Robustness Analysis of Counterfactual Explanations from Generative Models: An Empirical Study*

Kseniya Sahatova[1,2,*,†], Johannes De Smedt[1,†] and Xuefei Lu[2,†]

[1] KU Leuven, B-3000 Leuven, Belgium
[2] SKEMA Business School, 92156 Suresnes, France

Abstract

Counterfactual explanations are a key tool in explainable AI, offering insights into complex machine learning models by addressing *"What if?"* scenarios. While conventional methods for generating counterfactual explanations (CFEs) rely on computationally expensive optimization techniques, generative models such as GANs and VAEs have enabled faster CFE generation. However, their opaque nature raises concerns about trustworthiness, especially in high-stakes domains like healthcare and finance, where transparency and accountability are crucial. In this study, we benchmark existing methods for generating CFEs that apply generative models and connect them with a range of established metrics to assess robustness in both binary and multiclass image classification settings. Our analysis yields insights into the reliability of these approaches, while the proposed taxonomy organizes this rapidly evolving field through categorization based on CFE search methodologies.

Keywords

counterfactual explanations, explainability, robustness, image data

1. Introduction

The rapid development of generative models (GMs) aims to tackle complex tasks, such as identifying abnormalities in medical images, detecting fraud, and optimizing supply chains. As the demand for explainability and transparency in black-box AI models grows, counterfactual explanations (CFEs) have emerged as a tool in explainable AI that sheds light on decision-making processes by answering "What if?" question. CFEs generate plausible scenarios to provide alternative outcomes, enhancing model interpretability and trust. However, optimization-based methods for generating CFEs are computationally expensive [1]. The integration of GMs, such as Generative Adversarial Networks (GANs) [2] and Variational AutoEncoders (VAEs) [3], into the counterfactual explanation generation pipeline offers faster alternatives, although their opaque nature raises concerns about trustworthiness.

In high-stakes domains like healthcare, finance, and human behavior, black-box explainers require full transparency regarding both predictions and potential uncertainties. Attributes like proximity, sparsity, robustness, feasibility, and actionability are crucial. While extensive surveys ([1], [4], [5]) have categorized methods and benchmarks, they reveal that no single method satisfies all attributes simultaneously. Instead, methods balance these properties based on specific domain needs, enhancing the flexibility and adaptability of CFEs. Among these attributes, robustness, ensuring CFE stability despite input or model changes, remains underexplored for high-dimensional data like images. Robustness refers to the insensitivity of CFEs to small perturbations within guaranteed bounds, given that the model's prediction remains unchanged for the generated explanation. While significant progress has been made in evaluating the robustness of CFE generation methods [6], most work has focused on binary classification and tabular data.

In this research, we assess the robustness of methods based on generative models used for counterfactual explanations in a benchmark of three techniques with a lens on multiclass classification as most

Late-breaking work, Demos and Doctoral Consortium, colocated with the 3rd World Conference on eXplainable Artificial Intelligence: July 09–11, 2025, Istanbul, Turkey

[*] Corresponding author.

[†] These authors contributed equally.

kseniya.sahatova@kuleuven.be (K. Sahatova); johannes.desmedt@kuleuven.be (J. D. Smedt); xuefei.lu@skema.edu (X. Lu)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

of the robustness metrics in literature are primarily applied to binary classification tasks. Multiclass scenarios remain underexplored, where target class selection significantly impacts counterfactual quality due to varying distances in the data manifold and the proximity of decision boundaries. Additionally, we compare the stability of these methods with regard to task complexity, specifically binary and multiclass classification, and report the results based on the corresponding metrics. The considered research questions are as follows:

- **RQ 1:** Are the counterfactual explanations produced by generative models resistant to various forms of perturbations?
- **RQ 2:** Is the robustness of generated counterfactual explanations different in a multiclass setting compared to the binary classification setting?

2. Background and Related Works

In this section, the definition of counterfactual explanations for classification tasks is formalized and key components of the loss functions used by counterfactual explainers are outlined, which vary with model architecture and domain. For instance, GANs optimize an adversarial loss, while VAEs use an evidence lower bound. Further in the Related works subsection, we review recent work on GMs for CFEs and propose a taxonomy of these commonly employed pipelines.

Definition. A counterfactual explanation can be defined as follows. Given a predictive model f that maps the distribution of input data to a discrete class distribution, denoted as $f : X \to Y$, where $X \subseteq \mathbf{R}^d$ and $Y \in \{0, \ldots, n\}$, we define a counterfactual explanation for a factual data point x of the class y as $x_{cf} = G_{CF}(x, y, y'; f)$. G_{CF} is a generative model that can either output an explanation directly or a so-called difference mask, which must be applied to the factual data point. A valid counterfactual explanation satisfies the condition $f(x_{cf}) = y'$, where $y' \neq y$. Eq. 1 is the adapted framework proposed in [7] (Eq. 1,2) that we make more coherent with the introduced notation. The CFE algorithm is trained to minimize $L_{G_{CF}}$ as follows:

$$L_{G_{CF}}(x, x_{cf}, y') = \lambda_f L_f + \lambda_d L_d + \lambda_\chi L_\chi;\ \arg\min_{G_{CF}} L_{G_{CF}}(x, x_{cf}, y'), \tag{1}$$

where $L_f = d_f(f(x_{cf}), y')$ is a classification loss term, regularized by λ_f, encouraging the classifier's output on the counterfactual x_{cf} to be close to the desired class y'. $L_d = d(\cdot, \cdot)$ is a measure of the distance between the data point x and the counterfactual x_{cf} regularized with λ_d. L_χ is a generative model-dependent regularization term, penalizing out-of-distribution instances, typically formulated as an adversarial loss and/or a cycle-consistency loss. The number of components in the loss function may vary depending on the required properties for the generated CFEs.

Related Works. The development of GMs demonstrates not only their ability to produce high-quality synthetic data, but also their potential to create realistic and meaningful visual CFEs. Categorizing CFE methods based on GMs is challenging due to the complexity of models and their compound nature. Kirilenko et al. [8] outlined the literature on GMs for counterfactual explanations. In contrast, we focus on a higher-level classification based on CFE search mechanics rather than specific GMs. Our taxonomy aims to highlight emerging directions in integrating GMs for CFEs and should be seen as an extension, not a replacement, of existing benchmarks.

Latent space optimization/perturbation. The latent space of GMs offers a compact, adjustable representation for generating new instances. Applying simple linear interpolations in the latent space allows transformations of the latent vector [9], but these may not fully capture the complexities of decision boundaries learned by sophisticated classifiers. Singla et al. [10] apply walks with a fixed step size on the data manifold, which are then embedded in a low-dimensional space. C3LT [11] optimizes an external model to learn meaningful perturbations for steering predictions, and REVISE [12] uses constrained optimization with a pretrained VAE to modify the latent representation.

Disentanglement of latent space. Disentangling the latent space in GMs helps identify orthogonal factors that can be mapped to distinct, semantically meaningful concepts [13]. This approach can help to reveal biases in black-box models or data and enhance counterfactual explainability by detecting spurious correlations and enabling feature editing through factor manipulation. StyleGAN is used in [14] to extract human-understandable latent style vectors, concept disentanglers are employed in [15] to learn a predefined set of K concepts via cross-entropy loss, and Rotem et al. [16] enforce disentanglement by whitening the latent covariance matrix in an adversarial autoencoder.

Concept-based. Unlike tabular data, small image perturbations can lead to unrealistic adversarial examples instead of plausible explanations. GMs can facilitate operation at a more abstract conceptual level. STEEX [17] decomposes a latent vector into codes for semantic categories, while the work [18] encodes label-related concepts as binary latent variables, and Dominici et al. [19] combine a Concept Bottleneck Model with a VAE to model concept dependencies within a continuous latent space.

Residual Learning. Another widely adopted approach to generate CFEs involves learning differences or residuals that modify the initial input to achieve the desired result of the classifier. CounteRGAN [20] formalized residual GANs for identifying plausible CFEs. CX-GAN [21] generates discrepancy maps that transform abnormal instances into normal ones in a medical context without relying on a predictive model. COIN [22] applies a GAN conditioned on a flag for inpainting or removal of the abnormal region and a latent code. Van Looveren et al. [7] propose a framework for generating sparse, in-distribution CFEs for various data types, using a loss function tailored for desired properties of CFEs.

Diffusion-based. The limitations of GANs and VAEs have led to diffusion-based models for high-quality CFEs. Diffusion models reduce VAE blur while preserving quality and variability, which GANs struggle with. DiME [23] was among the first to use a diffusion model for explainability, combining an unconditional DDPM sampler with a guidance mechanism. [24] explored adversarial attacks to generate interpretable perturbations. LDCE [25] introduced a class-conditioned diffusion model with consensus guidance to filter misleading gradients.

3. Robustness of Counterfactual Explanations

Robustness property is particularly important due to its inherent trade-off with the proximity objective: proximity seeks minimal changes near the decision boundary, while robustness ensures explanations remain valid despite minor perturbations. We adopt the classification established by [6] for images, considering two groups of robustness to input and model changes that might affect the quality of explanations. Unlike other data modalities, images are highly sensitive to minor, often imperceptible perturbations that can generate adversarial examples, invalidating predictive outcomes. Therefore, the robustness of generative model-based methods for CFEs is highly relevant.

3.1. Input Changes

Local Instability (LI) Theoretical proofs and formalizations of robustness to input changes are present in [26], where it is defined as a measure of *local instability.* The authors opt for the L_1 norm without distance function restrictions in their experiments with tabular data and a handwritten digits dataset. The following Eq. 2 is a mathematical notation of local instability given by [26]:

$$\underset{x' \sim p_\epsilon(x)}{\mathbb{E}} \left[d(x_{\mathrm{cf}}, x'_{\mathrm{cf}}) \right] , \tag{2}$$

where x denotes the original instance, x_{cf} represents its CFE , x'_{cf} is the CFE for the perturbed instance x', p_ϵ represents the distribution of plausible perturbations around x, $d(\cdot)$ measures the similarity or distance between the CFEs of the original instance x and a perturbed instance x'.

Similarity estimation for images typically relies on L_p norms, which operate at the pixel level but ignore semantic differences like shapes and spatial correlations. Alternative metrics such as the Structural Similarity Index Measure (SSIM) and Peak Signal-to-Noise Ratio are commonly used in the related task of adversarial example generation [27]. In our experiments, SSIM [28] and L_1 norm are used.

Local Lipschitz Continuity Lipschitz continuity is a commonly used metric for evaluating the stability of post-hoc explanations [29]. It estimates the relative change in the output with respect to the variations in the input. Although it has been applied to assess the robustness of CFEs against model changes, it has not yet been utilized to evaluate robustness with respect to input dissimilarities. Eq. 3 below formalizes the estimation of local Lipschitz continuity of the explanations.

$$\hat{L}(x_i) = \max_{x_j \in N_\epsilon(x_i)} \frac{\|x_{cf_i} - x_{cf_j}\|_2}{\|x_i - x_j\|_2}, \tag{3}$$

where $N_\epsilon(x_i)$ represents an ϵ-ball centered at x_i, x_{cf_i} is its CFE, x_j and x_{cf_j} are an input instance sampled from the ϵ-ball and the corresponding generated CFE, respectively, where lower values of $\hat{L}(x_i)$ indicate more stable explanations.

3.2. Model Changes

This category of perturbations relates to variations in the predictive model f, which may be caused by model retraining, a common practice in real-world applications, alters decision boundaries and, consequently, the generated explanations. Prior studies have examined small perturbations caused by weight reinitialization or the removal of training data subsets [30].

Invalidation Rate (IR) In counterfactual consistency for deep neural networks, [31] suggests accounting for both the cost of generating stable explanations and the Lipschitz continuity of the predictive model in the vicinity of the counterfactual. The consistency of the explanations is measured by the Invalidation Rate (IR), given by:

$$\mathrm{IR}(x_{cf}, \theta) = \mathbb{E}_{\theta' \sim \tilde{\Theta}}[\mathbb{I}[f(x_{cf}; \theta') \neq f(x_{cf}; \theta)]], \tag{4}$$

where $f(x_{cf}; \theta)$ denotes the predictive model with parameters θ, and θ' represents the model parameters after variations in the training conditions.

Validity After Retraining (VaR) It is a commonly used evaluation metric that measures the percentage of CFEs that remain valid [30], i.e. belong to the same predicted class, under the retrained model. It can be considered the opposite of the IR. However, the latter provides only a relative estimation of the percentage of invalidated explanations after the model change introduced. In some cases, the counterfactual generation method itself might have a low validity. Comparing the validity of the perturbed predictive model with its initial counterpart can offer a broader perspective on the robustness and performance of the method. The validity can be defined as follows:

$$\text{Validity} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[f_{\theta'}(x_{cfi}) = t], \tag{5}$$

where n is a total number of counterfactual explanations, $f_{\theta\prime}(x_{cfi})$ is a prediction of the retrained model for the i-th counterfactual instance, and t is the target class.

Relaxed Stability (RS) The counterfactual stability metric was adapted for differentiable models based on the concept of local Lipschitz continuity in [30]. As exact Lipschitz estimation is often impractical, the authors derive a *relaxed stability* metric (Eq. 6), where the Lipschitz constant is approximated. The following properties are considered essential for generating robust explanations under naturally occurring model changes, enabling the proposed relaxation: (i) high model confidence for an input x_{cf}, denoted as $f(x_{cf})$; (ii) elevated values of $f(x'_{cf})$ for several points x'_{cf} in close proximity to x_{cf}; (iii) low variability in model outputs around x_{cf}.

$$\hat{R}_{k,\sigma^2}(x_{cf}, f) = \frac{1}{k} \sum_{x_{cf,i} \in N_{x_{cf},k}} \left(f_{\theta'}(x_{cf,i}) - |f_{\theta'}(x_{cf}) - f_{\theta'}(x_{cf,i})|\right), \tag{6}$$

where $N_{x_{cf},k}$ represents a set of k points sampled from a Gaussian distribution $\mathcal{N}(x_{cf}, \sigma^2 \mathbf{I}_d)$, with $\mathbf{I}_d$ being the identity matrix.

4. Experiments

4.1. Experimental Design

Datasets and classifiers. Most methods for generating CFEs focus on binary classification, where the target class is the opposite of the initial prediction. Multiclass scenarios, however, remain underexplored. In these cases, target class selection affects counterfactual quality, especially when the target class is farther in the data manifold, increasing explanation costs. For instance, class 7 is farther from class 8 than from class 0, based on cosine similarity and class centroid distances in the MNIST manifold. Additionally, multiple decision boundaries may challenge the stability of these explanations.

We use the MNIST dataset for experiments in binary and multiclass classification tasks. In the binary setting, class 1 is used for factual instances, with class 8 as the target for explanations. In the multiclass setting, class 8 remains the target while considering multiple initial classes. Factual instance classes are determined based on cosine similarity of features from the penultimate model layer, selecting the five closest classes (0, 3, 4, 5, and 9) to class 8. A simple CNN with 3 convolutional layers and max pooling achieved 99.95% accuracy in binary and 98.16% in multiclass classification.

Counterfactual explainers. The analyzed counterfactual explainers covered diverse approaches and model types, as detailed in Section 2, with code availability also considered. For example, STEEX [17] requires segmentation masks for training instances, CF-CBM [19] relies on annotated concepts in the data, and COIN is designed only for binary classification. Thus, REVISE [12], CounteRGAN [20], and C3LT [11] were selected for the initial experiments.

Robustness evaluation. For optimization-based methods like REVISE, robustness evaluation is computationally expensive, so we limited the number of factual instances to *k=100.* In contrast, CounteRGAN generates counterfactuals efficiently, allowing for more perturbed inputs. To estimate LI, we applied incremental Gaussian noise $\epsilon = \{0.001, 0.0025, 0.005, 0.0075, 0.01\}$, ensuring visual similarity. Estimating local Lipschitz continuity involves sampling from the ϵ-ball around an input and generating multiple explanations per sample. The number of sampled points around a given counterfactual explanation is set to 30 for REVISE, to reduce computational costs, and 50 for CounteRGAN and C3LT.

The metrics in Section 3.2 are evaluated on 10 perturbed CNN models, averaging results. Naturally occurring model changes, theoretically justified in [30], are introduced by reinitializing model weights with adjusted random seeds. We estimate IR and RS using explanations from the original test images.

Results. The primary results of our study are summarized below. To gain a better understanding of the quality of the generated explanations, we evaluate not only local instability w.r.t. (L_1) distance and *SSIM* but also the validity (the same Eq. 5) of the methods. The results of *LI* towards small input perturbations for both binary and multiclass settings are depicted in Fig. 1. The validity results indicate that REVISE is capable of generating CFEs that consistently lead the classifier to the same target output and remain stable across tested noise magnitudes. In contrast, CounteRGAN achieves only 43% validity at a perturbation level of 0.001, which further declines to 36% at a noise level of 0.01 in the multiclass classification task. For binary classification, CounteRGAN attains a maximum validity of 15% at a noise level of 0.001 and drops to 9%. C3LT maintains consistent validity across all multiclass experiments, with all generated explanations correctly classified as digit 8. However, in the binary setting, the validity of the algorithm is lower, reaching only 67%.

Regarding $LI(L_1)$ metric, REVISE exhibits stable average values even at higher noise levels, despite having a higher standard deviation. Furthermore, the distances between CFEs generated for original and perturbed inputs are greater in multiclass classification, presumably due to the richer semantics of the selected classes and, consequently, the larger number of transformed pixels during CFE generation. It is worth noting that the method optimizes the latent code of the perturbed input, which may converge along a different gradient path, potentially resulting in a different yet valid explanation. It can be seen in Fig. 1 (a), where $LI(SSIM)$ is rather low for both binary and multiclass tasks.

In contrast, explanations generated by CounteRGAN deviate more from the original explanations as the noise magnitude increases (Fig. 1 (b)). However, the average *LI* remains lower compared to REVISE in scenarios involving multiple classes. The method's poor performance in the binary setting

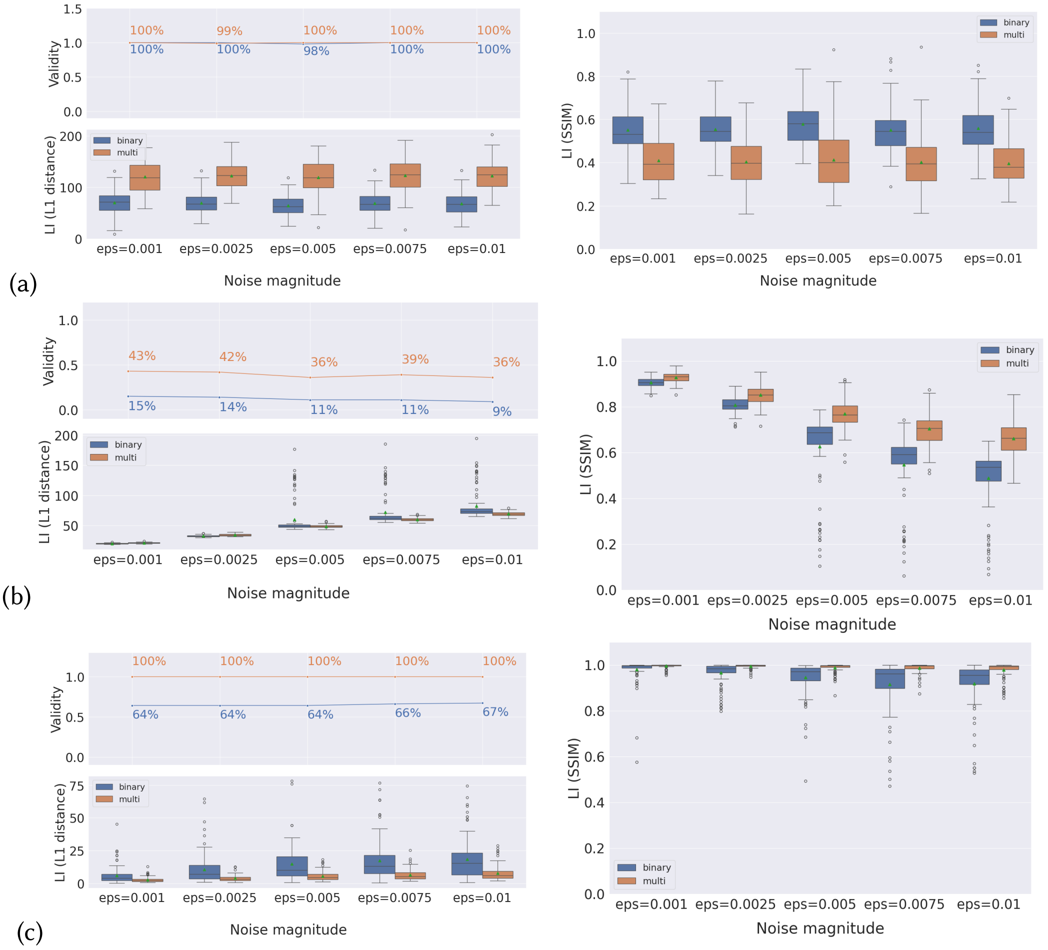

Figure 1: Validity and LI estimation w.r.t. L_1 and SSIM for REVISE (a), CounteRGAN (b), and C3LT (c) methods.

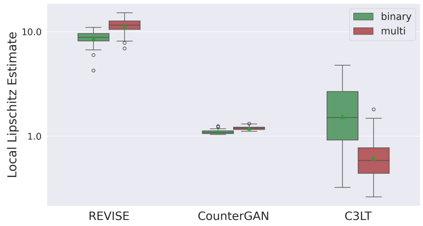

Figure 2: Estimates of Local Lipschitz Continuity.

	IV ↑	IR ↓	VaR ↑	RS ↑
Methods	Binary Classifiaction			
REVISE	1.0	0.195 (0.11)	**0.805** (0.11)	**0.787** (0.11)
CounteRGAN	0.15	0.015 (0.02)	0.135 (0.02)	0.75 (0.11)
C3LT	0.65	**0.005** (0.01)	0.645 (0.01)	0.644 (0.004)
Methods	Multiclass Classifiaction			
REVISE	1.0	0.046 (0.02)	0.954 (0.02)	0.924 (0.02)
CounteRGAN	0.47	0.093 (0.04)	0.37 (0.04)	0.613 (0.04)
C3LT	1.0	**0.0** (0.0)	**1.0** (0.0)	**1.0**(0.0)

Table 1: CFE robustness against model changes, mean (std).

is reflected in the results of $LI(SSIM)$, which indicate a deterioration in CounteRGAN's ability to generate perceptually similar explanations, which is less pronounced in the results of $LI(L_1)$. A similar pattern is observed for C3LT (Fig.1 (c)), although the average distances in all perturbations remain relatively low. Unlike REVISE that optimizes the latent code directly, C3LT uses this additional model g that learns the mapping of the given latent code of a factual class to the target class. The *local Lipschitz continuity* estimates are shown in Figure 2. This metric effectively reflects the conclusions on local instability, demonstrating a greater dissimilarity between counterfactuals generated by REVISE compared to those produced by CounteRGAN and C3LT.

Comparing the robustness of explanations against model changes, we additionally present the initial validity (IV) results of the unperturbed classifier in Table 1. REVISE shows a higher *IR* in binary settings, which is 0.195, than in multiclass - 0.046, consistent with *VaR* being the inverse of *IR*. The multiclass classifiers achieve an average *RS* of approximately 0.92, whereas binary classifiers attain only 0.79. CounteRGAN exhibits an inverse trend in terms of *IR*, with higher initial validity but greater invalidation in the multiclass setting. Nevertheless, the *RS* results reveal a different pattern: multiclass classifiers perform worse on the generated explanations validated by the unperturbed classifier. For this method, validity is initially low in either settings: 15% in the binary setting and 47% in the multiclass setting. C3LT provides only 65% of valid explanations in the binary problem, while reaching 100% in the multiclass task. The *IR* is 0 for the latter setting and constitutes only 0.001 for the former. However, the *RS* of slightly perturbed binary classifiers drops significantly compared to multiclass scenarios.

5. Conclusion

In this work, we present preliminary results on the robustness of counterfactual explanation methods based on generative models. Our proposed taxonomy provides structure to this rapidly evolving field by categorizing solutions according to their architectural properties. The evaluation reveals that robustness requires joint assessment across multiple metrics, with binary classification unexpectedly exhibiting greater fragility than multiclass scenarios despite its simpler decision boundaries. For RQ1, CounteRGAN and C3LT produced counterfactuals with greater deviation under perturbations, while REVISE better preserved explanation quality. However, perceptual analysis revealed that REVISE may still generate visually distinct explanations from minimally perturbed inputs, posing concerns in sensitive applications. For RQ2, multiclass tasks increased explanation costs due to more semantic features, reflected in REVISE's local instability. CounteRGAN and C3LT showed low validity in binary settings, declining with perturbations. All methods were less stable under model changes, with binary classification surprisingly showing higher invalidation rates. Overall, these findings reveal critical trade-offs between explanation quality, stability, and task complexity that must be addressed for reliable deployment in sensitive domains. The code is publicly available on GitHub [1].

Declaration on Generative AI

During the preparation of this work, the author(s) used Chat-GPT in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] R. Guidotti, Counterfactual explanations and how to find them: literature review and benchmarking, Data Mining and Knowledge Discovery 38 (2024) 2770–2824.

[2] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Advances in neural information processing systems 27 (2014).

[3] D. P. Kingma, M. Welling, et al., Auto-encoding variational bayes, 2013.

[4] S. Verma, V. Boonsanong, M. Hoang, K. E. Hines, J. P. Dickerson, C. Shah, Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review, 2022.

[5] A.-H. Karimi, G. Barthe, B. Schölkopf, I. Valera, A survey of algorithmic recourse: Contrastive explanations and consequential recommendations 55 (2022).

[6] J. Jiang, F. Leofante, A. Rago, F. Toni, Robust counterfactual explanations in machine learning: A survey, in: K. Larson (Ed.), Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24, International Joint Conferences on Artificial Intelligence Organization, 2024, pp. 8086–8094. Survey Track.

[7] A. Van Looveren, J. Klaise, G. Vacanti, O. Cobb, Conditional Generative Models for Counterfactual Explanations, 2021.

[1] https://github.com/kSahatova/CF-Robustness-Benchmark.git

[8] D. Kirilenko, P. Barbiero, M. Gjoreski, M. Luštrek, M. Langheinrich, Generative models for counterfactual explanations, View Article (2024).

[9] M. Y. Michelis, Q. Becker, On linear interpolation in the latent space of deep generative models, in: ICLR 2021 Workshop on Geometrical and Topological Representation Learning, 2021.

[10] S. Singla, B. Pollack, J. Chen, K. Batmanghelich, Explanation by Progressive Exaggeration, 2020.

[11] S. Khorram, L. Fuxin, Cycle-consistent counterfactuals by latent transformations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10203–10212.

[12] S. Joshi, O. Koyejo, W. Vijitbenjaronk, B. Kim, J. Ghosh, Towards realistic individual recourse and actionable explanations in black-box decision making systems, CoRR abs/1907.09615 (2019).

[13] A. Pandey, M. Fanuel, J. Schreurs, J. A. Suykens, Disentangled representation learning and generation with manifold optimization, Neural Computation 34 (2022) 2009–2036.

[14] S. Sankaranarayanan, T. Hartvigsen, L. Oakden-Rayner, M. Ghassemi, P. Isola, Real world relevance of generative counterfactual explanations, in: Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022, 2022.

[15] A. Ghandeharioun, B. Kim, C.-L. Li, B. Jou, B. Eoff, R. W. Picard, Dissect: Disentangled simultaneous explanations via concept traversals, arXiv preprint arXiv:2105.15164 (2021).

[16] O. Rotem, T. Schwartz, R. Maor, Y. Tauber, M. T. Shapiro, M. Meseguer, D. Gilboa, D. S. Seidman, A. Zaritsky, Visual interpretability of image-based classification models by generative latent space disentanglement applied to in vitro fertilization, Nature communications 15 (2024) 7390.

[17] P. Jacob, Zablocki, H. Ben-Younes, M. Chen, P. Pérez, M. Cord, STEEX: Steering Counterfactual Explanations with Semantics, 2022.

[18] I. Gat, G. Lorberbom, I. Schwartz, T. Hazan, Latent space explanation by intervention, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 679–687.

[19] G. Dominici, P. Barbiero, F. Giannini, M. Gjoreski, G. Marra, M. Langheinrich, Climbing the Ladder of Interpretability with Counterfactual Concept Bottleneck Models, 2024.

[20] D. Nemirovsky, N. Thiebaut, Y. Xu, A. Gupta, CounteRGAN: Generating counterfactuals for real-time recourse and interpretability using residual GANs, in: Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence, PMLR, 2022, pp. 1488–1497.

[21] T. Zia, Z. Nisar, S. Murtaza, Counterfactual Explanation and Instance-Generation using Cycle-Consistent Generative Adversarial Networks, 2023.

[22] D. Shvetsov, J. Ariva, M. Domnich, R. Vicente, D. Fishman, COIN: Counterfactual inpainting for weakly supervised semantic segmentation for medical images, 2024.

[23] G. Jeanneret, L. Simon, F. Jurie, Diffusion models for counterfactual explanations, in: Proceedings of the Asian conference on computer vision, 2022, pp. 858–876.

[24] G. Jeanneret, L. Simon, F. Jurie, Adversarial counterfactual visual explanations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 16425–16435.

[25] K. Farid, S. Schrodi, M. Argus, T. Brox, Latent diffusion counterfactual explanations, arXiv preprint arXiv:2310.06668 (2023).

[26] A. Artelt, V. Vaquet, R. Velioglu, F. Hinder, J. Brinkrolf, M. Schilling, B. Hammer, Evaluating Robustness of Counterfactual Explanations, 2021.

[27] M. Sharif, L. Bauer, M. K. Reiter, On the suitability of lp-norms for creating and preventing adversarial examples, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2018, pp. 1605–1613.

[28] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE transactions on image processing 13 (2004) 600–612.

[29] D. Alvarez-Melis, T. S. Jaakkola, On the robustness of interpretability methods, arXiv preprint arXiv:1806.08049 (2018).

[30] F. Hamman, E. Noorani, S. Mishra, D. Magazzeni, S. Dutta, Robust Counterfactual Explanations for Neural Networks With Probabilistic Guarantees, in: Proceedings of the 40th International Conference on Machine Learning, PMLR, 2023, pp. 12351–12367. ISSN: 2640-3498.

[31] E. Black, Z. Wang, M. Fredrikson, Consistent counterfactuals for deep models, in: International Conference on Learning Representations, 2022.