This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

SAND2022-12331C

Assessing the Fidelity of Explanations with Global Sensitivity Analysis

Michael R. Smith and Erin Acquesta
and Charles Smutz
Sandia National Laboratories
Albuquerque, New Mexico
{msmith4, eacques, csmutz}@sandia.gov

Ahmad Rushdi*
Stanford University
Stanford, California
rushdi@stanford.edu

Blake Moss
Sandia National Laboratories
Albuquerque, New Mexico
bmosss@sandia.gov

Abstract

Many explainability methods have been proposed as a means of understanding how a learned machine learning model makes decisions and as an important factor in responsible and ethical artificial intelligence. However, explainability methods often do not fully and accurately describe a model's decision process. We leverage the mathematical framework of global sensitivity analysis techniques to reveal deficiencies of explanation methods. We find that current explainaiblity methods fail to capture prediction uncertainty and make several simplifying assumptions that have significant ramifications on the accuracy of the resulting explanations. We show that the simplifying assumptions result in explanations that: (1) fail to model nonlinear interactions in the model *and (2)* misrepresent the importance of correlated features. *Experiments suggest that failing to capture nonlinear feature interaction has a larger impact on the accuracy of the explanations. Thus, as most state-of-the-art ML models have non-linear interactions and operate on correlated data, explanations should only be used with caution.*

Keywords: Explainability, Machine Learning, Artificial Intelligence, Fidelity, Sensitivity Analysis

1. Introduction

Artificial intelligence (AI) and Machine learning (ML) techniques are being used in increasingly more high-consequence applications such as malware detection [1], autonomous vehicles [2], and medical diagnoses [3]. Wide-spread adoption, though, is limited due to a recognized need to trust the models before they are deployed and integrated into larger systems. In response, several explainable AI (XAI) techniques have emerged [4]. However, we show that current XAI methods themselves often lack verifiable foundations and uncertainty quantification—leaving the end user or ML practitioner to decide *if the explanation itself is valid* and what to do with the provided information [5]. Many XAI methods have justified their approaches by examining how similar they are to how an end-user would explain a decision [6] with accompanying frameworks that measure the accuracy of explainability in the context of user-based explanations [7, 8]. While these have laid a foundation, reliance on human evaluation for the accuracy of explanations may bias explanations towards persuasive explanations rather than accurately describing the learned model [9] and can persuade users to accept incorrect model outputs [5]. Computational-based explanation accuracy or fidelity remains an open research question, with most prior work on XAI fidelity focusing on modified backpropagation and saliency-based methods [10, 11, 12].

In this paper, we examine the fidelity of post-hoc, model agnostic explainability methods using non-intuitive domains where the meaning is *not* obvious through visual inspection. We look to the validation and verification (V&V) principles that ensure the correctness of computational modeling and simulation in many science and engineering disciplines [13]. Uncertainty quantification (UQ) and sensitivity analysis (SA) are fundamental elements of most V&V practices and are often applied in tandem [14]. It is our intent to examine global SA (GSA) [15] methods that are well suited for data-driven UQ analysis with the goal of evaluating the credibility of current XAI methods and further developing credible explanations of an ML model. GSA methods share the objective with XAI of determining the most influential input on the model output and have similar implementations.

Our work builds on previous work highlighting several challenges in XAI. Specifically, Gosiewska and Biecek [16] caution against the use of additive explanations such as LIME and SHAP—specifically highlighting that they do not model feature interactions. Others have also highlighted this challenge for gradient

*Work done while at Sandia National Laboratories.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

based XAI methods [17]. Our work seeks to quantify these claims and demonstrate the impact of not quantifying feature interactions.

Other work also questions credibility of the explanations from XAI methods. Chen et al. [18] question the use of game-theoretic methods to provide explanations and argue that XAI methods should be application specific. Kumar et al. [19] highlight several mathematical problems with SHAP explanations and show that they do not suit human-centric goals of explainability. Frye et al. [20] highlight the problematic assumption of uncorrelated features showing that they give incorrect explanations, hide implicit model dependence on sensitive attributes and lead to ambiguous explanations. Other challenges with feature-importance-based explanations have also been highlighted such as low robustness of explanations (robustness assumes that similar inputs should have similar explanations) [21]. Our work maps current XAI methods into a mathematically-based framework, GSA, to identify the deficiencies of XAI and to bridge the work in the two communities.

Our primary contributions include: (1) a process leveraging GSA to evaluate the credibility of XAI methods, (2) developing a definition of explanation fidelity relative to a ground truth explanation found by a closed-form analytical solution using GSA, (3) exposing limitations of post-hoc XAI methods—specifically that nonlinear interactions in a model have larger ramifications on the fidelity of an explanation than input correlations, and (4) identifying research gaps that, if addressed, would result in higher-fidelity explanations. Highlighting these deficiencies helps channel research efforts to address them and increase the confident usage of XAI methods in high-consequence applications.

The rest of the paper is organized as follows. We first present preliminary background on GSA which motivates our approach. Section 3 discusses the trust of black box learned models and the connection between XAI with GSA. Section 4 presents our definition of explanation fidelity. We then use this definition of fidelity to empirically examine the fidelity of several XAI methods in a GSA framework in Section 5 before concluding in Section 6.

2. Global Sensitivity Analysis Methods

GSA apportions the influence that model parameter or input *uncertainties* have on the *uncertainty* of the model output [15]. GSA has a history of application to black box models in the science and engineering disciplines and is gaining more traction for systems modeling and policy support [22].

Let $\mathbf{x} \in \mathbb{R}^d$ represent an input vector, $y \in \mathbb{R}$ represent an output (label) and $f : \mathbf{x} \mapsto y$ represent a function that maps $\mathbf{x}$ to y. If we assume that the model parameters do not change, such as inference for an ML model, GSA proportions the uncertainty in each of the d input variables on the output uncertainty measured by a quantity of interest (QoI). Two common methods are: (1) variance-based Sobol' indices [23] and (2) game theoretic Shapley values [24]. Here, the output of GSA is a set of feature importance values $\Phi := \{\phi_j\}_{j=1}^d$ for each variable x_j.

Central to GSA is a proper design of experiments which has four core decisions [25]:

1. **Sampling the data.** How to sample the data to represent uncertainty in the inputs while preserving the statistical properties of the training data introducing only marginal standard error.

2. **Modeling Controlled and Uncontrolled Random Behavior.** Running sufficient replicates or trials to understand the behavior of the model

3. **Quantity of Interest.** measuring an appropriate metric—particularly for explainability purposes and capturing appropriate uncertainties.

4. **Proportioning Output Uncertainty to Input Uncertainties.** Methods to apportion the source of input uncertainty to the output uncertainty.

2.1. Sobol' Indices

Assuming that $\mathbf{x}$ is composed of d *mutually independent* random variables, and that the output y is a scalar, the high-dimensional model representation expands a multivariate function $y = f(\mathbf{x})$ as:

$$\begin{aligned} y =& f_0 + \sum_{j=1}^{d} f_j(x_j) + \sum_{k=j+1}^{d} f_{j,k}(x_j, x_k) + \cdots \\ &+ f_{1,2,\ldots,d}(x_1, x_2, \ldots, x_d) \end{aligned} \tag{1}$$

where x_j represents the j^{th} input variable, $\mathbb{E}$ denotes expectation, and

$$\begin{aligned} f_0 &= \mathbb{E}[y], \\ f_j(x_j) &= \mathbb{E}[y|x_j] - f_0, \\ f_{j,k}(x_j, x_k) &= \mathbb{E}[y|x_j, x_k] - f_0 - f_j - f_k, \\ &\cdots \end{aligned}$$

Further, using variance to measure uncertainty of the QoI, and assuming that $f(\mathbf{x})$ is square-integrable and

that each variable has finite variance, the variance of y from Equation 1 can be decomposed as:

$$Var(y) = \sum_{j=1}^{d} Var_{x_j}(\mathbb{E}_{\boldsymbol{x}_{\sim j}}[y|x_j]) + \left[\sum_{k=j+1}^{d} Var_{x_{jk}}(\mathbb{E}_{\boldsymbol{x}_{\sim jk}}[y|x_j, x_k]) - V_j - V_k\right] + \cdots \tag{2}$$

where $V_j := Var_{x_j}(\mathbb{E}_{\boldsymbol{x}_{\sim j}}[y|x_j])$ represents the variance contribution to y from input x_j alone. Thus, Sobol' indices provides a decomposition of the variance in y for each input variable as well as combinations of the input variables. First order Sobol' indices are computed as $S_j = \frac{V_j}{Var(y)}$ and can be extended to higher order indices as $S_{j,k,\ldots} = \frac{V_{j,k,\ldots}}{Var(y)}$. Due to the computational overhead of examining all possible subsets of features, Sobol' indices are not generally examined beyond the first, second, and total order indices; where total order indices quantify how much inputs or parameters contribute to the total variance on its own and through interactions with other inputs or parameters.

2.2. Shapley Values

From cooperative game theory, Shapley values proportion a global reward according to individual contributions in a team effort [24]. Let $\mathbb{S}$ be a subset of all input variables $\mathbb{M} := \{x_j\}_{j=1}^{d}$, and $v|_{\mathbb{S}}$ be a value function that approximates the QoI on the given subset of variables. Shapley values are defined as [26]:

$$\phi_j = \frac{1}{|\mathbb{M}|} \sum_{\mathbb{S} \subseteq \mathbb{M} \setminus \{x_j\}} \frac{|\mathbb{S}|!(|\mathbb{M}| - |\mathbb{S}|)! - 1}{|\mathbb{M}|!} \times \left(v|_{\mathbb{S} \cup \{x_j\}} - v|_{\mathbb{S}}\right), \forall \mathbb{S} \subset \mathbb{M} \tag{3}$$

In terms of feature importance, the impact each variable x_j is evaluated over all possible subsets $\mathbb{S}$. It is assumed that f is sufficiently complex such that a simpler surrogate model v is needed for computational feasibility. Shapley values have been used in GSA as a measure of variable importance [27, 28] and are theoretically bound by the first and total order Sobol' indices [29]. Similar to integrated gradients [30], Shapley values calculate the difference between the average and the actual output. The SHAP XAI method [31] is based on these values.

3. Trusting Learned Models

The need to trust learned models has been explored broadly in ML [32] and reinforcement learning [33], as well as for specific applications such as computer vision [34] and automotive software engineering [35]. However, most of these studies focus primarily on how robust ML models are to adversarial attacks or out-of-distribution data points. Additionally, with the increased usage of AI in many businesses, several maturity models have been put forward to assess if a learned model is ready to be deployed. Most of these focus on AI operations from a strategic and principled development point of view rather than on an examination of a learned model [36]. For example, guidelines from Ethical AI point out the need to provide explanations and transparency, but do not examine if the explanation reflects the underlying model [37]. As XAI methods have been proposed as a means of providing trust and verifying model behavior, we examine explanation fidelity from the perspective of GSA which have been used in V&V to ensure safety and increase trust in a model and argue that the fidelity of the explanations should also be used to establish credibility. Specifically, we examine how the four core design of experiments decisions from GSA (Section 2) are addressed in XAI.

3.1. Black Box Explanation Methods

There are several connections between GSA, specifically Sobol' indices and Shapely values, and perturbation feature importance methods. LIME [6] was one of the first methods to gain traction in explaining the predictions from an ML model as a means for building trust and examining that a model functions properly. The explanation is derived from a locally weighted linear regression model. To create this linear model, in its simplest form, the input space is randomly sampled and the sampled data points are passed through the model that is to be explained. Each sampled data point and its classification is weighted based on its distance from the data point to be explained. The linear model is then learned using a weighted linear regression algorithm. Explanations are then derived using the input values of the data point to be explained and the weights in the linear model.

SHAP [38] builds on cooperative game theory computing Shapley values (Equation 3), and appears to be the strongest theoretically grounded XAI method. SHAP calculates Shapley values using bootstrapping

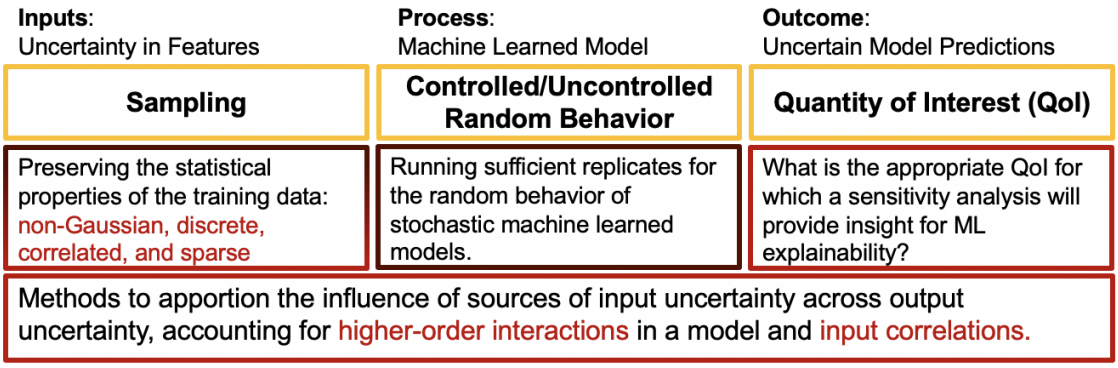


Figure 1. Mapping of GSA processes to XAI and highlighting current holes.

methods to replace a feature with noise to effectively remove it. To deal with high-dimensional data, SHAP is extended to Kernel SHAP where the subsets are weighted based on the weights of their contributions and also has several model specific implementations (e.g. TreeSHAP [39]).

3.2. Design of Experiments Decisions in XAI

As ML often deals with high dimensional data, computation feasibility is a major concern to XAI methods and impacts their design decisions. In this section we consider these decisions at a high-level. We then empirically analyze how they affect the fidelity of the explanations in the following sections.

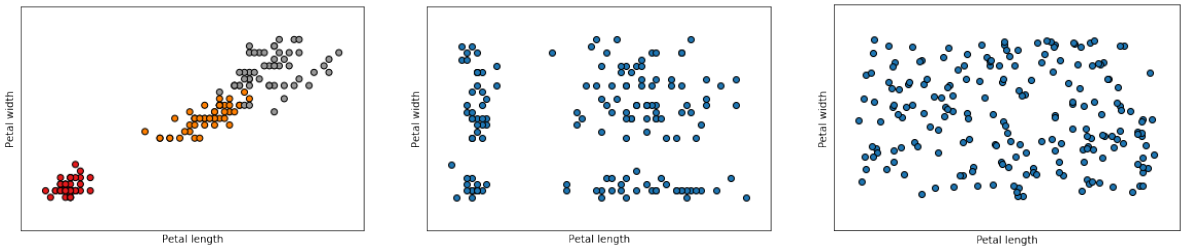
The overall process of GSA is mapped to the XAI paradigm in Figure 1 and is composed of four major design of experiment decisions. Examining XAI techniques through the lens of GSA, we have identified three primary areas where improvement could significantly enhance the fidelity of the explanations:

Sampling To make computation feasible, most XAI methods, including LIME and SHAP, assume mutual independence of input variables—ignoring all dependencies as part of their sampling strategy. Previously mentioned here and in other works [40], independence assumptions about the input cause incorrect feature importance values when correlations are present. The appeal of such a limiting assumption is motivated by avoiding the computational cost in modeling the full-joint probability distribution. There also exists a plethora of readily available expedient sampling techniques such as bootstrapping and independent random sampling that are commonly used. Proper care is rarely taking into consideration with regards to feasibility of the sampled data points resulting in these methods creating data points that are clearly out of distribution. For example, consider the third and fourth features (petal length and width) of the Iris dataset which are highly correlated to each other and with the class label. Figure 2 shows the original data points (Figure 2a) the bootstrapped sampled data points (Figure 2b) and random sampling (Figure 2c). Visually, it is clear that: (1) the original features are correlated, (2) the sampled data do not preserve the correlation between the data points and (3) therefore, the sampled data points are out-of-distribution from the training data. What does an explanation generated from these types of data points tell us about the model?

Future work needs to consider how to sample the data that preserves the statistical properties of the data. Sampling without care can introduce large amounts of standard error. Also, there is little consideration about the amount of replicates needed to properly model the input and output uncertainties.

Quantity of Interest For XAI, the output of the classifier or a confidence metric is often used, but is that really what is most important for explaining a prediction? Currently, model outputs, such as confidences, are often used. In XAI, this becomes challenging for classification problems as large amount of variances are sometimes needed to change the classification. Further, model confidence measures have been shown to be difficult to calibrate [41], they are often meaningless [42], and easily manipulated [43]. These results often do show which inputs are influential, but, as pointed out by Rudin [44], often just knowing *where* a model "looks" is not always sufficient for *why* a prediction was made. Future work should investigate other possibilities in QoIs that correlate with explanations.

Nonlinear Uncertainty Apportionment For computational expediency, current XAI methods assume a linear relationship between the output QoI and all input features—enabling the computation of expectations directly rather than from sampling [38]. For example, LIME uses a linear model to apportion



a b c

Figure 2. The results of sampling on the correlated petal width and petal length features from the iris data set showing a) the original dataset, b) the resampled dataset by perturbing only the petal width feature (bootstrapping), and c) the randomly sampled data.

the influence of the inputs on the outputs. Follow up work has proposed extending SHAP to model input dependencies (SHAP-Dep) [45] incurring an increase in computational complexity which severely limits its usability (the authors only examined up to 10 input features). We have found that the lack of ability to apportion the influences of input uncertainty across output uncertainty accounting for higher-order, nonlinear interaction in a model is the greatest challenge. This poses a significant hurdle to overcome as most state-of-the-art ML models are highly nonlinear and have high-order feature interactions within the model. Thus, not being able to model these interactions poses a significant hurdle to overcome in explaining ML models.

These issues are not isolated to XAI, but are also challenges in the GSA and UQ communities. Our hope is that techniques from GSA could be used to improve XAI. There are several promising research lines in GSA that could be leveraged in XAI including work from Sobol' indices that specifically address correlations [46] in certain situations and may be applicable for nonlinear interactions. Razavi et al. [22] provide a good overview of current research directions in GSA include perspectives from the GSA community on its use in ML. Appropriate QoIs could leverage recent advances in model output calibration [47] or leverage another metric which provides insights into the actual decision making process of the learned model, perhaps along the same lines as anchors [48] or the most influential data points such as prototypes and criticisms [49]. To our knowledge, addressing nonlinear uncertainty apportionment and appropriate QoIs for explainaiblity have not been explored in much depth.

4. What Constitutes the Fidelity of an Explanation?

In this section, we examine how the design decisions discussed above affect the fidelity of the explanation to the actual model. There are several measures that can be used to evaluate an explanation [50]. Here, we study *fidelity* which we define as the accuracy of the explanation to the underlying model. An explanation that has complete fidelity and completeness would describe the model in complete detail and would be the model itself. Therefore, to be useful to an end-user a trade-off exists between fidelity and completeness of an explanation to convey enough information to describe the decision process[1]. As noted earlier, reliance on human evaluation of fidelity may bias explanations towards persuasive explanations rather than accurately describing the learned model [9, 5]. Given this danger, our definition of fidelity is inspired by the need to determine what the learned model actually does to properly convey that to an end-user.

Here, we suggest definitions for an explanation and a mathematical notion of *fidelity* based on Shapley values. We denote the importance values estimated using XAI methods as $\hat{\Phi} i := \{\hat{\phi}_j\}_{j=1}^{d}$ by an explainability value function v. For consistency in notation, we denote the dependence of v and f on both $\mathbf{x}$ and the restriction to the subset of features $\mathbb{S}$ with the expressions $v|_{\mathbb{S}}(\mathbf{x})$ and $f|_{\mathbb{S}}(\mathbf{x})$. The validity of the explanation depends on how well $v|_{\mathbb{S}}$ approximates $f|_{\mathbb{S}}$ for all $\mathbb{S} \subseteq \mathbb{M}$. In practice, having a closed-form solution to Φ is not available or is restricted due to computational constraints, therefore $\hat{\Phi}$ is a numerical approximation of the analytical solution, Φ.

Definition 1 *An* explanation *is a subset* $\mathbb{S}^\star$ *of* $n \leq d$ *features with corresponding importance values* $\Phi_{\mathbb{S}^\star} :=$

[1]Note that this is a trade-off of current XAI methods. In the future XAI methods that have complete fidelity and completeness may be available.

$\{\phi_j\}_{x_j \in \mathbb{S}^\star}$ *from the set of all features* $\mathbb{M}$ *that have the greatest contributions to* $f(\mathbf{x})$*:*

$$\mathbb{S}^\star = \underset{\mathbb{S}\subseteq\mathbb{M},|\mathbb{S}|=n}{\arg\max} \sum_{x_j \in \mathbb{S}} \phi_j \tag{4}$$

where the ϕ_j correspond to Equation 3, such that the nominal feature values are defined by the realization of a data point $\mathbf{x}$.

Definition 2 *The* fidelity $\mathbb{F}_j$ of an explanation, $\mathbb{S}^\star$, of $\mathbf{x}$ for the j^{th} feature *is the complement to the absolute difference between the actual* ϕ_j *values and the estimated* $\hat{\phi}_j$ *values:*

$$\mathbb{F}_j(\mathbf{x}, v) = 1 - |\phi_j - \hat{\phi}_j|. \tag{5}$$

This definition assumes that the ϕ_j values are in the range of [0,1] which can easily be done by normalizing the values Φ. We subtract this quantity from one so that $\mathbb{F}_j$ near zero/one correspond to low/high fidelity. An aggregate score can also be obtained by summing the individual $\mathbb{F}_j$ scores, while care should be taken as the score could vary based on the number of features considered. This definition holds for a feature as well as a particular data point. The fidelity of an explanation may vary per feature and in different areas of the input space. Therefore, the fidelity $\mathbb{F}_j$ is dependent on how well $v|_{\mathbb{S}}(\mathbf{x})$ approximates $f|_S(\mathbf{x})$. Equation 5 is equivalently expressed as:

$$\mathbb{F}_j(\mathbf{x}, v, f) = 1 - \mid v|_{\mathbb{S}}(\mathbf{x}) - f|_S(\mathbf{x}) \mid \tag{6}$$

As defined in Equation 4, an explanation is a ranked feature list, therefore, v does not necessarily have to be equivalent to f to produce useful explanations, but it should be close enough to produce the same ordering and give an idea of the magnitude of each feature's importance. Therefore, while absolute difference can serve as an appropriate loss function, a metric such as the Kendall Tau metric could also measure the difference in the feature rankings. Practically, calculating S^* would require relearning $f|_S$ for all $\mathbb{S} \subseteq \mathbb{M}$ making it computationally infeasible for all but the most trivial problems, hence the need for and importance of v to accurately and efficiently approximate f. It should be understood how v differs from f and what uncertainty comes from the model-form error.

5. Empirical Examination of Explanation Fidelity

This section examines *the fidelity of explanations* for models with various properties. To establish ground truth explanations, we examine using closed-form analytical solution to Shapley values (Equation 3) on a synthetic model where we can manipulate the properties of the features and the model. We compare the global measures of LIME [6], SHAP [38], and SHAP-Dep [45] aggregating over the entire training set with: (1) permutation feature importance [51] on a random forest trained on data generated by the model using sklearn[2], (2) the GINI values from the same random forest model, (3-4) empirical and analytical Sobol indices (implemented in OpenTurns[3]), and (5-6) empirical Shapley values.

5.1. Synthetic Data

We use simple regression models with four input variables, with and without correlation, and with and without nonlinear feature interactions:

$$y = 2x_1 + 3x_2 + x_3 + x_4 \tag{7}$$

$$y = 2x_1 + 3x_2 + x_3x_4 \tag{8}$$

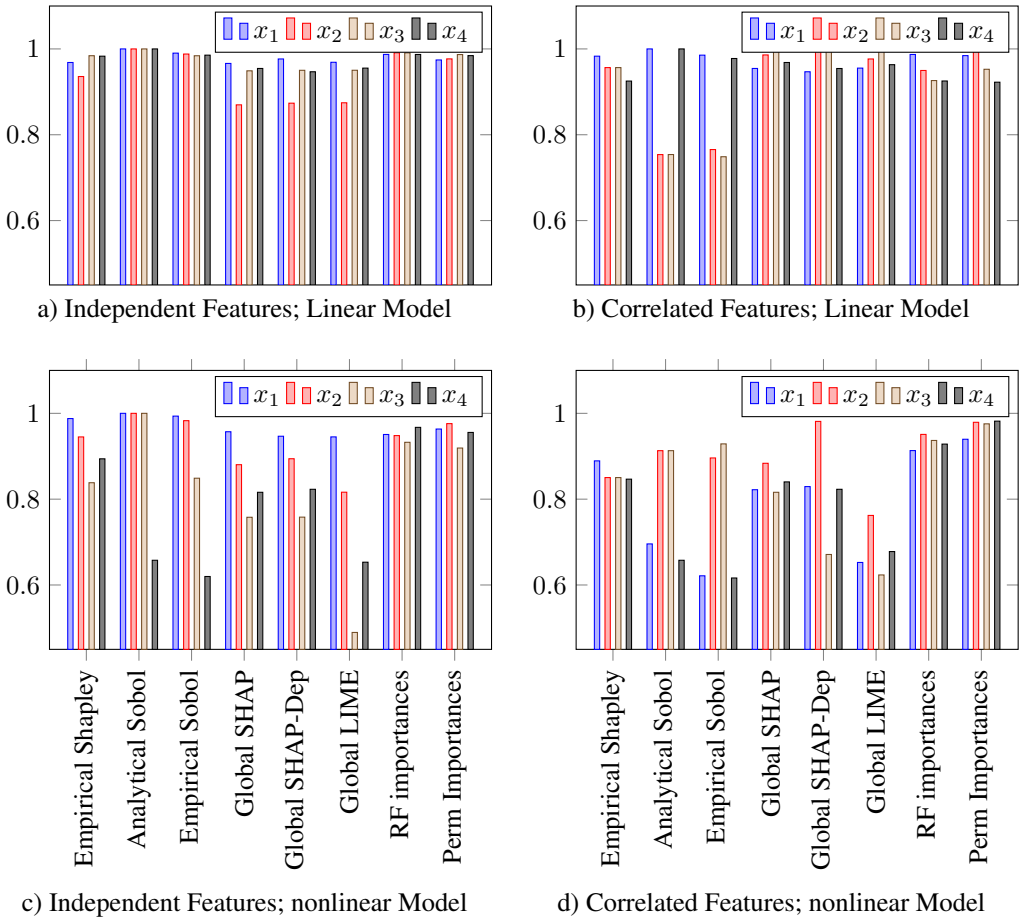
$$x_1 \sim \mathcal{N}(0,1), x_2 \sim \mathcal{N}(0,2), x_3 \sim \mathcal{N}(0,3), \quad x_4 \sim \mathcal{N}(0,4) \tag{9}$$

For correlation, we set x_1 and x_2 to be perfectly correlated as a worst-case scenario. We build on the traditional notion that greater weights and variance equate to a larger importance factor (in this case normalizing the data would negate this, but for demonstrative purposes we keep the different variances as variance is important for the GSA methods). Given the actual models and low number of features, we can calculate the closed-form solution to Shapley values considering *all* of the feature combinations. The explanation fidelity for each feature (Equation 5) is computed analytically using the Shapley value ϕ_j and the value from each XAI method as $\hat{\phi}_j$.

We create a dataset from each model with and without correlated features by sampling 1000 samples for each x_j and recording the resulting y from Equations 7 and 8. We then train a random regression forest on the data as our black box model (we tested several ML algorithms; the most important factor was the ability to model higher-order interactions, so models that make strong independence assumptions such as naïve Bayes had significantly different results). The fidelity of the explanations showing the differences between the Shapley values and the calculated importance for each feature are shown in Figure 3a-d for each combination of input independence/correlation and model linearity/nonlinearity. It is clear that while

[2]https://scikit-learn.org/

[3]https://pypi.org/project/openturns/



a) Independent Features; Linear Model

b) Correlated Features; Linear Model

c) Independent Features; nonlinear Model

d) Correlated Features; nonlinear Model

Figure 3. The fidelity of model explanations per feature. Here, a value of 1 is perfect fidelity. nonlinearity in the model has the greatest negative impact on fidelity

correlations have an impact on the fidelity, *nonlinear feature interactions produce even lower fidelity values*.

We make the following observations:

1. When the model and data meet the correct assumptions of feature independence and linearity (Figure 3a), the explanations best match the ground truth, although x_2 is consistently underestimated by LIME, SHAP and SHAP-Dep.

2. When only correlation is introduced (Figure 3b), the fidelity of each method degrades as expected—most significantly with the Sobol' methods.

3. Once nonlinear feature interactions are introduced (Figures 3c and d), the fidelity significantly decreases compared to feature correlation.

4. While SHAP-Dep is designed to work specifically with correlated features, we find that it shows lower model fidelity than SHAP in the presence of nonlinearity—this is probably partially due to the fact that we used treeSHAP as the underlying version of SHAP which is not available in SHAP-Dep. Additionally, the sampling complexity requirements are larger to model dependencies as demonstrated in the following section.

5. The tree-based importance methods consistently show high model fidelity and do exceptionally well in the case of correlated input features and nonlinear model interactions.

With these results, we strongly caution the use of SHAP, SHAP-Dep, LIME, and related approaches when the ML models include correlated features or nonlinear interactions. Of course, it should be noted that SHAP, SHAP-Dep and LIME are designed for local explanations for a specific data point. While the importance features from the random forest methods show the highest model fidelity, they only provide global feature importance—but do show robustness and

high fidelity explanations. One of the reasons why they perform exceptionally well is that they explicitly optimize the splits which separate the data. Studies examining both their benefits in identifying relevant features [52] and cautioning against possible bias [53] warrant a closer examination outside the scope of this paper. However, we will comment that random sampling and not using a linear dependence between the inputs and outputs may provide suggestions for improving other XAI methods.

6. Conclusion

While XAI has helped to provide insights into the learned models and offers some means of trust, we have demonstrated that strong assumptions that have been made for computational feasibility need to be considered when using XAI—especially in high-consequence applications and in models with nonlinear feature interactions. By casting XAI within the framework of GSA, we identified several holes and demonstrated that correlated input features and, more significantly, nonlinear model interactions have strong ramifications on the fidelity of XAI methods. Random forest models are surprisingly robust to these and may offer some paths for future work. New explainability methods are needed that are able to capture higher-order model interactions and can provide estimates of uncertainty. We only examined a small set of XAI methods. Future work could expand this notion given the ability to calculate a ground truth explanation value. Addressing these issues in neural networks will be even more difficult since the features of such networks are not well defined. Our motivation is that this analysis motivates others to pursue solutions to the gaps in XAI.

Acknowledgements

This article has been authored by an employee of National Technology & Engineering Solutions of Sandia,LLC under Contract No. DE-NA0003525 with the U.S. Department of Energy (DOE). The employee owns all right,title and interest in and to the article and is solely responsible for its contents. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this article or allow others to do so, for United States Government purposes. The DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan `https://www.energy.gov/downloads/doe-public-access-plan`.

References

[1] A. Shabtai, U. Kanonov, Y. Elovici, C. Glezer, and Y. Weiss, ""andromaly": a behavioral malware detection framework for android devices.," *Journal of Intelligent Information Systems*, vol. 38, no. 1, pp. 161–190, 2012.

[2] H. Etienne, "When AI ethics goes astray: A case study of autonomous vehicles," *Social Science Computer Review*, 2020.

[3] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, "Machine learning for medical imaging," *Radiographics*, vol. 37, no. 2, pp. 505–515, 2017.

[4] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[5] M. C. Stites, M. Nyre-Yu, B. Moss, C. Smutz, and M. R. Smith, "Sage advice? the impacts of explanations for machine learning models on human decision-making in spam detection," in *Proceedings of the 23rd International Conference on Human-Computer Interaction*, HCII '21, p. to appear, 2021.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?": Explaining the predictions of any classifier," in *Knowledge Discovery and Data Mining (KDD)*, 2016.

[7] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[8] I. Lage, E. Chen, J. He, M. Narayanan, B. Kim, S. Gershman, and F. Doshi-Velez, "An evaluation of the human-interpretability of explanation," *arXiv preprint arXiv:1902.00006*, 2019.

[9] B. Herman, "The promise and peril of human evaluation for model interpretability," *arXiv preprint arXiv:1711.07414*, p. 8, 2017.

[10] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," *Advances in Neural Information Processing Systems*, vol. 31, pp. 9505–9515, 2018.

[11] W. Nie, Y. Zhang, and A. Patel, "A theoretical explanation for perplexing behaviors of backpropagation-based visualizations," in *International Conference on Machine Learning*, pp. 3809–3818, PMLR, 2018.

[12] L. Sixt, M. Granz, and T. Landgraf, "When explanations lie: Why many modified bp attributions fail," in *International Conference on Machine Learning*, pp. 9046–9057, PMLR, 2020.

[13] R. C. Smith, *Uncertainty Quantification: Theory, Implementation, and Applications*. USA: Society for Industrial and Applied Mathematics, 2013.

[14] W. L. Oberkampf, M. Pilch, and T. G. Trucano, "Predictive capability maturity model for computational modeling and simulation," tech. rep., Sandia National Laboratories Albuquerque, NM, 2007.

[15] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola, *Global sensitivity analysis: the primer*. John Wiley, 2008.

[16] A. Gosiewska and P. Biecek, "Do not trust additive explanations," *arXiv preprint arXiv:1903.11420*, 2019.

[17] S. Lerman, C. Venuto, H. Kautz, and C. Xu, "Explaining local, global, and higher-order interactions in deep learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1224–1233, 2021.

[18] H. Chen, J. D. Janizek, S. Lundberg, and S.-I. Lee, "True to the model or true to the data?," *arXiv preprint arXiv:2006.16234*, 2020.

[19] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler, "Problems with shapley-value-based explanations as feature importance measures," in *International Conference on Machine Learning*, pp. 5491–5500, PMLR, 2020.

[20] C. Frye, D. de Mijolla, T. Begley, L. Cowton, M. Stanley, and I. Feige, "Shapley explainability on the data manifold," in *International Conference on Learning Representations*, 2020.

[21] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," *arXiv preprint arXiv:1806.08049*, 2018.

[22] S. Razavi, A. Jakeman, A. Saltelli, C. Prieur, B. Iooss, E. Borgonovo, E. Plischke, S. Lo Piano, T. Iwanaga, W. Becker, S. Tarantola, J. H. Guillaume, J. Jakeman, H. Gupta, N. Melillo, G. Rabitti, V. Chabridon, Q. Duan, X. Sun, S. Smith, R. Sheikholeslami, N. Hosseini, M. Asadzadeh, A. Puy, S. Kucherenko, and H. R. Maier, "The future of sensitivity analysis: An essential discipline for systems modeling and policy support," *Environmental Modelling & Software*, vol. 137, p. 104954, 2021.

[23] I. M. Sobol, "Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates," *Mathematics and computers in simulation*, vol. 55, no. 1-3, pp. 271–280, 2001.

[24] L. S. Shapley, "A value for n-person games," *Contributions to the Theory of Games*, vol. 2, no. 28, pp. 307–317, 1953.

[25] D. C. Montgomery, *Design and analysis of experiments*. John wiley & sons, 2017.

[26] E. Winter *et al.*, "The Shapley value," *Handbook of game theory with economic applications*, vol. 3, no. 2, pp. 2025–2054, 2002.

[27] A. B. Owen, "Sobol' indices and Shapley value," *SIAM/ASA Journal on Uncertainty Quantification*, vol. 2, no. 1, pp. 245–251, 2014.

[28] S. Lipovetsky and M. Conklin, "Analysis of regression in game theory approach," *Applied Stochastic Models in Business and Industry*, vol. 17, no. 4, pp. 319–330, 2001.

[29] B. Iooss and C. Prieur, "Shapley effects for sensitivity analysis with correlated inputs: Comparisons with sobol' indices, numerical estimation and applications," *International Journal for Uncertainty Quantification*, vol. 9, no. 5, 2019.

[30] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International Conference on Machine Learning*, pp. 3319–3328, PMLR, 2017.

[31] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *arXiv:1705.07874 [cs, stat]*, May 2017. arXiv: 1705.07874.

[32] I. Goodfellow and N. Papernot, "The challenge of verification and testing of machine learning," *Cleverhans-blog*, 2017.

[33] P. Van Wesel and A. E. Goodloe, "Challenges in the verification of reinforcement learning algorithms," *National Aeronautics and Space Administration, NASA STI Program*, 2017.

[34] V. Buhrmester, D. Münch, and M. Arens, "Analysis of explainers of black box deep neural networks for computer vision: A survey," *arXiv preprint arXiv:1911.12116*, 2019.

[35] M. Borg, C. Englund, K. Wnuk, B. Duran, C. Levandowski, S. Gao, Y. Tan, H. Kaijser, H. Lönn, and J. Törnqvist, "Safely entering the deep: A review of verification and validation for machine learning and a challenge elicitation in the automotive industry," *arXiv preprint arXiv:1812.05389*, 2018.

[36] K. R. C. Salveson, "The ai maturity framework," Tech. Rep. MSU-CSE-06-2, Element AI, 2020.

[37] The Institute for Ethical AI & Machine Learning, "The responsible machine learning principles." `https://ethical.institute/principles.html`. Accessed: 2020-04-12.

[38] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, pp. 4765–4774, 2017.

[39] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature machine intelligence*, vol. 2, no. 1, pp. 56–67, 2020.

[40] L. Toloşi and T. Lengauer, "Classification with correlated features: unreliability of feature ranking and solutions," *Bioinformatics*, vol. 27, no. 14, pp. 1986–1994, 2011.

[41] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.), vol. 70, pp. 1321–1330, PMLR, 06–11 Aug 2017.

[42] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," *Proceedings of the International Conference on Learning Representations*, 2019.

[43] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 427–436, 2015.

[44] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, pp. 206–215, 2019.

[45] K. Aas, M. Jullum, and A. Løland, "Explaining individual predictions when features are dependent: More accurate approximations to shapley values," *arXiv preprint arXiv:1903.10464*, 2019.

[46] J. Hart and P. A. Gremaud, "An approximation theoretic perspective of sobol'indices with dependent variables," *International Journal for Uncertainty Quantification*, vol. 8, no. 6, 2018.

[47] M. P. Naeini, G. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using bayesian binning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, 2015.

[48] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[49] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! criticism for interpretability," *Advances in neural information processing systems*, vol. 29, 2016.

[50] S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable ai systems," *arXiv preprint arXiv:1811.11839*, 2018.

[51] L. Breiman, "Random forests," *Machine Learning*, vol. 45, p. 5–32, Oct. 2001.

[52] B. H. Menze, B. M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, and F. A. Hamprecht, "A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data," *BMC bioinformatics*, vol. 10, no. 1, pp. 1–16, 2009.

[53] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC bioinformatics*, vol. 8, no. 1, pp. 1–21, 2007.