SAFARI: Versatile and Efficient Evaluations for Robustness of Interpretability

Wei Huang[1,2]   Xingyu Zhao[2,3]   Gaojie Jin[2,4]   Xiaowei Huang[2]

[1]Purple Mountain Laboratories   [2]University of Liverpool
[3]WMG, University of Warwick   [4]Institute of Software, CAS

{w.huang23,xingyu.zhao,g.jin3,xiaowei.huang}@liverpool.ac.uk

Abstract

Interpretability of Deep Learning (DL) is a barrier to trustworthy AI. Despite great efforts made by the Explainable AI (XAI) community, explanations lack robustness— indistinguishable input perturbations may lead to different XAI results. Thus, it is vital to assess how robust DL interpretability is, given an XAI method. In this paper, we identify several challenges that the state-of-the-art is unable to cope with collectively: i) existing metrics are not comprehensive; ii) XAI techniques are highly heterogeneous; iii) misinterpretations are normally rare events. To tackle these challenges, we introduce two black-box evaluation methods, concerning the worst-case interpretation discrepancy and a probabilistic notion of how robust in general, respectively. Genetic Algorithm (GA) with bespoke fitness function is used to solve constrained optimisation for efficient worstcase evaluation. Subset Simulation (SS), dedicated to estimate rare event probabilities, is used for evaluating overall robustness. Experiments show that the accuracy, sensitivity, and efficiency of our methods outperform the state-ofthe-arts. Finally, we demonstrate two applications of our methods: ranking robust XAI methods and selecting training schemes to improve both classification and interpretation robustness.

1. Introduction

A key impediment to the wide adoption of Deep Learning (DL) is its perceived lack of transparency. Explainable AI (XAI) is a research area that aims at providing the visibility into how a DL model makes decisions, and thus enables the use of DL in vision-based safety critical applications, such as autonomous driving [32], and medical image analysis [43]. Typically, XAI techniques visualise which input features are significant to the DL model's prediction via attribution maps [4, 21]. However, interpretations[1] suffer from the lack of robustness. Many works have shown

that a small perturbation can manipulate the interpretation while keeping model's prediction unchanged, e.g., [19, 25]. Moreover, there exists the misinterpretation of Adversarial Examples (AEs) [50], i.e., adversarial inputs are misclassified[2] by the DL model, but interpreted highly similarly to the benign counterparts. Fig. 1 illustrates examples of the aforementioned two types of misinterpretations. In this regard, it is vital to assess how robust the coupled DL model and XAI method are against input perturbations, which motivates this work.
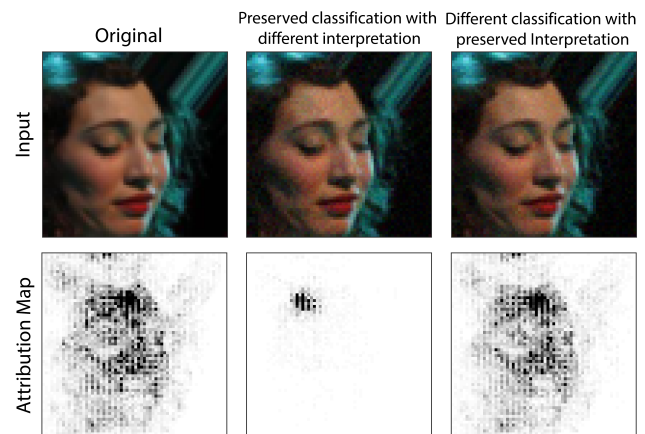

Figure 1: Two types of misinterpretations after perturbation

To answer the question, the first challenge we recognise is the lack of diverse evaluation metrics from the state-ofthe-art. Most of the existing works focus on adversarial attack [20] and defence [16, 42] on explanations, which essentially answer the *binary* question of whether there exist any adversarial interpretation in some perturbation distance. On the other hand, evaluation methods mainly study *worst-case* metrics, e.g., the *maximum* change in the resulting explanations when perturbations are made [3] and local

---

[1]Despite the subtle difference between interpretability and explainability, we use both terms interchangeably as attributes of DL models in this paper. However, as suggested in [31], we use the terms explanation/interpretation specifically for individual predictions.

[2]Without loss of generality, in this paper we assume the DL model is a classifier if with no further clarification.

Lipschitz continuity as the sensitivity to perturbations [48]. However, for *systematic* evaluation, we also need a notion of *how robust in general* the model is whenever a misinterpretation can be found (in line with the insight gained from evaluating classification robustness [45]). We introduce two metrics concerning the *worst-case* interpretation discrepancy and a probabilistic metric to calculate the *proportion* of misinterpretations in the local norm-ball around the original input, that complement each other from different perspectives.

Second, XAI techniques are so heterogeneous that no existing white-box evaluation methods are generic enough to be applicable to all common ones. That said, black-box methods, that only access inputs and outputs of the coupled DL model and XAI tool without requiring any internal information, are promising for all kinds of XAI techniques (including perturbation-based ones that are missing from current literature). Based on this insight, we design a Genetic Algorithm (GA) and a statistical Subset Simulation (SS) approach to estimate the aforementioned two robustness metrics, both of which are of black-box nature.

The third challenge we identified is that misinterpretations are normally rare events in a local norm-ball. Without white-box information like gradients, black-box methods have to leverage auxiliary information to detect such rare events efficiently. To this end, we design bespoke fitness functions in the GA (when solving the optimisation) and retrofit the established SS (dedicated to estimating rare event probabilities [6]) for efficient evaluation.

To the best of our knowledge, no state-of-the-art methods can collectively cope with the three pinpointed challenges like ours. To validate this claim, we conduct experiments to study the accuracy, sensitivity, and efficiency of our methods. Additionally, we develop two practical applications of our methods: i) We evaluate a wide range of XAI techniques and gain insights that no XAI technique is superior in terms of robustness to both types of adversarial attacks. ii) We discover a strong correlation between classification robustness and interpretation robustness through theoretical analysis (see Appx.) and empirical studies. We also identified the best training scheme to improve both aspects.

In summary, the key contributions of this paper include:

- Two diverse metrics, worst-case interpretation discrepancy and probabilistic interpretation robustness, complement each other as a versatile approach, allowing for a holistic evaluation of interpretation robustness.

- We introduce new methods based on GA and SS to estimate these two metrics. These methods are black-box and thus applicable to diverse XAI tools, enabling robustness evaluation of perturbation-based XAI techniques for the first time. Despite the rare occurrence

of misinterpretations, our GA and SS algorithms efficiently detect them.

- We demonstrate two practical applications of our methods: ranking robust XAI techniques and selecting training schemes to improve both classification and interpretation robustness.

2. Related Work

Evaluation of Interpretation Robustness: Existing evaluation metrics, proposed for interpretation robustness, only consider the misinterpretation when the prediction label of perturbed inputs remains unchanged [3]. [3] estimates the Local Lipschitz of interpretation, while [48] introduces the max-sensitivity and average-sensitivity of interpretation. Both of them use Simple Monte Carlo (SMC) sampling to estimate their metrics. [46] formally certify the robustness of gradient-based explanation by propagating a compact input or parameter set as symbolic intervals through the forwards and backwards computations of the neural network (NN). In [14], it defines the consistency as the probability that the inputs with the same interpretation have the same prediction label. However, their evaluation method is only applicable to tree ensemble models and tabular datasets, leaving the probabilistic estimation of misinterpretation for high-dimensional image datasets blank. Notably, toolsets/benchmarks [26, 1] for evaluating XAI techniques are emerging in the last two years. They are not specifically built for evaluating interpretation robustness, thus only concern the aforementioned metrics. That said, our metrics and their efficient estimators can be integrated into and complement those toolsets/benchmarks.

Adversarial Attack and Defence on Interpretation: Ghorbani et al. first introduce the notion of adversarial perturbation to NN interpretation [19]. Afterwards, several works are dedicated to generating indistinguishable inputs which have the same prediction label but substantially different interpretations [20, 39]. The theoretical analysis has shown that the lack of interpretation robustness is related to geometrical properties of NNs [15]. In [50], a new class of attack is proposed to fool the NN's prediction as well as the coupled XAI method. GA is introduced to manipulate SHAP in [8]. In [15], an upper bound on maximum changes of gradient-based interpretation is derived. The upper bound is proportional to the smooth parameter of the softplus activation function, which can be smoothed to improve the interpretation robustness. In [16], regularisation on training, like weight decay, and minimising the hessian of NNs are theoretically proved to be effective for training more robust NNs against interpretation manipulation. In [53], prior knowledge, e.g., from V&V evidence, is used in Bayesian surrogate models for more robust and consistent explanations. Specifically designed for perturbation-based

XAI tools, [10] devises defenses against adversarial attack. We refer readers to [9] for a recent survey on the topic.

3. Preliminaries

3.1. Feature-Attribution based XAI

While readers are referred to [4] for a review, we list common feature-attribution based XAI methods [51, 21] that are studied by this work.

Guided Backpropagation: It computes the gradient of output with respect to the input, but only the non-negative components of gradients are propagated to highlight the important pixels in the image [40].

Gradient $\times$ Input: The map $g(x) = x \odot \frac{\partial f(x)}{\partial x}$ is more preferable to gradient alone to leverage the sign and strength of input to improve the interpretation sharpness [38].

Integrated Gradients: Instead of calculating single derivative, this approach integrates the gradients from some baseline to its current input value $g(x) = (x - \bar{x}) \int_{\alpha=0}^{1} \frac{\partial f(\bar{x}+\alpha(x-\bar{x}))}{\partial x} d\alpha$, addressing the saturation and thresholding problems [41].

GradCAM: Gradient-weighted Class Activation Mapping (Grad-CAM) generates the visual explanation for convolutional neural network, using gradients flowing into the final convolutional layer to produce a coarse localization map, highlighting the relevant regions in the image for prediction[37].

Layer-wise Relevance Propagation (LRP): LRP operates by propagating the outputs $f(x)$ backwards, subject to the conservation rule [7]. Given neurons j and k in two consecutive layers, propagating relevance score R_k to neurons j in lower layer can be expressed as $R_j = \sum_k \frac{z_{jk}}{\sum_j z_{jk}} R_k$ where weight $z_{jk} = w_{jk} x_k$ is the weighted activation, representing the contribution of relevance neuron k makes to neuron j.

DeepLift: It is an improved version of LRP by considering changes in the neuron activation from the reference point when propagating the relevance scores [38]. Rescale rule is used to assign contribution scores to each neuron.

Perturbation-based: LIME trains an interpretable local surrogate model, such as liner regression model, by sampling points around the input sample and use the regression coefficients as interpretation results [35]. SHAP calculates the attribution based on Shapley Values from cooperative game theory [29]. It involves taking the permutation of input features and adding them one by one to the baseline. The output difference after adding input feature corresponds to its attribution.

3.2. Local Robustness of Interpretation

Analogous to the adversarial robustness of classification, interpretation can be fooled by adding perturbations to the input. The interpretation robustness is highly related to the robustness of classification, since the attribution map is produced based on some prediction class. Therefore, we first define the robustness of classification and then formalise the robustness of interpretation, using the following notations. Given an input seed x, we may find a norm ball $B(x, r)$ with the central point at x and radius r in L_p norm. We denote the prediction output of the DL model as the vector $f(x)$ with size equal to the total number of labels.

Classification robustness requires that DL model's prediction output should be invariant to the human imperceptible noise, which can be expressed through the prediction loss around an input seed x

$$J(f(x), f(x')) = \max_{i \neq y}(f_i(x') - f_y(x'))$$
$$y = \arg\max_i f_i(x), \quad x' \in B(x, r) \tag{1}$$

where $f_i(x')$ returns the probability of label i after input x' being processed by the DL model f. Note, $J \geqslant 0$ implies that x' is an AE. We then define the following indicator function for misclassification within the norm ball $B(x, r)$

$$I_c = \begin{cases} -1 & \text{if } J(f(x), f(x')) \geqslant 0 \\ 1 & \text{if } J(f(x), f(x')) < 0 \end{cases} \tag{2}$$

That is, $I_c = -1$ indicates misclassification, otherwise 1.

Previous works study two circumstances when small perturbation fools the interpretation $g(x)$, cf. Fig. 1 for examples. We use the interpretation discrepancy $\mathfrak{D}(g(x), g(x'))$ (defined later) to quantify the difference between the new interpretation $g(x')$ after perturbation and the reference $g(x)$, where $x' \in B(x, r)$. We then introduce two constants as thresholds, α and β, such that $\mathfrak{D} < \alpha$ represents consistent interpretations, while $\mathfrak{D} > \beta$ represents inconsistent interpretations[3]. Two misinterpretation regions within the norm ball $B(x, r)$ are then defined as

$$\widehat{F} = \{\mathfrak{D} > \beta \wedge J < 0\}, \quad \widetilde{F} = \{\mathfrak{D} < \alpha \wedge J \geqslant 0\} \tag{3}$$

$\widehat{F}$ represents preserved classification with different interpretation and $\widetilde{F}$ represents different classification with preserved interpretation, respectively. Note, α and β are hyperparameters that define the consistency notion of interpretations. They may vary case by case in the specific application

---

[3]When $\alpha \leqslant \mathfrak{D} \leqslant \beta$, it represents the case that we cannot clearly decide if the two interpretations are consistent or not.

context, representing the level of strictness required by the users on interpretation robustness. For example, if we use PCC (defined later) to quantify $\mathfrak{D}$, i.e. $\mathfrak{D}=1/\text{PCC}$, there is a rule of thumb [2] that PCC < 0.4 ($\beta = 1/0.4$) indicates inconsistent interpretations while PCC > 0.6 ($\alpha = 1/0.6$) represents consistent interpretations.

3.3. Interpretation Discrepancy Metrics

In order to quantify the visual discrepancy between the XAI results (i.e., attribution maps), there are several commonly used metrics, including Mean Square Error (MSE), Pearson Correlation Coefficient (PCC), and Structural Similarity Index Measure (SSIM) [15]. PCC and SSIM have the absolute values in $[0, 1]$. The smaller values indicate the larger discrepancy between two interpretations. MSE calculates the average squared differences, the value of which more close to 0 means higher similarity. Then, interpretation discrepancy $\mathfrak{D}$ can be expressed as

$$\mathfrak{D} = \frac{1}{\text{PCC}} \text{ or } \frac{1}{\text{SSIM}} \text{ or } \text{MSE} \qquad (4)$$

4. Worst Case Evaluation

The conventional way to evaluate robustness of classification is based on the *worst case* loss under the perturbation [49]. This underlines the adversarial attack and motivates the adversarial training. Similarly, *the worst case interpretation discrepancy between the original input and perturbed input may reflect the interpretation robustness.*

There are two types of misinterpretations after perturbation in a local region, cf. Eq. (3). Accordingly, two optimisations are formalised for the worst case interpretation discrepancy:

$$sol_{\widehat{F}} = \max_{x' \in B(x,r)} \mathfrak{D}(g(x), g(x')) \\ s.t. \ J(f(x), f(x')) < 0 \qquad (5)$$

$$sol_{\widetilde{F}} = \min_{x' \in B(x,r)} \mathfrak{D}(g(x), g(x')) \\ s.t. \ J(f(x), f(x')) \geqslant 0 \qquad (6)$$

That is, $sol_{\widehat{F}}$ corresponds to finding the largest interpretation discrepancy when perturbed input is still correctly classified. While $sol_{\widetilde{F}}$ is the minimum interpretation discrepancy between the AE x' and input seed x.

Previous works adopt white-box methods to solve the above optimisations for adversarial explanations [50, 19], in which case the DL model $f(x)$ and XAI method $g(x)$ are required to be fully accessible to their internal information. In addition, many XAI methods $g(x)$ are non-differentiable, and the strong assumptions (like smoothing gradient of ReLU non-linearity) are made to enable derivative-based optimisation. In contrast, Genetic Algorithm (GA) is a derivative-free method for solving both constrained and unconstrained optimisations, and has been successfully applied to the evaluation of classification robustness [12]. That motivates us to develop a black-box evaluation method for interpretation robustness based on GA. GA consists of 5 steps: *initialisation*, *selection*, *crossover*, *mutation*, and *termination*, the middle three of which are repeated until the convergence of fitness function values. We refer readers to Appx. 8.2 for more details of GA.

Initialisation: The population with N samples is initialized. Diversity of initial population could promise approximate global optimal [27]. Normally, we use the Gaussian distribution with the mean at input seed x, or a uniform distribution to generate a set of diverse perturbed inputs within the norm ball $B(x, r)$.

Selection: The core of GA is the design of fitness functions. Fitness function guides the selection of parents for latter operations. Considering the constrained optimization, we design the fitness function based on the superiority of feasible individuals to make distinction between feasible and infeasible solutions [34]. For the optimisation of Eq. (5), the constraint can be directly encoded as the indicator I_c into the fitness function

$$\mathcal{F}(x') = I_c \, \mathfrak{D}(g(x), g(x')) \qquad (7)$$

and $\mathfrak{D}(g(x), g(x'))$ is always none negative. All feasible individuals satisfying the constraint $J(f(x), f(x')) < 0$ will have $I_c = 1$, and $\mathcal{F} > 0$. If the constraint is violated, then $I_c = -1$, and $\mathcal{F} < 0$. In other words, the individuals violating the constraint will have smaller fitness values than the others and are suppressed during the evolution.

For the optimisation of Eq. (6), we note $J > 0$ is a rare event within the local region $B(x, r)$, as AEs are normally rare [45]. To accelerate the search in the feasible input space, we set two fitness functions $\mathcal{F}_1$ and $\mathcal{F}_2$. $\mathcal{F}_1$ increases the proportion of AEs in the population. On this basis, when over half amount of the population are AEs, $\mathcal{F}_2$ will guide the generation of adversarial explanations.

$$\mathcal{F}_1(x') = J(f(x), f(x')) \\ \mathcal{F}_2(x') = -I_c/\mathfrak{D}(g(x), g(x')) \qquad (8)$$

In $\mathcal{F}_2$, I_c also penalises the violation of constraints, which keeps the optimisation conditioned on AEs. Instead of directly selecting the best fitted individuals, we choose the fitness proportionate selection [28], which can maintain good diversity of population and avoid premature convergence. Then, the probability of selection p_i for each individual x'_i is formulated as

$$p_i = \frac{\mathcal{F}(x'_i)}{\sum_{j=1}^{N} \mathcal{F}(x'_j)} \qquad (9)$$

Crossover: The crossover operator will combine a pair of parents from last step to generate a pair of children, which share many of the characteristics from the parents. The half elements of parents are randomly exchanged.

Mutation: Some elements of children are randomly altered to add variance in the evolution. It should be noticed that the mutated samples should still fall into the norm ball $B(x, r)$. Finally, the children and parents will be the individuals for the next iteration.

Termination: GA terminates either when the allocated computation budget (maximum number of iterations) is depleted or the plateau is reached such that successive iterations no longer produce better results.

5. Probabilistic Evaluation

5.1. Probabilistic Metrics

In addition to the worst case evaluation, probabilistic evaluation based on statistical approaches is of the same practical interest—a lesson learnt from evaluating classification robustness [45, 44] and DL reliability [52, 17]. Thus, we study the *probability of misinterpretation* within $B(x, r)$, regarding the two types of misinterpretations[4] of the input image x under study:

$$P_F(x) = \int_{x' \in B(x,r)} \mathbb{1}_{x' \in F} \, q(x') \, dx', \quad F = \widehat{F} \text{ or } \widetilde{F} \tag{10}$$

where x' is a perturbed sample under the local distribution $q(x')$ (precisely the "input model" used by [45], when studying local probabilistic metric) in $B(x, r)$. $\mathbb{1}_{x' \in F}$ is equal to 1 when $x' \in F$ is true, 0 otherwise. Intuitively, Eq. (10) says, for the given input image x, if we generate an infinite set of perturbed samples locally (i.e., within a norm ball $B(x, r)$) according to the distribution q, then the proportion of those samples fall into the misinterpretation region F is defined as the proposed probabilistic metric.

5.2. Estimation by Subset Simulation

To estimate the two probabilistic metrics defined by Eq. (10), there are two challenges: i) misinterpretations represented by $\widetilde{F}$ and $\widehat{F}$ are arguably rare events (that confirmed empirically later in our experiments); ii) inputs of DL models are usually high dimensional data, like images. The first challenge requires sampling methods specifically designed for rare events rather than SMC (that is known to be inefficient for rare events). The second challenge rules out some commonly used advanced sampling methods, like importance sampling, as they may not be applicable to high dimensional data [5].

---

[4]Through out the paper, we use the shorthand notation F for either $\widehat{F}$ or $\widetilde{F}$, according to the context.

The well-established Subset Simulation (SS) can efficiently calculate the small failure probability in high dimensional space [6] and has been successfully applied to assessing classification robustness of DL models [45]. As a black-box method, it only involves the input and response of interest for calculation, thus generic to diverse XAI methods $g(x)$. The main idea of SS is introducing intermediate failure events so that the failure probability can be expressed as the product of larger conditional probabilities. Let $F = F_m \subset F_{m-1} \subset \cdots \subset F_2 \subset F_1$ be a sequence of increasing events so that $F_m = \bigcap_{i=1}^{m} F_i$. By conditional probability, we get

$$P_F := P(F_m) = P(\bigcap_{i=1}^{m} F_i) = P(F_1) \prod_{i=2}^{m} P(F_i | F_{i-1}) \tag{11}$$

The conditional probabilities of intermediate events involved in Eq. (11) can be chosen sufficiently large so that they can be efficiently estimated. For example, $P(F_1) = 1$, $P(F_i | F_{i-1}) = 0.1$, $i = 2, 3, 4, 5, 6$, then $P_F \approx 10^{-5}$ which is too small for efficient estimation by SMC sampling. In this section, we adapt SS for our problem as what follows.

5.2.1 Design of Intermediate Events

$\widehat{F}$ and $\widetilde{F}$ can be decomposed as the series of intermediate events through the expression of property functions J and $\mathfrak{D}$. For $\widehat{F}$, $J < 0$ is not rare for a well-trained DL model, representing the correctly classified input after perturbation. Thus, the intermediate events $\widehat{F}_{i-1}$ and $\widehat{F}_i$ can be chosen as

$$\widehat{F}_{i-1} = \{I_c \mathfrak{D} > \beta_{i-1}\}, \quad \widehat{F}_i = \{I_c \mathfrak{D} > \beta_i\}$$
$$\text{where} \quad \beta_{i-1} < \beta_i \leqslant \beta \tag{12}$$

such that $\widehat{F}_i \subset \widehat{F}_{i-1}$. I_c (in Eq. 2) encodes the constraint $J < 0$ as the sign of $\mathfrak{D}$.

In contrast, $J \geqslant 0$ in $\widetilde{F}$ represents the occurrence of AEs that are rare events, which cannot be directly expressed as the indicator I_c, since the random sampling within $B(x, r)$ cannot easily satisfy $J \geqslant 0$. Thus, for $\widetilde{F}$, $J \geqslant 0$ should be chosen as the critical intermediate event.

$$\widetilde{F}_j = \{J \geqslant 0\}, \quad \text{where} \quad 1 < j < m \tag{13}$$

For intermediate events $\widetilde{F}_{i-1}$ and $\widetilde{F}_i$, when $i < j$, we set

$$\widetilde{F}_{i-1} = \{J > \gamma_{i-1}\}, \quad \widetilde{F}_i = \{J > \gamma_i\}$$
$$\text{where} \quad \gamma_{i-1} < \gamma_i < 0 \tag{14}$$

such that $\widetilde{F}_j \subset \widetilde{F}_i \subset \widetilde{F}_{i-1}$. And for intermediate events $\widetilde{F}_{k-1}$ and $\widetilde{F}_k$, when $k - 1 > j$, we can set

$$\widetilde{F}_{k-1} = \{-I_c/\mathfrak{D} > 1/\alpha_{k-1}\}, \quad \widetilde{F}_k = \{-I_c/\mathfrak{D} > 1/\alpha_k\}$$
$$\text{where} \quad 0 < \alpha \leqslant \alpha_k < \alpha_{k-1} \tag{15}$$

such that $\widetilde{F}_k \subset \widetilde{F}_{k-1} \subset \widetilde{F}_j$.

5.2.2 Estimating Conditional Probabilities

Upon formally defined intermediate events, the question arises on how to set β_i, γ_i and α_i to make the conditional probability $P(F_i|F_{i-1})$ sufficiently large for estimation by a few simulations. Also, simulating new samples from F_i for estimating next conditional probability $P(F_{i+1}|F_i)$ is difficult due to the rarity of F_i. Therefore, the Markov Chain Monte Carlo sampling based on the Metropolis–Hastings (MH) algorithm is adopted. For simplicity, the intermediate event threshold is generally denoted as $L_i = \{\beta_i, \gamma_i, \alpha_i\}$.

5.2.3 Choices of Intermediate Event Threshold

Start from estimating $P(F_1)$, F_1 is chosen as the common event such that N samples are drawn from $q(\cdot)$ by SMC and all belong to F_1. A feasible way is setting the threshold of property function L_1 to $-\infty$, and $P(F_1) = 1$. For $i = 2, \cdots, m$, L_i affects the values of condition probabilities and hence the efficiency of SS. It is suggested that L_i is set adaptively to make $P(F_i|F_{i-1})$ approximately equals to ρ, and ρ is a hyper-parameter in SS (that takes a decimal less than 1 and normally $\rho = 0.1$ yields good efficiency, although it can be empirically optimised), i.e., $P(F_i|F_{i-1}) \approx \rho$. That is, at each iteration $i - 1$ when we simulate N samples, ρN samples should belong to F_i.

5.2.4 Simulating New Samples from $q(\cdot|F_i)$

At iteration $i = 2, \cdots, m - 1$, we already have ρN samples belonging to F_i and aim to simulate new samples to enlarge the set to N, so that the next conditional probability $P(F_{i+1}|F_i) = \frac{1}{N}\sum_{k=1}^{N} \mathbb{1}_{F_{i+1}}(x'_k)$ can be calculated. We can pick up an existing sample x' subject to the conditional distribution $q(\cdot|F_i)$, denoted as $x' \sim q(\cdot|F_i)$, and use the Metropolis Hastings (MH) algorithm to construct a Markov Chain. By running M steps of MH, the stationary distribution of the Markov Chain is $q(\cdot|F_i)$. Then new data $x'' \sim q(\cdot|F_i)$ can be sampled from the Markov Chain and added into the set. More details of the MH algorithm for SS are presented in Appx. 8.3.

5.2.5 Termination Condition and Returned Estimation

After the aforementioned steps, SS divides the problem of estimating a rare event probability into several simpler ones—a sequence of intermediate conditional probabilities as formulated in Eq. (11). The returned estimation $\bar{P}_F$ and coefficient of variation (c.o.v.) $\bar{\delta}$ (measuring the estimation error) are

$$\bar{P}_F = \prod_{i=1}^{m} \frac{1}{N} \sum_{k=1}^{N} \mathbb{1}_{F_i}(x'_k), \quad \bar{\delta}^2 \approx \sum_{i=1}^{m} \frac{1 - \bar{P}_{F_i}}{\bar{P}_{F_i} N}(1 + \lambda_i)$$

(16)

where $\lambda_i > 0$ represents the efficiency of the estimator using dependent samples drawn from the Markov Chain. For simplicity, we can assume $\lambda_i \approx 0$ when the number of steps M of MH is large [11]. Since each conditional probability $P(F_i|F_{i-1})$ approximately equals to ρ, then by Eq. (11), the returned estimation $\bar{P}_F \approx \rho^{m-1}$. m is the total number of intermediate event generated adaptively. The adaptive generation of intermediate events terminates when $\bar{P}_F < P_{\min}$, and $P_{\min}$ is a given termination threshold. More details of statistical properties of the estimator, like error bound, efficiency are presented in Appx. 8.3.

6. Experiments

6.1. Experiment Setup

We consider three public benchmark datasets , five XAI methods, and five training schemes in our experiments. The norm ball radius, deciding the oracle of robustness, is calculated with respect to the r separation property [47]. That is, $r = 0.3$ for MNIST, $r = 0.03$ for CIFAR10, and $r = 0.05$ for CelebA. More details of the DL models under study are presented in Appx. 8.5. For the probabilistic evaluation using SS, without loss of generality, we consider the uniform distribution as $q(x')$ within each norm ball. We compare $\mathfrak{D}$ = MSE, 1/PCC, and 1/SSIM for measuring interpretation discrepancy in Appx. 8.6, and find PCC is better to quantify the interpretation difference in our cases. Based on sensitivity analysis, we choose hyperparameters PCC thresholds $1/\beta = 0.4$, $1/\alpha = 0.6$, MH steps $M = 250$, $\rho = 0.1$, $\ln P_{\min} = -100$ for probabilistic evaluation, and population size $N = 1000$, number of iteration $itr = 500$ for the worst case evaluation by GA. Our tools and experiments are publicly available at https://github.com/havelhuang/Eval_XAI_Robustness.

6.2. Sensitivity to Hyper-Parameter Settings

We first investigate the sensitivity of objective function $\mathfrak{D}$ and constraint J (cf. Eq. (5) and (6)) to GA's population size and iteration numbers, as shown in Fig. 2. We observe from the 1st row that interpretation discrepancy measured by PCC (the red curve) quickly converge after 300 iterations with the satisfaction of constraint J (the blue curve), showing the effectiveness of our GA. From the 2nd row, we notice that the optimisation is not sensitive to population size, compared with the number of iterations, i.e., population size over 500 cannot make significant improvement to the optimisation. In addition, if the number of iterations is sufficiently large, the effect of population size on optimal solution is further diminished. We only present the results of one seed from CelebA, cf. Appx. 8.7 for more seeds from other datasets, while the general observation remains.

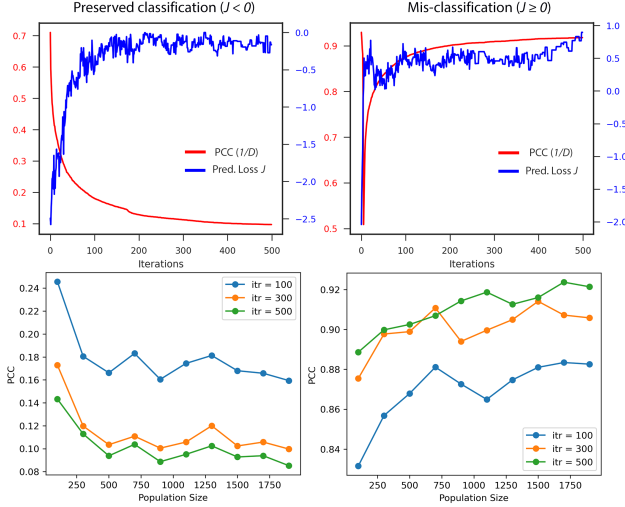Next, we study the sensitivity of SS accuracy to the number of MH steps M, varying the PCC threshold that defines

Figure 2: Sensitivity of objective $\mathfrak{D}$ and constraint J to GA's population size and iteration numbers. Each column represents a type of misinterpretation. 1st row: quickly converged GA objectives satisfying the constraint, with fixed population size of 1000 and varying iterations. 2nd row: GA solutions, with fixed iteration numbers and varying population size. A test seed (representing a norm ball) from CelebA is used; interpretation discrepancy $\mathfrak{D}$ is measured by $1/PCC$; "Gradient×Input" XAI method is studied.

the rarity level of misinterpretation events. In Fig. 3, we can calculate the difference $\Delta \ln P_F$ between SS estimations and the approximated ground truth (by SMC estimations using a sufficiently large number of samples[5]). The 1st row shows the overlapping of SS and SMC estimations (two red curves) and the reducing running time (the blue curve) when decreasing the rarity levels of misinterpretations (by controlling the PCC threshold). From the 2nd row we observe that, with increased MH steps M, the estimation accuracy of SS is significantly improved. In addition, the rarity of misinterpretation events determines the choice of M. E.g., if $\ln P_{\widehat{F}} = -3.87$ with $\widehat{F} = \{PCC < 0.4 \wedge J < 0\}$, then $M = 100$ already achieves high precision without additional sampling budget. Other parameters, e.g. the number of samples n and sample quantile ρ that are discussed in Appx. 8.7, are in general less sensitive than the number of MH steps M.

In summary, sensitivity analysis provides the basis of setting hyper-parameters in later experiments: 500 iterations and 1000 population size for GA, 250 MH steps for SS.

---

[5]We use 10^8 samples (for the specific seed) which can accurately estimate a small probability in natural logarithm around $-17 \sim -18$.
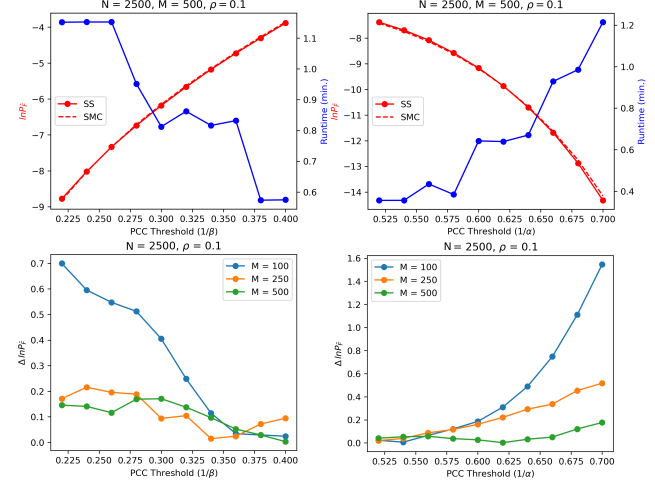

Figure 3: Each column represents a type of misinterpretation. 1st row: the probability of misinterpretation ($\ln P_F$) estimations returned by SS and approximated ground truth by SMC[5], varying the rarity of misinterpretations. Overlapping of two red curves shows high accuracy of SS. 2nd row: sensitivity of SS accuracy $\Delta \ln P_F$ to MH steps M, varying the rarity level of misinterpretation controlled by PCC threshold. A test seed from MNIST and "Gradient×Input" XAI method are used; Results are averaged over 10 runs.

6.3. Accuracy and Efficiency of Evaluation

We study the accuracy of our GA-based evaluation, comparing with state-of-the-art [3, 48]—they define the local Lipschitz ($SENS_{LIPS}$) and max-sensitivity ($SENS_{MAX}$) metrics for the maximum interpretation discrepancy, and empirically estimate the metrics using SMC sampling. For fair comparisons, we first choose MSE as the interpretation discrepancy metric in our fitness functions of GA, and then apply both GA and SMC to generate two populations of interpretations in which we calculate the three robustness metrics respectively and summarise in Table 1. We use 5×10^5 samples for both GA and SMC.

Table 1: Three worst case robustness metrics estimated by our GA and SMC, averaged over 100 test seeds. GA outperforms SMC (used by state-of-the-arts) w.r.t. all 3 metrics.

Dataset	GA			SMC		
	MSE ($sol_{\widehat{F}}$)	$SENS_{MAX}$	$SENS_{LIPS}$	MSE	$SENS_{MAX}$	$SENS_{LIPS}$
MNIST	**1.549**	**36.067**	**13.747**	0.271	15.226	2.772
CIFAR10	**42.436**	**328.147**	**314.861**	0.589	38.529	40.232
CelebA	**3.204**	**192.203**	**65.635**	0.013	11.298	3.563

As shown in Table 1, our GA-based estimator outperforms SMC in all of the three robustness metrics. Although the metrics of local Lipschitz and max-sensitivity are not explicitly encoded as optimisation objectives in our GA, GA

is still more effective and efficient to estimate those three extreme values than SMC. This is non-surprising, since all three metrics are compatible and essentially representing the same worst-case semantics. That said, our interpretation discrepancy metric complements $SENS_{LIPS}$ and $SENS_{MAX}$ (as the former is based on Lipschitz value while the latter defined only in L_2 norm), can be easily encoded in our GA.

In addition to the accuracy shown in Fig. 3, we compare the sample efficiency between SS and SMC by calculating the number of required simulations N_{SS} and N_{SMC} for achieving same estimation errors (measured by c.o.v. δ). As shown in Table 2, SS requires fewer samples, showing great advantage over SMC, cf. Appx. 8.3 for theoretical analysis.

Table 2: Sample efficiency of SS and SMC. In all six cases, SS requires fewer samples ($N_{SS} < N_{SMC}$) than SMC for achieving the same estimation errors δ^2. Each result is averaged over 10 seeds.

Dataset	F	$\ln P_F$	δ^2	N_{SS}	N_{SMC}
MNIST	$\widehat{F}$	-12.25	0.0184	**15000**	1.13×10^7
	$\widetilde{F}$	-24.63	0.0374	**27500**	1.34×10^{12}
CIFAR10	$\widehat{F}$	-0.79	0.0004	**2500**	2500
	$\widetilde{F}$	-33.54	0.0511	**40000**	7.22×10^{15}
CelebA	$\widehat{F}$	-31.43	0.0482	**35000**	9.29×10^{14}
	$\widetilde{F}$	-70.71	0.1090	**80000**	4.68×10^{31}

6.4. Evaluating XAI Methods

The first application of our methods is to draw insights on the robustness of common XAI techniques, from both the worst-case and probabilistic perspectives. Thanks to the black-box nature of GA and SS, our methods are applicable to diverse XAI tools, and we consider six popular ones in this section. In Appx. 8.8, we evaluate other XAI tools and discuss how the number of perturbed samples and image segmentation affect evaluation results on LIME and SHAP (which are missing from current literature).

We randomly sample 100 seeds from MNIST for evaluations, and summarise the statistics as box-and-whisker plots in Fig. 4. Based on the empirical results of Fig. 4, we may conclude: i) Perturbation-based XAI method also suffers from the lack of robustness. ii) for misinterpretation $\widehat{F}$—correct classification ($J < 0$) with inconsistent interpretation ($PCC < 0.4$), DeepLift and Integrated Gradients outperform others, while Guided Backprop and Gradient×Input are unrobust from both worst-case and probabilistic perspective; iii) for misinterpretation $\widetilde{F}$—wrong classification ($J \geqslant 0$) with persevered interpretation ($PCC > 0.6$), while all XAI methods perform similarly w.r.t. both metrics, LRP shows better robustness than others.

The empirical insights are as expected if we consider the mechanisms behind those XAI methods. For instance, considering $\widehat{F}$, DeepLift and Integrated Gradients are more ro-
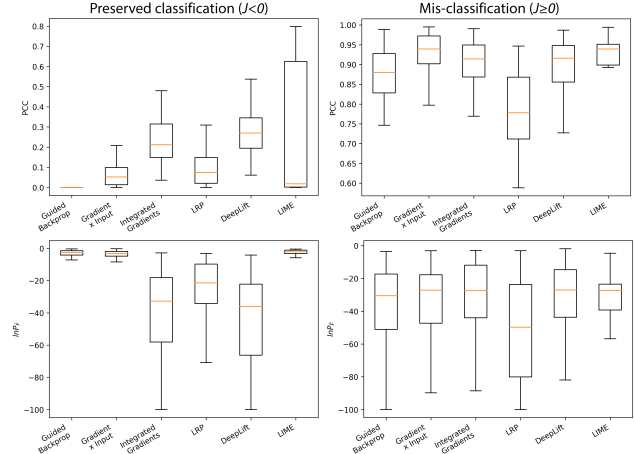

Figure 4: Worst-case (1st row) and probabilistic (2nd row) robustness evaluations of five XAI methods based on 100 random seeds from MNIST. Each column represents a type of misinterpretation—$\widehat{F}$ left and $\widetilde{F}$ right. For top-left plot, higher value means more robust; for all other plots, lower value means more robust.

bust, since they use the reference point to avoid the discontinuous gradients (large curvature) that mislead the attribution maps [38]. On the other hand, DeepLift and Integrated Gradients become vulnerable to $\widetilde{F}$. Because misclassification and misinterpretation are rare events, most perturbed inputs inside the norm ball have consistent interpretation with the seed. Consequently, the integration from the reference point which averages the attribution map over several points are prone to produce the consistent interpretations. See Appx. 8.8 for more discussions and experiments on CIFAR10 and CelebA dataset.

6.5. Evaluating Training Schemes

In this application, we study the effect of various training schemes on the interpretation robustness of DL models. In Appx. 8.1, we theoretically analyse the relation between classification robustness and interpretation robustness. The Prop. 1 shows that input hessian norm and input gradient norm are related to the change of classification loss and interpretation discrepancy. Thus, we add input gradient and input hessian regularisation terms to the training loss, and also consider the PGD-based adversarial training that improves classification robustness through minimising the maximal prediction loss in norm balls [22, 23]. Table 3 records the results.

In addition to the knowledge that input hessian can defence adversarial interpretation [16], we notice that it is significant and effective in improving both classification and interpretation robustness, than input gradient regularisation, confirming our Prop. 1. Moreover, we discover that adver-

Table 3: Evaluating classification (c) and interpretation ($\widehat{F}$ and $\widetilde{F}$) robustness of DL models, trained with input gradient norm regularisation (Grad. Reg.), input hessian norm regularisation (Hess. Reg.), both of them (Grad. + Hess. Reg.) and adversarial training (Adv. Train.). Results are averaged over 100 random seeds. Higher $sol_{\widehat{F}}$ means more robust, while for other metrics, the lower is the better.

Dataset	Model	Worst Case Evaluation			Probabilistic Evaluation		
		sol_c (J)	$sol_{\widehat{F}}$ (PCC)	$sol_{\widetilde{F}}$ (PCC)	$\ln P_c$	$\ln P_{\widehat{F}}$	$\ln P_{\widetilde{F}}$
MNIST	Org.	22.43	0.06	0.93	-24.28	-3.87	-31.47
	Grad. Reg	11.37	0.10	0.92	-31.51	-15.69	-44.96
	Hess. Reg.	10.59	0.17	0.90	-33.36	-21.27	-43.85
	Grad. + Hess.	10.04	0.20	0.90	-36.96	-23.79	-46.19
	Adv. Train.	**-0.16**	**0.21**	**0.59**	**-84.15**	**-28.67**	**-89.09**
CIFAR10	Org.	42.58	0.02	0.85	-31.55	-18.63	-71.46
	Grad. Reg	42.34	0.01	0.85	-27.31	-21.77	-65.75
	Hess. Reg.	8.99	0.08	0.81	-76.29	-99.20	-91.89
	Grad. + Hess.	8.47	0.06	0.81	-71.65	-98.49	-92.39
	Adv. Train.	**-0.67**	**0.25**	**0.80**	**-92.57**	**-100**	**-95.97**
CelebA	Org.	51.08	0.08	0.86	-13.77	-21.58	-70.82
	Grad. Reg	25.29	0.06	0.88	-45.52	-70.22	-83.26
	Hess. Reg.	18.71	0.09	0.86	-74.93	-100	**-95.85**
	Grad. + Hess.	25.41	0.06	0.88	-65.95	-100	-94.13
	Adv. Train.	**-0.45**	**0.55**	**0.81**	**-95.09**	**-100**	-95.58

sarial training is surprisingly effective at improving interpretation robustness, but at the price of dropping accuracy, cf. Appx. 8.5. This phenomenon reals the strong correlation between classification and interpretation robustness. That said, the improvement of classification robustness may lead to the improvement of interpretation robustness.

7. Discussion and Conclusion

In this work, we formalise two types of misinterpretations caused by adversarial perturbations on the input, and then introduce both worst-case and probabilistic metrics to study their interpretation robustness.

We first apply GA to solve the optimisation problem for estimating the worst-case metric, which outperforms state-of-the-arts in the category of black-box evaluations. Indeed, GA is metaheuristic and cannot be guaranteed with global-optimal solutions, which is common for most (if not all) black-box optimisation algorithms. To increase the likelihood of finding optimal solutions, we carefully design fitness functions and conduct extensive experiments on selecting hyperparameters to validate the effectiveness of our GA-based method. Notably, GA solves the challenges that most XAI methods are non-differentiable (e.g., perturbation-based LIME and SHAP) and heterogeneous, and it can optimise diverse problems formulated as discrete or continuous (multi-objective) functions, which allows for greater flexibility in formalising the distance between explanations.

We then formulate a statistical inference problem to estimate the probabilistic metric. The SS method was chosen to solve the problem, because: (1) it has the capability of es-

timating rare event probabilities; (2) it may efficiently deal with high dimensional data; (3) it comes with nice statistical properties—SS is asymptotically unbiased with N goes to infinity and provides statistical guarantees on estimation errors (e.g., the upper bound on fractional bias and c.o.v, cf. Prop. 2 and 3 in appendix).

In summary, we proposes two versatile and efficient evaluation methods for DL interpretation robustness. The versatility is twofold: (1) the proposed metrics are characterising robustness from both worst-case and probabilistic perspectives; (2) GA and SS are black-box methods thus generic to heterogeneous XAI methods. Considering the rare-event nature of misinterpretations, GA and SS show high efficiency in detecting them, thanks to the bespoke design of fitness functions in GA and encoding auxiliary information as intermediate events in SS. The conclusion is supported by our extensive experiments and two applications.

References

[1] Chirag Agarwal, Eshika Saxena, Satyapriya Krishna, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. Openxai: Towards a transparent evaluation of model explanations. *arXiv preprint arXiv:2206.11104*, 2022. 2

[2] Haldun Akoglu. User's guide to correlation coefficients. *Turkish journal of emergency medicine*, 18(3):91–93, 2018. 4

[3] David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018. 1, 2, 7

[4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020. 1, 3

[5] Siu-Kui Au and JL Beck. Important sampling in high dimensions. *Structural safety*, 25(2):139–163, 2003. 5

[6] Siu-Kui Au and James L Beck. Estimation of small failure probabilities in high dimensions by subset simulation. *Probabilistic engineering mechanics*, 16(4):263–277, 2001. 2, 5, 13, 14

[7] 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):e0130140, 2015. 3

[8] Hubert Baniecki and Przemyslaw Biecek. Manipulating shap via adversarial data perturbations. In *Proc. of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12907–12908, 2022. 2

[9] Hubert Baniecki and Przemyslaw Biecek. Adversarial attacks and defenses in explainable artificial intelligence: A survey. *IJCAI'23 Workshop on XAI*, 2023. 3

[10] Zachariah Carmichael and Walter J Scheirer. Unfooling perturbation-based post hoc explainers. In *Proc. of the 37th AAAI Conference on Artificial Intelligence (AAAI'23)*, 2022. 3

[11] Frédéric Cérou, Pierre Del Moral, Teddy Furon, and Arnaud Guyader. Sequential monte carlo for rare event estimation. *Statistics and computing*, 22(3):795–808, 2012. 6, 14, 15

[12] Jinyin Chen, Mengmeng Su, Shijing Shen, Hui Xiong, and Haibin Zheng. Poba-ga: Perturbation optimized black-box adversarial attacks via genetic algorithm. *Computers & Security*, 85:89–106, 2019. 4

[13] Jinyin Chen, Mengmeng Su, Shijing Shen, Hui Xiong, and Haibin Zheng. POBA-GA: perturbation optimized black-box adversarial attacks via genetic algorithm. *Comput. Secur.*, 85:89–106, 2019. 13

[14] Sanjoy Dasgupta, Nave Frost, and Michal Moshkovitz. Framework for evaluating faithfulness of local explanations. *arXiv preprint arXiv:2202.00734*, 2022. 2

[15] Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 4, 16

[16] Ann-Kathrin Dombrowski, Christopher J Anders, Klaus-Robert Müller, and Pan Kessel. Towards robust explanations for deep neural networks. *Pattern Recognition*, 121:108194, 2022. 1, 2, 8, 12

[17] Yi Dong, Wei Huang, Vibhav Bharti, Victoria Cox, Alec Banks, Sen Wang, Xingyu Zhao, Sven Schewe, and Xiaowei Huang. Reliability Assessment and Safety Arguments for Machine Learning Components in System Assurance. *ACM Trans. Embedded Computing Systems*, 2022. 5

[18] Andrew Gelman, Walter R Gilks, and Gareth O Roberts. Weak convergence and optimal scaling of random walk metropolis algorithms. *The annals of applied probability*, 7(1):110–120, 1997. 14

[19] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of Neural Networks Is Fragile. *Proc. of the AAAI Conference on Artificial Intelligence*, 33(01):3681–3688, 2019. 1, 2, 4

[20] Juyeon Heo, Sunghwan Joo, and Taesup Moon. Fooling neural network interpretations via adversarial model manipulation. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2

[21] Xiaowei Huang, Daniel Kroening, Wenjie Ruan, James Sharp, Youcheng Sun, Emese Thamo, Min Wu, and Xinping Yi. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence,

and interpretability. *Computer Science Review*, 37:100270, 2020. 1, 3

[22] Gaojie Jin, Xinping Yi, Wei Huang, Sven Schewe, and Xiaowei Huang. Enhancing adversarial training with second-order statistics of weights. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15273–15283, 2022. 8

[23] Gaojie Jin, Xinping Yi, Dengyu Wu, Ronghui Mu, and Xiaowei Huang. Randomized adversarial training via taylor expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16447–16457, 2023. 8

[24] Lambros S Katafygiotis and Konstantin M Zuev. Geometric insight into the challenges of solving high-dimensional reliability problems. *Probabilistic Engineering Mechanics*, 23(2-3):208–218, 2008. 14

[25] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (Un)reliability of Saliency Methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 267–280. Springer, Cham, 2019. 1

[26] 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, 2020. 2

[27] Abdullah Konak, David W Coit, and Alice E Smith. Multi-objective optimization using genetic algorithms: A tutorial. *Reliability engineering & system safety*, 91(9):992–1007, 2006. 4, 12

[28] Adam Lipowski and Dorota Lipowska. Roulette-wheel selection via stochastic acceptance. *Physica A: Statistical Mechanics and its Applications*, 391(6):2193–2196, 2012. 4, 13

[29] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017. 3

[30] Zbigniew Michalewicz and Marc Schoenauer. Evolutionary algorithms for constrained parameter optimization problems. *Evolutionary computation*, 4(1):1–32, 1996. 13

[31] Christoph Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.* E-book on leanpub.com, 2020. 1

[32] Daniel Omeiza, Helena Webb, Marina Jirotka, and Lars Kunze. Explanations in autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):10142–10162, 2021. 1

[33] Iason Papaioannou, Wolfgang Betz, Kilian Zwirglmaier, and Daniel Straub. Mcmc algorithms for subset simulation. *Probabilistic Engineering Mechanics*, 41:89–103, 2015. 14

[34] David Powell and Michael M. Skolnick. Using genetic algorithms in engineering design optimization with non-linear constraints. In *Proceedings of the 5th International Conference on Genetic Algorithms*, page 424–431, San Francisco, CA, USA, 1993. Morgan Kaufmann Publishers Inc. 4, 13

[35] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of

Any Classifier. In *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, New York, NY, USA, 2016. ACM. 3

[36] Gerhart Iwo Schueller, Helmuth J Pradlwarter, and Phaedon-Stelios Koutsourelakis. A critical appraisal of reliability estimation procedures for high dimensions. *Probabilistic engineering mechanics*, 19(4):463–474, 2004. 14

[37] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proc. of the IEEE Int. Conf. on Computer Vision*, pages 618–626, 2017. 3

[38] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017. 3, 8

[39] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020. 2, 17

[40] J Springenberg, Alexey Dosovitskiy, Thomas Brox, and M Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR (workshop track)*, 2015. 3

[41] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. 3

[42] Ruixiang Tang, Ninghao Liu, Fan Yang, Na Zou, and Xia Hu. Defense against explanation manipulation. *Frontiers in Big Data*, 5, 2022. 1

[43] Bas HM Van der Velden, Hugo J Kuijf, Kenneth GA Gilhuijs, and Max A Viergever. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, page 102470, 2022. 1

[44] Benjie Wang, Stefan Webb, and Tom Rainforth. Statistically robust neural network classification. In *Proc. of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of UAI'21 of *PMLR*, pages 1735–1745, 2021. 5

[45] Stefan Webb, Tom Rainforth, Yee Whye Teh, and M. Pawan Kumar. A statistical approach to assessing neural network robustness. In *7th Int. Conf. on Learning Representations (ICLR'19)*. OpenReview.net, 2019. 2, 4, 5

[46] Matthew Wicker, Juyeon Heo, Luca Costabello, and Adrian Weller. Robust explanation constraints for neural networks. *arXiv preprint arXiv:2212.08507*, 2022. 2

[47] Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Russ R Salakhutdinov, and Kamalika Chaudhuri. A closer look at accuracy vs. robustness. *Advances in neural information processing systems*, 33:8588–8601, 2020. 6

[48] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 7

[49] Fuxun Yu, Zhuwei Qin, Chenchen Liu, Liang Zhao, Yanzhi Wang, and Xiang Chen. Interpreting and evaluating neural network robustness. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 4199–4205, 2019. 4

[50] Xinyang Zhang, Ningfei Wang, Hua Shen, Shouling Ji, Xiapu Luo, and Ting Wang. Interpretable Deep Learning under Fire. In *29th USENIX Security Symposium*, USENIX Security 20. USENIX Association, Aug. 2020. 1, 2, 4

[51] Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2021. 3

[52] Xingyu Zhao, Wei Huang, Alec Banks, Victoria Cox, David Flynn, Sven Schewe, and Xiaowei Huang. Assessing the Reliability of Deep Learning Classifiers Through Robustness Evaluation and Operational Profiles. In *AISafety'21 Workshop at IJCAI'21*, volume 2916. ceur-ws.org, 2021. 5

[53] Xingyu Zhao, Wei Huang, Xiaowei Huang, Valentin Robu, and David Flynn. BayLIME: Bayesian local interpretable model-agnostic explanations. In *Proc. of the 37th Conference on Uncertainty in Artificial Intelligence*, volume 161 of *UAI'21*, pages 887–896. PMLR, 2021. 2, 17