Explaining the Model and Feature Dependencies by Decomposition of the Shapley Value*

Joran Michiels[a,*], Maarten De Vos[a,b], Johan Suykens[a]

[a]*Stadius group in the Department of Electrical Engineering, KU Leuven University, Kasteelpark Arenberg 10, 3001, Leuven, Belgium*
[b]*Department of Development and Regeneration, KU Leuven University, Herestraat 49, 3000, Leuven, Belgium*

Abstract

Shapley values have become one of the go-to methods to explain complex models to end-users. They provide a model agnostic post-hoc explanation with foundations in game theory: what is the worth of a player (in machine learning, a feature value) in the objective function (the output of the complex machine learning model). One downside is that they always require outputs of the model when some features are missing. These are usually computed by taking the expectation over the missing features. This however introduces a non-trivial choice: do we condition on the unknown features or not? In this paper we examine this question and claim that they represent two different explanations which are valid for different end-users: one that explains the model and one that explains the model combined with the feature dependencies in the data. We propose a new algorithmic approach to combine both explanations, removing the burden of choice and enhancing the explanatory power of Shapley values, and show that it achieves intuitive results on simple problems. We apply our method to two real-world datasets and discuss the explanations. Finally, we demonstrate how our method is either equivalent or superior to state-to-of-art Shapley value implementations while simultaneously allowing for increased insight into the model-data structure.

Keywords: Explainable AI, Feature attribution, End-user, Shapley value, Black box, Feature dependencies

1. Introduction

Black box machine learning methods have shown remarkable results on complex prediction tasks in the past years. Widespread adoption in practical settings is limited however. This is believed to be partly due to the lack of insights in the model's inner workings, making validating the model very difficult. A popular choice to gain insight in the model's decisions are model agnostic local post-hoc explanations. These explain one sample (*local*) and its corresponding prediction by sampling any (*model agnostic*) already trained model (*post-hoc*).

One local post-hoc explanation, the **SH**apley **A**dditive ex**P**lanation (SHAP) [14], has become increasingly popular in the last few years. Apart from its excellent Python module, this is likely due to the solid game-theoretic background of the method: other than the degree of approximation, nothing is left to the choice of the user, or so it seems. As has been recently noticed in related literature [9, 11, 22] the game-theoretic concept does not translate literally to machine learning models, leading to interpretation problems when features are dependent. It should be noted that similar problems can arise in other post-hoc explanations such as LIME [16]. They are however not covered in related papers (nor in our work). This is likely because uniqueness is one of the main selling points of SHAP, unlike other explanation methods.

In our opinion, central to the understanding of these interpretation problems is the distinction between the model and the result (the model combined with the data), and the corresponding explanations. We qualitatively argue that different settings can have different requirements for local post-hoc explanations with a simple example (inspired by [9] and [22]). Assume the following risk model f_r and dependent features:

$$\begin{gathered} f_r([X_1, X_2]) = X_1 \\ X_1, X_2 \quad \text{binary} \\ X_1 \not\perp\!\!\!\perp X_2 \end{gathered} \tag{1}$$

Let us imagine two different fictive settings with different feature definitions in Table 1. Although the assumed data distribution is simplified, it is known that for both settings the features tend to be associated [12, 17]. Given a certain sample, with SHAP we can explain its output by determining what effect every feature value has towards the output. In the first setting f_r is an indicator for recidivism

*This project has received funding from the Flemish Government (AI Research Program) and from the FWO ('Artificial Intelligence (AI) for data-driven personalised medicine', G0C9623N and 'Deep, personalized epileptic seizure detection', G0D8321N) and Leuven.AI Institute.

*Corresponding author

Email addresses: joran.michiels@esat.kuleuven.be (Joran Michiels), maarten.devos@kuleuven.be (Maarten De Vos), johan.suykens@esat.kuleuven.be (Johan Suykens)

Table 1: Different settings with different features

	X_1	X_2	$f_r(\cdot)$
court	recidivist	race	future recidivism
hospital	high HR	obesity	heart attack

of a convict. By Article 6 of the European Court of Human Rights, a convict $\boldsymbol{x}$ can demand an explanation of the algorithm involved in the courts decision. In this case the explanation should attribute $f_r([\boldsymbol{x}_1, \boldsymbol{x}_2])$ completely to $\boldsymbol{x}_1$, showing the absence of racial bias in the *model*. The second setting needs a different explanation. For patient $\boldsymbol{x}$ with a high heart attack risk, a good explanation should arguably attribute part of $f_r([\boldsymbol{x}_1, \boldsymbol{x}_2])$ to $\boldsymbol{x}_2$, since changing $\boldsymbol{x}_2$ by losing weight could decrease the heart rate, which in term reduces the odds of having a heart attack according to the model. Note that this last attribution is solely due to the dependency in the data.

Thus, although the model and data distribution are the same, different settings seem to require different kinds of explanations. The first explanation explains the model while the second tries to explain the model combined with the data, which we call the actual *result*. Preferably, a good method should account for both settings. In this paper, we show that a SHAP value for a feature can be split up into a direct (model corresponding) attribution and an attribution via other dependent features, allowing the end-user (e.g. convict or patient) to easily grasp the explanation of *both* the model and the result. Specifically, the contributions of this paper are:

- identification of the aforementioned distinction between explaining the model and explaining the result,
- an interpretation of a previously noticed decomposition [8] as a unification of the two most common Shapley value implementations [9, 1],
- a new model-agnostic implementation of Shapley values with accompanying plots that provide enhanced insights in model and data,
- experiments demonstrating the correctness of our approach and its superiority over the current state-of-art [10] to quantify the effect of dependent features.

2. Background

This section covers the necessary theoretical background of explanations based on Shapley values and their inherent problem with dependent features.

2.1. Shapley additive explanations

SHAP is a post-hoc (i.e. after training) method to interpret the output of a complex model for a particular sample $\boldsymbol{x}$ (note the bold font) [14]. It constructs a local approximation g of the original prediction model f around $\boldsymbol{x}$ and is of the form:

$$g\left(X'\right) = \phi_0 + \sum_{i=1}^{M} \phi_i X'_i \quad \text{with} \quad X' \in \{0,1\}^M \;\; \phi_i \in \mathbb{R}, \tag{2}$$

with X' the so-called simplified input of the original input X, and ϕ_i the actual SHAP values or feature attributions. To transform X' to X a local mapping function $h_{\boldsymbol{x}}$ is defined. In general any mapping $X = h_{\boldsymbol{x}}\left(X'\right)$ can be utilized, in practice X'_i will be defined as an indicator whether the particular input feature is present or not, i.e. for particular input $\boldsymbol{x}$, $X'_i = \mathbf{1}\left[X_i = \boldsymbol{x}_i\right]$. In this case M is equal to the number of features and the explanation effectively becomes an additive function on feature presence.

To select reasonable feature attributions ϕ_i for the explanation model, three properties are imposed.

1. **Local accuracy**:

$$f\left(\boldsymbol{x}\right) = g\left(\boldsymbol{x}'\right) \tag{3}$$

   i.e. at the sample to explain $\boldsymbol{x}$, the explanation matches the original model.

2. **Missingness**:

$$X'_i = 0 \implies \phi_i = 0, \tag{4}$$

   i.e. if the original input X already had a missing feature, its attribution is always zero. The equivalent form,

$$f(h_{\boldsymbol{x}}(x' \cup i)) = f(h_{\boldsymbol{x}}(x' \backslash i)) \;\forall x' \implies \phi_i = 0 \tag{5}$$

   with $x' \cup i$ meaning $x'_i = 1$ and with $x' \backslash i$ meaning $x'_i = 0$, is clearer.

3. **Consistency**:
   For any two models f_1 and f_2:

$$\begin{gathered} f_1\left(h_{\boldsymbol{x}}\left(x'\right)\right) - f_1\left(h_{\boldsymbol{x}}\left(x' \backslash i\right)\right) \\ \geq f_2\left(h_{\boldsymbol{x}}\left(x'\right)\right) - f_2\left(h_{\boldsymbol{x}}\left(x' \backslash i\right)\right) \;\; \forall x' \in \{0,1\}^M \\ \implies \phi_i\left(f_1, \boldsymbol{x}\right) \geq \phi_i(f_2, \boldsymbol{x}) \end{gathered} \tag{6}$$

   In words, this says that "if a model changes so that some simplified input's contribution increases or stays the same regardless of the other inputs, that input's attribution should not decrease" [14].

Only one[1] possible set of feature attributions satisfies these properties [14]:

$$\phi_i(f, \boldsymbol{x}) = \frac{1}{M!} \sum_{R} f\left(h_{\boldsymbol{x}}\left(S^R \cup \{i\}\right)\right) - f\left(h_{\boldsymbol{x}}\left(S^R\right)\right), \tag{7}$$

where R is a permutation of the ordering of simplified inputs and S^R the set of simplified inputs preceding i in

[1] [22] claim other properties are necessary for uniqueness.

this ordering. To make the formula less cluttered, we use $\left(S^R\right)$ to denote $\left(\left[X'_{S^R}=1, X'_{\bar{S}^R}=0\right]\right)$. The values ϕ_i are known as Shapley values in cooperative game theory [18]. In practice, when the simplified inputs correspond to feature presence, ϕ_i is the average over all orderings of the change in model output when adding feature i to the preceding, already 'known' features. Equation (7) requires the output of the model when features are missing: $f\left(h_{\boldsymbol{x}}\left(X'\right)\right)$ when $X' \neq \boldsymbol{x}'$. This involves some empirical expectation of $f\left(h_{\boldsymbol{x}}\left(X'\right)\right)$ over a set of training samples. What expectation to choose - conditioning on the missing features or not - will appear to be a controversial topic.

2.2. Missing features

The original SHAP paper [14] assumed a seemingly natural way to express the expectation of the model output with missing features: conditioning on known features S,

$$f\left(h_{\boldsymbol{x}}\left(X'\right)\right) = E\left[f(X)|X_S = \boldsymbol{x}_S\right]. \quad \text{(conditional SHAP)} \tag{8}$$

In practice, the exact data distribution is unknown. Therefore [14] proposed to use independent features. Then the expectation can be easily approximated by averaging over K background samples x^k:

$$E\left[f(X)|X_S = \boldsymbol{x}_S\right] \approx E\left[f\left(\left[\boldsymbol{x}_S, X_{\bar{S}}\right]\right)\right] \tag{9}$$

$$\approx \frac{1}{K}\sum_k f\left(\left[\boldsymbol{x}_S, x^k_{\bar{S}}\right]\right), \tag{10}$$

with missing features $\bar{S}$. A later contribution by the same authors [13] focussed on avoiding the feature independence assumption by exploiting the structure of tree models. Part of the community also followed this direction: [1, 2] estimate the conditional data distribution $p\left(X_{\bar{S}}|X_S\right)$ to better approximate (8) and [23] uses the data distribution assumed by probabilistic PCA to 'exactly' compute the SHAP value.

Some recent papers have tried to advocate a different estimation of the model output when features are missing. [9, 8, 6] analyse the missing feature problem from a causal perspective. The conditioning on known features is argued to be *interventional*, instead of *observational* (as was implicitly assumed in conditional SHAP). Computing interventional conditional distributions requires a causal graph over the features. [9] distinguishes between the true real world features $\tilde{X}$ and the model inputs X. There can be causal relations between the former but not between the latter. [8] does not make this distinction and the resulting SHAP values are arguably the true causal output effects. The authors therefore deemed their explanations *causal Shapley values.* To illustrate the difference with the approach of [9], a causal graph for a model with three input features is shown in Figure 1 for both approaches.

Contrary to causal Shapley values [8] the method of [9] does not require any knowledge of the causal relations between the (real world) features. The authors show that

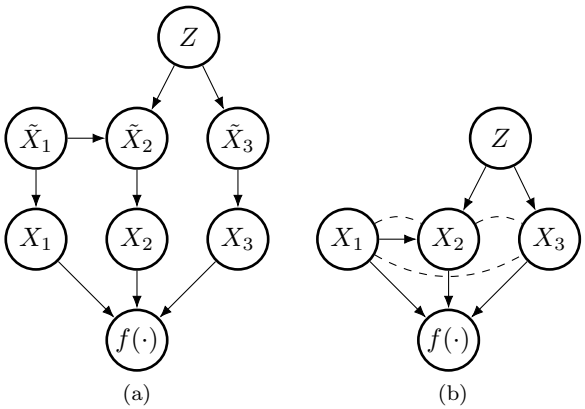


Figure 1: Example causal structures assumed by respectively interventional SHAP [9] (a) and causal Shapley values [8] (b). In (a) the $\tilde{X}_i$ are the real world features and X_i the model inputs. This distinction is not made in (b). For both cases Z is a latent confounder. The dotted lines are the observed statistical dependencies which are taken into account by conditional SHAP, and the new method proposed in this paper.

the particular causal structure (e.g Figure 1a) leads to a model output equivalent to the left hand side of (9):

$$f\left(h_{\boldsymbol{x}}\left(X'\right)\right) = E\left[f\left(\left[\boldsymbol{x}_S, X_{\bar{S}}\right]\right)\right]. \quad \text{(interventional SHAP)} \tag{11}$$

Without the weight of the coalition, the contribution for one permutation R, of feature i then becomes

$$\phi_{i,S^R} = E\left[f\left(\left[\boldsymbol{x}_{S^R}, \boldsymbol{x}_i, X_{\bar{S}^R_i}\right]\right)\right] - E\left[f\left(\left[\boldsymbol{x}_{S^R}, X_{\bar{S}^R}\right]\right)\right] \tag{12}$$

instead of

$$\begin{aligned}\phi_{i,S^R} = E\left[f\left(\left[\boldsymbol{x}_{S^R}, \boldsymbol{x}_i, X_{\bar{S}^R_i}\right]\right)|\boldsymbol{x}_{S^R}, \boldsymbol{x}_i\right] \\ -E\left[f([\boldsymbol{x}_{S^R}, X_{\bar{S}^R}])|\boldsymbol{x}_{S^R}\right]\end{aligned} \tag{13}$$

for conditional SHAP, with S^R_i the set S^R including feature i. In short, interventional SHAP breaks the dependencies between X_i and $X_{\bar{S}^R_i}$, while conditional SHAP allows for them to have an effect. It is important to note that interventional SHAP keeps the dependencies within $X_{\bar{S}^R_i}$ [9].

[22] uses the axiomatic approach to study the attribution problem. They test interventional and conditional SHAP against a list of desirable properties. Contrary to the properties imposed in Section 2.1, these are at the level of the model f not at the level of the model approximations ($f\left(h_{\boldsymbol{x}}\left(X'\right)\right)$ when $X' \neq \boldsymbol{x}'$), e.g. missingness (5) becomes 'dummy'.

Dummy:

$$f([x_i, x_{\setminus i}]) = f([d, x_{\setminus i}])\ \forall x, d \implies \phi_i = 0 \tag{14}$$

The authors show that conditional SHAP fails this property (and others), while interventional SHAP does not.

3. Related work

Recently [4] also suggested the same distinction as described in our introduction and motivate each of the approaches with a use case. Interventional SHAP seems the method of choice when a loan applicant wants to see what features directly increase their likelihood of being granted a loan. If we want to know more about the natural mechanism of gene expression in a patient with a type of blood cancer, the conditional SHAP value is preferred and identified more of the true underlying variables. They call this approach 'true to the data'. In contrast with this paper, no attempt to unify both types of SHAP values was undertaken.

Similar to our contribution, [8] decomposes the causal Shapley value into a direct and indirect effect and applies the decomposition to a theoretical example to show how different causal graphs result in different Shapley values. They do not provide a possible implementation and neither do they motivate using the decomposition in practice. Here we interpret this decomposition as a unification of conditional and interventional SHAP, and provide an implementation and illustrations to increase the understanding of the method. Furthermore we compare it with other state-of-the art Shapley implementations.

Shapley residuals [10] can capture limitations of regular Shapley value implementations. They are represented by residual vectors r_i measuring how distant all contributions for specific S ($\phi_{i,S}$) are from the contributions of the nearest inessential game. In an inessential game every player (in case of machine learning a feature value) contributes a fixed amount to the objective regardless of S. As shown in [10] the contributions of the nearest inessential game are the actual Shapley values ($\phi_{i,S} = \phi_i$). In case of conditional SHAP value, Shapley residuals quantify the effect of dependent features. Specifically, it is argued that the norm of r_i relates to the effect of dependencies between feature i and other features on the SHAP value. For interventional SHAP values, the residuals capture effects of feature interactions in the model (e.g. of the term X_1X_2 in model $f(X_1, X_2) = X_1 + X_2 + X_1X_2$). Explaining such feature interaction (also see Shapley-Interaction interaction indices [21]) is not the interest of this paper.

[24] also design a Shapley value implementation that provides a more involved explanation when dependencies between input features exist. In fact, they assign a contribution to every causal link (or edge in causal graph) while still adhering to the fundamental axioms described in Section 2.1. A connection with conditional SHAP is made but interventional SHAP values are not discussed. Similarly to causal Shapley values [8] a causal graph over the features is required.

Another important difference between interventional and conditional SHAP values is that the former evaluates the machine learning model on out of distribution samples. Intervening on feature values results in samples which do not represent the underlying data distribution (also called off-manifold). It is shown that this allows for adversarial attacks that can manipulate explanations to hide unwanted model biases [19]. While we acknowledge this issue, interventions still provide important insight into model as motivated in our introduction (e.g. for bias detection) and in [4]. For more information on the off-manifold problem we direct the reader to [5, 25]. In this paper, the focus will be on the meaningful usage and desired theoretic properties of SHAP values, and not on their robustness or efficiency.

4. Explaining model and result

The properties defined in [22] (such as (14)) quantify the model correspondence of the explanations. Conditional SHAP, not adhering to many of these properties, sacrifices this model correspondence for more result-centric feature attributions. These latter attributions arguably show the actual statistical relations between inputs and output (or result). Note that the relations may not be fully present in the data. The model fitting enforces some domain and/or data knowledge. Also, the complexity of the model might be limited for computational reasons, meaning that the 'true' relations in the data are not uncovered. On this point we do not completely agree with [4]. They mention that conditional SHAP values are true to the data and interventional SHAP values true to the model. As we argued, conditional SHAP values are both true to the data *and* true to the model.

Causal Shapley values [8] show the true *causal* relations between inputs and output, and thus also explain the result. The big difference is that conditional SHAP relies on association (or statistical dependence) and the causal Shapley value on causation. For example in Figure 1b X_2 and X_3 have a common cause Z and will be associated (note the dotted line) without a causal relationship. We agree with [8] that causal relations are more intuitive than observed statistical dependencies. However, in practice causal knowledge about the input features is often unavailable, hence the use of conditional SHAP in the rest of this paper.

It should be clear that in the introductory example interventional SHAP values are preferred in court; and in the hospital setting the conditional version would be better. Conditional SHAP also finds the feature's attribution through other dependent features. We can show that the feature's attribution can be split up into its 'direct' interventional effect and its 'indirect' dependent effect. This decomposition of the SHAP value can be used in both the court and hospital setting.

4.1. Decomposition of the SHAP value

Since our interest is decomposing the output effect into an interventional effect of $\boldsymbol{x}_i$ and its effect via the unknown features $\bar{S}_i$, we determine the effect of intervening on $\boldsymbol{x_i}$

when the features in S are already known. I.e. the effect of setting $X_i = \boldsymbol{x_i}$ (breaking X_i's dependencies) on $f_S([X_i, X_{\bar{S}_i}]) = f([\boldsymbol{x}_S, X_i, X_{\bar{S}_i}])$. This is equal to:

$$\begin{aligned} &E\left[f_S\left([\boldsymbol{x}_i, X_{\bar{S}_i}]\right)\right] - E\left[f_S(X_{\bar{S}})\right] \\ &= E\left[f\left([\boldsymbol{x}_S, \boldsymbol{x}_i, X_{\bar{S}_i}]\right)|\boldsymbol{x}_S\right] - E\left[f([\boldsymbol{x}_S, X_{\bar{S}}])|\boldsymbol{x}_S\right] \end{aligned} \tag{15}$$

This effect can be interpreted as a part of the contribution for one permutation R in case of the conditional SHAP value:

$$\begin{aligned} \phi_{i,S^R}(f,\boldsymbol{x}) =& E\left[f(X)|X_{S^R} = \boldsymbol{x}_{S^R}, Z_i = \boldsymbol{x}_i\right] \\ &- E\left[f(X)|X_{S^R} = \boldsymbol{x}_{S^R}\right] \\ =& E\left[f\left(\left[\boldsymbol{x}_{S^R}, \boldsymbol{x}_i, X_{\bar{S}_i^R}\right]\right)|\boldsymbol{x}_{S^R}, \boldsymbol{x}_i\right] \\ &- E\left[f([\boldsymbol{x}_{S^R}, X_{\bar{S}^R}])|\boldsymbol{x}_{S^R}\right] \\ =& \Big(E\left[f\left(\left[\boldsymbol{x}_{S^R}, \boldsymbol{x}_i, X_{\bar{S}_i^R}\right]\right)|\boldsymbol{x}_{S^R}\right] \\ &- E\left[f([\boldsymbol{x}_{S^R}, X_{\bar{S}^R}])|\boldsymbol{x}_{S^R}\right]\Big) \\ &+ \Big(E\left[f\left(\left[\boldsymbol{x}_{S^R}, \boldsymbol{x}_i, X_{\bar{S}_i^R}\right]\right)|\boldsymbol{x}_{S^R}, \boldsymbol{x}_i\right] \\ &- E\left[f\left(\left[\boldsymbol{x}_{S^R}, \boldsymbol{x}_i, X_{\bar{S}_i^R}\right]\right)|\boldsymbol{x}_{S^R}\right]\Big) \\ =& \phi_{i,S^R,int}(f,\boldsymbol{x}) + \phi_{i,S^R,dep}(f,\boldsymbol{x}). \end{aligned} \tag{16}$$

We have arrived at an equivalent decomposition as in [8] but without causal operations. From (16) we see that the contribution of a feature i (for a permutation R) is the sum of an interventional output effect of $\boldsymbol{x}_i$ conditioned on $\boldsymbol{x}_{S^R}$ and the output effect of knowing $\boldsymbol{x}_i$ on the unknown variables, again conditioned on $\boldsymbol{x}_{S^R}$. Notice the similarity between $\phi_{i,S^R,int}(f,\boldsymbol{x})$ and interventional SHAP (12): the dependencies between X_i and $X_{\bar{S}_i^R}$ are still broken. Consequently, the causal motivation behind interventional SHAP still holds and this decomposition successfully connects the interventional and conditional approach to dealing with missing features.

Note that above interpretation also holds for the Shapley value since that is an average over all $\phi_{i,S^R}(f,\boldsymbol{x})$. Thus, the conditional SHAP value can be split up into two SHAP value parts. Although these parts are not real SHAP values (there respective sums do not equal the original output values, i.e. no local accuracy), they arguably provide both a model and a result explanation.

As explained in the related work, Shapley residuals [10] for a particular model f and feature i can be interpreted as the geometric difference between the contributions $\phi_{i,S}$ for all S and the SHAP value ϕ_i, i.e. $r_i \in \mathbb{R}^{|S|}$ and $r_{i,S} = \phi_{i,S} - \phi_i$. Thus, they define a decomposition of similar form as Equation (16) which was not discussed in [10], namely:

$$\phi_{i,S^R}(f,\boldsymbol{x}) = \phi_i(f,\boldsymbol{x}) + r_{i,S^R} \tag{17}$$

This allows us to claim the following proposition.

Proposition 1. *The associated SHAP part of the Shapley residual is always zero.*

Proof. The SHAP part can be found by averaging (17) over all R:

$$\frac{1}{M!}\sum_R \phi_{i,S^R}(f,\boldsymbol{x}) = \phi_i(f,\boldsymbol{x}) + \frac{1}{M!}\sum_R r_{i,S^R}. \tag{18}$$

Since the left hand side is exactly equal to the SHAP value we arrive at:

$$\frac{1}{M!}\sum_R r_{i,S^R} = 0. \tag{19}$$

□

Because the associated SHAP part will always be zero, the concept of Shapley residuals can not be used to split the SHAP value into a sum of parts.

4.2. Theoretical validation

It is easily proven that the interventional part passes the dummy property (14), underlining its model correspondence. Also, for an independent distribution the interventional SHAP part coincides with the interventional SHAP value.

Proposition 2. *Interventional SHAP part passes the dummy property* (14)*, i.e.*

$$f([x_i, x_{\setminus i}]) = f([d, x_{\setminus i}])\ \forall x, d \implies \phi_{i,int} = 0 \tag{20}$$

Proof. Take feature i to be a missing feature:

$$f([x_i, x_{\setminus i}]) = f([d, x_{\setminus i}]) \quad \forall x, d. \tag{21}$$

Its feature value contribution for permutation R,

$$\begin{aligned} \phi_{i,S^R,int}(f,\boldsymbol{x}) =& E\left[f\left(\left[\boldsymbol{x}_{S^R}, \boldsymbol{x}_i, X_{\bar{S}_i^R}\right]\right)|\boldsymbol{x}_{S^R}\right] \\ &- E\left[f([\boldsymbol{x}_{S^R}, X_{\bar{S}^R}])|\boldsymbol{x}_{S^R}\right] \\ =& E\left[f\left(\left[\boldsymbol{x}_{S^R}, \boldsymbol{x}_i, X_{\bar{S}_i^R}\right]\right)|\boldsymbol{x}_{S^R}\right] \\ &- E\left[f\left(\left[\boldsymbol{x}_{S^R}, X_i, X_{\bar{S}_i^R}\right]\right)|\boldsymbol{x}_{S^R}\right] \end{aligned} \tag{22}$$

is zero since both terms are equal by (21). The actual SHAP part, an average over the former differences, is thus also zero. □

Proposition 3. *Our proposed decomposition* (16) *divides the model output effect among the model's additive components. Writing* $f(X) = \sum_{A\subseteq F} f(X_A)$ *then*

$$\begin{aligned} \phi_{i,int}(f,\boldsymbol{x}) &= \phi_{i,int}\left(\sum_{A\subseteq F|i\in A} f_A, \boldsymbol{x}\right) \\ \phi_{i,S^R,dep}(f,\boldsymbol{x}) &= \phi_{i,S^R,dep}\left(\sum_{A\subseteq F|S_R\subseteq A} f_A, \boldsymbol{x}\right) \end{aligned} \tag{23}$$

Proof. Using Equation (16) and knowing that the expectation of a sum of variables is equal to the sum of its individual expectations, we find that parts of the additive function cancel each other resulting in the claimed properties. □

We can gain extra intuition by considering a linear model: $f = \sum_i a_i X_i$. In this case, the interventional SHAP values ψ_i, interventional SHAP parts $\phi_{i,int}$ and dependent SHAP parts $\phi_{i,dep}$ for a sample $\boldsymbol{x}$ simplify to:

$$\psi_i = a_i(\boldsymbol{x}_i - E[X_i]) = f(\boldsymbol{x}) - f\left(\left[\boldsymbol{x}_{\setminus i}, E[X_i]\right]\right), \tag{24}$$

$$\begin{aligned}\phi_{i,int} =& \frac{1}{M!}\sum_R a_i(\boldsymbol{x}_i - E[X_i|\boldsymbol{x}_{S^R}]) \\ =& \frac{1}{M!}\sum_R f(\boldsymbol{x}) - f\left(\left[\boldsymbol{x}_{\setminus i}, E[X_i|\boldsymbol{x}_{S^R}]\right]\right),\end{aligned} \tag{25}$$

$$\phi_{i,dep} = \frac{1}{M!}\sum_R \sum_{j\in\bar{S}_i^R} a_j(E[X_j|\boldsymbol{x}_{S^R}, \boldsymbol{x}_i] - E[X_j|\boldsymbol{x}_{S^R}]). \tag{26}$$

First of all, notice that the interventional SHAP value corresponds with the change in model output when setting the respective feature to its average value. The interventional SHAP part, on the other hand, is the average over all permutations of the change in model output when intervening with a conditional mean. The conditioning is on the feature values appearing before i in the permutation. In one of the experiments below we will argue why the latter mean can be more realistic. Note that the above observations are only exact for a linear machine learning model. Lastly, as expected from Proposition 3, $\phi_{i,int}$ only depends on X_i's coefficient, while the dependent SHAP part $\phi_{i,dep}$ depends on all the other coefficients.

4.3. Implementation

Computing the SHAP values from (7) exactly is only feasible if the model has a few inputs, since it requires a sum over all possible orderings of inputs. To efficiently compute the SHAP value for models with more inputs, [14] proposes Kernel SHAP which interprets the computation of Equation (7) as performing a weighted linear regression. This is shown to be more sample-efficient than straightforwardly sampling the sum in (7). Computation of the SHAP parts can not be expressed as a weighted linear regression. Hence, to calculate the interventional SHAP part, we use a more rudimentary sampling approach as proposed by [20].

The complete procedure for a sample $\boldsymbol{x}$ is as follows:

1. obtain or estimate the conditional data distribution $p\left(X_{\bar{S}}|X_S\right)$;
2. compute the conditional SHAP values by applying Kernel SHAP, estimating $E\left[f(X)|X_S = \boldsymbol{x}_S\right] \approx \frac{1}{K_1}\sum_k f\left(\left[\boldsymbol{x}_S, x_{\bar{S}}^k\right]\right)$ with $x_{\bar{S}}^k$ sampled from $p\left(X_{\bar{S}}|X_S = \boldsymbol{x}_S\right)$;
3. compute the interventional SHAP parts:
   (a) sample K_2 feature permutations;
   (b) for each permutation R sample single data point $x_{\bar{S}^R}^1$ from $p\left(X_{\bar{S}^R}|X_{S^R} = \boldsymbol{x}_{S^R}\right)$;
   (c) $\phi_{i,int} \approx$ $\frac{1}{K_2}\sum_R f\left(\left[\boldsymbol{x}_{S^R}, \boldsymbol{x}_i, x_{\bar{S}_i^R}^1\right]\right) - f\left(\left[\boldsymbol{x}_{S^R}, x_{\bar{S}^R}^1\right]\right)$;
4. compute the dependent SHAP parts by subtracting the interventional parts from the SHAP values.

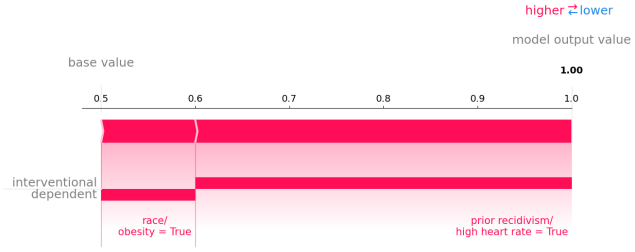


Figure 2: Force plot on sample from the introduction dataset. As can be seen, race or obesity has no direct interventional effect on the model output. They do have an indirect dependent effect via their dependent features, prior recidivism and high heart rate respectively.

Practically, K_1 and K_2 are increased until sufficient convergence is achieved. Due to the decreased sample efficiency [14], we advise K_2 to be taken bigger than K_1.

The above procedure can be applied for any conditional data distribution which can be sampled. In our experiments we used a multivariate Gaussian distribution and a Gaussian copula (as proposed by [1]). The complete algorithm in the case of a multivariate Gaussian data distribution is available in the Appendix. Like the original Kernel SHAP, our implementation is model agnostic and thus can be used with any model f including neural networks and support vector machines.

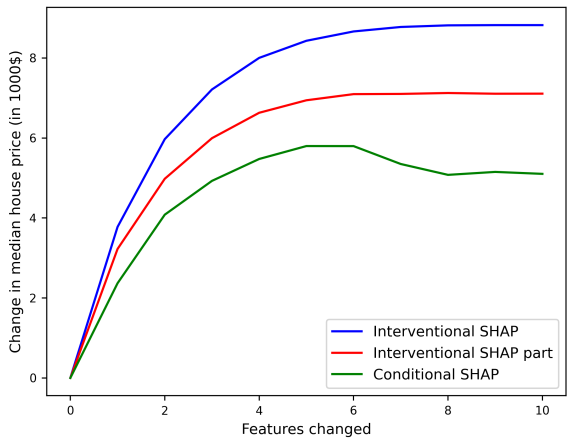
5. Experiments

Returning to the toy example in the introduction, with X_1 and X_2 both Bernoulli distributed with $p = 0.5$ and $P(X_1 = X_2) = 0.7$, we can exactly compute the SHAP parts. For the sample output $f_r(X_1 = 1, X_2 = 1) = 1$ they are shown as a force plot [15] in Figure 2. We see that the convict's race has no interventional effect. Whether the patient is obese does have an output effect via its dependent feature 'high heart rate' which is clearly shown in the explanation.

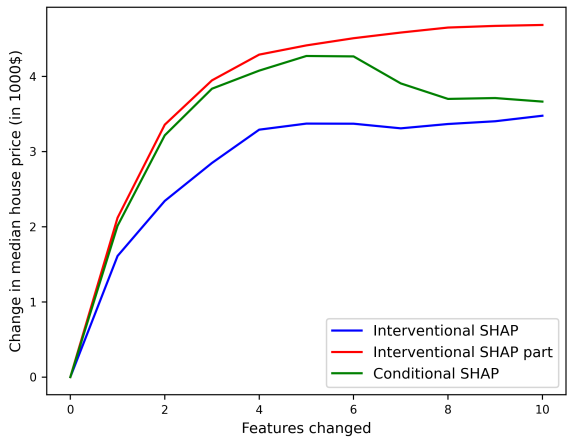
We continue this section by first comparing the interventional and dependent SHAP part to the state-of-the-art, whereafter the combined explanation is examined.

5.1. Interventional SHAP part

To underline the model correspondence of the interventional SHAP part, we conduct a first experiment. [4] trains a logistic regression model on a large dataset of loans and examines which explanation, interventional or conditional SHAP values, helped a loan applicant in decreasing their risk of default. In case of the Boston Housing dataset [7], we can consider a mayor who likes to increase the median house price in his town. To this extent, the town mayor can set the features with the most negative SHAP values to the mean feature value and check the change in median



(a)



(b)

Figure 3: Change in house price when imputing selected features with respectively the regular mean (a) and the mean conditioned on the unchanged features (b). A linear model is used to predict the house price. The values are computed for 200 random towns and averaged.

house price. In this experiment, the explanation is better if the median house price increases faster. Note that it is implicitly assumed that the features can be changed (unlike race for example). This is also assumed by [4].

In Figure 3a we show the change in house price when the features are selected by consulting, respectively, the interventional SHAP value, the interventional SHAP parts, the conditional SHAP values. Motivated by the theoretical observations in Section 4.2, a linear model (ordinary least squares) is used. Note that the effective model of [4] is also linear since the SHAP values are computed on the log odds. The dataset is assumed to be multivariate Gaussian.

It is also interesting to consider imputing with the mean *conditioned* on the (still) unchanged features. This

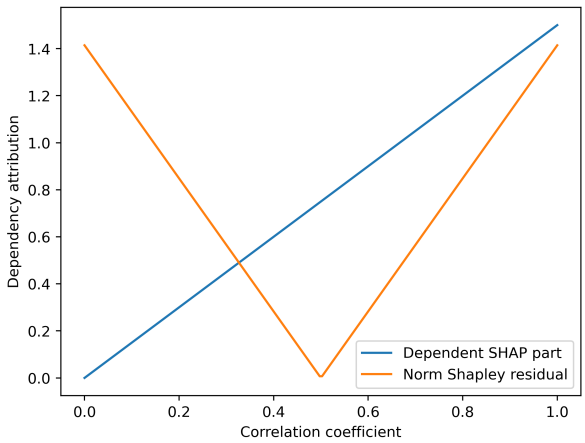


Figure 4: The dependency attribution of our method and the state-of-the-art [10] as a function of the correlation coefficient between two Gaussian variables in an additive model with feature interaction (Equation (27) with $a_{12} = 2$).

can be considered a more reasonable way to increase the median house price: how much we can realistically change a feature (e.g. decrease the crime rate) depends on many other features (e.g. the quality of the education in the town measured by the pupil-teacher ratio *PTRATIO*). The results are shown in Figure 3b.

The experiments are also repeated for a non-linear model (a random forest) and are shown in the Appendix.

5.2. Dependent SHAP part

Regarding attributions via dependent features we can compare our approach with Shapley residuals [10]. Consider the following two-dimensional additive model with feature interaction and data distribution:

$$\begin{aligned} f([X_1, X_2]) &= X_1 + X_2 + a_{12} X_1 X_2, \\ [X_1, X_2] &\sim \mathcal{N}\left([0,0], \begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix}\right). \end{aligned} \tag{27}$$

For a sample $[X_1 = 1, X_2 = 1]$ the conditional SHAP value and dependent SHAP part for $i \in \{1, 2\}$ are :

$$\begin{aligned} \phi_i &= 1 + 0.5 a_{12}(1 - \alpha), \\ \phi_{i,dep} &= 0.5(1 + a_{12})\alpha. \end{aligned} \tag{28}$$

The Shapley residual vector [10] and its norm are:

$$\begin{aligned} r_i &= \begin{bmatrix} 0.5a_{12} - (1 + 0.5a_{12})\alpha \\ -0.5a_{12} + (1 + 0.5a_{12})\alpha \end{bmatrix} \\ ||r_i|| &= \sqrt{2}|0.5a_{12} - (1 + 0.5a_{12})\alpha|. \end{aligned} \tag{29}$$

The relationship of both dependent attributions with respect to the correlation α is shown in Figure 4.

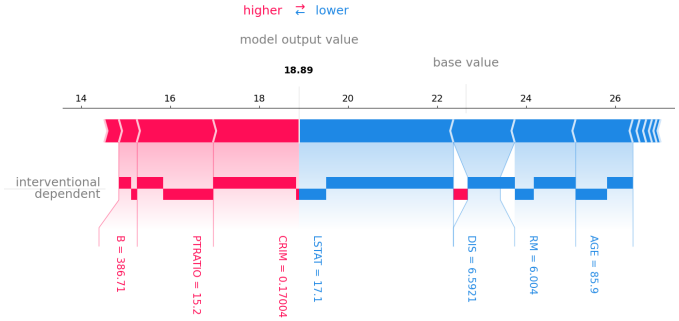


Figure 5: Force plot on sample from Boston Housing dataset

5.3. *Combined explanation*

We apply our method again to the Boston Housing dataset [7]. This time the model used is a high accuracy random forest and the dataset is assumed to be (multivariate) Gaussian distributed. Figure 5 depicts one complete explanation, showing a series of interesting local interventional *and* dependent effects. For example, the crime rate per capita ($CRIM$) being 0.17004 has a big positive effect on the median house price in this town, and the effect is largely interventional. This means that the $CRIM$ of this town has a strong effect on the output via its specific weights in the model. The pupil-teacher ratio $PTRATIO$ also has positive effect. A significant part of this effect is through its dependent features: it influences the distributions of the dependent features which in term increase the expected model output via their specific weights. We can also see that the model has a small racial bias: the interventional part of B is non-zero. Note that B is equal to $1000(b - 0.63)^2$ with b the black proportion of population [7].

Finally to show that our method generalizes to different datasets and their assumed distribution we apply the method to the Algerian Forest Fires dataset [3] and use a Gaussian copula to model the distribution [1]. We use a random forest to predict the probability of a forest fire based on temperature (T), relative humidity (RH), wind speed (Ws) and $Rain$. A complete explanation can be found in Figure 6a with a corresponding classic (interventional) SHAP explanation [15] in Figure 6b. Additionally, in the interest of studying the interventional and dependent connections between features and output, a partial correlation graph was constructed. This allows us to measure the degree of association between two variables, after removing the effect of all other controlling variables. Figure 7 shows the partial correlation graph for the Algerian Forest Fires dataset, where f_F is the log odds output of the trained model. This graph can be compared with the (normal) correlation between the features and their respective interventional and dependent SHAP parts in Table 2. All reported values are Spearman's rank correlations.

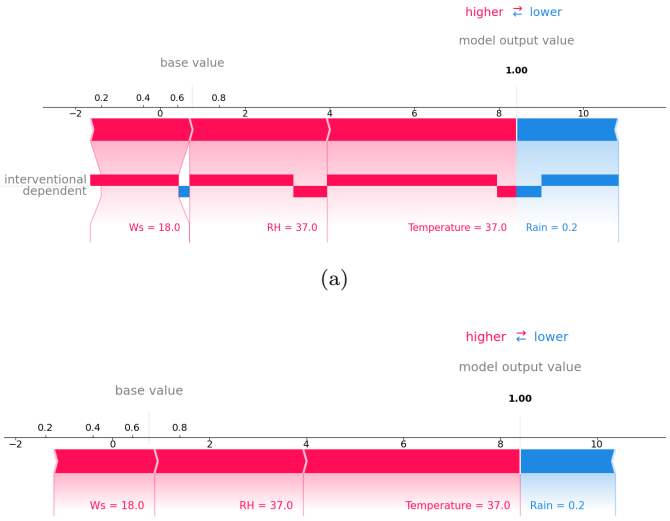


Figure 6: Force plot on a correctly classified sample of the Algerian Forest Fires dataset with our approach (a) and the approach of [15] (b) respectively. The top axis denotes the probability (for interpretability) and the one below denotes the log odds (on which the SHAP values and parts are actually computed).

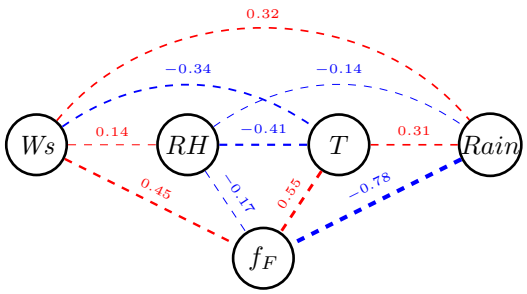


Figure 7: Partial correlation graph between features and output of the model f_F. Spearman's partial correlations are shown on the edges. Red and blue edges respectively show positive and negative correlation. The line width of the edge varies linear with the absolute value of the correlation.

Table 2: Correlations between feature values and their SHAP values

ϕ \ i	Ws	RH	T	$Rain$
$\phi_{i,int}(f_F)$	0.86	−0.69	0.79	−0.88
$\phi_{i,dep}(f_F)$	0.41	−0.29	−0.52	0.75

6. Discussion

The explanation for the toy example (Figure 2) is sufficient in both settings. In the court setting, the convict can clearly see the absence of racial bias. In the hospital setting, it is clear that the patient's obesity has an effect on the risk of a heart attack.

The other results should validate the interventional and dependent SHAP part and show their equivalence or superiority to the state-of-the-art. Secondly the combined explanation should explain both the model and the result, and should enhance interpretability for different settings and data distributions.

6.1. Interventional SHAP part

From Figure 3a we can see that on average selecting and mean imputing the features with the biggest negative interventional SHAP part is a better approach to increase the house price than using the conditional SHAP values. Using interventional SHAP values (not parts) still achieves bigger increases in house price, confirming the results of [4]. Both interventional approaches clearly outperform the conditional SHAP approach, confirming their model correspondence. The Appendix has additional experiments that show the significance of the results.

Figure 3a also shows that interventional SHAP values are more true to the model than interventional SHAP parts. This is as expected since for a linear model the exact interventional SHAP value is equal to the change in model output when imputing with the marginal mean (see (24)). If instead of the marginal, the conditional mean is used, interventional SHAP part comes out on top (Figure 3b). Again, because of (25), this is an expected result. Note that these observations are independent of the fitting procedure for the linear model: as long as the resulting model is linear, (24) and (25) will hold and the experiments should lead to the same conclusions. These results show that the interventional SHAP part is indeed a direct model (or interventional) effect and is more representative when the intervention is done with the arguably more intuitive conditional mean. Additionally it adheres to the dummy property (Proposition 2). Thus we can claim that our decomposition provides equivalent interventional insight as interventional SHAP [9].

For a non-linear model, (24) and (25) do not hold. Although the differences between the approaches are less significant in case of a random forest (see the Appendix), they are still in line with the results for a linear model.

6.2. Dependent SHAP part

Figure 4 shows that the dependent SHAP part increases linear with the correlation and is zero when the features are independent, as desired. The norm of the Shapley residual follows a less intuitive trajectory: it is not zero for independent features and does not monotonically increase with the correlation coefficient. Even more, it is zero at a certain non-zero correlation. Shapley residuals for conditional SHAP values also account for feature interactions in the model not only for interactions in the data, a observation not made in [10]. To effectively explain the model *and* the result, the dependent attribution should not attribute importance solely to feature interactions in the model (which Shapley residuals do for $\alpha = 0$). Therefore it is clear that our method is preferred. Furthermore, Shapley residuals can not be interpreted as additive parts of Shapley values: as claimed by Proposition 1 and confirmed by summing the components of r_i in Equation (29) this part would always be zero. This makes residuals unsuitable for intuitive representation through force plots, unlike our method.

6.3. Combined explanation

The explanation for the sample of the Algerian Forest Fire dataset (6a) allows to validate the model output: a lay person would agree with the large interventional effect of the temperature being 37 degrees Celsius. Furthermore, the combined explanation of interventional and dependent parts allow to motivate certain actions to change the expected result (fire or no fire). We can read that dropping water on this forest should directly decrease the probability of a fire (high interventional effects of relative humidity RH and $Rain$). A smaller (and possible delayed) but significant dependent effect is also expected. This information can not be distilled with the approach of [15] (Figure 6b).

The correlations between features and their SHAP values in Table 2 are supported by the partial correlations between features and output shown in Figure 7. Table 2 and Figure 7 both provide a limited view into the model-data structure: they are expected to show similarities but are not comparable in an absolute sense. The partial correlations between each input and the output f_F resemble the correlations between each input and its interventional SHAP part. Furthermore, the big positive correlation of $Rain$ with its dependent SHAP part (in contrast with its big negative correlation with its interventional SHAP part) can be attributed to positive partial correlations with other features (T, Ws) and their high positive partial correlation with the output f_F. These observations validate the decomposition and show how it can give more insight into the model-data structure than other SHAP implementations. Note that Shapley flow [24] would provide even more insight into the model and data dependencies, but their approach is less practical since it require a causal graph and the explanations can not be represented through the popular force plots [15].

In practice, our method can be used on any dataset (and model) regardless of its size and its number of features, and the computational cost can be controlled by the parameters of the method.

If no reasonable estimation of the data distribution can be made (either data driven or knowledge driven) and/or the interest is clearly in interventional output effects, interventional SHAP is the method of choice. If sufficient

causal knowledge is available it might be interesting to look into causal Shapley values [8] or Shapley flow [24]. In all other cases we advise to use our method, since there is only information gain with regards to (standard) conditional SHAP. An overview of the insights our approach provides compared to related approaches is given in Table 3.

7. Limitations

As is often the case with new explanation methods (as those proposed by [1], [22] and [9]), validation is mostly done through theory and (our own) intuition. Further work could consist of discussing the results and also the general method formulation with domain experts.

There are also some shortcomings of the method. First of all, the distinction between model and result might only be apparent to the machine learning researcher. The domain expert who is not familiar with machine learning or even statistics might not be interested in this distinction. Secondly, just like conditional SHAP and unlike interventional SHAP, our method requires conditionally sampled feature values. In our real data experiments we used a multivariate Gaussian distribution and a Gaussian copula. Other approaches are proposed by [1].

8. Conclusion

This paper started by motivating a new point of view on post-hoc explanations: an apparent distinction between model and result explanations. To remove this burden of choice, we derive a novel method, extending conditional SHAP, to combine and unify both explanations. Through theory and experiments we show that our method provides equivalent interventional insight as interventional SHAP while additionally explaining feature dependencies better that the existing state-of-the art. Finally, we contributed a novel Shapley value implementation and accompanying force plots suitable for a wide selection of settings, without requiring causal knowledge.

Code availability

Python code to calculate SHAP parts and corresponding plots is available on request.

Appendix A. Implementation with Gaussian data distribution

To compute the interventional and dependent SHAP parts, we need access to the conditional data distribution $p\left(X_{\bar{S}}|X_S = \boldsymbol{x}_S\right)$ with S a subset of the features. If we assume a multivariate Gaussian data distribution, the conditional distribution is also Gaussian with mean $\mu_{\bar{S}|S}$ and covariance matrix $\Sigma_{\bar{S}|S}$. Write

$$\Sigma = \begin{bmatrix} \Sigma_{SS} & \Sigma_{S\bar{S}} \\ \Sigma_{\bar{S}S} & \Sigma_{\bar{S}\bar{S}} \end{bmatrix} \text{ and } \mu = \begin{bmatrix} \mu_S \\ \mu_{\bar{S}} \end{bmatrix}, \tag{30}$$

then

$$\mu_{\bar{S}|S} = \mu_{\bar{S}} + \Sigma_{\bar{S}S}\Sigma_{SS}^{-1}\left(\boldsymbol{x}_S - \mu_S\right) \tag{31}$$

and

$$\Sigma_{\bar{S}|S} = \Sigma_{\bar{S}\bar{S}} - \Sigma_{\bar{S}S}\Sigma_{SS}^{-1}\Sigma_{S\bar{S}}. \tag{32}$$

Now the algorithm to compute the SHAP parts can be formulated:

1: **given** dataset $\mathbf{X}_{N\times M}$, sample to explain $\boldsymbol{x}$, model f, parameters K_1 and K_2
2: compute mean μ and covariance matrix Σ from $\mathbf{X}_{N\times M}$
3: compute the conditional SHAP values ϕ_i by applying Kernel SHAP, estimating every $E\left[f(X)|X_S = \boldsymbol{x}_S\right] \approx \frac{1}{K_1}\sum_k f\left(\left[\boldsymbol{x}_S, x_{\bar{S}}^k\right]\right)$ with $x_{\bar{S}}^k$ sampled from $\mathcal{N}(\mu_{\bar{S}|S}, \Sigma_{\bar{S}|S})$.
4: **for** $i = 0$ to M **do**
5: **for** $k = 0$ to K_2 **do**
6: $R \leftarrow$ Permutation$(\{1, 2, \dots M\})$
7: sample $x_{S^{\bar{R}}}^0$ from $\mathcal{N}(\mu_{S^{\bar{R}}|S^R}, \Sigma_{S^{\bar{R}}|S^R})$
8: **end for**
9: $\phi_{i,int} \leftarrow \frac{1}{K_2}\sum_R f\left(\left[\boldsymbol{x}_{S^R}, \boldsymbol{x}_i, x_{S_i^{\bar{R}}}^0\right]\right) - f\left(\left[\boldsymbol{x}_{S^R}, x_{S^{\bar{R}}}^0\right]\right)$
10: $\phi_{i,dep} \leftarrow \phi_i - \phi_{i,int}$
11: **end for**

A lengthy description of the Kernel SHAP algorithm is available in [1]. Lines 4 to 9 compute the interventional SHAP parts in the same way as the normal SHAP values were computed in [20].

Appendix B. Results non-linear model

Because of Equations (24) and (25), interpretation of the results with a linear model is more straightforward. For completeness' sake, we shown the results with a non-linear model here. The experiments of Figure 3 were repeated with a random forest in Figure 8. As discussed above, while the differences between the approaches are less significant, they are still in line with the results for a linear model.

Appendix C. Significance of results

Contrary to regular machine learning practice the averages in Figure 3 and Figure 8 are not accompanied with standard deviations. This was a deliberate choice. Since some towns are very difficult to improve while others can be improved quite easily, the standard deviations are very large and make the results seem insignificant (see Figure 9). A better metric to judge the significance is the per town difference of both interventional approaches with conditional SHAP depicted in Figure 10 (for a linear model) and Figure 11 (for a random forest). Here, in addition to

Table 3: Suitability of our decomposition and related Shapley implementations to simultaneously explain different important aspects of the model-data structure. The greyed-out methods require knowledge of the causal graph. In case of causal Shapley values the parentheses denote the lack of implementation. For Shapley residuals, it is clear from our experiment in Section 5.2 that they are not always suitable for explaining dependencies in the data.

	model	result (model+data)	interactions in model	dependencies in data
decomposition SHAP	✓	✓		✓
Shapley flow [24]	✓	✓		✓
causal Shapley values [8]	(✓)	✓		(✓)
Shapley residuals [10]			✓	(✓)
conditional SHAP [1]		✓		
interventional SHAP [9]	✓			

the mean we show the actual differences for some towns instead of a standard deviation because these values are far from Gaussian. We can see a significant difference between the approaches.

References

[1] Aas, K., Jullum, M., and Løland, A. (2021a). Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, 298:103502.

[2] Aas, K., Nagler, T., Jullum, M., and Løland, A. (2021b). Explaining predictive models using shapley values and non-parametric vine copulas. *Dependence Modeling*, 9(1):62–81.

[3] Abid, F. and Izeboudjen, N. (2019). Predicting forest fire in algeria using data mining techniques: Case study of the decision tree algorithm. In *International Conference on Advanced Intelligent Systems for Sustainable Development*, pages 363–370. Springer.

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

[5] Frye, C., de Mijolla, D., Begley, T., Cowton, L., Stanley, M., and Feige, I. (2020a). Shapley explainability on the data manifold. *arXiv preprint arXiv:2006.01272*.

[6] Frye, C., Rowat, C., and Feige, I. (2020b). Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in Neural Information Processing Systems*, 33:1229–1239.

[7] Harrison Jr, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1):81–102.

[8] Heskes, T., Sijben, E., Bucur, I. G., and Claassen, T. (2020). Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *Advances in Neural Information Processing Systems*, 33.

[9] Janzing, D., Minorics, L., and Blöbaum, P. (2020). Feature relevance quantification in explainable ai: A causal problem. In *International Conference on Artificial Intelligence and Statistics*, pages 2907–2916. PMLR.

[10] Kumar, I., Scheidegger, C., Venkatasubramanian, S., and Friedler, S. (2021). Shapley residuals: Quantifying the limits of the shapley value for explanations. *Advances in Neural Information Processing Systems*, 34:26598–26608.

[11] Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., and Friedler, S. (2020). Problems with shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning*, pages 5491–5500. PMLR.

[12] Lockwood, S. K., Nally, J. M., Ho, T., and Knutson, K. (2015). Racial disparities and similarities in post-release recidivism and employment among ex-prisoners with a different level of education. *Journal of Prison Education and Reentry*, 2(1):16–31.

[13] Lundberg, S. M., Erion, G. G., and Lee, S.-I. (2018a). Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*.

[14] Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774.

[15] Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K.-W., Newman, S.-F., Kim, J., et al. (2018b). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2(10):749.

[16] 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*, pages 1135–1144.

[17] Rossi, R. C., Vanderlei, L. C. M., Gonçalves, A. C. C. R., Vanderlei, F. M., Bernardo, A. F. B., Yamada, K. M. H., da Silva, N. T., and de Abreu, L. C. (2015). Impact of obesity on autonomic modulation, heart rate and blood pressure in obese young people. *Autonomic neuroscience*, 193:138–141.

[18] Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317.

[19] 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.

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

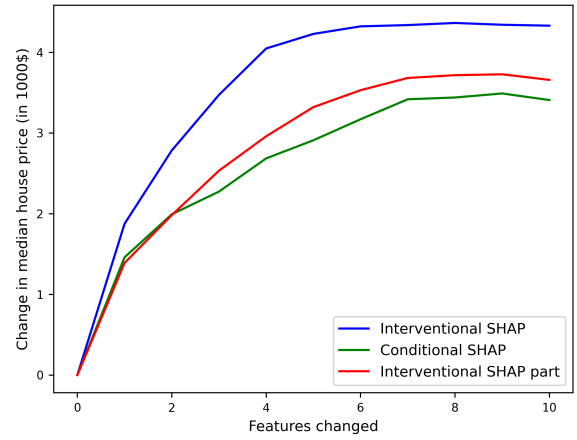
[21] Sundararajan, M., Dhamdhere, K., and Agarwal, A. (2020). The shapley taylor interaction index. In *International conference on machine learning*, pages 9259–9268. PMLR.

[22] Sundararajan, M. and Najmi, A. (2020). The many shapley values for model explanation. In *International Conference on Machine Learning*, pages 9269–9278. PMLR.

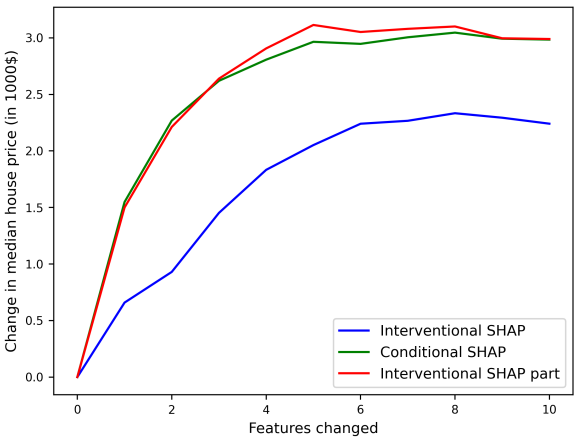
[23] Takeishi, N. (2019). Shapley values of reconstruction errors of pca for explaining anomaly detection. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 793–798. IEEE.

[24] Wang, J., Wiens, J., and Lundberg, S. (2021). Shapley flow: A graph-based approach to interpreting model predictions. In *International Conference on Artificial Intelligence and Statistics*, pages 721–729. PMLR.

[25] Yeh, C.-K., Lee, K.-Y., Liu, F., and Ravikumar, P. (2022). Threading the needle of on and off-manifold value functions for shapley explanations. In *International Conference on Artificial Intelligence and Statistics*, pages 1485–1502. PMLR.



(a)



(b)

Figure 8: Change in house price when imputing selected features with respectively the regular mean (a) and the mean conditioned on the unchanged features (b). A random forest is used to predict the house price. The values are computed for 200 random towns and averaged.

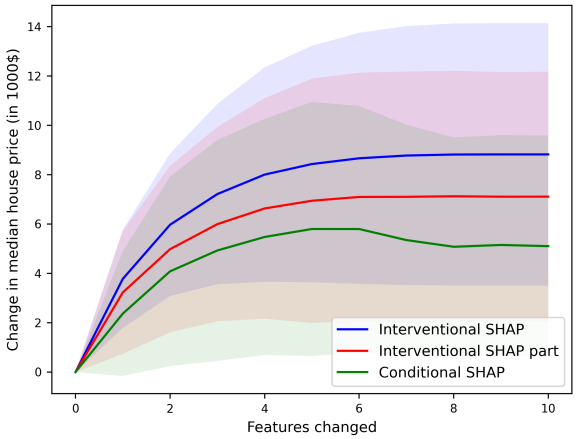
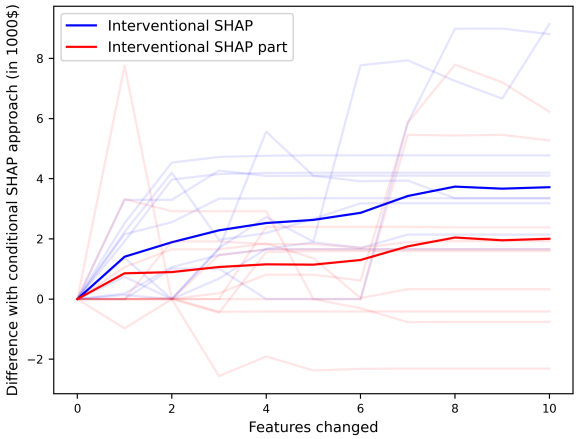
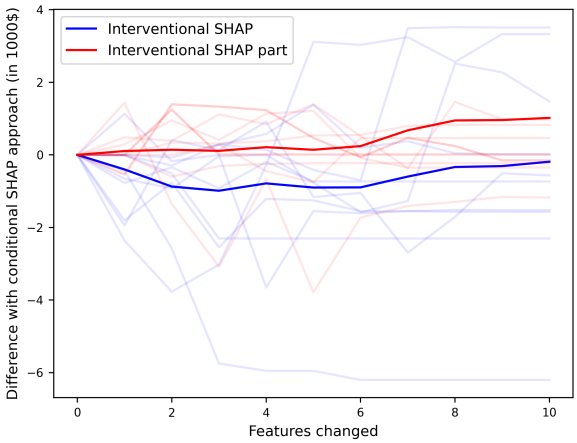


Figure 9: Change in house price when mean imputing selected features. A linear model is used to predict the house price. The values are computed for 200 random towns and averaged. The shaded region represents one standard deviation.

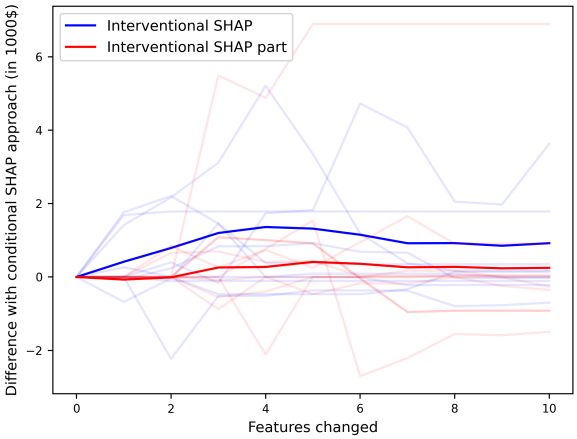


(a)

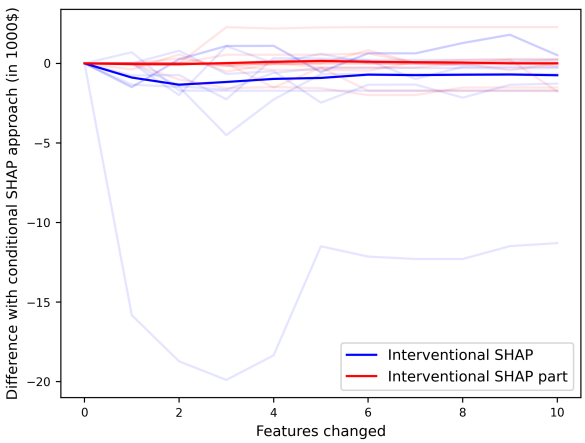


(b)

Figure 10: The per town difference of both interventional approaches with conditional SHAP when imputing selected features with respectively the regular mean (a) and the mean conditioned on the unchanged features (b). A linear model is used to predict the house price. The mean over 200 towns is a solid line. For ten randomly selected towns, actual differences are plotted (see-through lines).



(a)



(b)

Figure 11: The per town difference of both interventional approaches with conditional SHAP when imputing selected features with respectively the regular mean (a) and the mean conditioned on the unchanged features (b). A random forest is used to predict the house price. The mean over 200 towns is a solid line. For ten randomly selected towns, actual differences are plotted (see-through lines).