# Does Target Class Matter? A Reproducibility Study of Plausible Counterfactual Explanations for Image Classification

**Jasmin Kareem**[1,3,4], **Annelot W. Bosman**[2], **Meike Nauta**[3] and **Martijn C. Willemsen**[1,3], **Jan N. van Rijn**[2], **Maarten de Rijke**[4]
[1]Jheronimus Academy of Data Science
[2]LIACS, Leiden University
[3]Eindhoven University of Technology
[4]University of Amsterdam
{j.kareem, m.c.willemsen, m.nauta}@tue.nl, {a.w.bosman, j.n.van.rijn}@liacs.leidenuniv.nl, m.derijke@uva.nl

## Abstract

Counterfactual explanations have gained traction due to their contrastive and actionable nature: moving an outcome from the original class to an alternative target class. However, generating plausible and accurate counterfactuals remains challenging. This work presents a comprehensive empirical evaluation of multiple counterfactual explanation methods across all possible target classes on the MNIST and CIFAR-10 datasets, highlighting the target class as a critical but underexplored factor influencing counterfactual quality. To better assess plausibility, we introduce a novel evaluation method based on Moran's index, a spatial autocorrelation metric, which allows us to systematically identify and exclude structurally implausible counterfactuals that existing metrics may overlook. Our results show that methods often fail to generate counterfactuals for intended target classes due to timeouts, restrictive search spaces, or implementation issues, and that evaluating on only one target class provides an incomplete and potentially misleading picture of performance. Additionally, different plausibility metrics do not consistently agree, emphasising the need for more robust evaluation frameworks. In summary, we (i) identify the target class as a key overlooked dimension in counterfactual explanation performance, (ii) propose a novel plausibility assessment method using Moran's index, and (iii) provide actionable insights for the development of more generalisable counterfactual explanation methods.

## 1 Introduction

In recent years, the field of eXplainable Artificial Intelligence (XAI) has gained traction in making black-box models understandable [Nauta *et al.*, 2023]. As models and tasks have become increasingly complex, there is a growing need for post-hoc explanations that can accurately describe the inner workings of these models. A popular post-hoc explanation method is the counterfactual explanation, which asks 'what if' questions about model predictions: 'What if the inputs to my model had been different?'. In particular, counterfactual explanations are interested in minimally changing some factors to observe whether this small change would have affected the outcome. Whilst counterfactual explanations offer interesting insights into causal relations between a model's input and output on a per-instance level [Stepin *et al.*, 2021], the generated explanations cannot always be justified [Laugel *et al.*, 2019] and are not necessarily robust to model changes [Jiang *et al.*, 2023]. Additionally, prior theoretical work has shown that counterfactuals are always recourse sensitive and thus can never be fully robust [Fokkema *et al.*, 2023], exemplifying the need for rigorous evaluation.

A major challenge in XAI is that there is no agreed-upon strategy for evaluating the 'goodness' of an explanation [Hedström *et al.*, 2023; Nauta *et al.*, 2023]. Within counterfactual explanations specifically, there is no uniform and generally accepted set of evaluation techniques [Stepin *et al.*, 2021]. Benchmarks often use different implementations of the same algorithms and metrics [Karimi *et al.*, 2022; Le *et al.*, 2023], making it impossible to reliably compare methods. Furthermore, assessing counterfactual explanations using averages rather than performance gaps between subgroups can unfairly favour one method over another [Balagopalan *et al.*, 2022]. These trends risk further stifling the evaluation of counterfactual explanation methods and their use in practice.

Whilst there have been efforts to standardise counterfactual explanation methods for tabular data [Pawelczyk *et al.*, 2021; Verma *et al.*, 2020], analyse decision boundaries in time series classification [Baer *et al.*, 2025], and analyse post-hoc explanation methods in general [Bodria *et al.*, 2023; Klein *et al.*, 2024], the same level of analysis is lacking for counterfactual explanations in image classification. A notable exception is the work by Velazquez *et al.* [2023], where a wide variety of counterfactual explanation methods are evaluated. However, their results are based on a single synthetic dataset and do not analyse the complete scope of possible decision boundaries.

In this work, we address this gap by analysing the current state of counterfactual explanations for image classifiers and identifying key challenges affecting their evaluation and reli-

ability. We make the following contributions:

(1) We conduct a systematic empirical evaluation of five counterfactual explanation methods across two image classification datasets (MNIST and CIFAR-10), nine neural architectures, and all possible target classes – highlighting how the target class significantly affects counterfactual explanation quality. A detailed analysis of results across neural architectures is provided in the appendix.[1]

(2) We propose a novel plausibility evaluation method based on the spatial autocorrelation measure known as Moran's I [Moran, 1948], which enables the identification and exclusion of structurally implausible counterfactuals.

(3) We show that current methods often fail to generate counterfactuals for intended target classes due to timeouts, search space restrictions, or implementation issues.

(4) We demonstrate that existing plausibility metrics frequently disagree, underscoring the need for more robust evaluation tools.

Through our experiments,[1] we reproduce five different implementations of counterfactual explanations from Kenny and Keane [2021] and Van Looveren and Klaise [2021], and evaluate them using a variety of metrics, following a similar approach to Klein *et al.* [2024]. While we do not seek to add to the clutter of evaluation measures in XAI, we argue that Moran's I captures unique structural information relevant to plausibility and can complement existing metrics.

With our experimental design, we aim to answer the following research questions:

(RQ1) How do the different counterfactual explanation methods perform across a variety of evaluation metrics?

(RQ2) To what extent does the performance of different counterfactual explanation methods change for different decision boundaries?

(RQ3) To what extent do different implementations of plausibility agree with one another?

The paper is organised as follows. Section 2 provides background on counterfactual explanations and their evaluation. Section 3 introduces our proposed evaluation method. Section 4 details the experimental setup, including the counterfactual explanation methods and evaluation metrics. Section 5 presents the results, followed by discussion and conclusions in Section 6.

# 2 Background

## 2.1 Defining counterfactual explanations

To formally define counterfactual explanations, we must first consider the model task being explained. In image classification, the instance is the image being classified, and the perturbations are pixel changes within that instance. As broadly described by Guidotti [2022], given a classifier $f(x)$, where $x$ is the instance with decision $y = f(x)$, a counterfactual explanation $x'$ is a perturbed $x$ such that $f(x') \neq f(x)$, while minimising $\text{dist}(x', x)$ for some distance function $\text{dist}(\cdot, \cdot)$.

A counterfactual explainer may generate a set $C = \{x'_1, x'_2, \ldots, x'_h\}$ of examples, each encompassing different aspects of a single decision and relying on inherently different input values. Beyond this basic definition, counterfactual explanations are often evaluated against desirable properties: plausibility (the explanation should be realistic with respect to the original data distribution) [Laugel *et al.*, 2019], validity (satisfying $f(x') \neq f(x)$), minimality (no valid counterfactual $x''$ with a smaller distance to $x$ should exist), and similarity ($\text{dist}(x', x) < \epsilon$).

## 2.2 Counterfactual explanations for image classifiers

Since counterfactual explanations were initially described by Wachter *et al.* [2017] in a loan application setting, many adaptations have been developed to fit different data types. Initial applications in images proposed a minimum edit solution that greedily optimises towards a target image with the intended target class [Goyal *et al.*, 2019]. However, unlike in tabular data, the potential number of pixel changes is too large to generate an interpretable explanation. Thus, many approaches have turned to generative models such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) to ensure explanations lie close to the original image in the latent space [Guidotti *et al.*, 2019; Joshi *et al.*, 2019; Kenny and Keane, 2021; Liu *et al.*, 2019]. Other methods use prototypes [Van Looveren and Klaise, 2021], reinforcement learning [Samoilescu *et al.*, 2023], or energy-based modelling [Altmeyer *et al.*, 2024].

## 2.3 Evaluation metrics

The evaluation of counterfactual explanations remains an open problem. Nauta *et al.* [2023] provide an overview of metrics used in XAI research and their corresponding properties, known as the Co-12 properties. Several measures attempt to capture faithfulness, meaning how well the explanation reflects the model's behaviour. One approach is the faithfulness correlation metric introduced by Bhatt *et al.* [2020] for feature-based explanations. Other metrics incrementally flip pixels or change features to measure their effect on the model [Arya *et al.*, 2019; Bach *et al.*, 2015; Montavon *et al.*, 2018]. The main challenge in capturing faithfulness for counterfactual explanations is the lack of a ground truth, though synthetic datasets offer a possible solution [Liu *et al.*, 2021; Oramas *et al.*, 2019].

Plausibility is another commonly tested property, capturing how interpretable the explanation is and whether it is close to the original data distribution. Dhurandhar *et al.* [2018] train an autoencoder on the training dataset to ensure explanations remain close to the data manifold, and Van Looveren and Klaise [2021] take a similar approach with the IM1 and IM2 metrics. Kenny and Keane [2021] use Monte Carlo dropout for out-of-distribution detection, though this requires dropout layers in the explained network. The R% substitutability measure captures how well counterfactual explanations can

[1] Resources to support the reproducibility of this paper, including the appendix, are available at https://github.com/jasminkareem/CFX-benchmarking.

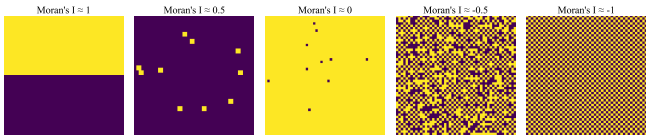



Figure 1: Illustration of spatial patterns corresponding to different values of Moran's I. From left to right, the images depict decreasing spatial autocorrelation: 1.0 (perfect positive autocorrelation, highly clustered), 0.5 (moderate clustering), 0 (random spatial pattern, no autocorrelation), –0.5 (moderate negative autocorrelation, alternating patterns), and –1.0 (perfect negative autocorrelation, complete spatial dispersion).


Figure 2: Original image with different counterfactual explanations produced by different methods, along with their MI scores. The original model prediction was a "7", and the target is a "3". In this case, the right most explanation has a lower MI score than the other explanations, mainly due to the checkerboard pattern in the bottom right corner.

substitute training data [Samangouei *et al.*, 2018], whilst Altmeyer *et al.* [2024] modify the plausibility metric of Guidotti [2022], measuring the distance of a counterfactual from its nearest neighbour in the target class.

## 3 Moran's Index as a Diagnostic Measure

As mentioned previously, there are many plausibility measures, which mostly try to capture how close a counterfactual explanation is to the original data distribution. Whilst this approach is conceptually sound, these methods often require generative models, and they may be hard to interpret. Thus, we propose using a simple statistic known as *Moran's I*. Global Moran's I is known as a measure of spatial autocorrelation and is often used in geographical analysis [Moran, 1948]. The values can range from $-1$ to 1. A simple illustrative example is provided in Figure 1 that visually shows these possible values. A Moran's I of 1 represents high spatial correlation, and a value close to 0 represents randomness. On the other hand, a Moran's I value of $-1$ indicates a perfect separation between each pixel. The index is generally defined as

$$I = \frac{n}{W} \cdot \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad (1)$$

where $n$ is the number of spatial units, in the case of images, these would be pixels. Then $x_i$ is the value of the variable of interest at location $i$ and $\bar{x}$ is the mean of the variable $x$ over all locations. Lastly, $w_{ij}$ is the spatial weight between location $i$ and location $j$ (typically from a chosen spatial weights matrix) and $W$ is the sum of all spatial weights, $W = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}$. Note that the spatial weight matrix decides which neighbouring features are important for $x_i$ and thus what is chosen can depend on what information is important in $x_i$. We choose to use the queen kernel, which counts all directions as possible neighbours. For example, in a $3 \times 3$ kernel, all 8 cells surrounding the cell of interest are considered neighbours and would be weighted with 1.

We assume that higher spatial autocorrelation relates to better interpretability of explanations, as illustrated in Figure 2. Thus, we propose using Moran's I (MI) as a diagnostic measure for explanation quality in images, as we believe it can filter out certain types of implausible explanations. We envision two ways of using it for explanations. The first is to simply compute the MI value of each explanation, where higher values are better. An alternative is to compute the absolute difference between the MI of the original image and the MI of the explanation. The reasoning for this is that the MI score might be more meaningful if we look at it relative to the original image. This means that lower values would be considered better.

## 4 Experimental Design

In this section, we describe the design of our experiments. First, we outline the networks that were trained and tested. Next, we detail the counterfactual explanation methods applied to interpret these models. Finally, we explain the evaluation metrics and procedures used to assess both model performance and the quality of the explanations.

### 4.1 Networks

To support a comprehensive evaluation, we train multiple architectures on two datasets. As our first dataset, we use MNIST [LeCun, 1998], as it is the most commonly used benchmark for counterfactual generation methods. The second dataset is CIFAR-10 [Krizhevsky, 2009].

The nine architectures used are summarised in Table 1 (see the appendix for implementation details). As some counterfactual explanation methods are written in PyTorch and others in Keras, we needed to translate the same neural architecture across these two different frameworks. For *mnist_mlp_relu_4_1024* we used an ONNX model to store the network weights, which could then be converted into either a Keras or Torch model using the learned weights. This led to many complications in aligning the weights to the network architecture in the corresponding package. For example, the method alibi-CF failed to produce counterfactual explanations for almost every instance and decision boundary, but this only occurred with the ONNX converted model. In every other case, we fully trained both Keras and PyTorch models separately, ensuring that the translated architecture was correct. Still, there were differences in test accuracy, the largest difference being 4.3% and 5.3% on the CNN and ResNet8 networks for CIFAR, respectively. For the other networks, this difference is either roughly equal to or less than 1%. Note that this means the underlying models were not necessarily perfectly the same for both Keras and Torch. Additional experiments analysing counterfactual explanation performance across these architectures are provided in the appendix.

### 4.2 Counterfactual explanation methods

We reproduce the Alibi-based counterfactual explanations [Klaise *et al.*, 2021; Van Looveren and Klaise, 2021] and

Table 1: Neural architectures and test performance on MNIST (M) and CIFAR-10 (C).

| Architecture | Torch | Keras | Data |
|---|---|---|---|
| MLP Relu_4_1024 | 0.9848 | 0.9848 | M |
| CNN [Kenny and Keane, 2021] | 0.9925 | 0.9934 | M |
| CNN [Van Looveren and Klaise, 2021] | 0.9720 | 0.9819 | M |
| LeNet5 [LeCun *et al.*, 1998] | 0.9726 | 0.9882 | M |
| ResNet8 | 0.9918 | 0.9924 | M |
| MLP [Schut *et al.*, 2021] | 0.9476 | 0.9659 | M |
| CNN [Kokhlikyan *et al.*, 2020] | 0.5779 | 0.6208 | C |
| ResNet8 | 0.7724 | 0.7196 | C |
| ResNet18 [He *et al.*, 2016] | 0.8387 | 0.8314 | C |

the PIECE method [Kenny and Keane, 2021] and its corresponding baselines. We chose to reproduce these methods because they represent different ways of generating counterfactual explanations. The main differentiating factor between all these methods is the loss function that is used to optimise towards a counterfactual. The most straightforward approach is the alibi-CF method, which is roughly based on the work of Wachter *et al.* [2017] and does not use any generative model to guide the explanation. Following this, the Min-Edit and C-Min-Edit both use a GAN $G$ and a latent representation of the original image $z$ in their optimisation step to get to the latent representation of the explanation $z'$. The main difference between these two is how they define the loss. The PIECE method also uses this same $z$ and $G$, but adds an additional factor using feature values in the original image that are less likely to occur in the target class, named exceptional features. Lastly, the alibi-Proto-CF method uses an autoencoder to add a prototype for each target class. A corresponding loss term is added to guide the explanation towards the target class and speed up the search process. We provide an overview of the optimisation step of each of the counterfactual explanation methods in Table 2.

Here $p_t$ denotes the target class probability and $f_t(\cdot)$ is the predicted class probability. The term $\lambda$ is a parameter that balances the trade-off between prediction accuracy and feature value similarity. The function $c(\cdot)$ denotes all layers of the classifier up until the penultimate layer and $f(\cdot)$ is the output of the classifier itself. The term $Y_{c'}$ represents the output of the classifier after the Softmax operation, yielding a probability vector corresponding to the target class $c'$. Lastly, for alibi-Proto-CF, $L_{\text{pred}}$ corresponds to the model's prediction function with scaling parameter $s$. The next two terms represent the elastic net regularisation, with parameter $\beta$, and $L_{AE}$ and $L_{\text{proto}}$ representing the autoencoder and prototype losses. There is a subtle distinction between the Alibi-based methods and the remaining approaches: in the latter, $z'$ denotes a latent vector, whereas in the Alibi methods, $L$ more generally represents a loss function.

These counterfactual explanation methods were chosen because Alibi is a better-known toolkit, often used as a baseline and should thus provide well-maintained functions. We include PIECE as we aim to reproduce it and it is a method focused on generating plausible counterfactual explanations. We generate counterfactual explanations for 100 instances on each potential target class.

Table 2: Overview of optimisation formulations used for counterfactual explanation methods in our experiments.

| Method (library) | Optimisation step |
|---|---|
| **alibi-CF (Keras)** | $L(x' \mid x) = \left(f_t(x') - p_t\right)^2 + \lambda L_1(x', x)$ |
| **Min-Edit (PyTorch)** | $z' = \arg\min_z \|f(c(G(z))) - Y_{c'}\|_2^2$ |
| **C-Min-Edit (PyTorch)** | $z' = \arg\min_z \max_\lambda \lambda \|f(c(G(z))) - Y_{c'}\|_2^2 + \text{dist}(c(G(z)), x)$ |
| **PIECE (PyTorch)** | $z' = \arg\min_z \|c(G(z)) - x'\|_2^2$ |
| **alibi-Proto-CF (Keras)** | $L = sL_{\text{pred}} + \beta L_1 + L_2 + L_{AE} + L_{\text{proto}}$ |

### 4.3 Evaluation metrics

We evaluate the counterfactual explanation methods using metrics for correctness, plausibility, size, and running time, which aligns with prior work on evaluating counterfactual explanations [Guidotti, 2022]. We focus on the following properties: correctness, success, size, running time and plausibility. The first metric is simply the time it takes to optimise a counterfactual explanation, referred to as **OT** in this paper. Then we are interested in computing the size of a counterfactual explanation. This is done using the $L_1$, $L_2$ and $L_\infty$ -norm between the original and the perturbed instance. Lastly, three proxy measures for plausibility are implemented, including our implementation of $MI$.

- **Validity/Correctness**: Does the new prediction of the counterfactual explanation change to the target class (excluding incorrect timeouts)?
- **Success**: How many counterfactual explanations can the method successfully produce (including timeouts)?
- **Success Ratio**: What fraction of the source-target pairs per image successfully produce a counterfactual explanation (including timeouts)?
- **Optimisation Time**: Time it took to run the method.
- **Size/Minimality**: $L_1$, $L_2$, and $L_\infty$ distance between $x$ and $x'$.
- **IM1/IM2** [Van Looveren and Klaise, 2021]: The IM1 and IM2 metrics are based on the reconstruction loss of autoencoders trained on each class:

$$\text{IM1}(x', y, y') = \frac{\|x' - AE_y(x')\|_2^2}{\|x' - AE_y(x')\|_2^2 + \epsilon} \tag{2}$$

$$\text{IM2}(x', y') = \frac{\|AE_y(x') - AE(x')\|_2^2}{\|x'\|_1 + \epsilon}, \tag{3}$$

where $AE_y$ represents the output vector of an autoencoder trained on a particular class $y$. We note that, based on prior research, IM2 is not necessarily interpretable [Mahajan *et al.*, 2019] and has been shown not to provide significantly different outcomes on out-of-distribution data [Schut *et al.*, 2021].

Table 3: Average performance of counterfactual explanation methods across all architectures for MNIST and CIFAR. Best-performing method per metric shown in **bold**. Values: mean ± std. (in parentheses). ST = success ratio; OT = optimisation time. *ST all* and *OT all* are over all explanations; other metrics are over correct counterfactuals only (those changing the prediction to the target class). Impl. = implausibility; MI = Moran's I. Both ΔMI (original vs. counterfactual) and MI(x') (counterfactual alone) are reported. ↑/↓ = higher/lower is better.

| Method | ST all↑ | OT all↓ | OT↓ | $L_2$↓ | IM1↓ | IM2↓ | Impl.↑ | ΔMI↓ | MI(x')↑ |
|---|---|---|---|---|---|---|---|---|---|
| **MNIST** | | | | | | | | | |
| C-Min-Edit | **.69** ±.27 | 33.82 ±26.05 | 20.67 ±16.63 | 15.70 ±3.83 | 3.31 ±9.82 | .109 ±.23 | .56 ±.06 | .08 ±.08 | .69 ±.11 |
| Min-Edit | .62 ±.28 | **19.43** ±15.78 | **9.63** ±9.85 | 15.84 ±3.79 | 3.46 ±10.35 | .108 ±.23 | .57 ±.06 | .08 ±.08 | .70 ±.11 |
| PIECE | .27 ±.24 | 28.48 ±13.86 | 10.62 ±11.19 | 15.49 ±4.03 | 3.03 ±10.57 | **.088** ±.20 | **.57** ±.06 | .08 ±.08 | **.70** ±.11 |
| alibi-CF | .61 ±.40 | 54.69 ±41.97 | 66.24 ±45.20 | **5.42** ±1.84 | 1.60 ±2.71 | .111 ±.20 | .50 ±.09 | **.06** ±.04 | .68 ±.09 |
| alibi-Proto-CF | .67 ±.29 | 111.39 ±58.34 | 96.01 ±47.39 | 23.71 ±8.31 | **1.08** ±.73 | .253 ±.16 | .14 ±.15 | .25 ±.23 | .50 ±.23 |
| **CIFAR** | | | | | | | | | |
| C-Min-Edit | **.99** ±.03 | 10.62 ±7.60 | 10.47 ±7.06 | 16.12 ±5.77 | 1.05 ±.24 | .0017 ±.0008 | .010 ±.02 | .05 ±.04 | .86 ±.08 |
| Min-Edit | .98 ±.06 | **6.83** ±6.93 | **6.42** ±5.98 | 16.13 ±5.77 | 1.05 ±.24 | .0017 ±.0008 | .010 ±.02 | .05 ±.04 | .86 ±.08 |
| PIECE | .75 ±.29 | 16.85 ±14.59 | 10.10 ±9.07 | **15.31** ±4.66 | 1.04 ±.24 | .0016 ±.0007 | **.011** ±.02 | .05 ±.04 | **.87** ±.07 |
| alibi-CF | .92 ±.18 | 107.99 ±77.33 | 112.16 ±78.70 | 32.12 ±6.41 | 1.08 ±.28 | **.0008** ±.0003 | −.006 ±.03 | **.01** ±.04 | .82 ±.09 |
| alibi-Proto-CF | .66 ±.18 | 298.58 ±101.32 | 287.69 ±91.23 | 36.70 ±8.97 | **1.02** ±.16 | .0040 ±.0040 | −.007 ±.02 | .35 ±.29 | .50 ±.31 |

- **(Im)-plausibility** [Altmeyer *et al.*, 2024]: Based on a definition by Guidotti [2022], implausibility is computed:

$$\text{impl}(x', X_{y_+}) = \frac{1}{|X_{y_+}|} \sum_{x \in X_{y_+}} \text{dist}(x', x), \qquad (4)$$

  where $X_{y_+}$ is a subsample of the training dataset for the class $y_+$. The distance function can be chosen; however, for images, the structural similarity method [Wang *et al.*, 2003] provides more insights into image similarity than the Euclidean norm would.
- **MI**: Moran's I of the explanation $MI(x')$. Alternatively, we compute $|MI(x) - MI(x')|$ or Δ MI as a potential variation of this metric.

We make a distinction between the *correctness* and the *success* of an explanation because an explanation is not necessarily incorrect, but just unsuccessful in finding a solution within the time limit. Had the explanation method been given more time, it may have produced something correct. Thus, we consider success to be a fairer evaluation of correctness.

## 5 Experimental Results

Our experimental results are structured according to our three research questions in Section 1. We first evaluate counterfactual explanation performance overall, then study the importance of target classes, and finally investigate correlations between plausibility metrics.

### 5.1 RQ1: Evaluating the performance of counterfactual explanation methods

The results addressing the performance of various counterfactual explanation methods across multiple evaluation metrics are presented in Table 3. Starting with validity, the success ratio is highest overall for C-Min-Edit. It is important to note that the table does not distinguish between success and correctness, particularly in scenarios involving incorrect instances that timeout. After excluding timed-out, incorrect, or both types of failed instances, both alibi methods consistently yielded correct counterfactual explanations, with one notable exception: nine incorrect counterfactual explanations from alibi-CF on *mnist_mlp_relu_4_1024*, all predicting the same class as the original label.

This raises the question: why do counterfactual explanation methods fail to produce correct results? Three primary reasons likely contribute. First, the method may time out, as each approach operates under a predefined iteration limit. Second, suboptimal hyperparameter settings might overly constrain the search space, as is particularly evident in C-Min-Edit, Min-Edit, and PIECE, where continuous increases in loss suggest the algorithm is stuck. Finally, bugs in the original source code can produce incorrect counterfactual explanations before a timeout is reached.

Regarding other aspects of Table 3, the smallest counterfactual explanations are generated by alibi-CF for MNIST and by PIECE for CIFAR. Optimisation time shows minimal

Table 4: A comparison of success and weak success as a percentage over the different methods across MNIST and CIFAR datasets.

| | MNIST | | CIFAR-10 | |
|---|---|---|---|---|
| **Method** | **Success** | **Weak** | **Success** | **Weak** |
| C-Min-Edit | 68.65 | 98.82 | 99.30 | 100.00 |
| Min-Edit | 61.69 | 97.64 | 98.48 | 100.00 |
| PIECE | 27.22 | 78.79 | 75.48 | 99.00 |
| alibi-CF | 60.55 | 76.60 | 91.74 | 100.00 |
| alibi-Proto-CF | 67.21 | 100.00 | 65.78 | 100.00 |

differences between all counterfactuals and correct ones only. Notably, the increased complexity of CIFAR significantly extends optimisation time for the alibi-based methods, whereas the other methods show improved efficiency. This is somewhat unexpected, but may be explained by the shared GAN architecture used across these methods: if the GAN performs better on CIFAR than on MNIST, this could positively impact performance on that dataset.

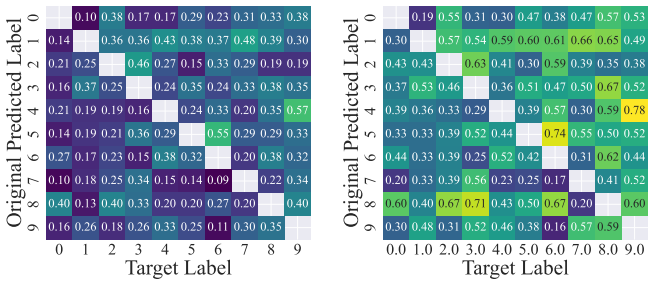
Regarding plausibility, PIECE demonstrates the best $MI(x')$ and implausibility scores across both datasets, with Min-Edit performing comparably. For the IM1 measure, alibi-Proto-CF yields the most favourable results. The simpler formulation $MI(x')$ proves more informative than $|MI(x) - MI(x')|$, as the latter produces substantially smaller variations across methods, making it less effective for distinguishing performance. This explains why alibi-CF has a high $MI(x) - MI(x')$ and the smallest size in MNIST: its explanations are likely very close to the original image and thus have a similar $MI$ score, without necessarily being more plausible. Interestingly, the method with the lowest $MI(x')$ overall is alibi-Proto-CF, which may relate to our anecdotal evidence in Figure 2.

In summary, there is no clear winner across metrics. However, the best-performing method per metric is mostly consistent across both datasets. The key insight is that effectiveness varies depending on which property is prioritised — size, optimisation time, or plausibility — as no single method excels across all criteria.

## 5.2 RQ2: Decision boundaries

To determine to what extent different decision boundaries impact explanation performance, Table 4 shows the percentage of successful and weakly successful explanations per method, where weak success requires only one successful decision boundary per image. It is clear that not all decision boundaries can be successfully reached: for instance, PIECE achieves only 27.22% full success on MNIST, whilst alibi-Proto-CF reaches 67.21%. Relaxing the success criterion leads to a substantial jump across all methods, with nearly every image yielding at least one successful counterfactual explanation.

For the more challenging CIFAR dataset, success rates are considerably higher overall, with most methods exceeding 90% full success. A notable exception is alibi-Proto-CF, which achieves only 65.78% full success despite reaching 100% weak success, suggesting it struggles to cross certain decision boundaries consistently. It should be noted, however, that higher success rates do not imply they are better, as




(a) Including timeouts (b) Excluding timeouts

Figure 3: Average Success for each original class and target class pair for the PIECE method on MNIST. Timeouts are either included in the average success calculation or filtered out.

the plausibility metrics in Table 3 are considerably lower for CIFAR overall.

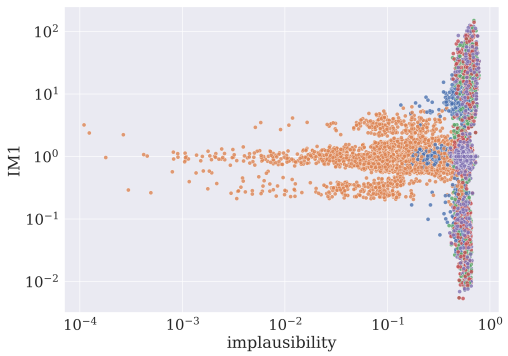
Additional evidence on specific original-to-target class transitions is presented in Figure 3 for the PIECE method, illustrating that certain class pairs are considerably easier to shift between. For example, moving from a 5 to a 6 or from a 4 to a 9 scores relatively well, whereas moving from a 0 to a 7 is far less successful — consistent with the intuition that visually and semantically similar classes are easier to transition between. Corresponding heatmaps for the remaining methods are provided in the appendix.

These results also highlight the influence of excluding unsuccessful timeouts. As shown in Figure 3b, removing timeouts leads to a consistent increase in average success across all decision boundaries, suggesting that a substantial portion of failures arises from methods exhausting their optimisation time. However, since success values do not reach 1 across all class pairs, additional factors — such as an overly constrained search space or code bugs — also contribute.

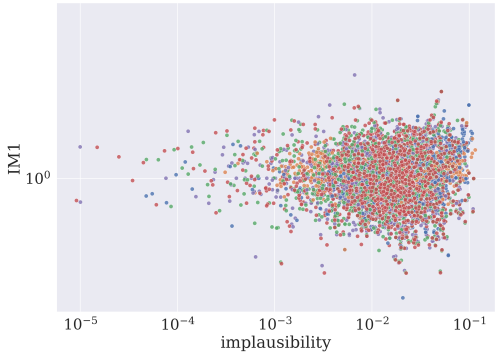
Taken together, our findings indicate that the choice of target decision boundary significantly influences counterfactual explanation performance, consistent with theoretical expectations: when the target boundary is too far from the original instance, finding a viable counterfactual within a constrained optimisation timeframe becomes increasingly infeasible. Some methods handle this better than others, and we advise practitioners to take this into consideration when evaluating counterfactual explanations.

## 5.3 RQ3: Plausibility metric agreement

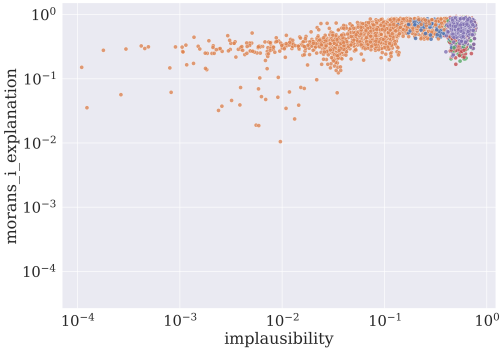
The final research question addresses whether the various plausibility measures agree with one another. Although all measures aim to assess plausibility, they are computed through fundamentally different approaches. As illustrated in Figure 4, there is no consistent alignment between the metrics. In subfigure (a), a distinct cluster is observed for alibi-Proto-CF, differentiating it from the other methods, but no clear patterns emerge beyond this. A similar observation holds for subfigure (b). The results in Table 3 suggest some potential agreement between implausibility and $MI(x')$, but no clear linear relationships are observed. In both subfigures (c) and (d), a distinct cluster associated with alibi-Proto-CF is visible, suggesting its explanations tend to be more distinctly implausible than those of other methods.



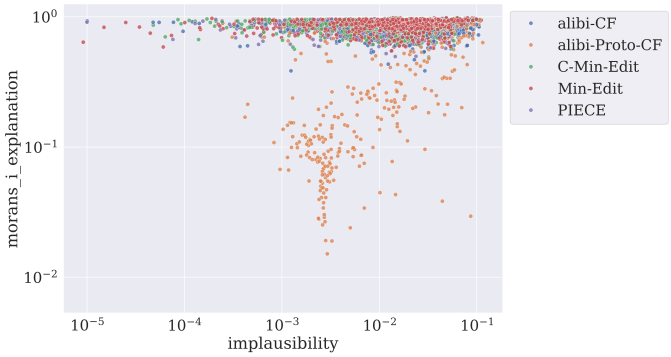

(a) IM1 against implausibility on MNIST with $r = 0.014$, $p = 0.037$, and $n = 22877$




(b) IM1 against implausibility on CIFAR with $r = -0.066$, $p < 0.001$, and $n = 12353$.




(c) $MI(x')$ against implausibility on MNIST with $r = 0.262$, $p < 0.001$, and $n = 22877$.




(d) $MI(x')$ against implausibility on CIFAR with $r = 0.17$, $p < 0.001$, and $n = 12353$.

Figure 4: Scatter plots showing agreement between plausibility measures for different explainability methods, colored by method. Spearman correlation coefficient values are shown in the subfigure captions, along with the corresponding p-values and sample sizes.

Overall, we find little agreement between plausibility measures, suggesting they capture different aspects of plausibility. Relying on a single plausibility metric is therefore insufficient, further supporting the notion that human validation may be essential for reliably assessing the plausibility of counterfactual explanations [Keane *et al.*, 2021].

## 6 Conclusion

In this study, we have provided an extensive analysis of counterfactual explanation methods in image classification, focusing on their performance, the influence of decision boundaries, and the agreement between different plausibility measures.

**Main findings.** Our results reveal several key insights. Methods like C-Min-Edit and PIECE demonstrate strong results in generating valid counterfactual explanations. However, no clear alignment between plausibility measures emerges, suggesting that different metrics capture different aspects of plausibility. Whilst our proposed spatial autocorrelation measure, Moran's I, shows similar average results to the implausibility measure, this agreement appears limited to the prototype-based method alibi-Proto-CF and does not generalise across other methods. The chosen target class plays a critical role in determining success rates, with some methods being more adept at navigating certain decision boundaries than others. Additional experiments further show that performance can also vary substantially across neural architectures, as detailed in the appendix.

**Implications.** Our work highlights that properly assessing counterfactual explanation methods requires evaluation across a variety of decision boundaries and neural architectures. By empirically evaluating counterfactual explanations across diverse experimental settings, we address broader calls for rigorous evaluation in XAI and provide a first step towards a benchmark for practitioners.

**Limitations.** Whilst our study covers 90 experimental settings, a more comprehensive analysis could include a larger variety of explanation methods, a deeper hyperparameter analysis, and additional datasets such as MedMNIST [Yang *et al.*, 2023] to strengthen generalisability.

**Future work.** As to future work, we want to generalise our findings in multiple directions. First, we plan to extend this work with a user study to evaluate explanation effectiveness and measure human agreement with plausibility metrics. Second, we plan to evaluate counterfactual explanation methods in other settings, such as in forecasting [Lucic *et al.*, 2019] and recommender systems [Ren *et al.*, 2017].

## Acknowledgments

We thank Maria Heuss for her early feedback on this project. This research was (partially) supported by the Dutch Research Council (NWO), under project numbers 024.004.022, NWA.1389.20.183, and KICH3.LTP.20.006, and the European Union under grant agreement No. 101201510 (UNITE). Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of their respective employers, funders and/or granting authorities.

## References

Patrick Altmeyer, Mojtaba Farmanbar, Arie van Deursen, and Cynthia C.S. Liem. Faithful model explanations through energy-constrained conformal counterfactuals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10829–10837, 2024.

Vijay Arya, Rachel K.E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques. *arXiv preprint arXiv:1909.03012*, 2019.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46, 07 2015.

Gregor Baer, Isel Grau, Chao Zhang, and Pieter Van Gorp. Why do class-dependent evaluation effects occur with time series feature attributions? A synthetic data investigation. In *arXiv preprint arXiv:2506.11790*, 2025.

Aparna Balagopalan, Haoran Zhang, Kimia Hamidieh, Thomas Hartvigsen, Frank Rudzicz, and Marzyeh Ghassemi. The road to explainability is paved with bias: Measuring the fairness of explanations. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022.

Umang Bhatt, Adrian Weller, and José M. F. Moura. Evaluating and aggregating feature-based model explanations. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 2020.

Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery*, 37(5):1719–1778, 2023.

Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. *Advances in Neural Information Processing Systems*, 31, 2018.

Hidde Fokkema, Rianne de Heide, and Tim van Erven. Attribution-based explanations that provide recourse cannot be robust. *Journal of Machine Learning Research*, 24(360):1–37, 2023.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 2376–2384, 2019.

Riccardo Guidotti, Anna Monreale, Stan Matwin, and Dino Pedreschi. Black box explanation by learning image exemplars in the latent feature space. In *ECML/PKDD*, 2019.

Riccardo Guidotti. Counterfactual explanations and how to find them: Literature review and benchmarking. *Data Mining and Knowledge Discovery*, pages 1–55, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M.-C. Höhne. Quantus: An explainable AI toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023.

Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. Formalising the robustness of counterfactual explanations for neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14901–14909, 2023.

Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*, 2019.

Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55(5):1–29, 2022.

Mark T. Keane, Eoin M. Kenny, Eoin Delaney, and Barry Smyth. If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4466–4474. International Joint Conferences on Artificial Intelligence Organization, 8 2021.

Eoin M. Kenny and Mark T. Keane. On generating plausible counterfactual and semi-factual explanations for deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11575–11585, 2021.

Janis Klaise, Arnaud Van Looveren, Giovanni Vacanti, and Alexandru Coca. Alibi explain: Algorithms for explaining machine learning models. *Journal of Machine Learning Research*, 22(181):1–7, 2021.

Lukas Klein, Carsten Lüth, Udo Schlegel, Till Bungert, Mennatallah El-Assady, and Paul Jäger. Navigating the maze of explainable AI: A systematic approach to evaluating methods and metrics. In *38th Annual Conference on Neural Information Processing Systems*, 2024.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for PyTorch. *arXiv preprint arXiv:2009.07896*, 2020.

Alex Krizhevsky. Learning multiple layers of features from tiny images. https://www.cs.toronto.edu/kriz/learning-features-2009-TR.pdf, 2009.

Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. The dangers of post-hoc interpretability: Unjustified counterfactual explanations. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.

Phuong Quynh Le, Meike Nauta, Van Bach Nguyen, Shreyasi Pathak, Jörg Schlötterer, and Christin Seifert. Benchmarking eXplainable AI - A survey on available toolkits and open challenges. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6665–6673, 2023.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yann LeCun. The MNIST database of handwritten digits. http://yann.lecun.com/exdb/mnist/, 1998.

Shusen Liu, Bhavya Kailkhura, Donald Loveland, and Yong Han. Generative counterfactual introspection for explainable deep learning. In *2019 IEEE Global Conference on Signal and Information Processing*, pages 1–5. IEEE, 2019.

Yang Liu, Sujay Khandagale, Colin White, and Willie Neiswanger. Synthetic benchmarks for scientific research in explainable machine learning. In *Advances in Neural Information Processing Systems Datasets Track*, 2021.

Ana Lucic, Maarten de Rijke, and Hinda Haned. Contrastive explanations for large errors in retail forecasting predictions through monte carlo simulations. In *XAI: IJCAI 2019 Workshop on Explainable Artificial Intelligence*, August 2019.

Divyat Mahajan, Chenhao Tan, and Amit Sharma. Preserving causal constraints in counterfactual explanations for machine learning classifiers. *arXiv preprint arXiv:1912.03277*, 2019.

Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018.

Patrick A.P. Moran. The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):243–251, 1948.

Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Computing Surveys*, 55(13s):1–42, 2023.

Jose Oramas, Kaili Wang, and Tinne Tuytelaars. Visual explanation by interpretation: Improving visual feedback capabilities of deep neural networks. In *International Conference on Learning Representations*, 2019.

Martin Pawelczyk, Sascha Bielawski, Johannes van den Heuvel, Tobias Richter, and Gjergji Kasneci. CARLA: A Python library to benchmark algorithmic recourse and counterfactual explanation algorithms. In *NeurIPS Datasets and Benchmarks*, 2021.

Zhaochun Ren, Shangsong Liang, Piji Li, Shuaiqiang Wang, and Maarten de Rijke. Social collaborative viewpoint regression with explainable recommendations. In *WSDM 2017: The 10th International Conference on Web Search and Data Mining*, pages 485–494, February 2017.

Pouya Samangouei, Ardavan Saeedi, Liam Nakagawa, and Nathan Silberman. ExplainGAN: Model explanation via decision boundary crossing transformations. In *Proceedings of the European Conference on Computer Vision*, pages 666–681, 2018.

Robert-Florian Samoilescu, Arnaud Van Looveren, and Janis Klaise. Model-agnostic and scalable counterfactual explanations via reinforcement learning. In *The AAAI 2023 Workshop on Representation Learning for Responsible Human-Centric AI (RHCAI)*, 2023.

Lisa Schut, Oscar Key, Rory Mc Grath, Luca Costabello, Bogdan Sacaleanu, Medb Corcoran, and Yarin Gal. Generating interpretable counterfactual explanations by implicit minimisation of epistemic and aleatoric uncertainties. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, pages 1756–1764, 2021.

Ilia Stepin, Jose M. Alonso, Alejandro Catala, and Martín Pereira-Fariña. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9:11974–12001, 2021.

Arnaud Van Looveren and Janis Klaise. Interpretable counterfactual explanations guided by prototypes. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 650–665. Springer, 2021.

Diego Velazquez, Pau Rodriguez, Alexandre Lacoste, Issam H Laradji, Xavier Roca, and Jordi Gonzàlez. Explaining visual counterfactual explainers. *Transactions on Machine Learning Research*, 2023.

Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan Hines, John Dickerson, and Chirag Shah. Counterfactual explanations and algorithmic recourses for machine learning: A review. *ACM Computing Surveys*, 2020.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31:841, 2017.

Zhou Wang, Eero P. Simoncelli, and Alan Conrad Bovik. Multiscale structural similarity for image quality assess-

ment. *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, 2:1398–1402 Vol.2, 2003.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. MedMNIST v2–A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 10(1):41, 2023.