



Fooling Partial Dependence via Data Poisoning

Hubert Baniecki^(✉), Wojciech Kretowicz, and Przemysław Biecek

Warsaw University of Technology, Warsaw, Poland

{hubert.baniecki.stud, wojciech.kretowicz.stud, przemyslaw.biecek}@pw.edu.pl

Abstract. Many methods have been developed to understand complex predictive models and high expectations are placed on post-hoc model explainability. It turns out that such explanations are not robust nor trustworthy, and they can be fooled. This paper presents techniques for attacking Partial Dependence (plots, profiles, PDP), which are among the most popular methods of explaining any predictive model trained on tabular data. We showcase that PD can be manipulated in an adversarial manner, which is alarming, especially in financial or medical applications where auditability became a must-have trait supporting black-box machine learning. The fooling is performed via poisoning the data to bend and shift explanations in the desired direction using genetic and gradient algorithms. We believe this to be the first work using a genetic algorithm for manipulating explanations, which is transferable as it generalizes both ways: in a model-agnostic and an explanation-agnostic manner.

Keywords: Explainable AI · Adversarial ML · Interpretability

1 Introduction

Although supervised machine learning became state-of-the-art solutions to many predictive problems, there is an emerging discussion on the underspecification of such methods which exhibits differing model behaviour in training and practical setting [14]. This is especially crucial when proper accountability for the systems supporting decisions is required by the domain [35, 38, 42]. Living with black-boxes, several explainability methods were presented to help us understand models' behaviour [5, 19, 23, 36, 40], many are designed specifically for deep neural networks [6, 34, 44, 49]. Explanations are widely used in practice through their (often estimation-based) implementations available to machine learning practitioners in various software contributions [4, 8, 11].

Nowadays, robustness and certainty become crucial when using explanations in the data science practice to understand black-box machine learning models; thus, facilitate rationale explanation, knowledge discovery and responsible decision-making [9, 22]. Notably, several studies evaluate explanations

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-26409-2_8.

© The Author(s) 2023

M.-R. Amini et al. (Eds.): ECML PKDD 2022, LNAI 13715, pp. 121–136, 2023.

https://doi.org/10.1007/978-3-031-26409-2_8

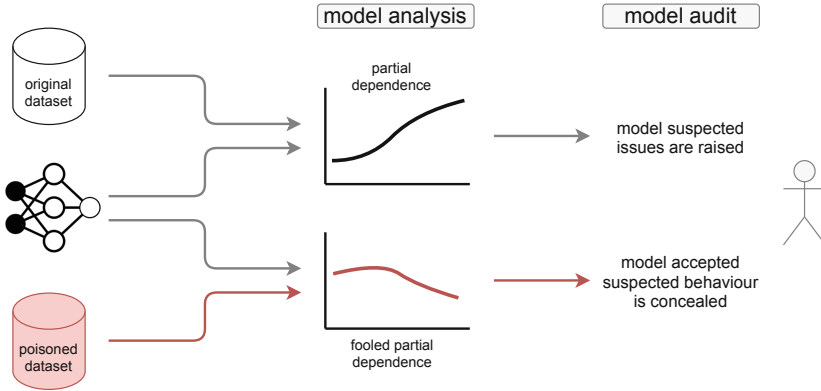


Fig. 1. Framework for fooling model explanations via data poisoning. The red color indicates the adversarial route, a potential security breach, which an attacker may use to manipulate the explanation. Researchers could use this method to provide a misleading rationale for a given phenomenon, while auditors may purposely conceal the suspected, e.g. biased or irresponsible, reasoning of a black-box model. (Color figure online)

[1, 2, 10, 27, 29, 51] showcasing their various flaws from which we perceive an existing robustness gap; in critical domains, one can call it a *security breach*. Apart from promoting wrong explanations, this phenomenon can be exploited to use adversarial attacks on model explanations to achieve manipulated results. In regulated areas, these types of attacks may be carried out to deceive an auditor.

Figure 1 illustrates a process in which the developer aims to conceal the undesired behaviour of the model by supplying a poisoned dataset for model audit. Not every explanation is equally good—just as models require proper performance validation, we need similar assessments of explainability methods. In this paper, we evaluate the robustness of Partial Dependence (PD) [19], moreover highlight the possibility of adversarial manipulation of PD (see Figs. 4 and 5 in the latter part of the paper). We summarize the contributions as follows:

- (1) We introduce a novel concept of using a *genetic algorithm* for manipulating model explanations. This allows for a convenient generalization of the attacks in a model-agnostic and explanation-agnostic manner, which is not the case for most of the related work. Moreover, we use a gradient algorithm to perform fooling via data poisoning efficiently for neural networks.
- (2) We explicitly target PD to highlight the potential of their adversarial manipulation, which was not done before. Our method provides a sanity check for the future use of PD by responsible machine learning practitioners. Evaluation of the constructed methods in extensive experiments shows that model complexity significantly affects the magnitude of the possible explanation manipulation.

2 Related Work

In the literature, there is a considerable amount of attacks on model explanations specific to deep neural networks [16, 21, 25, 30, 53]. At their core, they provide

various algorithms for fooling neural network interpretability and explainability, mainly of image-based predictions. Such explanations are commonly presented through saliency maps [45], where each model input is given its attribution to the prediction [6, 43, 44, 49]. Although explanations can be used to improve the adversarial robustness of machine learning models [37], we target explanations instead. When considering an explanation as a function of model and data, there is a possibility to change one of these variables to achieve a different result [54]. Heo et al. [25] and Dimanov et al. [15] propose fine-tuning a neural network to undermine its explainability capabilities and conceal model unfairness. The assumption is to alter the model’s parameters without a drop in performance, which can be achieved with an objective function minimizing the distance between explanations and an arbitrarily set target. Note that [15] indirectly change partial dependence while changing the model. Aivodji et al. [3] propose creating a surrogate model aiming to approximate the unfair black-box model and explain its predictions in a fair manner, e.g. with relative variable dependence.

Alternate idea is to fool fairness and explainability via data change since its (background) distribution greatly affects interpretation results [28, 30]. Solans et al. [48] and Fukuchi et al. [20] investigate concealing unfairness via data change by using gradient methods. Dombrowski et al. [16] propose an algorithm for saliency explanation manipulation using gradient-based data perturbations. In contrast, we introduce a genetic algorithm and focus on other machine learning predictive models trained on tabular data. Slack et al. [46] contributed adversarial attacks on post-hoc, model-agnostic explainability methods for *local-level* understanding; namely LIME [40] and SHAP [36]. The proposed framework provides a way to construct a biased classifier with safe explanations of the model’s predictions.

Since we focus on *global-level* explanations, instead, the results will modify a view of overall model behaviour, not specific to a single data point or image. Lakkaraju and Bastani [32] conducted a thought-provoking study on misleading effects of manipulated Model Understanding through Subspace Explanations (MUSE) [33], which provide arguments for why such research becomes crucial to achieve responsibility in machine learning use. Further, the robustness of neural networks [13, 50] and counterfactual explanations [47] have become important, as one wants to trust black-box models and extend their use to sensitive tasks. Our experiments further extend to global explanations the indication of Jia et al. [29] that there is a correlation between model complexity and explanation quality.

3 Partial Dependence

In this paper, we target one of the most popular explainability methods for tabular data, which at its core presents the expected value of the model’s predictions as a function of a selected variable. Partial Dependence, formerly introduced as *plots* by Friedman [19], show the expected value fixed over the marginal joint distribution of other variables. These values can be relatively easily estimated and are widely incorporated into various tools for model explainability [7, 8, 11, 24, 39]. The theoretical explanation has its practical estimator used to

compute the results, later visualized as a line plot showing the expected prediction for a given variable; also called *profiles* [12]. PD for model f and variable c in a random vector $\mathcal{X}$ is defined as $\mathcal{PD}_c(\mathcal{X}, z) := E_{\mathcal{X}_{-c}} [f(\mathcal{X}^{c=z})]$, where $\mathcal{X}^{c=z}$ stands for random vector $\mathcal{X}$, where c -th variable is replaced by value z . By $\mathcal{X}_{-c}$, we denote distribution of random vector $\mathcal{X}$ where c -th variable is set to a constant. We defined PD in point z as the expected value of model f given the c -th variable is set to z . The standard estimator of this value for data X is given by the following formula $\widehat{\mathcal{PD}}_c(X, z) := \frac{1}{N} \sum_{i=1}^N f(X_i^{c=z})$, where X_i is the i -th row of the matrix X and the previously mentioned symbols are used accordingly. To simplify the notation, we will use $\mathcal{PD}$, and omit z and c where context is clear.

4 Fooling Partial Dependence via Data Poisoning

Many explanations treat the dataset X as fixed; however, this is precisely a *single point of failure* on which we aim to conduct the attack. In what follows, we examine $\mathcal{PD}$ behaviour by looking at it as a function whose argument is an entire dataset. For example, if the dataset has N instances and P variables, then $\mathcal{PD}$ is treated as a function over $N \times P$ dimensions. Moreover, because of the complexity of black-box models, $\mathcal{PD}$ becomes an extremely high-dimensional space where variable interactions cause unpredictable behaviour. Explanations are computed using their estimators where a significant simplification may occur; thus, a slight shift of the dataset used to calculate $\mathcal{PD}$ may lead to unintended results (for example, see [26] and the references given there).

We aim to change the underlying dataset used to produce the explanation in a way to achieve the desired change in $\mathcal{PD}$. Figure 1 demonstrates the main threat of an adversarial attack on model explanation using data poisoning, which is *concealing the suspected behaviour of black-box models*. The critical assumption is that an adversary has a possibility to modify the dataset arbitrarily, e.g. in healthcare and financial audit, or research review. Even if this would not be the case, in modern machine learning, wherein practice dozens of variables are used to train complex models, such data shifts might be only a minor change that a person looking at a dataset or distribution will not be able to identify.

We approach fooling $\mathcal{PD}$ as an optimization problem for given criteria of attack efficiency, later called the attack loss. It originates from [16], where a similar loss function for manipulation of local-level model explanations for an image-based predictive task was introduced. This work introduces the attack loss that aims to change the output of a global-level explanation via data poisoning instead. The explanation’s weaknesses concerning data distribution and causal inference are exploited using two ways of optimizing the loss:

- **Genetic-based**¹ algorithm that does not make any assumption about the model’s structure – the black-box path from data inputs to the output prediction; thus, is model-agnostic. Further, we posit that for a vast number

¹ For convenience, we shorten the *algorithm based on the genetic algorithm* phrase to *genetic-based algorithm*.

- of explanations, clearly post-hoc global-level, the algorithm does not make assumption about their structure either; thus, becomes *explanation-agnostic*.
- **Gradient-based** algorithm that is specifically designed for models with differentiable outputs, e.g. neural networks [15, 16].

We discuss and evaluate two possible fooling strategies:

- **Targeted attack** changes the dataset to achieve the closest explanation result to the predefined desired function [16, 25].
- **Robustness check** aims to achieve the most distant model explanation from the original one by changing the dataset, which refers to the sanity check [1].

For practical reasons, we define the distance between the two calculated $\mathcal{PD}$ vectors as $\|x - y\| := \frac{1}{I} \sum_{i=1}^I (x_i - y_i)^2$, yet other distance measures may be used to extend the method.

4.1 Attack Loss

The intuition behind the attacks is to find a modified dataset that minimizes the attack loss. A changed dataset denoted as $X \in \mathbb{R}^{N \times P}$ is an argument of that function; hence, an optimal X is a result of the attack. Let $Z \subset \mathbb{R}$ be the set of points used to calculate the explanation. Let $T : Z \rightarrow \mathbb{R}$ be the target explanation; we write just T to denote a vector over whole Z . Let $g_c^Z : \mathbb{R}^{N \times P} \rightarrow \mathbb{R}^{|Z|}$ be the actual explanation calculated for points in Z ; we write g_c for simplicity. Finally, let $X' \in \mathbb{R}^{N \times P}$ be the original (constant) dataset. We define the attack loss as $\mathcal{L}(X) := \mathcal{L}^{g, s}(X)$, where g is the explanation to be fooled, and an objective is minimized depending on the strategy of the attack, denoted as s . The aim is to minimize $\mathcal{L}$ with respect to the dataset X used to calculate the explanation. We never change values of the explained variable c in the dataset.

In the **targeted attack**, we aim to **minimize** the distance between the target model behaviour T and the result of model explanation calculated on the changed dataset. We denote this strategy by t and define $\mathcal{L}^{g, t}(X) = \|g_c(X) - T\|$. Since we focus on a specific model-agnostic explanation, we substitute $\mathcal{PD}$ in place of g to obtain $\mathcal{L}^{\mathcal{PD}, t}(X) = \|\mathcal{PD}_c(X) - T\|$. This substitution can be generalized for various global-level model explanations, which rely on using a part of the dataset for computation.

In the **robustness check**, we aim to **maximize** the distance between the result of model explanation calculated on the original dataset $g_c(X')$, and the changed one; thus, minus sign is required. We denote this strategy by r and define $\mathcal{L}^{g, r}(X) = -\|g_c(X) - g_c(X')\|$. Accordingly, we substitute $\mathcal{PD}$ in place of g to obtain $\mathcal{L}^{\mathcal{PD}, r}(X) = -\|\mathcal{PD}_c(X) - \mathcal{PD}_c(X')\|$. Note that $\mathcal{L}^{g, s}$ may vary depending on the explanation used, specifically for $\mathcal{PD}$ it is useful to centre the explanation before calculating the distances, which is the default behaviour in our implementation: $\mathcal{L}^{\overline{\mathcal{PD}}, r}(X) = -\|\overline{\mathcal{PD}}_c(X) - \overline{\mathcal{PD}}_c(X')\|$, where $\overline{\mathcal{PD}}_c := \mathcal{PD}_c(X) - \frac{1}{|Z|} \sum_{z \in Z} \mathcal{PD}_c(X, z)$. We consider the second approach of

comparing explanations using centred $\overline{\mathcal{PD}}$, as it forces changes in the shape of the explanation, instead of promoting to shift the profile vertically while the shape changes insignificantly.

4.2 Genetic-Based Algorithm

We introduce a novel strategy for fooling explanations based on the genetic algorithm as it is a simple yet powerful method for real parameter optimization [52]. We do not encode genes conventionally but deliberately use this term to distinguish from other types of evolutionary algorithms [18]. The method will be invariant to the model’s definition and the considered explanations; thus, it becomes model-agnostic and explanation-agnostic. These traits are crucial when working with black-box machine learning as versatile solutions are convenient.

Fooling $\mathcal{PD}$ in both strategies include a similar genetic algorithm. The main idea is to define an individual as an instance of the dataset, iteratively perturb its values to achieve the desired explanation target, or perform the robustness check to observe the change. These individuals are initialized with a value of the original dataset X' to form a population. Subsequently, the initialization ends with mutating the individuals using a higher-than-default variance of perturbations. Then, in each iteration, they are randomly crossed, mutated, evaluated with the attack loss, and selected based on the loss values. **Crossover** swaps columns between individuals to produce new ones, which are then added to the population. The number of swapped columns can be randomized; also, the number of parents can be parameterized. **Mutation** adds Gaussian noise to the individuals using scaled standard deviations of the variables. It is possible to constrain the data change into the original range of variable values; also keep some variables unchanged. **Evaluation** calculates the loss for each individual, which requires to compute explanations for each dataset. **Selection** reduces the number of individuals using rank selection, and elitism to guarantee several best individuals to remain in the next population.

We considered the crossover through an exchange of rows between individuals, but it might drastically shift the datasets and move them apart. Additionally, a worthy mutation is to add or subtract whole numbers from the integer-encoded (categorical) variables. We further discuss the algorithm’s details in the Supplementary material. The introduced attack is model-invariant because no derivatives are needed for optimization, which allows evaluating explanations of black-box models. While we found this method a sufficient generalization of our framework, there is a possibility to perform a more efficient optimization assuming the prior knowledge concerning the structure of model and explanation.

4.3 Gradient-Based Algorithm

Gradient-based methods are state-of-the-art optimization approaches, especially in the domain of deep neural networks [34]. This algorithm’s main idea is to use gradient descent to optimize the attack loss, considering the differentiability of the model’s output with respect to input data. Such assumption allows for a faster and more accurate convergence into a local minima using one of

the stochastic optimizers; in our case, Adam [31]. Note that the differentiability assumption is with respect to input data, not with respect to the model’s parameters. We shall derive the gradients $\nabla_{X_{-c}} \mathcal{L}^{g, s}$ for fooling explanations based on their estimators, not the theoretical definitions. This is because the input data is assumed to be a random variable in a theoretical definition of $\mathcal{PD}$, making it impossible to calculate a derivative over the input dataset. In practice, we do not derive our method directly from the definition as the estimator produces the explanation.

Although we specifically consider the usage of neural networks because of their strong relation to differentiation, the algorithm’s theoretical derivation does not require this type of model. For brevity, we derive the theoretical definitions of gradients $\nabla_{X_{-c}} \mathcal{L}^{\mathcal{PD}, t}$, $\nabla_{X_{-c}} \mathcal{L}^{\mathcal{PD}, r}$, and $\nabla_{X_{-c}} \mathcal{L}^{\overline{\mathcal{PD}}, r}$ in the Supplementary material. Overall, the gradient-based algorithm is similar to the genetic-based algorithm in that we aim to iteratively change the dataset used to calculate the explanation. Nevertheless, its main assumption is that the model provides an interface for the differentiation of output with respect to the input, which is not the case for black-box models.

5 Experiments

Setup. We conduct experiments on two predictive tasks to evaluate the algorithms and conclude with illustrative scenario examples, which refer to the framework shown in Fig. 1. The first dataset is a synthetic regression problem that refers to the Friedman’s work [19] where inputs X are independent variables uniformly distributed on the interval $[0, 1]$, while the target y is created according to the formula: $y(X) = 10 \sin(\pi \cdot X_1 \cdot X_2) + 20(X_3 - 0.5)^2 + 10X_4 + 5X_5$. Only 5 variables are actually used to compute y , while the remaining are independent of. We refer to this dataset as **friedman** and target explanations of the variable X_1 . The second dataset is a real classification task from UCI [17], which has 5 continuous variables, 8 categorical variables, and an evenly-distributed binary target. We refer to this dataset as **heart** and target explanations of the variable **age**. Additionally, we set the discrete variables as constant during the performed fooling because we mainly rely on incremental change in the values of continuous variables, and categorical variables are out of the scope of this work.

Results. Figure 2 present the main result of the paper, which is that PD can be manipulated. We use the gradient-based algorithm to change the explanations of feedforward neural networks via data poisoning. The targeted attack aims to arbitrarily change the monotonicity of PD, which is evident in both predictive tasks. The robustness check finds the most distant explanation from the original one. We perform the fooling 30 times for each subplot, and the Y-axis denotes the model’s architecture: layers×neurons. We observe that PD explanations are especially vulnerable in complex models.

Next, we aim to evaluate the PD of various state-of-the-art machine learning models and their complexity levels; we denote: Linear Model (LM), Random Forest (RF), Gradient Boosting Machine (GBM), Decision Tree (DT),

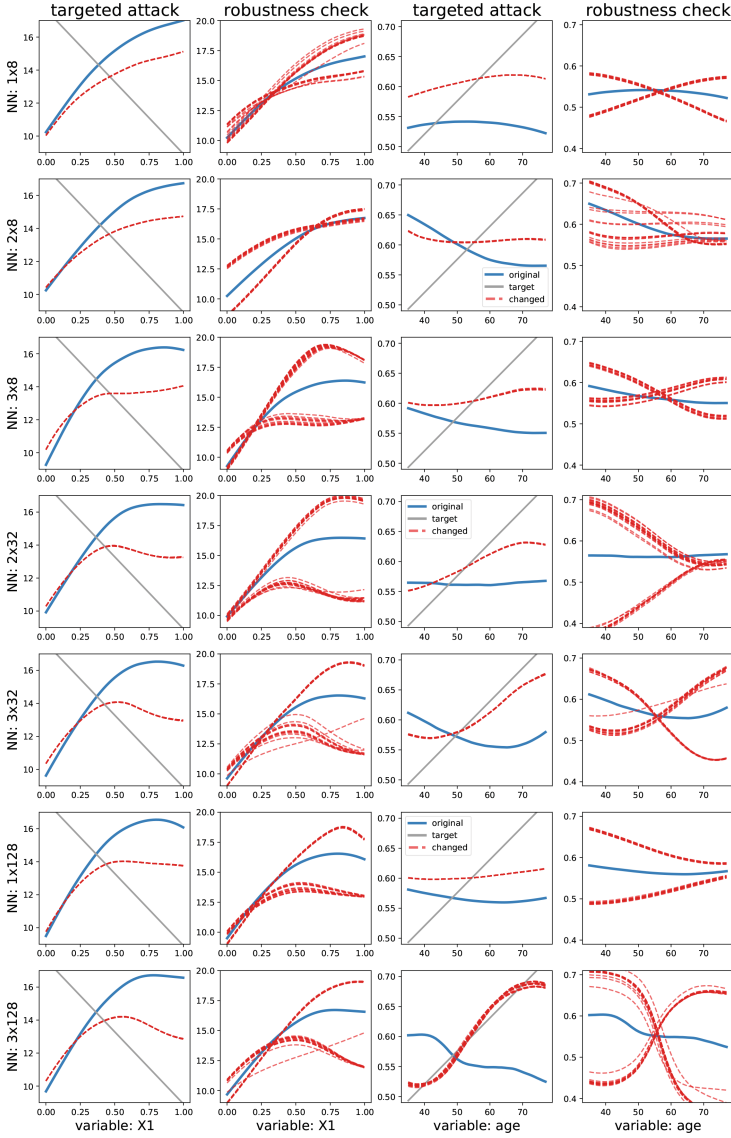


Fig. 2. Fooling Partial Dependence of neural network models (rows) fitted to the **friedman** and **heart** datasets (columns). We performed multiple randomly initiated gradient-based fooling algorithms on the explanations of variables X_1 and **age** respectively. The blue line denotes the original explanation, the red lines are the fooled explanations, and in the targeted attack, the grey line denotes the desired target. We observe that the explanations' vulnerability greatly increases with model complexity. Interestingly, the algorithm seems to converge to two contrary optima when no target is provided. (Color figure online)

K-Nearest Neighbours (KNN), feedforward Neural Network (NN). The model-agnostic nature of the genetic-based algorithm allows this comparison as it might be theoretically and/or practically impossible to differentiate the model’s output with respect to the input; thus, differentiate the explanations and loss.

Table 1. Attack loss values of the robustness checks for Partial Dependence of various machine learning models (**top**), and complexity levels of tree-ensembles (**bottom**). Each value corresponds to the scaled distance between the original explanation and the changed one. We perform the fooling 6 times and report the mean $\pm$ sd. We observe that the explanations’ vulnerability increases with GBM complexity.

Task \ Model							
	LM	RF	GBM	DT	KNN	NN	SVM
friedman	0 $\pm$ 0	152 $\pm$ 76	127 $\pm$ 71	332 $\pm$ 172	164 $\pm$ 61	269 $\pm$ 189	576 $\pm$ 580
heart	2 $\pm$ 3	20 $\pm$ 5	77 $\pm$ 28	798 $\pm$ 192	133 $\pm$ 21	501 $\pm$ 52	451 $\pm$ 25

Task \ Model	Trees						
		10	20	40	80	160	320
friedman	GBM	57 $\pm$ 12	114 $\pm$ 20	157 $\pm$ 37	176 $\pm$ 20	189 $\pm$ 8	210 $\pm$ 9
	RF	233 $\pm$ 22	219 $\pm$ 25	219 $\pm$ 9	201 $\pm$ 23	216 $\pm$ 13	209 $\pm$ 15
heart	GBM	1 $\pm$ 0	3 $\pm$ 1	29 $\pm$ 4	70 $\pm$ 24	152 $\pm$ 56	321 $\pm$ 95
	RF	62 $\pm$ 7	55 $\pm$ 3	29 $\pm$ 9	21 $\pm$ 6	14 $\pm$ 5	13 $\pm$ 2

Table 1 presents the results of robustness checks for Partial Dependence of various machine learning models and complexity levels. Each value corresponds to the distance between the original explanation and the changed one; multiplied by 10^3 in **friedman** and 10^6 in **heart** for clarity. We perform the checks 6 times and report the mean $\pm$ standard deviation. Note that we cannot compare the values between tasks, as their magnitudes depend on the prediction range. We found the explanations of NN, SVM and deep DT the most vulnerable to the fooling methods (**top** Table). In contrast, RF seems to provide robust explanations; thus, we further investigate the relationship between the tree-models’ complexity and the explanation stability (**bottom** Table) to conclude that an increasing complexity yields more vulnerable explanations, which is consistent with Fig. 2. We attribute the differences between the results for RF and GBM to the concept of bias-variance tradeoff. In some cases (**heart**, RF), explanations of too simple models become vulnerable too, since underfitted models may be as uncertain as overfitted ones.

Ablation Study. We further discuss the additional results that may be of interest to gain a broader context of this work. Figure 3 presents the distinction between the robustness check for centred Partial Dependence, which is the default algorithm, and the robustness check for *not centred* PD. We use the gradient-based algorithm to change the explanations of a 3 layers $\times$ 32 neurons ReLU neural

network and perform the fooling 30 times for each subplot. We observe that centring the explanation in the attack loss definition is necessary to achieve the change in explanation shape. Alternatively, the explanation shifts upwards or downwards by essentially changing the mean of prediction. This observation was consistent across most of the models despite their complexity.

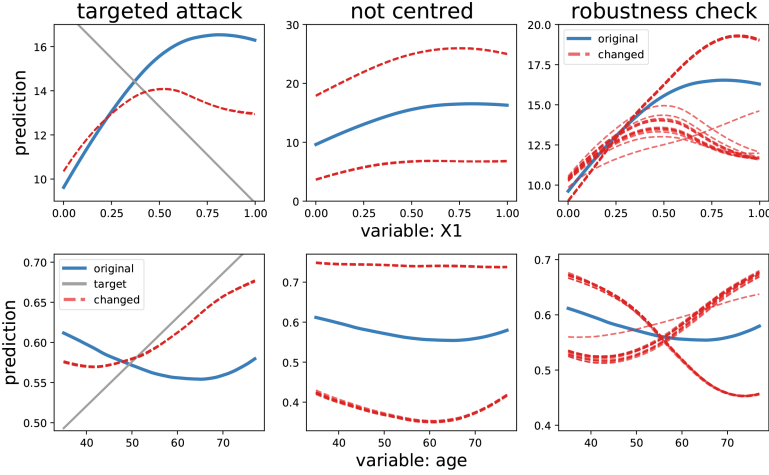


Fig. 3. Fooling Partial Dependence of a 3×32 neural network fitted to the **friedman** (**top** row) and **heart** (**bottom** row) datasets. We performed multiple randomly initiated gradient-based fooling algorithms on the explanations of variables X_1 and **age** respectively. We observe that centring PD is beneficial because it stops the manipulated explanation from shifting.

Table 2. Attack loss values of the robustness checks for PD of various ReLU neural networks. We add additional noise variables to the data before model fitting, e.g. **friedman**+2 denotes the referenced dataset with 2 additional variables sampled from the normal distribution. We perform the fooling 30 times and report the mean $\pm$ sd. We observe that the explanations’ vulnerability greatly increases with task complexity.

Task	NN						
	1×8	2×8	3×8	2×32	3×32	1×128	3×128
friedman	25 ± 3	33 ± 0	75 ± 24	100 ± 32	98 ± 42	54 ± 15	97 ± 50
friedman +1	31 ± 2	40 ± 4	50 ± 9	106 ± 40	115 ± 44	57 ± 15	114 ± 55
friedman +2	34 ± 1	40 ± 10	50 ± 22	106 ± 52	115 ± 50	50 ± 15	137 ± 66
friedman +4	46 ± 6	33 ± 0	83 ± 8	145 ± 31	163 ± 27	40 ± 5	140 ± 58
friedman +8	71 ± 9	47 ± 3	89 ± 15	204 ± 25	176 ± 25	39 ± 6	156 ± 34
heart	11 ± 0	8 ± 1	10 ± 0	32 ± 3	41 ± 5	6 ± 1	134 ± 14
heart +1	10 ± 1	17 ± 6	17 ± 2	44 ± 4	57 ± 13	6 ± 1	128 ± 8
heart +2	13 ± 1	31 ± 13	17 ± 5	63 ± 4	79 ± 10	14 ± 2	218 ± 82
heart +4	13 ± 1	21 ± 9	30 ± 17	113 ± 4	139 ± 60	29 ± 5	232 ± 36
heart +8	16 ± 0	28 ± 18	43 ± 20	125 ± 49	227 ± 28	25 ± 8	311 ± 283

Table 2 presents the impact of additional noise variables in data on the performed fooling. We observe that higher data dimensions favor vulnerable explanations (higher loss). The analogous results for targeted attack were consistent; however, showcased almost zero variance (partially observable in Figs. 2 and 3).

Adversarial Scenario. Following the framework shown in Fig. 1, we consider three stakeholders apparent in explainable machine learning: developer, auditor, and prediction recipients. Let us assume that the model predicting a heart attack should not take into account a patient’s `sex`; although, it might be a valuable predictor. An auditor analyses the model using Partial Dependence; therefore, the developer supplies a poisoned dataset for this task. Figure 5 presents two possible outcomes of model audit: concealed and suspected, which are unequivocally bound to the explanation result and dataset. In first, the model is unchanged while the stated assumption of even dependence between `sex` is concealed (equal to about 0.5); thus, the prediction recipients become vulnerable. Additionally, we supply an alternative scenario where the developer wants to provide evidence of model unfairness to raise suspicion (dependence for class 0 equal to about 0.7).

Supportive Scenario. In this work, we consider an equation of three variables: data, model, and explanation; thus, we poison the data to fool the explanation while the model remains unchanged. Figures 4 and 5 showcase an exemplary data shift occurring in the dataset after the attack where changing only a few explanatory variables results in bending PD. We present a moderate change in data distribution to introduce a concept of analysing such relationships for explanatory purposes, e.g. the first result might suggest that resting blood pressure and maximum heart rate contribute to the explanation of `age`; the second result suggests how these variables contribute to the explanation of `sex`. We conclude that the data shift is worth exploring to analyse variable interactions in models.

6 Limitations and Future Work

We find these results both alarming and informative yet proceed to discuss the limitations of the study. First is the assumption that, in an adversarial scenario, the auditor has no access to the original (unknown) data, e.g. in research or healthcare audit. While the detectability of fooling is worth analyzing, our work focuses not only on an adversarial manipulation of PD, as we sincerely hope such data poisoning is not occurring in practice. Even more, we aim to underline the crucial context of data distribution in the interpretation of explanations and introduce a new way of evaluating PD; black-box explanations by generalizing the methods with genetic-based optimization.

Another limitation is the size of the used datasets. We have engaged with larger datasets during experiments but were turned off by a contradictory view that increasing the dataset size might be considered as exaggerating the results. PD clearly becomes more complex with increasing data dimensions; moreover,

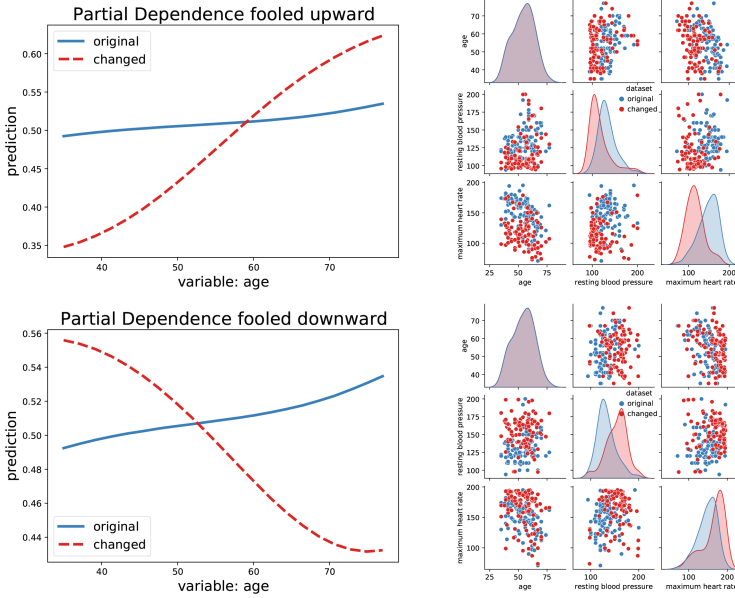


Fig. 4. Partial Dependence of **age** in the SVM model prediction of a heart attack (class 0). **Left:** Two manipulated explanations suggest an increasing or decreasing relationship between **age** and the predicted outcome depending on a desired outcome. **Right:** Distribution of the explained variable **age** and the two poisoned variables from the data, in which the remaining ten variables attributing to the explanation remain unchanged. The mean of the variables' Jensen-Shannon distance equals only 0.027 in the upward scenario and 0.021 in the downward scenario, which might seem like an insignificant change of the data distribution.

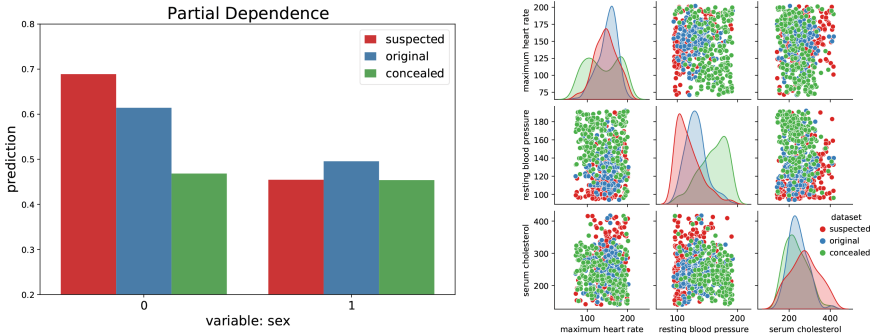


Fig. 5. Partial Dependence of **sex** in the SVM model prediction of a heart attack (class 0). **Left:** Two manipulated explanations present a suspected or concealed variable contribution into the predicted outcome. **Right:** Distribution of the three poisoned variables from the data, in which **sex** and the remaining nine variables attributing to the explanation remain unchanged. The mean of the variables' Jensen-Shannon distance equals only 0.023 in the suspected scenario and 0.026 in the concealed scenario.

higher-dimensional space should entail more possible ways of manipulation, which is evident in the ablation study. We note that in practice, the explanations might require 100–1000 observations for the estimation (e.g. kernel SHAP, PDP), hence the size of the datasets in this study. Finally, we omit datasets like Adult and COMPAS because they mainly consist of categorical variables.

Future Work. We foresee several directions for future work, e.g. evaluating the successor to PD – Accumulated Local Effects (ALE) [5]; although the practical estimation of ALE presents challenges. Second, the attack loss may be enhanced by regularization, e.g. penalty for substantial change in data or mean of model’s prediction, to achieve more meaningful fooling with less evidence. We focus in this work on univariate PD, but targeting bivariate PD can also be examined. Overall, the landscape of global-level, post-hoc model explanations is a broad domain, and the potential of a security breach in other methods, e.g. SHAP, should be further examined. Enhancements to the model-agnostic and explanation-agnostic genetic algorithm are thereby welcomed.

Another future direction would be to enhance the stability of PD. Rieger and Hansen [41] present a defence strategy against the attack via data change [16] by aggregating various explanations, which produces robust results without changing the model.

7 Conclusion and Impact

We highlight that Partial Dependence can be maliciously altered, e.g. bent and shifted, with adversarial data perturbations. The introduced genetic-based algorithm allows evaluating explanations of any black-box model. Experimental results on various models and their sizes showcase the hidden debt of model complexity related to explainable machine learning. Explanations of low-variance models prove to be robust to the manipulation, while very complex models should not be explained with PD as they become vulnerable to change in reference data distribution. Robustness checks lead to varied modifications of the explanations depending on the setting, e.g. may propose two opposite PD, which is why it is advised to perform the checks multiple times.

This work investigates the vulnerability of global-level, post-hoc model explainability from the adversarial setting standpoint, which refers to the responsibility and security of the artificial intelligence use. Possible manipulation of PD leads to the conclusion that explanations used to explain black-box machine learning may be considered black-box themselves. These explainability methods are undeniably useful through implementations in various popular software. However, just as machine learning models cannot be developed without extensive testing and understanding of their behaviour, their explanations cannot be used without critical thinking. We recommend ensuring the reliability of the explanation results through the introduced methods, which can also be used to study models behaviour under the data shift. Code for this work is available at <https://github.com/MI2DataLab/fooling-partial-dependence>.

Acknowledgements. This work was financially supported by the NCN Opus grant 2017/27/B/ST6/01307 and NCN Sonata Bis-9 grant 2019/34/E/ST6/00052.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: *NeurIPS* (2018)
2. Adebayo, J., Muelly, M., Liccardi, I., Kim, B.: Debugging tests for model explanations. In: *NeurIPS* (2020)
3. Aivodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., Tapp, A.: Fairwashing: the risk of rationalization. In: *ICML* (2019)
4. Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K.T., et al.: iNNvestigate neural networks! *J. Mach. Learn. Res.* **20**(93), 1–8 (2019)
5. Apley, D.W., Zhu, J.: Visualizing the effects of predictor variables in black box supervised learning models. *J. Roy. Stat. Soc. Ser. B (Stat. Methodol.)* **82**(4), 1059–1086 (2020)
6. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* **10**(7), 1–46 (2015)
7. Baniecki, H., Biecek, P.: modelStudio: interactive studio with explanations for ML predictive models. *J. Open Source Softw.* **4**(43), 1798 (2019)
8. Baniecki, H., Kretowicz, W., Piatyszek, P., Wisniewski, J., Biecek, P.: dalex: responsible machine learning with interactive explainability and fairness in Python. *J. Mach. Learn. Res.* **22**(214), 1–7 (2021)
9. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bannetot, A., Tabik, S., et al.: Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **58**, 82–115 (2020)
10. Bhatt, U., Weller, A., Moura, J.M.F.: Evaluating and aggregating feature-based model explanations. In: *IJCAI* (2020)
11. Biecek, P.: DALEX: explainers for complex predictive models in R. *J. Mach. Learn. Res.* **19**(84), 1–5 (2018)
12. Biecek, P., Burzykowski, T.: *Explanatory Model Analysis*. Chapman and Hall/CRC (2021)
13. Boopathy, A., et al.: Proper network interpretability helps adversarial robustness in classification. In: *ICML* (2020)
14. D’Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., et al.: Under-specification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395* (2020)
15. Dimanov, B., Bhatt, U., Jamnik, M., Weller, A.: You shouldn’t trust me: learning models which conceal unfairness from multiple explanation methods. In: *AAAI SafeAI* (2020)
16. Dombrowski, A.K., Alber, M., Anders, C., Ackermann, M., Müller, K.R., Kessel, P.: Explanations can be manipulated and geometry is to blame. In: *NeurIPS* (2019)
17. Dua, D., Graff, C.: *UCI Machine Learning Repository* (2017). <https://www.kaggle.com/ronitf/heart-disease-uci/version/1>
18. Elbeltagi, E., Hegazy, T., Grierson, D.: Comparison among five evolutionary-based optimization algorithms. *Adv. Eng. Inform.* **19**(1), 43–53 (2005)
19. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**(5), 1189–1232 (2001)

20. Fukuchi, K., Hara, S., Maehara, T.: Faking fairness via stealthily biased sampling. In: AAAI (2020)
21. Ghorbani, A., Abid, A., Zou, J.: Interpretation of neural networks is fragile. In: AAAI (2019)
22. Gill, N., Hall, P., Montgomery, K., Schmidt, N.: A responsible machine learning workflow with focus on interpretable models, post-hoc explanation, and discrimination testing. *Information* **11**(3), 137 (2020)
23. Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E.: Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. *J. Comput. Graph. Stat.* **24**(1), 44–65 (2015)
24. Greenwell, B.M.: pdp: an R package for constructing partial dependence plots. *R J.* **9**(1), 421–436 (2017)
25. Heo, J., Joo, S., Moon, T.: Fooling neural network interpretations via adversarial model manipulation. In: NeurIPS (2019)
26. Hooker, G.: Generalized functional ANOVA diagnostics for high-dimensional functions of dependent variables. *J. Comput. Graph. Stat.* **16**(3), 709–732 (2007)
27. Hooker, S., Erhan, D., Kindermans, P.J., Kim, B.: A benchmark for interpretability methods in deep neural networks. In: NeurIPS (2019)
28. Janzing, D., Minorics, L., Blöbaum, P.: Feature relevance quantification in explainable AI: a causal problem. In: AISTATS (2020)
29. Jia, Y., Frank, E., Pfahringer, B., Bifet, A., Lim, N.: Studying and exploiting the relationship between model accuracy and explanation quality. In: Oliver, N., Pérez-Cruz, F., Kramer, S., Read, J., Lozano, J.A. (eds.) ECML PKDD 2021. LNCS (LNAI), vol. 12976, pp. 699–714. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86520-7_43
30. Kindermans, P.-J., et al.: The (un)reliability of saliency methods. In: Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R. (eds.) Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. LNCS (LNAI), vol. 11700, pp. 267–280. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-28954-6_14
31. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: ICLR (2015)
32. Lakkaraju, H., Bastani, O.: “How do i fool you?”: Manipulating user trust via misleading black box explanations. In: AIES (2020)
33. Lakkaraju, H., Kamar, E., Caruana, R., Leskovec, J.: Faithful and customizable explanations of black box models. In: AIES (2019)
34. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**(7553), 436–444 (2015)
35. Lipton, Z.C.: The mythos of model interpretability. *Queue* **16**(3), 31–57 (2018)
36. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: NeurIPS (2017)
37. Mangla, P., Singh, V., Balasubramanian, V.N.: On saliency maps and adversarial robustness. In: Hutter, F., Kersting, K., Lijffijt, J., Valera, I. (eds.) ECML PKDD 2020. LNCS (LNAI), vol. 12458, pp. 272–288. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67661-2_17
38. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. *Artif. Intell.* **267**, 1–38 (2019)
39. Molnar, C., Casalicchio, G., Bischl, B.: iml: an R package for interpretable machine learning. *J. Open Source Softw.* **3**(26), 786 (2018)
40. Ribeiro, M.T., Singh, S., Guestrin, C.: “Why should i trust you?”: Explaining the predictions of any classifier. In: KDD (2016)

41. Rieger, L., Hansen, L.K.: A simple defense against adversarial attacks on heatmap explanations. In: ICML WHI (2020)
42. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019)
43. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* **128**(2), 336–359 (2020)
44. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: ICML (2017)
45. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps. In: ICLR (2014)
46. Slack, D., Hilgard, S., Jia, E., Singh, S., Lakkaraju, H.: Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. In: AIES (2020)
47. Slack, D., Hilgard, S., Lakkaraju, H., Singh, S.: Counterfactual explanations can be manipulated. In: NeurIPS (2021)
48. Solans, D., Biggio, B., Castillo, C.: Poisoning attacks on algorithmic fairness. In: Hutter, F., Kersting, K., Lijffijt, J., Valera, I. (eds.) ECML PKDD 2020. LNCS (LNAI), vol. 12457, pp. 162–177. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67658-2_10
49. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: ICML (2017)
50. Wang, Z., Wang, H., Ramkumar, S., Mardziel, P., Fredrikson, M., Datta, A.: Smoothed geometry for robust attribution. In: NeurIPS (2020)
51. Warnecke, A., Arp, D., Wressnegger, C., Rieck, K.: Evaluating explanation methods for deep learning in security. In: IEEE EuroS&P (2020)
52. Wright, A.H.: Genetic algorithms for real parameter optimization. *Found. Genet. Algorithms* **1**, 205–218 (1991)
53. Zhang, X., Wang, N., Shen, H., Ji, S., Luo, X., Wang, T.: Interpretable deep learning under fire. In: USENIX Security (2020)
54. Zhao, Q., Hastie, T.: Causal interpretations of black-box models. *J. Bus. Econ. Stat.* **39**(1), 272–281 (2019)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

