same class. For semi-factuals, staying close to the decision boundary and proximity to NUN are crucial considerations.

B. ALTER-FACTUAL EXPLANATION EVALUATIONS

Even though the alter-factual thinking concept is discussed in many psychology articles, it is relatively new in the XAI field. Mertes et al. [78] introduced alter-factual explanations in 2022. As this is the only application thus far and the authors relied solely on user evaluations, we currently lack information on the evaluation metrics used in the existing literature. However, similar to semi-factuals, the *proximity* metric and metrics measuring support from background data can be applied, in addition to metrics suitable for all instance-based explainers, as mentioned above. Moreover, since alter-factuals provide alternate scenarios that do not affect the decision-making process adversely, the *redundancy* metric can be utilized and a higher redundancy value indicates better results in that case.

V. CEval TOOLKIT

We provide an accompanying toolkit, a Python package, CEval[4], designed to facilitate the evaluation of counterfactual explanations. With a focus on adaptability, CEval toolkit can be easily adjusted for various use cases, providing users with a versatile solution. The package incorporates a comprehensive set of 14 implemented metrics, as detailed in Table 2 and Sect. III, ensuring a broad coverage of evaluation criteria. Metrics included in the package are carefully selected based on their applicability to diverse scenarios and their proven effectiveness in evaluating the quality of explanations. For instance, metrics that rely on specific explainers are omitted to ensure compatibility with explainers implemented in var-

[4] https://pypi.org/project/CEval/

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

TABLE 2. A summary table of quantitative evaluation metrics and their frequency in the literature (count), applicability to different types of counterfactual explainers (applicable), requirements to calculate them (requires), and summarized description (short description).

Metrics	**Count**	**Applicable**				**Requires**		**Short Description**
		Generated[3]	Existed[3]	Single[3]	Multi[3]	Data[3]	Model[3]	
Validity [1]	12	x	x	x	x	-	x	Whether the decision was altered.
Proximity [1]	24	x	x	x	x	-	-	Mean of feature-wise distance between the instance and its counterfactuals.
Sparsity [1]	15	x	x	x	x	-	-	Mean number of altered features between the instance and its counterfactuals.
# of counterfactuals [1]	1	x	x	-	x	-	-	Number of counterfactuals generated for an instance.
Diversity [1]	12	x	x	-	x	-	-	Mean proximity between counterfactuals generated for an instance.
$Diversity_{\_lcc}$ [1]	12	x	x	-	x	x	-	Diversity with class coverage coefficient.
yNN [1]	4	x	x	x	x	x	x	Amount of support that counterfactuals receive from positively classified background data.
Feasibility [1]	9	x	x	x	x	x	-	Mean proximity of the counterfactuals to their nearest observations in the background data.
kNLN Distance [1]	1	x	x	x	x	x	-	Mean distance of counterfactuals to their k-NLN.
Relative Distance [1]	4	x	-	x	x	x	-	The ratio of the mean distance between the instance and counterfactuals to the mean distance between the instance and its NUN.
Redundancy [1]	5	x	x	x	x	-	x	Mean count of unnecessary feature changes.
Robustness [2]	25	x	-	x	x	x	x	Mean proximity between the explanation of the instance and the explanation of a slightly perturbed version of the instance.
Plausibility [1]	12	x	-	x	x	x	-	The degree of credibility in the context.
Discriminative Power	3	x	x	-	x	x	-	The ability to differentiate two distinct classes through a naive approach.
Vulnerability	4	x	-	x	x	x	-	The extent of susceptibility to manipulations.
Complexity	6	x	x	x	x	x	x	Cost for the explainer to generate a single explanation.
Constraints [1]	9	x	x	x	x	-	-	Mean count of violated pre-defined constraints.
Coverage [2]	5	x	x	x	x	-	-	The ability to generate valid counterfactuals across various types of instances.

[1] This metric is implemented in the package.
[2] This metric requires explanation method/function to be calculated.
[3] Applicability to explainers that provide *generated:* counterfactuals that are generated, *existed:* counterfactuals that are selected from existing samples, *single:* only one counterfactual explanation per sample, *"multi":* many counterfactuals explanation per sample.

ious programming languages. Moreover, metrics are chosen for their clarity, interpretability, and ease of implementation, enhancing the toolkit's usability for both researchers and practitioners.

Notably, the package is compatible with explainers implemented in various languages, as it only requires the explanations themselves, making it a solid integration for users. Additionally, users can benefit from the examples provided, which enhance the accessibility and utility of the toolkit.

While the CEval toolkit offers a comprehensive set of evaluation metrics, it may only cover some possible aspects of counterfactual explanation evaluation. Users should be aware that the toolkit's effectiveness may vary depending on the specific characteristics of the explanations and the underlying data. Additionally, the toolkit's performance may be influenced by factors such as the complexity of the models being evaluated and the quality of the explanations themselves. As with any evaluation toolkit, users should exercise caution and consider the limitations of the metrics provided when interpreting the results.

VI. DISCUSSIONS AND FUTURE DIRECTIONS

In Section III, we introduced a suite of metrics with various approaches sourced from the existing literature. While enhancing and proposing slight modifications to some of these metrics in Sect. III, we strive to facilitate deeper reflections and discussions and offer valuable recommendations in this section.

We categorized these discussions into eight subcategories, respectively addressing: (I) the use of *validity as a criterion*, (II) interpretation of the level of *sparsity*, (III) exploration of the relationship between *proximity, plausibility, and robustness*, and determination of optimal values, (IV) the influence of *distance function selection* and identification of the most suitable function, (V) the implications of *robustness* across diverse application cases, (VI) common challenges associated with neighbor-based metrics such as *yNN, feasibility, and kNLN*, and the impact of neighbor selection, (VII) examination of the pros and cons of different approaches to the plausibility term and *plausibility maximization*, and (VIII) consideration of the effect of presenting a diverse set of hypothetical scenarios to users and strategies for determining the optimal *number of explanations*.

Validity as a criterion. Since the intrinsic promise of counterfactuals lies in their ability to provide counterexamples as explanations, validity can be set as a crucial prerequisite in explainers.

Sparsity. is a widely employed technique, and is often

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3410540

associated with the idea of changing the minimum number of features mean higher quality [40]. However, we aim to critically examine the question: 'Why do we need sparse counterfactuals?' In agreement with Virgolin and Fracaros [79], we acknowledge the potential benefits of sparsity in enhancing the quality of counterfactual explanations. However, it is important to note that relying solely on sparsity may not be sufficient to measure the overall quality of counterfactuals. Furthermore, we emphasize the significance of considering the specific use case, as blindly rejecting correlated features in certain application areas may lead to misinterpretations.

Proximity, plausibility, and robustness relationship. Using proximity as a primary metric is extensive in the literature. However, several works discuss and highlight various inquiries. Artelt et al. [62] and Dutta et al. [63] inquire about a potential trade-off between plausibility/robustness and proximity and question whether the closest counterfactuals are sufficiently robust and plausible. Artelt et al. endorse the use of plausibility over proximity, emphasizing the plausible counterfactuals are more robust and the instability of the closest counterfactuals due to their sensitivity to small perturbations. Similarly, diverse recent works [34], [50], [64] posit that counterfactuals lying on the data manifold may exhibit greater robustness than the closest counterfactuals. While the literature frequently explores the notion of a proximity-robustness or proximity-plausibility trade-off, particular works, including [22], [65], [66], propose algorithms designed to identify counterfactuals that are both close and robust. These works indicate the feasibility of generating close and robust explanations and proof on linear models. However, Delaney et al. [36] demonstrate that counterfactuals with high proximity might fall outside the data manifold, leading to implausibility. To the best of our knowledge, counterfactual generation is an optimization problem, and we agree with Artelt et al. [62] that generating stable and robust counterfactuals is still an open research problem. Moreover, tailoring the counterfactual generation process according to specific use cases and needs adds an additional layer of complexity to this challenge.

Distance function selection. The chosen distance function influences the effectiveness of evaluation or optimization. While common functions like Euclidean and Minkowsky are widely used, it is crucial to *consider the nature of the data*. For example, Guidotti et al.'s adapted distance function [2] and Gower distance support both nominal and numeric features, which is valuable in terms of multimodality. A practical approach involves using the same encoding method as the machine learning model combined with a distance metric that supports numeric features. Additionally, considering distance functions *with normalization or encoding* proves beneficial for the proximity-plausibility trade-off. *Domain expertise* is invaluable for explainers, and the involvement of a domain expert or domain-related data can contribute to defining a domain-specific distance function. Such tailored distance functions provide a nuanced and reliable solution for discussions surrounding the proximity-plausibility trade-off.

Robustness. Robustness is a frequently contemplated metric, with Jiang et al. [24] highlighting recent studies showing the potential lack of robustness of counterfactuals to changes in machine learning models. These discussions prompt questions about their reliability in real-world applications. In this paper, we used the definition of robustness that makes perturbations in the instance, generates counterfactuals for the perturbed instance, and assesses the distance between counterfactuals. However, we emphasize a concern regarding the perturbation of the instance, where the significance of perturbing specific features is negligible. Real-world applications often lack clear decision boundaries, and perturbing an instance may unexpectedly alter decisions, potentially leading to counterfactual generation for instances from another class. Recent discussions on the importance of feature contributions [26], [38] underline the significance of considering these perturbations thoughtfully, anticipating more favorable outcomes.

yNN, feasibility, and kNLN. The yNN, feasibility, and kNLN distance metrics make measurements over the nearest neighbors. Central to this assessment is the critical decision of determining the number of neighbors, which should be selected based on the specific application case. Also, for the yNN metric, a higher value nearing 1 is considered indicative of the surroundings of these points being explored by positively classified instances, suggesting a certain level of model understanding. However, the challenge lies in recognizing that class areas may not always exhibit consistent patterns in real-world applications. This fact raises questions about the appropriate selection of k, as an optimal choice should consider broad class regions and accommodate the intricacies of localized and specific decision boundaries. Achieving a balance between comprehensiveness and precision in neighbor selection becomes pivotal for robust counterfactual evaluation. Moreover, in the literature, feasibility is generally associated with connectedness and actionability. Therefore, combining constraint violations with a data-driven distributional approach might provide a more comprehensive assessment.

Plausibility maximization. The plausibility term is examined through four distinct approaches: *(I)* Being close to the decision boundary. In instances that have multiple decision boundaries around them, proximity to the target class decision boundary becomes crucial. Two key considerations arise: the availability of ground truth data and defining the threshold for proximity to the decision boundary. *(II)* Making changes in the actionable features. Utilizing feature importance techniques and predefined constraints is a common practice. In this paper, it is treated as the constraint violation metric. *(III)* Staying in the data manifold. A prevalent approach suggests that falling into the data distribution is a good indicator of plausibility [48], [50], [69], [70]. *(IV)* Originating from existing data samples. In low-density distribution, originating from existing data samples might cause implausibility. However, with a well-developed optimization method, this effect can be minimized, as in Bayrak and Bach's study [38]

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3410540

generating counterfactuals with a dynamic, iterative approach between the instance and NUN. While we acknowledge the significance of all four approaches, we assert that none individually defines plausibility adequately. Therefore, comprehensive consideration of these approaches and developing domain-aware explainers contributes to enhanced plausibility.

Number of explanations. Many methods aim to generate a set of diverse counterfactual explanations to explain an individual instance with many different approaches [21], [43], [45], [46]. Proposing users multiple hypothetical scenarios might be helpful, but we believe that counterfactuals should be generated on purpose, and sometimes, providing an excessive number of hypothetical examples might unintentionally disrupt the targeted reasoning process for users. In alignment with Karimi et al. [31], we argue that diversity should not be misconstrued as repetition. Removal of duplications, akin to considerations for validity, can also serve as a prerequisite in explainers.

VII. CONCLUSION

In this work, we conducted a comprehensive survey focusing on the evaluation of instance-based explanations, particularly aimed CE. Our primary objective was to lighten up the intricate landscape of CE evaluations, providing valuable insights for both XAI practitioners and researchers. Alongside offering a standardized notation for metrics and an accompanying Python toolkit, we extend our exploration to consider the potential applicability of metrics to alter-factuals and semi-factuals. Moreover, our work engages in critical discussions concerning the trade-offs and considerations inherent in metric selection, underscoring the necessity for a nuanced understanding of application-specific requirements. This contribution aims to heighten awareness regarding metric utilization and counterfactual generation method development.

Acknowledging the open nature of instance-based explanation evaluation, we propose future research directions that look into concepts such as causality and optimality, particularly in the context of instance-based explanations for multimodal data. Recognizing the inherent limitations of our efforts to meet broad community requirements, such as providing a framework for both qualitative and quantitative evaluations of explainers, our study strives to address discernible gaps, thereby laying the groundwork for further explorations.

APPENDIX.

CEval UTILIZATION AND EXPERIMENTS

This appendix serves as a concise guide for installing and using the package, offering straightforward instructions for users to instruct themselves quickly on its functionalities. Additionally, it presents reproducible experimental results showcasing the package's versatility across a set of open-source explainers, datasets, and models. These results underscore the package's effectiveness in various scenarios, providing users with valuable insights into its capabilities for supporting diverse XAI tasks.

The CEval package is available for download from PyPI and can be installed using the following command:

```
$ pip install CEval
```

To use the package, import it into your Python script:

```
from CEval import CEval
```

Create a CEval object (`evaluator`) and provide the following arguments: samples to explain (`X`), `label`, background data (`data`), `model`, `k`, distance function (`dist`), and a list of constraints:

```
evaluator = CEval(X, label, data, model, k, dist,
    constraints)
```

Add explainers with explanations generated for `X` to the `evaluator`. Provide explainer name, explanations, explainer type, and explanation mode arguments:

```
evaluator.add_explainer('Explainer-1', exp_res1,
    exp_type='generated-cf', mode='1to1')
evaluator.add_explainer('Explainer-2', exp_res2,
    exp_type='generated-cf', mode='1toN')
```

The CEval object (`evaluator`) maintains a DataFrame named `comparison_table` that stores the results of the evaluations for the added explainers:

```
display(evaluator.comparison_table)
```

To show the utilization of the CEval Toolkit, we conducted a set of reproducible experiments to evaluate explanations generated by utilizing three different CE methods on the Breast Cancer Dataset[5] and the South German Credit Dataset[6]. These datasets exhibit distinct characteristics. We used open-source CEs with different characteristics, CFNOW [80], DICE [21], and CFSHAP [81]. The experiments encompassed a comprehensive analysis of the performance of explainers on different machine learning models, including Random Forest (RF), and Gradient Boost (GBC and XGB), representing diverse applications.

Experiments were conducted with possible combinations of the explainers, models, and datasets. The quality of explanations was quantified through performance metrics implemented and facilitated by the CEval Toolkit, as presented in Table 3.

Indeed, it is essential to stress that the interpretation of these results should be context-dependent, considering the specific requirements and objectives of the users. For instance, when selecting an explainer for a Breast Cancer Data application where maintaining validity is essential, opting for an explainer like DICE may not be advisable despite its high

[5] https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

[6] https://archive.ics.uci.edu/dataset/522/south+german+credit

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3410540

TABLE 3. **Experimental results. Evaluation of CFNOW, DICE, and CFSHAP explanations.**

	Breast Cancer Dataset						Credit Dataset					
	CFNOW		DICE		CFSHAP		CFNOW		DICE		CFSHAP	
	RF	XGB	RF	XGB	RF	XGB	RF	GBC	RF	GBC	RF	GBC
validity	1.0	0.971	0.903	0.971	1.0	0.971	0.84	0.72	0.88	0.72	0.90	0.72
proximity	57.968	24.486	959.59	873.97	559.63	544.26	27.953	55.157	3177.4	3591.8	1932.2	1984.2
proximity$_{gower}$	0.052	0.052	0.056	0.056	0.052	0.052	0.113	0.113	0.093	0.093	0.113	0.113
sparsity	0.410	0.346	0.056	0.056	0.999	0.999	0.189	0.204	0.093	0.093	0.443	0.436
# of CF	-	-	5	5	-	-	-	-	5	5	-	-
diversity	-	-	0.002	0.001	-	-	-	-	0.0	0.0	-	-
diversity$_{lcc}$	-	-	0.003	0.001	-	-	-	-	0.0	0.0	-	-
yNN	0.109	0.143	0.517	0.536	0.646	0.646	0.0	0.0	0.0	0.0	0.0	0.0
feasibility	108.15	94.743	422.79	458.22	24.493	24.346	27.215	27.108	200.06	174.49	29.220	25.813
kNLN_dist	125.71	127.09	107.82	121.19	88.971	90.943	182.44	201.50	182.01	213.70	164.12	182.62
relative_dist	0.476	0.501	14.190	12.734	5.308	4.848	0.949	1.920	175.94	188.56	101.08	96.150
redundancy	7.400	6.400	3.851	2.846	29.400	29.086	1.120	1.48	0.468	0.52	6.980	7.80
plausibility	8.332	7.770	6.474	9.889	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0
constraint_violation	1.0	0.886	0.994	0.971	1.0	1.0	-	-	-	-	-	-

diversity and feasibility scores. Similarly, when selecting an explainer for Credit Data, CFNOW outperformed in proximity. However, if sparsity is more critical for that application, we cannot simply advise CFNOW because DICE performs significantly better in terms of sparsity, even though it may not excel in proximity. Consequently, users should carefully evaluate the trade-offs and select the most suitable explainer based on their unique use case and application needs.

REFERENCES

[1] F. Hvilshøj, A. Iosifidis, and I. Assent, "On quantitative evaluations of counterfactuals," *arXiv preprint arXiv:2111.00177*, 2021.

[2] R. Guidotti, "Counterfactual explanations and how to find them: literature review and benchmarking," *Data Mining and Knowledge Discovery*, pp. 1–55, 2022.

[3] R. Guidotti, "Evaluating local explanation methods on ground truth," *Artificial Intelligence*, vol. 291, p. 103428, 2021.

[4] G. Visani, E. Bagli, F. Chesani, A. Poluzzi, and D. Capuzzo, "Statistical stability indices for lime: Obtaining reliable explanations for machine learning models," *Journal of the Operational Research Society*, vol. 73, no. 1, pp. 91–101, 2022.

[5] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

[6] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.

[7] J. van der Waa, E. Nieuwburg, A. Cremers, and M. Neerincx, "Evaluating xai: A comparison of rule-based and example-based explanations," *Artificial Intelligence*, vol. 291, p. 103404, 2021.

[8] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, and C. Seifert, "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai," *ACM Computing Surveys*, 2022.

[9] M. Förster, M. Klier, K. Kluge, and I. Sigler, "Fostering human agency: A process for the design of user-centric xai systems," 2020.

[10] T. Laugel, M.-J. Lesot, C. Marsala, and M. Detyniecki, "Issues with post-hoc counterfactual explanations: a discussion," *arXiv preprint arXiv:1906.04774*, 2019.

[11] M. T. Keane, E. M. Kenny, E. Delaney, and B. Smyth, "If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual XAI techniques," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, International Joint Conferences on Artificial Intelligence Organization, Aug. 2021.

[12] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (xai)," *IEEE access*, vol. 6, pp. 52138–52160, 2018.

[13] M. Förster, M. Klier, K. Kluge, and I. Sigler, "Evaluating explainable artifical intelligence–what users really appreciate," 2020.

[14] A. Preece, D. Harborne, D. Braines, R. Tomsett, and S. Chakraborty, "Stakeholders in explainable ai," *arXiv preprint arXiv:1810.00184*, 2018.

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

[16] N. Maaroof, A. Moreno, A. Valls, M. Jabreel, and M. Szeląg, "A comparative study of two rule-based explanation methods for diabetic retinopathy risk assessment," *Applied Sciences*, vol. 12, no. 7, p. 3358, 2022.

[17] Z. Geng, M. Schleich, and D. Suciu, "Computing rule-based explanations by leveraging counterfactuals," *arXiv preprint arXiv:2210.17071*, 2022.

[18] A. REDELMEIER, M. JULLUM, K. AAS, and A. LØLAND, "Mcce: Monte carlo sampling of realistic counterfactual explanations," *stat*, vol. 1050, p. 18, 2021.

[19] M. Pawelczyk, S. Bielawski, J. v. d. Heuvel, T. Richter, and G. Kasneci, "Carla: a python library to benchmark algorithmic recourse and counterfactual explanation algorithms," *arXiv preprint arXiv:2108.00783*, 2021.

[20] X. Ye, R. Nair, and G. Durrett, "Connecting attributions and qa model behavior on realistic counterfactuals," *arXiv preprint arXiv:2104.04515*, 2021.

[21] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 607–617, 2020.

[22] K. Rawal and H. Lakkaraju, "Beyond individualized recourse: Interpretable and interactive summaries of actionable recourses," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12187–12198, 2020.

[23] D. Mahajan, C. Tan, and A. Sharma, "Preserving causal constraints in counterfactual explanations for machine learning classifiers," *arXiv preprint arXiv:1912.03277*, 2019.

[24] J. Jiang, F. Leofante, A. Rago, and F. Toni, "Formalising the robustness of counterfactual explanations for neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 14901–14909, 2023.

[25] T. Teofili, D. Firmani, N. Koudas, V. Martello, P. Merialdo, and D. Srivastava, "Effective explanations for entity resolution models," in *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pp. 2709–2721, IEEE, 2022.

[26] N. Wiratunga, A. Wijekoon, I. Nkisi-Orji, K. Martin, C. Palihawadana, and D. Corsar, "Actionable feature discovery in counterfactuals using feature relevance explainers.," CEUR Workshop Proceedings, 2021.

[27] P. Rasouli and I. C. Yu, "Analyzing and improving the robustness of tabular classifiers using counterfactual explanations," in *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 1286–1293, IEEE, 2021.

[28] F. Yang, S. S. Alva, J. Chen, and X. Hu, "Model-based counterfactual synthesizer for interpretation," in *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 1964–1974, 2021.

[29] B. Ustun, A. Spangher, and Y. Liu, "Actionable recourse in linear classification," in *Proceedings of the conference on fairness, accountability, and transparency*, pp. 10–19, 2019.

[30] I. Stepin, J. M. Alonso, A. Catala, and M. Pereira-Fariña, "Generation and evaluation of factual and counterfactual explanations for decision trees and

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3410540

fuzzy rule-based classifiers," in *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–8, IEEE, 2020.

[31] A.-H. Karimi, G. Barthe, B. Balle, and I. Valera, "Model-agnostic counterfactual explanations for consequential decisions," in *International Conference on Artificial Intelligence and Statistics*, pp. 895–905, PMLR, 2020.

[32] M. Pawelczyk, K. Broelemann, and G. Kasneci, "On counterfactual explanations under predictive multiplicity," in *Conference on Uncertainty in Artificial Intelligence*, pp. 809–818, PMLR, 2020.

[33] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.

[34] K. Kanamori, T. Takagi, K. Kobayashi, and H. Arimura, "Dace: Distribution-aware counterfactual explanation by mixed-integer linear optimization.," in *IJCAI*, pp. 2855–2862, 2020.

[35] A. Boopathy, S. Liu, G. Zhang, C. Liu, P.-Y. Chen, S. Chang, and L. Daniel, "Proper network interpretability helps adversarial robustness in classification," in *International Conference on Machine Learning*, pp. 1014–1023, PMLR, 2020.

[36] E. Delaney, D. Greene, and M. T. Keane, "Instance-based counterfactual explanations for time series classification," in *International Conference on Case-Based Reasoning*, pp. 32–47, Springer, 2021.

[37] S. Sharma, J. Henderson, and J. Ghosh, "CERTIFAI," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, ACM, Feb. 2020.

[38] B. Bayrak and K. Bach, "Pertcf: A perturbation-based counterfactual generation approach," in *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pp. 174–187, Springer, 2023.

[39] J. C. Gower, "A general coefficient of similarity and some of its properties," *Biometrics*, pp. 857–871, 1971.

[40] M. T. Keane and B. Smyth, "Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai)," in *Case-Based Reasoning Research and Development: 28th International Conference, ICCBR 2020, Salamanca, Spain, June 8–12, 2020, Proceedings 28*, pp. 163–178, Springer, 2020.

[41] P. Romashov, M. Gjoreski, K. Sokol, M. V. Martinez, and M. Langheinrich, "Baycon: Model-agnostic bayesian counterfactual generator," in *Proceedings of the 31st International Joint Conference on Artificial Intelligence, Vienna, Austria*, pp. 23–29, 2022.

[42] S. Verma, J. Dickerson, and K. Hines, "Counterfactual explanations for machine learning: Challenges revisited," *arXiv preprint arXiv:2106.07756*, 2021.

[43] C. Russell, "Efficient search for diverse coherent explanations," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 20–28, 2019.

[44] M. Downs, J. L. Chu, Y. Yacoby, F. Doshi-Velez, and W. Pan, "Cruds: Counterfactual recourse using disentangled subspaces," *ICML WHI*, vol. 2020, pp. 1–23, 2020.

[45] S. S. Hada and M. Á. Carreira-Perpiñán, "Exploring counterfactual explanations for classification and regression trees," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 489–504, Springer, 2021.

[46] K. Mohammadi, A.-H. Karimi, G. Barthe, and I. Valera, "Scaling guarantees for nearest counterfactual explanations," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 177–187, 2021.

[47] A. Kulesza, B. Taskar, *et al.*, "Determinantal point processes for machine learning," *Foundations and Trends® in Machine Learning*, vol. 5, no. 2–3, pp. 123–286, 2012.

[48] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki, "The dangers of post-hoc interpretability: Unjustified counterfactual explanations," 2019.

[49] S. Dandl, C. Molnar, M. Binder, and B. Bischl, "Multi-objective counterfactual explanations," in *International Conference on Parallel Problem Solving from Nature*, pp. 448–469, Springer, 2020.

[50] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, and P. Flach, "Face: feasible and actionable counterfactual explanations," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 344–350, 2020.

[51] B. Smyth and M. T. Keane, "A few good counterfactuals: generating interpretable, plausible and diverse counterfactual explanations," in *International Conference on Case-Based Reasoning*, pp. 18–32, Springer, 2022.

[52] S. C. Smith and S. Ramamoorthy, "Counterfactual explanation and causal inference in service of robustness in robot control," in *2020 Joint IEEE 10th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pp. 1–8, IEEE, 2020.

[53] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee, "Counterfactual visual explanations," in *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, pp. 2376–2384, PMLR, 09–15 Jun 2019.

[54] H. Lakkaraju, N. Arsov, and O. Bastani, "Robust and stable black box explanations," in *International Conference on Machine Learning*, pp. 5628–5638, PMLR, 2020.

[55] E. Ates, B. Aksar, V. J. Leung, and A. K. Coskun, "Counterfactual explanations for multivariate time series," in *2021 International Conference on Applied Artificial Intelligence (ICAPAI)*, pp. 1–8, IEEE, 2021.

[56] S. Mishra, S. Dutta, J. Long, and D. Magazzeni, "A survey on the robustness of feature importance and counterfactual explanations," *arXiv preprint arXiv:2111.00358*, 2021.

[57] R. Guidotti and S. Ruggieri, "On the stability of interpretable models," in *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2019.

[58] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," *arXiv preprint arXiv:1806.08049*, 2018.

[59] D. Alvarez Melis and T. Jaakkola, "Towards robust interpretability with self-explaining neural networks," *Advances in neural information processing systems*, vol. 31, 2018.

[60] C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D. I. Inouye, and P. K. Ravikumar, "On the (in) fidelity and sensitivity of explanations," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[61] J. Adebayo, M. Muelly, I. Liccardi, and B. Kim, "Debugging tests for model explanations," *arXiv preprint arXiv:2011.05429*, 2020.

[62] A. Artelt, V. Vaquet, R. Velioglu, F. Hinder, J. Brinkrolf, M. Schilling, and B. Hammer, "Evaluating robustness of counterfactual explanations," in *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 01–09, IEEE, 2021.

[63] S. Dutta, J. Long, S. Mishra, C. Tilli, and D. Magazzeni, "Robust counterfactual explanations for tree-based ensembles," in *International Conference on Machine Learning*, pp. 5742–5756, PMLR, 2022.

[64] M. Pawelczyk, K. Broelemann, and G. Kasneci, "Learning model-agnostic counterfactual explanations for tabular data," in *Proceedings of the web conference 2020*, pp. 3126–3132, 2020.

[65] S. Upadhyay, S. Joshi, and H. Lakkaraju, "Towards robust and reliable algorithmic recourse," *Advances in Neural Information Processing Systems*, vol. 34, pp. 16926–16937, 2021.

[66] E. Black, Z. Wang, M. Fredrikson, and A. Datta, "Consistent counterfactuals for deep models," *arXiv preprint arXiv:2110.03109*, 2021.

[67] D. Slack, A. Hilgard, H. Lakkaraju, and S. Singh, "Counterfactual explanations can be manipulated," *Advances in neural information processing systems*, vol. 34, pp. 62–75, 2021.

[68] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.

[69] E. M. Kenny and M. T. Keane, "On generating plausible counterfactual and semi-factual explanations for deep learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 11575–11585, 2021.

[70] A. V. Looveren and J. Klaise, "Interpretable counterfactual explanations guided by prototypes," *CoRR*, vol. abs/1907.02584, 2019.

[71] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 93–104, 2000.

[72] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! criticism for interpretability," *Advances in neural information processing systems*, vol. 29, 2016.

[73] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki, "Unjustified classification regions and counterfactual explanations in machine learning," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II*, pp. 37–54, Springer, 2020.

[74] L. Lei and E. J. Candès, "Conformal inference of counterfactuals and individual treatment effects," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 83, no. 5, pp. 911–938, 2021.

[75] B. Bayrak and K. Bach, "A twin xcbr system using supportive and contrastive explanations," in *ICCBR 2023 Workshop Proceedings*, CEUR Workshop Proceedings, 2023.

[76] S. Aryal and M. T. Keane, "Even if explanations: Prior work, desiderata & benchmarks for semi-factual xai," *arXiv preprint arXiv:2301.11970*, 2023.

[77] E. M. Kenny and W. Huang, "The utility of" even if..." semifactual explanation to optimise positive outcomes," *arXiv preprint arXiv:2310.18937*, 2023.

[78] S. Mertes, C. Karle, T. Huber, K. Weitz, R. Schlagowski, and E. André, "Alterfactual explanations–the relevance of irrelevance for explaining ai systems," *arXiv preprint arXiv:2207.09374*, 2022.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3410540

[79] M. Virgolin and S. Fracaros, ‘‘On the robustness of sparse counterfactual explanations to adverse perturbations,’’ *Artificial Intelligence*, vol. 316, p. 103840, 2023.

[80] R. M. B. de Oliveira, K. Sörensen, and D. Martens, “A model-agnostic and data-independent tabu search algorithm to generate counterfactuals for tabular, image, and text data,” *European Journal of Operational Research*, 2023.

[81] E. Albini, J. Long, D. Dervovic, and D. Magazzeni, ‘‘Counterfactual Shapley Additive Explanations,’’ in *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’22, pp. 1054–1070, 2022.

BETÜL BAYRAK Betül is a Ph.D. candidate in Computer Science at the Norwegian University of Science and Technology (NTNU). They hold a B.S. in Computer Science from Gazi University (2017) and an M.Sc. from Çankaya University (2020). Betül’s current research focuses on eXplainable AI (XAI) and instance-based explanations.

KERSTIN BACH Kerstin is professor in Artificial Intelligence at the Norwegian University of Science and Technology (NTNU) and research director of the Norwegian Research Center for AI Innovation (NorwAI). She started her career at the German research center for AI (DFKI) developing decision support systems for various industries. Since then she’s has been responsible for an open-source tool, called myCBR, that has been used in numerous research and industry projects across Europe. In the past years her research focus is the development of AI prototypes in healthcare, intelligent sensing, and the explainability and trustworthiness of AI systems. She was manager of one EU H2020 research grant, selfBACK, whichs results are currently exploited in a number of european healthcare systems. Moreover, she has been involved in developing the European ecosystem for AI (AI4Europe platform) and an active member of the European AI community organising workshop, conference and symposia. Currently, she is the technical lead of a number of interdisciplinary projects that work on understanding healthcare data and developing novel healthcare services.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4