Testing Most Distant Neighbor (MDN) Variants for Semi-Factual Explanations in XAI

Saugat Aryal[1,2,*], Mark T. Keane[1,2]

[1]School of Computer Science, University College Dublin, Dublin, Ireland
[2]Insight Centre for Data Analytics, Dublin, Ireland

Abstract

Recently, Semi-factual explanations have gained popularity in the eXplainable AI (XAI) community. They provide *"even if"* justifications to indicate what key input features could change without changing the outcome. Although several methods have now been proposed to compute semi-factuals, the instance-based, Most Distant Neighbors (MDN) method has emerged in recent comprehensive tests to be quite competitive, even though was originally proposed as a naive benchmark. However, on some metrics MDN comes bottom of the class (e.g., sparsity). In this paper, we explore nine variants of the MDN method to determine whether its performance on some metrics can be improved relative to older methods by performing comprehensive tests on key metrics. The results show that there are MDN variants that perform better on some key metrics, but that some of the historical methods still do better.

Keywords

XAI, XCBR, Semi-Factual, Most Distant Neighbor

1. Introduction

Research on eXplainable AI (XAI) has received significant attention in recent years as it aims to shed light on the internal workings of the black-box AI models. As AI systems are being increasingly deployed in critical, high-stakes decision-scenarios (such as healthcare and finance), XAI methods play a key role in improving their interpretability and fostering human trust. Very recently, semi-factual explanations have emerged as a new explanation strategy [1, 2, 3], emerging from the exploration of counterfactual explanations. Semi-factuals provide explanations using *"even if"* reasoning, to highlight the key input-features that can be mutated without changing a predictive outcome. They are different from their counterfactual counterparts which provide explanations using *"only if"* reasoning, to highlight the key input-features that could be mutated to change a predictive outcome [4, 5, 6]. Although semi-factuals have been less studied, they have the potential to be as useful as counterfactuals due to their application in several areas such as healthcare diagnosis [7], decision space analysis [8, 9], data augmentation [10], causal analysis [3] and explanations for positive-outcomes [2].

Semi-factual as explanations has close ties with explanatory case-based reasoning (XCBR) research as their origin dates back to early 2000s. In these works, semi-factuals were used in the form of *a fortiori* arguments to provide "more" convincing explanations [11, 12, 13]. The researchers noted that in some scenarios, a case-based, nearest neighbor which was farther away from the query and closer to decision-boundary could provide a better explanation than the actual nearest neighbor (see Fig. 1). For example, consider a patient John who has a fever of 100.5°F and the decision system recommends release from the hospital. Instead of explaining this decision using Billy (his nearest neighbor) who was released with the fever of 100.1°F, it could be better to explain it using Matt (a more distant neighbour and a semi-factual) who was released with the fever of 102°F; the reasoning being that Matt is a more extreme case than John and Billy, so if it is ok to release Matt, it must be ok to release John. Accordingly, the received wisdom is that the further a semi-factual is from the query, the better it is as an explanatory case. Hence, semi-factuals offer novel perspectives on explaining the CBR systems.

ICCBR XCBR'24: Workshop on Case-Based Reasoning for the Explanation of Intelligent Systems at ICCBR2024, July 1, 2024, Mérida, Mexico

*Corresponding author.
saugat.aryal@ucdconnect.ie (S. Aryal); mark.keane@ucd.ie (M. T. Keane)

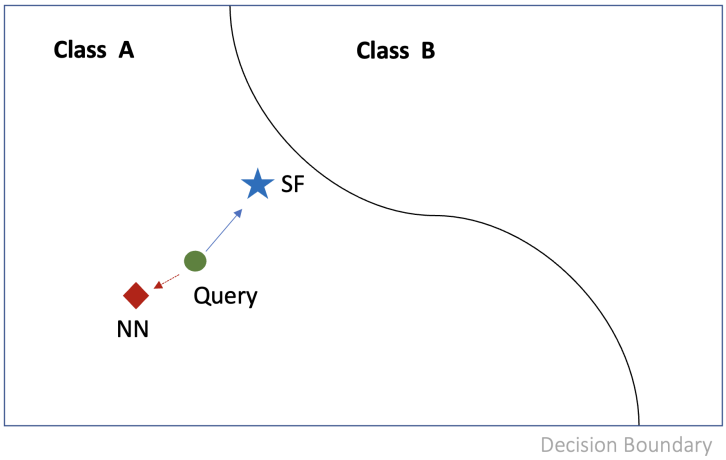

Figure 1: Consider a *Query* which lies in Class A. *NN* is the Nearest Neighbor of the Query while *SF* is the Semi-Factual which lies farther away from Query and closer to the decision boundary.

Kenny & Keane [3] advanced a recent counterfactual model that happened as a side-effect to generate useful semi-factuals as well. They pointed out that they were merely echoing very early XAI research in CBR, a literature that advanced a suite of models such as the Local Region model [13], and KLEOR [12] which had several variants using different similarity measures. Aryal & Keane [1] reviewed the history of this work and identified a rapid expansion of semi-factual work in both Machine Learning (e.g., augmentation, reject options) and XAI (e.g., explanation, interpretability). They also pointed semi-factuals lacked a naive model, by analogy to Nearest Unlike Neighbours (NUNs) for counterfactuals; NUNs are known data-points in a contrasting class to the query class that are, perhaps, the simplest counterfactual model one can design. Aryal & Keane argued that the equivalent entity to NUNs for semi-factuals would be Most Distant Neighbors (MDNs); namely, known data-points in the query class that are the furthest instances from a given query along its dimensions.

The underlying principle behind MDN is to balance a trade-off between two components: one that maximizes key feature-value differences and another that minimizes the number or sparsity of non-key features. As such, MDNs inherently realizes most of the computational desiderata for a "good" semi-factual explanation [1]. Moreover, although it was proposed as a naive benchmark against which more sophisticated models could be compared; this model turned out to be very competitive compared to historical CBR methods [1]. These findings position MDNs as a promising direction for further research. The model, however, has its own caveats that warrant further exploration.

The mean ranks of benchmark semi-factual methods reported by Aryal & Keane shown in Figure 2 indicates the overall success of these methods. The methods were evaluated on key evaluation metrics that attempt to assess semi-factuals for their computational characteristics (detailed in Section 4.1). The results suggest that MDN is able to find semi-factuals that are farthest away from the query in both feature (Query-to-SF Distance) and instance space (Query-to-SF kNN). However, it falls behind on three measures: distance to the query's class distribution (SF to Query-Class Distance), distance to the NUN (SF-to-NUN Distance) and Sparsity. It is desirable that semi-factual lies close to the data-manifold of query class. However, since MDN is finding the most distant semi-factual it may as well be selecting outliers lying on the edge of distribution making it far from the manifold. Similarly, semi-factuals are expected to lie close to the decision boundary and hence, close to the NUN. However, since MDN could be identifying outliers, they could lie farther from decision space and NUN. Finally, fewer feature differences is desired between query and semi-factual. However, MDN could be internally prioritising more on maximizing the feature-value difference at the cost of higher feature differences to yield poor sparsity results. Moreover, when measuring sparsity, the continuous features are considered "same" if the values fall within the range obtained by using a threshold of "20% of standard deviation". The threshold represents a hyperparameter which is empirically selected and lacks sufficient experiments.

These shortcomings motivate the present study, which aims to address the limitations of MDN and

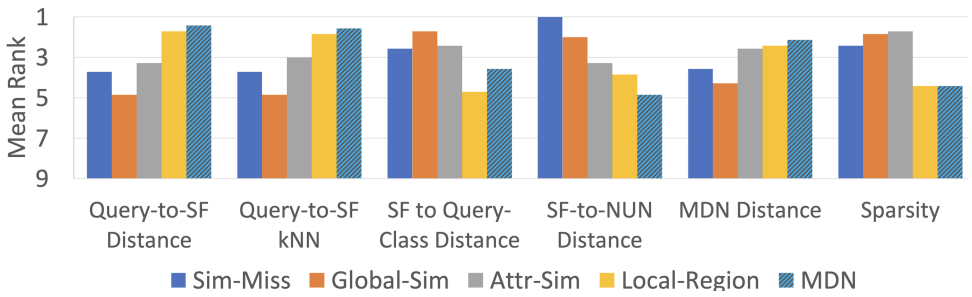

Figure 2: Mean Ranks of Benchmark Semi-Factual Methods on 6 Evaluation Measures across 7 Datasets. Adapted from Aryal & Keane [1]

improve their overall performance. As such, this work advances the research on MDNs for semi-factuals and makes several novel contributions:

- Introduction of two new MDN variants optimized to overcome its initial limitations.
- A comprehensive analysis of MDN variants in different threshold settings along with historic CBR methods.
- A thorough discussion on the use of MDNs for Semi-factual explanations and directions for future development in this area.

2. Background

MDNs are instance-based semi-factual explanation methods, in that, they use an existing datapoint to explain the query's prediction. They were recently compared with XCBR methods that are also inherently instance-based (KLEOR-variants [12] and Local-Region [13]) in the benchmarking study by Aryal & Keane [1]. In this section, we formally introduce the original MDN, KLEOR and Local-Region methods used in this study.

2.1. Most Distant Neighbor (MDN)

Aryal & Keane [1] proposed a novel method, MDN which reflects many of the desiderata for a semi-factual explanation. MDN works by finding an instance in the query-class which is farther away from it on some key-feature(s) while also being similar to it on other non-key features. As such, it finds the most distant neighbor of the query to be deemed as the semi-factual.

The algorithm first partitions the query-class instances into two sets: Higher and Lower Set based on the feature-values being higher or lower than the query along the selected feature-dimension. It then uses a custom distance function, *Semi-Factual Scoring (sfs)* which scores the candidate instances based on two components: extremity of feature-values on selected key-dimension and sparsity across remaining dimensions, as follows:

$$sfs(q, S, F) = \frac{diff(q_f, x_f)}{diff_{max}(q_f, S_f)} + \frac{same(q, x)}{F} \tag{1}$$

where S is Higher/Lower Set and $x \in S$, *diff()* gives the feature-value difference on key-feature, f, and $diff_{max}()$ is the maximum feature-value difference for that key-feature in the Higher/Lower Set, *same()* measures the number of similar non-key features between q and x, F is the total number of features. The instance with the highest *sfs* score along each dimension is selected as the best-feature-MDN independently. Finally, the best of the best-feature-MDNs across all dimensions is selected to be the semi-factual for the query as shown in Fig 3.

$$\text{SF}_{\text{MDN}}(q, S) = \arg\max_{x \in S} sfs(x) \tag{2}$$

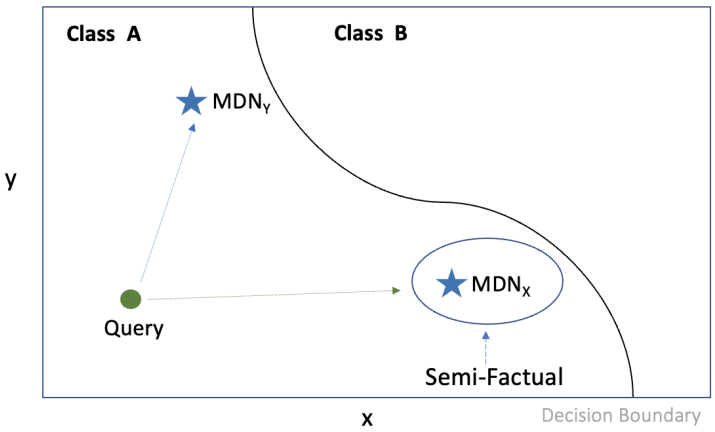

Figure 3: Consider a 2-dimensional feature space with features, **x** and **y**. MDN_X is the Most Distant Neighbor along **x** for Query while MDN_Y is the MDN along **y**. MDN_X achieves higher *sfs()* score than MDN_Y as it is farther from the query along **x** while also being most similar to query along **y**. Hence, MDN_X is selected as the best MDN and hence, Semi-Factual for the Query and **x** is chosen as the final key-feature.

2.2. Knowledge-Light based Explanation-Oriented Retrieval (KLEOR)

In the CBR-era of *a fortiori* reasoning, Cummins & Bridge [12] proposed a similarity-based approach to retrieve such explanations. They compute similarity between the instances in the query-class and the Nearest Unlike Neighbor (NUN) to find the best semi-factual for a given query. The method relies on the intuition that a semi-factual to a query is more similar to its NUN and hence use them as a guide. KLEOR has three different variants based on how the decision boundary partitions the feature space.

The first variant, *Sim-Miss*, selects a query-class instance which is most similar to the NUN to be the semi-factual as:

$$\mathrm{SF}_{\text{Sim-Miss}}(q, nun, N) = \arg\max_{x \epsilon N} Sim(x, nun) \tag{3}$$

where q is the query, x is the instance, N represents the set of all instances in the same class as the query, *nun* is the NUN, and *Sim* computes the Euclidean similarity in the feature space. It is the most naive variant as it reasons that the decision boundary neatly divides the feature space. The second variant, *Global-Sim* method considers complex decision boundaries which induces discontinuous feature-spaces. Hence, it imposes an additional similarity constraint which requires the semi-factual to lie between q and *nun* as:

$$\mathrm{SF}_{\text{Global-Sim}}(q, nun, N) = \arg\max_{x \epsilon N} Sim(x, nun) + Sim(q, x) > Sim(q, nun) \tag{4}$$

Finally, the third variant, *Attr-Sim* is a sophisticated version of *Global-Sim* in that it considers all the feature-dimensions. It computes similarities across each feature-attribute, ensuring that the semi-factual lies between the q and *nun* across the majority of features:

$$\mathrm{SF}_{\text{Attr-Sim}}(q, nun, N) = \arg\max_{x \epsilon N} Sim(x, nun) + \max_{a \epsilon F} count[Sim(q_a, x_a) > Sim(q_a, nun_a)] \tag{5}$$

where F is the feature-dimension set and a is a feature-attribute.

2.3. Local-Region Model

Nugent *et al.* [13] proposed another method for obtaining such explanation called Local-Region. This method analyses the local region around the query using a surrogate model, specifically logistic regression model. The surrogate model is built using subset of instances surrounding the query and hence essentially captures the local decision-space around it (akin to LIME [14]). The locally-trained

logistic regression model is then used to select a nearest neighbor with marginal-probability to be identified as the semi-factual explanation as:

$$\mathrm{SF}_{\text{Local-Region}}(q, N) = \arg\min_{x \epsilon N} LR(x) \tag{6}$$

where, *N* is the set of candidate neighbors and *LR()* is the local logistic regression model providing the probability score.

The intuition behind this approach is that a good semi-factual explanation should lie locally close to query but also as distant from it and near to its local decision boundary.

3. MDN Variants

The original naive MDN (henceforth, MDNv1) had few limitations even though it showed competitive results along some measures. The model uses a custom *sfs()* function to score each candidate instance based on their relative distance as well as closeness to the query. The scoring function however, can be modified to fine-tune the interplay between these two entities to optimize the behaviour of MDNv1. Along this line, we introduce two new MDN variants which uses novel scoring functions.

3.1. Sparse-MDNs

The baseline MDNv1 constitutes of two components aiming to find the furthest instance from query on some key-feature value while also being similar on other features. The two objectives are equally weighted such that the scoring function balances both the criteria to find the best semi-factual. However, since, MDNv1 performed relatively poorly in the sparsity metric (see Figure 5 in [1] and Figure 2), we propose Sparse-MDNs (MDNv2), which prioritises the similarity between non-key features. Essentially, we introduce a regularizer in the original *sfs()* function, which penalizes the algorithm for finding semi-factuals with higher feature-differences, thus promoting sparse explanations. We modify the scoring function by weighing it with the proportion of features that are "not same" between the query and the instance. Hence, instances with higher number of similar features will be assigned high scores to obtain sparse MDNs.

$$sfs_{v2}(q, S, F) = \frac{1}{F - same(q, x)} * \left(\frac{diff(q_f, x_f)}{diff_{max}(q_f, S_f)} + \frac{same(q, x)}{F}\right) \tag{7}$$

$$\mathrm{SF}_{\text{MDNv2}}(q, S) = \arg\max_{x \in S} sfs_{v2}(x) \tag{8}$$

3.2. Dist-MDNs

In both MDNv1 and MDNv2, the similarity between non-key features between query and instances using *same()* involves a direct comparison of their values. Specifically, the function checks if the values are identical in case of categorical features, while the continuous features are considered same if they fall within a predefined threshold range. However, it is not always straight-forward to determine the optimal threshold and it may vary across different features. Hence, we propose Dist-MDNs (MDNv3) where we modify the scoring function to compute similarity directly in the feature space as:

$$sfs_{v3}(q, S) = \frac{diff(q_f, x_f)}{diff_{max}(q_f, S_f)} * \frac{1}{dist(q_{nf}, x_{nf})} \tag{9}$$

where *dist()* computes the L_2-norm distance and q_{nf} and x_{nf} represents the query and instance with only non-key features (i.e excluding the key-feature in consideration) respectively. The final semi-factual is selected based on the highest scoring function as:

$$\mathrm{SF}_{\text{MDNv3}}(q, S) = \arg\max_{x \in S} sfs_{v3}(x) \tag{10}$$

4. Experimental Setup

We conduct a thorough comprehensive analysis of MDN variants (MDNv1, MDNv2 and MDNv3) along with the historic CBR methods (KLEOR and Local-Region) on benchmark metrics and datasets.

4.1. Evaluation Measures

We adopt the metrics proposed by Aryal & Keane [1] to evaluate the relative performance of the semi-factual methods.

- **Query-to-SF Distance.** The L_2-norm distance from the query to the semi-factual where higher scores are better as the semi-factual is preferred to be farther from the query.
- **Query-to-SF kNN (%).** It measures the amount of instances lying between query and semi-factual with respect to the total instances expressed as percentage. The higher scores are preferred again as the semi-factual is expected to be furthest instance from the query.
- **SF-to-Query-Class Distance.** It is a distributional measure which leverages Mahalanobis distance [15] to compute the closeness of semi-factual to the query's class distribution. Mahalanobis distance considers the variances and correlations between features to provide a more accurate measure of distance in the multi-dimensional space. Lower values are preferred which indicates that the semi-factual lies close to the query-class manifold.
- **SF-to-NUN Distance.** The L_2-norm distance from semi-factual to the NUN, where lower scores are better as the semi-factual is closer to the class boundary.
- **Sparsity (%).** It measures the percentage of semi-factuals obtained by the methods with single feature-difference with the query. Higher scores is desired for semi-factuals to be easily comprehended [16].

4.2. Method

We implement MDNv1 and MDNv2 with different threshold values $\alpha \in [20, 40, 60, 80]$ to comprehensively analyze their impact. The KLEOR variants were implemented using 3-NN model. The surrogate model in the Local-Region model was built using a minimum of 200 instances from each class. All the 13 methods were evaluated on 5 metrics across 7 benchmark tabular datasets commonly used in the XAI; namely, Adult Income (D1), Blood Alcohol (D2), Default Credit Card (D3), Diabetes (D4), German Credit (D5), HELOC (D6) and Lending Club (D7). We performed leave-one-out cross-validation to evaluate each method on each dataset. All the experiments were performed in Python 3.9 environment on a Ubuntu 23.10 server with AMD EPYC 7443 24-Core Processor.

4.3. Results & Discussion

The scores for 9 MDN variants, 3 KLEOR-methods and Local-Region on 5 metrics across 7 datasets is shown in Table 1 and 2. Figure 4 summarizes the overall performance of the methods as mean ranks. The results indicate that the new MDN variants (MDN_v2 and MDNv_3) show improved performance on the initial limitations of MDNv_1. Specifically, they achieve better scores in *SF-to-Query-Class Distance*, *SF-to-NUN Distance* and *Sparsity* measures. However, this improvement comes at the expense of their diminishing performance in *Query-to-SF Distance* and *Query-to-SF kNN* metrics. Nevertheless, MDNv2_20 achieved the highest rank overall aggregated across all the metrics and datasets closely followed by MDNv1_20. The older, Attr-Sim emerged as the third best overall and the best among the CBR methods.

MDNv2 and MDNv3 did not show any significant improvements on *Query-to-SF Distance* and *Query-to-SF kNN(%)* measures where MDNv1 still scored the best. However, they outperformed KLEOR-variants and showed competitive results on par with Local-Region. The KLEOR methods scored the best on *SF-to-Query-Class Distance* and *SF-to-NUN Distance* metrics. MDNv2 and MDNv3, however showed improved results on these measures relative to original MDNv1. In similar vein, the new variants,

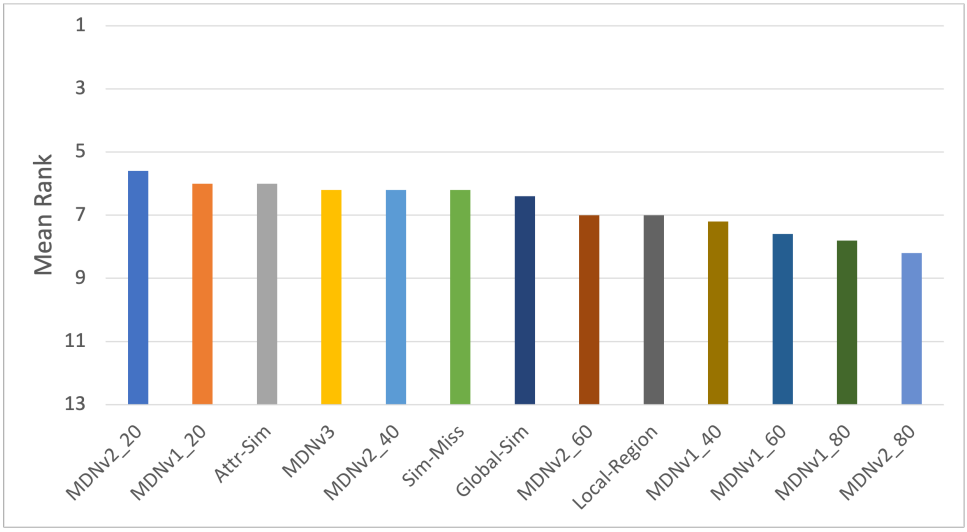

Figure 4: Mean Ranks of Each Method aggregated across Five Metrics and Seven Datasets.

specifically, MDNv2_20 and MDNv3 again performed best on *Sparsity* metric with improved results compared to MDNv1 while also outscoring KLEOR methods. The thresholds exhibited a subtle impact across all the measures. It can be observed that lower thresholds (20 and 40) tend to optimize their relative scores, whereas, higher thresholds (60 and 80) have the diminishing effect.

In summary, MDNs and its variants show best results in majority of the metrics while the historical methods still perform better on others.

Table 1
Sparsity values for each method across seven datasets. Higher scores indicate best performance. MDNv1_20 represents MDNv1 using a threshold of $\pm 20\%$ and so on.

Method	D1	D2	D3	D4	D5	D6	D7
MDNv1_20	2.95	0.2	0.06	0.01	0.02	0.01	0.21
MDNv1_40	2.95	0.2	0.06	0.01	0.02	0.01	0.21
MDNv1_60	2.24	0.01	0.05	0.01	0.02	0.02	0.11
MDNv1_80	1.74	0.01	0.04	0.02	0.02	0.01	0.07
MDNv2_20	13.88	4	0.21	0.02	0.4	0.01	2.23
MDNv2_40	4.51	0.7	0.1	0.02	0.4	0.01	0.6
MDNv2_60	3.13	0.02	0.05	0.02	0.4	0.01	0.26
MDNv2_80	2.1	0.02	0.05	0.02	0.4	0.01	0.17
MDNv3	27.71	6.15	0.31	0.02	0.02	0.01	0.24
Sim-Miss	6.51	1.95	0.12	0.01	0.14	0.01	0.23
Global-Sim	7.54	3.96	0.27	0.02	0.3	0.01	0.5
Attr-Sim	13.48	1.69	0.09	0.02	0.2	0.01	0.72
Local-Region	0.16	0.21	0.01	0.02	0.01	0.01	0.01

5. Conclusion

Semi-factual explanations are gaining considerable attention as a new paradigm in the XAI research. The recently proposed MDNs have shown promising results to serve as strong candidate, although they have their own limitations. In this work, we thoroughly analyzed nine different variants of MDNs along with four historic CBR-based methods on key evaluation metrics. The results suggest that MDNs could be optimized to improve their original performance, however still falls behind the old methods on some metrics. Nevertheless, MDNs for semi-factual explanations is a promising area of further research.

Table 2
Raw scores for each method on the four metrics across seven datasets. The arrows after metrics indicate the direction of best scores.

Metrics	Method	D1	D2	D3	D4	D5	D6	D7
Query-to-SF Distance (↑)	MDNv1_20	0.392	0.172	0.262	0.634	0.829	0.758	0.737
	MDNv1_40	0.371	0.484	0.4288	0.573	0.940	0.723	0.381
	MDNv1_60	0.364	0.488	0.415	0.539	0.935	0.692	0.356
	MDNv1_80	0.356	0.489	0.415	0.531	0.938	0.677	0.351
	MDNv2_20	0.271	0.350	0.211	0.533	0.905	0.652	0.181
	MDNv2_40	0.247	0.378	0.165	0.445	0.878	0.560	0.127
	MDNv2_60	0.246	0.447	0.159	0.385	0.855	0.483	0.157
	MDNv2_80	0.252	0.454	0.170	0.361	0.854	0.449	0.164
	MDNv3	0.236	0.131	0.111	0.386	0.518	0.328	0.344
	Sim-Miss	0.148	0.1695	0.085	0.352	0.455	0.322	0.055
	Global-Sim	0.135	0.146	0.073	0.297	0.385	0.262	0.046
	Attr-Sim	0.169	0.140	0.243	0.349	0.706	0.445	0.143
	Local-Region	0.346	0.279	0.206	0.666	0.910	0.490	0.1440
Query-to-SF kNN (%) (↑)	MDNv1_20	17.10	12.53	18.84	38.71	28.45	33.15	62.70
	MDNv1_40	14.42	33.9	25.15	32.54	36.18	31.13	29.43
	MDNv1_60	13.69	33.08	23.4	29.54	35.45	28.27	26.38
	MDNv1_80	12.71	32.96	23.29	28.24	35.38	27.49	25.82
	MDNv2_20	9.25	24.43	14.76	30.73	33.66	25.30	12.93
	MDNv2_40	7.09	25.47	9.25	23.48	32.56	18.62	5.21
	MDNv2_60	6.51	29.25	7.82	17.76	30.87	12.95	7.69
	MDNv2_80	6.04	29.1	8.25	15.13	30.59	10.18	7.99
	MDNv3	5.67	6.52	5.23	13.98	12.57	4.24	25.42
	Sim-Miss	2.49	11.71	1.13	15.91	4.49	2.42	0.30
	Global-Sim	1.97	9.92	0.816	10.24	2.72	1.03	0.22
	Attr-Sim	4.77	8.10	20.00	15.22	23.85	11.68	8.96
	Local-Region	11.99	21.51	5.07	52.85	39.79	12.09	2.50
SF-to-Query-Class Distance (↓)	MDNv1_20	3.25	2.15	3.24	2.80	4.14	4.37	2.85
	MDNv1_40	5.32	3.32	9.70	2.88	6.85	4.29	7.66
	MDNv1_60	5.33	3.38	9.64	2.88	6.85	4.29	7.56
	MDNv1_80	5.31	3.42	9.61	2.90	6.83	4.25	7.64
	MDNv2_20	4.22	2.73	3.47	2.63	6.59	4.08	3.25
	MDNv2_40	4.22	2.90	3.44	2.44	6.49	3.90	3.07
	MDNv2_60	4.27	3.22	3.52	2.41	6.40	3.83	3.42
	MDNv2_80	4.37	3.29	3.64	2.36	6.38	3.80	3.66
	MDNv3	3.18	2.29	3.44	2.72	4.02	3.89	2.77
	Sim-Miss	2.86	2.33	3.26	2.09	3.97	3.40	2.37
	Global-Sim	2.86	2.25	3.26	2.06	3.95	3.42	2.37
	Attr-Sim	2.93	2.10	2.93	2.13	4.30	3.49	2.39
	Local-Region	3.59	2.65	3.90	2.80	4.33	3.69	3.02
SF-to-NUN Distance (↓)	MDNv1_20	0.390	0.251	0.273	0.625	0.811	0.799	0.734
	MDNv1_40	0.373	0.489	0.428	0.578	0.946	0.713	0.380
	MDNv1_60	0.367	0.493	0.414	0.558	0.942	0.685	0.355
	MDNv1_80	0.359	0.491	0.414	0.551	0.943	0.674	0.351
	MDNv2_20	0.292	0.393	0.216	0.536	0.923	0.647	0.186
	MDNv2_40	0.271	0.422	0.172	0.481	0.899	0.565	0.134
	MDNv2_60	0.270	0.477	0.166	0.441	0.880	0.500	0.162
	MDNv2_80	0.273	0.482	0.176	0.418	0.879	0.473	0.168
	MDNv3	0.268	0.239	0.127	0.455	0.620	0.383	0.349
	Sim-Miss	0.041	0.057	0.035	0.194	0.289	0.200	0.021
	Global-Sim	0.053	0.068	0.044	0.227	0.335	0.236	0.028
	Attr-Sim	0.191	0.108	0.236	0.306	0.748	0.419	0.147
	Local-Region	0.318	0.225	0.200	0.582	0.864	0.473	0.144

Acknowledgments

This work has emerged from research conducted with the financial support of Science Foundation Ireland (SFI) to the Insight Centre for Data Analytics under Grant Number 12/RC/2289 P2.

References

[1] S. Aryal, M. T. Keane, Even if explanations: Prior work, desiderata & benchmarks for semi-factual xai, in: IJCAI-23, 2023, pp. 6526–6535. URL: https://doi.org/10.24963/ijcai.2023/732. doi:10.24963/ijcai.2023/732.

[2] E. Kenny, W. Huang, The utility of "even if" semifactual explanation to optimise positive outcomes, Advances in Neural Information Processing Systems 36 (2024).

[3] E. M. Kenny, M. T. Keane, On generating plausible counterfactual and semi-factual explanations for deep learning, in: Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI-21), 2021, pp. 11575–11585.

[4] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38.

[5] M. T. Keane, E. M. Kenny, E. Delaney, B. Smyth, If only we had better counterfactual explanations, in: Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI-21), 2021.

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

[7] A. Vats, A. Mohammed, M. Pedersen, N. Wiratunga, This changes to that: Combining causal and non-causal explanations to generate disease progression in capsule endoscopy, arXiv preprint arXiv:2212.02506 (2022).

[8] A. Artelt, B. Hammer, "even if..."–diverse semifactual explanations of reject, arXiv preprint arXiv:2207.01898 (2022).

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

[10] J. Lu, L. Yang, B. Mac Namee, Y. Zhang, A rationale-centric framework for human-in-the-loop machine learning, arXiv preprint arXiv:2203.12918 (2022).

[11] C. Nugent, P. Cunningham, D. Doyle, The best way to instil confidence is by being right, in: International Conference on Case-Based Reasoning, Springer, 2005, pp. 368–381.

[12] L. Cummins, D. Bridge, Kleor: A knowledge lite approach to explanation oriented retrieval, Computing and Informatics 25 (2006) 173–193.

[13] C. Nugent, D. Doyle, P. Cunningham, Gaining insight through case-based explanation, Journal of Intelligent Info Systems 32 (2009) 267–295.

[14] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD-16, 2016, pp. 1135–1144.

[15] M. P. Chandra, et al., On the generalised distance in statistics, in: Proceedings of the National Institute of Sciences of India, volume 2, 1936, pp. 49–55.

[16] M. T. Keane, B. Smyth, Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai), in: Proceedings of the 28th International Conference on Case-Based Reasoning (ICCBR-20), Springer, 2020, pp. 163–178.