A. Reproducibility

To access the full version of the appendix, navigate to: https://ignacystepka.com/projects/betarce.html

A.1. Code Availability

To ensure reproducibility and enable further experimentation with BETARCE, we make the source code publicly available on GitHub: https://github.com/istepka/betarce.

A.2. Datasets

Wine, Breast Cancer, Car eval, and Rice datasets were obtained from https://archive.ics.uci.edu/. Diabetes dataset was sourced from https://www.kaggle.com/datasets/mathchi/diabetes-data-set. HELOC (fico) dataset from https://community.fico.com/s/explainable-machine-learning-challenge?tabset-158d9=d157e.

B. Proof of Theorem 1

Assume that $\hat{\delta}$ follows a priori Beta distribution with parameters a, b, therefore the a priori probability of $P(\hat{\delta})$ has the following probability density function:

$$\begin{aligned} f(\hat{\delta}; a, b) &= \frac{1}{\mathrm{B}(a,b)} \hat{\delta}^{a-1}(1-\hat{\delta})^{b-1} \\ &\propto \hat{\delta}^{a-1}(1-\hat{\delta})^{b-1} \end{aligned} \tag{12}$$

where the beta function B is a normalization constant.

Algorithm 1 samples k random variables $X_i = \mathbb{1}_{M'(x^{cf})==y^{cf}}$ from the space of admissible model changes. Applying the Bayes theorem, we obtain the following a posteriori distribution:

$$\begin{aligned} f(\hat{\delta}|\mathbf{X}, a, b) &\propto f(\mathbf{X}|\hat{\delta}; a, b) f(\hat{\delta}; a, b) \\ &\propto \left(\prod_{i=1}^{k} \hat{\delta}^{x_i}(1-\hat{\delta})^{1-x_i} \right) \hat{\delta}^{a-1}(1-\hat{\delta})^{b-1} \\ &\propto \hat{\delta}^{\sum x_i + a - 1}(1-\hat{\delta})^{k - \sum x_i + b - 1} \end{aligned} \tag{13}$$

Using $z = \sum_{i=1}^{k} x_i$ to denote the number of times the counterfactual was robust for the sampled model, we obtain

$$\begin{aligned} f(\hat{\delta}|\mathbf{X}; a, b) &\propto \hat{\delta}^{z+a-1}(1-\hat{\delta})^{k-z+b-1} \\ &= f(\hat{\delta}; a+z, b+(k-z)) \end{aligned} \tag{14}$$

which is exactly the Beta distribution. Note that Algorithm 1 adds 1 to a every time the counterfactual is robust to the sampled model, so effectively adds z to a during the entire execution. Similarly, $k - z$ is added to b. Therefore, Algorithm 1 estimates the posterior distribution of $P(\hat{\delta})$.

According to Definition 4, (δ, α)-robust counterfactual satisfies the following condition:

$$P(\hat{\delta} > \delta) > \alpha$$

Applying Eq. 14, we obtain:

$$\begin{aligned} P(\hat{\delta} > \delta) &= \int_{\delta}^{1} f(\hat{\delta}; a+z, b+(k-z))\, d\hat{\delta} \\ &= 1 - F_{Beta}(\delta) \end{aligned} \tag{15}$$

where F_{Beta} is the cumulative distribution function of Beta distribution.

$$\begin{aligned} P(\hat{\delta} > \delta) > \alpha &\Rightarrow 1 - F_{Beta}(\delta) > \alpha \\ &\Rightarrow F_{Beta}(\delta) < 1 - \alpha \\ &\Rightarrow F^{-1}_{Beta}(1-\alpha) > \delta \end{aligned} \tag{16}$$

which is consistent with line 10 of Algorithm 1.

C. Proof of Theorem 2

In this section, we provide a proof of the Theorem 2.

C.1. Background

The cumulative distribution function (CDF) of a probability distribution is a function describing the following relationship:

$$F(x) = P(X \leq x) = u \tag{17}$$

where X is a random variable, x is a real number, and u is a probability between 0 and 1. The inverse cumulative distribution function (inverse CDF), also known as the quantile function, is used to find the value x for a given probability u:

$$F^{-1}(u) = x \tag{18}$$

This function returns the value x such that the probability of the random variable X being less than or equal to x is u.

C.2. Proof

Let:

- $n + m = k$ and $n > m$, where $n, m, k \in \mathbb{Z}_+$
- $a = b$ where $a, b \in \mathbb{R}_+$ be a priori parameters of the Beta distribution: $Beta(a, b)$.

We begin by stating Lemma 3 which asserts that for any α greater than 0.5, the CDF of a Beta distribution will always be greater if its first parameter is greater than the second one.

Lemma 3.

$$\forall_{x \in (0.5, 1]} F_{Beta(a+n, b+m)}(x) > F_{Beta(a+m, b+n)}(x) \tag{19}$$

In order to prove Lemma 3, we first simplify the underlying equations in a following way:

$$\begin{aligned} & F_{Beta(a+n,b+m)}(x) > F_{Beta(a+m,b+n)}(x) \\ = \frac{B(x;a+n,b+m)}{B(a+n,b+m)} > \frac{B(x;a+m,b+n)}{B(a+m,b+n)} & \mid \text{Beta func. symmetry} \\ = B(x;a+n,b+m) > B(x;a+m,b+n) & \\ = \int_0^x t^{a+n-1}(1-t)^{b+m-1}dt > \int_0^x t^{a+m-1}(1-t)^{b+n-1}dt & \end{aligned} \tag{20}$$

WLOG, for simplicity of notation, we can assume $a = b = 1$. Therefore, the equation above simplifies to:

$$\int_0^x t^n(1-t)^m dt > \int_0^x t^m(1-t)^n dt \tag{21}$$

From that, it is sufficient to show that the left-hand function is strictly greater than the right-hand function in the integrated domain.

Lemma 4.

$$\forall_{t\in(0.5,1]} \forall_{n,m\in\mathbb{Z}_+, n>m} \; t^n(1-t)^m > t^m(1-t)^n \tag{22}$$

Proof: Lemma 4 can be proven via simple arithmetic manipulations:

$$\begin{aligned} t^n(1-t)^m &> t^m(1-t)^n \\ = t^{n-m}(1-t)^m &> (1-t)^n \\ = t^{n-m} &> (1-t)^{n-m} \\ \implies t &> (1-t) \\ = t &> 0.5 \end{aligned} \tag{23}$$

QED

This completes the proof of Lemma 4

The result of Lemma 4, $t > 0.5$, finishes the proof for Lemma 3, which in turn proves Theorem 2, because the highest attainable value of CDF is at $1-\alpha$, where $\alpha > 0.5$

D. Examining BETARCE parameters in detail

In the main paper, we briefly outlined the relationship between parameters in BETARCE . Remember, BETARCE relies on three internal parameters that impact its performance: δ, representing the lower bound for the probability of robustness; α, indicating the method's confidence level; and k, denoting the number of estimators. Their interplay is defined by the following equation, also featured in the paper:

$$\delta_{max} = F^{-1}_{Beta(a+k,b)}(1-\alpha) \tag{24}$$

This equation offers an intuitive approach to determining the parameters based on practical application requirements. The maximum achievable δ (and consequently (δ,α)-robustness) is constrained by the number of estimators k and the selected confidence level α.

The interpretation of this equation is straightforward: $F^{-1}(1-\alpha)$ identifies the lower bound of robustness at $1-\alpha$. The inverse Cumulative Distribution Function (F^{-1}) is derived from the estimated Beta distribution $Beta_{(a+k,b)}$, with a and b representing default priors of the distribution. Here, k is added to the a parameter of the distribution, as it contributes to the right-skewness of the distribution.

To provide a clearer understanding, below we present a visual representation of how parameters in the Beta distribution influence its shape:

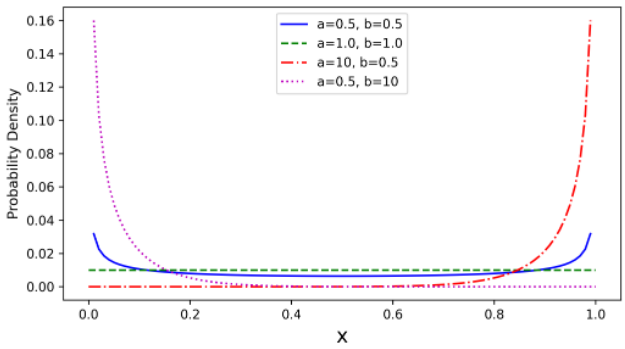


Figure 5. Some priors of Beta distribution

Increasing the a parameter skews the distribution to the right, while altering the b parameter skews it to the left. Therefore, adding k to a identifies the most optimistic (positively skewed) distribution obtainable with the given parameters. Consequently, this facilitates the calculation of the most optimistic lower bound that can be attained: δ_{max}. The proof for this statement is in Sec. C.

Fig. 5 visualizes the shape of noninformative Jeffreys prior used in the paper: (0.5, 0.5). This prior is a U-shaped distribution, with slightly denser tails. Another plausible option was to utilize a prior of (1.0, 1.0), resulting in a uniform distribution.

Below, we provide a plot illustrating the relationship between all these parameters:

Furthermore, we include an auxiliary table (Tab. 2 containing precomputed δ_{max} values (assuming priors equal to 0.5) to facilitate parameter selection in BETARCE for the reader:

E. Experimental setup

In this section, we provide more details on the implementation of experiments.

E.1. General

For all experiments, we utilized a 3-fold cross-validation approach, with 2 folds allocated for training and a single fold for evaluation. During evaluation on each fold, we randomly sampled 30 instances for the generation of robust

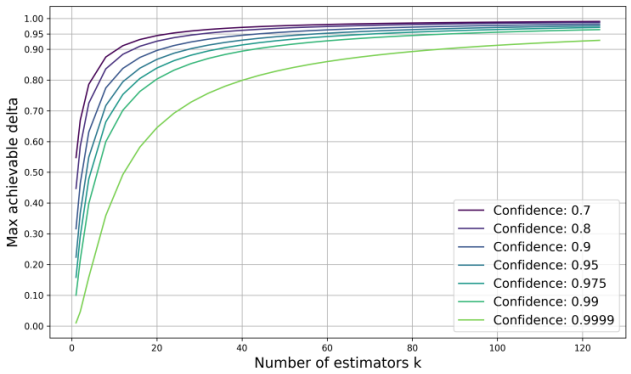


Figure 6. Max achievable δ values as a function of k and α.

Table 2. A table of ready-to-use parameter settings. The columns stand for α values, rows for k, and cells for δ.

k α	0.7	0.8	0.9	0.95	0.975	0.99	0.999
1	0.387	0.316	0.224	0.158	0.112	0.071	0.022
2	0.531	0.464	0.368	0.292	0.232	0.171	0.079
4	0.684	0.631	0.549	0.478	0.416	0.347	0.219
12	0.864	0.838	0.794	0.753	0.714	0.665	0.557
20	0.914	0.896	0.867	0.839	0.812	0.777	0.696
28	0.937	0.924	0.902	0.881	0.860	0.833	0.769
36	0.950	0.940	0.922	0.905	0.888	0.867	0.814
44	0.959	0.950	0.936	0.921	0.907	0.889	0.845
52	0.965	0.957	0.945	0.933	0.921	0.905	0.866
60	0.969	0.963	0.952	0.941	0.931	0.917	0.883
68	0.973	0.967	0.958	0.948	0.938	0.926	0.896
76	0.976	0.971	0.962	0.953	0.945	0.934	0.906
84	0.978	0.973	0.965	0.958	0.950	0.940	0.914
92	0.980	0.976	0.968	0.961	0.954	0.945	0.922
100	0.981	0.977	0.971	0.964	0.958	0.949	0.928
108	0.983	0.979	0.973	0.967	0.961	0.953	0.933
116	0.984	0.981	0.975	0.969	0.963	0.956	0.937
124	0.985	0.982	0.976	0.971	0.966	0.958	0.941

counterfactuals and then assessed the *Empirical Robustness* on 30 new models (from the space of admissible model changes).

For each model, we randomly split the training data into 80-20 *train-validation* sets to facilitate model training and parameter tuning.

E.2. Datasets

Below, we provide information about the datasets used in our study.

Dataset	Rows	Columns	Imbalance Ratio
HELOC	2502	24	1.66
Wine	6497	12	1.73
Diabetes	768	9	1.87
Breast Cancer	569	31	1.68
Car eval	1728	6	2.34
Rice	3810	7	1.34

Preprocessing of these datasets involved dropping rows containing missing values and performing min-max normalization.

E.3. Hyperparameters of the Baselines

Below, we present the hyperparameters that were searched for every end-to-end CFE generation method, both the standard and robust ones:

- **DICE**
  - Diversity Weight: {0.05, 0.1, 0.2}
  - Proximity Weight: {0.05, 0.1, 0.2}
  - Sparsity Weight: {0.05, 0.1, 0.2}
- **FACE**
  - Fraction: {0.1, 0.3, 0.5}
  - Mode: {knn, epsilon}
- **RBR**
  - Max Distance: 1.0
  - Num Samples: 100
  - Delta Plus: {0.0, 0.1, 0.2}
  - Epsilon OP: 0.0
  - Epsilon PE: 0.0
  - Sigma: {0.5, 1.0, 1.5}
  - Perturb Radius (synthesis): {0.1, 0.2, 0.3}
- **ROAR**
  - Delta Max: {0.01, 0.05, 0.1}
  - Learning Rate (LR): {0.01, 0.05, 0.1}
  - Norm: {1, 2}
- **ROBX**
  - N: 1000
  - τ: {0.4, 0.5, 0.6, 0.7, 0.8}
  - Variances: {0.1, 0.01}

For all visualizations, we chose the hyperparameter configuration that provided the highest empirical robustness to ensure a fair comparison. The only exception is the post-hoc method ROBX , as it is also crucial to evaluate the distance to the base counterfactual in such methods. Therefore, to highlight different aspects of ROBX , we included two distinct settings in all comparisons: one optimized for empirical robustness and another that strikes a balance with a good distance to the base CFE.

E.4. Models

In our experiments, we utilize three models as the core black-boxes: a neural network (NN), LightGBM, and logistic regression (LR). These models are implemented using torch, lightgbm, and sklearn, respectively. The validation sets were employed for early stopping in the NN and as the evaluation set for LightGBM. Detailed specifications are provided below:

E.4.1 Neural Network

Parameter	Fixed hparams	Hparams to Vary
Layers	3	3-5
Neurons per layer	128	64-256
Activations	ReLU	
Terminal activation	Sigmoid	
Optimizer	Adam	
Learning rate	1e-3	
Loss	BCE	
Early stopping	5	
Dropout	0.4	
Batch size	128	
Seed	42	

E.4.2 LightGBM

Parameter	Fixed hparams	Hparams to Vary
No. of leaves	15	10-20
No. of estimators	30	15-40
Min. child samples	20	10-20
Subsample	0.8	0.5-1.0 (freq: 0.1)
Objective	binary	
Seed	42	

E.4.3 Logistic Regression

Parameter	Fixed hparams	Hparams to Vary
solver	lbfgs	lbfgs, newton-cg, sag
penalty	l2	l2, none
max_iter	100	50 - 200
C	1	0.1 - 1.0
seed	42	

F. BETARCE intrinsic analysis

In this section, we expand on the analysis presented in the main body of the paper regarding the impact of BETARCE parameters on various aspects of the method's performance.

F.1. Credible intervals for robustness

In this section, we present the full version of Fig. 2 from the main paper with GROWINGSPHERES as a base explainer (Fig. 7), and we also include an additional plot with DICE (Fig. 8) serving as a base explainer.

Figure 7. With GROWINGSPHERES as a base counterfactual explanation.

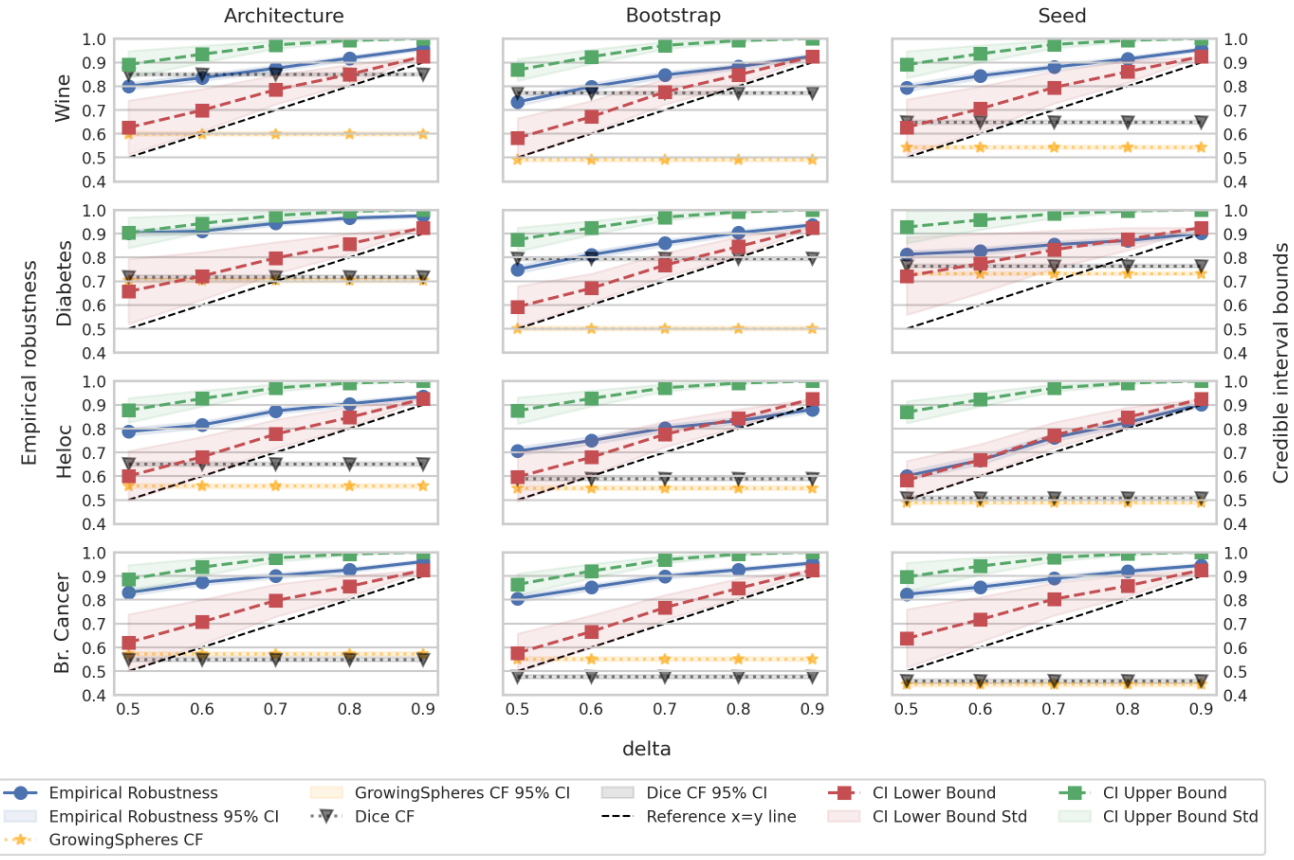
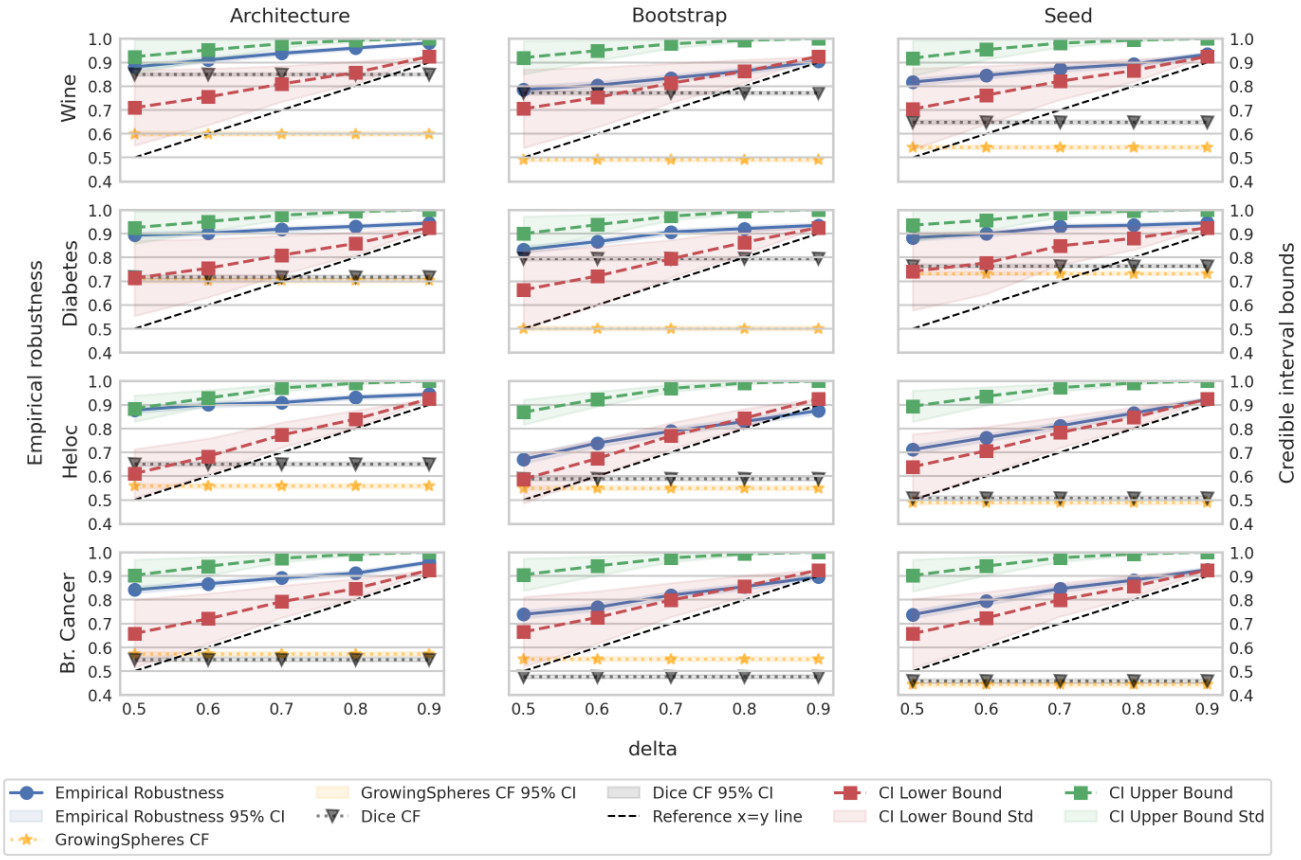


Figure 8. With DICE as a base counterfactual explanation.



F.2. Exploring the impact of the confidence parameter

The parameter α reflects the overall confidence in the estimates provided by our method. Here, we briefly look into how different α values influence the model's performance.

Our first analysis juxtaposes α with *Empirical Robustness* across three δ (Fig. 9).

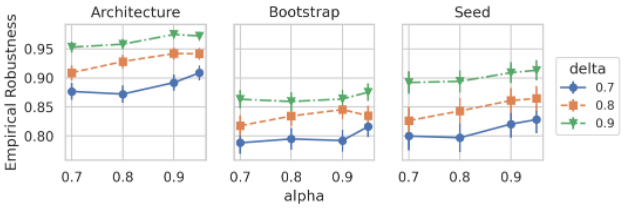


Figure 9. Empirical robustness as a function of α and δ

As observed, the average *Empirical Robustness* shows a slight increase with higher confidence values. This aligns with the notion that greater prediction confidence leads to a more secure robustness estimate, consequently yielding a higher average *Empirical Robustness*.

The subsequent visualization illustrates this enhanced security with higher α values, as indicated by the greater distance between the blue line and the red line, representing the lower bound of the credible interval:

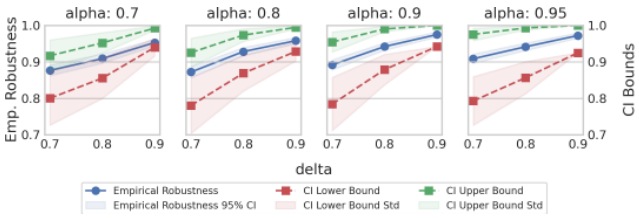


Figure 10. The relationship of empirical robustness and α confidence and its placement credible intervals

F.3. Exploring the impact of the number of estimators

In this section, we examine how varying the number of estimators, denoted as k, affects the performance of BETARCE . As depicted in the Fig. 11 below, increasing k results in narrower credible intervals, indicating a higher level of confidence in the robustness range.

This outcome is anticipated because a higher value of k allows for more combinations of parameters a and b to form the Beta distribution. Consequently, the distribution becomes more flexible, enabling a better fit to the empirical distribution.

The next plot (Fig. 12) illustrates the relationship of the k parameter and the *Empirical Robustness*.

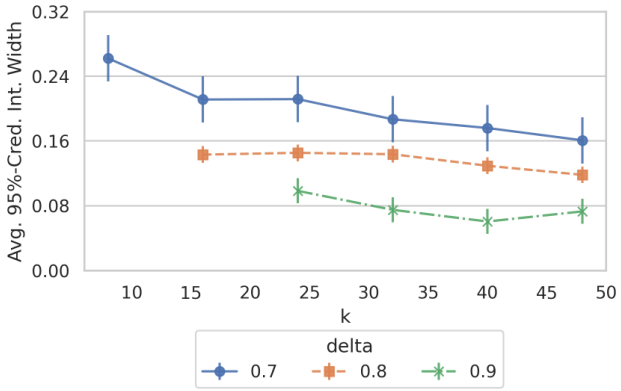


Figure 11. The impact of parameter k on the average credible interval width

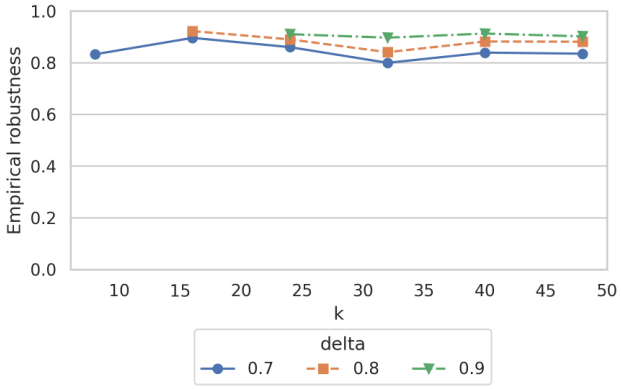


Figure 12. The impact of parameter k on the average empirical robustness

As observed, the average *Empirical Robustness* does not seem to highly depend on the number of estimators.

From these empirical experiments, the conclusion that k increases the *Empirical Robustness* cannot be drawn. Therefore, our recommendation is to use the lowest possible k which allows for realizing desired δ, using introduced for that purpose equation (Eq. 7).

F.4. Investigating the generalization capabilities across different experiment types

In this section, we conduct an experimental analysis to investigate how BETARCE performs when its admissible model space contains different model change types than those encountered during deployment. Specifically, we sample from an admissible model space that does not overlap with the one used for evaluation.

The results of these experiments are presented in two tables (Tables 3 and 4). Each table shows the results for a different base CFE method, averaged across four datasets, with $\delta = 0.9$ and $\alpha = 0.95$.

Table 3. Empirical Robustness of BETARCE with GROWINGSPHERES as the base CFE generation method. The results are averaged over all datasets.

	Generalization		
Original	Architecture	Bootstrap	Seed
Architecture	0.913 ± 0.007	0.865 ± 0.009	0.923 ± 0.007
Bootstrap	0.939 ± 0.006	0.877 ± 0.008	0.909 ± 0.007
Seed	0.927 ± 0.007	0.866 ± 0.009	0.890 ± 0.008

Table 4. Empirical Robustness of BETARCE with DICE as the base CFE generation method. The results are averaged over all datasets.

	Generalization		
Original	Architecture	Bootstrap	Seed
Architecture	0.937 ± 0.005	0.875 ± 0.007	0.930 ± 0.005
Bootstrap	0.927 ± 0.005	0.847 ± 0.007	0.913 ± 0.006
Seed	0.929 ± 0.005	0.805 ± 0.008	0.918 ± 0.005

The diagonal in the table is the normal, in-distribution setting, while all the other cells contain generalizations. As observed, even though the changes are out-of-distribution, BETARCE still robustifies counterfactuals to a satisfiable extent. It is worth to note, that the probabilistic bounds do not hold for out-of-distribution changes, but from the practical perspective it is useful to generalize well for such changes, which BETARCE seems to do well.

The diagonal in the table represents the normal, in-distribution setting, while all other cells contain generalizations. As observed, even though the changes are out-of-distribution, BETARCE still robustifies counterfactuals to a satisfactory extent. It is worth noting that the probabilistic bounds do not hold for out-of-distribution changes, but from a practical perspective, it is useful to generalize well for such changes, which BETARCE seems to be able to accomplish.

G. Computational Complexity

The computational complexity of BETARCE is determined primarily by two factors: (1) the complexity of the chosen optimization algorithm and (2) the number of estimators (k) used for bootstrap robustness verification. The latter directly affects the complexity of evaluating the objective function's constraints, as each evaluation requires querying k estimators. Consequently, k inference calls are introduced as a constant multiplier to the overall complexity.

The optimization algorithm employed to solve Eq. 6 plays the most significant role in determining BETARCE's complexity. In this paper, we utilize GROWINGSPHERES for optimization. While this algorithm does not offer theoretical guarantees regarding the complexity of finding an optimal or ϵ-optimal solution, its complexity is heavily influenced by η and n hyperparameters.

First, η controls the granularity of the iterative expansion of the sphere's perimeter. Second, n specifies the number of samples evaluated on the perimeter during each iteration. When combined with BETARCE, each step of GROWINGSPHERES requires $n \cdot k + 1$ model evaluations. Here, k estimators perform inference on each of the n samples to compute the robustness term (Eq. 5), while the additional $+1$ accounts for the evaluation of the validity term (Eq. 4).

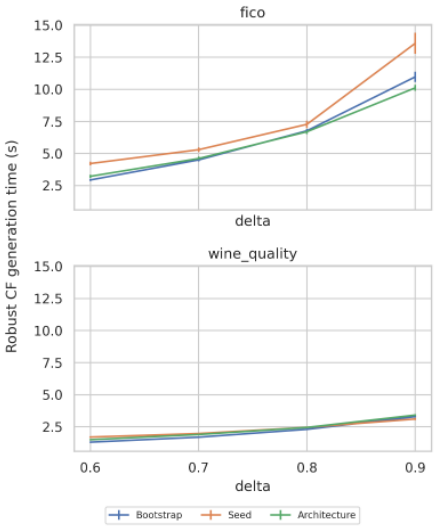


Figure 13. Robust counterfactual generation time as a function of δ.

H. Step-by-step visualization of BETARCE algorithm

In this section, we provide an intuitive visualization of how BETARCE works when integrated with GROWINGSPHERES . Fig. 14 illustrates eight consecutive steps of the BETARCE algorithm for a simple example.

First, a base counterfactual explanation is generated for a given input (see Fig. 14a). Next, we move to the warmup stage of the GROWINGSPHERES search algorithm, as described in Alg. 2 (lines 2-6). In particular, first, in Fig. 14b five candidates are sampled from a large sphere (line 2), second, in Fig. 14c robustness and validity are evaluated (line 3). Since there were both valid and robust examples in the sphere, its radius is halved (line 4). Next, Fig. 14d and Fig. 14e show similar steps for a sphere with smaller radius (line 5), but this time, no valid and robust examples were found. Thus, the warmup stage of GROWINGSPHERES is over.

In the next figures, we show an iteration of the search procedure (lines 7-13), where candidates are sampled from a region between increasing lower and upper radius. Fig. 14f illustrates the sampling of candidates (line 8) and Fig. 14g shows the candidates being evaluated (line 9). Since two robust and valid examples were found, the algorithm is terminated (lines 9-13) and the closest counterfactual (Fig. 14h) is returned as the robust counterfactual explanation (line 14).

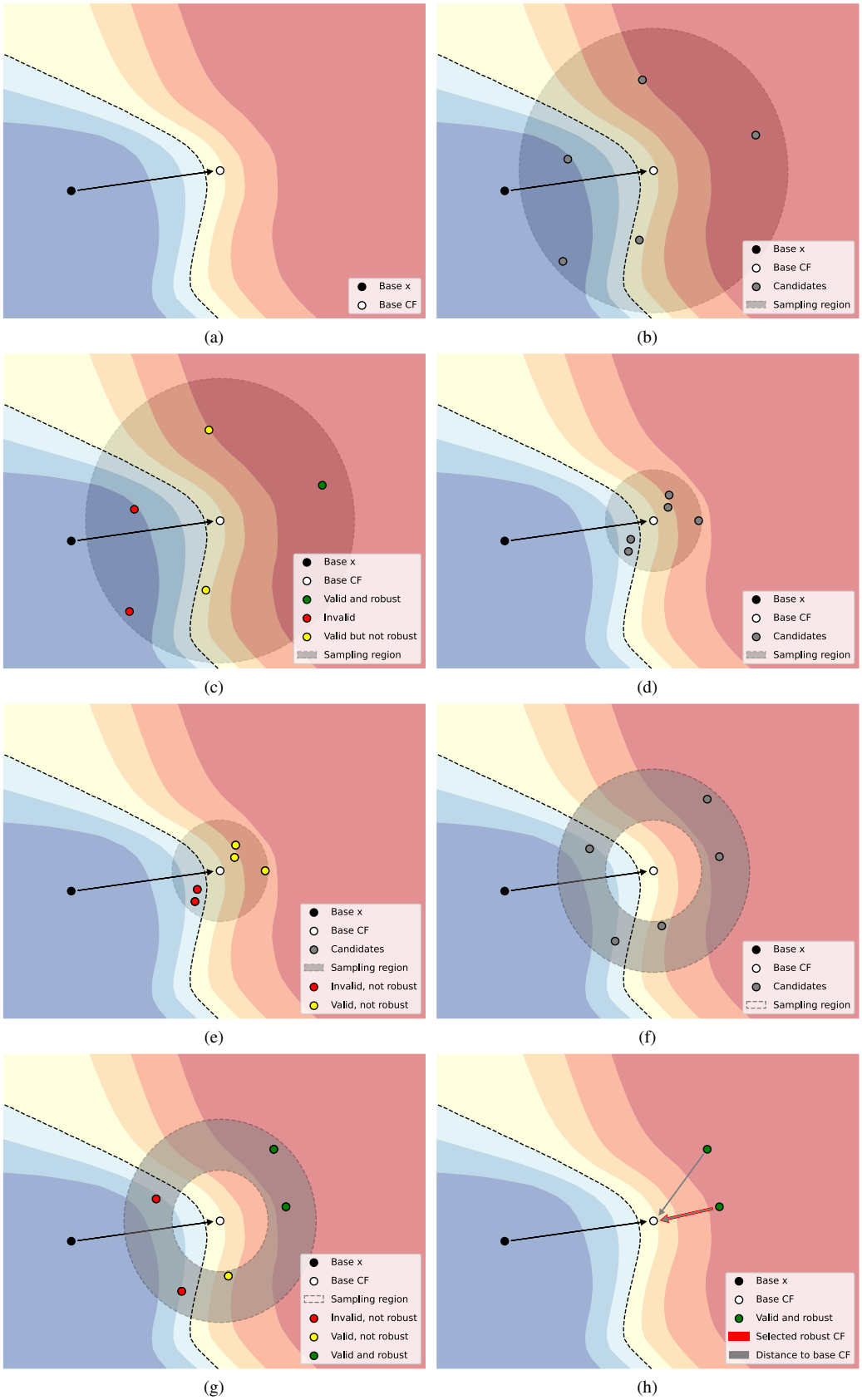


Figure 14. A visualization of BETARCE with GROWINGSPHERES search algorithm.

I. The relationship between *Distance to Base* and other metrics

In this section, we briefly share the empirical observation that the *Distance to Base* metric is negatively correlated with the degradation of CFE performance metrics, focusing on different properties of CFE. Hence, we believe that post-hoc methods which introduce smaller changes to the original CFE (in terms of *Distance to Base*) are generally better at preserving its properties. Fig. 15 illustrates how proximity metrics (both L1 and L2) deteriorate on average as *Distance to Base* increases. We also note that this deterioration is significantly less prominent with the *Plausibility* metric.

To support these findings, we calculated correlation coefficients in Tab. 5.

Table 5. Correlation between *distance to base* and other metrics. All p-values are less than 0.001.

Metric	Correlation type	rho
Proximity L1	Pearson	-0.31
Proximity L1	Spearman	-0.70
Proximity L2	Pearson	-0.31
Proximity L2	Spearman	-0.73
Plausibility	Pearson	-0.31
Plausibility	Spearman	-0.02
Average Pearson		-0.31
Average Spearman		-0.48

J. Background on Bernoulli and Beta Distributions

Before getting into the specifics of BETARCE parameters, it's crucial to understand the foundational distributions underlying our method: the Bernoulli distribution and the Beta distribution.

J.1. Bernoulli Distribution

The Bernoulli distribution is a discrete probability distribution for a random variable that takes only two values, typically 0 and 1. It's often used to model binary outcomes, such as success/failure or yes/no scenarios. The probability density function (PDF) of a Bernoulli distribution is given by:

$$P(X = x) = p^x(1-p)^{1-x}, \quad x \in \{0, 1\} \tag{25}$$

where p is the probability of success (i.e., $X = 1$).

J.2. Beta Distribution

The Beta distribution is a continuous probability distribution defined on the interval [0, 1]. It's characterized by two shape parameters, a and b, which control its shape. The PDF of a Beta distribution is:

$$f(x; a, b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a, b)}, \quad 0 \leq x \leq 1 \tag{26}$$

where $B(\alpha, \beta)$ is the Beta function.

J.3. Conjugate Relationship and Conjugate Priors

In Bayesian statistics, a conjugate prior is a prior distribution that, when combined with the likelihood function, yields a posterior distribution of the same family as the prior. This property is particularly useful for computational and analytical reasons.

The Beta distribution is the conjugate prior for the Bernoulli distribution. To understand this intuitively:

- Imagine we're trying to estimate the probability p of a coin landing heads.
- Our prior belief about p is represented by a Beta distribution, Beta(a, b).
- We then observe a series of coin flips (Bernoulli trials).
- After observing these trials, our updated belief (the posterior) about p is still a Beta distribution, just with updated parameters.

Mathematically, this relationship is expressed as:

$$\text{Prior: } p \sim \text{Beta}(\alpha, \beta) \tag{27}$$

$$\text{Likelihood: } X|p \sim \text{Bernoulli}(p) \tag{28}$$

$$\text{Posterior: } p|X \sim \text{Beta}(\alpha + \sum x_i, \beta + n - \sum x_i) \tag{29}$$

where n is the number of observations and $\sum x_i$ is the number of successes (heads).

K. Comparative Analysis

In this section, we present the comprehensive results from the comparative study detailed in Sec. 4.4. Parameters used in a given method are listed next to this method's name; for ROBX these are τ and variance, while for BETARCE – δ ($\alpha = 0.9$). The values in each cell represent the mean $\pm$ standard error. The column **Type** sorts the methods by categories. The abbreviations Btsr and Arch used next to BETARCE in the **Type** column stand for Bootstrap and Architecture, respectively. Below present results for three base models: neural network (NN), LightGBM, logistic regression (LR).

- **NN**: Tab. 6 provides the extended version of Tab. 1, including results across all datasets, with GROWINGSPHERES used as the base CFE method. Additionally, in Tab. 7, we present the results for when DICE generates base CFEs.

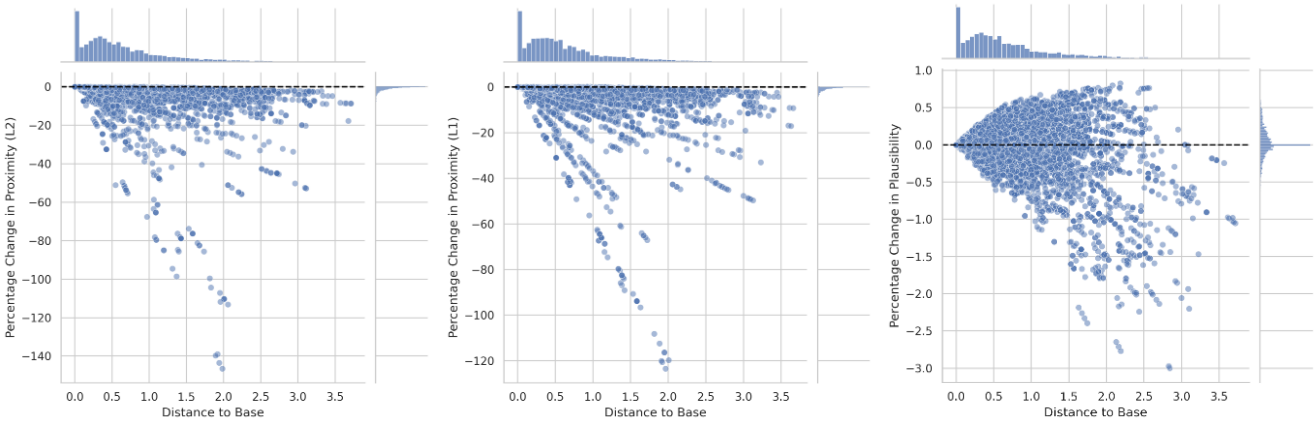


Figure 15. A plot of the correlation between *distance to base* and other metrics.

- **LightGBM**: Tab. 8 contains the results for the scenario where LightGBM is the underlying black-box model, and GROWINGSPHERES is employed as the base CFE generation method.
- **LR**: Tab. 9 similarly provides results for GROWINGSPHERES as a base CFE generation method, but with logistic regressor employed as a black-box model.

Table 6. Comparative study results when the underlying model is a neural network and both ROBX and BETARCE are using GROWINGSPHERES as the base counterfactual explainer

Dataset	Type	Method	Metrics				Empirical Robustness		
			Dist. to Base ↓	Proximity L1 ↓	Proximity L2 ↓	Plausibility ↓	Architecture ↑	Bootstrap ↑	Seed ↑
Diabetes	Standard CFEs	DICE	-	1.002 ± 0.001	0.645 ± 0.001	0.499 ± 0.001	0.916 ± 0.002	0.889 ± 0.003	0.916 ± 0.003
		GROWINGSPHERES	-	0.800 ± 0.001	0.345 ± 0.001	0.358 ± 0.001	0.939 ± 0.003	0.852 ± 0.003	0.853 ± 0.003
		FACE	-	0.880 ± 0.001	0.401 ± 0.001	0.248 ± 0.001	0.869 ± 0.005	0.694 ± 0.006	0.721 ± 0.006
	Robust end-to-end	RBR	-	0.714 ± 0.001	0.339 ± 0.001	0.319 ± 0.001	0.618 ± 0.006	0.617 ± 0.005	0.576 ± 0.005
		ROAR	-	10.887 ± 0.001	4.703 ± 0.001	4.424 ± 0.001	0.415 ± 0.005	0.417 ± 0.005	0.408 ± 0.005
	Robust post-hoc	ROBX (0.5,0.1)	1.224 ± 0.001	1.432 ± 0.001	0.651 ± 0.001	0.324 ± 0.001	0.998 ± 0.001	0.947 ± 0.004	0.969 ± 0.004
		ROBX (0.5,0.01)	0.429 ± 0.001	0.748 ± 0.001	0.339 ± 0.001	0.289 ± 0.001	0.970 ± 0.003	0.872 ± 0.006	0.922 ± 0.005
	BETARCE Arch	BETARCE (0.8)	0.488 ± 0.013	0.870 ± 0.013	0.382 ± 0.005	0.372 ± 0.004	0.966 ± 0.004	-	-
		BETARCE (0.9)	0.607 ± 0.014	0.953 ± 0.013	0.420 ± 0.006	0.378 ± 0.004	0.975 ± 0.004	-	-
	BETARCE Btsr	BETARCE (0.8)	0.445 ± 0.006	0.840 ± 0.008	0.359 ± 0.003	0.359 ± 0.002	-	0.903 ± 0.006	-
		BETARCE (0.9)	0.583 ± 0.007	0.949 ± 0.008	0.407 ± 0.003	0.369 ± 0.002	-	0.937 ± 0.005	-
	BETARCE Seed	BETARCE (0.8)	0.247 ± 0.006	0.813 ± 0.008	0.346 ± 0.003	0.350 ± 0.002	-	-	0.871 ± 0.006
		BETARCE (0.9)	0.315 ± 0.006	0.862 ± 0.008	0.367 ± 0.003	0.353 ± 0.002	-	-	0.902 ± 0.006
HELOC	Standard CFEs	DICE	-	3.190 ± 0.004	1.163 ± 0.001	1.003 ± 0.001	0.912 ± 0.002	0.781 ± 0.004	0.815 ± 0.003
		GROWINGSPHERES	-	2.782 ± 0.003	0.717 ± 0.001	0.773 ± 0.001	0.862 ± 0.003	0.794 ± 0.003	0.752 ± 0.004
		FACE	-	2.254 ± 0.001	0.659 ± 0.001	0.441 ± 0.001	0.829 ± 0.005	0.717 ± 0.006	0.717 ± 0.006
	Robust end-to-end	RBR	-	1.682 ± 0.001	0.505 ± 0.001	0.468 ± 0.001	0.754 ± 0.005	0.690 ± 0.005	0.706 ± 0.005
		ROAR	-	19.803 ± 0.001	5.427 ± 0.001	4.786 ± 0.001	0.591 ± 0.005	0.51 ± 0.005	0.588 ± 0.005
	Robust post-hoc	ROBX (0.5,0.01)	1.145 ± 0.002	2.341 ± 0.001	0.636 ± 0.001	0.598 ± 0.001	0.939 ± 0.005	0.814 ± 0.007	0.890 ± 0.006
		ROBX (0.5,0.1)	3.548 ± 0.005	3.938 ± 0.004	1.144 ± 0.001	0.575 ± 0.001	0.991 ± 0.002	0.957 ± 0.004	0.955 ± 0.005
	BETARCE Arch	BETARCE (0.8)	1.538 ± 0.049	2.912 ± 0.053	0.749 ± 0.014	0.802 ± 0.011	0.904 ± 0.006	-	-
		BETARCE (0.9)	1.697 ± 0.031	2.927 ± 0.036	0.753 ± 0.009	0.783 ± 0.007	0.935 ± 0.005	-	-
	BETARCE Btsr	BETARCE (0.8)	2.288 ± 0.041	3.451 ± 0.044	0.889 ± 0.011	0.859 ± 0.008	-	0.833 ± 0.007	-
		BETARCE (0.9)	3.547 ± 0.071	4.501 ± 0.073	1.156 ± 0.019	1.044 ± 0.015	-	0.880 ± 0.006	-
	BETARCE Seed	BETARCE (0.8)	1.420 ± 0.021	2.526 ± 0.028	0.653 ± 0.007	0.726 ± 0.004	-	-	0.826 ± 0.007
		BETARCE (0.9)	1.927 ± 0.030	2.906 ± 0.035	0.750 ± 0.009	0.776 ± 0.006	-	-	0.902 ± 0.006
Wine	Standard CFEs	DICE	-	0.888 ± 0.001	0.549 ± 0.001	0.413 ± 0.001	0.934 ± 0.002	0.839 ± 0.003	0.873 ± 0.003
		GROWINGSPHERES	-	0.474 ± 0.001	0.174 ± 0.001	0.211 ± 0.001	0.877 ± 0.003	0.837 ± 0.003	0.877 ± 0.003
		FACE	-	0.54 ± 0.001	0.216 ± 0.001	0.131 ± 0.001	0.78 ± 0.003	0.747 ± 0.003	0.783 ± 0.003
	Robust end-to-end	RBR	-	0.508 ± 0.001	0.199 ± 0.001,	0.176 ± 0.001	0.749 ± 0.002	0.73 ± 0.002	0.764 ± 0.002
		ROAR	-	15.768 ± 0.001	5.607 ± 0.001	5.26 ± 0.001	0.734 ± 0.002	0.755 ± 0.002	0.727 ± 0.002
	Robust post-hoc	ROBX (0.5,0.01)	0.564 ± 0.001	0.674 ± 0.001	0.293 ± 0.001	0.159 ± 0.001	1.000 ± 0.001	0.979 ± 0.003	0.997 ± 0.001
		ROBX (0.5,0.1)	1.318 ± 0.002	1.378 ± 0.002	0.593 ± 0.001	0.236 ± 0.001	1.000 ± 0.001	0.970 ± 0.004	1.000 ± 0.001
	Arch	BetaRCE(0.8)	0.330 ± 0.005	0.525 ± 0.006	0.192 ± 0.002	0.219 ± 0.002	0.918 ± 0.005	-	-
		BetaRCE(0.9)	0.471 ± 0.006	0.647 ± 0.007	0.237 ± 0.003	0.242 ± 0.002	0.959 ± 0.004	-	-
	Btsp	BetaRCE(0.8)	0.335 ± 0.005	0.513 ± 0.006	0.188 ± 0.002	0.215 ± 0.002	-	0.882 ± 0.006	-
		BetaRCE(0.9)	0.455 ± 0.006	0.610 ± 0.006	0.224 ± 0.002	0.229 ± 0.002	-	0.926 ± 0.005	-
	Seed	BetaRCE(0.8)	0.308 ± 0.005	0.480 ± 0.006	0.177 ± 0.002	0.210 ± 0.002	-	-	0.914 ± 0.005
		BetaRCE(0.9)	0.399 ± 0.006	0.555 ± 0.006	0.205 ± 0.002	0.217 ± 0.002	-	-	0.954 ± 0.004
Breast Cancer	Standard CFEs	DICE	-	3.004 ± 0.003	1.157 ± 0.001	1.113 ± 0.001	0.894 ± 0.003	0.816 ± 0.003	0.838 ± 0.003
		GROWINGSPHERES	-	3.811 ± 0.001	0.864 ± 0.001	0.949 ± 0.001	0.898 ± 0.003	0.886 ± 0.003	0.886 ± 0.003
		FACE	-	3.427 ± 0.001	0.785 ± 0.001	0.416 ± 0.001	0.93 ± 0.002	0.868 ± 0.003	0.905 ± 0.003
	Robust end-to-end	RBR	-	2.653 ± 0.001	0.617 ± 0.001	0.547 ± 0.001	0.377 ± 0.002	0.343 ± 0.002	0.352 ± 0.002
		ROAR	-	9.271 ± 0.001	2.057 ± 0.001	1.517 ± 0.001	0.386 ± 0.002	0.384 ± 0.002	0.378 ± 0.002
	Robust post-hoc	ROBX (0.5,0.01)	1.251 ± 0.001	3.135 ± 0.001	0.706 ± 0.001	0.728 ± 0.001	0.848 ± 0.007	0.957 ± 0.004	0.846 ± 0.007
		ROBX (0.5,0.1)	3.163 ± 0.001	3.956 ± 0.001	0.876 ± 0.001	0.563 ± 0.001	0.997 ± 0.001	1.000 ± 0.001	0.996 ± 0.001
	Arch	BetaRCE(0.8)	1.684 ± 0.026	4.108 ± 0.04	0.916 ± 0.009	0.969 ± 0.007	0.925 ± 0.005	-	-
		BetaRCE(0.9)	2.058 ± 0.028	4.324 ± 0.042	0.964 ± 0.009	0.996 ± 0.007	0.961 ± 0.004	-	-
	Btsp	BetaRCE(0.8)	1.528 ± 0.024	3.937 ± 0.041	0.901 ± 0.009	0.996 ± 0.008	-	0.926 ± 0.005	-
		BetaRCE(0.9)	1.965 ± 0.026	4.176 ± 0.042	0.954 ± 0.01	1.024 ± 0.008	-	0.955 ± 0.004	-
	Seed	BetaRCE(0.8)	1.749 ± 0.028	3.788 ± 0.039	0.863 ± 0.009	0.940 ± 0.007	-	-	0.920 ± 0.005
		BetaRCE(0.9)	2.115 ± 0.032	4.000 ± 0.042	0.91 ± 0.009	0.967 ± 0.007	-	-	0.945 ± 0.004
Car eval	Standard CFEs	DICE	-	1.189 ± 0.001	0.899 ± 0.001	0.469 ± 0.001	0.866 ± 0.003	0.800 ± 0.004	0.825 ± 0.004
		GROWINGSPHERES	-	0.965 ± 0.001	0.469 ± 0.001	0.487 ± 0.001	0.592 ± 0.006	0.566 ± 0.005	0.608 ± 0.005
		FACE	-	0.977 ± 0.001	0.633 ± 0.001	0.441 ± 0.001	0.761 ± 0.004	0.826 ± 0.004	0.766 ± 0.004
	Robust end-to-end	RBR	-	0.822 ± 0.001	0.505 ± 0.001	0.460 ± 0.001	0.603 ± 0.001	0.649 ± 0.001	0.661 ± 0.001
		ROAR	-	1.637 ± 0.001	0.739 ± 0.001	0.746 ± 0.001	0.277 ± 0.001	0.170 ± 0.001	0.329 ± 0.002
	Robust post-hoc	RobX(0.5,0.01)	0.254 ± 0.002	1.110 ± 0.013	0.573 ± 0.006	0.474 ± 0.001	0.844 ± 0.007	0.842 ± 0.007	0.903 ± 0.006
		RobX(0.5,0.1)	0.895 ± 0.006	1.604 ± 0.015	0.876 ± 0.007	0.456 ± 0.001	0.980 ± 0.003	0.995 ± 0.001	0.994 ± 0.001
	BETARCE Arch	BETARCE (0.8)	0.277 ± 0.014	0.948 ± 0.031	0.456 ± 0.014	0.474 ± 0.002	0.817 ± 0.021	-	-
		BETARCE (0.9)	0.353 ± 0.018	0.966 ± 0.027	0.481 ± 0.013	0.474 ± 0.002	0.850 ± 0.020	-	-
	BETARCE Btsr	BETARCE (0.8)	0.279 ± 0.005	1.148 ± 0.013	0.561 ± 0.006	0.487 ± 0.001	-	0.925 ± 0.005	-
		BETARCE (0.9)	0.324 ± 0.006	1.172 ± 0.013	0.573 ± 0.006	0.485 ± 0.001	-	0.948 ± 0.004	-
	BETARCE Seed	BETARCE (0.8)	0.27 ± 0.005	1.154 ± 0.013	0.561 ± 0.006	0.487 ± 0.001	-	-	0.939 ± 0.005
		BETARCE (0.9)	0.331 ± 0.005	1.193 ± 0.013	0.581 ± 0.006	0.484 ± 0.001	-	-	0.972 ± 0.003
Rice	Standard CFEs	DICE	-	0.905 ± 0.001	0.681 ± 0.001	0.493 ± 0.001	0.791 ± 0.004	0.804 ± 0.004	0.751 ± 0.004
		GROWINGSPHERES	-	0.863 ± 0.001	0.391 ± 0.001	0.250 ± 0.001	0.615 ± 0.005	0.669 ± 0.005	0.530 ± 0.005
		FACE	-	0.805 ± 0.001	0.341 ± 0.001	0.076 ± 0.001	0.619 ± 0.005	0.611 ± 0.005	0.566 ± 0.005
	Robust end-to-end	RBR	-	0.890 ± 0.001	0.396 ± 0.001	0.213 ± 0.001	0.413 ± 0.001	0.421 ± 0.001	0.425 ± 0.001
		ROAR	-	1.974 ± 0.001	0.813 ± 0.001	0.606 ± 0.001	0.231 ± 0.001	0.340 ± 0.002	0.284 ± 0.001
	Robust post-hoc	RobX(0.5,0.01)	0.413 ± 0.004	1.036 ± 0.008	0.445 ± 0.004	0.141 ± 0.002	0.940 ± 0.005	0.979 ± 0.003	0.953 ± 0.004
		RobX(0.5,0.1)	1.021 ± 0.005	1.627 ± 0.009	0.694 ± 0.004	0.101 ± 0.001	1.000 ± 0.001	1.000 ± 0.001	1.000 ± 0.001
	BETARCE Arch	BETARCE (0.8)	0.143 ± 0.003	1.063 ± 0.013	0.487 ± 0.006	0.245 ± 0.001	0.821 ± 0.021	-	-
		BETARCE (0.9)	0.205 ± 0.003	1.070 ± 0.013	0.491 ± 0.006	0.250 ± 0.001	0.850 ± 0.020	-	-
	BETARCE Btsr	BETARCE (0.8)	0.138 ± 0.003	1.167 ± 0.009	0.591 ± 0.004	0.265 ± 0.001	-	0.895 ± 0.004	-
		BETARCE (0.9)	0.187 ± 0.004	1.172 ± 0.009	0.593 ± 0.004	0.266 ± 0.001	-	0.923 ± 0.003	-
	BETARCE Seed	BETARCE (0.8)	0.286 ± 0.005	1.156 ± 0.009	0.587 ± 0.004	0.261 ± 0.001	-	-	0.919 ± 0.004
		BETARCE (0.9)	0.345 ± 0.005	1.162 ± 0.009	0.589 ± 0.004	0.263 ± 0.001	-	-	0.953 ± 0.003

Table 7. Comparative study results when the underlying model is a neural network and both ROBX and BETARCE are using DICE as the base counterfactual explainer.

Dataset	Type	Method	Metrics				Empirical Robustness		
			Dist. to Base ↓	Proximity L1 ↓	Proximity L2 ↓	Plausibility ↓	Architecture ↑	Bootstrap ↑	Seed ↑
Diabetes	Standard CFEs	DICE	-	0.872 ± 0.001	0.685 ± 0.001	0.49 ± 0.001	0.866 ± 0.001	0.7 ± 0.002	0.745 ± 0.002
		GROWINGSPHERES	-	0.596 ± 0.001	0.257 ± 0.001	0.335 ± 0.001	0.726 ± 0.009	0.639 ± 0.011	0.552 ± 0.01
		FACE	-	0.846 ± 0.001	0.39 ± 0.001	0.248 ± 0.001	0.864 ± 0.003	0.692 ± 0.004	0.726 ± 0.004
	Robust end-to-end	RBR	-	0.718 ± 0.001	0.339 ± 0.001	0.318 ± 0.001	0.606 ± 0.002	0.594 ± 0.002	0.569 ± 0.002
		ROAR	-	5.533 ± 0.001	2.58 ± 0.001	2.389 ± 0.001	0.346 ± 0.002	0.36 ± 0.002	0.346 ± 0.002
	Robust post-hoc	ROBX (0.5,0.01)	0.6 ± 0.001	0.796 ± 0.001	0.377 ± 0.001	0.274 ± 0.001	0.982 ± 0.002	0.765 ± 0.006	0.816 ± 0.006
		ROBX (0.6,0.01)	0.814 ± 0.001	0.989 ± 0.001	0.47 ± 0.001	0.29 ± 0.001	0.996 ± 0.001	0.823 ± 0.006	0.873 ± 0.005
	BETARCE Arch	BETARCE (0.8)	0.338 ± 0.004	0.655 ± 0.006	0.286 ± 0.003	0.329 ± 0.002	0.928 ± 0.006	-	-
		BETARCE (0.9)	0.432 ± 0.005	0.73 ± 0.006	0.318 ± 0.003	0.339 ± 0.002	0.953 ± 0.005	-	-
	BETARCE Btsr	BETARCE (0.8)	0.523 ± 0.005	0.792 ± 0.006	0.347 ± 0.003	0.343 ± 0.002	-	0.878 ± 0.006	-
		BETARCE (0.9)	0.705 ± 0.005	0.949 ± 0.007	0.414 ± 0.003	0.368 ± 0.002	-	0.886 ± 0.006	-
	BETARCE Seed	BETARCE (0.8)	0.307 ± 0.004	0.624 ± 0.006	0.276 ± 0.003	0.33 ± 0.002	-	-	0.87 ± 0.008
		BETARCE (0.9)	0.406 ± 0.005	0.703 ± 0.007	0.31 ± 0.003	0.34 ± 0.002	-	-	0.884 ± 0.007
HELOC	Standard CFEs	DICE	-	1.241 ± 0.001	0.9 ± 0.001	0.855 ± 0.001	0.602 ± 0.002	0.56 ± 0.002	0.589 ± 0.002
		GROWINGSPHERES	-	1.946 ± 0.001	0.504 ± 0.001	0.674 ± 0.001	0.543 ± 0.01	0.556 ± 0.01	0.467 ± 0.01
		FACE	-	2.235 ± 0.001	0.653 ± 0.001	0.439 ± 0.001	0.826 ± 0.003	0.712 ± 0.004	0.707 ± 0.004
	Robust end-to-end	RBR	-	1.658 ± 0.001	0.496 ± 0.001	0.466 ± 0.001	0.759 ± 0.002	0.664 ± 0.002	0.633 ± 0.002
		ROAR	-	9.129 ± 0.001	2.515 ± 0.001	2.015 ± 0.001	0.35 ± 0.002	0.369 ± 0.002	0.365 ± 0.002
	Robust post-hoc	ROBX (0.5,0.01)	1.761 ± 0.001	1.943 ± 0.001	0.572 ± 0.001	0.473 ± 0.001	0.919 ± 0.004	0.762 ± 0.006	0.859 ± 0.005
		ROBX (0.6,0.01)	2.52 ± 0.001	2.59 ± 0.001	0.763 ± 0.001	0.461 ± 0.001	0.987 ± 0.002	0.886 ± 0.005	0.966 ± 0.003
	BETARCE Arch	BETARCE (0.8)	1.989 ± 0.054	2.486 ± 0.063	0.638 ± 0.016	0.78 ± 0.011	0.874 ± 0.006	-	-
		BETARCE (0.9)	2.797 ± 0.076	3.193 ± 0.084	0.819 ± 0.021	0.895 ± 0.016	0.929 ± 0.005	-	-
	BETARCE Btsr	BETARCE (0.8)	2.511 ± 0.038	3.046 ± 0.044	0.78 ± 0.011	0.831 ± 0.008	-	0.77 ± 0.008	-
		BETARCE (0.9)	3.793 ± 0.054	4.228 ± 0.059	1.08 ± 0.015	1.028 ± 0.012	-	0.807 ± 0.008	-
	BETARCE Seed	BETARCE (0.8)	1.978 ± 0.034	2.438 ± 0.036	0.629 ± 0.009	0.76 ± 0.006	-	-	0.927 ± 0.005
		BETARCE (0.9)	2.813 ± 0.049	3.192 ± 0.05	0.821 ± 0.013	0.885 ± 0.009	-	-	0.95 ± 0.004
Wine	Standard CFEs	DICE	-	0.674 ± 0.001	0.556 ± 0.001	0.433 ± 0.001	0.781 ± 0.002	0.719 ± 0.002	0.749 ± 0.002
		GROWINGSPHERES	-	0.294 ± 0.001	0.108 ± 0.001	0.187 ± 0.001	0.539 ± 0.01	0.526 ± 0.01	0.525 ± 0.01
		FACE	-	0.528 ± 0.001	0.21 ± 0.001	0.132 ± 0.001	0.78 ± 0.003	0.747 ± 0.003	0.783 ± 0.003
	Robust end-to-end	RBR	-	0.506 ± 0.001	0.198 ± 0.001	0.175 ± 0.001	0.749 ± 0.002	0.73 ± 0.002	0.764 ± 0.002
		ROAR	-	8.395 ± 0.001	3.19 ± 0.001	2.859 ± 0.001	0.734 ± 0.002	0.755 ± 0.002	0.727 ± 0.002
	Robust post-hoc	ROBX (0.5,0.01)	0.546 ± 0.001	0.641 ± 0.001	0.284 ± 0.001	0.16 ± 0.001	0.935 ± 0.004	0.902 ± 0.004	0.924 ± 0.004
		ROBX (0.6,0.01)	0.733 ± 0.001	0.815 ± 0.001	0.374 ± 0.001	0.156 ± 0.001	0.95 ± 0.003	0.931 ± 0.004	0.968 ± 0.003
	Arch	BETARCE (0.8)	0.342 ± 0.005	0.55 ± 0.005	0.206 ± 0.002	0.238 ± 0.002	0.884 ± 0.006	-	-
		BETARCE (0.9)	0.435 ± 0.006	0.622 ± 0.006	0.233 ± 0.002	0.249 ± 0.002	0.909 ± 0.006	-	-
	Btsp	BETARCE (0.8)	0.528 ± 0.005	0.701 ± 0.005	0.265 ± 0.002	0.257 ± 0.002	-	0.829 ± 0.007	-
		BETARCE (0.9)	0.678 ± 0.005	0.831 ± 0.006	0.315 ± 0.002	0.277 ± 0.002	-	0.847 ± 0.007	-
	Seed	BETARCE (0.8)	0.281 ± 0.004	0.476 ± 0.004	0.179 ± 0.002	0.222 ± 0.002	-	-	0.875 ± 0.006
		BETARCE (0.9)	0.418 ± 0.005	0.585 ± 0.005	0.219 ± 0.002	0.238 ± 0.002	-	-	0.906 ± 0.006
Breast Cancer	Standard CFEs	DICE	-	1.623 ± 0.001	1.016 ± 0.001	1.056 ± 0.001	0.559 ± 0.002	0.596 ± 0.002	0.505 ± 0.002
		GROWINGSPHERES	-	3.086 ± 0.003	0.701 ± 0.001	0.853 ± 0.001	0.543 ± 0.01	0.537 ± 0.01	0.472 ± 0.01
		FACE	-	3.427 ± 0.001	0.785 ± 0.001	0.416 ± 0.001	0.93 ± 0.002	0.868 ± 0.003	0.905 ± 0.003
	Robust end-to-end	RBR	-	2.653 ± 0.001	0.617 ± 0.001	0.547 ± 0.001	0.377 ± 0.002	0.343 ± 0.002	0.352 ± 0.002
		ROAR	-	9.271 ± 0.001	2.057 ± 0.001	1.517 ± 0.001	0.386 ± 0.002	0.384 ± 0.002	0.378 ± 0.002
	Robust post-hoc	ROBX (0.5,0.01)	2.849 ± 0.001	3.116 ± 0.001	0.71 ± 0.001	0.471 ± 0.001	0.904 ± 0.004	0.891 ± 0.005	0.873 ± 0.005
		ROBX (0.6,0.01)	3.321 ± 0.001	3.474 ± 0.001	0.792 ± 0.001	0.443 ± 0.001	0.955 ± 0.003	0.952 ± 0.003	0.919 ± 0.004
	Arch	BETARCE (0.8)	1.868 ± 0.05	3.336 ± 0.059	0.752 ± 0.013	0.868 ± 0.011	0.949 ± 0.004	-	-
		BETARCE (0.9)	2.547 ± 0.065	3.822 ± 0.072	0.858 ± 0.016	0.94 ± 0.013	0.961 ± 0.004	-	-
	Btsp	BETARCE (0.8)	4.707 ± 0.096	5.454 ± 0.105	1.213 ± 0.023	1.211 ± 0.02	-	0.845 ± 0.019	-
		BETARCE (0.9)	6.831 ± 0.137	7.412 ± 0.144	1.642 ± 0.032	1.552 ± 0.027	-	0.85 ± 0.019	-
	Seed	BETARCE (0.8)	2.813 ± 0.059	3.66 ± 0.056	0.824 ± 0.013	0.927 ± 0.01	-	-	0.894 ± 0.006
		BETARCE (0.9)	3.555 ± 0.067	4.269 ± 0.065	0.957 ± 0.014	1.023 ± 0.012	-	-	0.923 ± 0.005
Car eval	Standard CFEs	DICE	-	1.189 ± 0.0	0.899 ± 0.0	0.469 ± 0.0	0.866 ± 0.003	0.8 ± 0.003	0.825 ± 0.003
		GROWINGSPHERES	-	0.965 ± 0.001	0.469 ± 0.0	0.487 ± 0.0	0.592 ± 0.005	0.566 ± 0.004	0.608 ± 0.004
		FACE	-	0.977 ± 0.0	0.633 ± 0.0	0.441 ± 0.0	0.761 ± 0.003	0.826 ± 0.003	0.766 ± 0.003
	Robust end-to-end	RBR	-	0.822 ± 0.0	0.505 ± 0.0	0.46 ± 0.0	0.603 ± 0.001	0.649 ± 0.001	0.661 ± 0.001
		ROAR	-	1.637 ± 0.0	0.739 ± 0.0	0.746 ± 0.0	0.277 ± 0.001	0.17 ± 0.001	0.329 ± 0.001
	Robust post-hoc	ROBX (0.5,0.01)	0.062 ± 0.002	1.194 ± 0.007	0.88 ± 0.004	0.468 ± 0.0	0.903 ± 0.004	0.874 ± 0.005	0.936 ± 0.003
		ROBX (0.5,0.1)	0.528 ± 0.007	1.59 ± 0.009	1.011 ± 0.004	0.457 ± 0.0	0.987 ± 0.002	0.991 ± 0.001	0.994 ± 0.001
	BETARCE Arch	BETARCE (0.8)	0.062 ± 0.003	1.197 ± 0.008	0.872 ± 0.005	0.467 ± 0.0	0.966 ± 0.002	-	-
		BETARCE (0.9)	0.083 ± 0.003	1.217 ± 0.008	0.88 ± 0.005	0.468 ± 0.0	0.98 ± 0.002	-	-
	BETARCE Btsr	BETARCE (0.8)	0.133 ± 0.004	1.331 ± 0.008	0.942 ± 0.005	0.472 ± 0.001	-	0.977 ± 0.002	-
		BETARCE (0.9)	0.156 ± 0.004	1.34 ± 0.008	0.94 ± 0.005	0.47 ± 0.001	-	0.988 ± 0.001	-
	BETARCE Seed	BETARCE (0.8)	0.103 ± 0.004	1.34 ± 0.008	0.944 ± 0.005	0.469 ± 0.0	-	-	0.944 ± 0.003
		BETARCE (0.9)	0.143 ± 0.004	1.37 ± 0.008	0.948 ± 0.005	0.469 ± 0.0	-	-	0.964 ± 0.003
Rice	Standard CFEs	DICE	-	0.905 ± 0.0	0.681 ± 0.0	0.493 ± 0.0	0.791 ± 0.003	0.804 ± 0.003	0.751 ± 0.003
		GROWINGSPHERES	-	0.863 ± 0.0	0.391 ± 0.0	0.25 ± 0.0	0.615 ± 0.004	0.669 ± 0.004	0.53 ± 0.004
		FACE	-	0.805 ± 0.0	0.341 ± 0.0	0.076 ± 0.0	0.619 ± 0.004	0.611 ± 0.004	0.566 ± 0.004
	Robust end-to-end	RBR	-	0.89 ± 0.0	0.396 ± 0.0	0.213 ± 0.0	0.413 ± 0.001	0.421 ± 0.001	0.425 ± 0.001
		ROAR	-	1.974 ± 0.0	0.813 ± 0.0	0.606 ± 0.0	0.231 ± 0.001	0.34 ± 0.001	0.284 ± 0.001
	Robust post-hoc	ROBX (0.5,0.01)	0.286 ± 0.006	1.031 ± 0.005	0.647 ± 0.003	0.393 ± 0.003	0.907 ± 0.004	0.959 ± 0.003	0.889 ± 0.005
		ROBX (0.5,0.1)	0.503 ± 0.007	1.423 ± 0.006	0.865 ± 0.003	0.415 ± 0.001	0.979 ± 0.002	0.989 ± 0.002	0.993 ± 0.001
	BETARCE Arch	BETARCE (0.8)	0.256 ± 0.006	1.022 ± 0.005	0.642 ± 0.004	0.392 ± 0.003	0.962 ± 0.002	-	-
		BETARCE (0.9)	0.276 ± 0.006	1.035 ± 0.005	0.648 ± 0.004	0.393 ± 0.003	0.978 ± 0.002	-	-
	BETARCE Btsr	BETARCE (0.8)	0.306 ± 0.006	1.044 ± 0.005	0.65 ± 0.004	0.392 ± 0.003	-	0.984 ± 0.002	-
		BETARCE (0.9)	0.326 ± 0.006	1.051 ± 0.005	0.654 ± 0.004	0.392 ± 0.003	-	0.992 ± 0.001	-
	BETARCE Seed	BETARCE (0.8)	0.276 ± 0.006	1.045 ± 0.005	0.649 ± 0.004	0.392 ± 0.003	-	-	0.899 ± 0.005
		BETARCE (0.9)	0.296 ± 0.006	1.053 ± 0.005	0.653 ± 0.004	0.392 ± 0.003	-	-	0.909 ± 0.005

Table 8. Comparative study results when the underlying model is LightGBM and both ROBX and BETARCE are using GROWINGSPHERES as the base counterfactual explainer. Note, there is no seed experiment presented for LightGBM as different seeds were yielding the same models.

Dataset	Type	Method	Metrics				Empirical Robustness	
			Dist. to Base ↓	Proximity L1 ↓	Proximity L2 ↓	Plausibility ↓	Architecture ↑	Bootstrap ↑
Diabetes	Standard CFEs	DICE	-	0.872 ± 0.001	0.685 ± 0.001	0.49 ± 0.001	0.866 ± 0.001	0.7 ± 0.002
		GROWINGSPHERES	-	0.596 ± 0.001	0.257 ± 0.001	0.335 ± 0.001	0.726 ± 0.009	0.639 ± 0.011
		FACE	-	0.846 ± 0.001	0.39 ± 0.001	0.248 ± 0.001	0.864 ± 0.003	0.692 ± 0.004
	Robust end-to-end	RBR	-	0.718 ± 0.001	0.339 ± 0.001	0.318 ± 0.001	0.606 ± 0.002	0.594 ± 0.002
		ROAR	-	5.533 ± 0.001	2.58 ± 0.001	2.389 ± 0.001	0.346 ± 0.002	0.36 ± 0.002
	Robust post-hoc	ROBX (0.5,0.01)	0.6 ± 0.001	0.796 ± 0.001	0.377 ± 0.001	0.274 ± 0.001	0.934 ± 0.004	0.878 ± 0.005
		ROBX (0.6,0.01)	0.814 ± 0.001	0.989 ± 0.001	0.47 ± 0.001	0.29 ± 0.001	0.998 ± 0.001	0.961 ± 0.003
	BETARCE Arch	BETARCE (0.8)	0.338 ± 0.004	0.655 ± 0.006	0.286 ± 0.003	0.329 ± 0.002	0.848 ± 0.005	-
		BETARCE (0.9)	0.432 ± 0.005	0.73 ± 0.006	0.318 ± 0.003	0.339 ± 0.002	0.888 ± 0.005	-
	BETARCE Btsr	BETARCE (0.8)	0.523 ± 0.005	0.792 ± 0.006	0.347 ± 0.003	0.343 ± 0.002	-	0.902 ± 0.005
		BETARCE (0.9)	0.705 ± 0.005	0.949 ± 0.007	0.414 ± 0.003	0.368 ± 0.002	-	0.955 ± 0.003
HELOC	Standard CFEs	DICE	-	1.241 ± 0.001	0.9 ± 0.001	0.855 ± 0.001	0.602 ± 0.002	0.56 ± 0.002
		GROWINGSPHERES	-	1.946 ± 0.001	0.504 ± 0.001	0.674 ± 0.001	0.543 ± 0.01	0.556 ± 0.01
		FACE	-	2.235 ± 0.001	0.653 ± 0.001	0.439 ± 0.001	0.826 ± 0.003	0.712 ± 0.004
	Robust end-to-end	RBR	-	1.658 ± 0.001	0.496 ± 0.001	0.466 ± 0.001	0.759 ± 0.002	0.664 ± 0.002
		ROAR	-	9.129 ± 0.001	2.515 ± 0.001	2.015 ± 0.001	0.35 ± 0.002	0.369 ± 0.002
	Robust post-hoc	ROBX (0.5,0.01)	1.761 ± 0.001	1.943 ± 0.001	0.572 ± 0.001	0.473 ± 0.001	0.914 ± 0.004	0.82 ± 0.006
		ROBX (0.6,0.01)	2.52 ± 0.001	2.59 ± 0.001	0.763 ± 0.001	0.461 ± 0.001	1.0 ± 0.001	0.962 ± 0.003
	BETARCE Arch	BETARCE (0.8)	1.989 ± 0.054	2.486 ± 0.063	0.638 ± 0.016	0.78 ± 0.011	0.867 ± 0.009	-
		BETARCE (0.9)	2.797 ± 0.076	3.193 ± 0.084	0.819 ± 0.021	0.895 ± 0.016	0.889 ± 0.008	-
	BETARCE Btsr	BETARCE (0.8)	2.511 ± 0.038	3.046 ± 0.044	0.78 ± 0.011	0.831 ± 0.008	-	0.875 ± 0.009
		BETARCE (0.9)	3.793 ± 0.054	4.228 ± 0.059	1.08 ± 0.015	1.028 ± 0.012	-	0.94 ± 0.006
Wine	Standard CFEs	DICE	-	0.674 ± 0.001	0.556 ± 0.001	0.433 ± 0.001	0.781 ± 0.002	0.719 ± 0.002
		GROWINGSPHERES	-	0.294 ± 0.001	0.108 ± 0.001	0.187 ± 0.001	0.539 ± 0.01	0.526 ± 0.01
		FACE	-	0.528 ± 0.001	0.21 ± 0.001	0.132 ± 0.001	0.78 ± 0.003	0.747 ± 0.003
	Robust end-to-end	RBR	-	0.506 ± 0.001	0.198 ± 0.001	0.175 ± 0.001	0.749 ± 0.002	0.73 ± 0.002
		ROAR	-	8.395 ± 0.001	3.19 ± 0.001	2.859 ± 0.001	0.734 ± 0.002	0.755 ± 0.002
	Robust post-hoc	ROBX (0.5,0.1)	0.546 ± 0.001	0.641 ± 0.001	0.284 ± 0.001	0.16 ± 0.001	0.984 ± 0.002	0.948 ± 0.003
		ROBX (0.6,0.1)	0.733 ± 0.001	0.815 ± 0.001	0.374 ± 0.001	0.156 ± 0.001	1.0 ± 0.001	0.994 ± 0.001
	BETARCE Arch	BETARCE (0.8)	0.342 ± 0.005	0.55 ± 0.005	0.206 ± 0.002	0.238 ± 0.002	0.893 ± 0.005	-
		BETARCE (0.9)	0.435 ± 0.006	0.622 ± 0.006	0.233 ± 0.002	0.249 ± 0.002	0.928 ± 0.004	-
	BETARCE Btsr	BETARCE (0.8)	0.528 ± 0.005	0.701 ± 0.005	0.265 ± 0.002	0.257 ± 0.002	-	0.936 ± 0.004
		BETARCE (0.9)	0.678 ± 0.005	0.831 ± 0.006	0.315 ± 0.002	0.277 ± 0.002	-	0.97 ± 0.003
Breast Cancer	Standard CFEs	DICE	-	1.623 ± 0.001	1.016 ± 0.001	1.056 ± 0.001	0.559 ± 0.002	0.596 ± 0.002
		GROWINGSPHERES	-	3.086 ± 0.003	0.701 ± 0.001	0.853 ± 0.001	0.543 ± 0.01	0.537 ± 0.01
		FACE	-	3.427 ± 0.001	0.785 ± 0.001	0.416 ± 0.001	0.93 ± 0.002	0.868 ± 0.003
	Robust end-to-end	RBR	-	2.653 ± 0.001	0.617 ± 0.001	0.547 ± 0.001	0.377 ± 0.002	0.343 ± 0.002
		ROAR	-	9.271 ± 0.001	2.057 ± 0.001	1.517 ± 0.001	0.386 ± 0.002	0.384 ± 0.002
	Robust post-hoc	ROBX (0.5,0.1)	2.849 ± 0.001	3.116 ± 0.001	0.71 ± 0.001	0.471 ± 0.001	0.959 ± 0.003	0.904 ± 0.004
		ROBX (0.6,0.1)	3.321 ± 0.001	3.474 ± 0.001	0.792 ± 0.001	0.443 ± 0.001	0.997 ± 0.001	0.971 ± 0.002
	BETARCE Arch	BETARCE (0.8)	1.868 ± 0.05	3.336 ± 0.059	0.752 ± 0.013	0.868 ± 0.011	0.902 ± 0.008	-
		BETARCE (0.9)	2.547 ± 0.065	3.822 ± 0.072	0.858 ± 0.016	0.94 ± 0.013	0.936 ± 0.007	-
	BETARCE Btsr	BETARCE (0.8)	4.707 ± 0.096	5.454 ± 0.105	1.213 ± 0.023	1.211 ± 0.02	-	0.931 ± 0.009
		BETARCE (0.9)	6.831 ± 0.137	7.412 ± 0.144	1.642 ± 0.032	1.552 ± 0.027	-	0.964 ± 0.006

Table 9. Comparative study results when the underlying model is logistic regression and both ROBX and BETARCE are using DICE as the base counterfactual explainer.

Dataset	Type	Method	Metrics				Empirical Robustness		
			Dist. to Base ↓	Proximity L1 ↓	Proximity L2 ↓	Plausibility ↓	Architecture ↑	Bootstrap ↑	Seed ↑
Breast Cancer	Standard CFEs	GROWINGSPHERES	-	4.53 ± 0.002	1.029 ± 0.000	1.044 ± 0.000	0.595 ± 0.009	0.576 ± 0.010	0.704 ± 0.009
	Robust end-to-end	RBR	-	2.906 ± 0.000	0.68 ± 0.000	0.503 ± 0.000	0.457 ± 0.002	0.520 ± 0.002	0.528 ± 0.002
		ROAR	-	0.295 ± 0.002	0.062 ± 0.000	0.498 ± 0.000	0.079 ± 0.001	0.047 ± 0.001	0.020 ± 0.001
	Robust post-hoc	ROBX (0.5,0.01)	1.141 ± 0.016	3.986 ± 0.041	0.905 ± 0.009	0.832 ± 0.008	0.782 ± 0.008	0.769 ± 0.008	0.944 ± 0.004
		ROBX (0.5,0.1)	3.170 ± 0.022	4.293 ± 0.031	0.963 ± 0.007	0.582 ± 0.005	0.987 ± 0.002	0.984 ± 0.002	0.999 ± 0.001
	BETARCE Arch	BETARCE (0.8,0.9)	1.649 ± 0.025	5.083 ± 0.046	1.152 ± 0.010	1.104 ± 0.008	0.937 ± 0.005	-	-
		BETARCE (0.9,0.9)	2.092 ± 0.027	5.302 ± 0.047	1.202 ± 0.011	1.127 ± 0.008	0.986 ± 0.002	-	-
	BETARCE Btsr	BETARCE (0.8,0.9)	1.529 ± 0.026	5.166 ± 0.047	1.173 ± 0.010	1.134 ± 0.008	-	0.957 ± 0.004	-
		BETARCE (0.9,0.9)	1.751 ± 0.028	5.275 ± 0.047	1.197 ± 0.011	1.142 ± 0.008	-	0.988 ± 0.002	-
	BETARCE Seed	BETARCE (0.8,0.9)	0.557 ± 0.014	5.006 ± 0.053	1.135 ± 0.012	1.096 ± 0.009	-	-	0.956 ± 0.004
		BETARCE (0.9,0.9)	0.748 ± 0.016	5.075 ± 0.053	1.153 ± 0.012	1.103 ± 0.009	-	-	0.991 ± 0.002
Car Evaluation	Standard CFEs	GROWINGSPHERES	-	0.514 ± 0.000	0.506 ± 0.000	1.057 ± 0.001	0.635 ± 0.009	0.626 ± 0.009	0.662 ± 0.009
	Robust end-to-end	RBR	-	0.881 ± 0.000	0.573 ± 0.000	0.475 ± 0.000	0.643 ± 0.002	0.658 ± 0.002	0.600 ± 0.002
		ROAR	-	2.323 ± 0.000	1.018 ± 0.000	0.912 ± 0.000	0.391 ± 0.002	0.424 ± 0.002	0.422 ± 0.002
	Robust post-hoc	ROBX (0.5,0.01)	0.264 ± 0.003	1.129 ± 0.013	0.586 ± 0.006	0.485 ± 0.001	0.920 ± 0.005	0.879 ± 0.006	0.949 ± 0.004
		ROBX (0.5,0.1)	0.774 ± 0.004	1.455 ± 0.012	0.814 ± 0.006	0.461 ± 0.001	1.000 ± 0.000	0.991 ± 0.002	1.000 ± 0.000
	BETARCE Arch	BETARCE (0.8,0.9)	0.184 ± 0.003	1.164 ± 0.014	0.579 ± 0.007	0.513 ± 0.001	0.948 ± 0.004	-	-
		BETARCE (0.9,0.9)	0.210 ± 0.003	1.192 ± 0.014	0.593 ± 0.007	0.513 ± 0.001	0.984 ± 0.002	-	-
	BETARCE Btsr	BETARCE (0.8,0.9)	0.226 ± 0.004	1.313 ± 0.016	0.633 ± 0.007	0.520 ± 0.002	-	0.949 ± 0.004	-
		BETARCE (0.9,0.9)	0.281 ± 0.005	1.355 ± 0.016	0.654 ± 0.007	0.522 ± 0.002	-	0.980 ± 0.003	-
	BETARCE Seed	BETARCE (0.8,0.9)	0.158 ± 0.003	1.124 ± 0.014	0.543 ± 0.007	0.504 ± 0.001	-	-	0.953 ± 0.004
		BETARCE (0.9,0.9)	0.181 ± 0.003	1.147 ± 0.014	0.553 ± 0.007	0.505 ± 0.001	-	-	0.987 ± 0.002
Diabetes	Standard CFEs	GROWINGSPHERES	-	0.545 ± 0.003	0.234 ± 0.001	0.310 ± 0.001	0.543 ± 0.010	0.535 ± 0.035	0.410 ± 0.009
	Robust end-to-end	RBR	-	0.772 ± 0.000	0.367 ± 0.000	0.311 ± 0.000	0.536 ± 0.002	0.751 ± 0.003	0.550 ± 0.002
		ROAR	-	0.766 ± 0.002	0.303 ± 0.001	0.443 ± 0.000	0.306 ± 0.002	0.112 ± 0.002	0.156 ± 0.002
	Robust post-hoc	ROBX (0.5,0.01)	0.405 ± 0.003	0.838 ± 0.007	0.372 ± 0.003	0.310 ± 0.002	0.711 ± 0.009	0.990 ± 0.007	0.718 ± 0.009
		ROBX (0.5,0.1)	1.130 ± 0.006	1.422 ± 0.007	0.664 ± 0.003	0.390 ± 0.002	0.903 ± 0.006	1.000 ± 0.000	0.905 ± 0.006
	BETARCE Arch	BETARCE (0.8,0.9)	0.789 ± 0.010	1.225 ± 0.012	0.517 ± 0.005	0.443 ± 0.003	0.925 ± 0.005	-	-
		BETARCE (0.9,0.9)	1.068 ± 0.012	1.501 ± 0.013	0.626 ± 0.005	0.499 ± 0.003	0.977 ± 0.003	-	-
	BETARCE Btsr	BETARCE (0.8,0.9)	0.192 ± 0.010	0.535 ± 0.011	0.230 ± 0.005	0.283 ± 0.004	-	0.940 ± 0.017	-
		BETARCE (0.9,0.9)	0.200 ± 0.007	0.542 ± 0.013	0.239 ± 0.006	0.285 ± 0.004	-	0.980 ± 0.010	-
	BETARCE Seed	BETARCE (0.8,0.9)	0.746 ± 0.014	1.155 ± 0.016	0.483 ± 0.006	0.449 ± 0.004	-	-	0.957 ± 0.004
		BETARCE (0.9,0.9)	0.786 ± 0.014	1.184 ± 0.015	0.503 ± 0.006	0.460 ± 0.004	-	-	0.987 ± 0.002
Rice	Standard CFEs	GROWINGSPHERES	-	0.992 ± 0.000	0.447 ± 0.000	0.246 ± 0.000	0.684 ± 0.009	0.807 ± 0.008	0.733 ± 0.009
	Robust end-to-end	RBR	-	0.932 ± 0.000	0.420 ± 0.000	0.193 ± 0.000	0.414 ± 0.002	0.409 ± 0.002	0.424 ± 0.002
		ROAR	-	1.562 ± 0.000	0.665 ± 0.000	0.414 ± 0.000	0.230 ± 0.002	0.190 ± 0.002	0.253 ± 0.002
	Robust post-hoc	ROBX (0.5,0.01)	0.285 ± 0.003	1.084 ± 0.009	0.469 ± 0.004	0.153 ± 0.002	0.891 ± 0.006	0.997 ± 0.001	0.984 ± 0.002
		ROBX (0.5,0.1)	0.735 ± 0.003	1.511 ± 0.008	0.644 ± 0.003	0.097 ± 0.001	0.989 ± 0.002	1.000 ± 0.000	1.000 ± 0.000
	BETARCE Arch	BETARCE (0.8,0.9)	0.247 ± 0.005	1.106 ± 0.010	0.490 ± 0.004	0.239 ± 0.002	0.937 ± 0.005	-	-
		BETARCE (0.9,0.9)	0.369 ± 0.006	1.158 ± 0.011	0.513 ± 0.005	0.242 ± 0.002	0.982 ± 0.003	-	-
	BETARCE Btsr	BETARCE (0.8,0.9)	0.056 ± 0.002	1.039 ± 0.010	0.469 ± 0.004	0.250 ± 0.002	-	0.969 ± 0.003	-
		BETARCE (0.9,0.9)	0.078 ± 0.002	1.056 ± 0.009	0.476 ± 0.004	0.251 ± 0.002	-	0.991 ± 0.002	-
	BETARCE Seed	BETARCE (0.8,0.9)	0.079 ± 0.002	1.029 ± 0.010	0.461 ± 0.004	0.246 ± 0.002	-	-	0.963 ± 0.004
		BETARCE (0.9,0.9)	0.110 ± 0.002	1.045 ± 0.010	0.468 ± 0.004	0.245 ± 0.002	-	-	0.988 ± 0.002