Towards Rigorous Interpretations: a Formalisation of Feature Attribution

Darius Afchar [1 2] **Romain Hennequin** [1] **Vincent Guigue** [2]

Abstract

Feature attribution is often loosely presented as the process of selecting a subset of relevant features as a rationale of a prediction. Task-dependent by nature, precise definitions of "relevance" encountered in the literature are however not always consistent. This lack of clarity stems from the fact that we usually do not have access to any notion of ground-truth attribution and from a more general debate on what good interpretations are. In this paper we propose to formalise feature selection/attribution based on the concept of relaxed functional dependence. In particular, we extend our notions to the instance-wise setting and derive necessary properties for candidate selection solutions, while leaving room for task-dependence. By computing ground-truth attributions on synthetic datasets, we evaluate many state-of-the-art attribution methods and show that, even when optimised, some fail to verify the proposed properties and provide wrong solutions.

1. Introduction

As the adoption of intelligent algorithms of growing complexity is becoming ubiquitous in our everyday lives, concerns have consequently emerged about the lack of transparency and need for interpretability of these methods (Parliament, 2016). Interpretability is unfortunately somewhat ill-defined and ill-evaluated (Doshi-Velez and Kim, 2017; Lipton, 2018), partly because a wide range of concepts are encompassed under the same label. One can think of the protean purposes of interpretations: *informativeness*, *causality*, *fairness*, *interactivity*, *trust*, etc (Tintarev and Masthoff, 2007; Arrieta et al., 2020). Nonetheless, there is a consensus on the fact that interpretations stem from a notion of *incompleteness* and aim at boosting *human understandability*. But viewing understandability in an holistic manner requires controlling and disentangling every aspect of a model prediction, ranging from how the data is inherently structured, to what priors are induced by a model architecture, to what impact can a design choice to present explanations have on a target audience and in a particular setting.

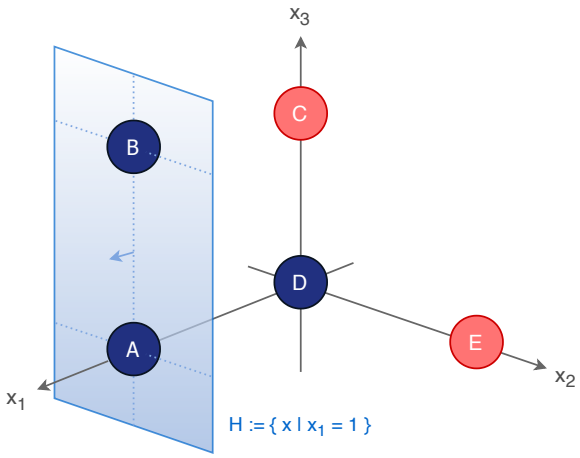

Figure 1. **Intuition of the formalisation** Example prediction task with five points $p_A, ..., p_E$ in $\mathbb{R}^3$ and binary labels blue/red. Note that, to correctly label point p_A as *blue*, it is sufficient to know that the point has its coordinate $(p_A)_1 = 1$. Indeed, the incomplete view that $x_1 = 1$, may lead to confuse points p_A and p_B, but not the determination of their label (*blue*). We say that p_A functionally depends on X_1. Since all points in H symmetrically have the same label, they also share the same dependence in X_1; by comparison, this does not hold for other points at $x_1 = 0$. We will see that this symmetry argument is a necessary property for any instance-wise feature selection candidate solution.

This may explain why many methods have resorted to proxy measures of interpretability and have proposed list of general requirements for interpretations - *e.g.* (Ribeiro et al., 2016; Lundberg and Lee, 2017; Sundararajan et al., 2017), that are sometimes confirmed by user-studies: *e.g.* sparsity is widely considered a general desiderata of interpretation. This process is not always successful (Rudin, 2019). In fact, many recent works (Adebayo et al., 2018; Serrano and Smith, 2019; Kindermans et al., 2019; Dombrowski et al., 2019; Sixt et al., 2020; Kumar et al., 2020) tend to suggest that well-established interpretation methods may

[1]Deezer Research, Paris, France [2]LIP6, Paris, France. Correspondence to: Darius Afchar <research@deezer.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

not provide much understandability after further inspection, while being coherent with their self-defined interpretation criteria. Additionally, Kaur et al. (2020) showed that several popular methods may be misused by practitioners with lack of consideration for methods' assumptions relevance or requirements or application domain, and thus be prone to confirmation biases. Such blunders are not new in the field of interpretable machine learning, which is why Doshi-Velez and Kim (2017) had advocated for rigorous formalisation so as to avoid any subjective definition, vague evaluations and practitioners misuses, such as what had already been done in the subfields of fairness or privacy.

In this paper we propose to formalise a popular class of interpretation methods that we find lacks clarity: **feature attribution**. Feature attribution/importance aims at providing a rationale for the association of target values to input instances; where target values may correspond to a model's predictions – enabling the inspection of its behaviour, or observed true labels – to interpret data. To do that, all attributions tasks can be decomposed into two subproblems: **(1)** providing a scoring function that represents the *responsibility* of a feature or group of features in the association to a given value, then **(2)** returning a parsimonious subset of features as a rationale of the association, using the scores. The rationale can either apply to all instances – *global attribution*, allowing to discard noisy and redundant features (Tibshirani, 1996), or be computed locally – *instance-wise attribution*. The concept of *responsibility* is however task-dependent and varies widely between methods, and the relevance of the returned minimal features is sometimes ill-evaluated, if evaluated at all. In particular in the instance-wise setting, and unlike global attribution, we will show that checking prediction performances from selected features is not sufficient to ensure that the correct rationale was found.

That said, ground-truth knowledge of input responsibility is not usually available in any form for collected data. Furthermore, evaluations on real data often come with the hardship of disentangling interpretation errors from prediction errors (Dinu et al., 2020).

That is why we propose to study in detail an informed scenario, for which we know everything about the input distribution p_X and target distribution $p_{Y|X}$. Specifically, we generate synthetic supervised tasks and abstract models from the task by replacing them with optimal distributions or mappings[1]. Doing so, we are able to derive ground-truth rationales and critically assess the interpretation capabilities of many attribution methods. Our vision is that if a method fails at providing relevant attributions given this ideal and noise-controlled distribution of the data, this should be worrisome for real-world applications.

[1]For instance, $\mathbb{E}_{Y|X}[Y|X]$ for regression tasks with normal priors and $\arg\max_c p_{Y|X}(y = c \mid X)$ for categorical tasks.

Our contributions are the following:

1. We propose a formalisation of selection and attribution based on functional dependence and derive necessary properties to extend them to the instance-wise setting;
2. We rigorously evaluate feature selections of many state-of-the-art methods on generated data and show that only a few of them achieve satisfying performances;
3. We show that our proposed necessary properties allow to evaluate estimated selections quality without having access to ground-truth solutions.

2. Feature attribution formalisation

We start by defining some notations. As mentioned, we study a supervised prediction interpretation setting: let us denote by $x \in \mathcal{X}$ an input sample and $y \in \mathcal{Y}$ its associated label or continuous value. We suppose $\mathcal{X} \subseteq \mathbb{R}^n$ and denote $[n] = \{1, ...n\}$ the set of input indexes. The attribution problem is that for a given sample x and for all subsets $I \subset [n]$, we first want to estimate a value $\text{attr}_I(x)$ that represents the *responsibility* of $(x_k)_{k\in I}$ in the observed association of x to y, and then return a minimal responsible subset using all the values $(\text{attr}_J(x))_{J\subset[n]}$.

The issue is that responsibility, sometimes referred to as *relevance* or *importance*, is ill-defined. There are however two principles that are shared across all attribution methods that will guide us in our formalisation. First, since interpretations depend on their application field and target audience, **responsibility is task-specific (*P1*)**. For instance, it is sometimes relevant to have a notion of negative responsibility – *e.g.* in sentiment prediction tasks to find words that flip the meaning of a sentence; and sometimes not – *e.g.* for a recommender system using an implicit feedback dataset where negative interactions are not meaningful (Hu et al., 2008). The second principle lies in the binary distinction between *null* and *non-null* responsibilities: a null value indicates a subset of variables that has nothing to do with the association of x to y; a non-null one does, to some task-specific extent. **Responsibilities should enable to distinguish contributing and non-contributing features (*P2*)**. Splitting input features into a minimal subset of contributing features versus non-contributing others is called the **feature selection problem** (Natarajan, 1995; Blum and Langley, 1997). We argue that selection should always be implied by attribution, and by contraposition, that an attribution method that does not allow to return a correct selection solution should be questioned.

In the rest of the section, we first formalise the notion of *contributing subset of features* from (*P2*) using the concept of functional dependence. In particular, we will extend

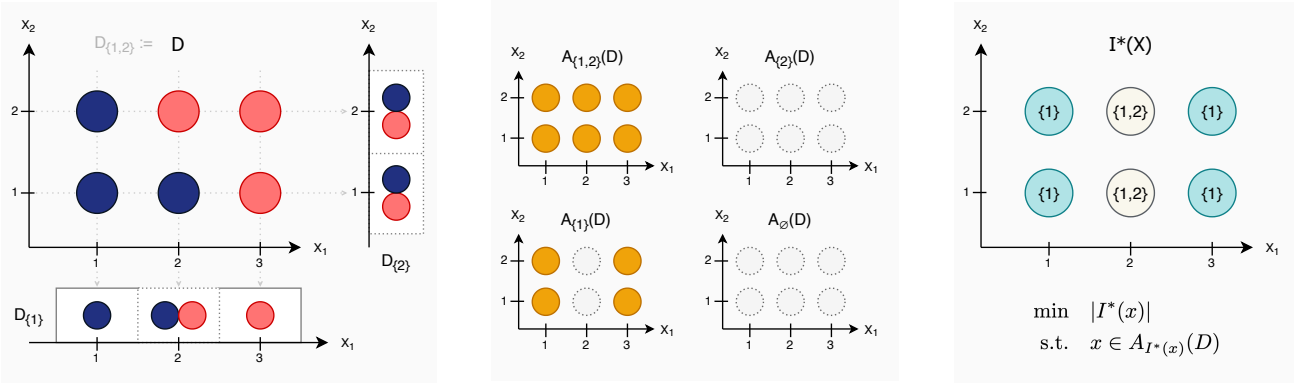

Figure 2. **Example of instance-wise selection derivation** From left to right: we are given a relation D from $\mathcal{X} = [3] \times [2]$ to $\mathcal{Y} = \{\text{blue}, \text{red}\}$, for simplicity, we assume that it defines unique associations (*ie.* D is a function), we compute its associated projected relations D_I; then its functionality domains $A_I(D)$; and finally derive the instance-wise selection solution $I^*(x)$.

our notion to the instance-wise setting which lacks formalism. Then, for (*P1*), we propose to see *responsibility* as its probabilistic relaxation and derive task-specific examples.

Now, to properly define what contribution means, we come back to the definition of a function and answer the following question: *what does it mean for a function to depend or not on a set of variables?*

2.1. Background on functionality

In set theory, the notion of function is built from the concept of *binary relation*. We adapt the definition and notation from Hamilton (1982).

Definition 1 (Binary relation)**.** *A binary relation from a set $\mathcal{X}$ to a set $\mathcal{Y}$ is a subset of the Cartesian product of the two sets. If R is such a relation and $(x, y) \in R$, we say that x is related to y and for convenience we may write xRy.*

What differentiates a relation from a function is that multiple outcomes can be in the image of a single input element of a relation. The second difference is that some points from $\mathcal{X}$ may not have been related to any point in $\mathcal{Y}$. Hence the following definition:

Definition 2 (Function)**.** *A partial function f is a binary relation that is single-valued. For all $x \in \mathcal{X}$, $y, z \in \mathcal{Y}^2$:*

$$((x, y) \in f) \ \wedge \ ((x, z) \in f) \ \Rightarrow \ y = z$$

To obtain a function, we additionally require this partial function f to be left-total:

$$\forall x \in \mathcal{X}, \ \exists y \in \mathcal{Y}, \ (x, y) \in f$$

When these two conditions are met, we can write the familiar expression $f(x)$ that denotes for all points of $\mathcal{X}$ the existing and unique element $y \in \mathcal{Y}$ such that xfy.

These two definitions are the starting point of our formalisation. We consider a given dataset of samples and associated labels. We want to express it as a dependence between the given input and associated labels. By definition, a dataset induces a binary relation between an input set $\mathcal{X}$ and a target set $\mathcal{Y}$ (continuous or discrete, it does not matter at this point). We denote it D. Without loss of generality, we assume D to be left-total, or reduce $\mathcal{X}$ accordingly. The single-value condition in definition 2 tells us when it is possible or not to uniquely assign a target label/value to a point in space, hence creating a functional dependence, *i.e.* given a dataset, this point is always related to a specific target, it *implies* it. For a given binary relation R, we define the subset of the domain $\mathcal{X}$ such that this condition is met:

$$A(R) = \{x \in \mathcal{X} \mid \forall y, z \in \mathcal{Y}^2, \ xRy \wedge xRz \Rightarrow y = z\}$$

By construction, our dataset D with its domain restricted to $A(D)$ is a function, meaning that points of D are *uniquely associated* on $A(D)$. By contrast, all points in $\bar{A}(D)$ are such that multiple target labels are related to a single input, the information given by the sole point position is intrinsically not sufficient to predict or assign a label. This is aside from the probabilistic considerations we will have in 2.4.

2.2. Subset functionality and selection

For selection though, we are interested in *defining dependence to only a subset among the input dimensions*. To do that, we first ignore some features by setting them to zero. There, it is convenient to use canonical projections. We denote $\mathcal{X} \subseteq \mathbb{R}^n$, $(\vec{e_1}, ...\vec{e_n})$ the canonical base of $\mathbb{R}^n$ and for a subset of indices $I \subset [n]$ the canonical projection on those indices $P_I(x) = \text{proj}(x, \{\vec{e_k} \mid k \in I\})$, and then define for a relation R, its relation with projected domain R_I:

$$R_I = (P_I \times \text{Id}_{\mathcal{Y}})(R)$$

For our dataset D, D_I is the dataset such that all input features with indices not in I are set to zero[2]. As a result, multiple points in the domain of D, with potentially different labels, may be collapsed into a single representative in D_I, thus killing the functionality property they may have verified in $A(D)$. On the point-wise level, the construction of a projected relation R_I implies that if xRy then $(P_I x)R_I y$, and reciprocally if $x_I R_I y$, there exists an antecedent x such that xRy and $P_I x = x_I$. We refer to figure 2 for a simple example to reason about the different concepts introduced in this section.

We now extend the previous definition of functional domain A to the case where we only consider subsets of features.

Definition 3. *For a given relation $R \subset \mathcal{X} \times \mathcal{Y}$, a subset of indices $I \subset [n]$ and R_I its projection to I, $A_I(R) \subset \mathcal{X}$ is the subset such that for all $x \in \mathcal{X}$, $y, y' \in \mathcal{Y}$, $x_I = P_I x$,*

$$x \in A_I(R) \Leftrightarrow (x_I R_I y \wedge x_I R_I y' \Rightarrow y = y')$$

Or equivalently, x is in $A_I(R)$ if and only if

$$\forall x' \in \mathcal{X} \text{ s.t. } P_I x' = P_I x,\ xRy \wedge x'Ry' \Rightarrow y = y'$$

Proof. $(P_I x' = P_I x = x_I) \wedge (xRy) \wedge (x'Ry') \Leftrightarrow (x_I R_I y) \wedge (x_I R_I y')$ □

By construction, D_I with its domain restricted to $P_I(A_I(D))$ is a function. Or said differently, for a given subset of indices I, for all points $x \in A_I(D)$, a target label y can be uniquely associated to x by the mere knowledge of its subset I of features. By comparison to definition 2, the only added condition is that the single-valueness must be verified not only by x but also all points with the same projection as x on I.

Now, once we have computed all 2^n domain subsets $A_I(D)$, the selection problem is formulated as the task of finding minimal subsets of input indices that all points functionally depend on. Which leads to two possible settings:

Problem 1 (Global subset selection)**.** *Given a relation R, find a subset of indices $I^* \subset [n]$ that minimises*

$$\min_{J \subset [n]} \textit{Card}(J)$$
$$\textit{s.t.} \quad \forall x,\ x \in A_J(R)$$

Problem 2 (Instance-wise subset selection)**.** *Given a relation R, for all $x \in \mathcal{X}$, find a local subset of indices $I^*(x) \subset [n]$ that minimises*

$$\min_{J \subset [n]} \textit{Card}(J)$$
$$\textit{s.t.} \quad x \in A_J(R)$$

[2]Because we know the subset I we project on, we can distinguish a *data* zero in I from the *ignoring* zeros of $\bar{I}$.

Note that it is not assured that these minima are unique, which is not problematic and rather natural, for instance when some input features are correlated.

Our derived definition of dependence/contribution and *global* selection coincides with Blum and Langley (1997). In the rest on the paper, we study its *instance-wise* extension, for it is the most difficult case with the largest risk of providing degenerate explanations if not done carefully.

2.3. Necessary properties of instance-wise dependence

The above definitions allow us to derive properties a given instance-wise selection solution $\hat{I}(x)$ should verify.

Property 1 (Complementary dependence)**.** *If a point depends on a subset of indices, all point in directions in the complement of this subset have the same dependence : for $x \in \mathcal{X}$, if there exists $I \subset [n]$ such that $x \in A_I(R)$, then for all $x' \in \mathcal{X}$ such that $P_I x' = P_I x$, one has $x' \in A_I(R)$.*

Proof. $[(P_I(x'), y') \in R_I] \wedge [(P_I(x'), y'') \in R_I] = [(P_I(x), y') \in R_I] \wedge [(P_I(x), y'') \in R_I] \Rightarrow y' = y''$ □

This property is illustrated in figure 1. We will see in the experiment section 4 that this property is not verified by some widely used attribution methods.

Property 2 (Dependence hierarchy)**.** *Any point that depends on a subset also depends on its parent subsets : $I \subset J \Rightarrow A_I(R) \subset A_J(R)$.*

Proof. $R_I = ((P_I \times \mathrm{Id}_{\mathcal{Y}}) \circ (P_J \times \mathrm{Id}_{\mathcal{Y}}))(R)$, thus $(P_J x R_J y) \wedge (P_J x' R_J y') \Rightarrow (P_I x R_I y) \wedge (P_I x' R_I y') \Rightarrow y = y'$ □

2.4. Attribution as relaxed functional dependence

We have formalised the notion of binary feature contributions for the selection task in quite an unrealistic case where we could find a perfect dependence. We now propose to *frame attribution values as its probabilistic relaxation*. Indeed, there are several reasons we may want to adopt a probabilistic framework and relax functional dependence:

- Real-data is noisy, we only have access to a sample of it, and may wish to control a certainty of dependence;
- For continuous $\mathcal{Y}$, we may tolerate having several outcomes for $x \in \bar{A}_I(R)$ but that are close to one another; and for categorical $\mathcal{Y}$, a small stochasticity of label;
- Generally, we want to accurately model probable associations of input and target label/values while minimising the weight of rare and out-of-distribution points.

Instead of a dataset D, we now consider probabilistic densities p_X and $p_{Y|X}$ on $\mathcal{X}$ and $\mathcal{Y}$ with their usual associated input and target random variables X and Y. We relax our notion to *approximate functional dependence*. At that point, we have to consider task-dependency as there is no one-relaxation-fits-all rule (*P1*). Attribution values should however still allow to differentiate between relevant and non-relevant subset of features to be meaningful (*P2*). As a general framework, we first define an attribution relaxation $\text{attr}_I(x)$ for all subsets $I \subset [n]$ and all samples $x \sim X$, and we then create the link to selection with a comparison to a chosen threshold parameter η. For instance, we could choose that all subsets of features with absolute attribution value higher than η should be selected. We can not define an encompassing comparison mechanism, the implication mechanism from attribution to selection is part of the relaxation elaboration and directly translates the meaning of the degree of approximation we choose with η. We give some examples of attribution relaxation to clarify this framework.

Regression setting Let Y be continuous, *e.g.* $\mathcal{Y} = \mathbb{R}$, and the function we want to interpret be the mean mapping $f(x) = \mathbb{E}[Y \mid X = x]$. To define an instance-wise responsibility measure g_I that will imply functional dependence on I, we can use the conditional variance:

$$g_I(x) = \text{Var}_{X|X_I}[Y \mid X_I = P_I(x) = x_I] \tag{1}$$

where X_I denotes the projected random variable $P_I(X)$. We verify that $g_I(x) = 0$ if and only if for all samples (x', y') such that $P_I x' = x_I$, the associated value y' is equal to the conditional mean $\mathbb{E}_{X_{\bar{I}}|X_I}[Y \mid X_I = x_I]$, hence verifying $x \in A_I(f)$ and thus (*P2*) in the perfect setting.

In the literature, it is more usual that attribution values near zero denote independence to a subset. To do that, we could use the reciprocal notion of *precision*: $\text{attr}_I(x) = 1/g_I(x)$. When the precision is low, the samples with common features on the indices I are spread, it is thus not possible to assign a value that will be representative enough of these points. When precision is high, the mean value will be a relevant predictor of the points, we can state that we have a dependence to I *with a given precision/variance*.

With this measure and for a given variance threshold η, we have obtained **approximated functionality domains**:

$$A_I^\eta(f) = \{x \in \mathcal{X} \mid |\text{attr}_I(x)| \geq 1/\eta\} \tag{2}$$

again, we verify that $A_I^0(f) = A_I(f)$ (*P2*).

To fix ideas through a simple application example, let us consider a bidimensional uniform input $X = (X_1, X_2)$ on $\mathcal{X} = [-1, 1]^2$, and Y such that,

$$p_X = p_{X_1} p_{X_2} = 1/4$$
$$Y = X_1 + \alpha X_2, \quad |\alpha| < 1$$

which corresponds to a deterministic identity mapping from X_1 to Y with a small tilt effect from X_2 with coefficient α. Then, for all $x_1 \in [-1, 1]$,

$$\text{Var}_{X_2|X_1}[Y \mid X_1 = x_1] = \frac{1}{2}\int_{-1}^{1} (\alpha t)^2 dt = \alpha^2/3$$

for a given variance threshold η, the attribution measure (1) states that if $\alpha \leq \sqrt{3\eta}$, the target variable Y can be approximated with $Y' = X_1$, i.e. X_2 is ignored and only X_1 is responsible for Y.

Similarly, let us have $Y = X_1 + \epsilon$, with ϵ a noise variable following $\mathcal{N}(0, \sigma^2)$. Because of the noise, there is no region of the domain where samples of Y can be uniquely determined on a set of variables. But when $\sigma^2 \leq \eta$, the noise can be ignored and this distribution can be approximated by the univariate distribution of $Y' = X_1$ with variance η.

With the attribution measure (1), we were able to relax dependence to a probabilistic framework allowing to control noise and small feature effects, and yield approximate feature contribution. Our choice of relaxation through the conditional variance works well when Y is assumed to follow a normal law $\mathcal{N}(\mu(X), \sigma(X)^2)$. This is of course not the only possible attribution measure, in particular if we want to study more than the mean effects f we chose.

Classification setting When Y takes values in the set of n labels $c_1, \dots c_n$. It seems natural to define an attribution measure as the probability of assigning the label with maximum probability.

$$P_I^c(x) = \mathbb{P}(Y = c \mid X_I = P_I x)$$
$$\text{attr}_I(x) = \max_c P_I^c(x) \tag{3}$$
$$A_I^\eta(f) = \{x \in \mathcal{X} \mid \text{attr}_I(x) \geq 1 - \eta\} \tag{4}$$

The attribution value is bounded between $1/n$ (uniform) and 1 (deterministic label). These responsibilities have a nice interpretation since they directly represents the proportion of samples in the same class when conditioning on the variables in I. Adjusting η also means that we control the error on the prediction of a class for these samples.

In the perfect setting, for $\eta = 0$ we check that $x \in A_I^0(f) \Rightarrow \text{attr}_I(x) = 1 \Rightarrow x \in A_I(f)$, thus (*P2*). In the imperfect setting, our function f under study is noisy, **the goal is to tune η to maximise the verification of (*P2*)**, which we will evaluate in section 4.

Alternatively, it may be more relevant to take in consideration all labels probabilities with an entropy measure:

$$\text{attr}_I(x) = 1 - \sum_c \frac{P_I^c \ln(P_I^c)}{\ln(1/n)} \tag{5}$$

We have normalised the entropy to obtain a value between 0 (uniform label distribution) and 1 (deterministic label).

3. Related methods

We present classic and state-of-the-art selection/attribution methods in the light of the formalism we propose, and with a specific focus on instance-wise methods. Due to size constraints, it is impossible to present all variations of assumptions and clever solutions of these methods, we will thus only present four general ideas that, we think, constitute the bulk of research on instance-wise feature attribution.

3.1. Mixture of restricted experts

The first thing we have to mention is that the attribution relaxation (1) we introduced in the context of regression is strongly inspired by the success of the classical *analysis of variance* diagnostics and its more recent formulation of *weighted functional ANOVA* (Hooker, 2007) that decomposes $\mathcal{L}_2$ functions into the sum of all n-variate subfunctions under a hierarchical orthogonality constraint, weighted by the data distribution. Given that one takeaway of our paper will be that we have to consider the full input distribution for relevant interpretations, not just local information, we should have been happy with weighted fANOVA. Specifically, one key consequence of fANOVA is that the overall variance can be decomposed as a sum of variance from each subfunction, and hence each input subset. However, this decomposition is made identifiable through an *integration-to-zero* constraint on the subfunctions, allowing to formulate global selection criteria but not to distinguish the non-null instance-wise contributions we seek from any centering effects (see Supplementary A).

Another idea, similar in spirit to fANOVA, is to try to directly learn a mixture of n-variate functions. Since there is a potential exponential number of subfunctions, one approximation making training tractable is to consider only summed univariate contributions – *e.g. GAM* (Hastie and Tibshirani, 1990); or interactions up to a fixed order – *e.g. GA^{2}M* (Lou et al., 2013), *NIT* (Tsang et al., 2018); or with a fixed structure – *e.g. Archipelago* (Tsang et al., 2020), *InterpretableNN* (Afchar and Hennequin, 2020). The key advantage of mixture models is that they disentangle the different orders of interaction effects. In our formulation of dependence, no distinction can for instance be made between $f(x) = x_1 + x_2$ and $f(x) = x_1 x_2$ with a uniform input distribution. This may be useful in some applications. But conversely, and beyond the trivial limitation that these models provide solutions within a restricted candidate space, additive models strongly suffer from an identifiability issue and can produce contradictory interpretations. Identifiability can be achieved with fANOVA-like regularisation (Lengerich et al., 2020), but we have argued that this does not allow to obtain exact attribution in an instance-wise setting. This effect gets worst with high-order interactions and redundant or correlated features. Meanwhile, our attribution formalisation allows to distinguish multiple possible candidate solutions, hence isolating redundancies, but at the cost of interaction hierarchical decomposability. We may assert that both approaches are complementary.

3.2. Proxy models

A large body of work on instance-wise attribution circumvents the above tractability issue by providing proxy measurements of attribution. Two large class of methods are **gradient-based** analysis – *e.g.* saliency methods (Simonyan et al., 2014), *SmoothGrad* (Smilkov et al., 2017), ... ; and **baseline-comparison** methods – *e.g. LIME* (Ribeiro et al., 2016), *SHAP* (Lundberg and Lee, 2017), ... ; the line between these two classes is fuzzy – *e.g. Integrated Gradient* (Sundararajan et al., 2017), *Expected Gradient* (Erion et al., 2019). Again, we will not discuss the profusion of variations but only their general spirit. For good meta-analysis on a unification of these methods, we recommend (Covert et al., 2020) and (Sundararajan and Najmi, 2020). Nevertheless, the underlying principle behind the computation of a gradient as an indication of feature contribution can be found in its simplest form in Friedman and Popescu (2008). In substance, it says that *a function $F(x)$ is said to exhibit an interaction between k variables with indexes $I = (i_1, \dots i_k)$ if* $\mathbb{E}_X[\partial^k F / \partial x_{i_1} \dots \partial x_{i_k}]^2 > 0$, meaning that the difference in value of $F(x)$ as a result of changing some variables of I depends on the remaining variables of I. Beyond noise considerations that may create nuisance interactions, this approach is rather sound for global selection. Problems occur in its extension to the instance-wise setting when $\mathbb{E}_X$ is dropped without any further considerations. This is the foundation of saliency methods and subsequent papers have focused on providing gradient estimates that proved robust to noise. To adopt the same formalism as before, we could write those gradient-based selection measures in the general form:

$$G_I(f) = \{x \in \mathcal{X} \mid (\partial^{|I|} f(x) / \partial X_I)^2 > 0\} \qquad (6)$$

with f a function. For a relaxed formulations for attribution, many aspects have to be considered to provide a relevant estimate for the derivatives for a given task, we will not discuss them here and assume an ideal favorable setting where this measure is available.

Baseline-comparisons methods, in the spirit of counterfactual reasoning, determine the extent to which a function output differs from an output considered "neutral" – the baseline. Many choices exist to model the baseline, a common one is to estimate a conditional expectation. We may formalise them in the general form:

$$C_I(f) = \{x \in \mathcal{X} \mid f(x) \neq \mathbb{E}[f(X) \mid X_I = P_I x]\} \qquad (7)$$

choosing another baseline, as $f(X_I, \mathbb{E}_{\bar{I}}(X_{\bar{I}}))$ (Lundberg and Lee, 2017) does not change our discussion.

To link these two subsets with previous notions, we introduce the following subset of $\mathcal{X}$:

$$B_I(f) = \{x \mid \exists x',\ P_{\bar{I}}x' = P_{\bar{I}}x,\ f(x) \neq f(x')\} \quad (8)$$

i.e. the set of points x for which when fixing the I features, there is still a alternate value for f. This notion is reminiscent of the functionality property in the subsets (A_I), and indeed we have the trivial connection $B_I = \overline{A_{\bar{I}}}$. Then, for gradient-based methods, having a finite non-null gradient implies that there exists a neighborhood such that there exists distinct values for f, and hence $G_I \subset B_I$. But gradient methods miss some cases, for instance if f is constant in the neighborhood of x but vary further away, x will not be included in G_I. Similarly, we have $C_I \subset B_I$: to find a probable point that is different from an average, there must exists points with different value that counterweight its deviation from the mean. Note that the case of improbable points can be handled with a restriction of $\mathcal{X}$. C_I also misses some points of B_I, if a point is associated with the baseline target value, there still may be other points with the same projection on I and with different labels. Thus,

$$A_I \subset \overline{C_{\bar{I}}} \quad (9) \qquad A_I \subset \overline{G_{\bar{I}}} \quad (10)$$

Gradient-based and baseline-comparison proxies are **linked to the formalisation we derive and provide upper bounds for functionality domains**. In section 4 we quantify how good these two approximations are.

3.3. Selector-predictors

A final recent idea is to try to incorporate and learn the instance-wise selection task during training (Chen et al., 2018; Yoon et al., 2019; Arik and Pfister, 2019; Yamada et al., 2020). These techniques have been referred to as *selector-predictor* (Camburu, 2020). The idea is to use two models: a *selector* $\text{Sel} : \mathcal{X} \mapsto \{0,1\}^n$ whose goal is to determine a map S of the most-relevant features for each point; and a *predictor* $\text{Pred} : \mathcal{X} \mapsto \mathcal{Y}$ acting as the usual prediction model of Y with the twist that it takes $X \odot S$ as input. The training objective varies between methods but the general spirit is to maximise the performances of $\text{Pred}(X \odot \text{Sel}(X))$ at predicting Y while either minimising the number of selected features in $\text{Sel}(X)$ or ensuring the constraint that $k < n$ features are selected. A first issue is that most of these methods are only evaluated on performance-degradation metrics or on rather global synthetic selection tasks, which do not truly evaluate instance-wise interpretations. A second, more alarming, issue is that the selector model is completely free and prone to degenerate selection solutions (see Supplementary B). In particular, the selector does not verify properties 1 and 2.

3.4. Relational database connections

We should lastly mention that we found our formalisation to resemble the concept of *functional dependency* from relational database theory (Armstrong, 1974). Our simple categorical attribution (3) is strikingly similar to (Kivinen and Mannila, 1995). But the purpose is not interpretation and in this latter field, global multi-dependence among all columns of a table are sought, differently from between a subset of the input and a designated output, and, to our knowledge, not in an instance-wise manner.

4. Experiments

Armed with a formalism, we generate synthetic distributions with instance-wise ground-truth selections to evaluate attributions methods approximate selection performances and check their solution structure. All generated data, implementations and evaluations methods are available and fully reproducible at our paper code repository [3].

4.1. Synthetic tasks with ground-truth selections

In this section we first explain how, from a desired selection random variable S^*, we are able to build a distribution $p_{X,Y}$ with a given selection solution S^*, *i.e.*

$$S^* = \arg\min_{I \subset [n]} X \in A_I(p_{Y|X})$$

note that A_I depends on p_X. As most selection methods do not handle multiple minimal solution well, we restrict our study to the case with unique selection minimum.

We consider the following simple generative process to draw the data: we uniformly sample from a finite list of points $(c_1, ...c_m) \in \mathcal{X}$ – we call *centroids* – with an associated binary label y_j in $\mathcal{Y} = \{0,1\}$. This is our **perfect-dependence** distribution $p_{X,Y}$:

$$C \sim \mathcal{U}\{1, ...m\}$$
$$\mathbb{P}(X = c_j, Y = y_j) = \mathbb{P}(C = j)$$

As we are in a binary case, interpreting $p_{Y|X}$ can be reduced to the study of the optimal mapping $f = \mathbb{P}(Y = 1|X = x)$. Since we want to assign a unique selection subset $S^*(x) \in [n]$ to each point, we need to ensure that $S^*(x)$ is indeed the minimal subset such that $x \in A_{S^*(x)}(f)$. To do that, we choose the centroids in order to have neighbors with opposite labels in each direction of $S^*(c_j)$ exclusively, so that we know that $c_j \in A_{S^*(c_j)}(f)$, and that for all $J \subset [n]$ such that $J \cap S^*(c_j) \neq \emptyset$, we have $c_j \in B_J(f)$. An example is shown in figure 3.

To have a continuous distribution and allow gradient computations, we then replace our discrete points with normal

[3] Source code at github.com/deezer/functional_attribution

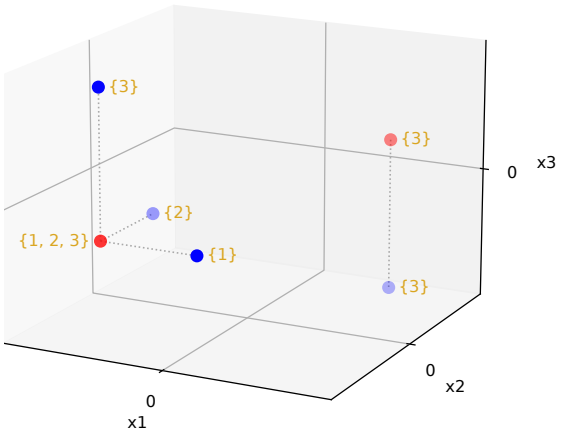

Figure 3. Example of generated distribution $p_{X,Y}$ from a list of six centroids in $\mathbb{R}^3$, with associated labels in blue (0) and red (1), and corresponding unique selection solution $S^*(x)$ in yellow. This example can be found in our dataset under the name `task_3_19`.

distributions with fixed variance σ^2. We obtain a familiar Gaussian mixture distribution $p'_{X,Y}$:

$$p'_{X|C}(x \mid c_j) = \mathcal{N}(x; c_j, \sigma^2)$$
$$p'_{X,Y}(x, y) = \sum_{j=1}^{m} p_C(c_j) p'_{X|C}(x \mid c_j) \delta_{y=y_j}$$

The dependencies are now **imperfect**, we evaluate the capacities of attribution methods to return S^* given $p'_{X,Y}$ and the imperfect optimal mapping $f' = p'_{Y|X}(y = 1 \mid x)$.

With this principle, we are able to generate synthetic distributions of any dimension, with unique selection ground-truth of any dimension. The full generation algorithm, more details and examples are given in the Supplementary C.

4.2. Considered methods

With no requirement to learn a mapping from X to Y, or to make prior assumption on the selection space, many methods collapse into one. Additive mixture of experts methods can all be summarised as the evaluation of *GAM*, *GA^{2}M* with added pairwise interactions, ... up to *GA$^\infty$M* that considers all possible input subset restrictions, thus with an exponential complexity. Note that *GA$^\infty$M* is equivalent to the weighted *fANOVA* without the *integration-to-zero* that hampered instance-wise selection. Generalised additive models do not directly define attribution values, but since their finality is to estimate $\mathbb{E}[Y|X_I]$, we can use the relaxation (3) we derived in section 2.4. We thus dub them with a *"attr"* prefix to underline the modification of the original models. We analyse two supplementary recent methods that we deemed sufficiently different from generalised additive models: *InterpretableNN* (Afchar and Hennequin, 2020), based on *GA$^\infty$M* with a custom selection mechanism inspired by boosting; and *Archipelago* (Tsang et al., 2020), based on *GA^{2}M*, that merges found pairwise dependence using a union-find algorithm to yield disjoint subset selection candidate with a quadratic complexity. Among proxy methods, we evaluate *LIME* (Ribeiro et al., 2016) in both categorical (*Cat.*) and continuous (*Cont.*) configurations; all gradient-based methods cited in 3.2; the sampled classic shapley value estimation (Štrumbelj and Kononenko, 2014) – $\mathbb{E}(f')$, and the baseline approximation introduced in SHAP (Lundberg and Lee, 2017) – $f'(\mathbb{E})$. The selector-predictors are the only methods for which we have to sample from $p'_{X,Y}$ and train two neural networks, we evaluate *L2X* (Chen et al., 2018) with a fixed number of sampled selection dimensions, and *INVASE* (Yoon et al., 2019) that notably replaces this constraint with a Lagrangian penalty in its objective.

4.3. Methods evaluation

We generate **1000 supervised tasks with ground-truth unique univariate selections** – $S^*(c_j)$ is a singleton for all centroids; and **1000 tasks with unique multivariate selections** – $S^*(c_j)$ has a cardinality $k(c_j)$ and is chosen among $\binom{n}{k(c_j)}$ possible subsets. We additionally generate 100 multivariate tasks to tune η for each method. The input space dimension is gradually raised from $\mathbb{R}^2$ to $\mathbb{R}^{11}$, leading up to 2^{11} possible selection subset candidates per centroid.

Our results for univariate selection are given in table 1. *Archipelago*, *InterpretableNN* and *attr-GA^{k}M* methods are all equivalent when returning univariate solutions. We use the standard accuracy metric between the predicted $\hat{S}$ and ground-truth selection S^* on each centroid. Only generalised additive models equipped with the attribution measure (3) and shapley-based methods solve the tasks perfectly. We note that this latter method counter-part, *SHAP*, with the baseline choice $f'(X_I, \mathbb{E}_{\bar{I}}(X_{\bar{I}}))$ especially underperforms despite its complexity. This had already been noticed (Slack et al., 2020) and is due to the fact that the baseline requires out-of-distribution evaluations of f'. For fairness, we also include a performance evaluation (*Acc**) leveraging the prior knowledge that the ground-truth solutions are singletons – *i.e.* selecting the singleton of maximum responsibility.

In table 2, we show the results for selection tasks with ground-truth selections subsets of any cardinality, which is particularly more difficult. The best performing models are still the generalised additive-based and shapley-value-based models. It must be noted that, with synthetic distributions, all methods have access in $O(1)$ to $p'(Y = 1 \mid X_I = x_I)$ for all subset I, whereas we have to let selector-predictors methods learn it from scratch to properly evaluate their selector, hence their high computation time T. Additive model methods are all derived from *GA$^\infty$M* and use caching for faster inferences, we thus only display order of magnitude

Table 1. **Feature selection performance on 1000 univariate tasks** of attributions methods under study, with 95% confidence interval indicators and total computation time T.

Method	Acc (%)	Acc* (%)	T (h:m:s)
LIME (Cat.)	32.4 ± 1.8	61.9 ± 0.8	0:00:54
LIME (Cont.)	10.6 ± 1.0	43.1 ± 0.9	0:00:47
attr-GAM	**100**	**100**	0:00:10
Shapley ($\mathbb{E}(f)$)	**100**	**100**	0:05:36
SHAP ($f(\mathbb{E})$)	23.1 ± 1.2	37.9 ± 1.2	0:05:39
Gradient	33.5 ± 1.0	87.8 ± 0.9	0:00:02
Grad$\times$Input	32.2 ± 1.0	88.4 ± 0.9	0:00:02
Integrated Grad.	38.5 ± 0.9	80.6 ± 1.2	0:00:05
Expected Grad.	45.8 ± 0.9	63.3 ± 0.8	0:00:20
attr-GA$^\infty$M	**100**	**100**	0:01:10
L2X	51.8 ± 1.1	52.5 ± 1.0	19:33:37
INVASE	26.5 ± 1.1	35.5 ± 1.1	45:36:49

for T. *INVASE* particularly underperforms, we believe that this may be magnified by the difficult tuning of its sparsity-inducing penalty term (see Supplementary D).

Table 2. **Feature selection performance on 1000 multivariate tasks** for attributions methods under study.

Method	Acc (%)	T (h:m:s)
LIME (Cat.)	16.2 ± 1.3	0:05:54
LIME (Cont.)	27.4 ± 1.6	0:05:47
attr-GAM	24.5 ± 1.5	0:00:25
Shapley ($\mathbb{E}(f)$)	74.3 ± 1.1	0:16:29
SHAP ($f(\mathbb{E})$)	15.7 ± 1.3	0:17:41
Gradient	26.5 ± 1.5	0:00:04
Gradient$\times$Input	22.6 ± 1.5	0:00:04
Integrated Gradient	18.5 ± 1.4	0:00:24
Expected Gradient	21.4 ± 1.4	0:03:42
attr-GA$^\infty$M	$\mathbf{81.7 \pm 1.1}$	0:17:44*
attr-GA^{2}M	52.5 ± 1.8	$\ll *$
attr-GA^{3}M	74.1 ± 1.3	$< *$
attr-GA^{4}M	81.2 ± 1.1	$< *$
InterpretableNN	79.7 ± 1.2	$\simeq *$
Archipelago	70.2 ± 1.1	$\simeq *$
L2X	23.7 ± 1.6	32:53:16
INVASE	7.4 ± 0.9	44:15:44

4.4. Necessary property evaluation

We finally link the performance differences we observe, to the necessary properties we derived in section 2.3 and check whether the methods provide well structured selection solutions. Using the predicted selections of the multivariate tasks, we compute the ratio of centroids verifying property 1. To do that, we leverage property 2: 1 is verified for a centroid c_j *iff* for a given selection $\hat{S}(c_j)$, for every centroids c_k such that $P_{\hat{S}(c_j)}(c_j) = P_{\hat{S}(c_j)}(c_k)$, we have $\hat{S}(c_k) \subseteq \hat{S}(c_j)$.

The results are presented in table 3. Strikingly, we observe a correlation between the property-verification and feature selection performances ($\rho = 0.88$). This is a strong indication that **well structured selections solutions** with regards to all points in the input distribution **tend to also be better performing**. Interestingly, we must underline that computing the property verification rate **does not require to have access to ground-truth selections**, which opens the door for further applications to real-data tasks. We must however emphasize that property 1 is not sufficient: *e.g.* globally returning all or no input features for all points as selection is a perfectly structured solution according to 1, but rarely an optimal one. Property 1 is necessary but not sufficient. It must help design better instance-wise attribution methods that intrinsically verify it, as *GA$^\infty$M*, but should not be a criterion to maximise.

Table 3. **Ratio of points verifying property 1 on 1000 multivariate tasks**. Note that this does not require any ground-truth label, only the proposed selection solution is analysed.

Method	Property verification rate (%)
LIME (Cat.)	29.9 ± 1.7
LIME (Cont.)	46.6 ± 1.6
attr-GAM	61.5 ± 1.1
Shapley ($\mathbb{E}(f)$)	79.5 ± 1.1
SHAP ($f(\mathbb{E})$)	23.7 ± 1.5
Gradient	61.6 ± 1.3
Gradient $\times$ Input	54.5 ± 1.3
Integrated Gradient	39.7 ± 1.5
Expected Gradient	41.8 ± 1.5
attr-GA$^\infty$M	$\mathbf{92.9 \pm 0.6}$
attr-GA^{2}M	63.7 ± 1.4
attr-GA^{3}M	81.2 ± 1.4
attr-GA^{4}M	90.7 ± 1.1
InterpretableNN	86.9 ± 0.9
Archipelago	88.8 ± 0.7
L2X	37.5 ± 1.6
INVASE	61.3 ± 1.7

5. Conclusion

The growing interest in *interpretable machine learning* and the profusion of recent feature attribution methods has motivated us to take a step back and propose a rigorous formalisation of often vaguely defined concepts in this field. Though to some extent task-dependent, we argue that all these methods can be analysed through an irreducible component: feature selection. Doing so, we could evaluate many state-of-the-art methods on rigorously derived ground-truth rationales, and we have derived provable necessary properties that any computed interpretations must verify – which is not the case for some popular methods. Our future directions involve using our relaxation framework to derive good attribution measures for specific applications and building new efficient and well-formulated attribution models.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I. J., Hardt, M., and Kim, B. (2018). Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems 31, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 9525–9536.

Afchar, D. and Hennequin, R. (2020). Making neural networks interpretable with attribution: Application to implicit signals prediction. In *RecSys 2020: Fourteenth ACM Conference on Recommender Systems, Virtual Event, Brazil, September 22-26, 2020*, pages 220–229. ACM.

Arik, S. O. and Pfister, T. (2019). Tabnet: Attentive interpretable tabular learning. *arXiv preprint arXiv:1908.07442*.

Armstrong, W. W. (1974). Dependency structures of data base relationships. In *IFIP congress*, volume 74, pages 580–583. Geneva, Switzerland.

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

Blum, A. L. and Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial intelligence*, 97(1-2):245–271.

Camburu, O.-M. (2020). Explaining deep neural networks. *arXiv preprint arXiv:2010.01496*.

Chen, J., Song, L., Wainwright, M. J., and Jordan, M. I. (2018). Learning to explain: An information-theoretic perspective on model interpretation. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80, pages 882–891. PMLR.

Covert, I., Lundberg, S., and Lee, S.-I. (2020). Feature removal is a unifying principle for model explanation methods. *NeurIPS 2020 ML-Retrospectives, Surveys & Meta-Analyses Workshop*.

Dinu, J., Bigham, J., and Kolter, J. Z. (2020). Challenging common interpretability assumptions in feature attribution explanations. *NeurIPS 2020 ML-Retrospectives, Surveys & Meta-Analyses Workshop*.

Dombrowski, A., Alber, M., Anders, C. J., Ackermann, M., Müller, K., and Kessel, P. (2019). Explanations can be manipulated and geometry is to blame. In *Advances in Neural Information Processing Systems 32, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13567–13578.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Erion, G., Janizek, J. D., Sturmfels, P., Lundberg, S., and Lee, S.-I. (2019). Learning explainable models using attribution priors. *arXiv preprint arXiv:1906.10670*.

Friedman, J. H. and Popescu, B. E. (2008). Predictive learning via rule ensembles. *Ann. Appl. Stat.*, 2(3):916–954.

Hamilton, A. G. (1982). *Numbers, sets and axioms: the apparatus of mathematics*. Cambridge University Press.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*, volume 43. CRC press.

Hooker, G. (2007). Generalized functional anova diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 16(3):709–732.

Hu, Y., Koren, Y., and Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining*, pages 263–272. Ieee.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H. M., and Vaughan, J. W. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, pages 1–14. ACM.

Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., and Kim, B. (2019). The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 267–280.

Kivinen, J. and Mannila, H. (1995). Approximate inference of functional dependencies from relations. *Theoretical Computer Science*, 149(1):129–149.

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., and Friedler, S. A. (2020). Problems with shapley-value-based explanations as feature importance measures. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119, pages 5491–5500. PMLR.

Lengerich, B., Tan, S., Chang, C., Hooker, G., and Caruana, R. (2020). Purifying interaction effects with the functional ANOVA: an efficient algorithm for recovering

identifiable additive models. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108, pages 2402–2412. PMLR.

Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, 16(3):31–57.

Lou, Y., Caruana, R., Gehrke, J., and Hooker, G. (2013). Accurate intelligible models with pairwise interactions. In *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*, pages 623–631.

Lundberg, S. M. and Lee, S. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774.

Natarajan, B. K. (1995). Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234.

Parliament, E. U. (2016). Regulation (eu) 2016/679.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Serrano, S. and Smith, N. A. (2019). Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. *Workshop, ICLR*.

Sixt, L., Granz, M., and Landgraf, T. (2020). When explanations lie: Why many modified BP attributions fail. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119, pages 9046–9057. PMLR.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). 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.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. *Workshop on Visualization for Deep Learning, ICML*.

Štrumbelj, E. and Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3):647–665.

Sundararajan, M. and Najmi, A. (2020). The many shapley values for model explanation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119, pages 9269–9278. PMLR.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70, pages 3319–3328. PMLR.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tintarev, N. and Masthoff, J. (2007). A survey of explanations in recommender systems. In *2007 IEEE 23rd international conference on data engineering workshop*, pages 801–810. IEEE.

Tsang, M., Liu, H., Purushotham, S., Murali, P., and Liu, Y. (2018). Neural interaction transparency (NIT): disentangling learned interactions for improved interpretability. In *Advances in Neural Information Processing Systems 31, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 5809–5818.

Tsang, M., Rambhatla, S., and Liu, Y. (2020). How does this interaction affect me? interpretable attribution for feature interactions. *Advances in Neural Information Processing Systems*.

Yamada, Y., Lindenbaum, O., Negahban, S., and Kluger, Y. (2020). Feature selection using stochastic gates. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119, pages 10648–10659. PMLR.

Yoon, J., Jordon, J., and van der Schaar, M. (2019). INVASE: instance-wise variable selection using neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.